BIOE249: Computational Functional Genomics Final Project,

Study of the differences in gene expression in Aorta and Heart cells from mice.

Hugo Hakem, Meng BIOE 24' – 12/14/2023.

I Introduction and Objectives

The purpose of this project is to study the differences in gene expression in aorta and heart cells from mice using python. The data used to serve this purpose are single cell transcriptome issued from the Tabula Muris database [1],[2]. In the Tabula Muris database are gathered cells from 20 different organs and tissues. The choice of the Aorta and the Heart is motivated by the will to understand what genetically differentiate those two tissues intrinsically connected. To interpret the results obtained, additional datasets have been downloaded from the Tabula Muris database: Metadata, and Annotations, which gather biological information on the cells.

Different methods of dimension reduction have been tested such as PCA, UMAP and tSNE. The batch correction effect has been studied. Described in [4], the batch effect account for differences induced in single cell analysis due to technical factor such as the humidity in the experiment room, the experimenter, the subject difference and so on. These technical factors must be corrected as it may hide true biological differences. The python tutorial can be found in [5]. The question whether the batch correction must be performed before or after dimension reduction has been answered. Different methods of clustering have been studied among: Phenotype, KMeans and Spectral algorithms. Finally different statistic test has been performed to find the best genes of identified clusters using the 'ttest' option and 'difference' option of the scprep.stats.differential_expression python command.

Overall, this project has been run using code extracted from HW3 and HW4 from the Berkeley BIOE249: Computational Functional Genomics taught by the professor Liana Lareau.

II Methods and Results

II.1 Data presentation and pre-processing

The raw data of aorta and heart cells from the Tabula Muris database [1],[2] are shown in **Figure II.1** after merging those two-database containing respectively 1113 cells and 6003 cells in rows and 23433 genes in columns. The metadata and Annotation DataFrame are not shown but the most important features of those two is that: Metadata contains the plate.barcode (the batch where the cell come from), subtissue of the cells (where does it come from) and that Annotation contains the cell_onthology_class (which kind of cell, the cell is) and tissue (whether the cell comes from the aorta or heart).

	0610005C13Rik	0610007C21Rik	0610007L01Rik	0610007N19Rik	0610007P08Rik	0610007P14Rik
A21.MAA000594.3_8_M.1.1_Aorta	0.0	406.0	714.0	0.0	358.0	0.0
D1.MAA000594.3_8_M.1.1_Aorta	0.0	0.0	0.0	0.0	0.0	0.0
F8.MAA000594.3_8_M.1.1_Aorta	0.0	0.0	1.0	0.0	0.0	0.0
H11.MAA000594.3_8_M.1.1_Aorta	0.0	0.0	0.0	0.0	0.0	0.0
N15.MAA000594.3_8_M.1.1_Aorta	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...
A7.MAA100097.3_39_F.1.1_Heart	0.0	0.0	0.0	0.0	0.0	0.0
D6.MAA100097.3_39_F.1.1_Heart	0.0	147.0	116.0	0.0	0.0	0.0
F1.MAA100097.3_38_F.1.1_Heart	0.0	0.0	0.0	0.0	0.0	0.0
A9.MAA100097.3_39_F.1.1_Heart	0.0	0.0	0.0	0.0	0.0	0.0
D8.MAA100097.3_39_F.1.1_Heart	0.0	100.0	0.0	31.0	2.0	1.0

7115 rows × 23433 columns

Figure II.1.1: Raw dataset of single cell transcriptome of Aorta and Heart, (Cells-rows), (Genes-col).

To analyse those cells, filtering are recquired to remove unrelevant cells and genes and make the remaining cells comparable with each other. Unrelevant cells are qualified as having a too big library which may be due to the fusion of the genetic material of two cells containing in the same droplet during the drop-seq analysis. Or having a too small one, probably indicator of a diying cell. Genes with a too small expression are also removed because too little information is available. Finally the gene count are normalized by library size and transform with a square root to make cells and genes more comparable with each others. Overall, after the data pre-processing, 4268 cells and 17849 genes remain.

II.2 Dimension reduction

In HW3, different dimension reduction methods have been tested: PCA, UMAP fed with PCA, and Phate fed with PCA. It happens that Phate does not add any additional information compared to the two first methods. However, as the tSNE methods looks to be frequently used in the literature, this method has been tested in this project using [3].

