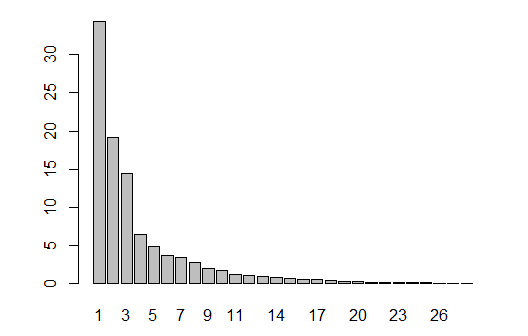
Principal Component Analysis

Dataset Monocytes

This dataset contains 29 individuals and 50 variables, one quantitative variable and nine qualitative ones are considered as supplementary.

### 1. Inertia distribution

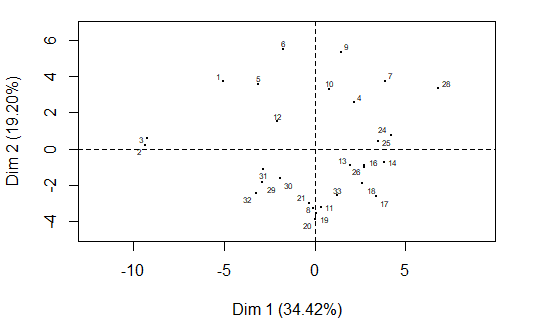
The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied. The first two dimensions of PCA express **53.62%** of the total dataset inertia; that means that 53.62% of the individuals (or variables) cloud total 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 **21.4%**, 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 1160 data tables of equivalent size based on a normal distribution).



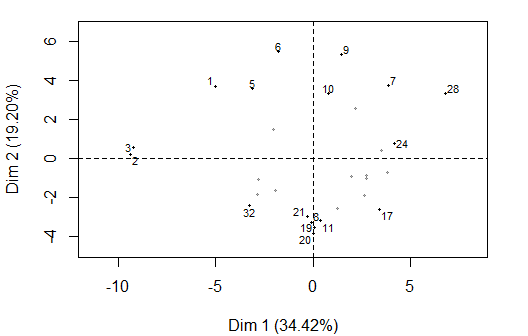
**Figure 2 - Decomposition of the total inertia by the PCA components.**

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 (68.08% against 30.04%). 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.*

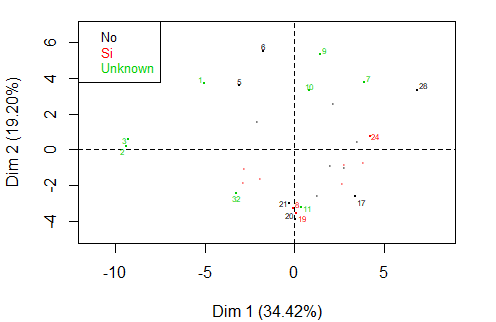


**Figure 3.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

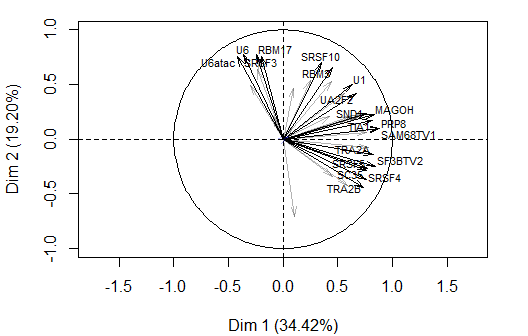
The Wilks test p-value indicates which variable factors explain best the distance between individuals.

ComplicObstet CMIT Hta Sexo dsDNA   
 0.03143352 0.21550352 0.27632881 0.42358915 0.47115210   
 ACAIgM Trombosis B2GPl ACAIgG   
 0.57423251 0.84312541 0.91848781 0.95833763

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 colored by their categories in 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 *28*, *9*, *7* and *24* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *3* and *2* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *28*, *9*, *7* and *24* stand (characterized by a positive coordinate on the axis) is sharing:

* high values for variables like *UA2F2*, *SF3BTV1*, *SRFSF9*, *TIA1*, *U1*, *RBM22*, *PTB*, *PRP8*, *SAM68TV1* and *SF3BTV2* (variables are sorted from the strongest).

The group in which the individuals *3* and *2* stand (characterized by a negative coordinate on the axis) is sharing:

* low values for variables like *SAM68TV1*, *PTB*, *PRP8*, *SRSF6*, *CUGBP*, *SRSF10*, *RBM3*, *SF3BTV1*, *SRSF4* and *MAGOH* (variables are sorted from the weakest).

