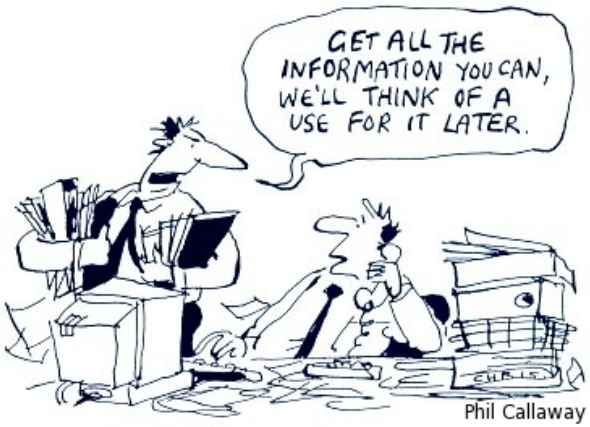
Unravelling 'omics' data with the mixOmics R package

Illustration on several studies

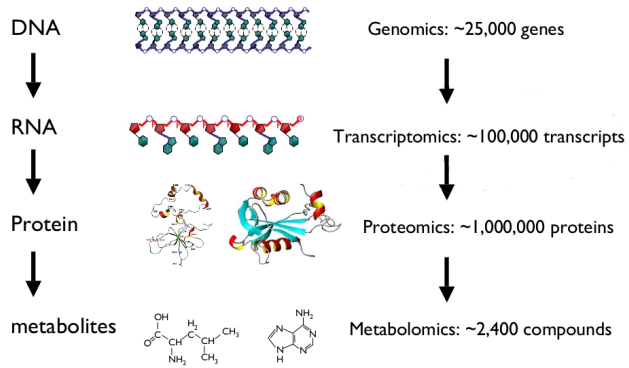
Kim-Anh Lê Cao

Queensland Facility for Advanced Bioinformatics
The University of Queensland

The issue with integrative systems biology



- Unlimited quantity of data ($n << p$ problem)
- Data from multiple sources



$\rightarrow$ Efficient and biologically relevant statistical methodologies are needed to combine the information in these heterogeneous data sets.

Single Omics analysis

- Do we observe a 'natural' separation between the different groups of patients?
- Can we identify potential biomarker candidates predicting the status of the patients?

Integrative Omics analysis

- Can we identify a subset of correlated genes and proteins from matching data sets?
- Can we predict the abundance of a protein given the expression of a small subset of genes?
- Do two matching omics data set contain the same information?

The data

n = number of patients
p, q = number of biological features (genes, proteins ..)

Single Omics analysis

- one omic data set X(n x p)
- for a supervised analysis, Y vector indicating the class of the patients

Integrative Omics analysis

- two matching omics data sets (measured on the same patients)
- X (n x p) and Z (n x q)

Linear multivariate approaches enable:

- Dimension reduction
  $\rightarrow$ project the data in a smaller subspace
- To handle multicollinear, irrelevant, missing variables
- To capture experimental annd biological variation

