# Cardiomyocyte classification using single cell RNA-seq from human fetal hearts

Ioannis Anastopoulos
TIM 209 Final Project

## Introduction

In recent years high throughput sequencing has revolutionized genomics. Scientists are now able to sequence multiple samples at great detail with minimal cost. This advance in sequencing technologies has allowed geneticists to study human disease, and the characteristics of what makes each individual different from any other, in much greater detail. Take for example the 1000 Genomes project, which has sequenced DNA from 1000 individuals and has reported on SNPs (Single Nucleotide Polymorphisms) that contribute to distinct phenotypes among individuals. Before going any deeper into sequencing and how it allows us to study biology, lets first understand how the genome works.

In brief, the cell contains all its genomic information at its center: the nucleus. The nucleus is surrounded by a double lipid bilayer, which controls traffic of various molecules. In the nucleus is where DNA lives. DNA is the molecule that contains the genetic code of each cell. DNA, via a process called transcription, expresses genes via RNA. While DNA is a double strand molecule, RNA is single stranded. Typically, RNA crosses the nuclear membrane into the cytosol of the cell, where it becomes translated into proteins, via a process called translation. Proteins are responsible for any characteristic that is different between individuals, for example heir and eye color. The whole process can be describes succinctly as follows:

DNA → RNA → Protein

**Figure 1** that describes the above process:

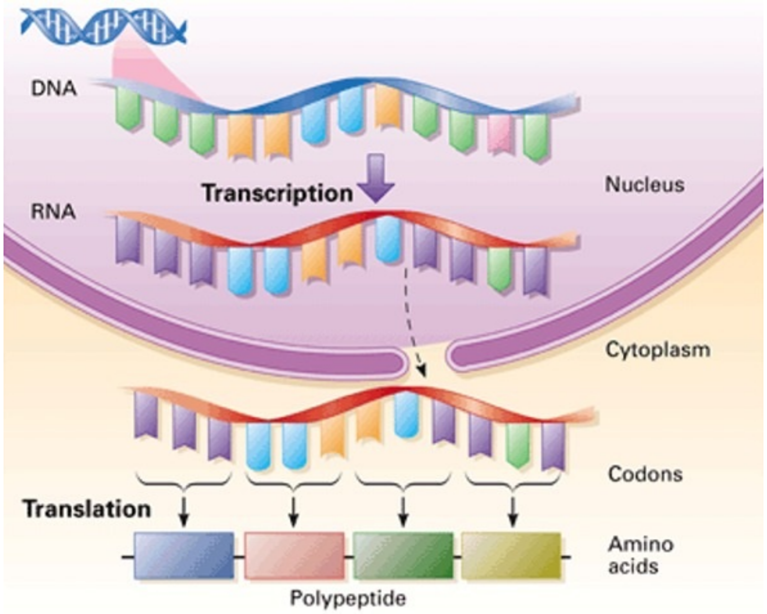


*Figure 1: DNA conversion to RNA and eventually protein. The central dogma of biology*

In this report, the main focus will be on RNA. Recent advances in sequencing technologies have allowed us to be able to sequence the transcriptome (everything that is constituted by RNA) of every single cell in a tissue. Typically, the field has been sequencing DNA/RNA form bulk tissue. What this means is that a piece of tissue (heart, tumor, skin, etc) is obtained, and is processed via a chemical process that exposes the molecules that live in the nucleus of each cell. This amalgam of different cells from the tissue is sequencing simultatneously. While this technology has allowed scientists to make new advances in many fields, for example cancer biology, it is still limited by a number of factors.

Bulk sequencing has been incredible useful in the field of cancer research. In the last decade, many discoveries have allowed clinicians to better understand various types of cancrer, and treat it better. Take for example the discovery of the gene *BRCA*, which is responsible to a large percentage of breast cancers. Another example is the mutation *BRAF* V600E, which is present in >60% of melanoma patients. Additionally, scientists have been able to characterize the genetic causes of more rare diseases such as uveal melanoma. The shortcoming, however, of such sequencing technology, as many studies have shown, is that it is unable to capture the heterogeneity of the tissues we study. Each cell in a tissue does not have a uniform expression of all its genes, because each cell is in a different location and it therefore responds to different stimuli, causing it to express different levels of the same gene. This is where single cell RNA sequencing holds a lot of promise.

Single cell RNA sequencing is able to extract, and sequence the RNA molecules of each cell in the tissue we study. This has led to the foundation of large consortia, such as the Human Cell Atlas (HCA), where UCSC plays a central role. The whole of the HCA is to sequence every cell in the human body, with various labs contributing in the project.

