Rishabh Lenka

BIOE 190 Final Project

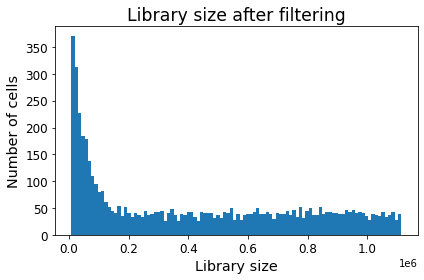
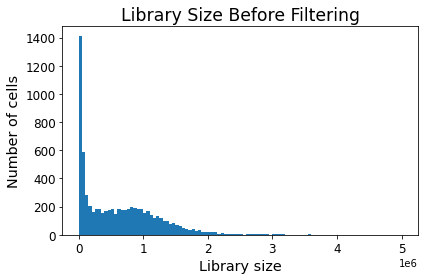
December 1, 2022

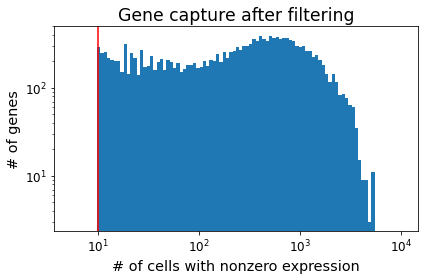
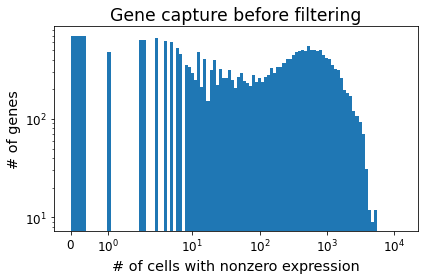
Analyzing the Differential Expression of Acta2 in the Heart and the Aorta

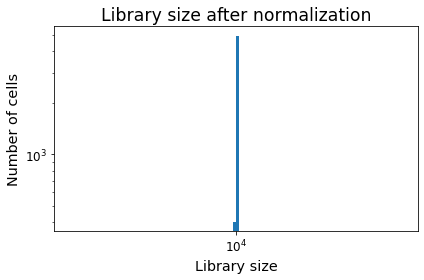
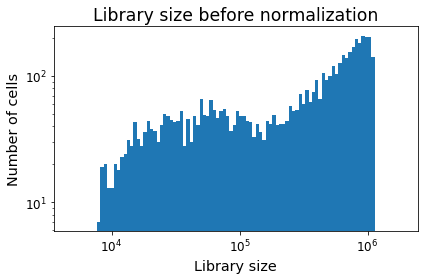
In this project, I downloaded datafiles from the Tabula Muris project and chose the Heart-counts and Aorta-counts datasets to analyze the populations of certain genes in both areas of the body. This data is useful for analyzing the relationships between organs, transcribing their genetic history and allowing medical professionals and biomedical scientists to triangulate and explore new parts of the genome.

The first step involved was to import numerous packages like sklearn, pandas, and phenograph, which allow researchers to use the powerful Python interface in gene analysis. Next, the datasets from heart and aorta were combined into a single matrix.

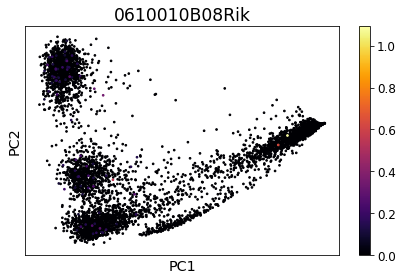
The second stage involved pre-processing. Here, I cut out cells that were either too large or too small and would skew the data. This was done by visualizing the library size distribution and choosing percentile cutoffs. Afterwards, genes that rarely appeared in the dataset were filtered out and then the entire dataset was normalized. This is done by dividing every cell by its library size and then rescaling all to 10,000. Finally, the data was log-transformed so that it became easier to interpret patterns in the data and reduce skew.



