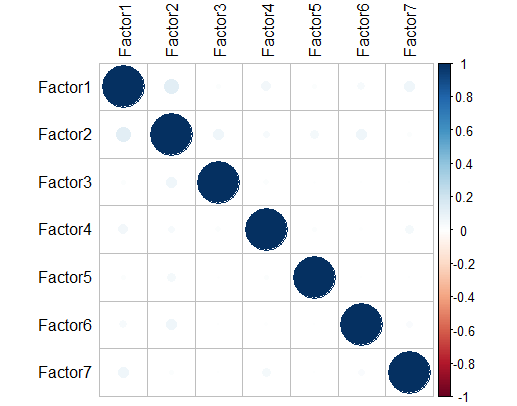
**MOFA model**

**#POSTPROCESSING MOFA2model**

**#Factors correlation analysis**

plot\_factor\_cor(MOFAobject)



#No significant interdependencies between the factors found, we’re good. Factors show no significant correlation.

**#Variance Break Down**

**# Explained Variance Decomposition by Factor** (the explained variance summarises the sources of variation from the observed factors)

r2$r2\_total

$group1

mRNA Mutations Methylation

35.05060 18.81650 24.23956

r2$r2\_per\_factor

$group1

mRNA Mutations Methylation

Factor1 7.684249 12.155580724 20.61634867

Factor2 4.981513 4.561314302 0.73184614

Factor3 7.940887 0.004667873 0.03862314

Factor4 2.589175 2.222432294 2.64019505

Factor5 7.024012 0.065890488 0.05910500

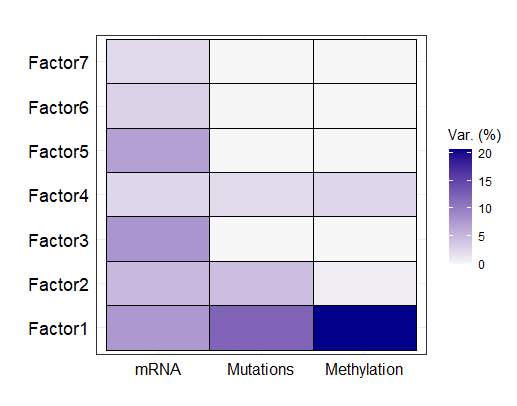
Factor6 2.945433 0.186174323 0.04387541

Factor7 2.366642 0.010565205 0.02307561

varExpPlot <- plot\_variance\_explained(MOFAobject, censor = 0.15 )

varExpPlot

**Explained variance by Factor and data modality(-omics)**



Conclusions:

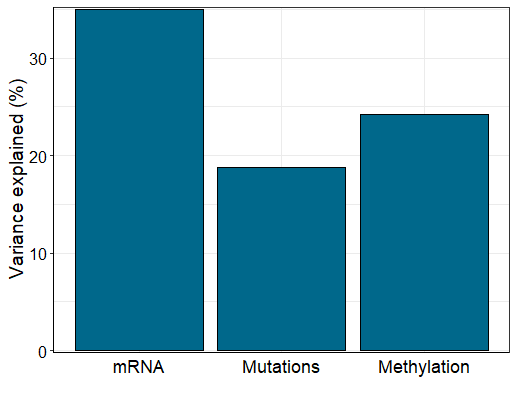
Factor1 represent strong diversity across all omics modalities (looks like, it is supreme for the disease etiology).

Factor3 and 5 express strong variance related solely to mRNA.

Factor4 retains variance in all omics

plot\_variance\_explained(MOFAobject, plot\_total = T)[[2]]

**Total explained variation per data modality**



#Higher percentages of the variance indicate a stronger association (means that we make better predictions).

Conclusuions:

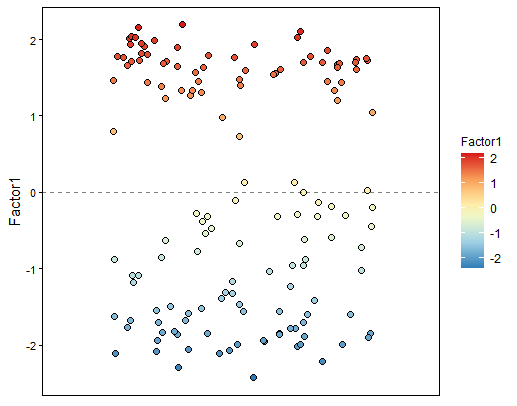
Extracted 7 Factors explain: 35% mRNA variance, 24,2% DNA methylation profile variance, 18,8% Mutations variance.