This report will focus on data that UCSC and the Stuart lab is working on. The data consists of human fetal heart cells, and their RNA sequencing. These cells have been collected from various developmental points, and from various locations in the heart. The goal is to be able to characterize the main type of cells in the heart, the cardiomyoctes, and to be able to suggest treatment to people that are affected by diseases in these cells. Here, I am attempting to build two types of classifiers:

1. A classifier that will be able to distinguish cardiomyocyte cells from non-cardiomyocyte cells
2. A classifier that will be able to distinguish cardiomyocyte cells of the atria, from cardiomyocyte cells of the ventricles, the two major regions of the heart.

**Figure 2**, outlines the major regions of the heart, along with the processi for scRNA-seq.

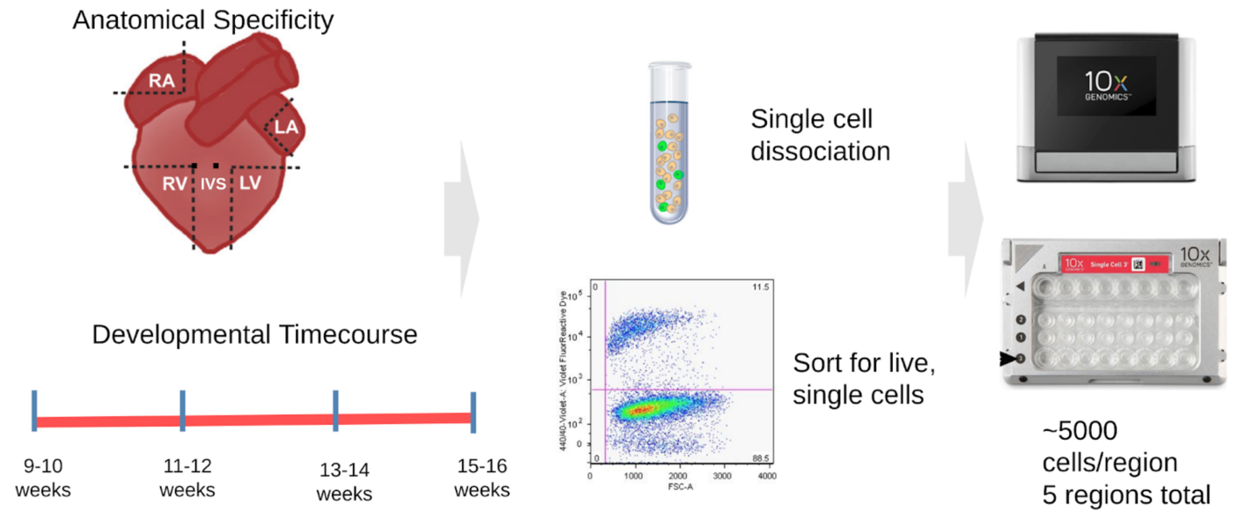


*Figure 2*: *Regions of the heart, developmental time at which samples were taken, and the process for RNA sequencing*

Data preprocessing

The RNA-seq data that were obtained the expression of each gene was normalized according to standard practices in the field. On top of that, in order for the classifiers to better understand the data, genes that have low variance were removed from the dataset. This is because a gene that does not display significant variation among cells, can either be considered a housekeeping gene, meaning a gene that is responsible for maintaining the cell alive, by carrying our basic functions, or a gene that is silenced, meaning it is just not expressed. Such genes give low information as to the characteristics of each cell, so they are removed from the dataset.Secondly, unsupervised clustering was carried out in two levels:

1. All cells were clustered and the marker gene *TNNT2*, that is known to be a marker gene for cardiomyocytes was used to identify **cardiomyocyte-like cells**.
2. The cardiomyocyte-like cells that were identified in step 1 were further subclustered, and colored by the region that they came from.

