# Integration example 6: Partial Least Squares



Sonia Tarazona starazona@cipf.es

Genomics of Gene Expression Group



23-27 January, 2017 MIAGE course

#### PLS-like methods



- PLS regression
- Sparse PLS
- o N-PLS
- o O2-PLS
- PLS-DA



#### Principal Component Analysis (PCA)

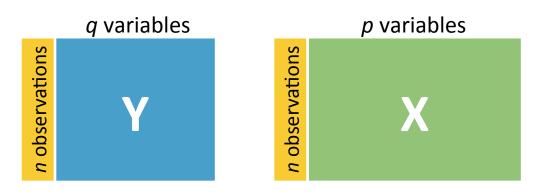


# p variables X

- Goal: Reduce the number of variables describing the variability in X
- PCA is a multivariate dimension-reduction method that transforms data from potentially correlated variables into uncorrelated principal components (PCs)
- $\times$  X = TP<sup>t</sup> + E
- Loadings (P) = Weights of original variables in PCs
- Scores (T) = Weights of observations in PCs
- \* Residuals (E)
- ♦ Iterative process → Different algorithms

#### Principal Component Regression (PCR)





- Goal: Predicting Y from X
- Classical regression models: p < n
- PCR can handle p >> n !!
- ♦ PCR is a multivariate dimension-reduction method that first computes PCA on X matrix and then uses the resulting PCs (scores T) as regressors on Y  $\rightarrow$  Y = TB + E
- BUT the PCs are chosen to explain X rather than Y so it could happen that they are not really relevant for Y.

#### Partial Least Squares (PLS)

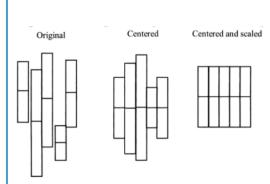


# q variables p variables X

- o Goal: Predicting Y from X and describing their common structure → Y = XB + E
- Classical regression models: p < n
- PLS can handle p >> n !!
- PLS is a multivariate dimension-reduction method that maximizes the covariance between Y and X by finding linear combinations of the variables (latent variables) from both
- $X = TP^t + E$ ;  $Y = UQ^t + F$
- ♦ Loadings (P and Q) = Weights of original variables in latent variables
- Scores (T and U) = Weights of observations in latent variables
- \* Residuals (E and F)
- ♦ Iterative process → Different algorithms

#### Scaling data for multivariate methods

- Centering
  - Subtracting the variable mean from the variable values
- Scaling (or weighting)
  - Dividing the variable by its standard deviation
  - Variables become comparable when they are measured in different units and therefore a variable with different range will not dominate the others
- Autoscaling: mean centering + scaling



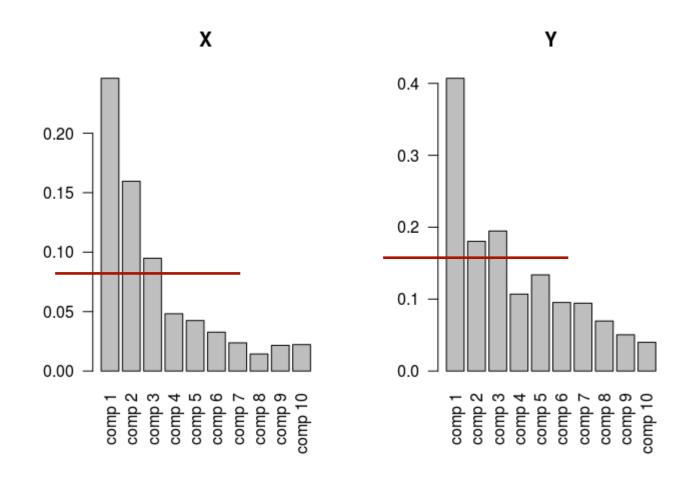
- Warning! Most of the multivariate R functions apply autoscaling by default.
- Some times, e.g. genes, all variables are measured in the same units and it is interesting to allow them for having different variability
- A preliminary exploration of data (e.g. with PCA) can help to choose the best option.
- Use scale() R function to prepare the data prior to multivariate analysis instead of letting the multivariate method to perform the scaling.

