Final Project Report - CMSE 410/890: Bioinformatics and Computational Biology
Exploring the SingleR pipeline for annotation of single cell RNA sequencing data
Gary Zhang
Dr. Arjun Krishnan

i. Problem and Goals

Fibrosis is a tissue disorder that is characterized by pathology where normal tissue is replaced with connective tissue. This results in the formation of scar tissue formation which damages the function and architecture of the organ that it affects. Currently, diseases that cause fibrosis are of great importance to study since we currently have little approved treatments for many fibrosing disease. My project focuses on a paper titled "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage". This paper was published in 2019 in Nature Immunology and focuses on a exploring the different molecular mechanisms that exist in lung fibrosis and discovery of specific cell types that drive fibrosis. The authors do this by incorporating a single cell annotating program called SingleR (single cell recognition).

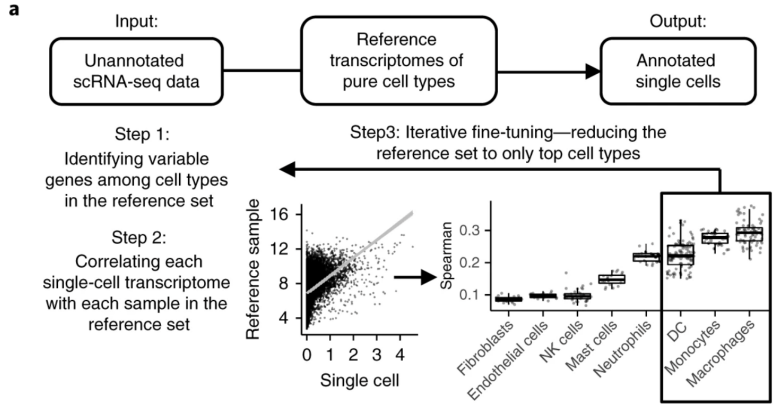


Diagram of the single cell computational pipeline Source: (Dvir Aran et al, 2019)

In the beginning, the major goal of my project was to replicate the annotation created by SingleR on datasets that were used in the paper. Since SingleR uses a reference database to annotate single cell RNA sequencing data, another goal of my project was to compare annotations from different reference databases.

My main conceptional/technical motivation for choosing this problem was the use of single cell RNA sequencing data. Traditional bulk sequencing produces a homogenous mixture of many different cell types and only serves to give a general expression level for a whole tissue. Single cell RNA sequencing differs because it allows for the specific expression level of each cell to be measured. By measuring the expression levels for each individual cell, they can be categorized into different cell types which can lead to interesting discoveries regarding disease pathology.

During the time course of this computational project, my scopes evolved quite frequently. At the beginning of the semester, I was mainly focused on downloading all the correct datasets and getting all the packages in order. I was also focused on learning more about single cell RNA sequencing and how to analyze different datasets with cell cluster tools such as Seurat and SingleR. Towards the middle of the

semester, I worked on recreating the plots shown in the paper using R and the raw data that was given. At the end of the semester, most of my effort has been placed on using the program SingleR from the paper and comparing how using different variables such as the referenced database can affect the results that it produces.

ii. Datasets

Most of the datasets include single cell RNA sequencing data. Initially I intended on using just the datasets that were used in the SingleR paper which mainly consisted of scRNAseq data from mouse bone marrow derived dendritic cells and fibroblasts. Later, I utilized a well-known single cell RNA sequencing package that contained a collection of public scRNAseq data sets that were in SingleCellExperiment objects. Because SingleR utilizes reference databases to annotate scRNAseq data, I also had to install the CellDex packages which is a collection of reference datasets that have labels for different specific cell types.

I was able to obtain all of the data that I wanted since they were all public datasets and available to install as a package or available through NCBI GEO(Gene Expression Omnibus).

