Principal Component Analysis

Dataset DMSOforPCA217

This dataset contains 234 individuals and 22 variables, 9 quantitative variables are considered as illustrative, 4 qualitative variables are considered as illustrative.

### 1. Study of the outliers

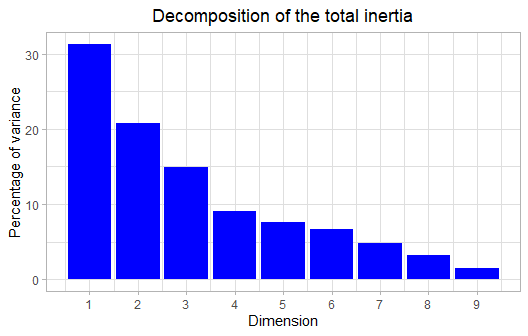
The analysis of the graphs does not detect any outlier.

### 2. Inertia distribution

The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied.

The first two dimensions of analyse express **52.14%** of the total dataset inertia ; that means that 52.14% of the individuals (or variables) cloud total variability is explained by the plane. This percentage is relatively high and thus the first plane well represents the data variability. This value is strongly greater than the reference value that equals **29.35%**, the variability explained by this plane is thus highly significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 11902 data tables of equivalent size on the basis of a normal distribution).

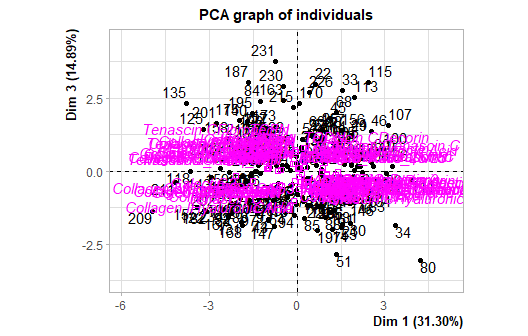
From these observations, it should be better to also interpret the dimensions greater or equal to the third one.



**Figure 2 - Decomposition of the total inertia**

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 3 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (67.03% against 42%). This observation suggests that only these axis are carrying a real information. As a consequence, the description will stand to these axis.

### 3. Description of the plane 1:2

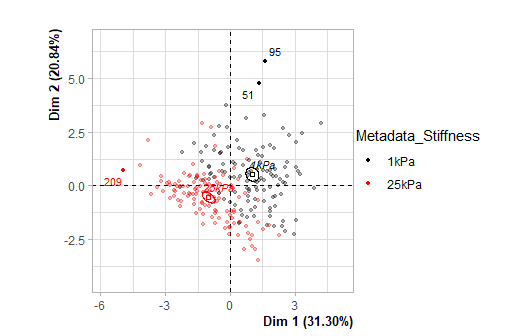


**Figure 3.1 - Individuals factor map (PCA)**

The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

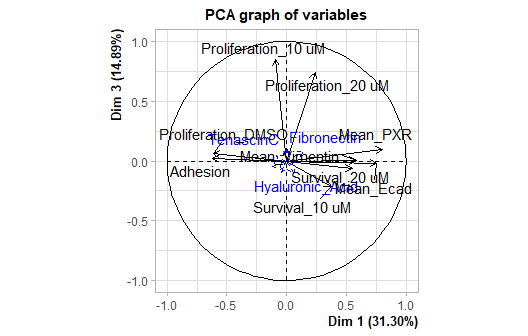
## Metadata\_Stiffness Metadata\_Ratio Metadata\_ECMs Abb\_ECM   
## 2.649745e-37 1.714267e-01 4.748250e-01 4.748250e-01

The best qualitative variable to illustrate the distance between individuals on this plane is : *Metadata\_Stiffness*.

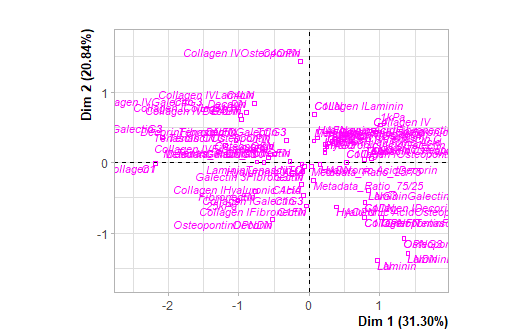


**Figure 3.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.* *The individuals are coloured after their category for the variable* Metadata\_Stiffness.