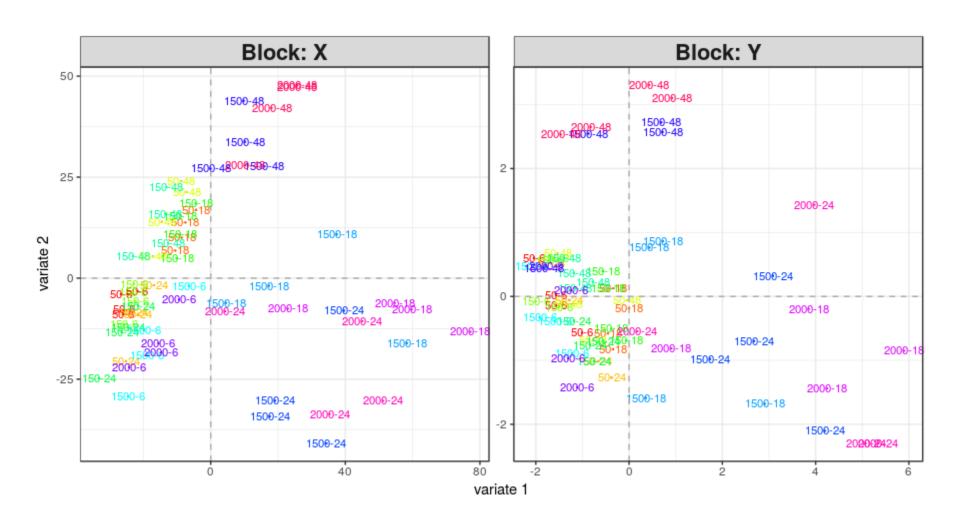


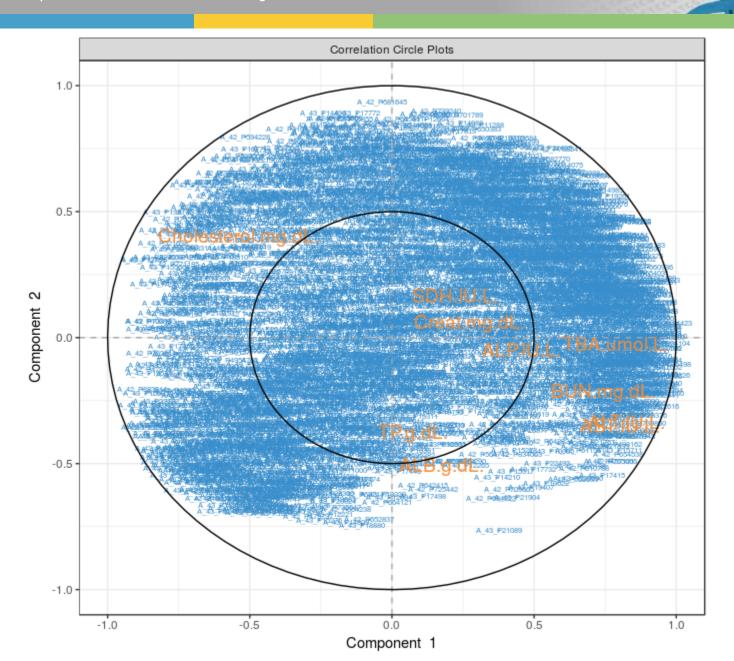
- HandsOn/IE6\_PLS/IE6\_Examples.R
- Study from Bushel et al., 2007 (taken from mixOmics R library)
- Organism: rat
  - 64 male rats
  - Samples from liver after necropsy
- Types of data:
  - Expression of 3,116 genes
  - Measurements for 10 clinical variables (markers for liver injury)
- Experimental design:
  - Doses of acetaminophen: 50 mg/kg (non-toxic), 150 mg/kg (non-toxic), 1500 mg/kg (moderately toxic), 2000 mg/kg (severely toxic)
  - o 6, 18, 24 and 48 hours after exposure