**Factor1 Profile**

Factor is a linear combination of initial features and represents a source of data variability. On the image below, Factor values are located around the zero axis. Points with positive and negative signs represent contrasting phenotypes. A larger absolute value correlates with a more expressed biological effect.

plot\_factor(MOFAobject, factors = 1, color\_by = "Factor1")

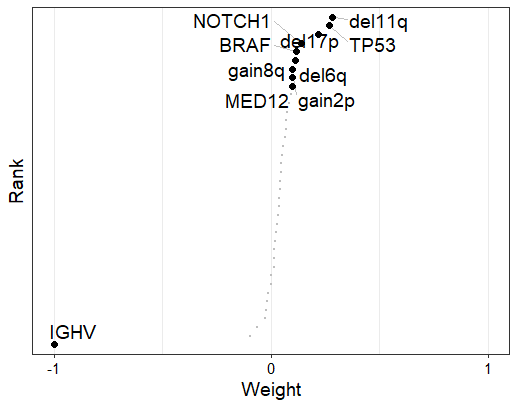
**Factor1 values**



**Feature Weights**

are the grades of the interconnection of each feature with each factor. A larger absolute feature weight indicates a stronger association of the feature with the factor. A sign interprets the effect direction.

**Factor1 Somatic Mutations**



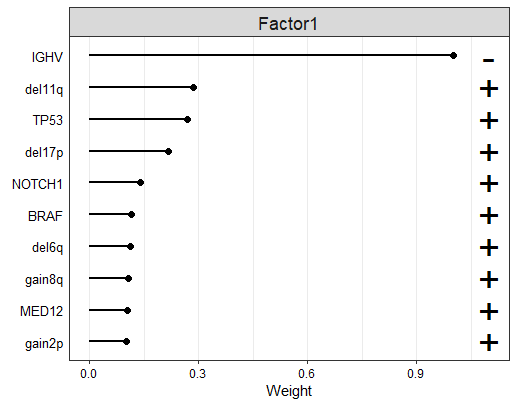
Conclusions:

Most features are allocated around zero weight vertical line, indicating that have weak association with Factor1.

IGHV gene mutation clearly stands out.

Let’s create a plot to display features with the corresponding weight sign on the right.

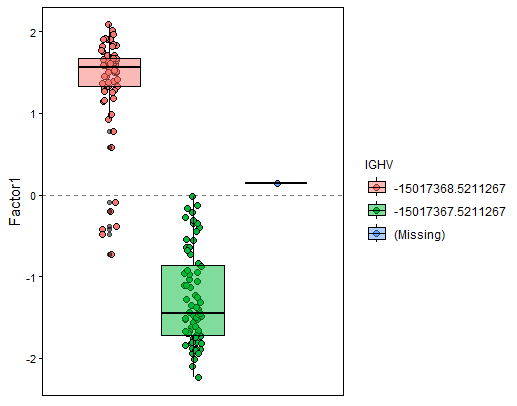
**Factor1 somatic mutations top feature weights**



IGHV has a negative weight. Samples with positive Factor1 values have IGHV mutation whereas samples with negative F1 values do not have the IGHV mutation. To confirm this, let’s plot the Factor values and color the IGHV mutation status.

plot\_factor(MOFAobject, factors = 1, color\_by = "IGHV", add\_violin = TRUE, dodge = TRUE)

**IGHV mutation correlation with Factor1 values**

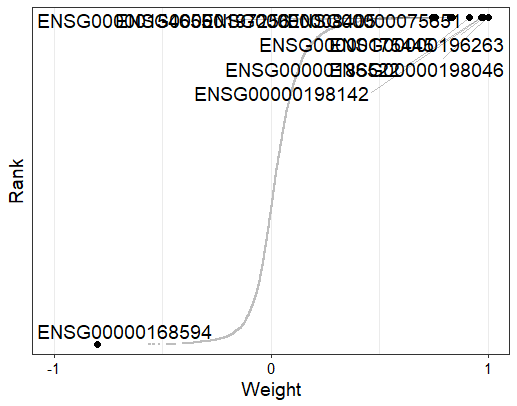


Warning message:

In plot\_factor(MOFAobject, factors = 1, color\_by = "IGHV", add\_violin = TRUE, :

Warning: some 'color\_by' groups have only one observation, violin plots cannot be added. Adding boxplots instead...

**Factor1 mRNA Expression**



Conclusions:

A substantial number of genes with positive weights and genes with negative weight (for instance, ENSG00000168594 (ADAM29) are present.

Likely, genes with large positive mRna expression values are more expressed in the samples with IGHV mutation; genes with large negative values are more expressed in the samples without the IGHV mutation.

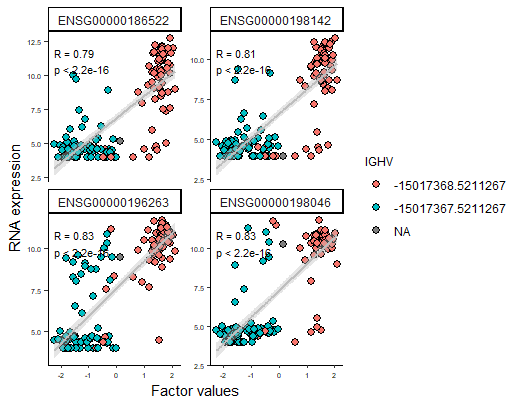
Let’s verify the last assumption.

