LDATA2010 Information visualization

Report of the project

Prepared by : Group R - Vermeulen Corentin and Rghioui Mehdi
Student number : 25271800 - 29912100
Instructor : Lee John
Date : 16 December 2022

Table des matières

1 Introduction

The goal of this project is to create a decent user interface in order to have a good visualization about the large data set "transcriptomics_data.csv". We implemented some methods to analyse these data and find links between the variables (genes and cell type). Our work is divided in three parts : A basic exploration of the data to have an idea of what is going on, then we have some clustering methods to discover groups of genes, and finally some dimensionality reduction techniques.

2 Basic exploration

2.1 Scatter plot

The first graph is a basic scatter plot where you can freely choose the x and the y axis among all the genes or the cell type. We let an other choice to the user by subsetting the data to some cell types. It is also surrounded by boxplot in order to have an idea of the distribution of the different variables.

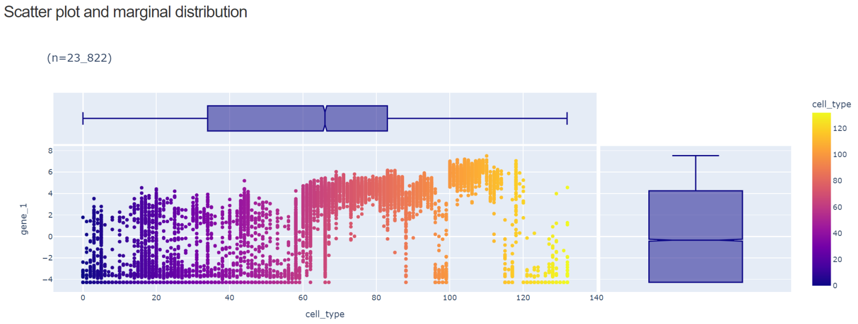
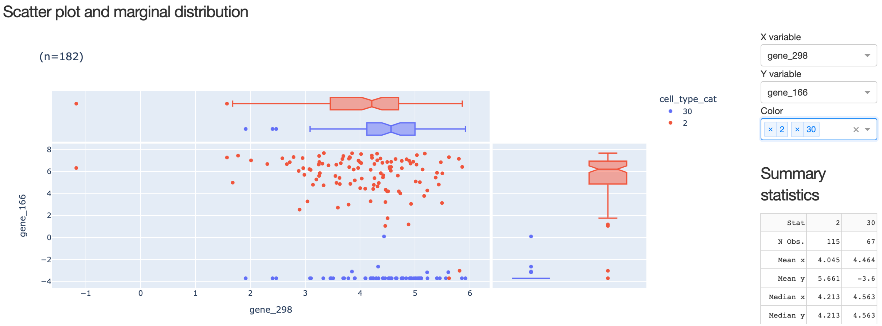


FIGURE 1 – Scatter plot with x as cell_type

When you select `cell_type` as the x variables, we can see the cell_type distribution for the y axis gene as on figure 1.



X variable: gene_298

Y variable: gene_166

Color: × 2 × 30

Summary statistics

Stat	2	30
N Obs.	115	67
Mean x	4.045	4.464
Mean y	5.661	-3.6
Median x	4.213	4.563
Median y	4.213	4.563

FIGURE 2 – Genes values on x and y axis, data subsetted to two cell types

The figure 2 shows how the different gene expression for the selected cell_types.

Moreover, a basic summary statistics table is created, based on the choices made beforehand. We can find in it the count of observations, the mean, median, standard error, minimal and maximal observation for the selected variables.

2.2 Boxplot

This boxplot allows the user to compare the distribution of multiple genes simultaneously as shown on figure 3