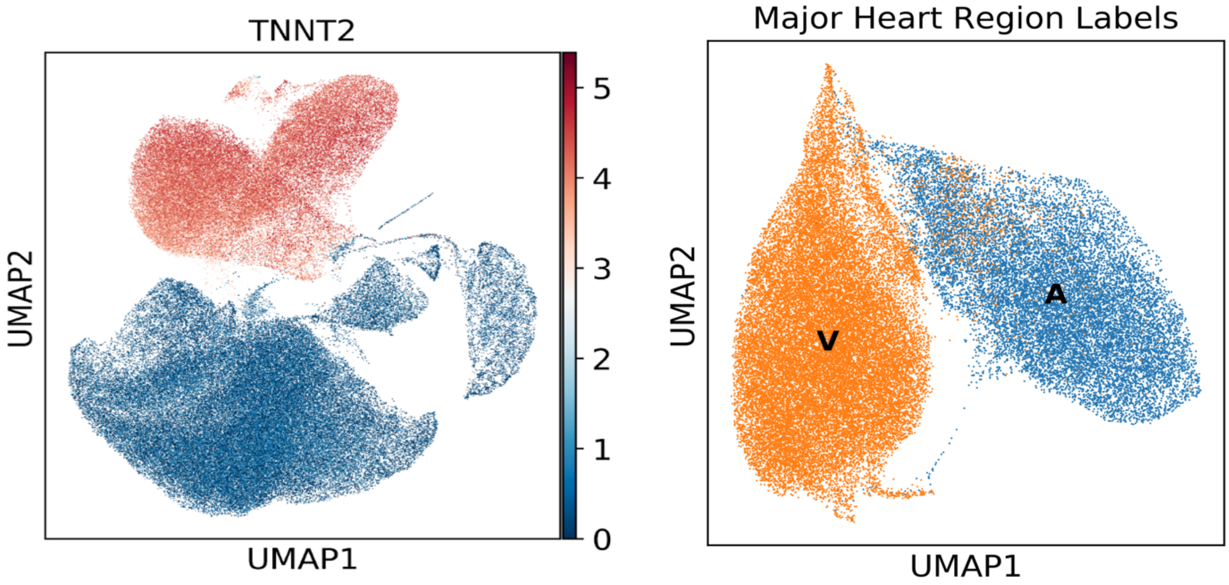


***Figure*** 3. Left: cardiomyocyte-like cells colored by the expression of their marker gene. Right: sub-cluster of the identigied cardiomyocyte-like cells, colored by the region they came from

**Figure 3** shows the unsupervised clustering results carried out using UMAP. It is clear from unsupervised clustering that the transcriptomic profile of each cell, meaning the patert of gene expression it displays, is indivative of the type, as well as the location they came from. These results were used to label the cells in our data, and those labels were used in the classifiers presented below.

Methods

It should be noted, that due to the large size of the dataset, not many models could be tested.

- **Cardiomyocyte classification:**
  The classifiers used for cardiomyocyte classification are the following: logistic regression, random forest and KNN. For logistic regression lasso regularization was used, along with the Adam optimizer. The best learning rate, and lambda for the lasso were determined via 10 fold cross validation over a space of possible values for both.

  For random forest the best parameters were determined based on 10 fold cross validation over a space of possible combinations. The space of possible parameters examined was the following:

| Parameter | Space |
|---|---|
| # of trees | 10, 20, 30, 40 trees |
| Bootstrap | True, False |
| Evaluation criterion | Gini, Entropy |
| Max depth of each tree | 3, 4 splits |
| Minimum samples required to split an internal node | 2, 3, 10 samples |

For knn a range of 1 – 20 neighbors was determined via 10-fold cross validation, and the best score was determined by average MSE.

- **<u>Ventricular and Atrial Cardiomyocyte Classification</u>**
  The classifiers used for cardiomyocyte classification are the following: logistic regression, Neural Net and random forest.

  For logistic regression lasso regularization was used, along with the Adam optimizer. The best learning rate, and lambda for the lasso were determined via 10 fold cross validation over a space of possible values for both.

  For the neural net, the following architecture was used:

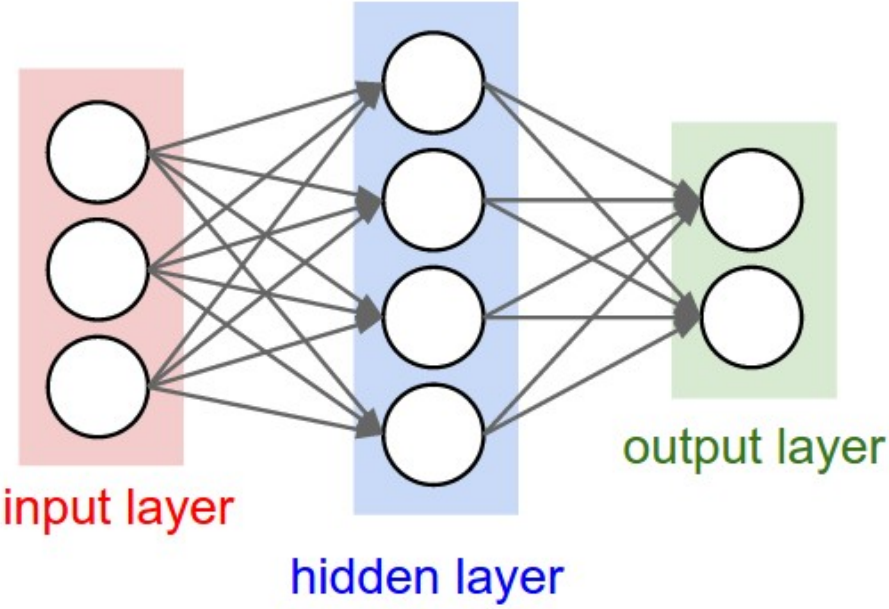


***Figure 4***: *neural net architecture used for the cardiomyocyte region classfier*

The hidden nodes, and regularization of the layer were determined via 10-fold cross validation. Interestingly, both lasso and dropout did produce better results, compared to when they are not used. Glorot uniform initialization of weights was used. As for the activation function, and the loss function, tanh was used in the hidden layer, and binary cross entropy was used respectively. On the output layer the sigmoid activation function was used. More specifically:

| Parameter | Space |
|---|---|
| Learning rate space | np.linspace(0,0.0005, 5) |
| Hidden dimension space | 500 ,600 |