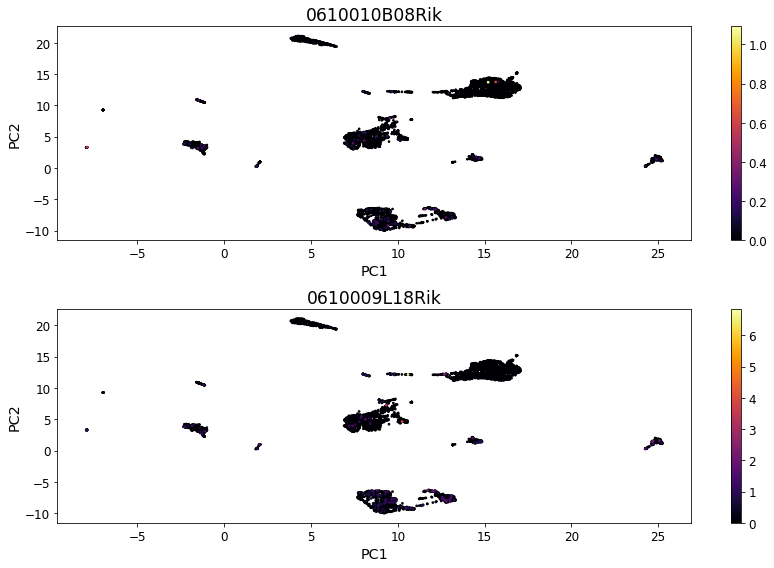




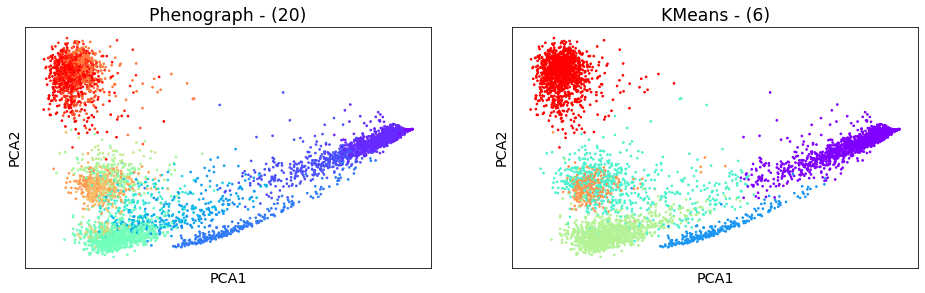
Next, I picked a gene to plot after running Principle Components Analysis (PCA), the most simplistic form of dimension reduction. Dimension reduction is a technique in which unnecessary or redundant variables with low statistical significance are eliminated, and only variables that highly impact the dataset remain. This is done to isolate the variables that cause the most change and to make it easier for computer processing and visualization.



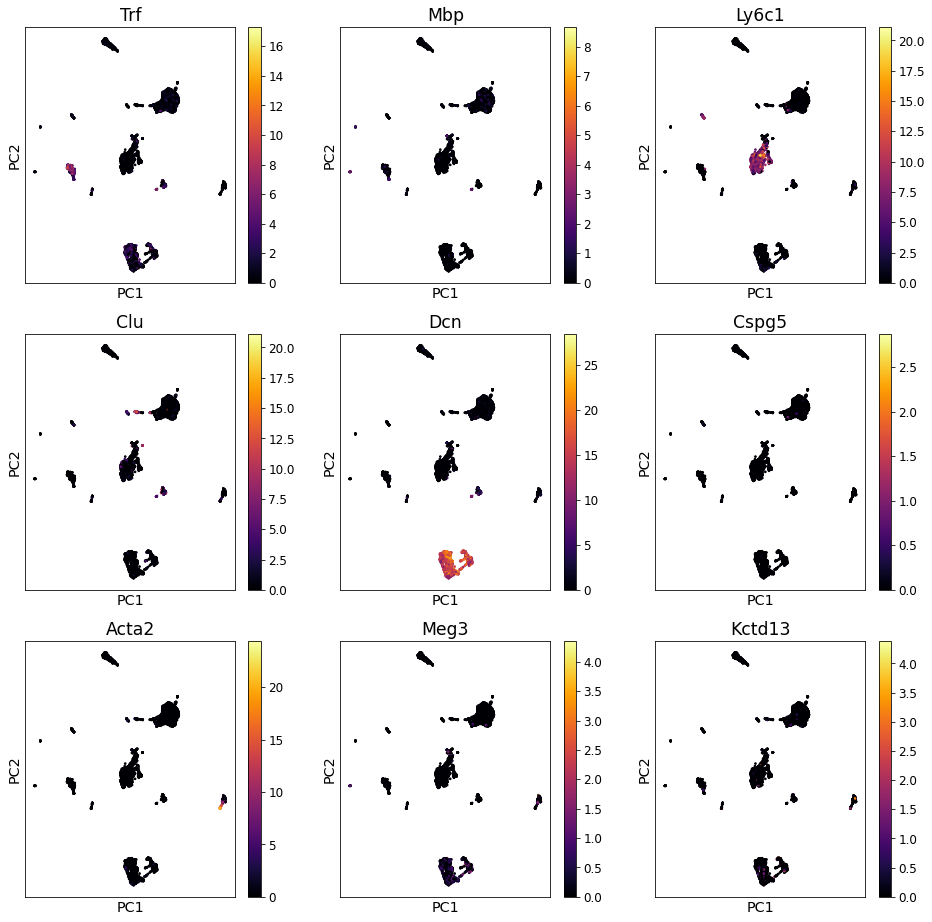
Then, dimension reduction was also done with UMAP in order to divide the data into separate clusters. The data was visualized so that PCA and UMAP reductions could be compared. Afterwards, two marker genes, 0610010B08Rik and 0610009L18Rik, were chosen to plot the UMAP coordinates.