Moreover, UMAP on the filtered data have been compared to UMAP fed with PCA, and tSNE on filtered data compared to tSNE fed with PCA and the results are gathered **Figure II.2.1.**

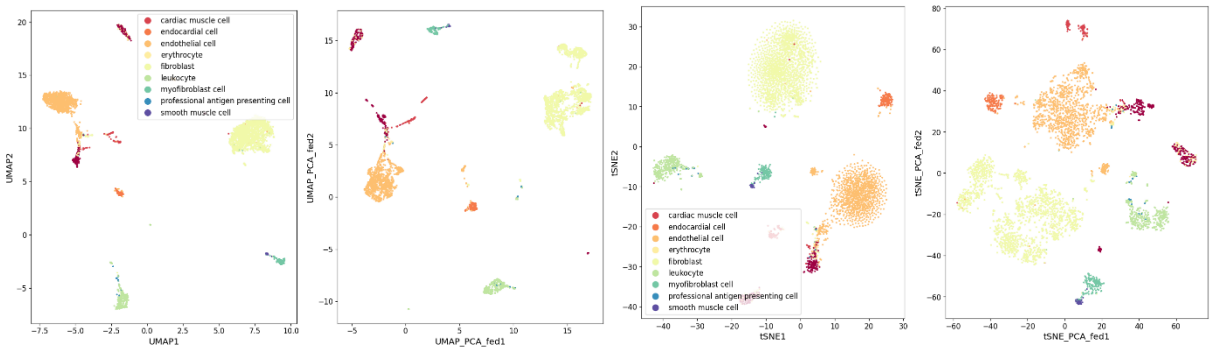

Figure II.2.1: UMAP vs UMAP(PCA fed) and tSNE vs tSNE(PCA fed), colored in cell_onthology_class.

Feeding UMAP or tSNE with PCA happens to be faster to run and provide better cluster as well. Indeed each clusters are a little bit more fragmented, which is not too important to lead to conclude that there are clear differences but it may be relevant enough to identify sub clusters with small biological distinction. In addition, tSNE(PCA fed) is the overall preferred method, as the clusters are more distinguishable and may therefore lead to easier interpretation.

II.3 Batch effect correction

To correct the batch effect, two methods has been tested: either performing the harmonypy [5] algorithm on the data with their dimension already reduced, either doing it on the filtered data and therefore doing the dimension reduction. What it is expected is that the sample from a same batch or with a same plate.barcode originally gathered around same spot get after correction more mixed with each other. Indeed, having a same batch in the same area is not relevant as it enlightens technical factor differences and not biological differences. The results are gathered in **Figure II.2.1**.

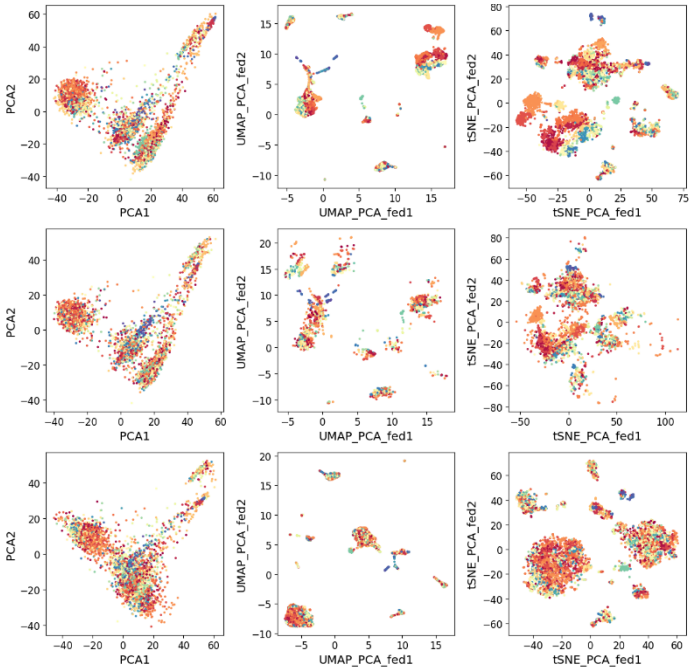

Figure II.3.1: Comparison of different projection techniques, without batch correction (top) with but after reduction (middle) and with but before reduction (bottom), (Coloration in Plate.barcode).

