1- Correlation between numeric variables

The Pearson's correlation coefficient was used to determine the correlation between two quantitative variables. The following table shows the small, medium and large correlations between the genetic factors and numeric clinical factors. The genetic factors are ordered from higher to lower correlation values. We can see that large correlations were not detected, except in the case of PCA3exp with RNU11. All scatter plots can be consulted in the folder "*Correlations*". The scatter plots did not show good relations…

Clinical factor	**Correlation**		
	Small (.1 to .3 , -0.1 to -0.3)	**Medium (.3 to .5 -0.3 to -0.5)**	**Large (.5 to 1.0 -0.5 to -1.0)**
BMI	Positive: RBM22, SRSF3,PRPF8, RBM3, SF3B1,PRPF40A, NOVA1,RAVER1, SRRM1 Negative: U2AF2, snRNP200, MAGOH, U4ATAC, RNU12	Positive: SF3B1tv1 Negative: SFPQ	
Age	Positive: SRRM1, RNU11, SRSF3, SF3B1tv1 Negative: RAVER1, SRSF6, RBM3	Positive: RNU12, SF3B1	
Gleason score	Positive: SRRM4, U4ATAC, RAVER1, snRNP200, MAGOH, KHDRSB1, PRPF40A, RBM3 Negative: SF3B1, SF3B1tv1, RNU12		
PSA	Positive: snRNP200, U2AF2, RBM3, SRSF6 Negative: U4ATAC	Positive: RNU12,	
PSAexp	Positive: SFPQ, RBM3, SF3B1, U4ATAC, KHDRSB1,	Positive: NOVA1, SRRM1, SRSF3	

	PRPF8 Negative: PRPF40A, RAVER1	Negative: SRRM4	
PCA3 exp	Positive: RNU12, SRSF6, U2AF2, KHDRSB1, PRPF8 Negative: PRPF40A, SRRM4, SFPQ	Positive: SF3B1tv1, U4ATAC, snRNP200	Positive: RNU11
sst5TMD4exp	Positive: KHDRSB1, RNU12, SRRM1, NOVA1, RBM22, SRSF3, RBM3 Negative: PRPF40A SFPQ	Positive: RNU11, SRRM4, PRPF8, U2AF2, snRNP200, U4ATAC	
In1Ghrelinexp	Positive: SFPQ, RNU12, SRRM4, PRPF40A, MAGOH, KHDRSB1 Negative: SF3B1tv1, snRNP200, RNU11, U2AF2, SRSF6, RBM22	Positive: SRSF3, RBM3,	
Arexp	Negative: SRRM4,SRSF3, MAGOH,U2AF2, RNU11, KHDRSB1,SF3B1, snRNP200, PRPF40A	Negative: U4ATAC	

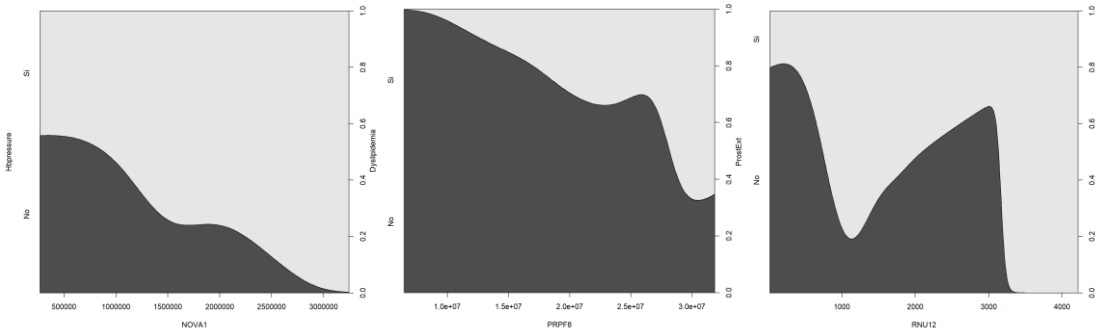
2- Association between numeric variables and nominal ones

Eta squared is a measure of effect size, it represents the proportion of variance in Y explained by X, and it can detect non-linear correlations. The following table shows the small, medium and large correlations between the genetic factors and nominal clinical factors. The genetic factors are ordered from higher to lower correlation values. No large correlations were found.

