Using Single Cell RNA Sequencing to Study Pulmonary Arterial Hypertension

PATH 828

Mack Sell

April 14th, 2022

Abstract:

A single cell RNA-seq dataset generated from rat models of pulmonary arterial hypertension was preprocessed and analyzed with the intention of identifying differentially expressed genes across different timepoints of the disease. In the process of accomplishing this, the dataset was visualized using a variety of plots, filtered for outliers, normalized, log transformed, and clustered using K-means clustering. Some of the clusters were then labelled based on the types of cells they contained. Supervised learning models were then compared and used for classifying cells from two of these clusters. Lastly, the Mann-Whitney U Test was applied to compare a natural killer cells across two timepoints in disease.

Introduction

Pulmonary Arterial Hypertension (PAH) is characterized by obstructive vascular remodelling caused by the proliferation of pulmonary endothelial and vascular smooth muscle cells. The result of this remodelling is a narrowing of the pulmonary arterioles, increased pulmonary vascular resistance, and eventual death by right heart failure. There is a strong link between immune dysfunction and PAH but how exactly altered cellular immunity leads to the pathological vascular remodelling is not well understood. The goal of this project is to develop a data analysis pipeline for single cell RNA-sequencing data so that mouse and rat models of PAH can be studied. This will help to identify the link between cellular immunity and pathological vascular remodelling in the lungs. Specifically, a focus will be placed on studying the differential gene expression in NK cells across multiple timepoints of disease. In the rest of this section, a brief overview will be provided of single cell RNA sequencing and the dataset being used for this project.

Single Cell Sequencing

In order to perform single cell RNA sequencing, the animal tissue of interest must be harvested, digested into single cells, barcoded, sequenced, then aligned to form a count matrix.

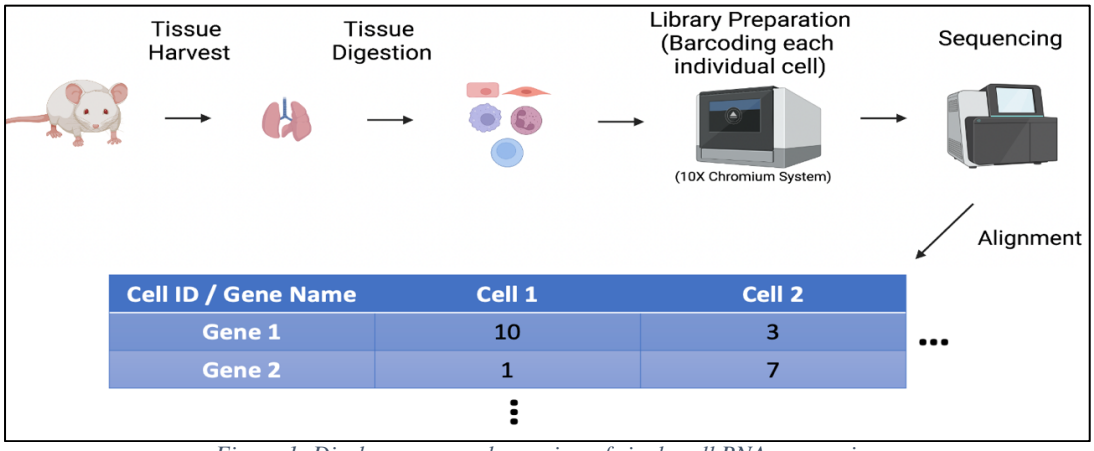

Figure 1: Displays a general overview of single cell RNA sequencing

This allows the gene expression of cells in the tissue to be analyzed. This is advantageous compared to bulk RNA-seq since it allows for the identification of individual cell types present in the tissue. However, what if it is necessary to compare animals across multiple conditions? Sequencing samples from different conditions separately using single cell RNA-seq and then combining the data afterwards will result in batch effects. The solution to this is sample multiplexing.

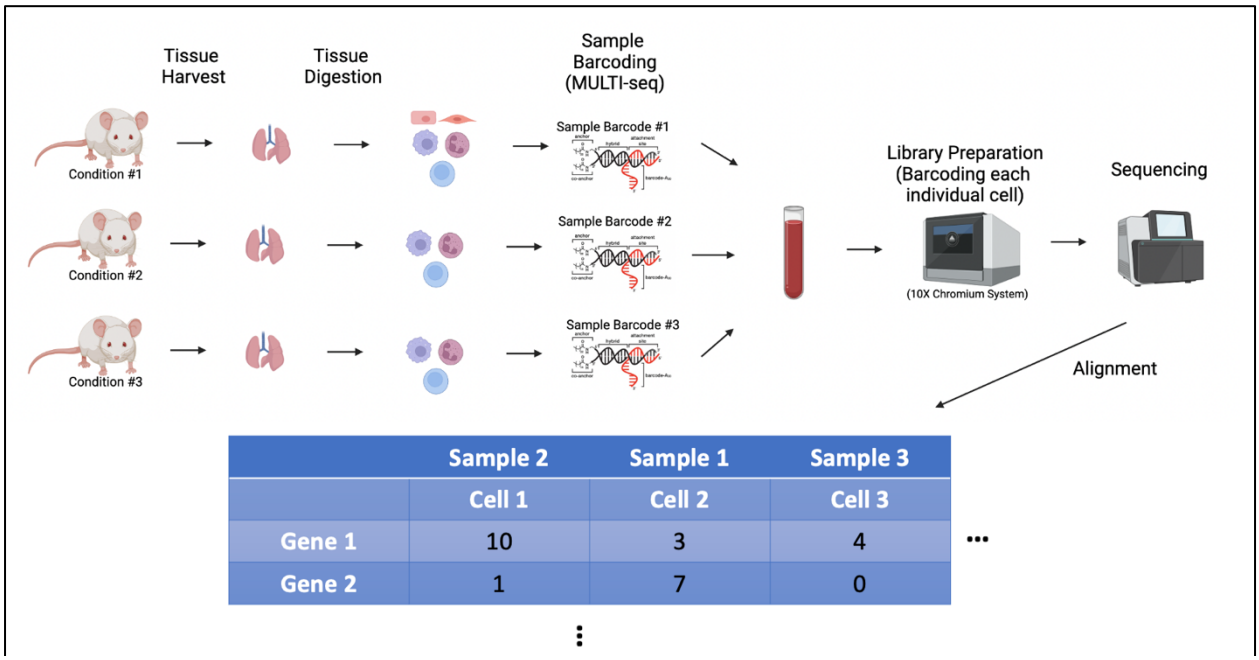

Figure 2: Displays a general overview of sample barcoding/multiplexing in combination with single cell RNA sequencing.

Sample multiplexing is when the cells from each sample are given a sample barcode before they receive a cell barcode. This allows the cells from multiple samples to be sequenced all at the same time without any risk of mixing up which cell is from which sample. Multiplexing is beneficial because it is more efficient and reduces the risk of batch effects in the data. Thus, sample barcoded, single cell RNA-seq data allows a scientist to know the gene expression for each cell and the original sample that the cell came from.

Dataset

The dataset for this project was provided by the Stewart Lab out of the Ottawa Hospital Research Institute. It consists of data derived from the lungs of 9 rats. These rats have received different levels of exposure to sugen and hypoxic environments, which are the disease-causing conditions. Out of the 9 animals, there was 1 animal under sugen and hypoxic conditions for 1 week, 3 weeks, 5 weeks, and 8 weeks, as well as 1 animal used as a control. In addition to this, 2 animals were exposed to only hypoxic conditions for 1 week and 2 animals were given only sugen for 1 week. For the purpose of this course project, I am only going to compare the samples from the control, the 1 week, and the 3 week animals. Beyond this course, I will perform these steps using the other timepoints as well.

