I. Introduction

The paper that I selected to replicate images from is titled "Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy". Heart failure is one of the most common public health concerns and the researchers wanted to utilize RNA Sequencing of Single cells to study this disease. The motivation behind this paper is to perform molecular phenotyping on the common causes of heart failure that are Dilated Cardiomyopathy, Hypertrophic Cardiomyopathy and understand the underlying mechanism and disease progression better. In this paper researchers performed single nucleus RNA sequencing of samples from the left ventricles of 44 individuals suffering from Dilated Cardiomyopathy, Hypertrophic Cardiomyopathy, and unaffected individuals. After quality control and preprocessing, they obtained 592,689 nuclei samples and studied their expression data.

II. Data Analysis

a) Dendrogram-Principle

A dendrogram is a diagram which represents a tree and clusters many numbers of points based on their similarities. It is mostly created as the output from hierarchical clustering. The main use of dendrogram is to allocate objects to clusters from a scatterplot. The points on the dendrogram which share the same node are more closely related than other points. Further the difference in the height of the branches of the dendrogram can also tell us about the dissimilarity of the nodes. A greater difference in height indicated more dissimilarity.

In this paper they used the UMAP data to construct the dendrogram based on the similarity of the centers of the clusters. The UMAP was constructed using Leiden clustering on the nuclei expression data. All data that the researchers obtained was processed using Scanpy module. The procedure for clustering used in the paper was as follows

1. Selecting 2000 highly variable genes with *sc.pp.highly_variable_genes(flavor = 'seurat_v3')*
2. Normalizing cell count data *sc.pp.normalize(1e4)* followed by *sc.pp.log1p()*.
3. Expression of the highly variable genes was scaled down to unit variance and zero mean with *sc.pp.scale()* which is another form of normalization.
4. The top 50 principal components were visualized with PCA analysis using *sc.tl.pca()* which was implemented in pytorch.
5. The neighborhood graph was constructed with n=15 neighbors, using all 50 principal components using the line *sc.pp.neighbors(n_neighbors = 15).* This is the data upon which the UMAP is generated.
6. The UMAP was done on the resulting neighborhood *sc.tl.umap(min_dist = 0.2)*. This data was clustered with Leiden clustering with a resolution of 0.6 to cluster nuclei into 21 groups

In the paper they performed hierarchical clustering using the Scanpy function *scanpy.pl.dendrogram* to find the similarity between the center of the clusters

b) Methodology

Since the Scanpy module uses anndata file format which requires lot of computing power on a shared computing cluster to work, I utilized the SciPy's dendrogram function to create the plot based on the available data. The data which was made available to me from the plots was the x and y coordinates from the UMAP for each individual data point. I performed the clustering on this 2-dimensional data set to plot the dendrogram. Firstly, I loaded the text file using *pd.read_cs*v and extracted the data points to store them in a separate data frame. Looping through each of the 21 clusters, I found the center of the datapoints by using *df.mean()* and stored it in a separate array using the method *df.to_numpy(copy=True).* This is the data upon which the linkage matrices and the dendrogram is generated.

The distance matrix is created and stored in a variable using the method *sch.linkage(method = 'ward', metric = 'euclidean')*. The dendrogram is created from this distance matrix using the method *sch.dendrogram(leaf_rotation = 90)*. Since I am using a different library to perform clustering, I experimented with a variety of clustering methods

such as 'single', 'complete', 'average', 'weighted', 'centroid', 'median', 'ward'. I found that average clustering method with the Euclidean distance measurement produced the most similar dendrogram to the one produced in the paper.

c) Challenges

One specific challenge that I faced when replicating this image was that the groupby() method did not work when I wanted to split the samples in the UMAP data frame according to their clusters. It did not find the mean centers of all the clusters. To overcome this, I had to perform it in a different way through iteration. First, I converted all values in the 'X' and 'Y' columns of the dataframe from strings to float values. Then through iterating through the dataframe for each individual cluster and extracting the rows of the cluster I performed the method mean() to find the center points of the cluster. This was then stored in another data frame.

