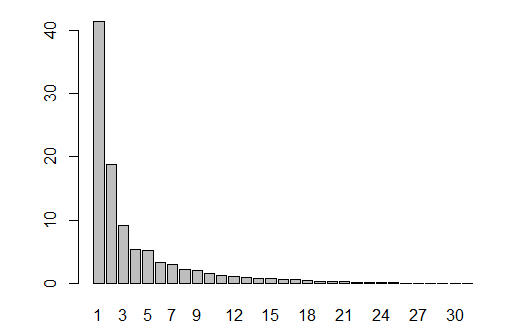
Principal Component Analysis

Dataset Lymphocytes

This dataset contains 32 individuals and 44 variables, one quantitative variable and nine qualitative ones are considered as supplementary.

### 1. Inertia distribution

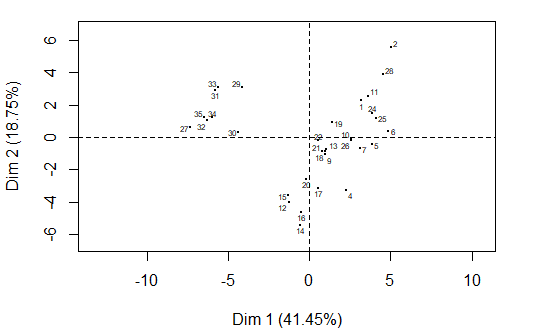
The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied. The first two dimensions of PCA express **60.2%** of the total dataset inertia; that means that 60.2% of the individuals (or variables) cloud total 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 **22.01%**, 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 1211 data tables of equivalent size based on a normal distribution).



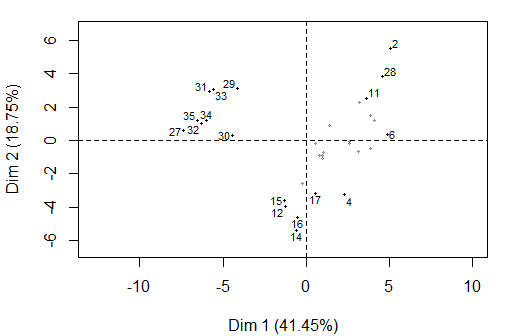
**Figure 2 - Decomposition of the total inertia on the components of the PCA**

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 (69.31% against 30.86%). 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.*

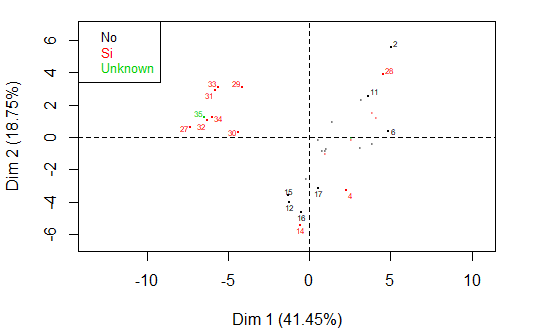


**Figure 3.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

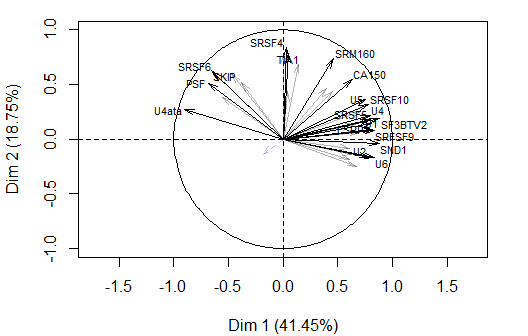
The Wilks test p-value indicates which variable factors explain best the distance between individuals.

Trombosis ComplicObstet ACAIgG ACAIgM CMIT   
 0.09060277 0.13410878 0.18407872 0.35502267 0.40403469   
 Sexo B2GPl dsDNA Hta   
 0.67573120 0.77571890 0.87239885 0.92862078

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

****

**Figure 3.2 - 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* Trombosis.



**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 *28*, *2*, *6* and *11* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *35*, *31*, *27*, *33*, *32*, *34* and *29* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *28*, *2*, *6* and *11* stand (characterized by a positive coordinate on the axis) is sharing:

* high values for variables like *SF3BTV2*, *U4*, *U5*, *SRSF5*, *CA150*, *TRA2A*, *U1*, *ESRP1*, *SRFSF9* and *RBM3* (variables are sorted from the strongest).
* low values for the variables *PSF*, *Edad* and *U4ata* (variables are sorted from the weakest).