I did have to change my data plan between the mid-term presentation and the end. Originally, I was focused on analyzing the mouse bone marrow scRNAseq data however, since the paper was published, the format of the dataset has become incompatible with many of the latest pipelines (especially Seurat). Although I could have tried to install an older version of the pipeline or even filter and edit the matrices of the original data, I felt that these would lead to more potential problems and bottleneck the progress of my project. What I did instead was switch to using public single cell RNA sequencing datasets that would allow me to test the capabilities of SingleR and its annotation accuracy. I used a Bioconductor package that contained a compilation of different types of PBMC data from 10x Genomics. This provided data in the form of "SingleCellExperiment" objects which allowed me to easily integrate the SingleR pipeline onto them.

```
class: SingleCellExperiment
dim: 33694 8381
metadata(0):
assays(1): counts
rownames(33694): ENSG00000243485 ENSG00000237613 ... ENSG00000277475 ENSG00000268674
rowData names(3): ENSEMBL_ID Symbol_TENx Symbol
colnames: NULL
colData names(11): Sample Barcode ... Individual Date_published
reducedDimNames(0):
altExpNames(0):
```

Description of a subset of Single Cell data from the PBMC data package

Here is an example of the PBMC data I used. They contain a SingleCellExperiment class consisting of the different genes as rownames.

iii. Computational Approach

My original plan was to evaluate the results that I get from the methods/software, I would compare what I produced to the figures produced by the paper. The original goal of this project is the replicate/verify the results in the paper and produce reproducible documentation of the process. Specifically, I wanted to

make sure that the graphs of clustered scRNA-seq data match the ones presented in Figure 1c of the paper.

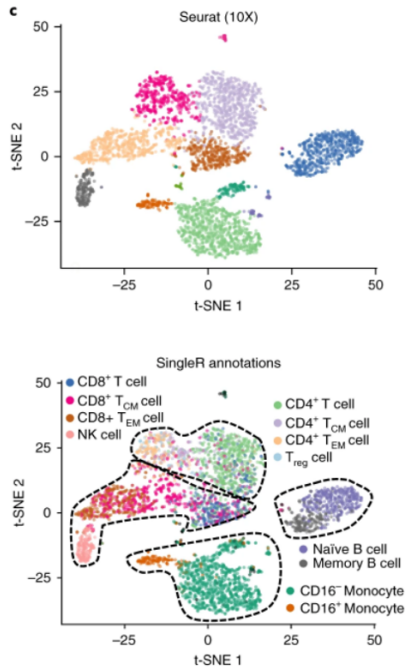


Figure 1c taken from the paper compares the annotation made by Seurat to the annotations provided by SingleR. SingleR is able to make more specific and accurate annotations of the different cellular population that exist. Citation (Dvir Aran et al / 2019)

Creation of this figure should prove that the alignment and correct dataset are being used. Comparison would also have been made with the annotations done by SingleR to verify that the methods of the paper were accurately reproduced. Finally, I would evaluate whether I am able to reach the same conclusion about profibrotic macrophage discovery that the paper reached using the data I have analyzed.
I originally proposed to download raw single-cell RNA sequencing data, align the data using Bowtie, run QC on the data, then cluster the data into cell types using Seurat and SingleR annotations. Here is a flowchart of what I proposed.

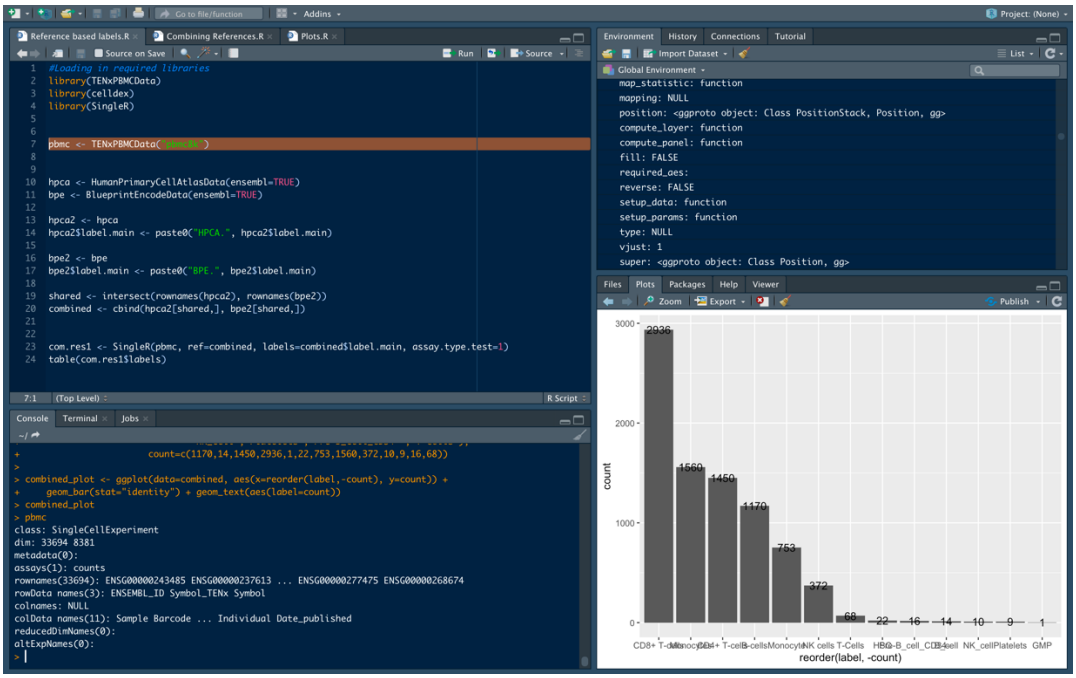
What I quickly discovered, is that running the pipeline was challenging for me to integrate. For Bowtie alignment and clustering/annotation, I would needed to use the HPCC since it is a computationally intensive process. Furthermore, the paper that demonstrates SingleR did not have much detail about how to use Bowtie and how they filtered and quality controlled the data beforehand. After I found that starting from raw data to a finished product was challenging, I then attempted to use preprocessed data found on the SingleR project github. When I tried to use Seurat to cluster the data, I found that the data format and