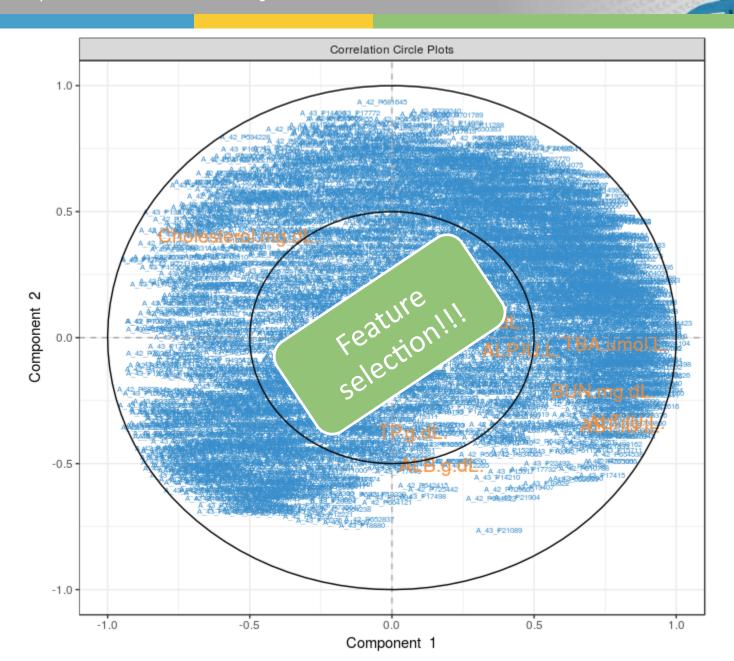


#### Choosing the number of components $\rightarrow$ ncomp = 3









#### PLS performance



#### How to measure the prediction error:

- R<sup>2</sup>: Variance of Y explained by the PLS model
- Residual Sum of Squares (RSS):  $RSS_h^k = \sum_{i=1}^n (y_i^k \hat{y}_{hi}^k)^2$
- PRediction Error Sum of Squares (PRESS) or Mean Square Error of Prediction (MSEP):  $PRESS_h^k = \sum_{i=1}^n (y_i^k \hat{y}_{h(-i)}^k)^2$
- Root Mean Square Error of Prediction (RMSEP)
- Q<sub>h</sub><sup>2</sup>: Marginal contribution of latent variable h (or score vector) to the predictive power of the model  $Q_h^2 = 1 \frac{\sum_{k=1}^q PRESS_{kh}}{\sum_{k=1}^q RSS_{k(h-1)}}$

Rule: Latent variable h is selected if  $Q_h^2 \geq (1 - 0.95^2) = 0.0975$ 

#### Cross-validation



- Model validation allows for assessing whether the generated model will be suitable for prediction given a new independent data set.
- Validation = Divide the data into two subsets: training set and testing set → Requires a large number of observations
- Small number of observations → Cross-validation (CV): Multiple rounds of validation → Prediction error is averaged for all the rounds
  - Leave-one-out CV (LOOCV): Training set = n-1 observations,
     Test set = 1 observation → Recommended for small sample sizes
  - K-fold CV: The data are divided in k equal sized subsets.
     Training set = k-1 subsets, Test set = 1 subset

# Feature selection in PLS

#### Feature selection in PLS



- Based on loadings, that measure the relative importance of the variables in the model.
- VIP (Variable Importance in Projection) measures the importance of X-variables in the model for each component (computed from the loadings for each component).
  - "Greater than one" rule is generally used
- Permutation strategies.
- Elastic Net regularization (Zou and Hastie, 2005).
- Sparse PLS (Lê Cao et al., 2008)
- O ...

Tarazona et al. *Variable selection for multifactorial genomic data*. Chemometrics and Intelligent Laboratory Systems, 2012

#### Sparse PLS (*Lê Cao et al., 2008*)



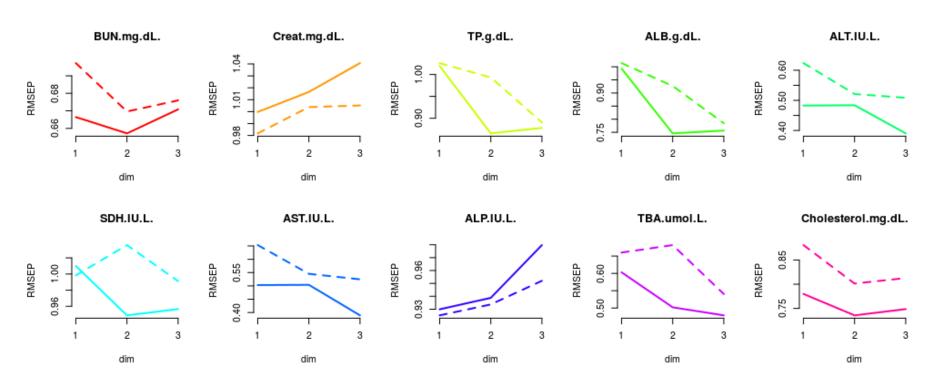


- Idea: Select the X and Y variables that are contributing most to the PLS model
- How: Penalizing loadings -> Lasso penalization
- The user chooses the number of variables to be retained = number of variables with non-zero loading for each component and for X and Y
- Implemented in mixOmics package

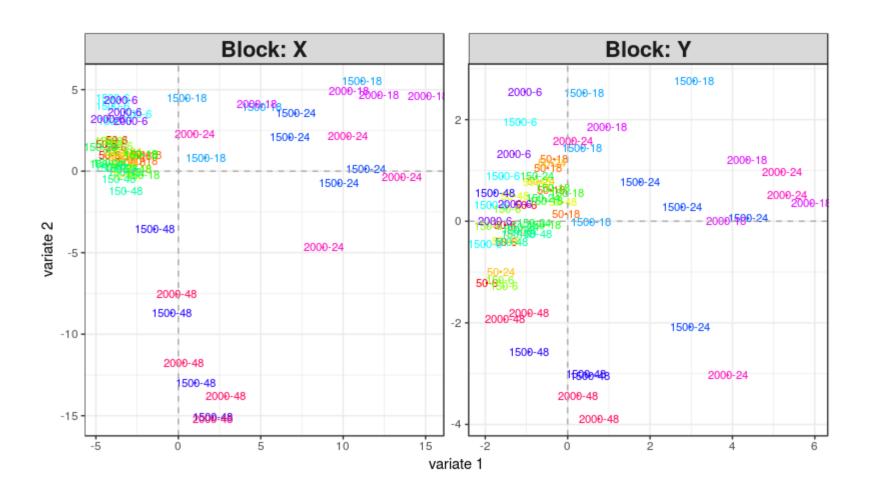
#### Example on sPLS



- HandsOn/IE6\_PLS/IE6\_Examples.R
- Prediction error tends to be lower for sPLS (full line) than for PLS (dashed line)
- This plot can be used to either choose the number of components or the number of features to retain