For random forest the best parameters were determined based on 10 fold cross validation over a space of possible combinations. The space of possible parameters examined was the following:

| Parameter | Space |
|---|---|
| # of trees | 10, 20, 30, 40 trees |
| Bootstrap | True, False |
| Evaluation criterion | Gini, Entropy |
| Max depth of each tree | 3, 4 splits |
| Minimum samples required to split an internal node | 2, 3, 10 samples |

For all models the evaluation was done by computing the following:

- F scores for each label
- ROC curves for each label
- Confusion matrices for each classifier
- 10-fold CV accuracy score, with the exception of KNN, where MSE was used

## Results

- **Cardiomyocyte Classifiers**

For logistic regression the best lasso lambda chosen was 0.01, and the best learning rate was chosen to be **0.0025.**

For random forest, the best parameters were the following:

- bootstrap: **False**
- criterion: **Gini**
- max depth of each tree : **4 splits**
- min samples split : **2**
- number of trees : **40**

| **Model** | **F score** | **CV acc** |
|---|---|---|
| **Logistic Regression** | **CM-like:** 0.968 | **96%** |
| | **Non CM-like**: 0.984 | |
| **Random Forest** | **CM-like**: 0.987 | **98%** |
| | **Non CM-like**: 0.993 | |
| **KNN** | Failed | Failed |

*__Table 1__: Summary of cardiomyocyte classifiers*

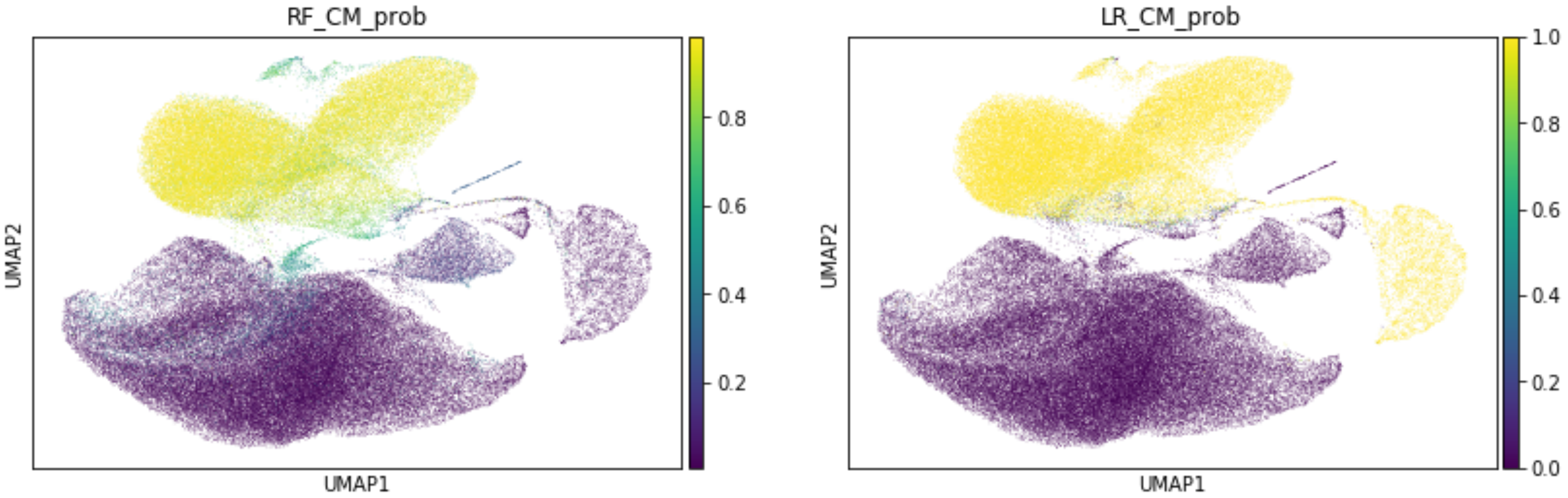


*__Figure 5__: UMAP plots with the probability of being a cardiomyocyte as determined from each classifier. Left: probabilities computed by random forest. Right: probabilities computed by logistic regression*

Figure 5 shows the probability of being a cardiomyocyte computed by each of the classifiers used. Compare figure 5 to the left panel of Figure 3. The region colored by the known marker gene for cardiomyocytes most closely relates to the probabilities from the random forest classifier. Logistic regression seems to have more false positive cells. This can also be observed in **Figure 5.** The AUC for random forest is higher compared to logistic regression. The confusion matrices for the two models can also confirm the above observation. The total number of false positives for random forest are 320, compared to 780 for logistic regression.