The batch get mixed with each others after the batch correction which proves that the batch correction is effective. The second method, with batch correction on the filtered data and therefore dimension reduction is prefered as the obtained cluster for UMAP_(PCA fed) and tSNE_(PCA fed) are more distinguishable than with the first method.

II.4 Clustering methods

As in HW3 and HW4, Phenotype, KMeans and Spectral algorithms have been run and here again, KMeans has been the preferred method once optimized with a rational choice of number of clusters obtained with the elbow strength curve. Indeed, the number of clusters obtained is not too important like with Phenotype which allows to not over interpret the data, and the cluster are more relevant than with the Spectral algorithm.

II.5 Biological interpretation and Gene analysis of clusters

In **Figure II.5.1** and **Table II.5.1** (interpreted from **Figure II.5.1**) are gathered biological interpretation of clusters obtained through simple coloration of the tSNE (PCA fed) projection according informations from metadata and annotation. The precision of the tSNE (PCA fed) projection can be appreciated as enlighten clusters well match the annotation writtent by the scientist from the Tabula Muris database (see cell_onthology_class).

Two methods have been tested to identify best genes associated to each clusters. The question was whether it is better to use the 'ttest' option or the standard 'difference' option of the scprep.stats.differential_expression python command. It happens that gene screening was better with the 'difference' option as the genes obtained with this method were more specific for each cluster. However, some reserve must be made, this method is not perfect, and even by choosing for each gene which cluster is the most relevant according the 'difference' method, and then selecting the gene for each cluster with the best score, it happens that some gene are not specific enough. To illustrates this, see **Figure II.5.2**.

Cluster	0	1	2	3	4
Cell class	Cardiac muscle cell	//	//	Endothelial cell	Endocardial & endothelial cell
subtissue	Cardiomyocyte	Aorta	Aorta	Left & right ventricule	Mix

5	6	7	8	9
Myofibroblast & smooth muscle cell	Leukocyte	Fibroblast	Fibroblast	Fibroblast
Mix	Mix	Mix of left & right atrium	Right atrium	Left atrium

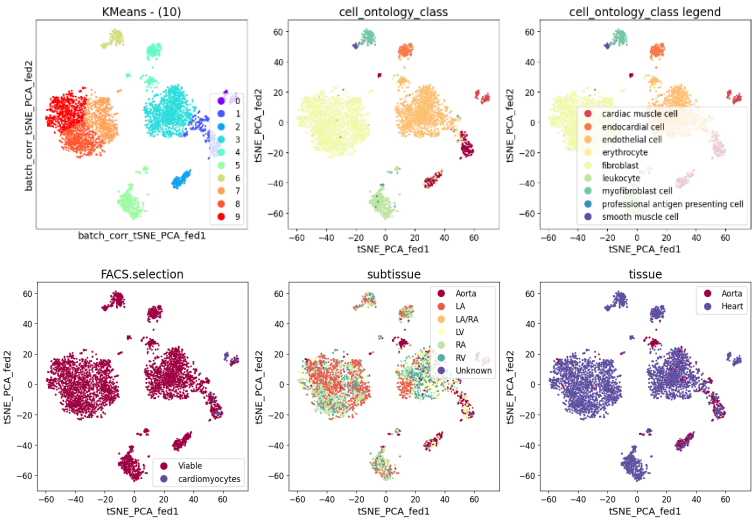

Figure II.5.1 and Table II.5.1: Biological interpretation of the clusters from tSNE (PCA_fed) projection