Clinical factor	**Correlations**		
	Small (0.02-0.13)	**Medium (0.13-0.26)**	**Large (>0.26)**
HB pressure	PRPF40A, SRRM4,	NOVA1	

	KHDRSB1, U4ATAC, RBM3, RBM22, SFPQ, RNU11		
Diabetes	U4ATAC, SRSF3, RBM3, SF3B1tv1, PRPF8, SRRM1		
Dyslipidemia	KHDRSB1, NOVA1, RBM3, snRNP200, RNU11, MAGOH, RAVER1, SRRM4,	PRPF8	
ProstExt	RBM22, SFPQ, SRRM4, RNU11, PRPF40A, SRSF6, RBM3, SRRM1	RNU12	
PerineuralInv	PRPF40A, U4ATAC, RBM22, RNU11, RNU12, MAGOH		

The following conditional density plots show the relationship between the genetic factors and clinical ones. We only show the medium correlations; the rest of graphs are in the folder "*Correlations*". The results show that patients with higher values in the factor NOVA1 are more likely to present an HB pressure. Also, patients with higher values in the factor RNU12 are more likely to have a prost. ext.

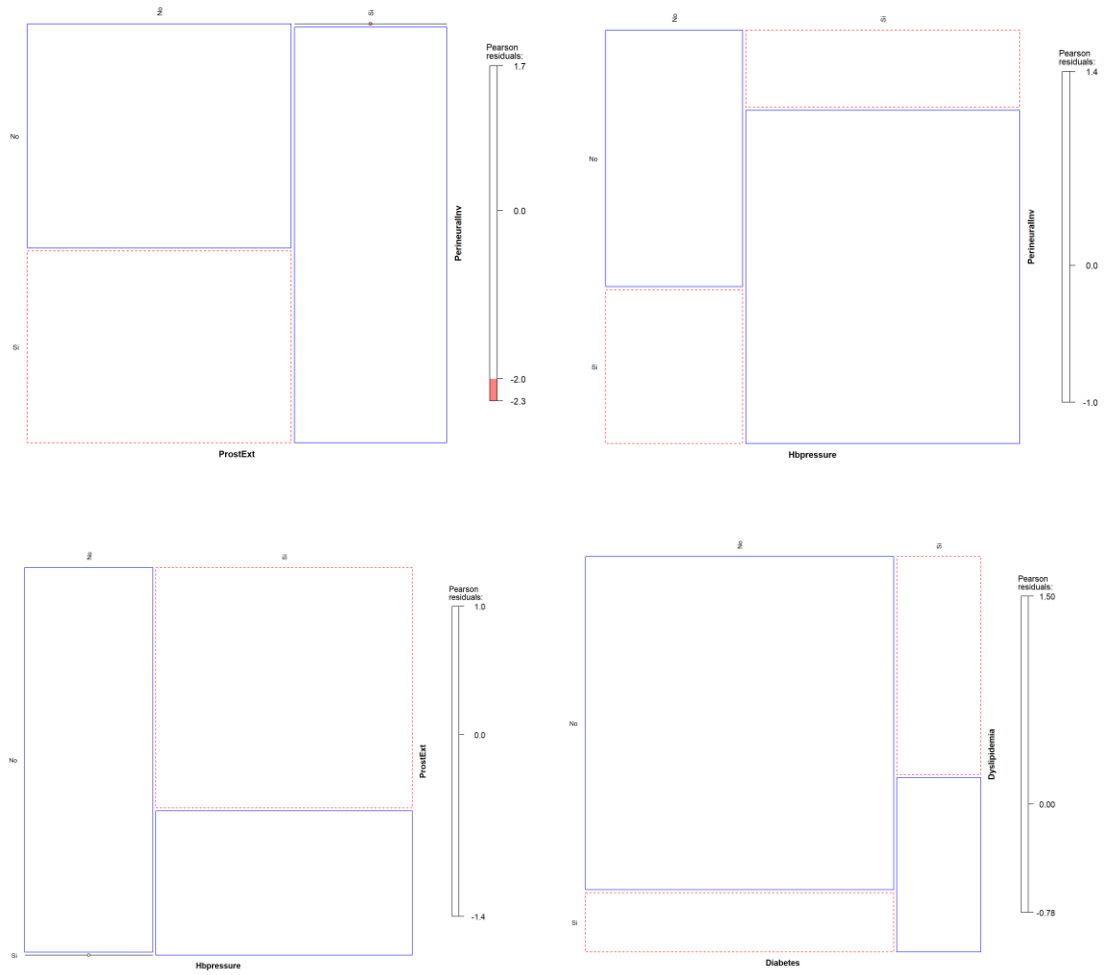


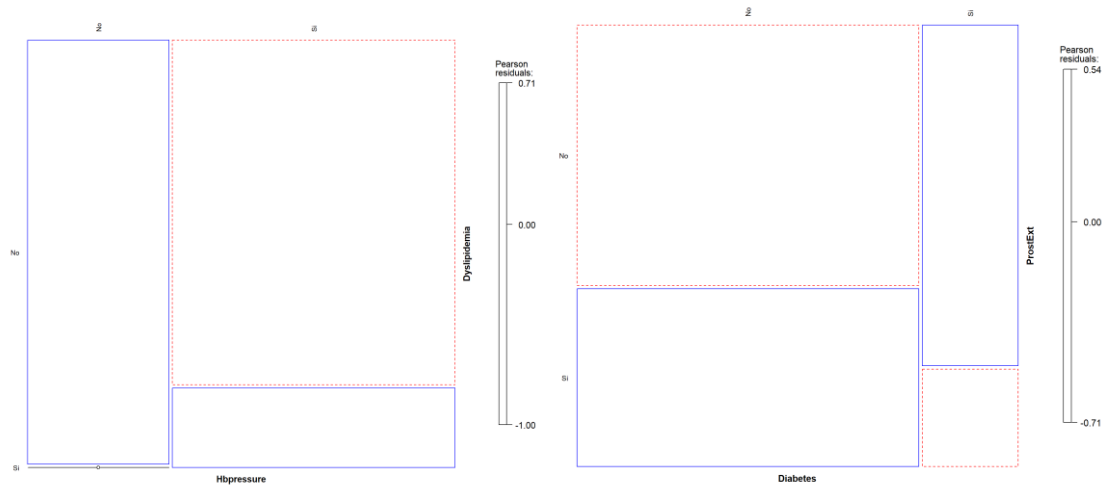
3- Association between nominal clinical factors