The **dimension 2** opposes individuals such as *5*, *1* and *6* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *20* and *21* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

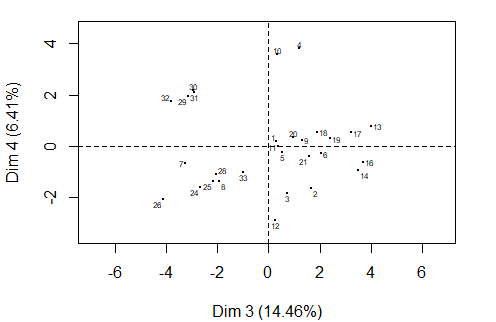
The group in which the individuals *5*, *1* and *6* stand (characterized by a positive coordinate on the axis) is sharing:

* high values for the variables *U2*, *U4* and *ESRP1* (variables are sorted from the strongest).

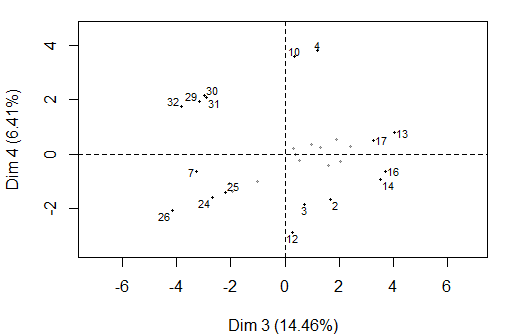
The group in which the individuals *20* and *21* stand (characterized by a negative coordinate on the axis) is sharing:

* high values for the variables *U4ata*, *SRSF1* and *PSF* (variables are sorted from the strongest).
* low values for the variables *U2AF1*, *U6*, *U12*, *U4*, *RBM17* and *SRSF3* (variables are sorted from the weakest).

### 3. Description of the dimension 3



**Figure 4.1 - Individuals factor map (PCA)** *All individuals.*

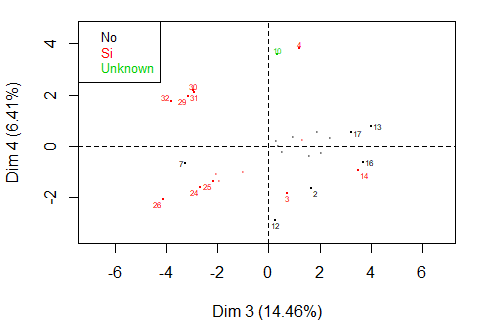


**Figure 4.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

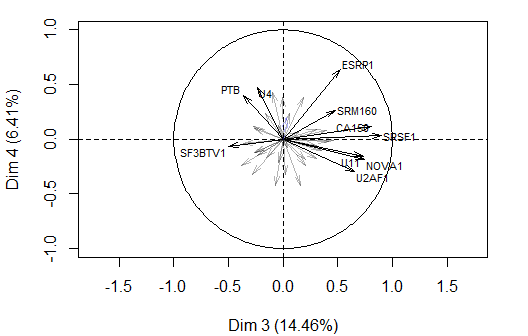
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

Trombosis B2GPl Hta ACAIgM dsDNA   
 0.004548155 0.299820286 0.303787210 0.326714008 0.354182811   
ComplicObstet CMIT ACAIgG Sexo   
 0.361776452 0.430936074 0.759197278 0.899732177

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



**Figure 4.2 - Individuals factor map (PCA)** *The individuals are colored by their categories in the variable* Trombosis.



**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 *13*, *16* and *14* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *30*, *26*, *32*, *29*, *31*, *24* and *25* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *13*, *16* and *14* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for variables like *SF3BTV2*, *SRSF4*, *U11*, *U5*, *SRSF1*, *U2AF1*, *TRA2A*, *SRM160*, *SRSF5* and *CA150* (variables are sorted from the strongest).

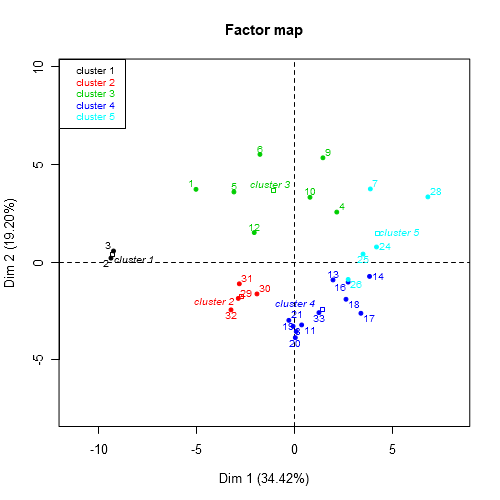
The group in which the individuals *30*, *32*, *29* and *31* stand (characterized by a negative coordinate on the axis) is sharing:

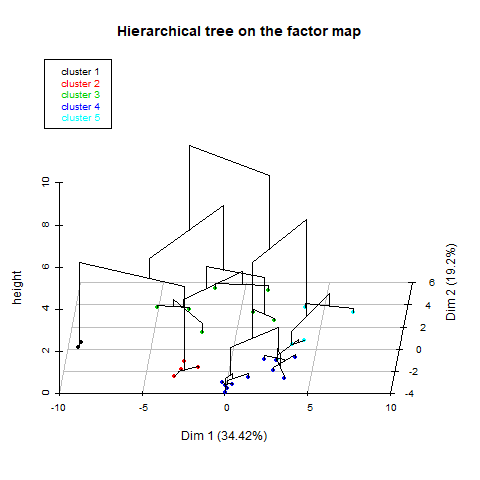
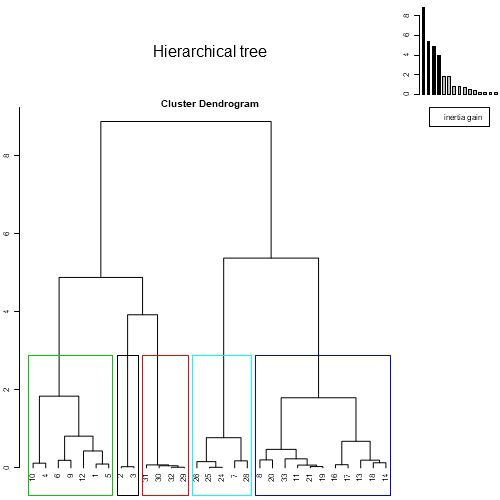
* high values for the variable *U4ata*.
* low values for variables like *NOVA1*, *CUGBP*, *U1*, *SRSF4*, *SKIP*, *SRSF1*, *SRM160*, *U5*, *PSF* and *SRFSF9* (variables are sorted from the weakest).

The group in which the individuals *26*, *24* and *25* stands (characterized by a negative coordinate on the axis) is sharing:

* high values for the variables *SF3BTV1*, *SRSF6*, *SAM68TV1*, *TIA1*, *RBM3*, *U1*, *MAGOH*, *SC35* and *PSF* (variables are sorted from the strongest).
* low values for the variables *ESRP1*, *U6atac* and *U4ata* (variables are sorted from the weakest).

### 5. Classification





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

The **cluster 1** is made of individuals such as *2* and *3*. This group is characterized by:

* high values for the variables *RBM17*, *U6atac*, *NOVA1*, *U6* and *SRSF3* (variables are sorted from the strongest).
* low values for variables like *CUGBP*, *TIA1*, *TRA2B*, *RBM45*, *MAGOH*, *RBM3*, *SC35*, *SRSF6*, *SRM160* and *SF3BTV1* (variables are sorted from the weakest).

The **cluster 2** is made of individuals such as *32*. This group is characterized by:

* high values for the variable *U4ata*.
* low values for variables like *SRM160*, *TRA2A*, *SKIP*, *RBM17*, *U5*, *SRSF10*, *PSF*, *RBM45*, *NOVA1* and *MAGOH* (variables are sorted from the weakest).

The **cluster 3** is made of individuals such as *1*, *5*, *6*, *9* and *10*. This group is characterized by:

* high values for variables like *U6atac*, *U6*, *U2*, *SRSF10*, *U4*, *RBM17*, *NOVA1*, *CUGBP*, *ESRP1* and *RBM3* (variables are sorted from the strongest).
* low values for the variables *SKIP*, *SRSF5*, *SC35*, *U4ata* and *TRA2B* (variables are sorted from the weakest).

The **cluster 4** is made of individuals such as *8*, *11*, *17*, *19*, *20* and *21*. This group is characterized by:

* high values for variables like *SKIP*, *SRSF4*, *SF3BTV2*, *SRSF1*, *TRA2B*, *U5*, *TRA2A*, *SRSF5*, *U4ata* and *PSF* (variables are sorted from the strongest).
* low values for the variables *U6*, *U12*, *U6atac*, *RBM17*, *U2*, *SRSF3* and *U4* (variables are sorted from the weakest).

The **cluster 5** is made of individuals such as *7*, *24* and *28*. This group is characterized by:

* high values for variables like *SF3BTV1*, *TIA1*, *SAM68TV1*, *SRSF6*, *MAGOH*, *RBM22*, *SRFSF9*, *U1*, *PRP8* and *PTB* (variables are sorted from the strongest).
* low values for the variables *U4ata*, *SRSF1* and *ESRP1* (variables are sorted from the weakest).

- The cluster 1 is linked to the category CMIT=Unknown. The 100% of the individuals belonging to this cluster has an unknown CMIT, which represents the 33.33% of the population having an unknown CMIT.

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

CMIT=Unknown 33.33333 100 20.68966 0.03694581 2.086362

- The cluster 2 is linked to the category Thrombosis=Si. The 100% of the individuals belonging to this cluster has thrombosis, which represents the 28.57% of the population having thrombosis.

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

Trombosis=Si 28.57143 100 48.27586 0.04214559 2.03208

- Neither individual belonging to the cluster 3 presents a ComplicObset.

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

ComplicObstet=Si 0 0 31.03448 0.04966747 -1.962817

- There are not categories describing the clusters 4 and 5.