**Factor1 Molecular Signature Clustering**

Let’s make a heat map of Factor values (x-axis) versus gene expression values (y-axis). Samples are colored by IGHV status – is the mutation present(red) or not (blue).

The function plot\_data\_scatter generates a scatterplot of Factor 1 values (x-axis) versus expression values (y-axis) for the top 4 genes with largest positive weight. Samples are coloured by IGHV status:

**Factor values (positive weight)**



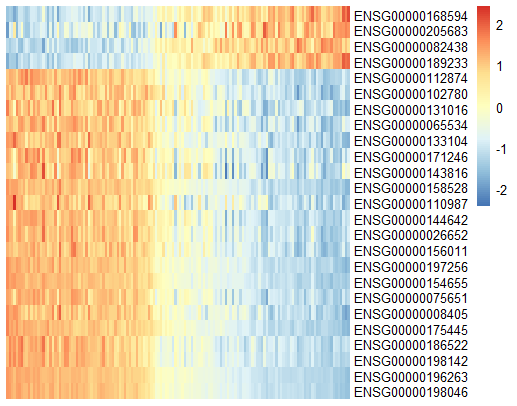
**Factor values (negative weight)**

Immagine che contiene mappa

Descrizione generata automaticamente

An alternative visualisation is to use a heatmap:

**Denoised heatmap for the clusters of Factor1 gene expression**

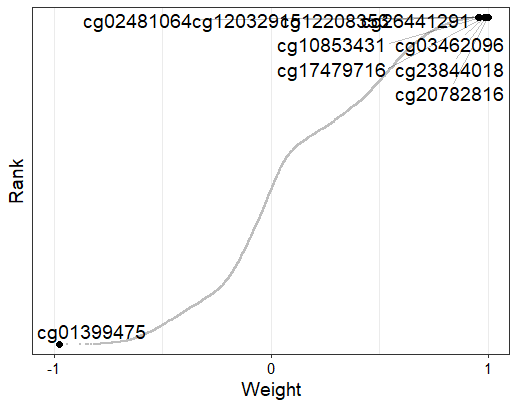


**Conclusion:**

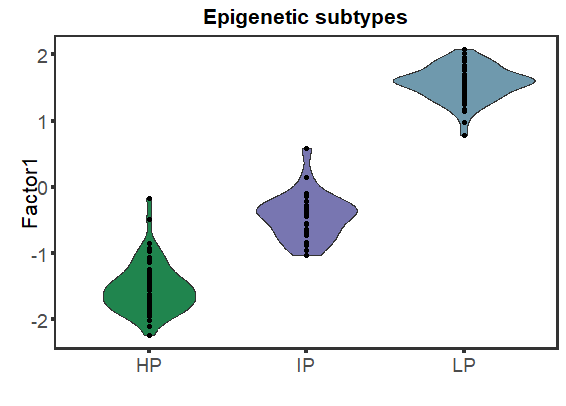
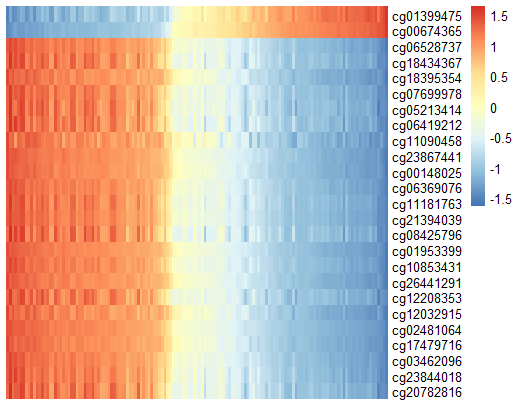
Clusters with different gene expressions are clearly visible which proves the assumption.

Associations between Factor 1 and epigenetic subtypes

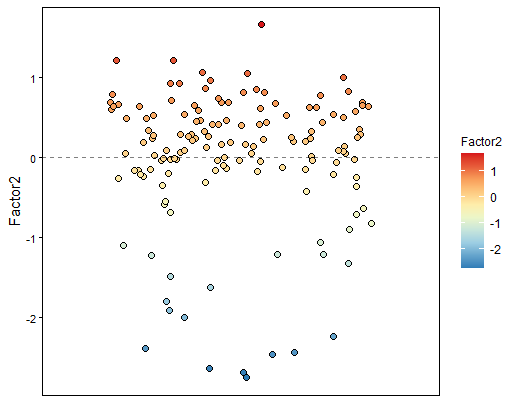
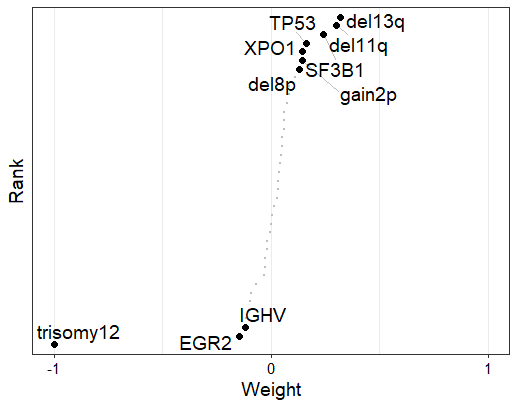
Methylation profile