Furthermore, thanks to the linear nature of the logistic regression classifier, we are able to extract the most important genes for this classification, given by the magnitude of the coefficient for each predictor (genes) (**Figure 8**). This makes logistic regression that much more valuable for this problem. On the other hand, we are able to extract the same information from the random forest package, however, the algorithm has fitted many trees, with a different order and combination of genes in each split. This makes interpreting the model in this case a little harder.

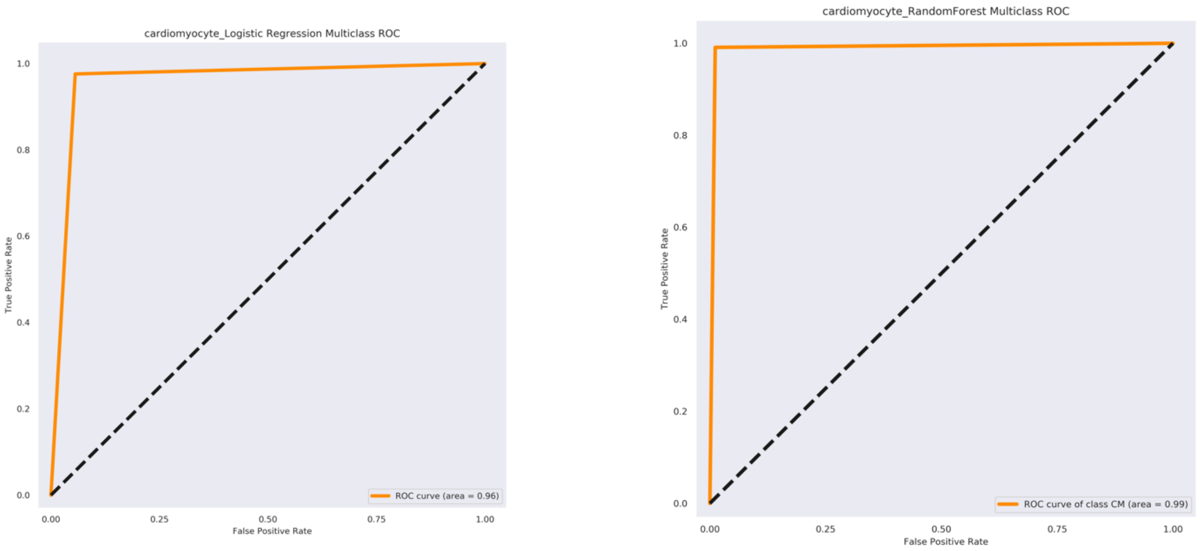
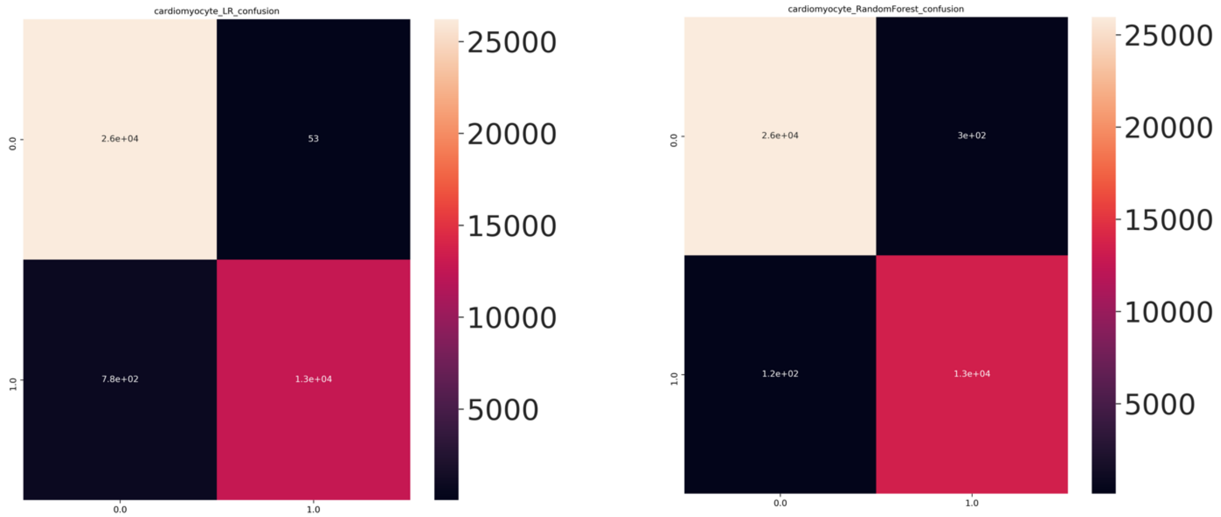
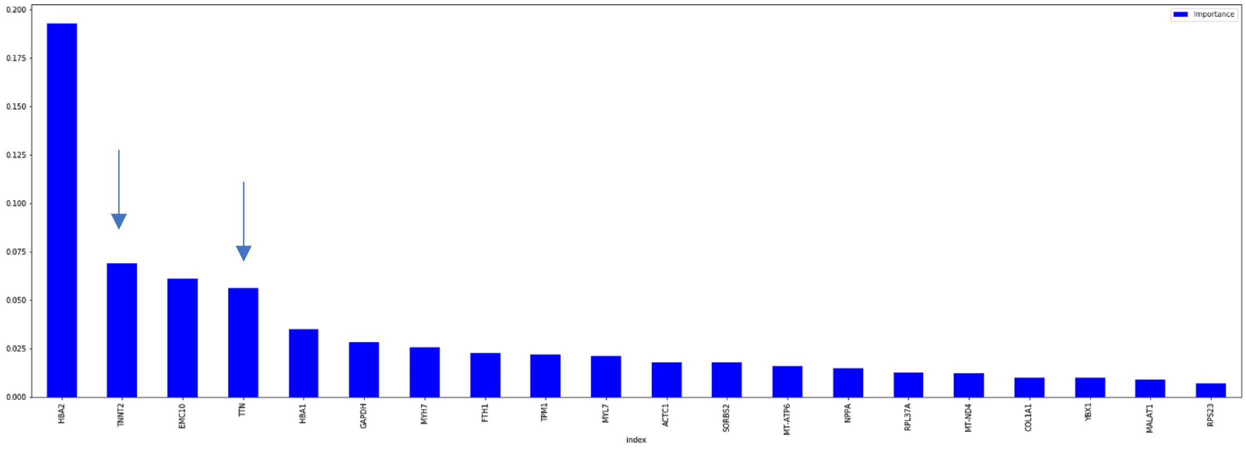


