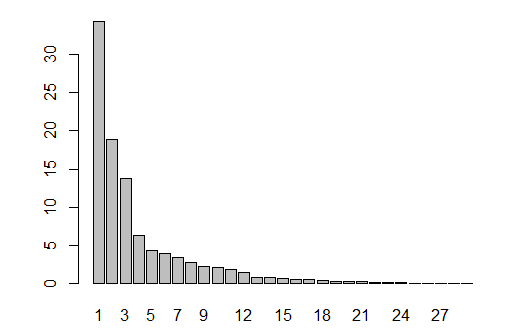
Principal Component Analysis

Dataset Neutrophils

This dataset contains 32 individuals and 39 variables, 1 quantitative variable and nine qualitative variables are considered as supplementary.

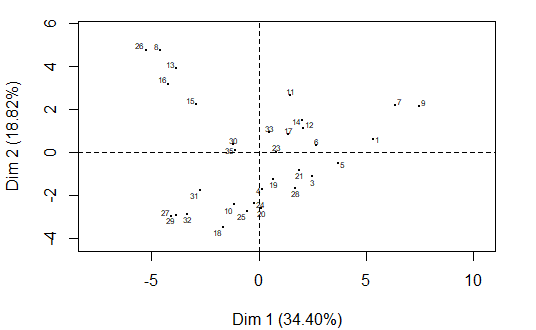
### 1. Inertia distribution

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 PCA express **53.22%** of the total dataset inertia; that means that 53.22% 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 **23.57%**, 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 1234 data tables of equivalent size based on a normal distribution).

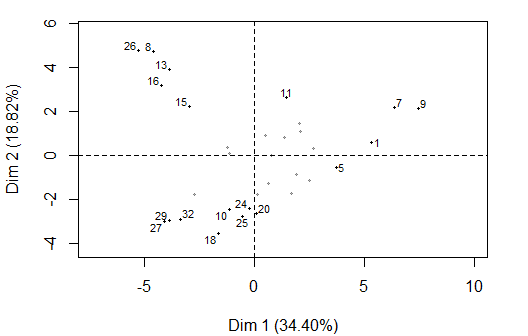
**Figure 2 - Decomposition of the total inertia on the components of the PCA**

An estimation of the right number of axis to interpret suggests restricting the analysis to the description of the first 3 axis. These axes present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (67.02% against 32.88%). This observation suggests that only these axes are carrying a real information. Therefore, the description will stand to these axes.

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



**Figure 3.1 - Individuals factor map (PCA)** *All individuals.*

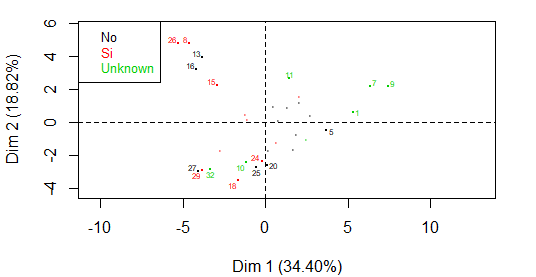


**Figure 3.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

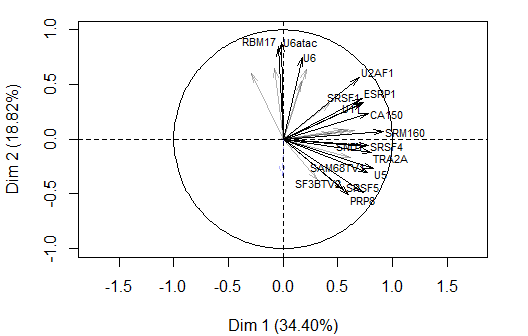
The Wilks test p-value indicates which variable factors explain best the distance between individuals.

ComplicObstet CMIT ACAIgG dsDNA Sexo   
 0.04852218 0.20502182 0.29729638 0.38751730 0.39525821   
 Trombosis Hta B2GPl ACAIgM   
 0.40927895 0.41276073 0.78792775 0.81509379

The best qualitative variable to illustrate the distance between individuals on this plane is *ComplicObstet*.



**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* ComplicObstet.



**Figure 3.3 - Variables factor map (PCA)** *The labeled variables are those the best shown on the plane.*

The **dimension 1** opposes individuals such as *9*, *7*, *1* and *5* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *13*, *26*, *8*, *16* and *15* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *9*, *7*, *1* and *5* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for variables like *U11*, *SRM160*, *TRA2A*, *SRSF1*, *CA150*, *ESRP1*, *U5*, *SRSF4*, *U4* and *SND1* (variables are sorted from the strongest).

The group in which the individuals *13*, *26*, *8*, *16* and *15* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variables *NOVA1*, *SRSF3*, *U6atac*, *RBM17* and *U6* (variables are sorted from the strongest).
* low values for the variables *SF3BTV2*, *SF3BTV1*, *U4*, *SND1*, *SRSF4*, *SAM68TV1*, *PRP8*, *U5* and *SRSF5* (variables are sorted from the weakest).