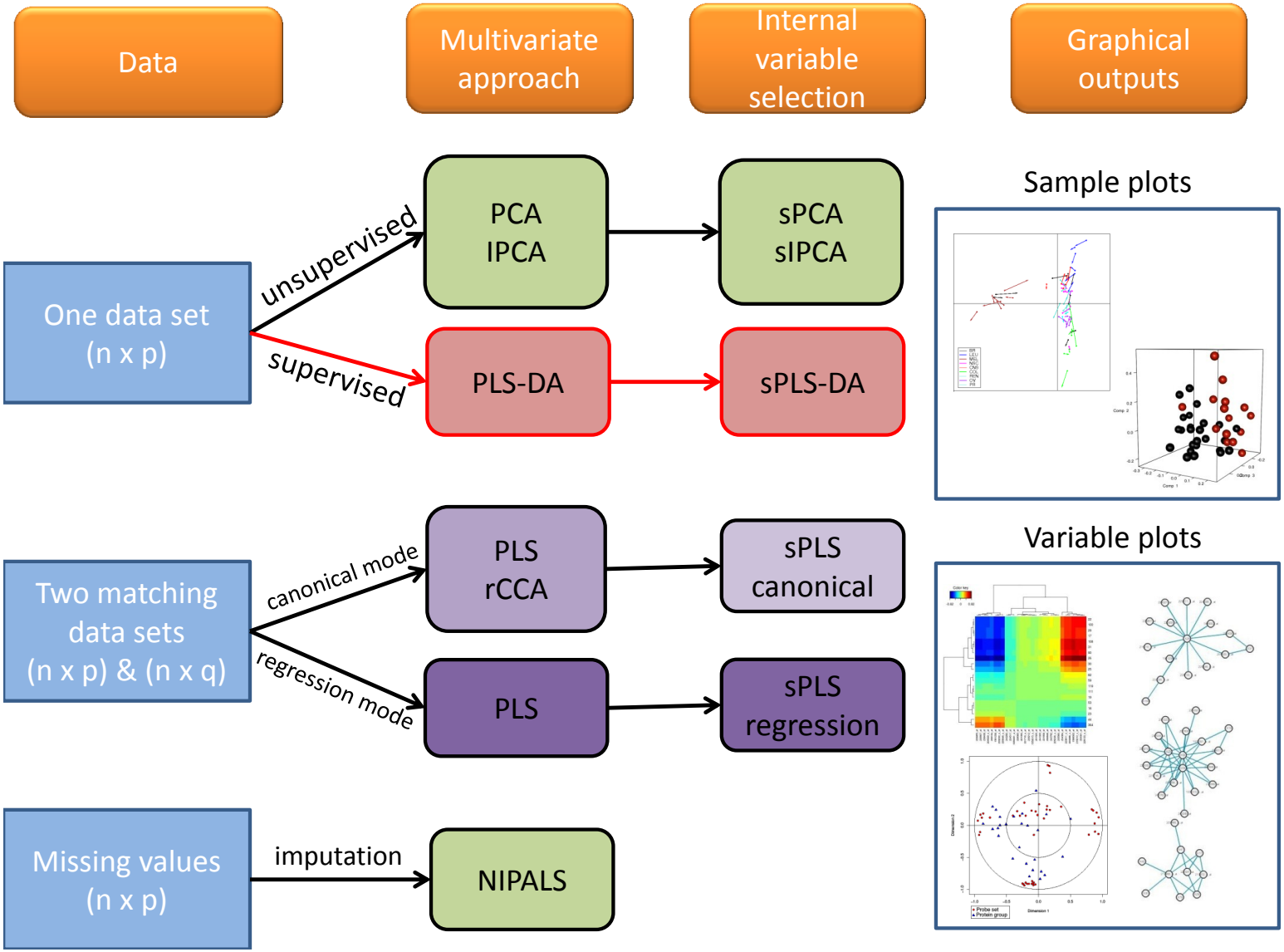
In particular, in `mixOmics`, focus is on:

- Data integration
- Variable selection
- Computationally efficient methodologies for large biological data sets
- Interpretable graphical outputs

mixOmics is an R package dedicated to the exploration and the integrative analyses of high dimensional biological data sets.

- Website
  - R tutorials
  - Newsletter
- Web Interface
  - User friendly interface
  - Comprehensive results page

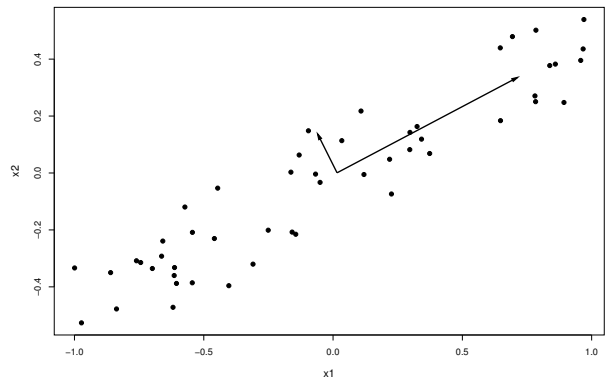
-**Lê Cao et al. (2009)** integRomics/**mixOmics**: an R package to unravel relationships between two omics data sets, *Bioinformatics*


Data
Multivariate approach
Internal variable selection
Graphical outputs
One data set (n x p)
unsupervised
supervised
PCA IPCA
PLS-DA
sPCA sIPCA
sPLS-DA
Two matching data sets (n x p) & (n x q)
canonical mode
regression mode
PLS rCCA
PLS
sPLS canonical
sPLS regression
Missing values (n x p)
imputation
NIPALS
Sample plots
Variable plots

Principal Component Analysis: PCA

Seek the best directions in the data that account for most of the variability

→ principal components: artificial variables that are linear combinations of the original variables:



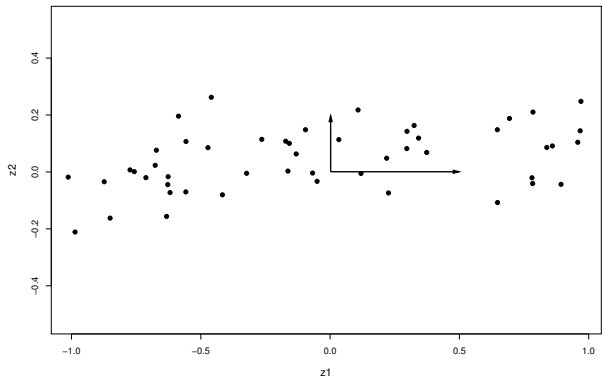
$$\underset{(n)}{\boldsymbol{c}} = \underset{(nxp)}{\boldsymbol{X}} \quad \underset{(p)}{\boldsymbol{v}}$$

- $\boldsymbol{c}$ is a linear function of the elements of $\boldsymbol{X}$ having maximal variance
- $\boldsymbol{v}$ is called the associated loading vector

Principal components cont.

The new PCs form a vectorial subspace of dimension $< p$

Project the data on these new axes.

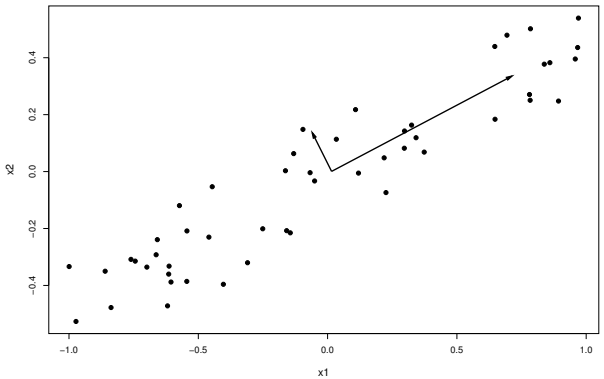


$\rightarrow$ approximate representation of the data points in a lower dimensional space

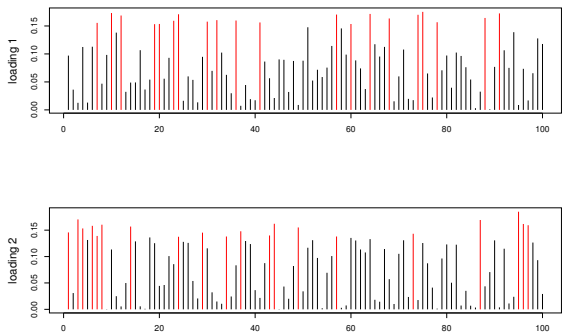
Problem:
Interpretation difficult with very large number of (possibly) irrelevant variables

sparse Principal Component Analysis: sPCA

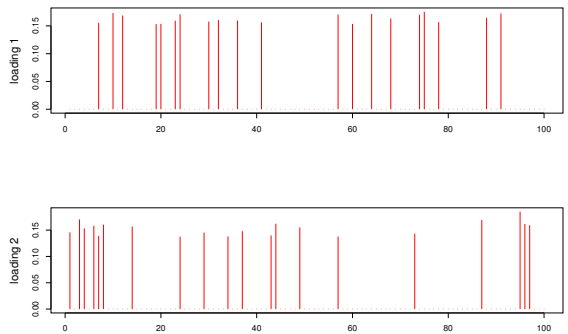
Principal components



loading vectors (PCA)



sparse loading vectors (sPCA)



The principal components are linear combinations of the original variables, variables weights are defined in the associated loading vectors.

sparse PCA computes the sparse loading vectors to remove irrelevant variables using lasso penalizations (**Shen & Huang 2008**, *J. Multivariate Analysis*).