The dataset for this project consisted of raw reads FASTQ files and the count matrix. The count matrix was generated using the Cell Ranger software that is part of the system used for single cell barcoding/library preparation called the 10X genomics system. The count matrix has individual genes as row labels and cells as column labels. This is sample barcoded, single cell RNA-seq data so each mRNA that is successfully processed through this system will have a unique molecular identifier (UMI) that is unique to it, a cell barcode that it shares with other mRNA from the same cell, and a sample barcode that it shares with all other mRNA from the same sample. Thus, each column in the count matrix displays the number of unique molecular identifiers (UMI) being expressed for each gene in a particular single cell.

The count matrix is not normalized or preprocessed. The sample that each cell belongs to has not been identified yet in the provided data. This has to be done using a package written in R that was made by the creators of the sample barcoding method called MULTI-seq. All the files in the

provided data came labelled as Pool A and Pool B files. Both Pool A and Pool B are from the same mixture of digested rat lung tissue that is composed of sample barcoded cells from all the samples. The mixture was divided into two separate pools (Pool A and B) due to the capacity limits of each well in the 10x Chromium System. This system is used for single cell barcode labelling prior to sequencing. As a result, Pool A and Pool B were processed in the exact same way and are from the same source but they were run-in different wells due to well capacity limits. The data from the two pools must be combined in R.

For this project, I am only going to focus on the work completed after the sample barcodes have been aligned/ the samples of each cell have been identified. This is because the MULTI-seq R package is the only way I currently know how to align the sample barcodes and this topic is somewhat outside the scope of the course project. Here is a general overview of the dataset composed of just the control, week 1, and week 3 samples:

Table 1: Displays general stats about the features and expression of cells in the dataset.

Total Cells	5050 Cells
Total Genes in Count Matrix	16381 genes
Average Total Number of Features Expressed:	962 genes per cell
Average Total Number of Transcript Counts:	2335 counts per cell

Methods

Sample Barcoding

As mentioned earlier, sample multiplexing was conducted using MULTI-seq. This method uses lipid-tagged oligonucleotides which adhere to the cell membrane of each cell. Part of their structure is a known sample barcode. Since it is known, it can be identified later in order to trace a transcript and cell back to its original sample.

Cell Barcoding

Single cell barcoding is accomplished using the 10x Genomics Chromium System. In this system, single cells eventually encounter and adhere to a gel bead. These bead-cell combinations are then trapped in individual oil droplets where they are referred to as Gel-Beads-In-Emulsions (GEMs). In each GEM is a single cell adhered to a gel bead and it is now isolated from other single cells. It is at this point that the cell is lysed. The beads are coated with barcoded primers that each have a unique molecular identifier (UMI) incorporated into it, which attach to all the free mRNA that was released from the cell's lysed membrane. The mRNA are then reverse transcribed into cDNA then sequenced. Thus, in addition to the cDNA sequence, there will be a unique molecular identifier and cell barcode that was also sequenced, thereby allowing unique mRNA transcripts to be identified along with what cell they came from.

Computational

The MULTI-seq method for sample multiplexing has a corresponding software package in R that was written specifically for it called deMULTIplex. This package aligns the sample barcodes associated with each cell to the known sample barcodes that are associated with a particular timepoint/sample. This project began in R because this was the only option available for

identifying which sample each cell came from. This is outside the scope of the course project, however, it is important to note that two forms of quality control occurred in R. The first is that cells expressing less than 200 counts or 200 features were filtered out. This is because cells expressing low counts or features are likely unhealthy and dying cells. Additionally, cells that were expressing high concentrations (25% and greater) levels of mitochondrial genes were filtered out. Again, this is because high mitochondrial gene expression is an indication that a cell is dying. All of the work in R was done using a package called Seurat that is specifically for working with single cell RNA-seq datasets. This package has special data structures that make it easier to manipulate the massive count matrices that are associated with this data. Once the samples that each cell was associated with were identified using the deMULTIplex package, the cell count matrix was moved over to MATLAB. This was a slow process due to navigating the Seurat data structures in R.

Once in MATLAB, the tools used were more reflective of the PATH 828 course. Scatter plots, box plots, and histograms were used to visualize the data, identify outliers, and filter them out. A t-SNE plot was used for dimensionality reduction in conjunction with K-means clusters. The K-value was selected using elbow and silhouette plots. Cell markers were then used to identify the clusters containing natural killer and endothelial cells. These cells were isolated and used in classification. The classification learner app in MATLAB was used to test a variety of different supervised learning methods. Lastly, SPSS was used to compare natural killer cells at the Week 1 and Week 3 timepoint. Specifically, the Mann-Whitney U test was used. Before it was used however, feature selection was performed using fscmrmr to try and reduce the number of features being compared between timepoints. Each of these steps and the reasoning behind them will be explained in more detail in the following section.

Results

Visualizations and Quality Control

Comparison of the total counts and total genes/features expressed in each cell is a good way of identifying outliers. The total counts should increase linearly with the total features being expressed. If a cell deviates from this pattern, then it indicates that there was likely a technical problem.

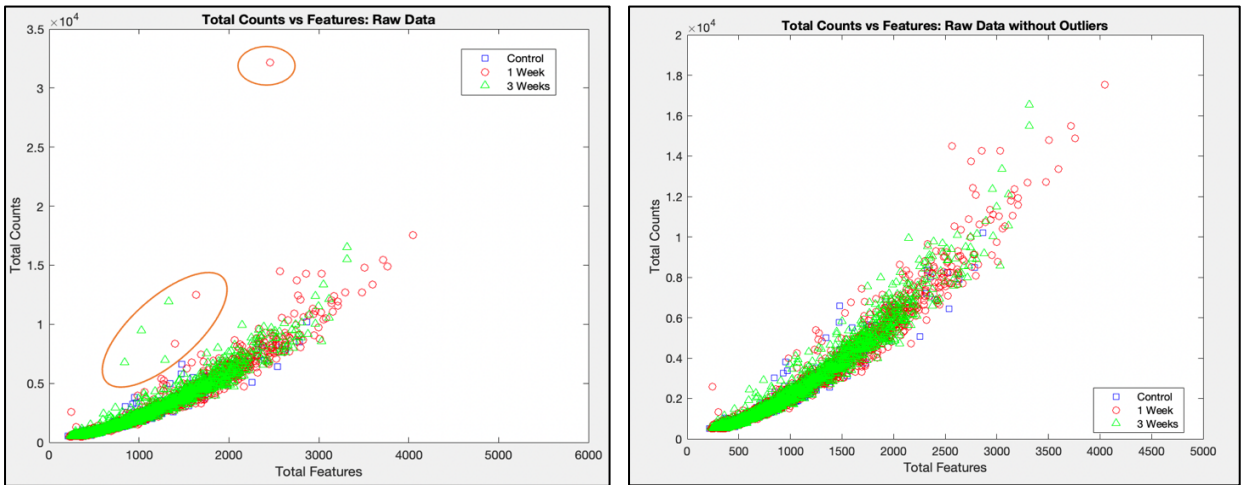

Figure 3: Displays the relationship between the total counts per cell and the total features per cell with labels indicating the samples/timepoints they are from. The points circled in orange in the left figure are suspected outliers. They are removed in the right figure.

The cells that are circled in orange are those that deviate from the expected linear pattern. High total counts and/or high total features can indicate that the cell is a doublet. This means that two cells were accidentally captured and barcoded together. So, a single barcode, which should only represent one cell, actually represents two cells and it shows twice the expression. Based on this, the cells circled in orange are likely doublets. The indices of the cells circled in orange were recorded so they could be investigated in more depth. It turned out that all of these cells were expressing a single gene in an abnormally high amount. For example, the cell that is circled in orange near the top of Figure 7 is expressing one gene 23,700 times when its total expression is 32,000 counts. This is a clear indicator of a technical problem because it is unusual for two thirds

of a cells gene expression to be from a single gene. Thus, the cells circled in orange were identified as outliers and removed. Next, scatter plots of only total counts and only total features were created to search for batch effects in the data.