Cramer's V is the most popular of the chi-square-based measures of nominal association because it gives good norming from 0 to 1 regardless of table size. In practice, you may find that a Cramer's V of .10 provides a good minimum threshold for suggesting there is a substantive relationship between two nominal variables. The following table shows the pairs of variables that have a Cramer's V greater than 0.10. The rest of the pairs were discarded.

Pair	Cramer's V
ProstExt vs PerineuralInv	0.547
Hbpressure vs PerineuralInv	0.437
Hbpressure-ProstExt	0.408
Diabetes vs Dyslipidemia	0.293
Hbpressure vs Dyslipidemia	0.267
Diabetes vs ProstExt	0.158

The following mosaic plots show the distributions of the category levels between the nominal variables. As a matter of example, the results show that a major number of patients that don't have diabetes also don't have dyslipidemia.



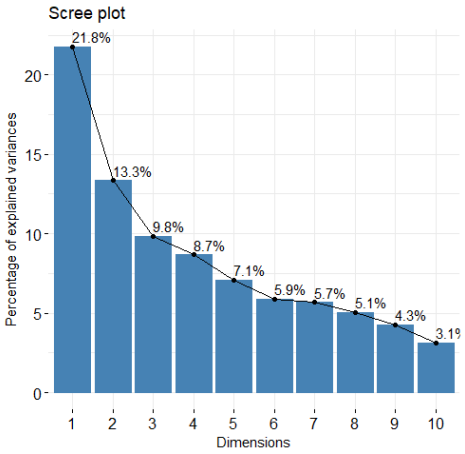


4- Principal Component Analysis

The dataset contains 42 individuals and 34 variables, where 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 dataset.