Seurat version that was used to create the paper was not compatible with the latest versions. This led me to attempt to download the correct Seurat version (version 2.2 instead of version 4) and recluster the data. After repeated attempts, I kept running into various errors that required editing the matrix of the single cell data. Eventually, I switched to using more updated scRNAseq data that was also organized into SingelCellExperiment classes to progress. I downloaded the SingleR pipeline from Bioconductor and used R to analyze data and make plots.

Throughout my project, I attempted to use various environments to do my data analysis. I originally started with just using terminal. What I found was that I was not yet familiar enough with using unix and the terminal environments to efficiently work on the project. I often found it hard to backtrack and navigate between different files due to unfamiliarity. Another preference issue was that I couldn't see directly what my plots look like and would always have to save a PDF version of a plot to view it. Although all these issues are easily solved by making myself more familiar with different unix commands and practice, I was often frustrated by having to google how to perform simple tasks.

As an alternative to using terminal, I tried to use Jupyter Notebook and Google Colab for my project. In previous CMSE classes, most of the work and projects were done on Jupyter Notebook so I was more familiar with the format. Notebooks also make it easier to trouble-shoot code and test different aspects of a code. Since most of my programming would be done using the R language, I had to attempt to integrate a usable R environment into Jupyter Notebook or Google Colab. These notebooks are mainly designed for python, and I found that using R was not efficient and produced many issues relating to version control (specifically these notebooks were not compatible by default to the latest versions of R)

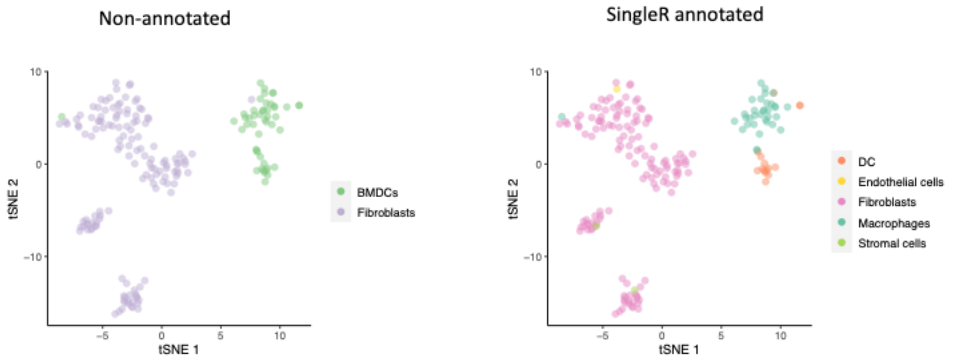
What I eventually settled on using was R studio. R-studio was easy for me to use since it was very beginner friendly and allowed me to efficiently navigate different aspects of the project. I found R-studio to be the most preferential tool for me since it allowed me to test scripts in the console while writing them. What this means is that I can easily test individual lines of code and troubleshoot/debug. Also, R-studio can show the different objects I have created and the plot in the same window. Here is an example of the workflow that I ended up using.