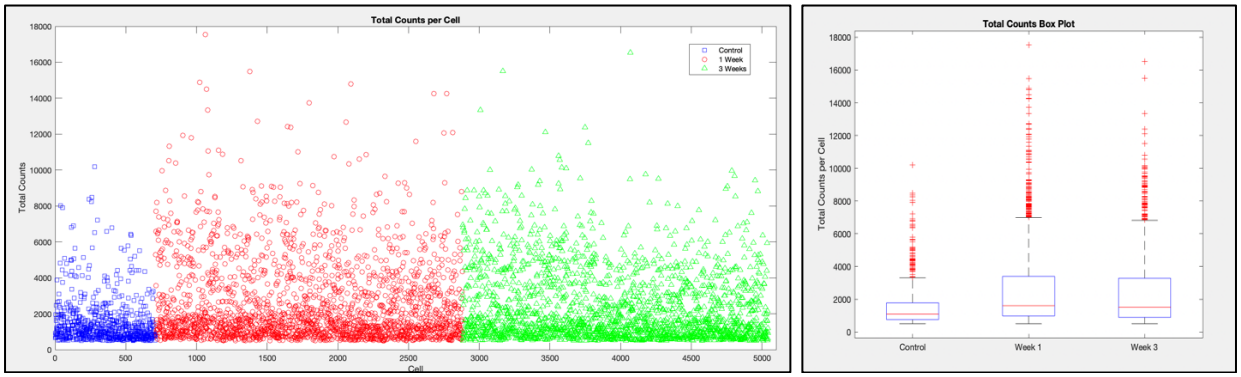

Figure 4: Displays the total counts for each cell with labels for each time point (left) and the boxplots of the total counts for each timepoint.

The scatter plot of total counts highlights the significantly lower number of cells in the control sample. This is technically a batch effect because there was likely less tissue used for this sample compared to the others. Having fewer cells is okay if the gene expression in those cells is still representative of the sample/timepoint. However, the control sample does appear to show differences in the range of total counts when compared to the other samples. The box plot displays that the interquartile range of the control sample's total counts is around half the interquartile range of the other two samples. This is further evidence that there may be batch effects occurring.

The cells near the top of the scatter plot that look like they could be outliers were investigated. Their indices were recorded, and expression assessed in a similar manner as earlier. Specifically, a ratio was calculated for many of these cells. This ratio was the number of counts of the highest expressed gene divided by the total counts for that cell. If this ratio is particularly high say 30% or more then there is likely a technical problem because it is abnormal for a cell to express so much of only a single gene. Out of the genes flagged as possible outliers, only one was identified as expressing an abnormal amount of a single gene and was removed.

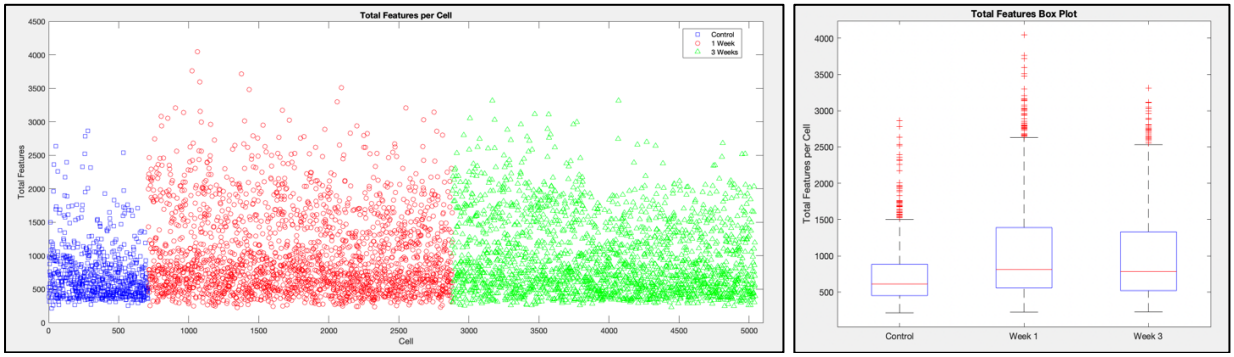

Figure 5: Displays the total features for each cell with labels for each time point (left) and the boxplots of the total counts for each timepoint.

The total features scatter and box plot provides a similar perspective as the previous plot. The range of total features expressed by the control is smaller than that of the other two features. Altogether, the control sample has fewer cells, are expressing smaller number of features, and are expressing a smaller amount of those features. Now the question becomes, are these differences between the control and other samples because of batch effects or because the other samples are in process of developing the disease? This digested lung tissue contains multiple cell types and not all cells should be affected by the development of disease so seeing such a large shift in features and total counts over all the cells in the sample is most likely due to batch effects. As a result, going forward only the samples from Week 1 and Week 3 will be used. This is unfortunate because the control sample is important to the experiment, but these do appear to be batch effects. Next, histograms will be made to visualize the number of counts typically found for a gene.

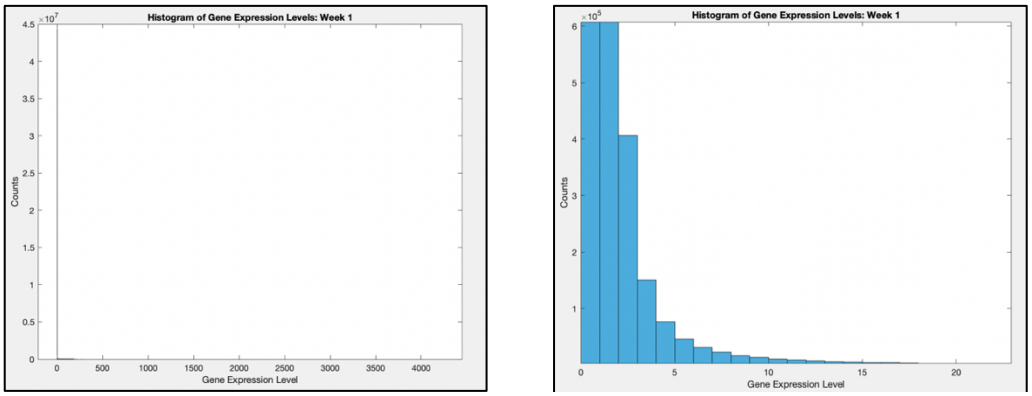

Figure 6: Displays the number of genes that are expressed at particular levels for the week 1 sample. The plot on the right is a zoomed in version of the plot on the left.

Only the histogram for the week 1 sample was included because the histogram for week 3 looks very similar. Single cell RNA-seq datasets have an enormous number of 0 counts. This is because there are over 16,000 genes being tested for and a single cell in this data set is on average expressing just under 1,000 genes. Based on the histogram, when a gene is being expressed it is most likely getting fewer that 5 counts. Next, a boxplot was made of all data.

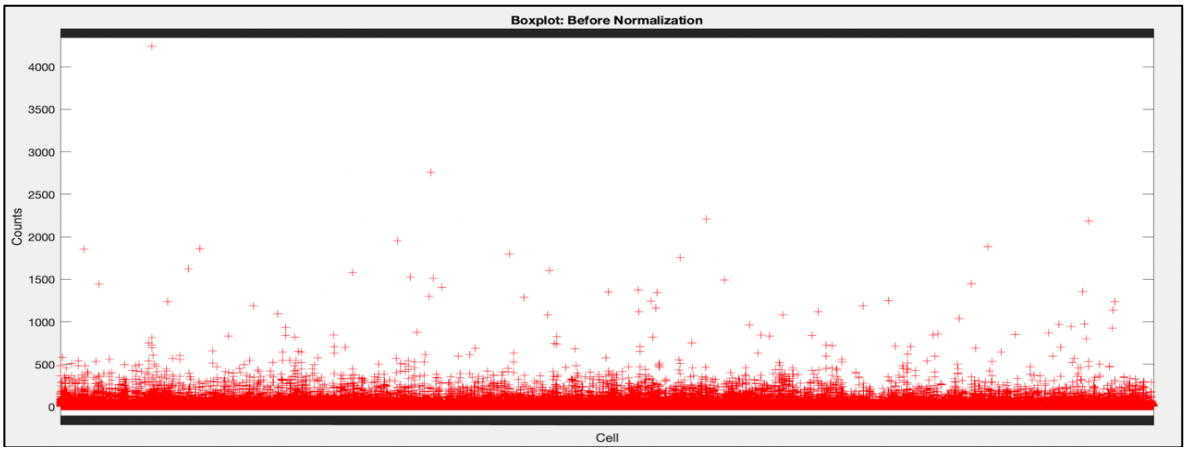

Figure 7: Displays boxplots for the entire dataset.