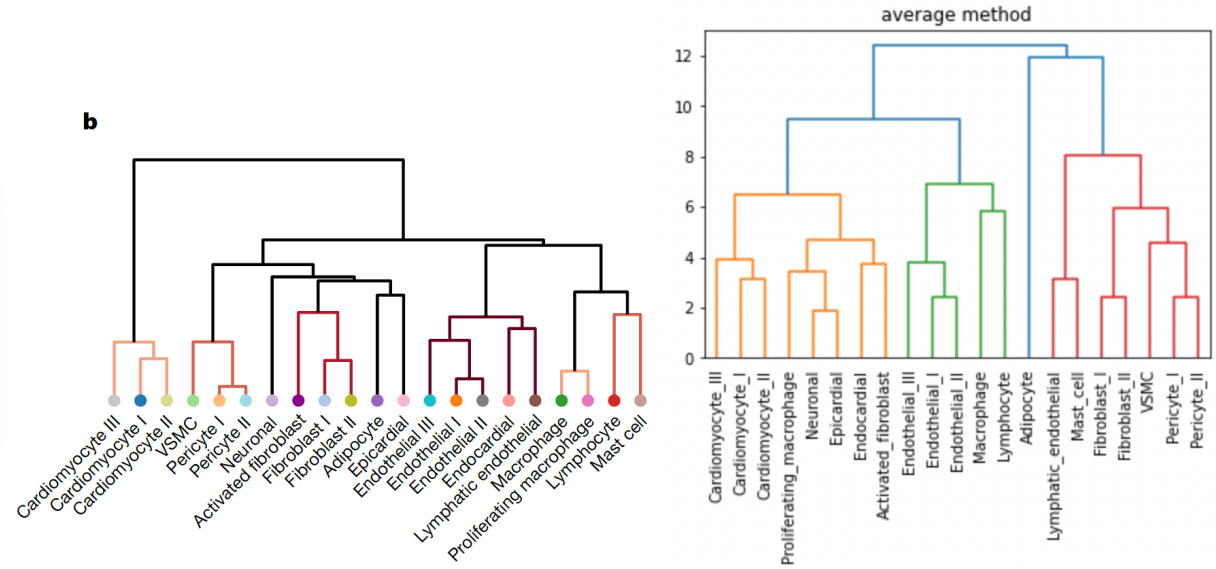

Figure1. (Left) Dendrogram created in the paper to demonstrate similarity of cluster centroids with Leiden Clustering (Right) Dendrogram created with Python's SciPy Function to demonstrate the hierarchical clustering of the cluster centroids.

d) Observations

The difference in the clustering of the dendrogram can be attributed to the difference in the methods used for clustering. The researchers used Leiden clustering while I used average method of clustering which produced similar clustering results on the first level of the dendrogram. Based on these 21 clusters we can infer the transcriptional similarity of the different cell types. In the paper this clustering method was also applied to generate the heatmap. Since this clustering method shows the similarity of the cell types, when used to generate the heatmap of transcriptional similarity, we will be able to make suitable biological observations of which genes are expressed in different cell types which can be used to make conclusions about the etiology of heart disease.

III. Data analysis-Heatmap

a) Principle

Heatmap is a powerful tool which is used for the visual display of expression data obtained from microarrays or NGS. Heatmap is a graphical representation of data that uses a system of color-coding in representing different values contained in a matrix.

Heatmap is a useful in analyzing and visualizing multi-dimensional datasets, is generally utilized in studying samples with gene expression data It is mainly to locate hidden groups among analyzed genes. It is also used in associating experimental conditions to gene expression patterns. This is done with combining heatmap with

clustering methods which groups genes based on the similarity of their expression patterns. The data in the heat map is usually displayed in a grid wherein each row represents a gene is on the y axis while each column representing a sample is on the x-axis. The changes in gene expression are represented by a color gradient.

In this paper they have done hierarchical clustering on the cell types and showed the heatmap of the 2000 most highly expressed genes were selected based on the enrichment values in biological processes based on gene ontology. The expression values of the genes were Z-score normalized to infer beneficial information from the heatmap.

b) Methodology

The expression data for the genes were provided in the dataset with column as the clusters and rows as the different genes. This data was read as a data frame and the values were converted into an array which can be read by the seaborn cluster map function. Clustering was only performed on the cell type clusters and not performed on the genes which is like the paper. The resultant heatmap was modified to be similar to the graph on the paper.