To conclude, I found R-Studio to be the easiest to work with given the project context (coding in R). I found it challenging to work with raw scRNAseq data and ended up using a package that contained organized scRNAseq data into SingleCellExperiment objects that were directly compatible with SingleR. This led me to shift in data source allowed me to implement the SingleR pipeline and perform analysis such as the ones found in the key findings section below.

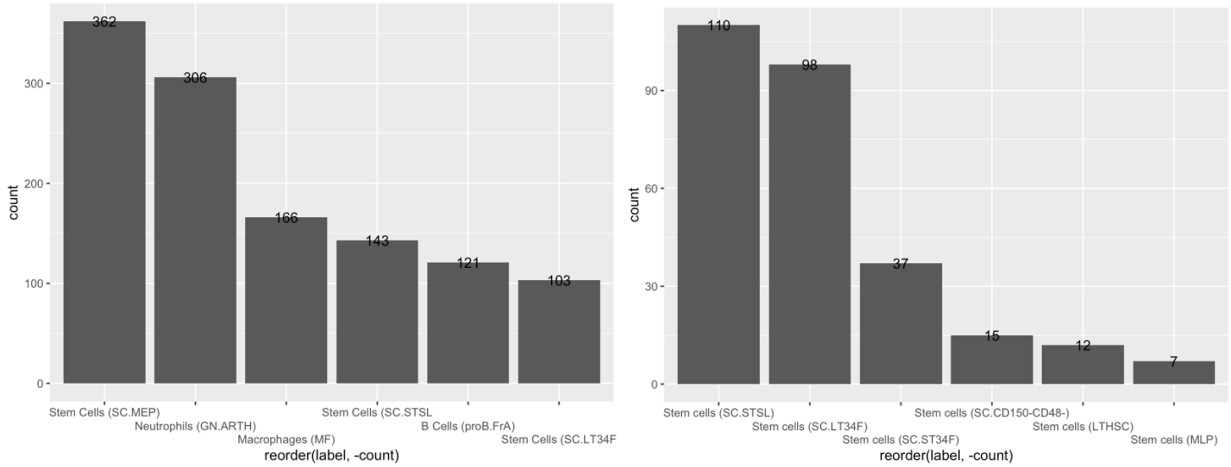
<u>iv. Summary of key findings and take homes</u>

Using the processed and preclustered data from the SingleR project GitHub, I was able to create figure 1b which consists of a t-SNE plot of the scRNAseq data from fibroblasts and BDMCs. As shown, without annotations, there are two cell types that are identified (BMDCs and Fibroblasts) after SingleR annotation is applied, there are many more cell types that can be identified which include DC, endothelial cells, Fibroblasts, Macrophages, and Stromal cells. What these figures are showing is the SingleR annotation can allow for more descriptive representation of the cell types that exist within the population. This also shows that the documentation of the project is reliable and reproducible.



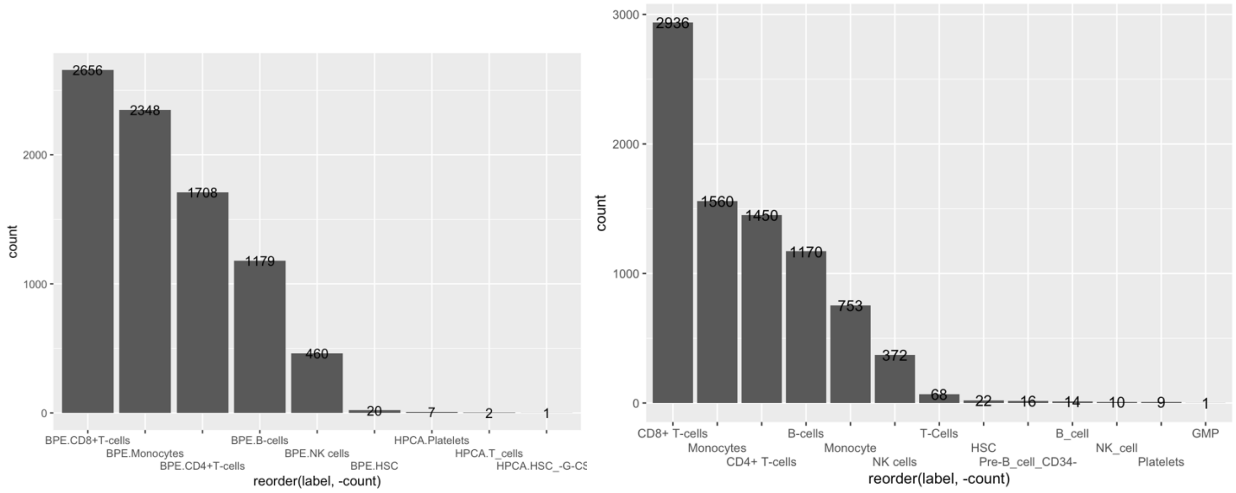
t-SNE plot of fibroblasts and BDMCs. The left plot shows non-annotated cluster while the plot shows clusters annotated by SingleR. Citation:Dvir Aran et al / 2019

