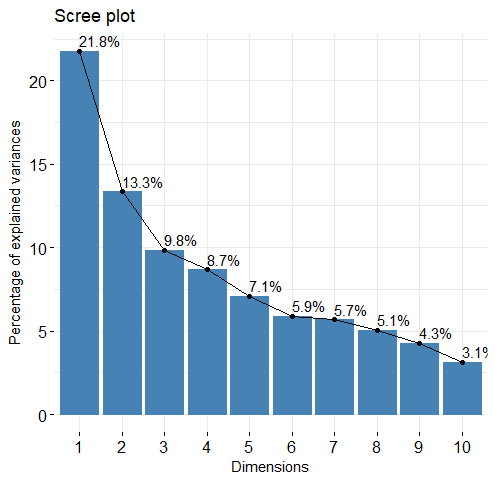
Principal Component Analysis

The dataset contains 42 individuals and 34 variables, 9 quantitative variables and 5 qualitative ones are considered as supplementary. The PCA is constructed over the genetic factors, and the rest of variables are used to describe the data.

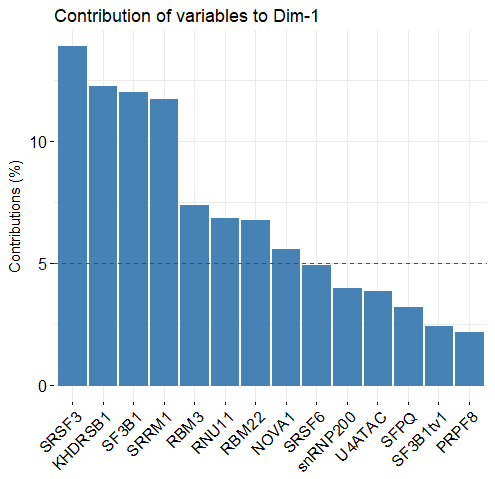
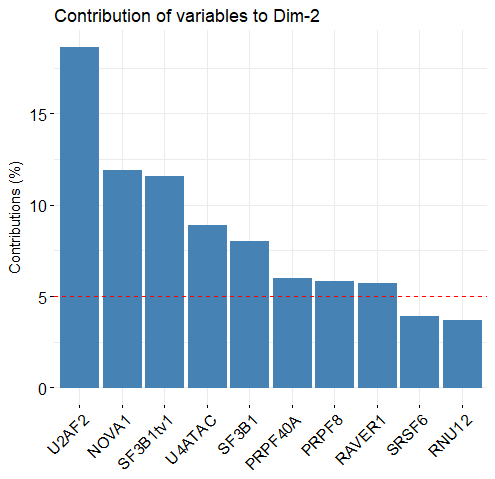
### 1. Inertia distribution

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 **35.11%** of the total inertia; that means that 35.11% of the individuals (or variables) cloud total variability is explained by the plane. This is an intermediate percentage and the first plane represents a part of the data variability. This value is greater than the reference value that equals **25.55%**, the variability explained by this plane is thus significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 1214 data tables of equivalent size based on a normal distribution). This observation suggests that only these axes are carrying a real information. As a consequence, the description will stand to these axes.

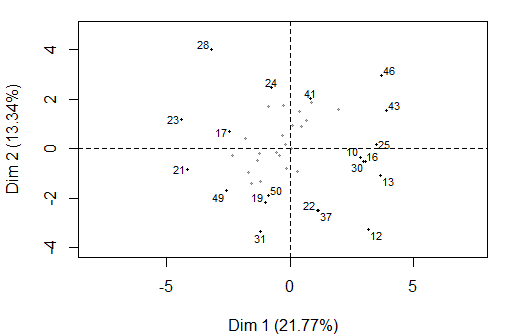


**Figure 2 - Decomposition of the total inertia on the PCA components**

The following figures show the contribution of the variables to each dimension.

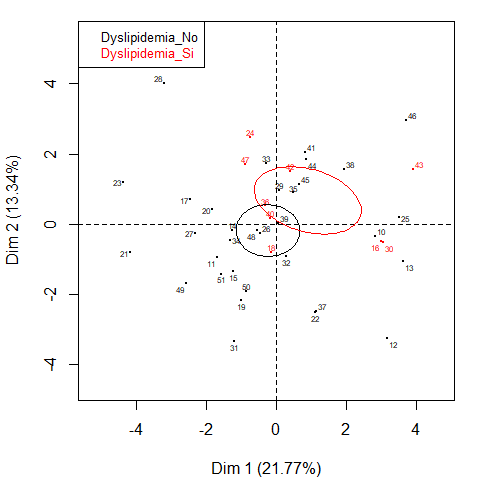
### 2. Description of the plane 1:2



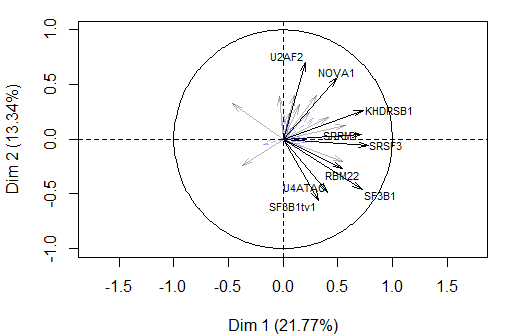
**Figure 3.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

The Wilks test p-value indicates which variable factors explain best the distance between individuals. The best qualitative variable to illustrate the distance between individuals on this plane is *Dyslipidemia*.