The data was then normalized for sequencing depth using relative frequency normalization. This was performed instead of TPM normalization because the 10x genomics library preparation/cell barcoding technique makes it unnecessary to normalize for gene-length [1]. 10x genomics advises not to normalize based on gene length because their method differs from traditional RNA-seq.

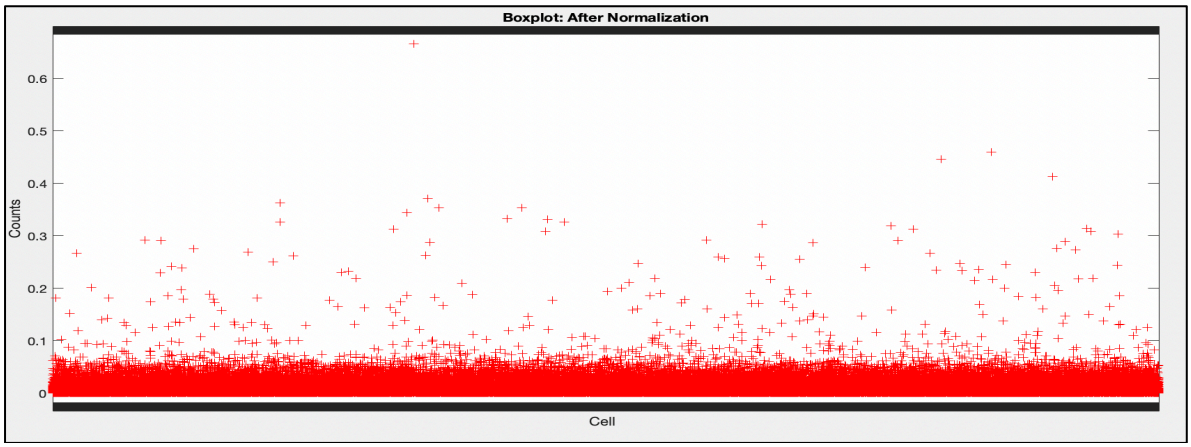

Figure 8: Boxplots of the entire normalized dataset.

The dataset was then log2 transformed:

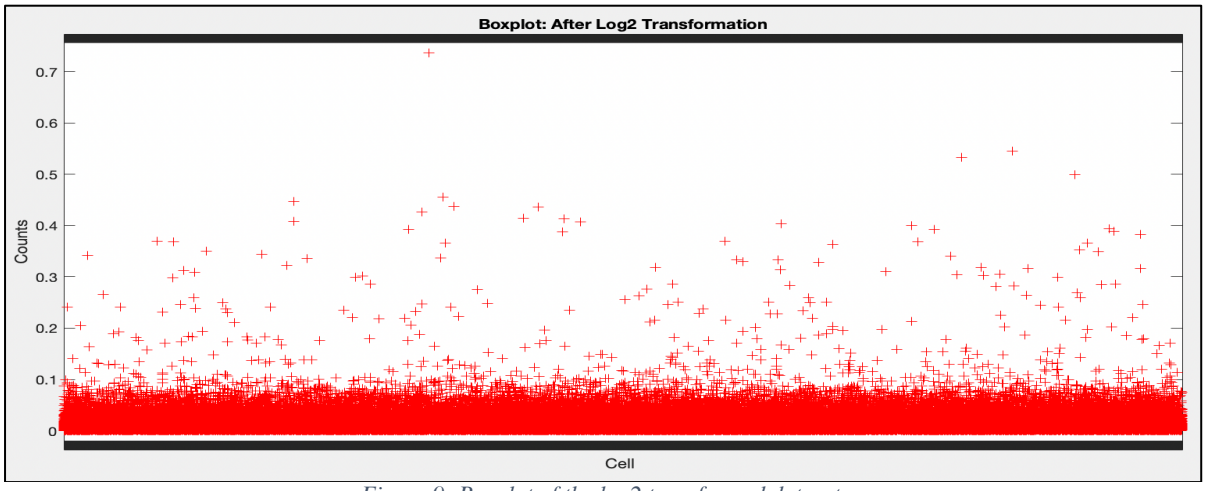

Figure 9: Boxplot of the log2 transformed dataset.

After these changes to the dataset the box of these box and whisker plots still remains at 0. This is understandable since on average only 1/16 of all the features are being expressed in a single cell with all the other features being 0. Filtering of the lowly expressed genes/features would eventually display more aligned boxplots. However, care must be taken when filtering because the absence of gene expression can be as important in determining cell type as the presence of expression of genes. As a result, around 6000 lowly expressed genes were filtered out.

Unsupervised Learning

First, cells that are expressing the gene called Ncr1 and the gene called Vwf were labelled in the data. Ncr1 is a cell marker for natural killer cells and Vwf is a cell marker for some types of endothelial cells. This will be used eventually for identifying the clusters formed from unsupervised learning. Next, t-SNE was used for reducing the dimensionality of the dataset. Without tuning any hyperparameters, the t-SNE plot looks okay but it can be improved.

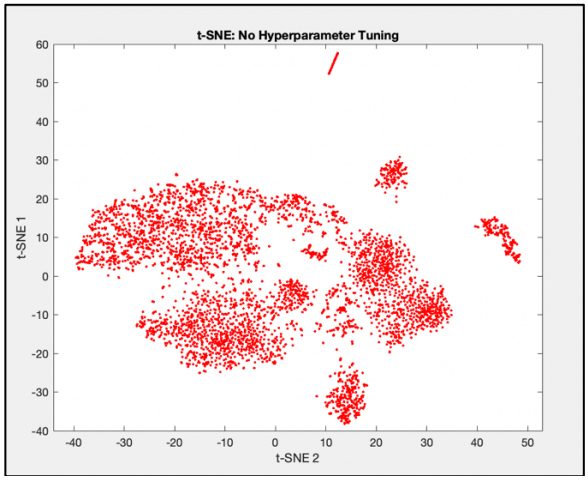

Figure 10: t-SNE plot without any hyperparameter tuning.

It is common for PCA to be used in single cell RNA-seq data sets as a first round of dimensionality reduction before t-SNE. The t-SNE function in MATLAB has an optional input argument that performs PCA and uses a specified number of the components for the t-SNE reduction. The following plot was made to help select the number of components to use.

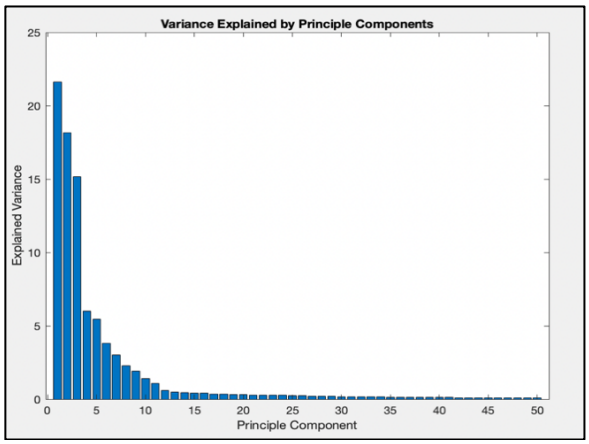

Figure 11: The percentage of variance explained by each principal component.