*__Figure 6__: ROC plots of the models used to classify cardiomyocytes. Left:  ROC and AUC of logistic regression. Right; ROC and AUC of random forest.*



*<u>**Figure 7**</u>: Confusion matrices. Left: confusion matric for logistic regression. Right: confusion matrix for random forest.*



*<u>**Figure 8**</u>: Waterfall plot of the top 20 important genes as determined by the logistic regression coefficients. The arrows point to known marker genes for cardiomyocytes, with TNNT2 being the most improtnat*

- **Ventricular vs Atrial cardiomyocyte classifiers**

| Model | F score | CV acc |
|---|---|---|
| **Logistic Regression** | **Ventricular CM**: 0.932 | **91%** |
| | **Atrial CM**: 0.871 | |
| **Random Forest** | **Ventricular CM**: 0.910 | **88%** |
| | **Atrial CM**: 0.782 | |
| **Neural Net** | **Ventricular CM**: XX | XX |
| | **Atrial CM**: XX | |

For logistic regression the best lasso lambda chosen was **0.000125**, and the best learning rate was chosen to be **0.000125.**

For random forest, the best parameters were the following:

- bootstrap: **False**
- criterion: **Gini**
- max depth of each tree : **4 splits**
- min samples split : **3**
- number of trees : **40**

Since logistic regression did so much better than random forest, and is more interpretable than the neural net, its probability on the UMAP form the subclustered cardiomyocytes was plotted for confirmation (**Figure 9**). Note that in terms of interpretability, since logistic regression is a linear function of the dimensionality of the data (genes in this case) it is possible to plot the most important genes as determined by the coefficients of each predictor (gene) (**Figure 10**). This is something that is not as easily achievable with the neural net.