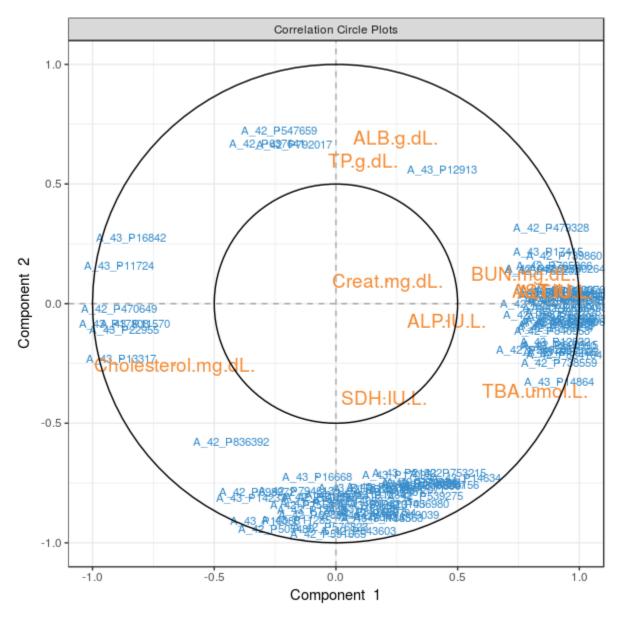


#### sPLS example: Samples plot

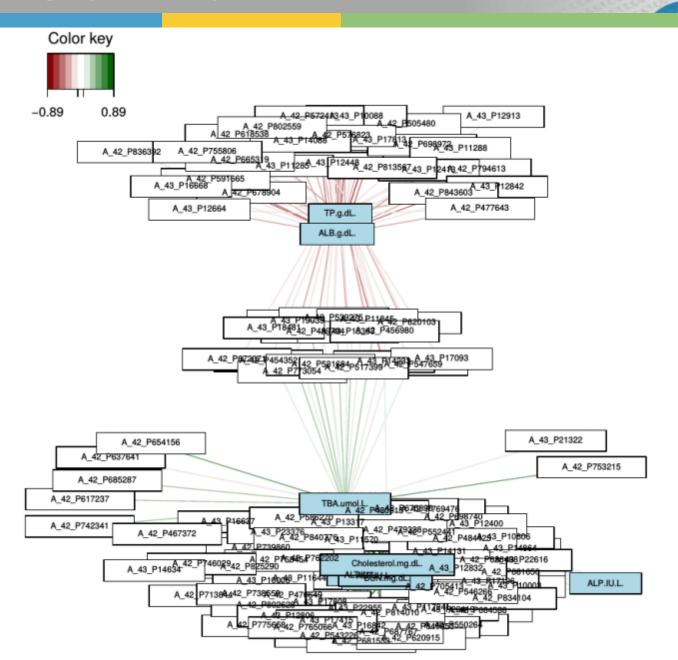


#### sPLS: Loadings plot

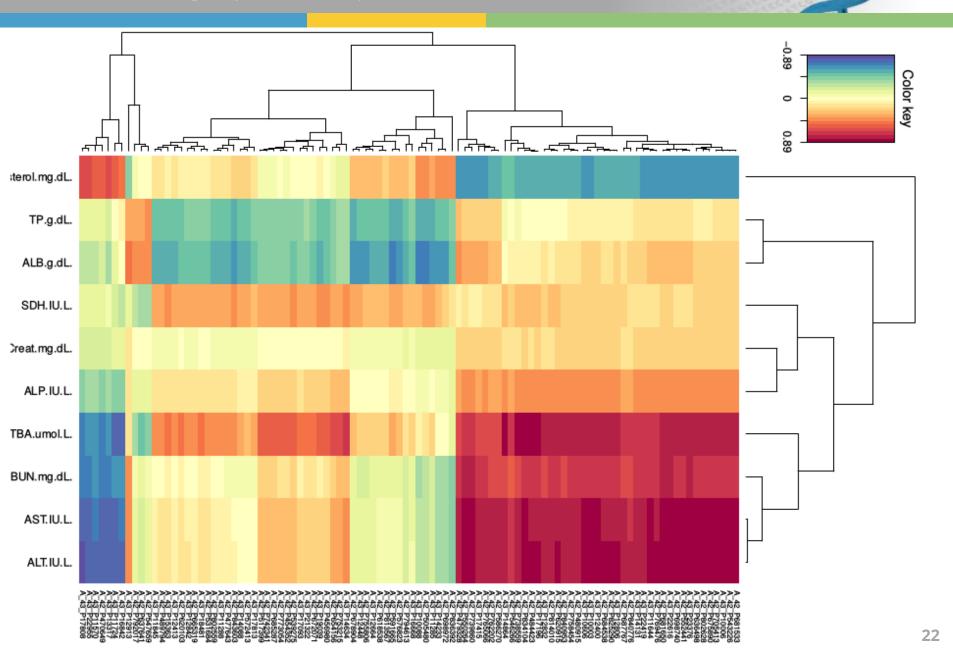




#### sPLS: Other graphical representations



### sPLS: Other graphical representations













- Dataset for Escherichia coli K-12 strains (Ishii et al., 2007).
- o mRNA, protein, and metabolite concentrations
- Knockout strains in which a single enzyme gene is knocked out in each strain + Wild type
- Datasets are normalized, and the effects of differences due to the sampling dates and arrays were eliminated. However, there are still missing values in some datasets.

HandsOn/IE\_PLS/EcoliData







- 1. Missing value treatment
  - The following tasks can be performed with missing values (accepted by mixOmics functions)
  - However, make sure you do not have entire rows or columns with missing values
- 2. Explore **expression** and **protein** data, and decide the best scaling option
  - Boxplots, PCA, etc.
- 3. Generate an appropriate PLS model
  - Number of components?
  - Sparse or not?
- 4. Conclusions from PLS model

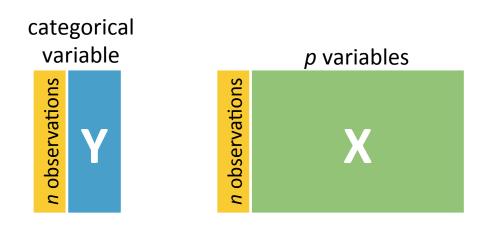