There is an elbow at the 11th principal component in the bar graph, thus, 11 was select as the number of PCA components used in the t-SNE dimensionality reduction. The other hyperparameter available for tuning is the perplexity. Generally, perplexity is supposed to increase with the size of the dataset. Trial and error was used to select for perplexity. A t-SNE plot was made for incrementally larger perplexities. The value selected was 43 as it yielded a plot that looks the most similar to those created with similar datasets in other papers [2].

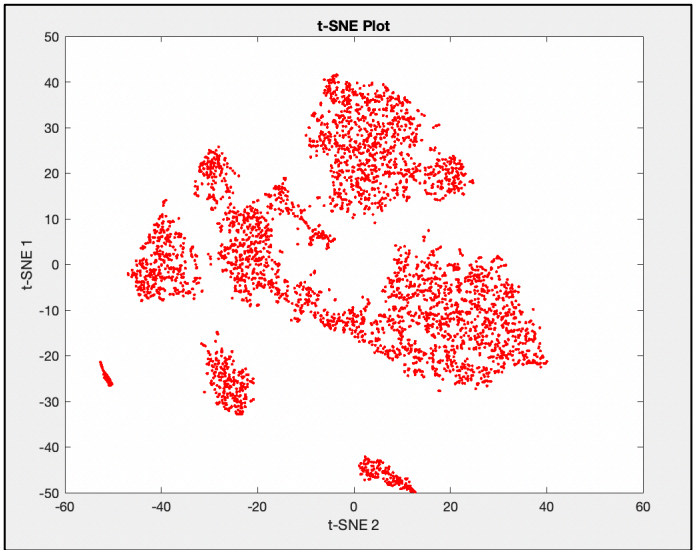

Figure 12: Displays the t-SNE after adjusting the PCA components and perplexity.

The method of unsupervised learning attempted was K-means clustering. The hyperparameters K and the distance measure used were changed to achieve the best clustering. Similar datasets have identified around 20 clusters in their clustering [2]. However, this is not an obvious choice based on the original data alone, so it is necessary to try a variety of K values and use silhouette plots and an elbow plot to determine the best one.

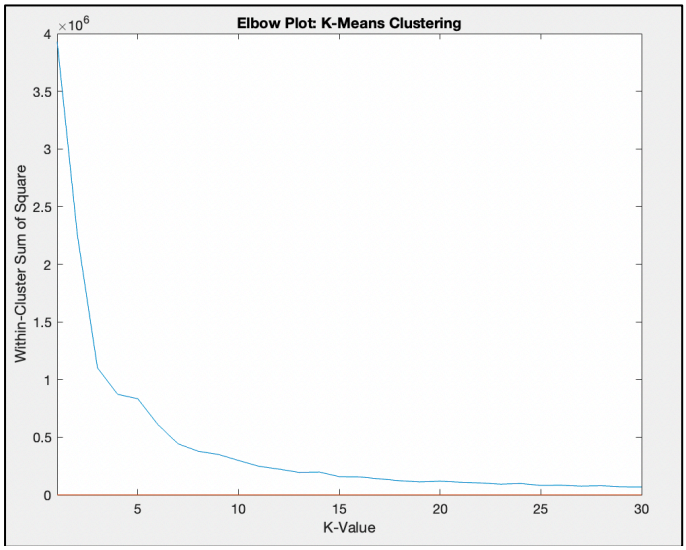

Figure 13: Displays an elbow plot for different k-values.

The y-axis of this plot is the “within-cluster sum of squares”, so a smaller value indicates tighter clusters. Based on the elbow plot, a K-value around 15 would be a good choice. Silhouette plots were made for K-values between 10 and 20 to aid in selecting a K-value.

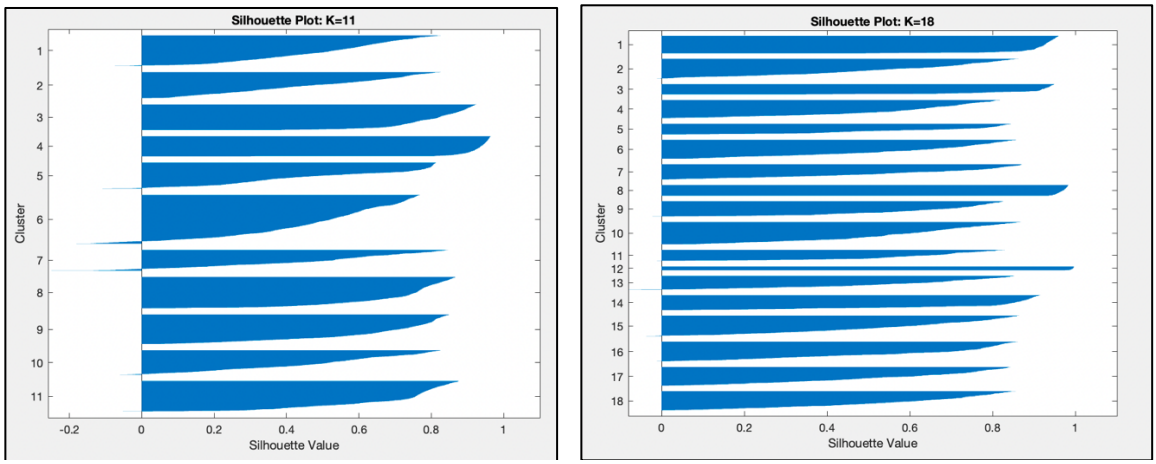

Figure 14: Displays the silhouette plots when K is 11 and 18.

The desired outcome is that the average distance of a point to other points in its own cluster is much greater than the average distance of that same point to points in other clusters. This ideal situation occurs as the silhouette coefficient approaches 1. The negative values indicate points that have not been assigned well to their current cluster. None of the silhouette plots had no negative values so it was a matter of selecting the plot that had the least of them. The best silhouette plot occurred when the K-value was 18 compared to the less desirable K = 11. This value can now be used to generate a t-SNE plot with the labelled 18 clusters.

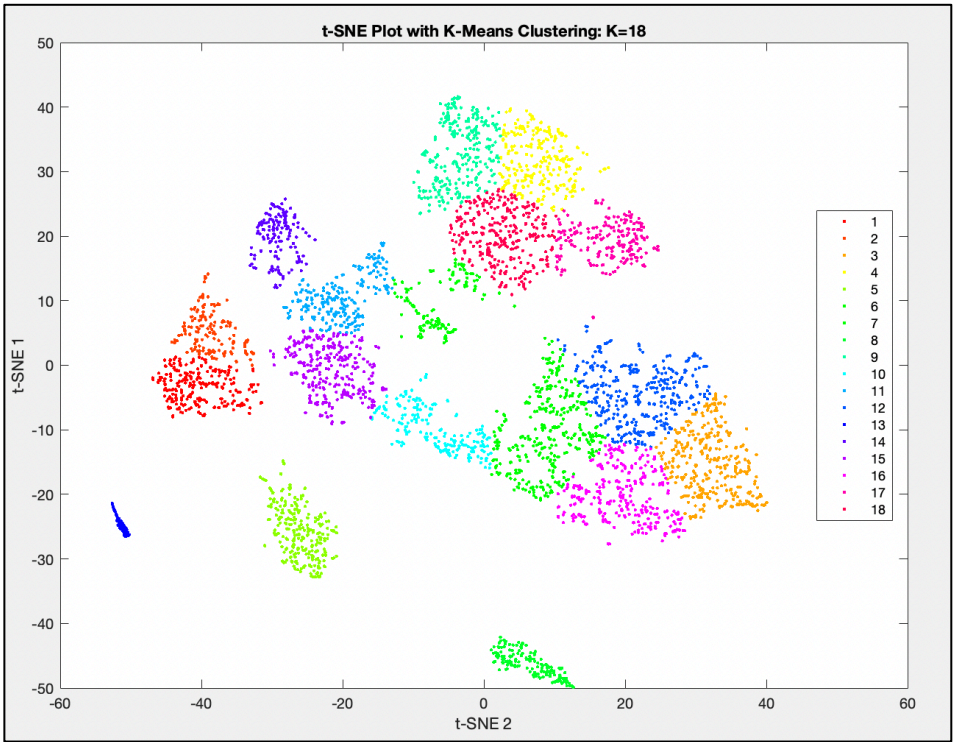

Figure 15: Displays a t-SNE plot of the data with the 18 K-means clusters labels.