To test the accuracy in using SingleR to annotate cell types and assign labels, I applied a dataset that contained sorted stem cells specifically HSCs. Since the data is sorted, this could test if SingleR can accurately label the one cell type that exists. Before any quality control is done to the dataset, there are multiple different immune cells detected such as neutrophils and macrophages. Although these cells may be related to the HSCs, ,the accuracy is not what we would expect since the data should have been sorted to only contain one cell type. After quality control and removing of a batch of low-quality data, SingleR can annotate the cell types more like we would expect (All cells identified are stem cells). What this suggests is that low quality reads or batches can greatly interfere with SingleR annotations which makes it important to perform quality control for accurate results.



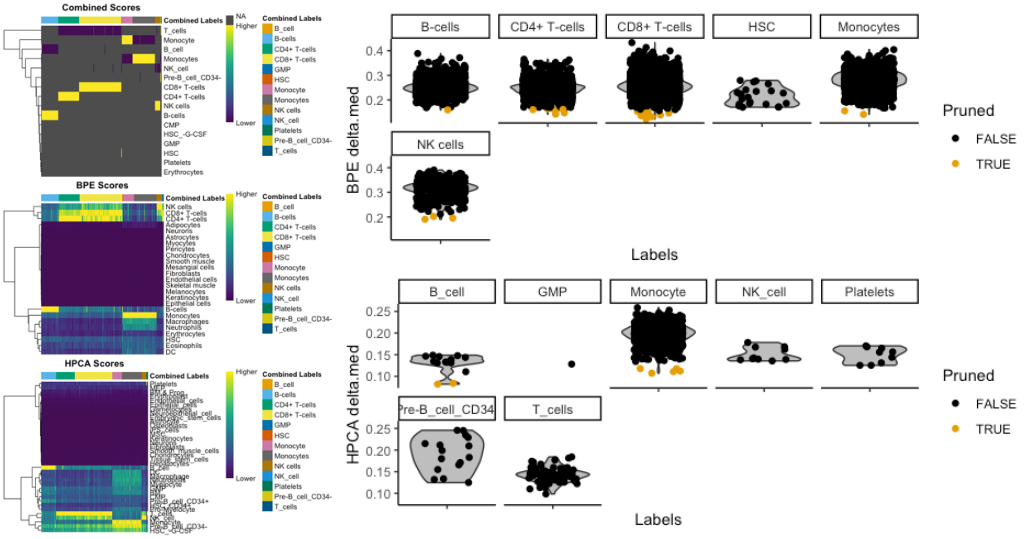
Graph on the left shows the counts for different cell types annotated before quality control. Graph on the right shows counts after quality control. Citation:SingleRBook

Since SingleR utilizes a reference database to annotate single cell RNA sequencing data, one of the major goals of the project was the compare different references for annotation of a dataset. Upon reading SingleR documentation, there were apparently multiple strategies to combine the inferences from multiple different reference datasets. The two main ways I focused on were using reference specific labels or combined labels. Reference specific labels would result in the dataset being annotated with labels from both references that were being used. This would create different labels for the same cell type based off which reference the label from. Combined labels would separately annotate the dataset based off a different reference and simply choose which label had the highest scores to create a combined representation of highest scores on one dataset. To test these methods, I implemented two references (BPE and HPCA) on a PBMC dataset from the PBMC 10X genomics package. What I found is that both methods produced similar results as shown by the figure. What this suggests is that both methods of using multiple references can be used however different situations or problems could possibly warrant the use of one strategy over the other.