PLS-DA



#### PLS Discriminant Analysis (PLS-DA)





- Goal: Predicting Y from X, where Y is a categorical variable
- Y is recoded as a dummy block matrix
- The number of components is usually set to k-1, where k is the number of classes in the categorical variable Y
- Sparse PLS-DA (Lê Cao, BMC Bioinformatics 2011) allows for variable selection and it is implemented in mixOmics R package

#### PLS-DA example: Small Round Blue Cell Tumors



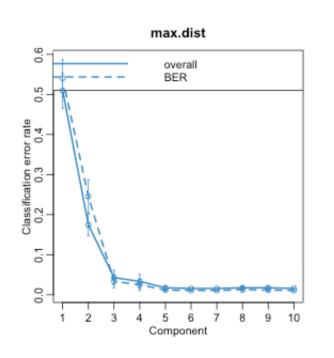


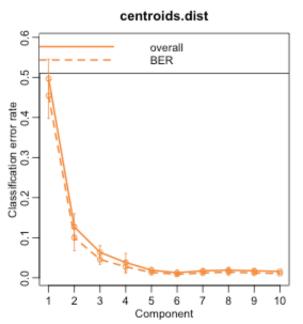


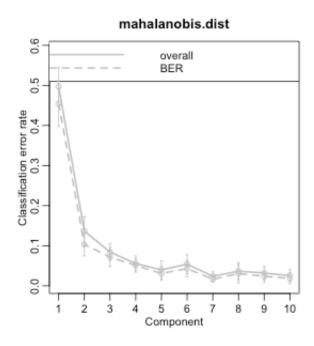
- Study from Kahn et al., 2001
- Expression levels for 2,308 genes
- 63 samples distributed in 4 classes
  - 8 Burkitt Lymphoma (BL)
  - 23 Ewing Sarcoma (EWS)
  - 12 neuroblastoma (NB)
  - 20 rhabdomyosarcoma (RMS)

#### Choosing the number of components

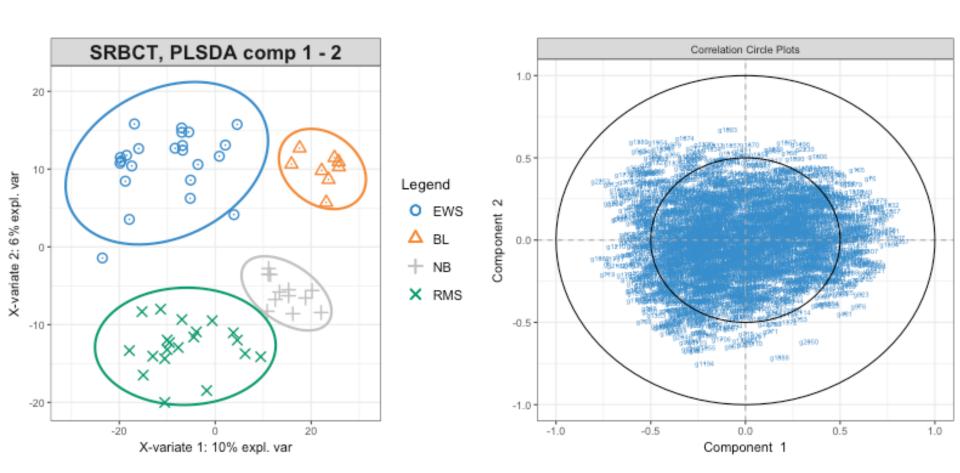
- 1. Compute the model with a high number of components, e.g. ncomp = 10.
- 2. Perform cross-validation.
- Look at the error rate results to choose the number of components.
  - BER = Balanced Error Rate → Used when the number of samples per group is not balanced