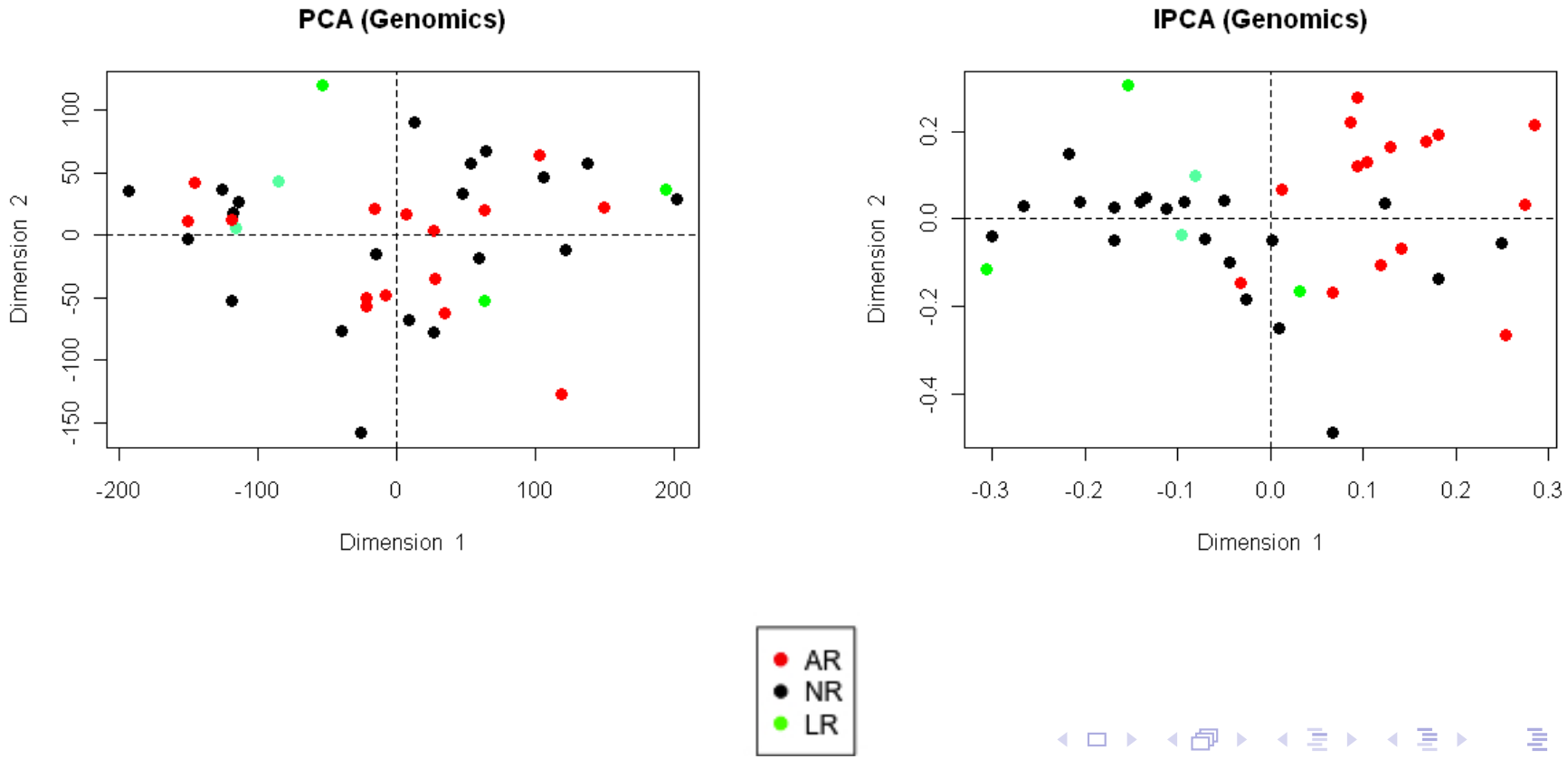
IPCA is based on Independent Component Analysis (ICA):

- assumes non Gaussian data distribution ($\neq$ PCA).
- 'blind source' signal separation.
- seeks for a set of independent components ($\neq$ PCA).
- Combines the advantages of both PCA and ICA.
- The PCA loadings are transformed via ICA to obtain independent loading vectors and independent principal components.
- sparse IPCA also developed to select the variable contributing to the independent loading vectors

Yao, F. Coquery, J. and Lê Cao, K-A. 2012 Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets, *BMC Bioinformatics*.

Illustration of PCA and IPCA

Sample representation for the kidney transplant study (3 groups of rejection status patients)



PLS - Discriminant Analysis

- Similarly to Linear Discriminant Analysis, classical PLS-DA looks for the best components to separate the sample groups.
- As opposed to PCA/IPCA methods, it is a supervised approach.
- In addition to this, sPLS-DA searches for discriminative variables that can help separating the sample groups.
- Evaluation of the discriminative power of the selected variables using external data sets or cross-validation.

Lê Cao K-A., Boitard S. and Besse P. (2011) Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems, *BMC Bioinformatics*, 12:253.