The group in which the individuals *35*, *31*, *27*, *33*, *32*, *34* and *29* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variables *PSF*, *U4ata*, *SKIP*, *SRSF6*, *U12*, *U6atac* and *SC35* (variables are sorted from the strongest).
* low values for variables like *RBM3*, *U5*, *U4*, *TRA2A*, *SRSF5*, *SRSF10*, *U11*, *SRFSF9*, *SAM68TV1* and *U2* (variables are sorted from the weakest).

The **dimension 2** opposes individuals such as *28*, *2*, *6* and *11* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *14*, *16*, *12* and *15* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

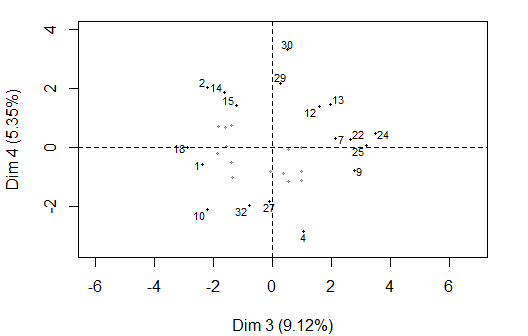
The group in which the individuals *28*, *2*, *6* and *11* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for variables like *SF3BTV2*, *U4*, *U5*, *SRSF5*, *CA150*, *TRA2A*, *U1*, *ESRP1*, *SRFSF9* and *RBM3* (variables are sorted from the strongest).
* low values for the variables *PSF*, *Edad* and *U4ata* (variables are sorted from the weakest).

The group in which the individuals *14*, *16*, *12* and *15* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variable *NOVA1*.
* low values for variables like *TRA2B*, *CA150*, *PSF*, *SKIP*, *U12*, *MAGOH*, *SF3BTV1*, *TIA1*, *U2AF1* and *SRSF6* (variables are sorted from the weakest).

### 3. Description of the dimension 3

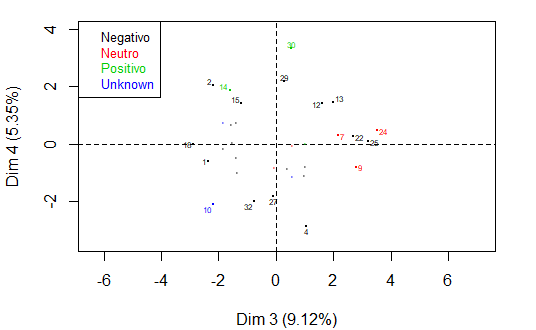


**Figure 4.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

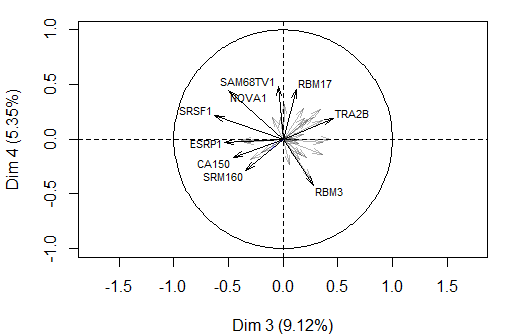
The Wilks test p-value indicates which variable factors explain best the distance between individuals.

ACAIgM Trombosis ACAIgG ComplicObstet B2GPl   
 0.04412756 0.15323873 0.26738479 0.42844807 0.46806339   
 Sexo CMIT dsDNA Hta   
 0.49584211 0.52970164 0.62373683 0.94334213

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

****

**Figure 4.2 - 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* ACAIgM.



**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 *9*, *22*, *13*, *24*, *25* and *7* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *18*, *10* and *1* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

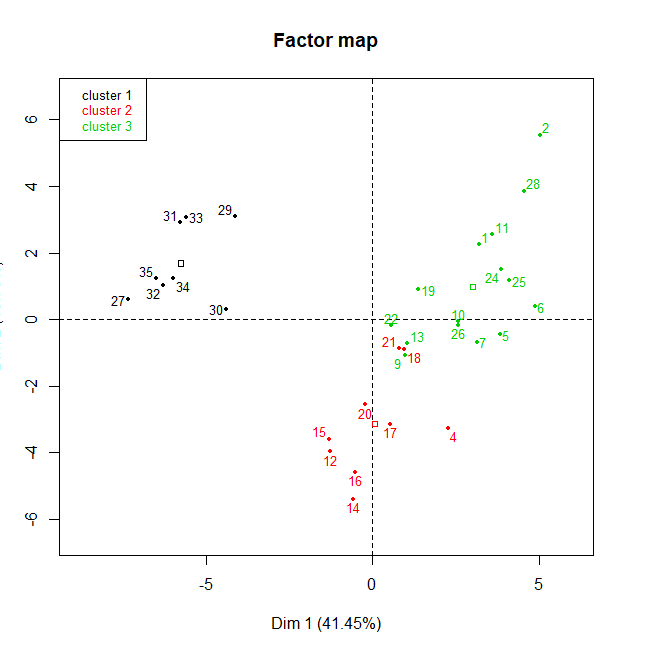
The group in which the individuals *9*, *22*, *13*, *24*, *25* and *7* stand (characterized by a positive coordinate on the axis) is sharing :