The left plot shows the labels created using reference specific labels. PBMC dataset was annotated with BPE(BlueprintEncodeData), and HPCA(HumanPrimaryCellAtlasData). The right plot shows the labels created using the combined method which annotated the dataset with BPE and HPCA separately and then combined the highest scores. Citation:SingleRBook

To continue comparing the use of multiple references with SingleR, I used the diagnostic plot function that came with the SingleR package. I was able to create a heatmap of scores from from the different references and the scores of the combined results. What this shows is that there are different scores assigned to each label in the 2 different references used (BPE and HPCA). The heatmap of the combined results shows the scores that are computed for labels in individual references. The scores that are individual to each reference are the only scores computed by the combined score which is why they are depicted as NA on the heatmap. Using the delta distribution function in the SingleR package, I am also able to visualize the distribution of the change of scores across the dataset for the different cells that were assigned a label in the combined section. This shows the change in the individual references for the BPE and HPCA datasets. Based on the figure, what can be suggested is that most of the cell types do not have a combined score.



The figure on the left shows a heatmap of the scores assigned to each cell in the PBMC dataset from BPE and HPCA references. The figure on the right shows the data in scores for cells in the PBMC dataset after they are combined. Citation:SingleRBook

Code and Data Availability: https://github.com/garyzhang01/CMSE410.git

v. Challenges

In the beginning of the semester, my main goal was to compare clustering with Seurat and annotations made by SingleR which would lead to the reproduction of figure 1c from the SingleR paper. Specifically, my goals were to

-Download dataset from the paper
-Download all relevant software and set up
-Obtain access to the high-performance computing grid
-Align sc-RNA-seq data using BowTie software
-Produce clustered results using Seurat, produce graph and compare the results to the results shown in the figures of the paper
-Annotate results with SingleR and produced annotated graph of clustered results

Early on, I anticipated challenges regarding version control. These were the major technical challenges that I faced working on this project. I felt that I was not experienced enough with coding in R to address these issues. Some practical challenges I faced when working on this project were that it was difficult to find time every week to dedicate to working on the project. Also, inexperience in computational biology resulted in spending more time and becoming more frustrated about certain problems than need be.

vi. Reflection and future directions

If I were able to start this project I would focus more attention on learning how to be more efficient with using Unix and how to install and update different packages/version control. I think my biggest struggle with the project was that my technical abilities were not aligned with the problems that I wanted to answer and the questions that I had. For example, one of the biggest challenges with the project that I mentioned was that I had trouble with the different versions of Seurat and incompatibility with scRNAseq data. If I had more technical experience/knowledge, I believe I would be able to solve this issue and even possibly edit the original raw dataset matrices to be compatible with the latest versions of Seurat and SingleR. For my problem and approach, I would have been more active in asking for help and possibly contacting the authors of the SingleR paper for more direction and tips for using their pipeline to answer interesting questions. Finally, in terms of code, I would make sure to be more organized and more actively document the different troubleshooting that I did and record how I came to a solution.

For a future direction, I would be interested in applying the SingleR pipeline onto scRNAseq data of a specific disease progression event (other than fibrosis like described in the SingleR paper) to see if there are any interesting cell types or subtypes that could be identified.

vii. Acknowledgements

Thank you to Dr. Arjun Krishnan for being supportive and offering advice and feedback throughout the project timeline. I would also like to acknowledge the authors who developed the SingleRBook which provided very detailed documentation about different applications and ways to analyze SingleR annotations.

viii. References

Aran, D., Looney, A.P., Liu, L. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20,** 163–172 (2019). https://doi.org/10.1038/s41590-018-0276-y

Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
Seurat: https://satijalab.org/seurat/
SingleR: https://bioconductor.org/packages/release/bioc/html/SingleR.html

Glossary

SingleR: Annotation pipeline for single cell RNA sequencing data that annotates data based off a reference database
scRNAseq: Single cell RNA sequencing data which defers from bulk sequencing by providing the gene expression profile of each cell as opposed to an average expression profile.
Main Article referenced: https://www.nature.com/articles/s41590-018-0276-y#data-availability
NCBI source that contained many raw scRNAseq datasets:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111664
Github for SingleR project which helped a lot in understanding SingleR and utilizing functions for graphs: https://github.com/dviraran/SingleR
Bioconductor SingleRBook which helped me create a lot of the code and learn more about SingleR, especially how to incorporate the use of multiple references:
http://bioconductor.org/books/devel/SingleRBook/using-multiple-references.html
Resource for refreshing myself on how to graph in R: http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization
Bioconductor link with easy to use formatted scRNAseq data:
https://bioconductor.org/packages/release/data/experiment/html/TENxPBMCData.html