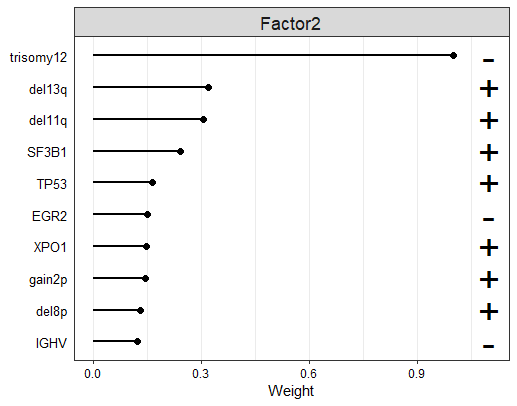
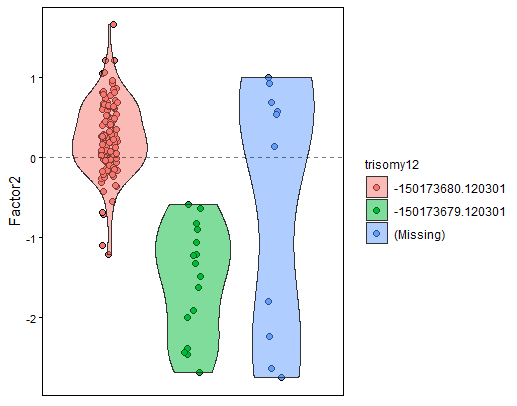
Immagine che contiene tavolo

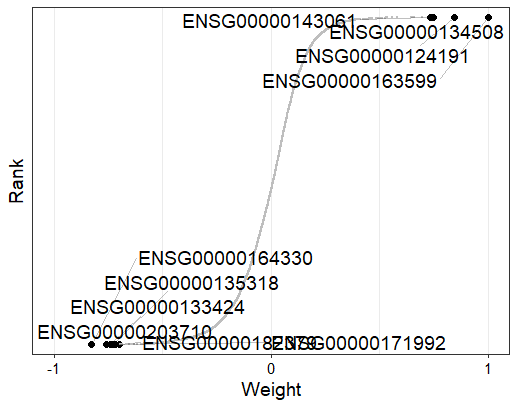
Descrizione generata automaticamente



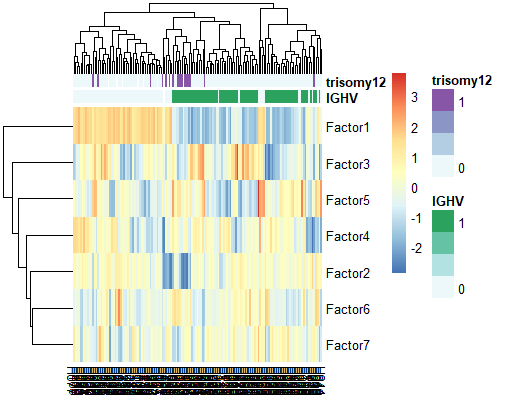
**Factor2 profile**

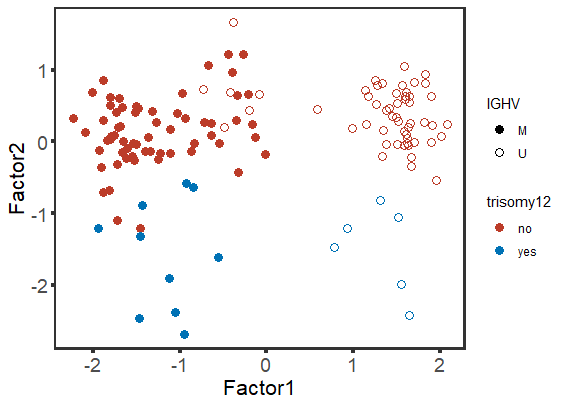


**Factor heatmap**



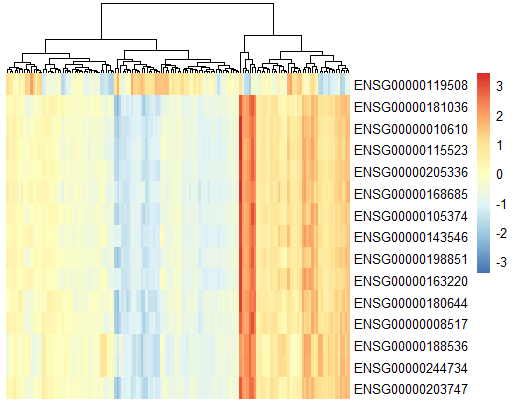
**Inspection of combinations of F1 and F2**

Characterization the etiology of the two main Factors, let’s explore them together:

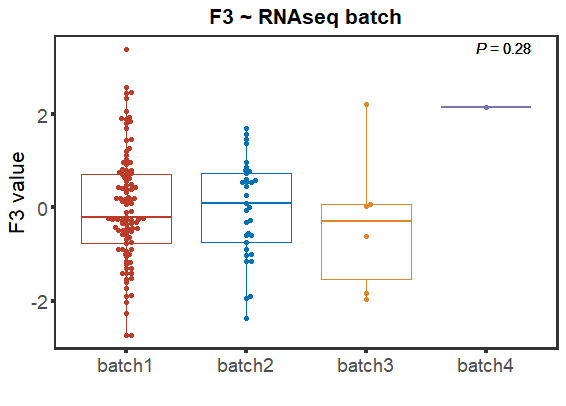


This plot is extremely important. It classifies the patients into four different subgroups depending on their (multi-omic) molecular profile.

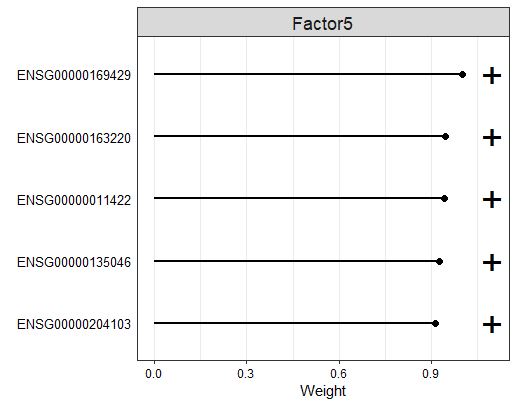
**Factor 3 profile – only explain mRNA expression variation**



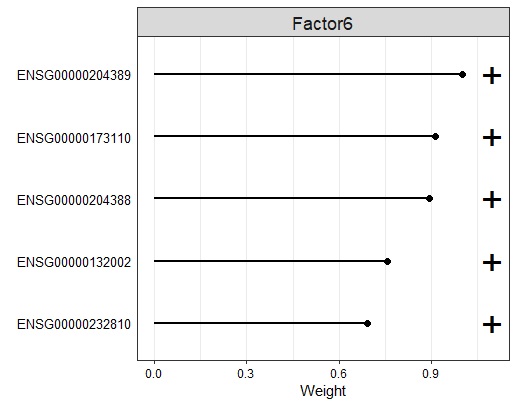
**F3 – RNAseq batch**



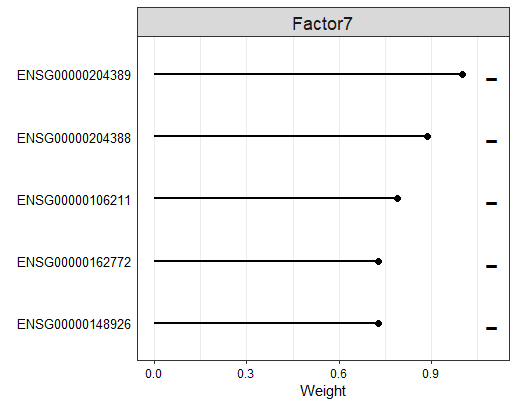
**Factor 5 profile**



**Factor 6 profile**



**Factor 7 profile**



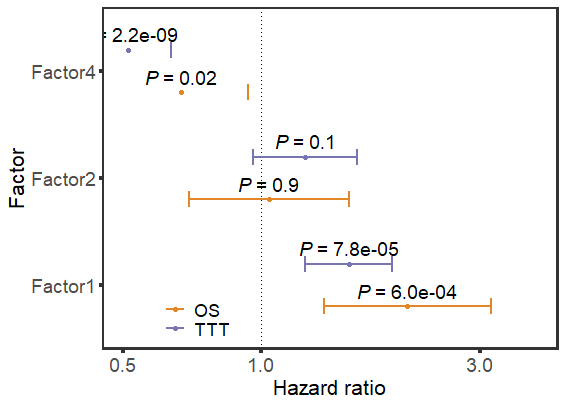
**Associations of latent factors to clinical behaviour**

The factors inferred by MOFA2 can be related to clinical outcomes such as TTT or OS. As this type of data is censored (not for all samples we have already observed the event) we will use [Cox models](http://www.bandolier.org.uk/painres/download/whatis/COX_MODEL.pdf) for this purpose. In a Cox proportional hazards model we model the hazard of an event ocurring (e.g. death or treatment) as a function of some covariates (here the factors). If a factor has a influence on the OS or TTT it will receive a high absolute coefficient in the Cox model. In particular:

* If the coefficient is positive, samples with large factor values have an increased hazard (of death or treatment) compared to samples with small factor values.
* If the coefficient is negative, samples with small factor values have an increased hazard compared to samples with a large factor values.