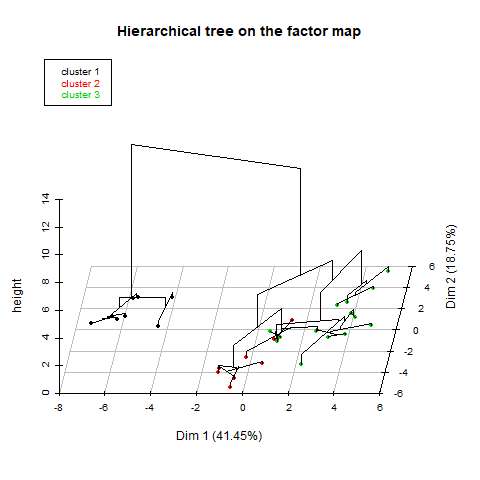
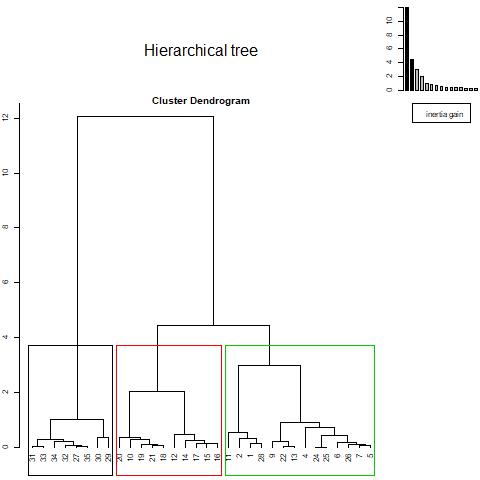
* high values for the variables *SRSF3*, *TRA2B*, *U2*, *SF3BTV2*, *SRSF5*, *SRFSF9* and *SC35* (variables are sorted from the strongest).
* low values for the variables *SRSF4*, *CA150* and *SRSF1* (variables are sorted from the weakest).

The group in which the individuals *18*, *10* and *1* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variables *ESRP1*, *SRSF1*, *CA150*, *SRM160*, *SRSF4*, *NOVA1* and *U2AF1* (variables are sorted from the strongest).

### 4. Classification





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

The **cluster 1** is made of individuals such as *27*, *29*, *30*, *31*, *32*, *33*, *34* and *35*. This group is characterized by:

* high values for the variables *U4ata*, *SRSF6*, *PSF*, *SKIP*, *U12*, *U6atac* and *SC35* (variables are sorted from the strongest).
* low values for variables like *CA150*, *TRA2A*, *RBM17*, *RBM3*, *U5*, *SRSF10*, *U4*, *U1*, *SAM68TV1* and *SRSF5* (variables are sorted from the weakest).

The **cluster 2** is made of individuals such as *4*, *12*, *14*, *15*, *16* and *17*. This group is characterized by:

* high values for the variable *NOVA1*.
* low values for variables like *SRSF4*, *TRA2B*, *SC35*, *SRM160*, *MAGOH*, *PSF*, *U2AF1*, *SKIP*, *SRSF6* and *SF3BTV1* (variables are sorted from the weakest).

The **cluster 3** is made of individuals such as *2*, *6*, *11* and *28*. This group is characterized by:

* high values for variables like *SF3BTV2*, *U4*, *SRSF5*, *SRFSF9*, *U5*, *RBM3*, *U1*, *SRSF10*, *TRA2A* and *SND1* (variables are sorted from the strongest).
* low values for the variables *U6atac* and *U4ata* (variables are sorted from the weakest).

- The cluster 1 is linked to the category Thrombosis=Si. The 87.5 % of the individuals belonging to this cluster has thrombosis, which represents the 50% of the population with thrombosis.

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

Trombosis=Si 50 87.5 43.75 0.006444197 2.724279

Trombosis=No 0 0.0 50.00 0.001223582 -3.233325

- The cluster 2 is not described by any category.

- The 40% of the individuals belonging to cluster 3 has an unknown ComplicObset, which represents the 85% of the population which has this feature.

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

ComplicObstet=Unknown 85.71429 40 21.875 0.02910255 2.182095