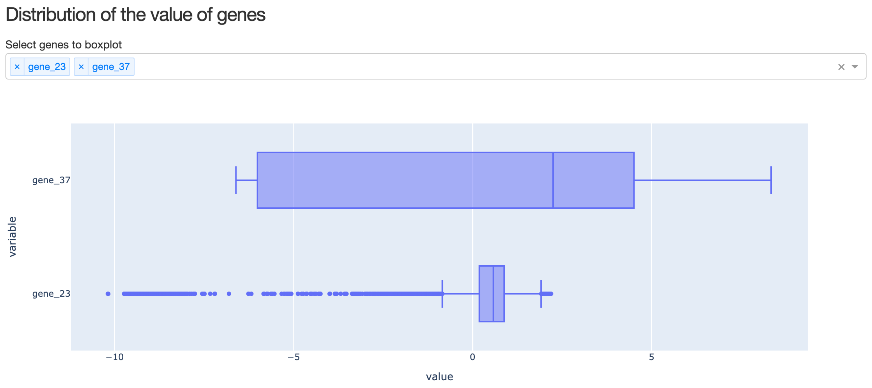
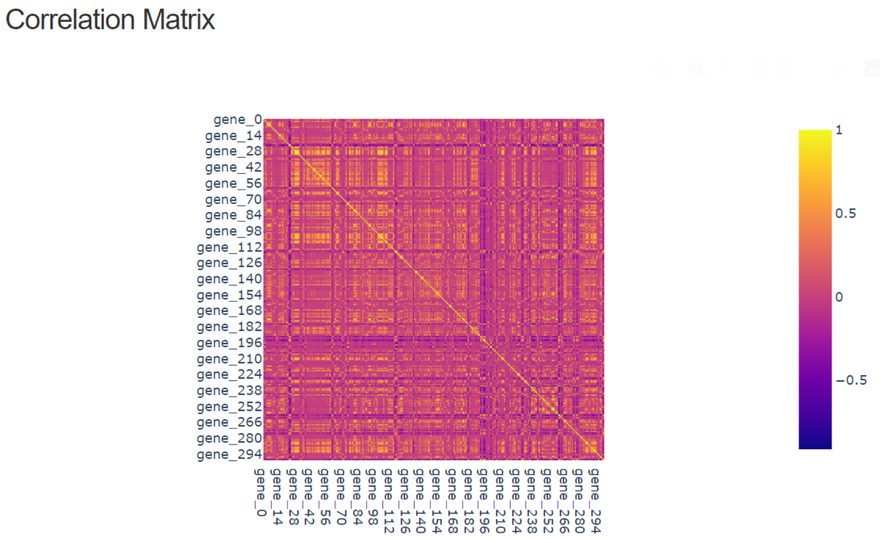


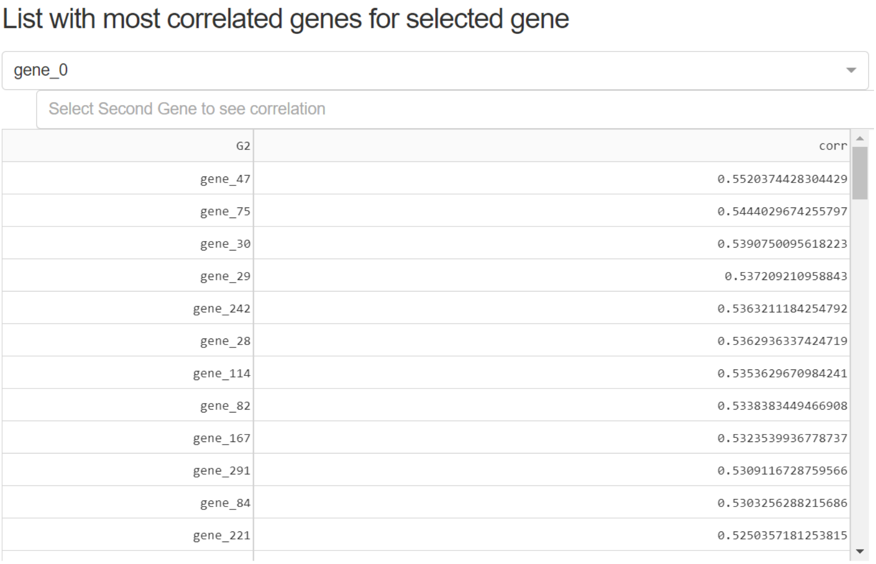
FIGURE 3 – Box plot comparing gene 37 and gene 23 distribution

2.3 Correlation matrix

A correlation matrix is a good way to represent correlations between all the variables. As we have a lot of genes in our data set, we decided to make a heat map to represent this correlation matrix, but even that is not very visible, so don't hesitate to zoom in.



In addition, we provided a reactive table allowing to see the most correlated genes. There are two dropdowns to manage what to display.
If you select 'all' on the first one and nothing in the second one, it will display the most correlated genes by descending order of absolute value.
If you select a gene on the first dropdown and nothing for the second one it will display the most correlated genes by descending order of absolute value for this selected gene.
If you select a gene in each dropdown, it will display the correlation between theses two genes.