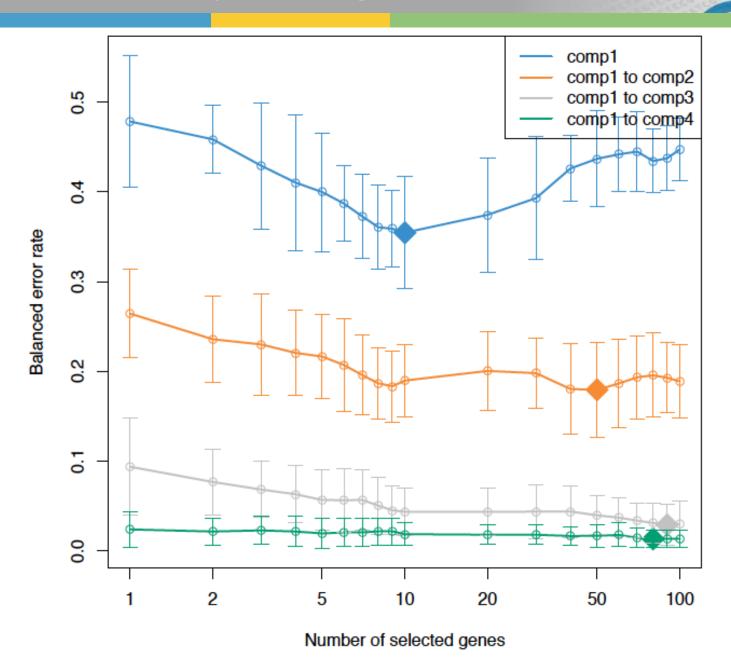




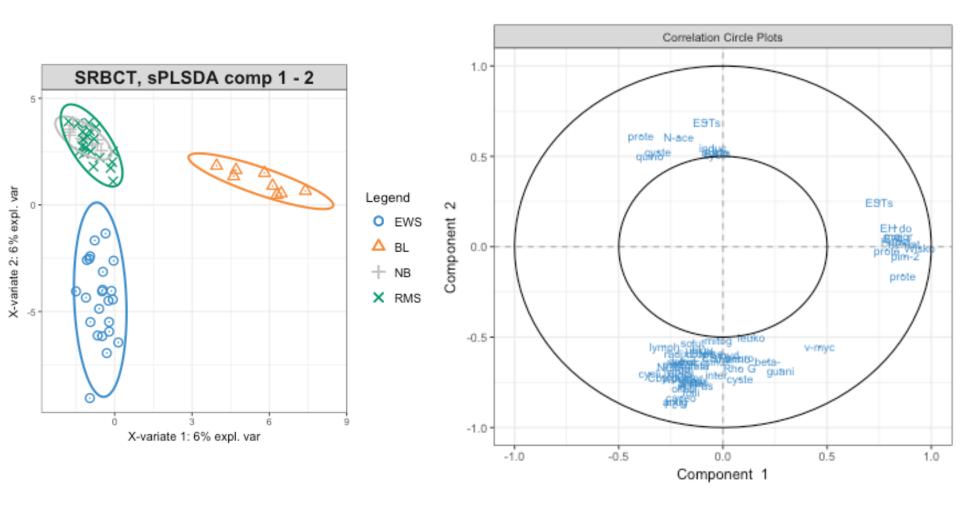
#### PLS-DA example: Model results



#### Sparse PLS-DA example: Tuning the model



# Sparse PLS-DA example: Final model



#### PLS-DA for multi-omics integration

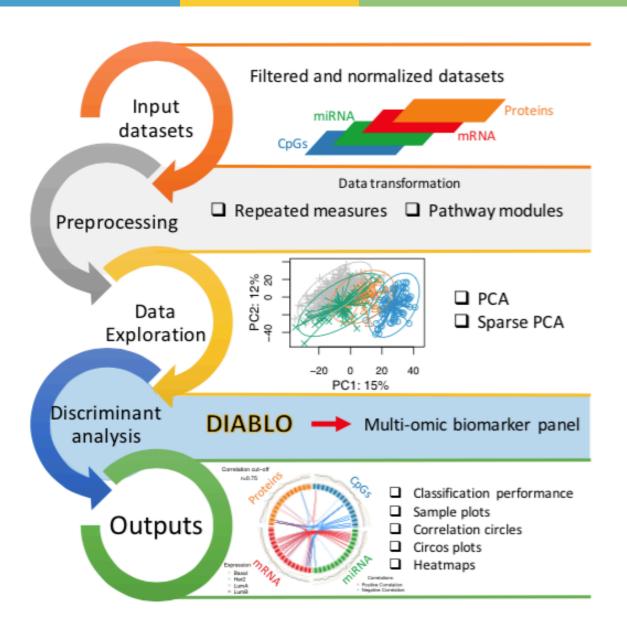


# Different omics measured on the same samples > Multi-omic signature to predict outcome

- Concatenate omics in X + PLS-DA
  - Different number of features for each omic → Unbalanced signature
- Model-based integration (e.g. ensemble classifiers)
- DIABLO (mixOmics)
  - http://www.biorxiv.org/content/biorxiv/early/ 2016/08/03/067611.full.pdf
  - Categorical Y + different omic data X<sub>1</sub>, X<sub>2</sub>, etc.
  - Still requires a minimum sample size in order to have train and test data
  - Recommended to pre-filter omic data: variability filter, features with significant changes, PCA, sparse PCA,...

#### DIABLO (mixOmics R package)







#### DIABLO example: Breast TCGA







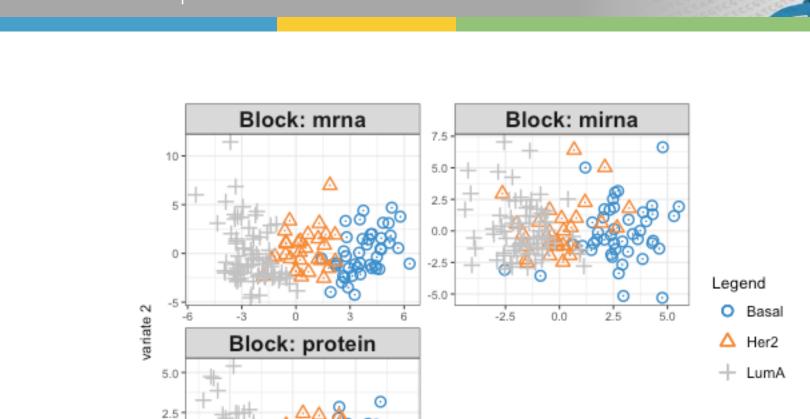
- Data from The Cancer Genome Atlas (<u>https://cancergenome.nih.gov/</u>)
- Omic data types:
  - o mRNAs
  - miRNAs
  - proteins
- 220 samples from 3 cancer subtypes
  - Basal
  - Her2
  - LumA