Now, the clusters that include Ncr1 and Vwf cells can be identified. An easy was to do this is to create a bar chart displaying the cluster labels for all the Ncr1+ and Vwf+ cells.

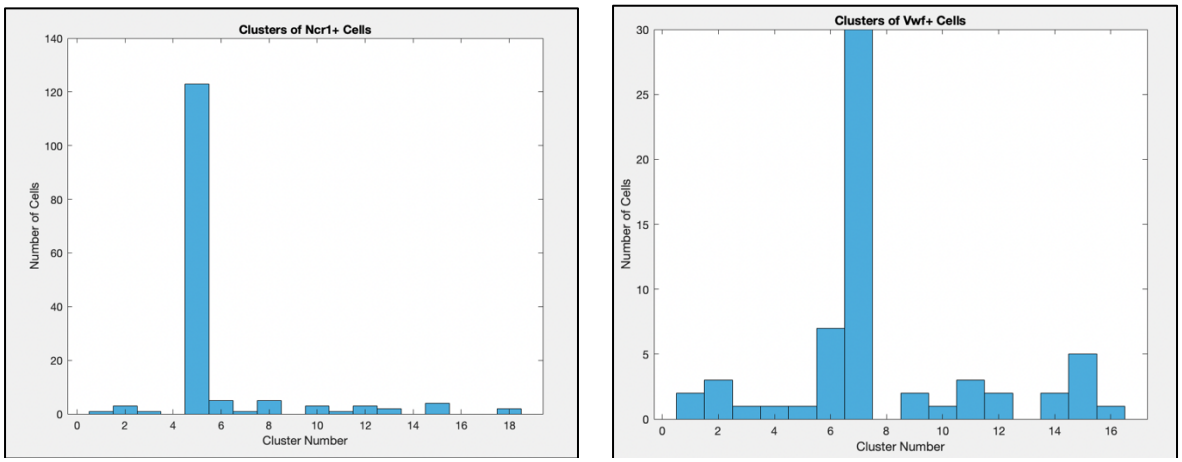

Figure 16: Displays histograms of the cluster labels assigned to Ncr1+ and Vwf+ cells.

It is clear from these histograms that the natural killer cells (Ncr+ cells) are found in cluster 5 and the endothelial cells (Vwf+ cells) are found in cluster 7. These clusters are marked on the following t-SNE plot to make their location easier to identify.

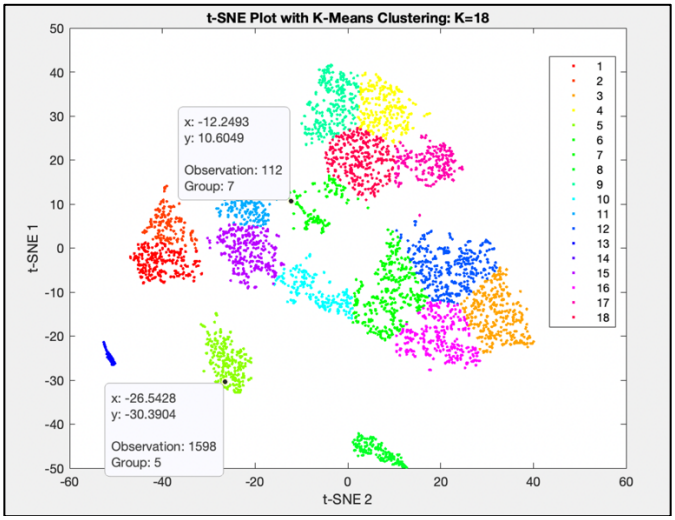

Figure 17: Displays the K-means clustered t-SNE plot with added labels specifically for clusters 5 and 7.

The identified natural killer and endothelial cells were separated from the rest of the data for use in the supervised learning portion of this project.

Supervised Learning

Now, the Ncr1 feature must be removed from the NK cell data and the Vwf feature must be removed from the endothelial cell data so that supervised learning can be performed without the models relying on the genes that were used to separate out the data originally. The dataset was then split into 80% for a training set and 20% for a testing set. The classification learner app was then used to train and test multiple supervised learning models to classify between the NK cells and the endothelial cells. For cross validation, 10 folds were selected to begin with and the value was adjusted between 5 and 10 to see if it would impact the training accuracy. Overall, 10 folds was found to be the best choice.

First, neural network models were trained. They were found to have consistently high accuracies close to 90% with the best being the Medium Neural Network at 90.7%. No attempt was made to improve these accuracies by experimenting with the misclassification costs. This is because if the

training accuracy is too high then the model is likely overfitting the dataset and will not perform well on the test set or other datasets.

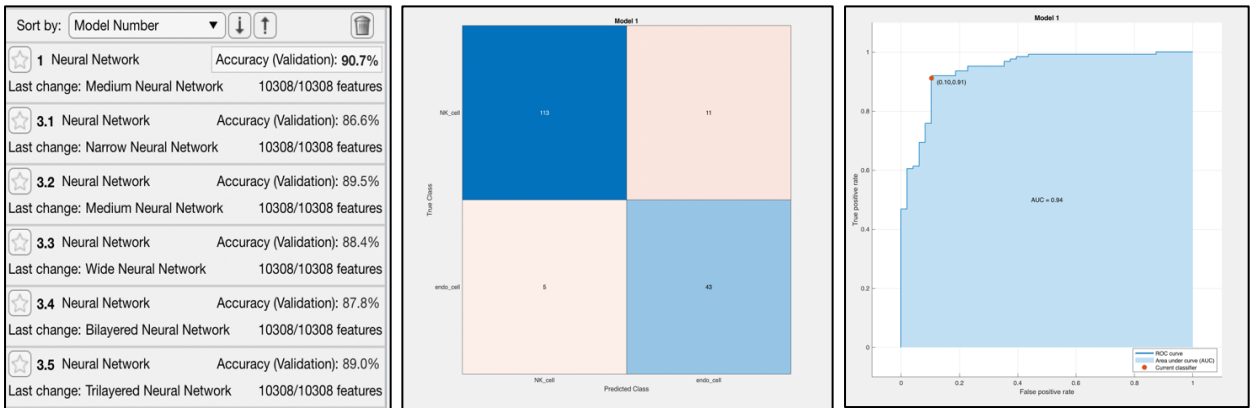

Figure 18: Displays the types of neural networks trained as well as the confusion matrix and ROC curve for the highest performing model.

Scatter plots were excluded from the report because the normalized and log2 transformed counts are quite small and the classification learner app was unable to plot them successfully. The sensitivity for the medium neural network is 91%. This means that 91% of NK cells were successfully classified. The specificity is 90% meaning that 90% of the endothelial cells were successfully classified. The axes on the ROC curve displays the true positive rate (or sensitivity) and the false positive rate, which is the proportion of endothelial cells that were incorrectly classified as NK cells. The red point marked on the ROC curve is the point at which the percentage of NK cells that are successfully classified is maximized without sacrificing a greater number of misclassifications of the endothelial cells.