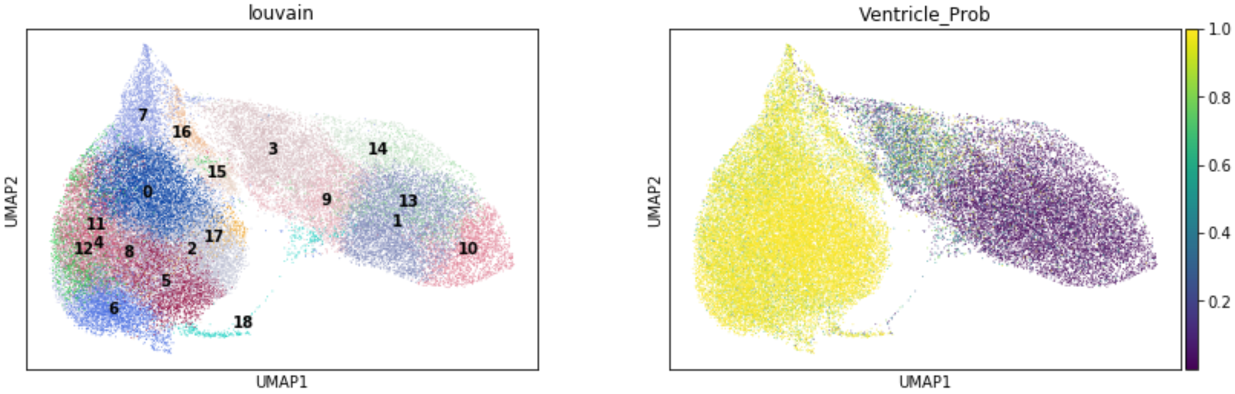


***Figure 9**: Ventricular and Atrial cardiomyocytes sub-clusters. Left: original clustering, with the left cluster indicating ventricular cardiomyocytes. Right: sub clusters colors by their probability of being a ventricular cardiomyocyte as determined by logistic regression*

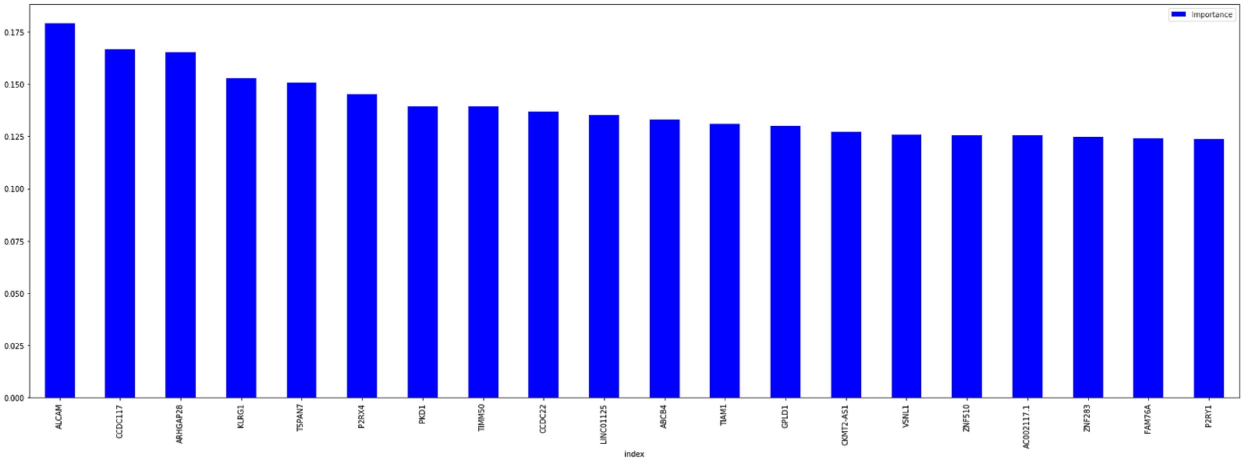


***Figure 10**: Waterfall plot of the top 20 important genes as determined by the logistic regression coefficients.*

In terms of performance the AUC of logistic regression was 0.97, compared to 0.89 of random forest. Indicating a much better classifier with a low false positive rate (**Figure 11**). The confusion matrices in **Figure 12** tell the same story. Logistic regression's false positives are 810, whereas random forest's false positives are 1500.