**Univariate cox regression**

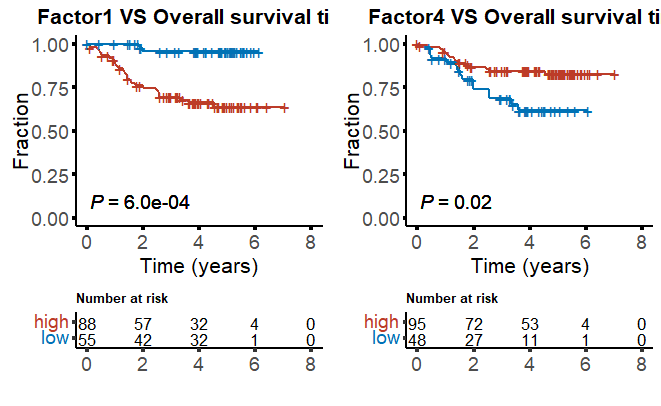
**Plot p values and hazard ratios**



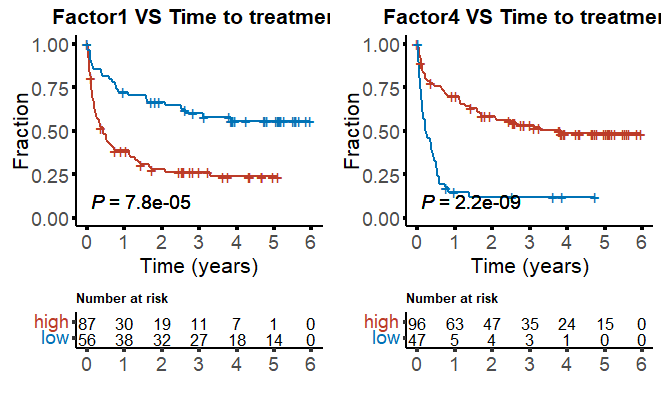
**Karlan-Meiler plots**

For illustration purposes we split the samples based on the factor values into two groups using the maximally selected rank statistics from the maxstat R package and plot the Kaplan Meier plots per group.

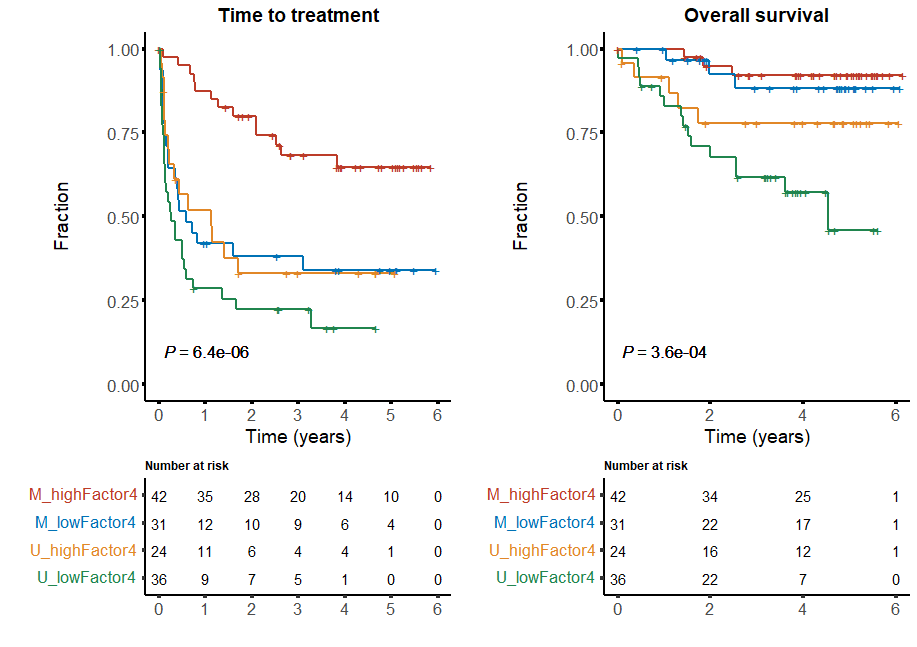
KM plot for overall survival (OS)