List with most correlated genes for selected gene

gene_0

Select Second Gene to see correlation

G2	corr
gene_47	0.5520374428304429
gene_75	0.5444029674255797
gene_30	0.5390750095618223
gene_29	0.537209210958843
gene_242	0.5363211184254792
gene_28	0.5362936337424719
gene_114	0.5353629670984241
gene_82	0.5338383449466908
gene_167	0.5323539936778737
gene_291	0.5309116728759566
gene_84	0.5303256288215686
gene_221	0.5250357181253815

3 Clustering

3.1 Hierarchical clustering

The first method we implemented is the hierarchical clustering. This part aims at helping find the relevant number of cluster in K-Means method. This is why we decided to render theses two figures side by side. Since constructing a dendrogram with plotly takes a very long time we decided to save a static images of the dendrogram and to just display it. This way we avoid the computation time at the cost of reactivity.

3.2 K-means clustering

This part shows the K-means clustering on the PCA principal components. Note that the K-means were computed on the high dimensional data, the PCA only allows to visualize it on two dimensions.

Drag the slider to change K, the number of clusters.

Note : this plot can take some time to update due to computation time, please be patient

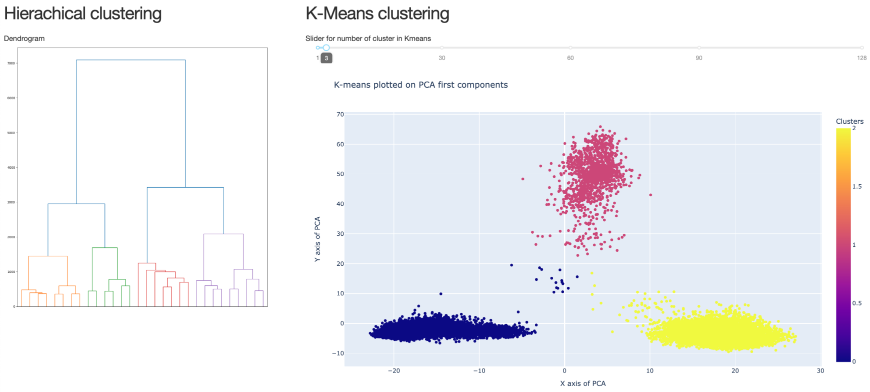
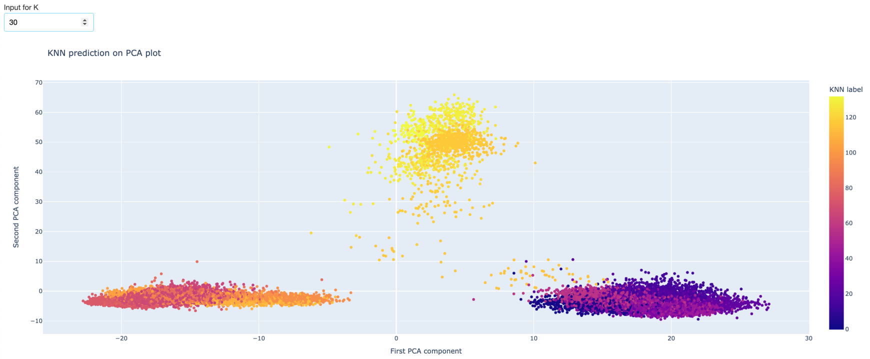


FIGURE 4 – Dendrogram and K-means side by sides

3.3 K-nearest neighbors

We choose to implement the K-nearest neighbors as the density based algorithm because this method is well known and easy to understand. You can change the value of the hyper parameters K. By default it's $\sqrt{n}$.
Note : this plot can take some time to update due to computation time, please be patient


Input for K
30

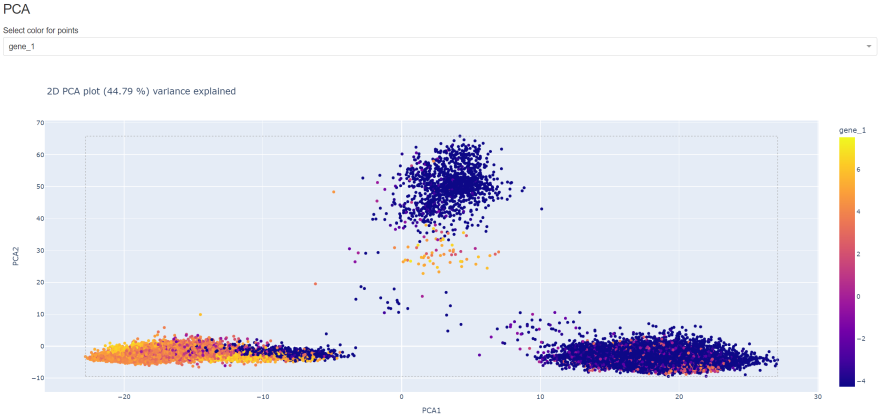
KNN prediction on PCA plot

FIGURE 5 – KNN plots with K=30

4 Dimension reduction

4.1 PCA

For this part, we have implemented two techniques. The first one is the PCA, the principle component analysis. It is a linear method which will represent the data in only two dimensions preserving the distance between points. The user can choose the color representation. Either the cell type or the expression value of a certain gene as shown on figure 6