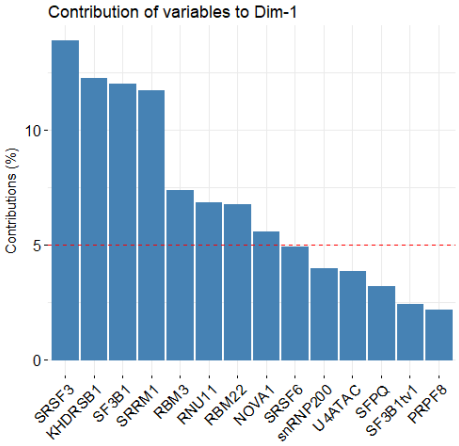
4.1. Inertia distribution

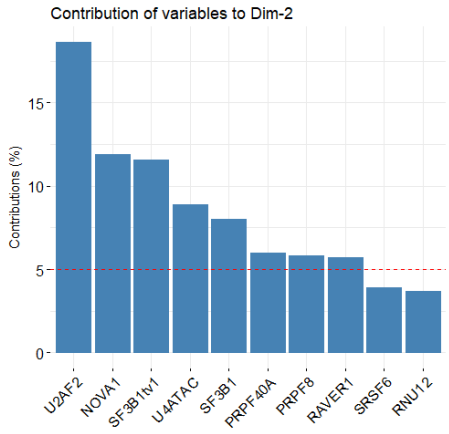
The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied. The first two dimensions of PCA express **35.11%** of the total inertia; that means that 35.11% of the individuals (or variables) cloud total variability 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. Therefore, the description will stand to these axes.



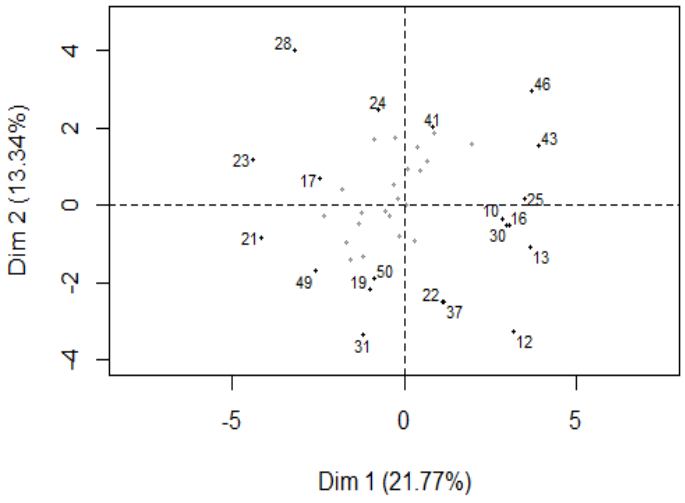
Decomposition of the total inertia on the PCA components