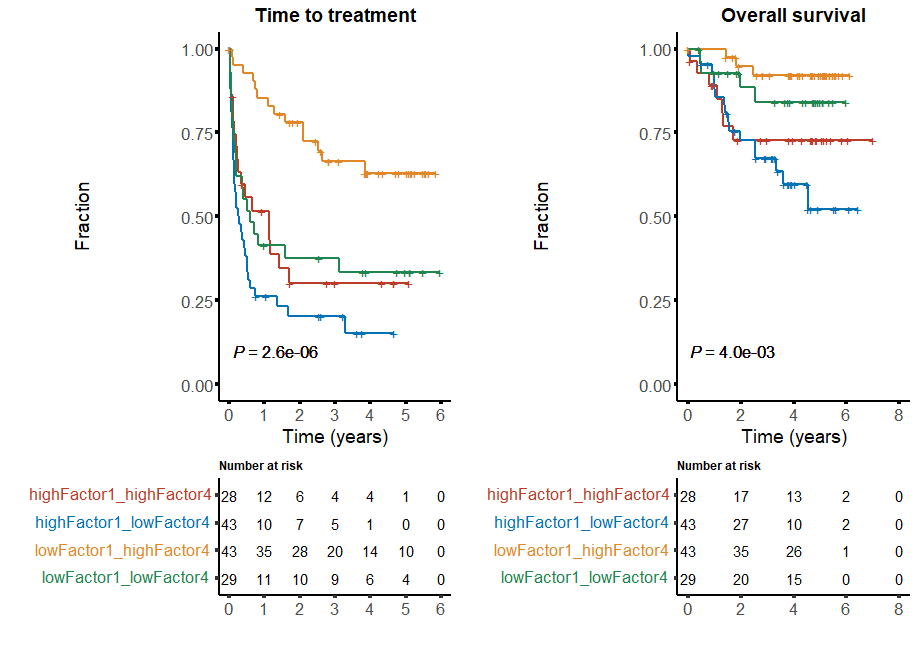
KM plot for time to treatment (TTT)



**KM plot for subgroup defined by IGHV status and median latent factor values**



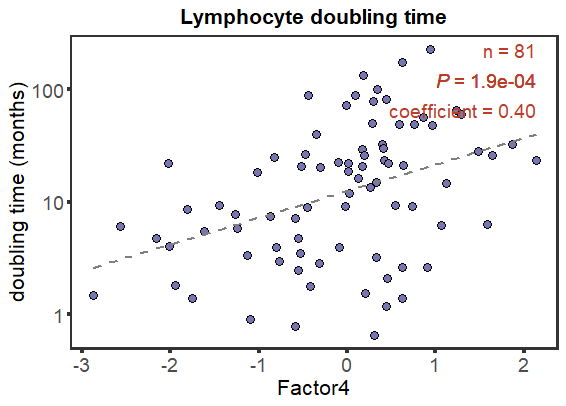
**KM plot for subgroup defined by median latent factor values of F1 and F4**



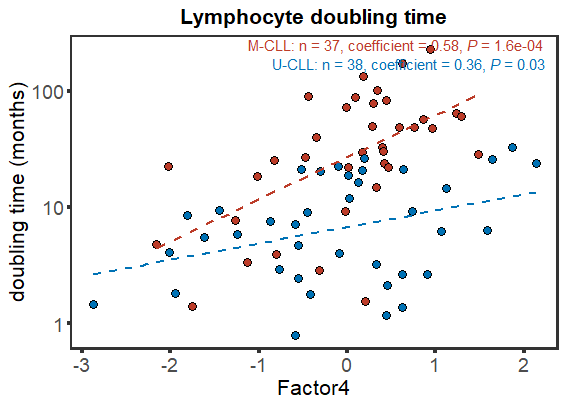
Correlation between F4 and Lymphocyte doubling time

Univariate test

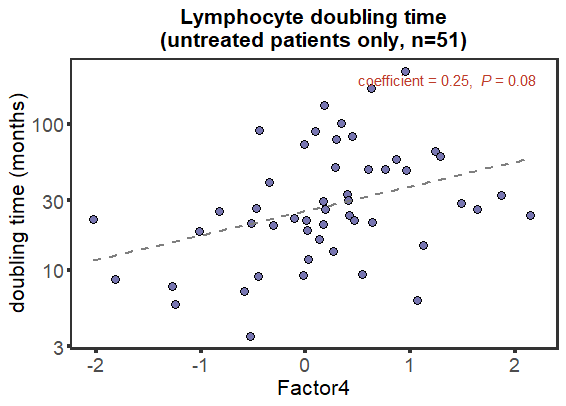
Pearson’s correlation test



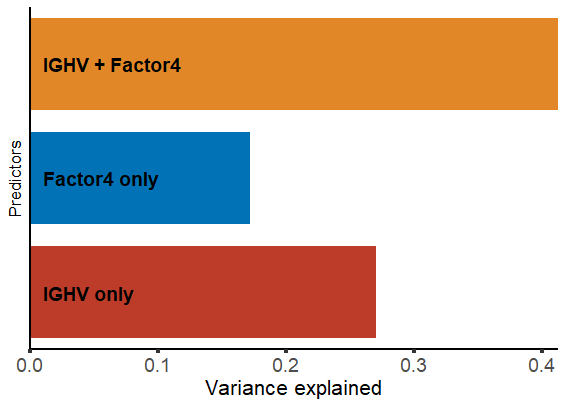
Scatter plot of correlations, stratified by IGHV