Illustration of sPLS-DA

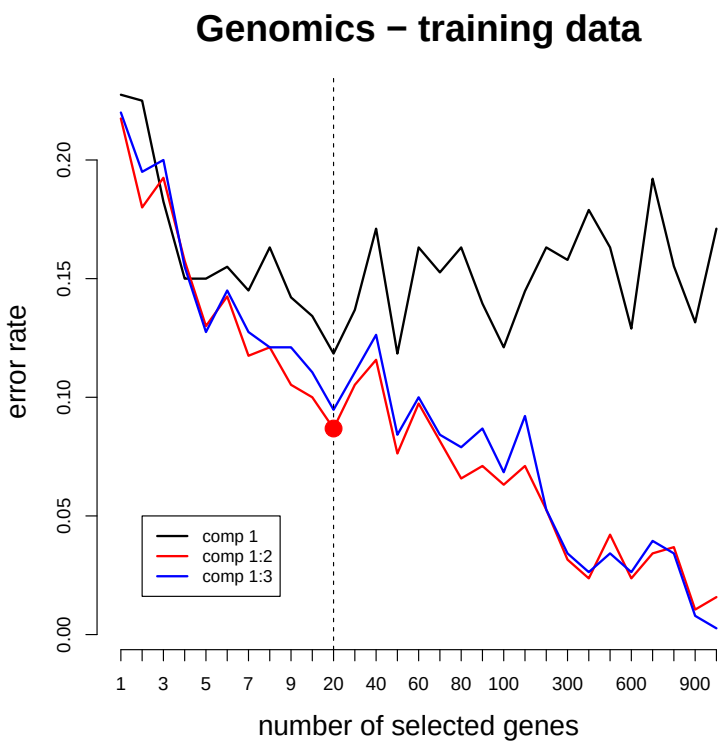


Figure: Tuning the number of var. to select with cross-validation

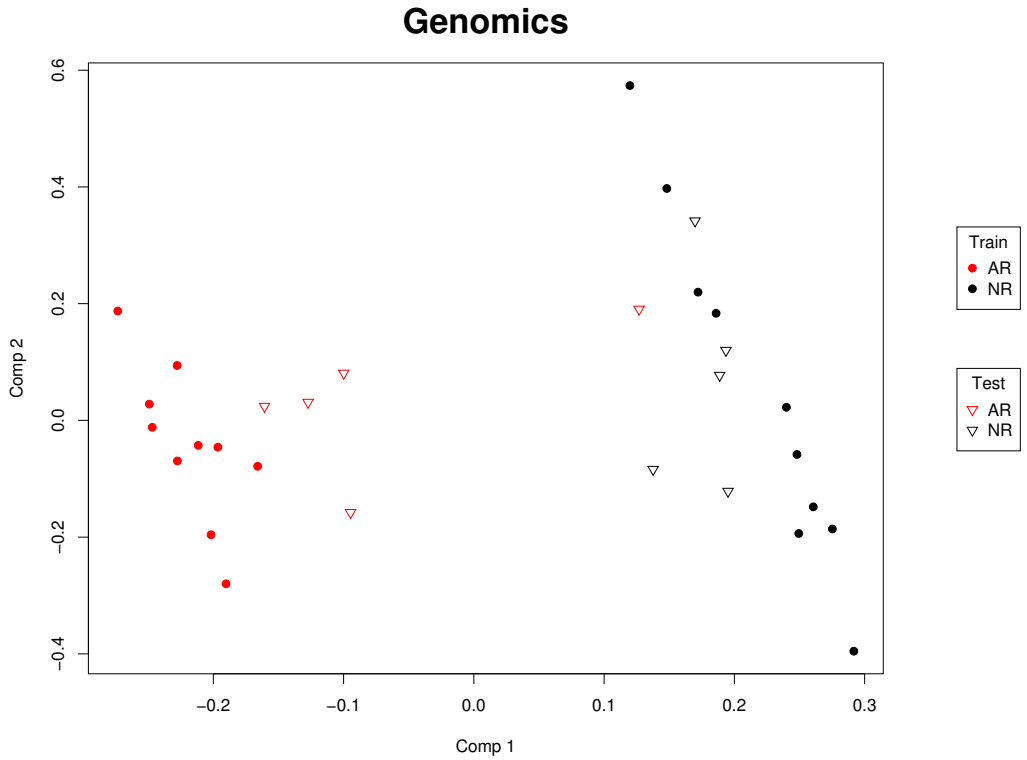


Figure: Predicting the class of the test set samples

Aims:

- unravel the correlation stucture between two data sets
- select co-regulated biological entities across samples

$\rightarrow$ select and integrate in a one step procedure the different types of data

- Partial Least Squares regression maximises the covariance between each linear combination (components) associated to each data set
- sparse PLS has been developed to include variable selection from both data sets
- Two modes are proposed to model the relationship between the two data sets ('regression' and 'canonical')

Lê Cao et al. (2008), *SAGMB*, **Lê Cao et al. (2009)**, *BMC Bioinformatics*

Illustration of sparse PLS: sample plot

sPLS aims at selecting correlated variables (genes, proteins) across the same samples by performing a multivariate regression.
Regression: explain the protein abundance w.r.t the gene expression "$\Rightarrow$ relationship".