0.0

-2.5

-2.5

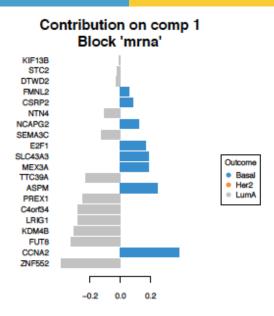
0.0

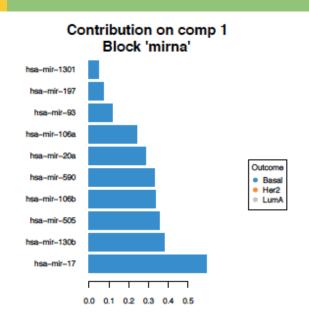


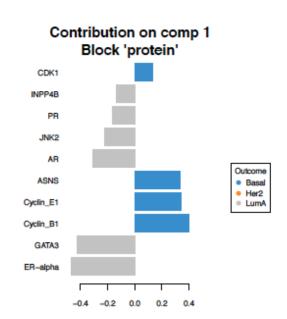
2.5

5.0

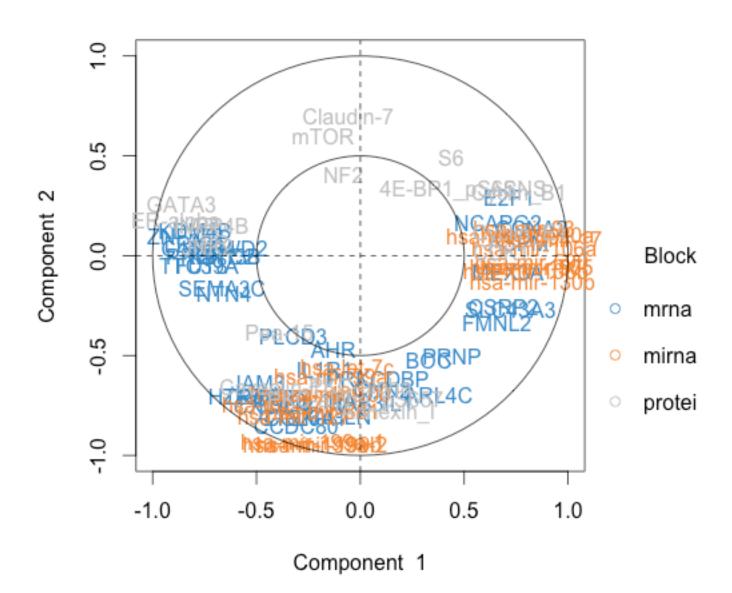
variate 1



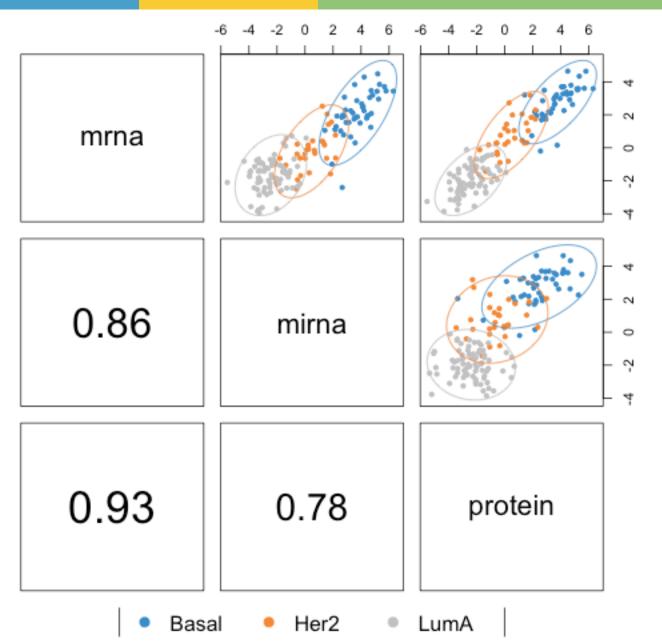




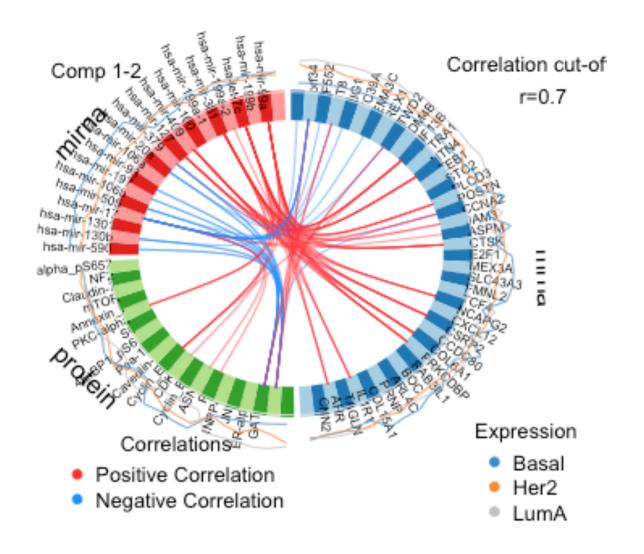
#### **Correlation Circle Plots**



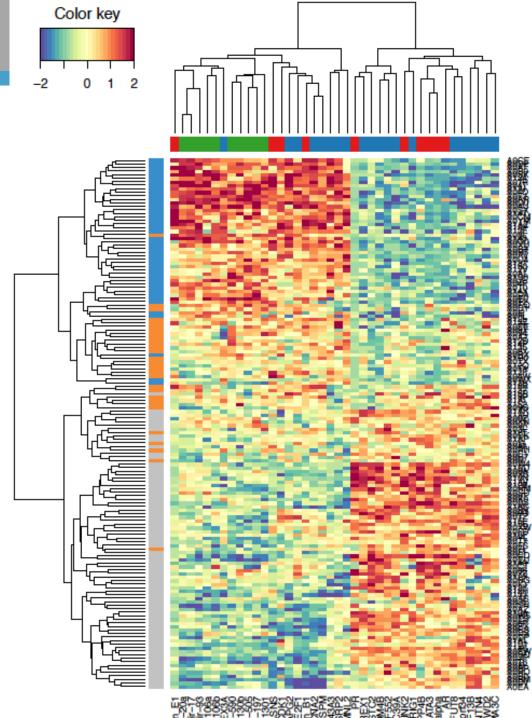








#### DIABLO example:



#### Rows

- Basal
- Her2
- LumA

#### Columns

- mrna
- mirna
- protein



#### Glioblastoma data from TCGA



- Data from STAegRa R package: data("STATegRa\_S3")
- Gene expression, miRNA expression
- Processed data from TCGA (Van Deun et al., 2012)
  - o Glioblastoma multiforme (GBM) → Tumor subtypes:
    - Classical
    - Mesenchymal
    - Neural
    - Proneural
  - o 600 genes with significant changes
  - 300 miRNAs with significant changes.



#### Glioblastoma data from TCGA





- 1. Select the train data and apply DIABLO to obtain a multiomic signature to predict the different cancer subtypes.
  - Which is the optimal number of components?
  - Which is the optimal number of variables to keep from each data set?
- 2. Generate the appropriate plots for the final model and interpret the results.
- Apply the obtained signature to the test data and discuss how well the fitted model predicts the cancer subtype.