Correlations in untreated patients only



Variance explained for lymphocyte doubling time

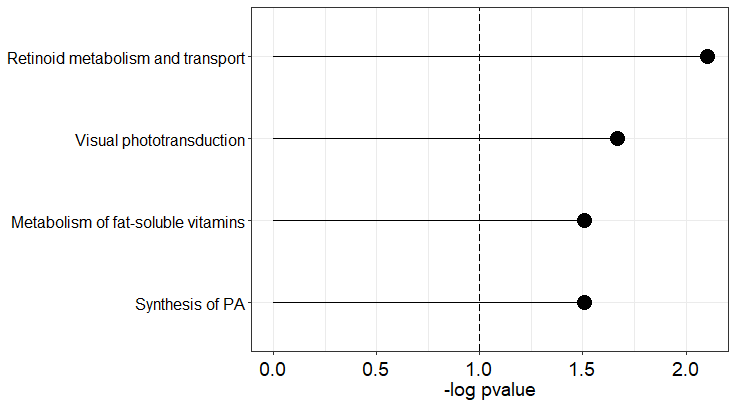
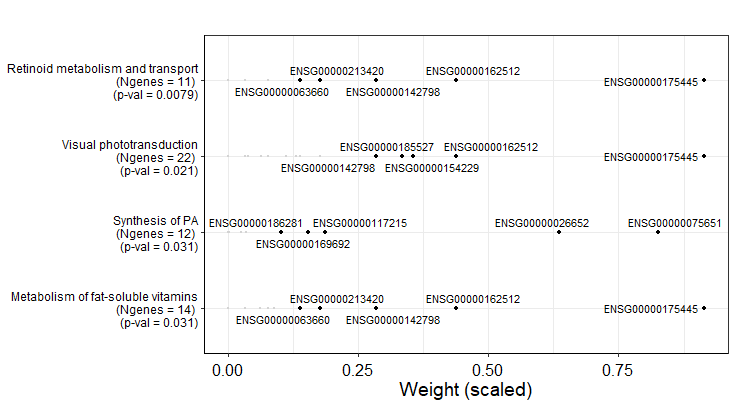


**Gene enrichment analysis**

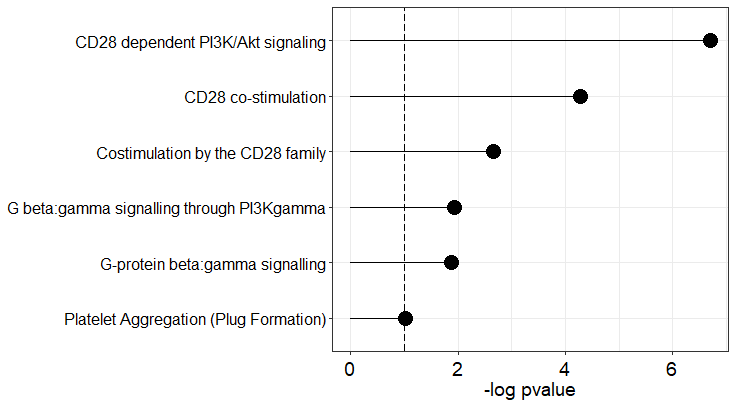
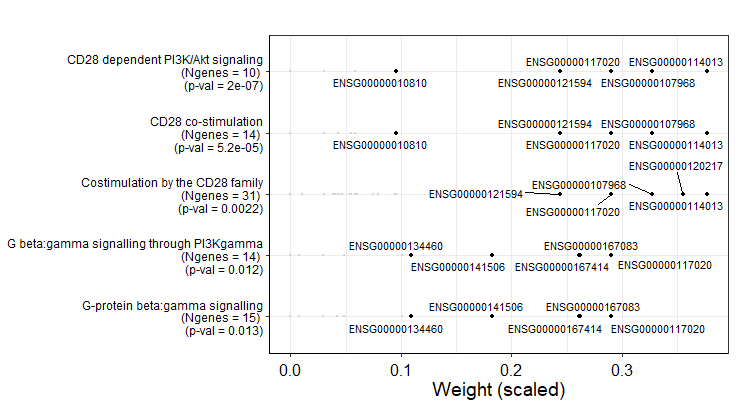
In addition to exploring the individual weights for each factor, we can use enrichment analysis to look for significant associations of factors to genesets. Here, we use the Reactome genesets for illustrations.

**Factor 1**

**GSEA on positive weights**

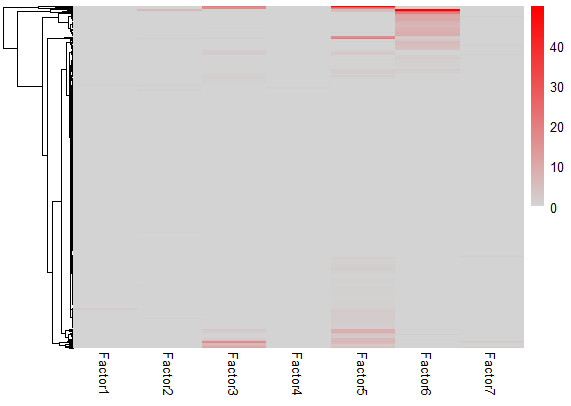
 

**GSEA on negative weights**

**Enrichment heatmap**

**GSEA on positive weights**



**GSEA on negative weights**