- The latent variables (components) are determined based on the selected genes and proteins
  $\rightarrow$ give more insight into the samples similarities.
- Unsupervised approach

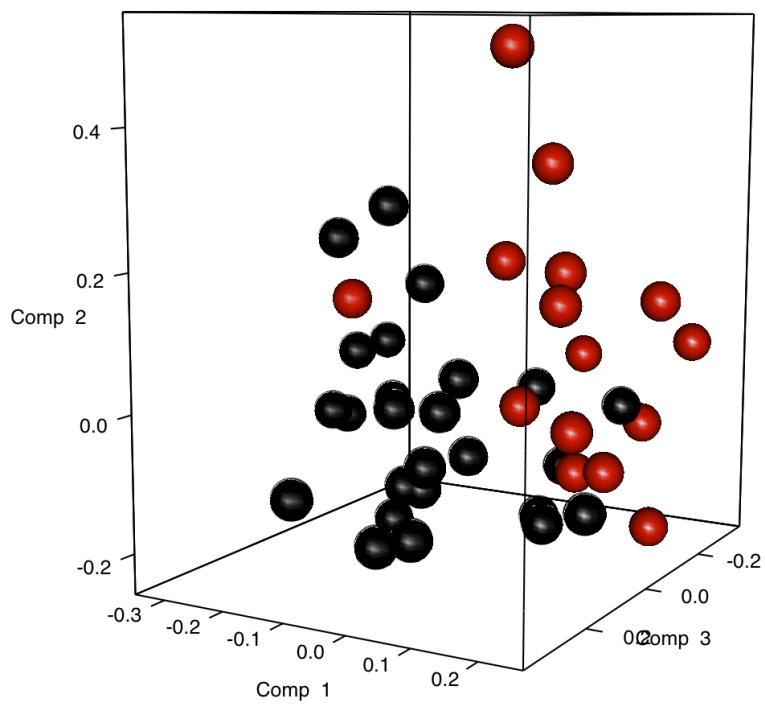
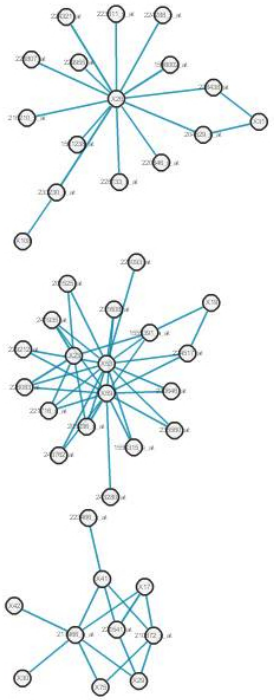


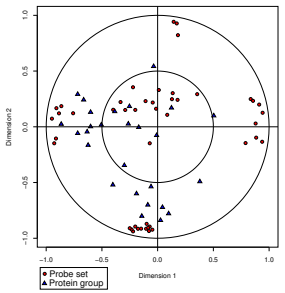
Illustration of sparsePLS: variable plot

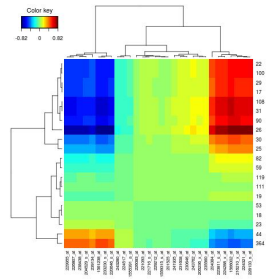


Relevance networks are bipartite graphs directly inferred from the sPLS components.

Some other insightful graphical outputs:

- correlation circle plots
- clustered image maps





González I., Lê Cao K.-A., Davis, M.D. and Déjean S. Visualising association between paired 'omics' data sets. *In revision.*

Illustration of sPLS canonical mode

Selects correlated variables across the same samples and highlights the correlation structure between the two data sets.

Canonical mode: "$\Leftrightarrow$ relationship"