Next, the SVM models were trained and had accuracies that ranged from 52.9% up to 87.2% for the quadratic SVM. The misclassification cost was adjusted to inflict a heavier punishment for the misclassification of NK cells in the Medium Gaussian SVM, which was the SVM with the lowest accuracy. This surprisingly ended up lowering the training accuracy to 46% even though many

NK Cells were misclassified in the model. As an experiment, 12 PCA components were used to see if the original accuracy could be improved but this did not work as all cells ended up being predicted as NK cells. Upon further investigation, some of the other SVM models that appear to perform decently (around 70% accuracy) are actually just classifying all cells as NK cells. This yields a fairly good accuracy since there is a higher proportion of NK cells in the dataset. In the end, the Quadradic SVM was the best choice of the SVMs. The sensitivity and specificity of the quadratic SVM are 96% and 65% respectively.

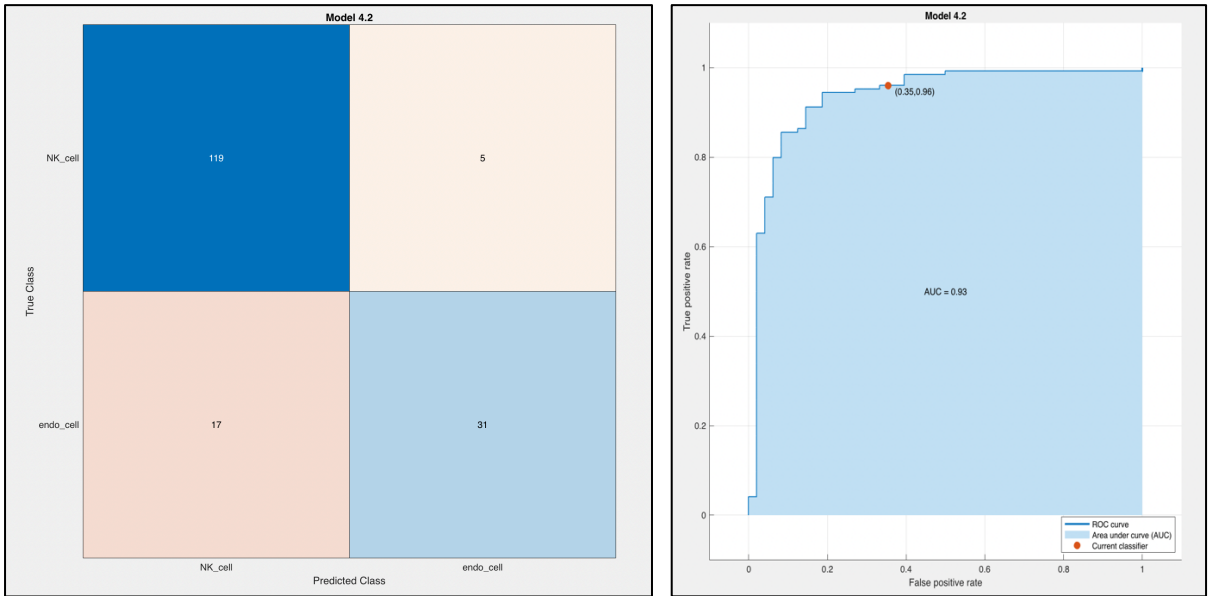

Figure 19: Displays the confusion matrix and the ROC curve for the Quadradic SVM.

The accuracies of the KNNs were mostly in the 70% range but they also tended to predict all or most cells as NK cells like the the SVM models. The highest performing KNN was the Cosine KNN at 86% accuracy. In an attempt to improve this model, the misclassification costs were adjusted, and the number of nearest neighbours was increased to 15 in the advanced KNN tuning settings. These changes made little difference to the model accuracy.

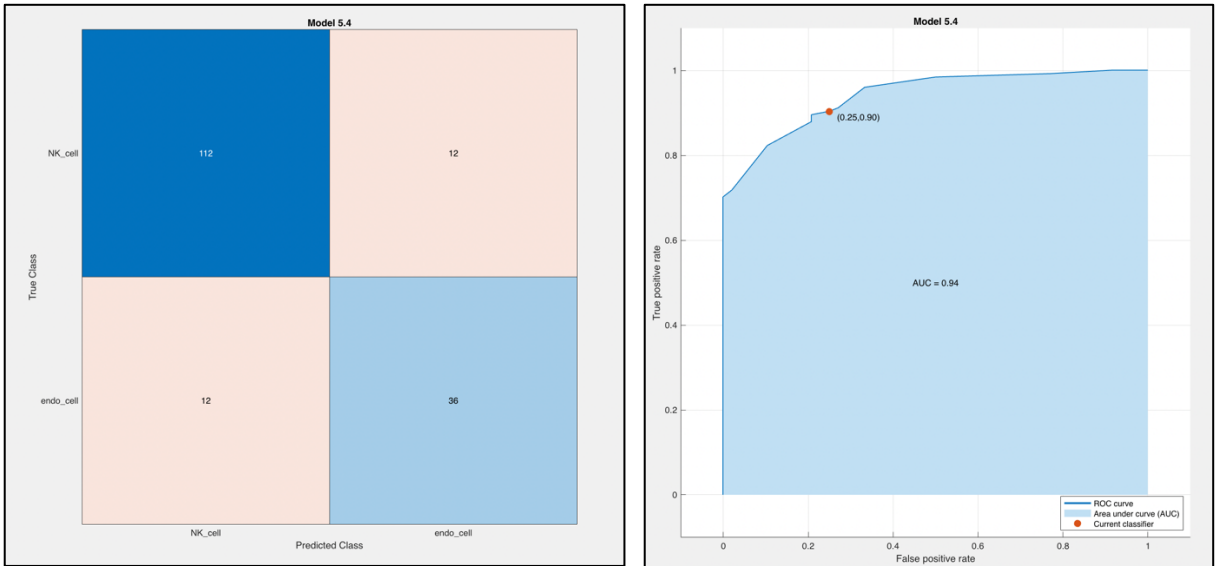

Figure 20: Displays the confusion matrix and the ROC curve for the Cosine KNN.

The sensitivity and specificity of the Cosine KNN are 90% and 75% respectively. A logistic regression was also attempted but it only yielded an accuracy of 65.7%. The Medium Neural Network, the Quadradic SVM, and the Cosine KNN were then all used on the test dataset. All three models achieved the same score of 88.4% on the test dataset. This result as well as a comparison between the top 3 models will be done in the discussion.

Statistical Tests

The Mann-Whitney U Test (also called the Wilcoxon Rank Sum Test) is a non-parametric way of checking for a difference between two independent groups such as the gene expression of NK cells at the Week 1 timepoint and the NK cells at the Week 3 timepoint. This test checks if there is a difference in the ranked sum between the groups instead of the checking for a difference in mean. The ranked sum is calculated by ordering the cells from smallest gene expression to largest gene expression then summing the ranks for each independent group / sample /timepoint. This means the data does not have to be normally distributed to compare across groups. The null hypothesis is that there is no difference in the sum of the rankings in the two groups. The alternative hypothesis is that there is a difference between the ranked sum of the two groups.

Before performing any statistical tests, features that were not expressed in either the Week 1 or Week 2 NK cells were filtered out leaving around 6400 features. This is still too many features to be able to perform the Mann-Whitney test. This is because p-values must be adjusted using either the Bonferroni method or the Benjamini-Hochberg procedure to account for the number of comparisons being made between independent samples/timepoints. Performing the Mann-Whitney test was attempted using this many features/genes in R just to try it but all the adjusted p-values ended up indicating that the null hypothesis should be accepted. This was because they were all undergoing a large correction due to the 6400 comparisons.

In order to overcome this obstacle, the feature selection method fscmrmr was used to further reduce the number of features being used in comparison across groups.