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





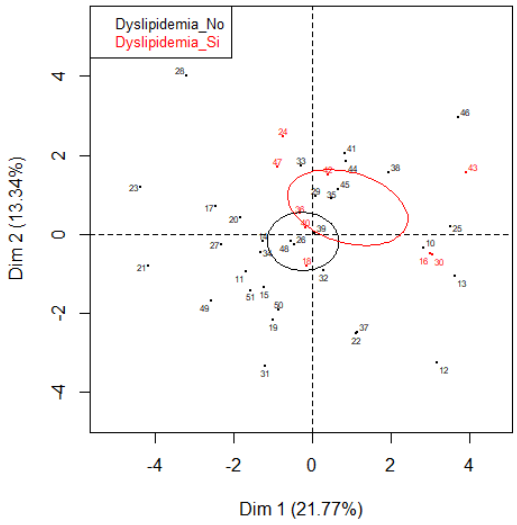
4.2. Description of the plane 1:2



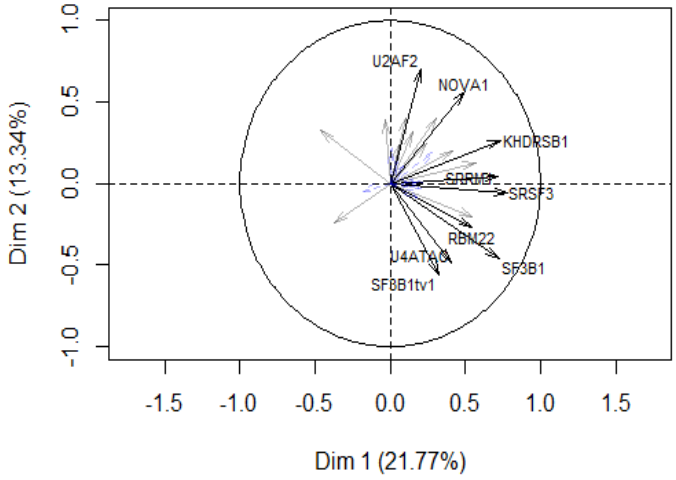
Individuals factor map (PCA) *The labeled individuals are those with the higher contribution to the plane construction.*

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



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.



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 Sections 1, 2 and 3), it is confirmed that the clinical factor sst5TMD4exp is positively 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).

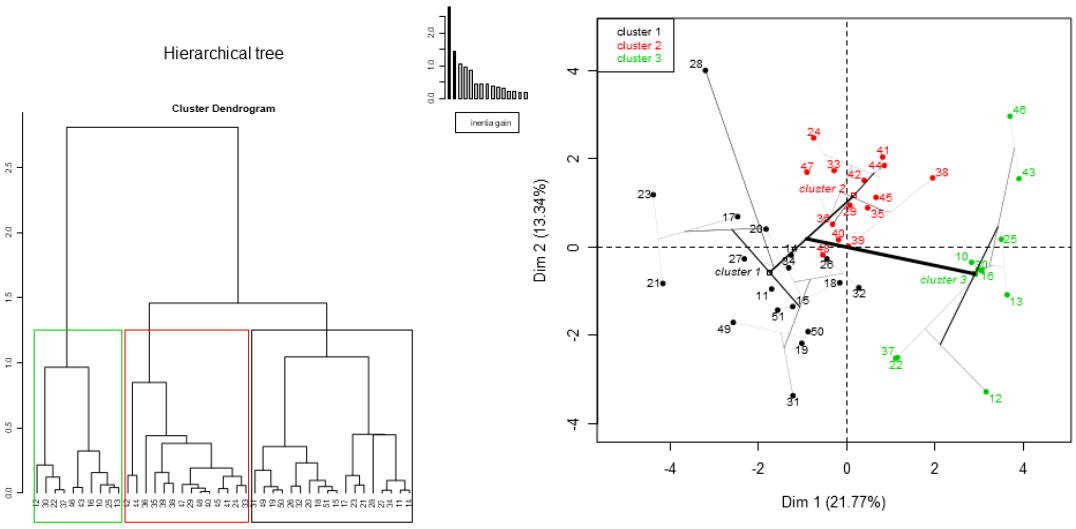
The group in which the individuals *23*, *28*, *21* and *24* stand (characterized by a positive coordinate on the axis) is characterized by:

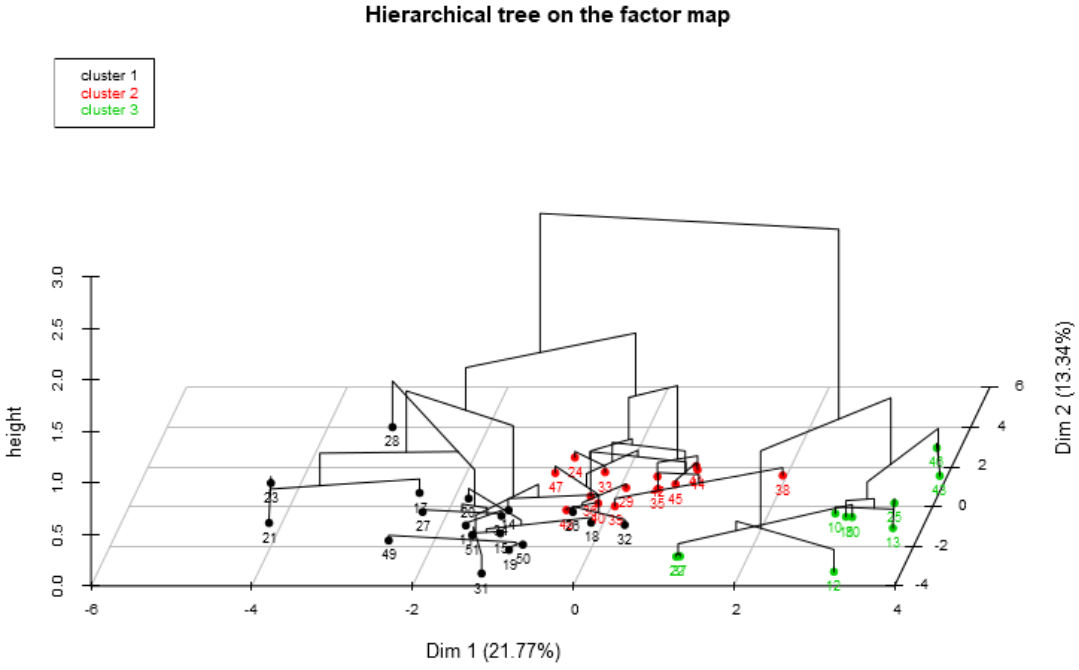
- high values for the variable *SRSF6*.
- low values for the variables *SRRM1*, *RBM22*, *RNU11*, *SRSF3*, *U4ATAC*, *SF3B1tv1* and *SF3B1*.

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).

4.3. Clustering





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

The genetic factors that best characterize the partition are (ordered from those that best characterized the partition moving to those that only partially characterize the partition, but they are still significant.)

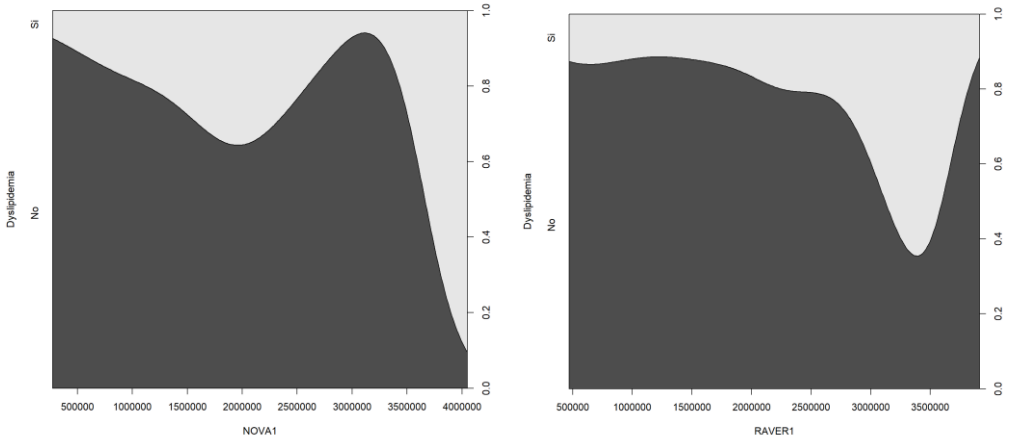
	Eta2	P-value
SF3B1	0.4918375	1.849486e-06
KHDRSB1	0.4570550	6.725709e-06
SRSF3	0.4527727	7.839200e-06
U2AF2	0.3583873	1.745108e-04
SRRM1	0.3461662	2.521237e-04
SF3B1tv1	0.3395279	3.070172e-04
NOVA1	0.2809801	1.608599e-03
RBM22	0.2563000	3.106397e-03
RNU11	0.2327760	5.701417e-03
snRNP200	0.2267564	6.640007e-03
U4ATAC	0.2259639	6.773981e-03
RBM3	0.2149384	8.925293e-03