Dyslipidemia Hbpressure Diabetes PerineuralInv ProstExt   
 0.1375136 0.1574256 0.7245308 0.8830162 0.9164418

****

**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 category for the variable* Dyslipidemia.



**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 *46*, *12*, *43*, *25*, *13*, *30* and *10* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *23*, *28*, *21* and *24* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *46*, *12*, *43*, *25*, *13*, *30* and *10* stand (characterized by a positive coordinate on the axis) is characterized by:

* high values for the variables *KHDRSB1*, *SRSF3*, *SRRM1*, *SF3B1*, *NOVA1*, *RBM3*, *snRNP200*, *sst5TMD4exp* and *RNU11*.
* low values for the variable *SFPQ*.

According to the preliminary study regarding the correlations between variables (see Section …), it is confirmed that the clinical factor *sst5TMD4exp* is positive correlated to *KHDRSB1, SRSF3, SRRM1, NOVA1, RBM3, snRNP200 and RNU11, and it is negatively correlated to SFPQ.*

The group in which the individuals *23*, *28*, *21* and *24* stand (characterized by a negative coordinate on the axis) is characterized by:

* high values for the variable *SRSF6*.
* low values for the variables *SRRM1*, *RBM22*, *RNU11*, *SRSF3*, *U4ATAC*, *SF3B1tv1* and *SF3B1*.

The **dimension 2** opposes individuals such as *23*, *28*, *21* and *24* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *49*, *31*, *22*, *19*, *37* and *50* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

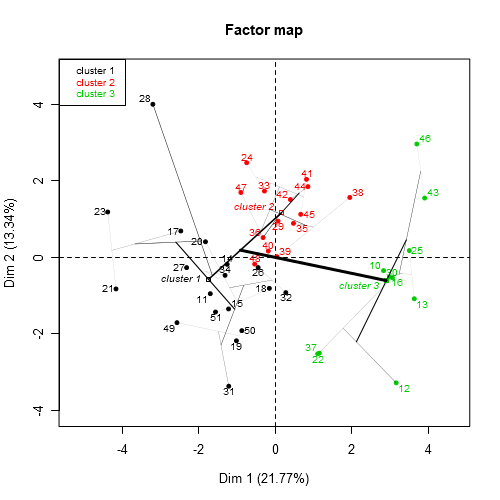
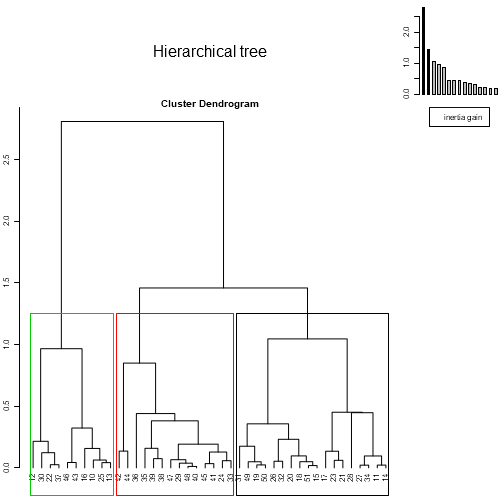
The group in which the individuals *23*, *28*, *21* and *24* stand (characterized by a positive coordinate on the axis) is characterized by:

* high values for the variable *SRSF6*.
* low values for the variables *SRRM1*, *RBM22*, *RNU11*, *SRSF3*, *U4ATAC*, *SF3B1tv1* and *SF3B1*.

The group in which the individuals *49*, *31*, *22*, *19*, *37* and *50* stand (characterized by a negative coordinate on the axis) is characterized by:

* low values for the variables *NOVA1* and *U2AF2* (variables are sorted from the weakest).

### 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 *17*, *19*, *21*, *23*, *31*, *49* and *50*. This group is characterized by:

* high values for the variable *Arexp*.
* low values for the variables *RAVER1*, *SRRM4*, *RNU11*, *SRRM1*, *RBM3*, *SRSF3*, *snRNP200*, *NOVA1*, *U2AF2* and *KHDRSB1*.

According to the preliminary study regarding the correlations between variables (see Section …), it is confirmed that the clinical factor *Arexp* is negatively correlated to *SRRM4, RNU11, SRSF3, snRNP200, U2AF2 and KHDRSB1.*

The **cluster 2** is made of individuals such as *24*, *28* and *41*. This group is characterized by:

* high values for the variables *U2AF2*, *snRNP200*, *PSA*, *RNU12* and *RAVER1*.
* low values for the variables *SF3B1* and *SF3B1tv1*.

According to the preliminary study regarding the correlations between variables (see Section …), it is confirmed that the clinical factor *PSA* is positively correlated to *RNU12, snRNP200 and U2AF2 .*

The **cluster 3** is made of individuals such as *10*, *12*, *13*, *16*, *22*, *25*, *30*, *37*, *43* and *46*. This group is characterized by:

* high values for the variables *SF3B1*, *SRSF3*, *SRRM1*, *RBM22*, *KHDRSB1*, *U4ATAC*, *SF3B1tv1*, *RNU11*, *RBM3* and *NOVA1*.

Cla/Mod Mod/Cla Global p.value  
Dyslipidemia=Dyslipidemia\_No 51.51515 94.444444 78.57143 0.03555456  
Dyslipidemia=Dyslipidemia\_Si 11.11111 5.555556 21.42857 0.03555456  
 v.test  
Dyslipidemia=Dyslipidemia\_No 2.101985  
Dyslipidemia=Dyslipidemia\_Si -2.101985