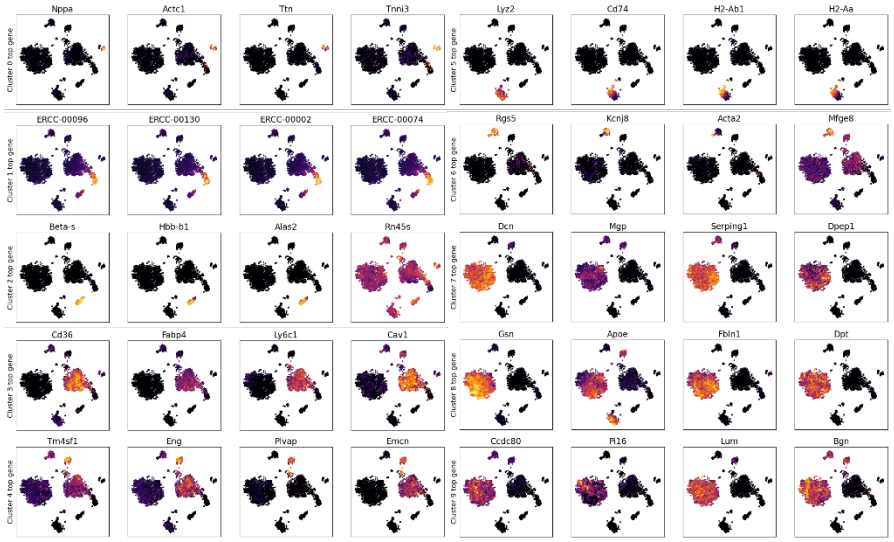

Figure II.5.1: The four genes having the best score with the 'difference' method, for each cluster. (The brighter a dot is, the more impportant the gene expression is for the cell.)

Jitterplot have been performed for some genes to offer a different way of visualisation and can be found in the Notebook.

1.

III Discussion and Conclusion

The goal of this project was to study the difference in term of gene expression in Aorta and Heart cells from mice. Each cluster are specific for a type of cell and different subtissue as seen in **Table II.5.1**. Without detail every cluster, a reason which might explain the similarity of 7,8,9 yet differences, is that all of these clusters indicate fibroblast cells, so all of them are similar but some differences can be observed due to the fact that it comes from different localization in the heart.

Similarly, without detail all of the genes, let's just consider the gene found for cluster 0 associated to cardiac muscle cell / cardiomyocytes and the cluster 2 associated to the Aorta to see if the method of gene identification is relevant. For the cluster 0, a relevant gene identified was Nppa, which happen to be primarily expressed in cardiomyocytes [6]. For the cluster 2, Beta-s, Hbb-b1 and Alas2 gens happen to be hemoglobin gene [7]. As Aorta is the biggest channel for blood, those gene should be relevant for this tissue. Therefore, the literature support the finding and valid the method used.

To conclude, this project was also the opportunity to test different method to perform analysis over single cell transcriptome dataset. The best practice found is following this scheme: Data filtering / Batch effect correction / Dimension reduction with tSNE (PCA fed) / KMeans / relevant gene identification with 'difference' statistical method.

IV References

[1] *Tabula Muris*. https://tabula-muris.ds.czbiohub.org/. Viewed the 12th of december 2023.

[2] *Single-Cell RNA-Seq Data from Smart-Seq2 Sequencing of FACS Sorted Cells (V2)*. figshare, 20 september 2018. */gshare.com*, https://doi.org/10.6084/m9.figshare.5829687.v8. Viewed the 12th of december 2023.

[3] *Krishnaswamy Lab. Github.com*, https://github.com/KrishnaswamyLab/SingleCellWorkshop/blob/master/Visualizing_retinal_bipolar_dataset_revised_notebook.ipynb. Viewed the 12th of december 2023.

[4] « Batch Effect Correction ». 10x Genomics, https://www.10xgenomics.com/resources/analysis-guides/introduction-batch-effect-correction. Viewed the 12th of december 2023.

[5] *Harmonypy. Github.com*, https://github.com/slowkow/harmonypy. Viewed the 12th of december 2023.

[6] Song, Wei, et al. « Atrial Natriuretic Peptide in Cardiovascular Biology and Disease (NPPA) ». *Gene*, vol. 569, nº 1, september 2015, p. 1-6. *PubMed Central*, https://doi.org/10.1016/j.gene.2015.06.029.

[7] Stankiewicz, Adrian M., et al. « Social stress increases expression of hemoglobin genes in mouse prefrontal cortex ». BMC Neuroscience, vol. 15, december 2014, p. 130. PubMed Central, https://doi.org/10.1186/s12868-014-0130-6.