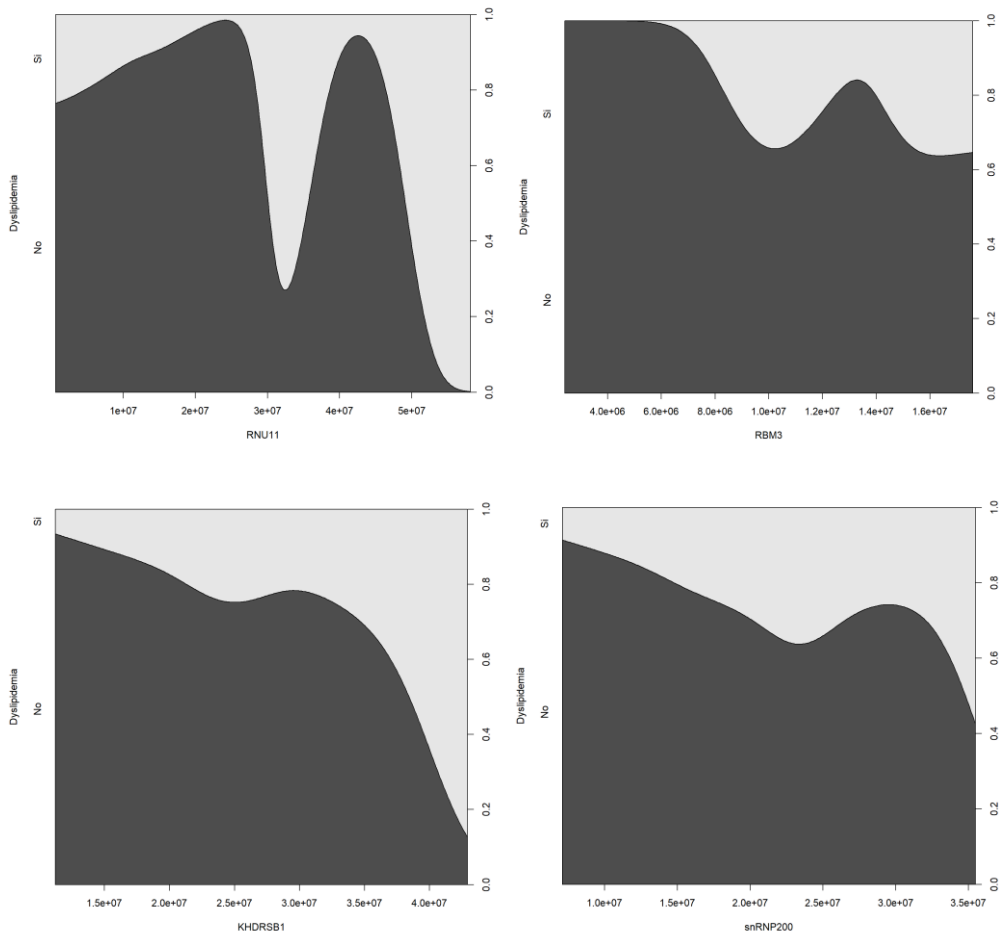
The **cluster 1** is made of individuals such as *17*, *19*, *21*, *23*, *31*, *49* and *50*. This group is characterized by:

- 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 individuals that belongs to cluster1 have a mean of Arexp equal to 12692.500, whereas the average over all individuals is 8944.216.

The best qualitative variable that is significantly describing the cluster1 is Dyslipidemia with a p-value of 0.03555456. This result matches with some previous results (see Section 1, 2, and 3) where we found that Dyslipidemia was associated to NOVA1, RAVER1, SRRM4, RNU11, RBM3, snRNP200, and KHDRSB1. Most of the individuals of cluster1 doest not have Dyslipidemia. The following conditional plots showed us that individuals with low values in the factors NOVA1, RAVER1, SRRM4, RNU11, RBM3, snRNP200, and KHDRSB1 are more likely to do not have Dyslipidemia.





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 individuals that belongs to cluster2 have a mean of PSA equal to 5.313571e+02, whereas the average over all individuals is 2.628571e+02.

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*.