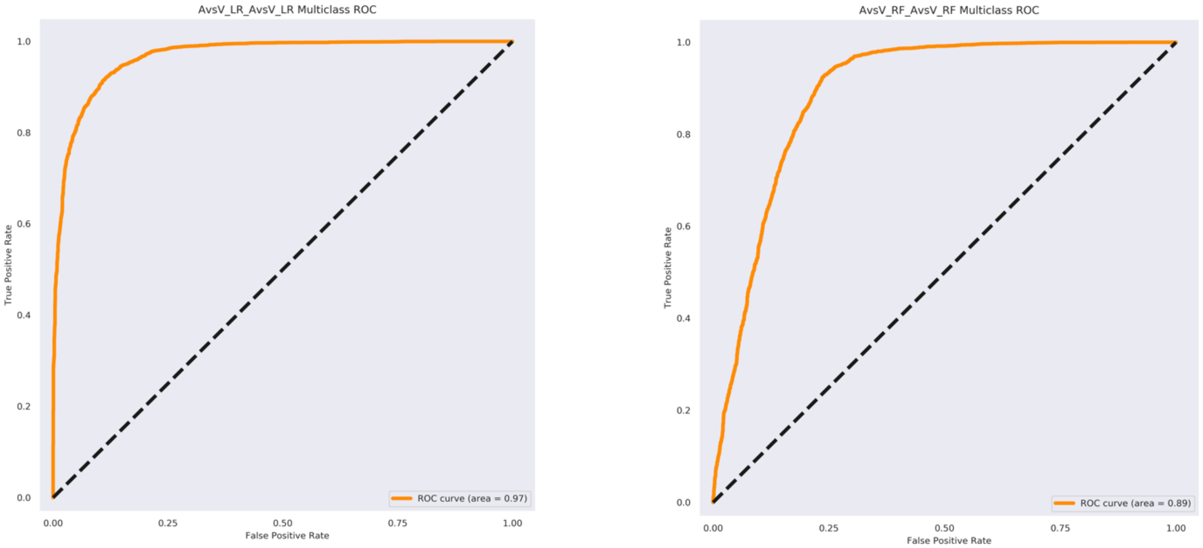


***Figure 11***: *ROC plots of the models used to classify cardiomyocytes. Left:  ROC and AUC of logistic regression. Right; ROC and AUC of random forest.*

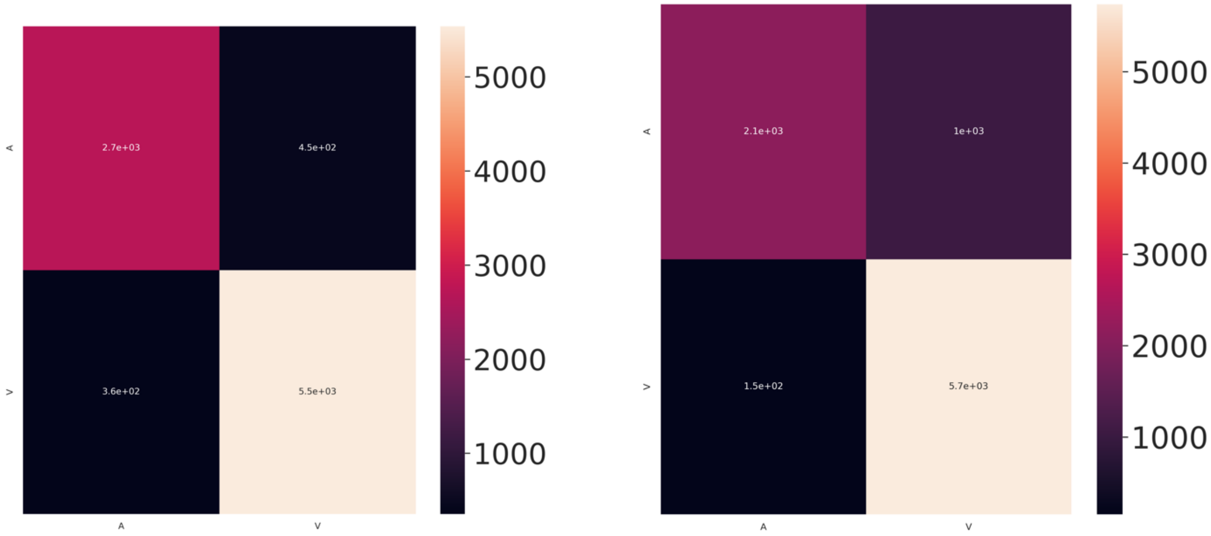


***Figure 12***: *Confusion matrices of regional classifiers. Left: confusion matric for logistic regression. Right: confusion matrix for random forest.*

## Conclusions

For cardiomyocyte classification, random forest did the best, whereas for the regional classifiers, the neural net did the best, but logistic regression is much preferred in this setting given its much higher interpretability, and not that lower performance compared to the neural net.

The random forest algorithm performed the best for the cardiomyocyte classification problem, however its interpretability in this context is not great. Because the algorithm user the square root of the number of predictors everytime a tree is made in the forest, the order of genes used is not known. Even though the algorithm can give you the importance of each gene used in the prediction, the way the trees use the genes and the order in which the genes are used, are very important. Tree extraction and analysis is required for this sort of interpretation of the model. On the other hand, logistic regression, given to its linear nature, and almost equal performance, would be the most preferred model in the case, because we are able to extract the most important genes, given the determined coefficients the algorithm assigns to each coefficient for each gene.

The neural net with the tanh activation function is able to capture a lot more non – linearity compared to logistic regression. This is the main reason why the neural net outperforms all other models used for the regional classifiers. However, immediate interpretability of the model is near impossible, as the original predictors have been transformed to latent variables of much lower dimension (500 latent variables compared to 28k genes). Unless some package, such as SHAPly, is used to interpret the impact of the original predictors, these latent variables are not interpretable. Again, logistic regression would be the preferred model here for the reasons outlines above.