## Warning: ggrepel: 6 unlabeled data points (too many overlaps). Consider increasing max.overlaps



**Figure 3.3 - Variables factor map (PCA)** *The variables in black are considered as active whereas those in blue are illustrative.* *The labeled variables are those the best shown on the plane.*



**Figure 3.4 - Qualitative factor map (PCA)** *The labeled factors are those the best shown on the plane.*

The **dimension 1** opposes individuals such as *95* and *51* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *209* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group 1 (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Mean\_Vimentin*, *Mean\_PXR*, *Mean\_Ecad*, *Fibronectin* and *Proliferation\_DMSO* (variables are sorted from the strongest).
* low values for the variables *Adhesion*, *Survival\_20.uM*, *CollagenIV*, *Survival\_10.uM* and *Proliferation\_20.uM* (variables are sorted from the weakest).

The group in which the individuals *95* and *51* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Survival\_20.uM*, *Survival\_10.uM*, *Proliferation\_20.uM*, *Mean\_Ecad* and *Mean\_PXR* (variables are sorted from the strongest).
* low values for the variables *Proliferation\_DMSO*, *Decorin* and *Adhesion* (variables are sorted from the weakest).

The group in which the individual *209* stands (characterized by a negative coordinate on the axis) is sharing :

* high values for the variables *Adhesion*, *Proliferation\_DMSO* and *CollagenIV* (variables are sorted from the strongest).
* low values for the variables *Mean\_Ecad*, *Mean\_PXR*, *Mean\_Vimentin*, *Survival\_20.uM*, *Survival\_10.uM*, *Proliferation\_20.uM* and *Laminin* (variables are sorted from the weakest).

Note that the variables *Galectin 3* and *G3* are highly correlated with this dimension (respective correlation of 0.91, 0.91). These variables could therefore summarize themselve the dimension 1.

The **dimension 2** opposes individuals such as *95* and *51* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals characterized by a strongly negative coordinate on the axis (to the bottom of the graph).

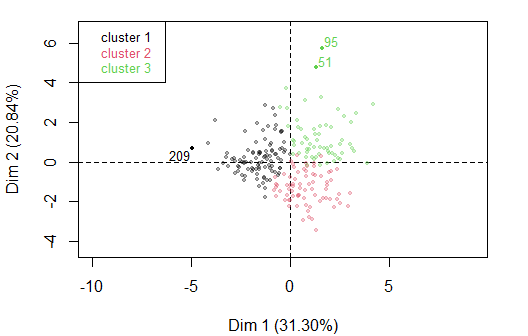
The group in which the individuals *95* and *51* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Survival\_20.uM*, *Survival\_10.uM*, *Proliferation\_20.uM*, *Mean\_Ecad* and *Mean\_PXR* (variables are sorted from the strongest).
* low values for the variables *Proliferation\_DMSO*, *Decorin* and *Adhesion* (variables are sorted from the weakest).

The group 2 (characterized by a negative coordinate on the axis) is sharing :

* high values for the variables *Mean\_Vimentin*, *Mean\_PXR*, *Mean\_Ecad*, *Fibronectin* and *Proliferation\_DMSO* (variables are sorted from the strongest).
* low values for the variables *Adhesion*, *Survival\_20.uM*, *CollagenIV*, *Survival\_10.uM* and *Proliferation\_20.uM* (variables are sorted from the weakest).

### 4. Classification



**Figure 4 - Ascending Hierarchical Classification of the individuals.** *The classification made on individuals reveals 3 clusters.*

The **cluster 1** is made of individuals such as *209*. This group is characterized by :

* high values for the variables *Adhesion*, *Proliferation\_DMSO*, *CollagenIV* and *Proliferation\_10.uM* (variables are sorted from the strongest).
* low values for the variables *Mean\_PXR*, *Mean\_Ecad*, *Mean\_Vimentin*, *Survival\_20.uM*, *Survival\_10.uM* and *Laminin* (variables are sorted from the weakest).

The **cluster 2** is made of individuals sharing :

* high values for the variables *Mean\_Vimentin*, *Mean\_PXR*, *Mean\_Ecad* and *Proliferation\_DMSO* (variables are sorted from the strongest).
* low values for the variables *Adhesion*, *Proliferation\_20.uM*, *Survival\_20.uM*, *CollagenIV*, *Proliferation\_10.uM* and *Survival\_10.uM* (variables are sorted from the weakest).

The **cluster 3** is made of individuals such as *51* and *95*. This group is characterized by :

* high values for the variables *Survival\_20.uM*, *Survival\_10.uM*, *Proliferation\_20.uM*, *Mean\_Ecad*, *Mean\_PXR* and *Osteopontin* (variables are sorted from the strongest).
* low values for the variables *Proliferation\_DMSO*, *Adhesion* and *Decorin* (variables are sorted from the weakest).

## Annexes