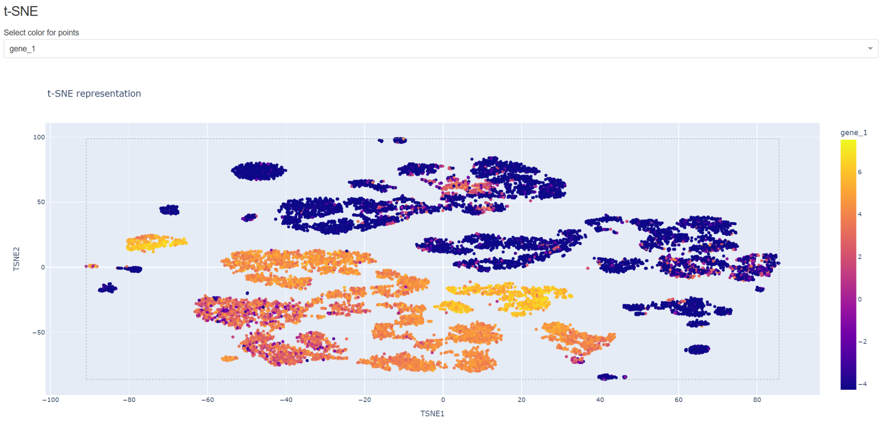

PCA
Select color for points
gene_1

2D PCA plot (44.79 %) variance explained

FIGURE 6 –

4.2 T-SNE

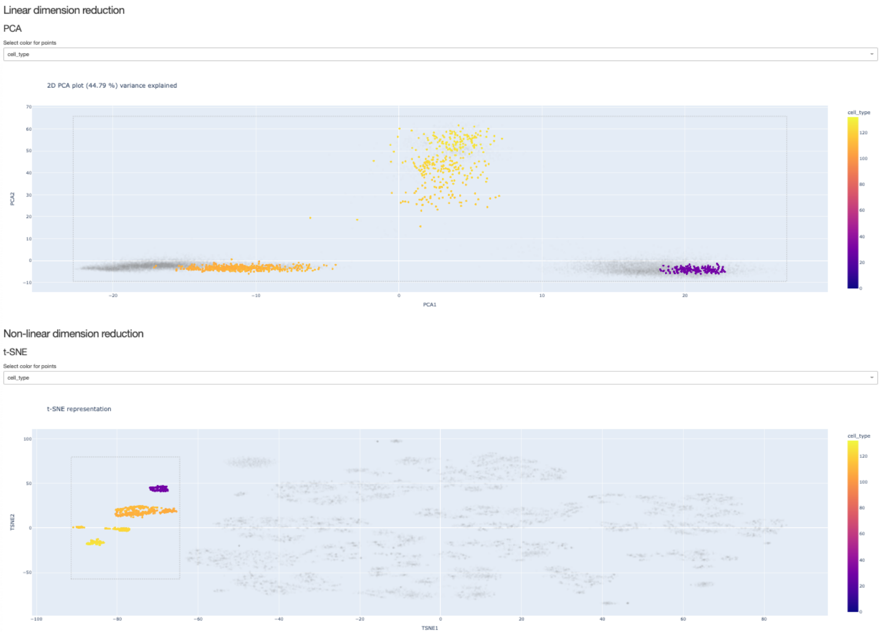
The second technique is T-SNE, the t-distributed stochastic neighbor embedding. It is a non-linear dimension reduction method where the local neighbourhood of the points is preserved. The user can also tune the color of points as for PCA.


t-SNE

Select color for points

gene_1

These graphs are interactive with each others : If we select points in one plot, it will automatically update the second one and show the same observations as shown on figure 4.2


Linear dimension reduction

PCA

Select color for points

cell_type

Non-linear dimension reduction

t-SNE

Select color for points

cell_type

FIGURE 7 – Demo of cross filtering

5 Conclusion

This was a little guide to help you use easily use our data visualization tools. Of course some it can be improved, for example allow to tune the t-SNE parameters, reorder the correlation heatmap, allow to visualize the clustering either on PCA or on t-sne 2D plots to avoid the redundant information between clustering and dimension reduction tabs from this software.