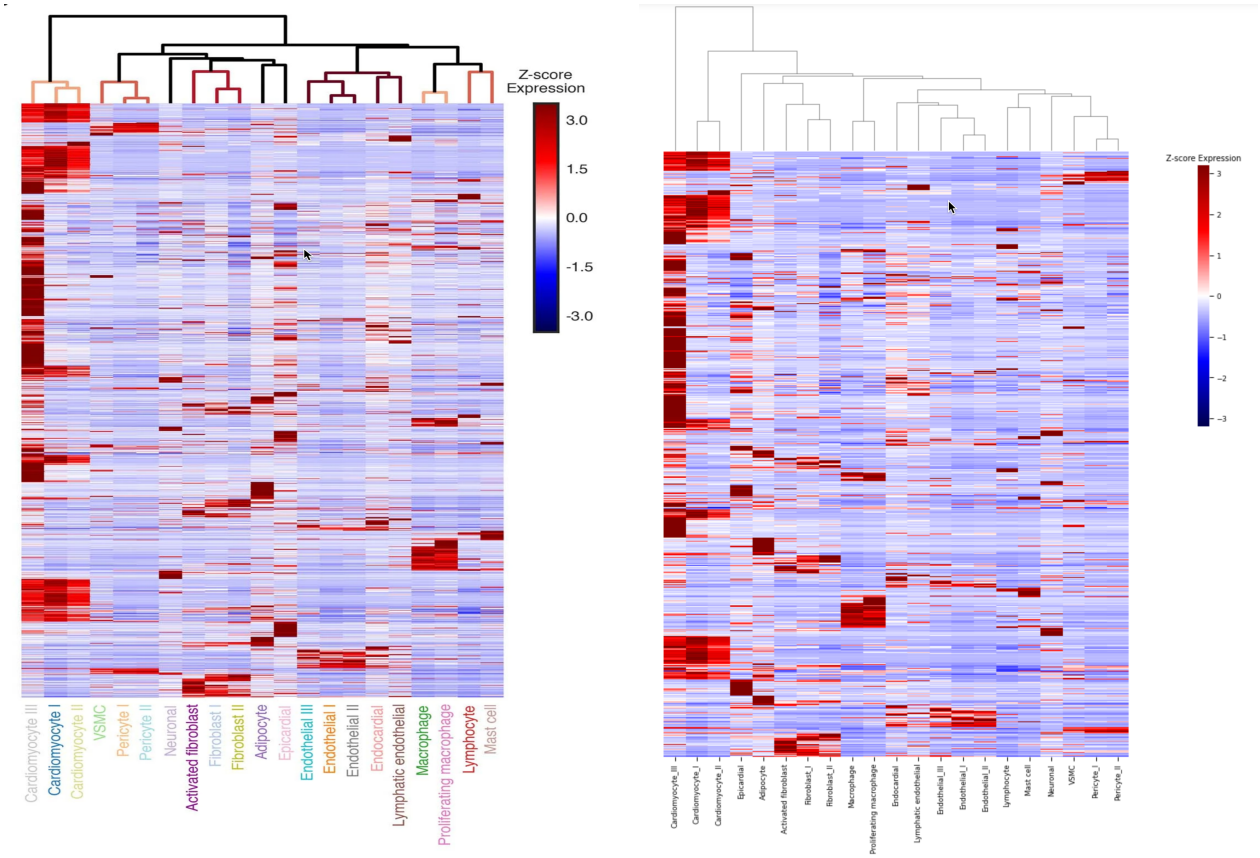

Figure 2. (Left) The heatmap depicting the top 2000 most differentially expressed genes on y-axis and cell type clusters on X-axis. (Right) Heatmap generated using the Seaborn's Clustermap function with clustering only performed on the cell type along X-axis.

c) Observation

In this cluster map, based on the data, which was already given, clustering was not done along the Y-axis but still we are able to locate some pattern to the gene expression data. However, if clustering was also done on the Y-axis, I observed that we can see a better correlation between the genes more commonly expressed between clusters. Normalization of the expression scores also allows us to better make observations in the graph. Since the methods used for the clustering of the x-axis are different, there is a difference in the 2 graphs. I implemented a similar approach which I followed for the heatmap, where I ran the cluster map function with multiple methods to

find out which method could produce a graph most like the image shown in the paper. I found this clustering method to be ‘Ward’ with ‘Euclidean’ Distance.

V. Conclusion

There are many different methods available to perform analysis on Bioinformatics data. Each of these methods produce a different output, due to the way in which the underlying algorithm works. In such cases, the method in which we choose to perform our analysis contributes a great deal to the inference that we want to make. So the decision on which software, statistical package or method that we implement in research is important and has to be optimized.

VI. References

1. Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open-Source Software, 6(60), 3021. https://doi.org/10.21105/joss. 03021
2. Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357-362 (2020). DOI: 10.1038/s41586-020-2649-2. (Publisher link).
3. McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).
4. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007
5. Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0
6. Chaffin, M., Papangeli, I., Simonson, B. et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. Nature 608, 174–180 (2022). https://doi.org/10.1038/s41586-022-04817-8