The **dimension 2** opposes individuals such as *13*, *26*, *8*, *16* and *15* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *27*, *32*, *29* and *10* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

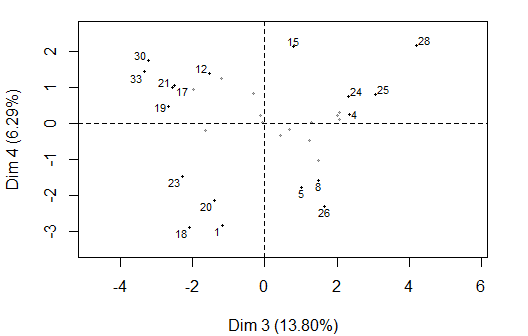
The group in which the individuals *13*, *26*, *8*, *16* and *15* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *NOVA1*, *SRSF3*, *U6atac*, *RBM17* and *U6* (variables are sorted from the strongest).
* low values for the variables *SF3BTV2*, *SF3BTV1*, *U4*, *SND1*, *SRSF4*, *SAM68TV1*, *PRP8*, *U5* and *SRSF5* (variables are sorted from the weakest).

The group in which the individuals *27*, *32*, *29* and *10* stand (characterized by a negative coordinate on the axis) is sharing :

* low values for variables like *U12*, *SRSF1*, *U11*, *ESRP1*, *SRSF10*, *NOVA1*, *U2*, *RBM3*, *RBM17* and *U6atac* (variables are sorted from the weakest).

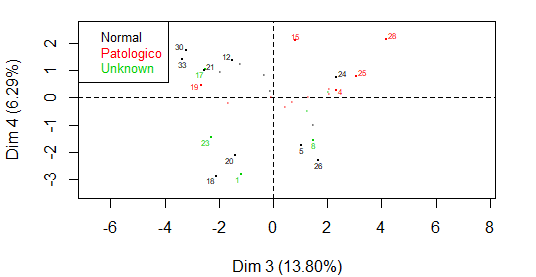
### 3. Description of the dimension 3



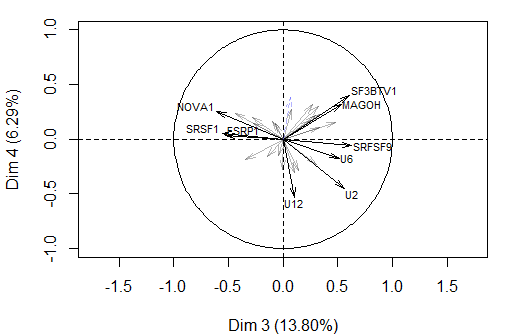
**Figure 4.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

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).

CMIT Trombosis ACAIgM Sexo dsDNA   
 0.05272994 0.28285609 0.48083513 0.62195596 0.70222317   
 ACAIgG Hta ComplicObstet B2GPl   
 0.71195729 0.71508726 0.80407425 0.99330617

The best qualitative variable to illustrate the distance between individuals on this plane is *CMIT*.

**Figure 4.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.* *The individuals are coloured by their categories in the variable* CMIT.



**Figure 4.3 - Variables factor map (PCA)** *The labeled variables are those the best shown on the plane.*

The **dimension 3** opposes individuals such as *28*, *4*, *25* and *24* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *19*, *33*, *30*, *21* and *17* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

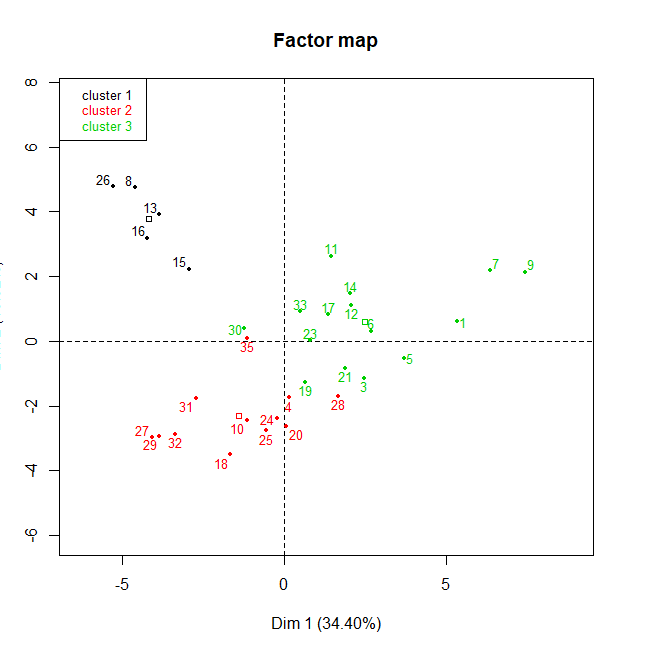
The group in which the individuals *28*, *4*, *25* and *24* stand (characterized by a positive coordinate on the axis) is sharing:

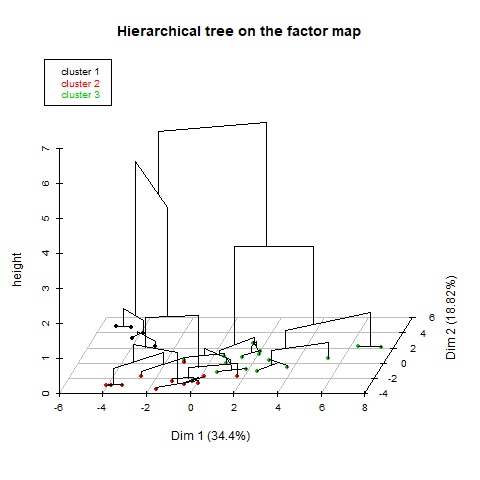
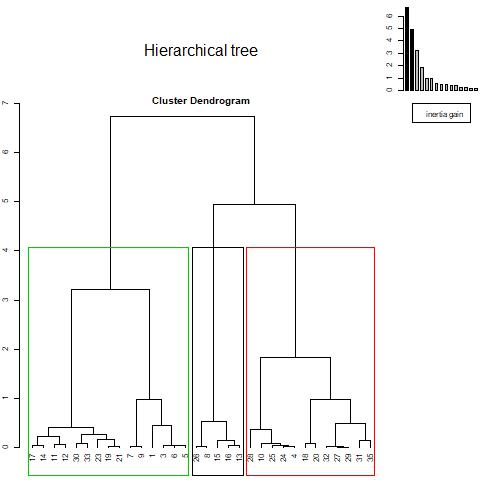
* high values for the variables *SAM68TV1*, *SF3BTV1*, *MAGOH*, *PRP8*, *SRFSF9*, *U2* and *U6* (variables are sorted from the strongest).
* low values for the variables *SRSF4*, *SRM160*, *NOVA1*, *ESRP1*, *SRSF1* and *U11* (variables are sorted from the weakest).

The group in which the individuals *19*, *33*, *30*, *21* and *17* stand (characterized by a negative coordinate on the axis) is sharing:

* high values for the variables *ESRP1*, *SRM160*, *SRSF1*, *NOVA1*, *U2AF1*, *SRSF4*, *U11* and *U6atac* (variables are sorted from the strongest).
* low values for the variable *U2*.

### 5. Classification





**Figure 5 - Hierarchical Classification of the individuals.** *The classification made on individuals reveals 3 possible clusters.*

The **cluster 1** is made of individuals such as *8*, *13*, *15*, *16* and *26*. This group is characterized by:

* high values for the variables *SRSF3*, *RBM17*, *U6*, *U6atac*, *U2* and *NOVA1* (variables are sorted from the strongest).
* low values for the variables *UA2F2*, *TRA2A*, *SF3BTV2*, *SND1*, *SRM160*, *SAM68TV1*, *SRSF4*, *U5*, *PRP8* and *SRSF5* (variables are sorted from the weakest).

The **cluster 2** is made of individuals such as *10*, *18*, *20*, *24*, *25*, *27*, *29* and *32*. This group is characterized by:

* low values for variables like *RBM3*, *RBM17*, *SRM160*, *U6*, *CA150*, *NOVA1*, *U6atac*, *U11*, *SRSF1* and *ESRP1* (variables are sorted from the weakest).

The **cluster 3** is made of individuals such as *1*, *5*, *7*, *9* and *11*. This group is characterized by:

* high values for variables like *SRSF1*, *SRM160*, *ESRP1*, *U11*, *SRSF4*, *U2AF1*, *CA150*, *TRA2A*, *SND1* and *U5* (variables are sorted from the strongest).

- There are no categories that describe the cluster 1.

-The cluster 2 is linked to Hta= Si, the 50% of the individuals belonging to this cluster has Hta, which represents the 66.66% of the population having Hta. Neither individual belonging to cluster 2 has an unknown CMIT. Only the 16% of individuals belonging to this cluster do not have thrombosis, which represents the 13% of the population that do not have thrombosis.

Cla/Mod Mod/Cla Global p.value v.test

Hta=Si 66.66667 50.00000 28.125 0.04900602 1.968539

CMIT=Unknown 0.00000 0.00000 18.750 0.04277239 -2.025926

Trombosis=No 13.33333 16.66667 46.875 0.01074302 -2.550949

-The cluster 3 is linked to Hta= No, the 80% of the individuals belonging to this cluster do not have Hta, which represents the 63.15% of the population that does not has Hta. The 100% of individuals with positive ACAIgG belong to this cluster. The 66.66% of the individuals that do not have thrombosis belong to this cluster.

Cla/Mod Mod/Cla Global p.value v.test

Hta=No 63.15789 80.000000 59.375 0.03350341 2.125998

ACAIgG=Positivo 100.00000 26.666667 12.500 0.03795884 2.075299

Trombosis=No 66.66667 66.666667 46.875 0.04547782 2.000208

Hta=Si 11.11111 6.666667 28.125 0.01473396 -2.438853