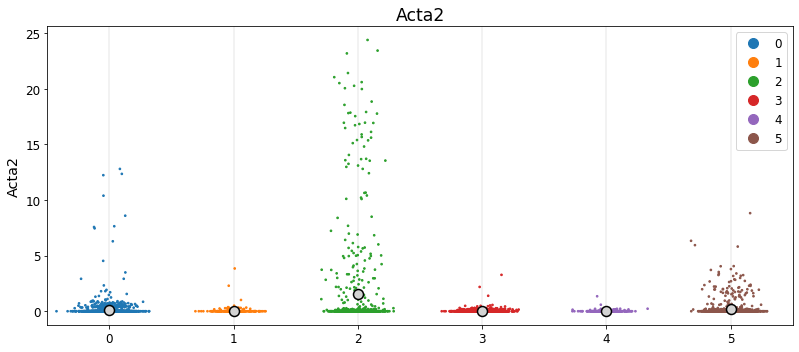
Next, I ran Phenograph and KMeans clustering algorithms on the data in order to identify the observations that were most similar to each other, and then mapped them by PCA coordinates so that the gene expressions would be tied to the clusters.



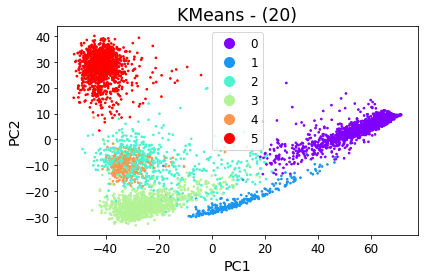
Finally, I chose 9 genes to plot using the UMAP data and received this visualization.

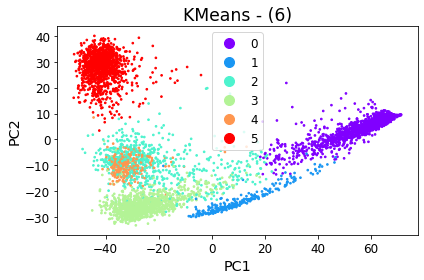


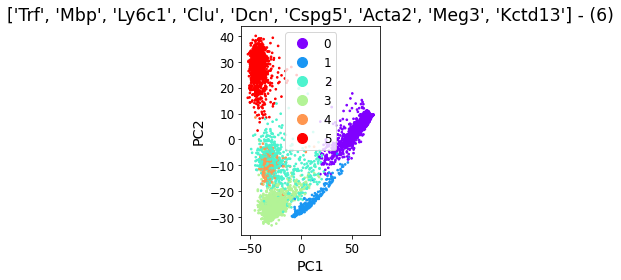
Next, a KMeans algorithm was used to create jitterplots of Acta2, Meg3 and Mbp genes in order to find out how often they were expressed in the genemaps. After this, I did analysis on Acta2 and found that it was a gene responsible for the development of smooth muscle in the body. Because my datasets were the heart and aorta, it made intuitive sense that Acta2 was found in high density, because these muscles automatically work to keep blood pumping inside the body at all times. I also found that it matched the FACS.selection.



Finally, I visualized KMeans graphs of the cell types and the clusters.







In the very last portion of the project, I ran a t-test in order to see if there were any significant changes in gene expression throughout the datasets. This test used a mean-difference formula with rank-sum statistics. What I found was that the DCN, Serping1 and Gsn genes had the highest difference in the abundance of gene transcripts. This is important data because it allows scientists and doctors to visualize the differences in the healthy and diseased states of specific organs, allowing for better treatment options and immediate recognition of problems.

The first major limitation of this project I found was that in order to have high quality results, you must first have high quality datasets from which to map the genes. Even small errors in the datasets can lead to errors in the later stages. Adding on to this concept, mistakes during data preprocessing can magnify into large errors because not only are we reducing the amount of data available to analyze, but it is also possible for human researchers to overcorrect and remove too much information during this stage. The larger the amount of data used, the more accurate the analysis will also be.

It is also possible that there are better clustering algorithms and dimensional reduction techniques that can be used in this study instead of PCA, UMAP, KMeans and phenographs. It would be best to rerun the project with multiple techniques to approximate the best results.

Overall, this project taught and hammered in the fundamentals of RNA sequencing, including but not limited to, Data Preprocessing, dimensional reduction, clustering, and differential expression. I was able to learn how to use these different techniques to modify, visualize and analyze RNA data.