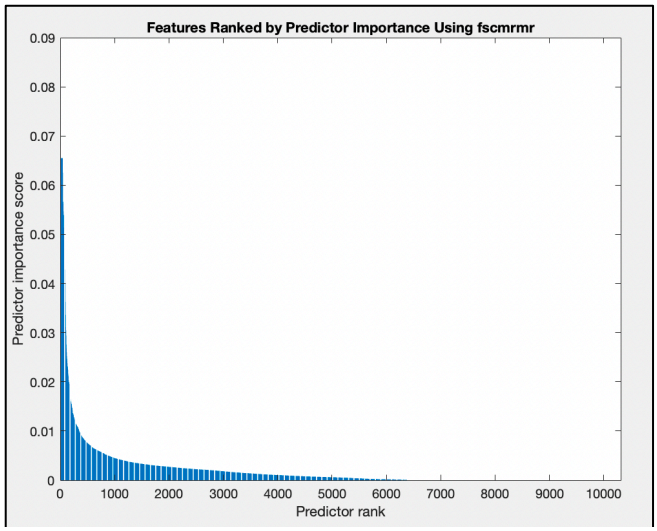

Figure 21: Displays the ranked features based on predictor importance determined by fscmrmr.

The features above the elbow at around 300 were selected and used in the Mann-Whitney test to try and overcome the effect of having so many features on the adjusted p-value. The Mann-Whitney test was performed in SPSS. The 1 week samples were marked as 0 and the 3 week samples were marked as 1. After completing the test, there were p-values less than 0.05 associated with some features. These features will now be reported.

Gene Name	Sig.	Sample	N	Mean Rank	Sum of Ranks	Z	Effect Size [r]
Hopx	<0.01	0	94	89.09	8374.00	-4.3	-0.34757
		1	60	59.35	3561.00		
		Total	154				
Prpf38a	0.002	0	94	81.33	7645.00	-2.9	-0.23124
		1	60	71.50	4290.00		
		Total	154				
Heatr1	0.021	0	94	80.05	7525.00	-2.3	-0.18632
		1	60	73.50	4410.00		
		Total	154				
Akr7a2	0.002	0	94	74.50	7003.00	-3.1	-0.25114
		1	60	82.20	4932.00		
		Total	154				
Asl	0.003	0	94	72.87	6850.00	-3	-0.23872
		1	60	84.75	5085.00		
		Total	154				
Aven	0.02	0	94	73.67	6925.00	-2.3	-0.18814
		1	60	83.50	5010.00		
		Total	154				
Orai2	:0.01	0	94	81.21	7633.50	-2.6	-0.20864
		1	60	71.69	4301.50		
		Total	154				
Zbtb44	0.021	0	94	80.05	7525.00	-2.3	-0.18632
		1	60	73.50	4410.00		
		Total	154				
Unc13d	0.021	0	94	80.05	7525.00	-2.3	-0.18632
		1	60	73.50	4410.00		
		Total	154				
Ubl4a	0.009	0	94	80.69	7585.00	-2.6	-0.2097
		1	60	72.50	4350.00		
		Total	154				
Tm9sf1	0.032	0	94	80.24	7543.00	-2.1	-0.17252
		1	60	73.20	4392.00		
		Total	154				
Lonp2	0.014	0	94	80.37	7555.00	-2.5	-0.19828
		1	60	73.00	4380.00		
		Total	154				
Efr3a	<.001	0	94	70.15	6594.00	-3.9	-0.31131
		1	60	89.02	5341.00		
		Total	154				
Ccdc58	0.021	0	94	80.05	7525.00	-2.3	-0.18632
		1	60	73.50	4410.00		
		Total	154				
Nelfe	0.009	0	94	74.80	7031.00	-2.6	-0.21011
		1	60	81.73	4904.00		
		Total	154				

Figure 22: Displays the Mann-Whitney test output values that have significance values less than 0.05.

Discussion

Visualization and Quality Control

If I were to perform the preprocessing again steps again then I would calculate the ratio that I used for checking if the outliers were overexpressing a single gene on every single cell. I think this would catch other cells that are expressing abnormally high amount of a single gene. The decision to abandon the control sample was a difficult one but I feel it was justified. Some of the other timepoints provided such as 5 and 8 Weeks could be used in the future to confirm that this was a good decision. If they have similar scatter and box plots to the 1 Week and 3 Week then I will know that abandoning the control sample was the right decision.

Unsupervised Learning

A good amount of time was spent adjusting the perplexity and number of principle components used in the t-SNE. In the end, the impact seems to only have been minimal. Regardless, the approach taken to selecting values for each of these parameters was a logical one. It was surprising that none of the silhouette plots were perfect. All the silhouette plots at least had a very small amount of negative silhouette values.

Supervised Learning

The Medium Neural Network is the best out of the three top classification models presented in the results section. All three models have a similar AUC, but the Medium Neural Network and Cosine KNN are slightly better than the Quadradic SVM. Models with higher AUC are better so the Quadradic SVM is a slightly worse model. In the training dataset, the Medium Neural Network only misclassified 16 cells in total while the Cosine KNN misclassified 24. Additionally, the

Medium Neural Network specificity is substantially higher at 90% compared to 75% for the Cosine KNN. This means that more endothelial cells are successfully classified in the Medium Neural Network. Both models had similar sensitivities. The also performed the same on the test dataset, which seems odd. This result should be viewed with skepticism. The data loaded into the program was definitely correct. Perhaps, the test dataset should have been larger and maybe the data should have been divided into 70% training and 30% testing instead. Thus, more weight should be placed on the other information and the training accuracies. Taking in all this information, the Medium Neural Network is the better model.

Statistical Tests

From the Mann-Whitney U Test, there were approximately 40 features with p-values less than 0.05, however, many of them had little difference in their mean ranks. The genes displayed in the above Figure 26 are those that had a low p-value and a larger difference in mean ranks of the different timepoints/samples. According to the IBM SPSS guide, the Mann-Whitney results should be reported with the z score and effect size even though this is a non-parametric test that does not require the dataset to be normally distributed. The bigger concern with these results is that it is not clear from the documentation whether SPSS is automatically adjusting the p-value or significance for multiple comparisons. If SPSS is displaying the adjusted p-values then everything is fine. However, if it is not then all these p-values will be very large after an adjustment with the Bonferroni method. In this case, it would mean multiplying each by 300 for the 300 comparisons made. This would result in none of the results rejecting the null hypothesis.

Future Work

It would be valuable experience to align the original FASTQ files to the genome using the techniques taught in class to verify that it was done correctly in the count matrix that was provided. Additionally, further investigation needs to be done of the MULTI-Seq package that was used to align the sample barcodes. A deeper understanding needs to also be formed of the Mann-Whitney U Test in SPSS to confirm whether the p-values found are adjusted or not. Once this is accomplished, the other samples can be analyzed in a similar way.

Conclusion

Single cell RNA-sequencing is growing in popularity due to its ability to identify the cell types being expressed in tissues. However, the technique does not come without its challenges. Two of these challenges are working with the enormous matrices and handling the fact that most of the values in these large matrices are zero. This project has been a valuable experience in solidifying my understanding of R and MATLAB as well as establishing my knowledge of unsupervised and supervised learning models. I look forward to future projects with similar datasets!

References

[1] *Should I calculate TPM, RPKM or FPKM ... - 10X Genomics*. (n.d.). Retrieved April 10, 2022, from https://kb.10xgenomics.com/hc/en-us/articles/115003684783-Should-I-calculate-TPM-RPKM-or-FPKM-instead-of-counts-for-10x-Genomics-data-

[2] Godoy, R. S., Cook, D. P., Cober, N. D., McCourt, E., Deng, Y., Wang, L., Schlosser, K., Rowe, K., & Stewart, D. J. (2021). Novel apelin-expressing gCap endothelial stem-like cells orchestrate lung microvascular repair. https://doi.org/10.1101/2021.07.12.452061