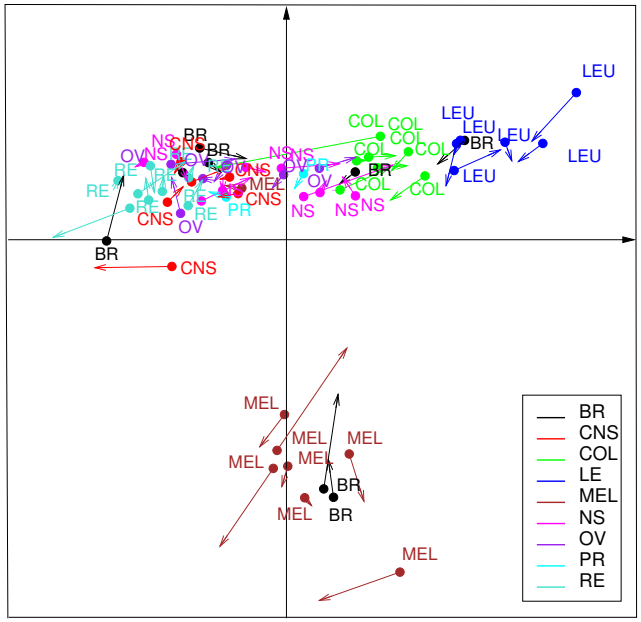


Figure: Arrow plot to highlight the similarties between 2 data sets

Cross-over design

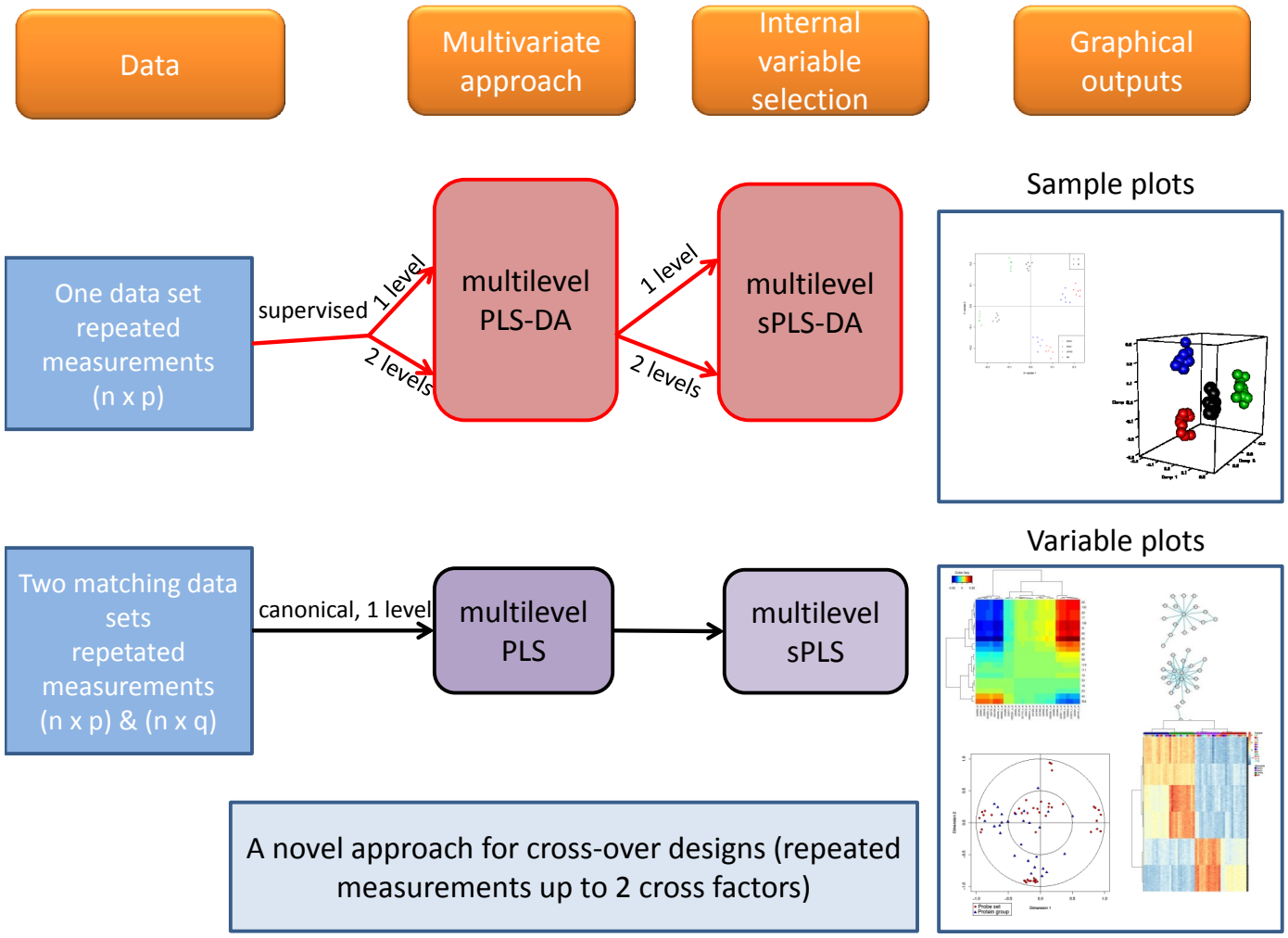
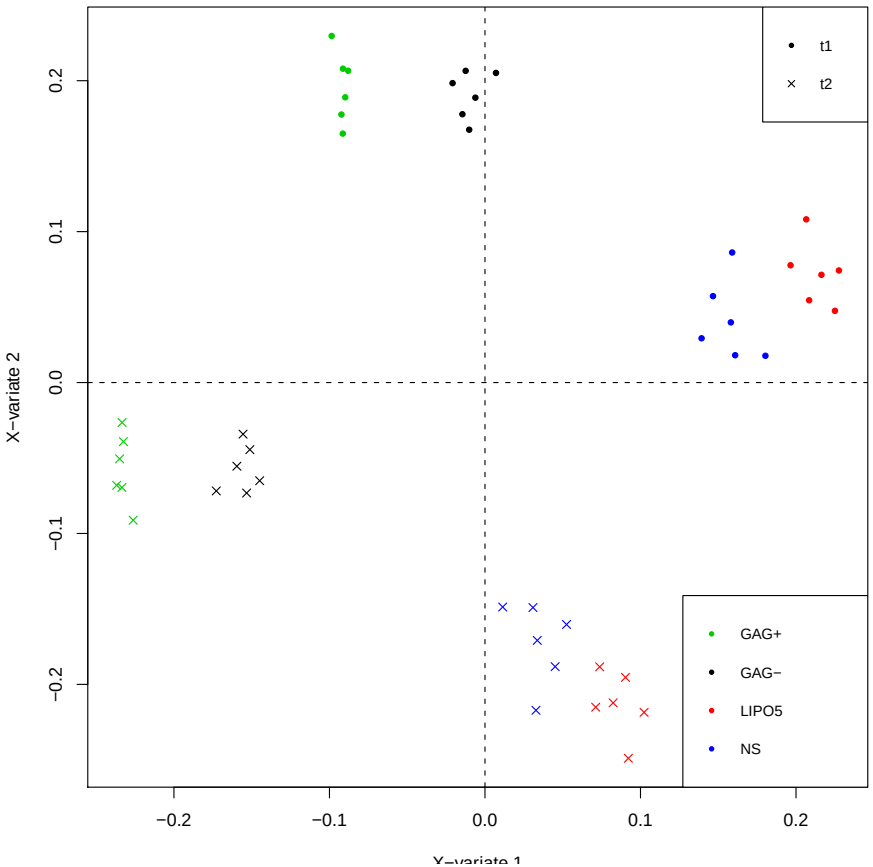
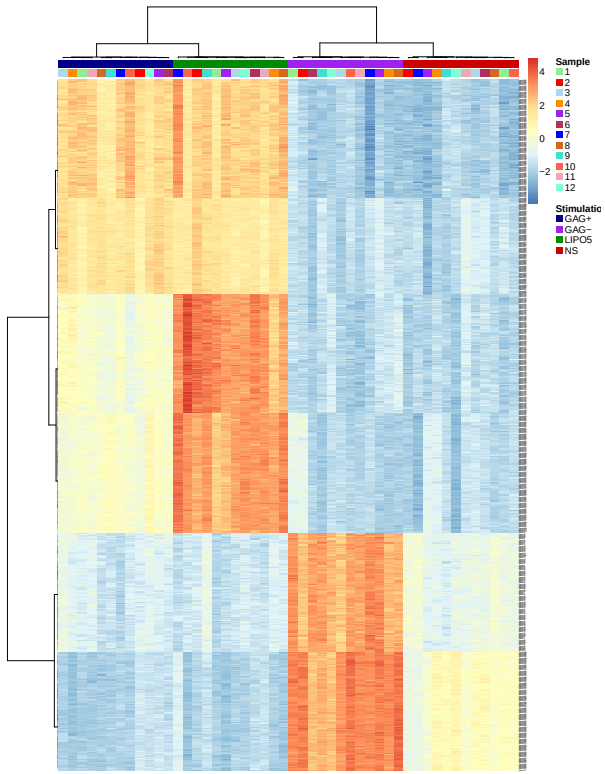


Illustration of multilevel sPLS-DA





Conclusions

The exploratory and integrative approaches are:

- flexible and can answer various types of questions.
- can highlight the potential of the data.
- enable to generate new hypotheses to be further investigated.

Future work includes:

- Cross-platform comparison
- Integration of multiple data sets (unsupervised and supervised)
- Time-course experiments

Acknowledgements

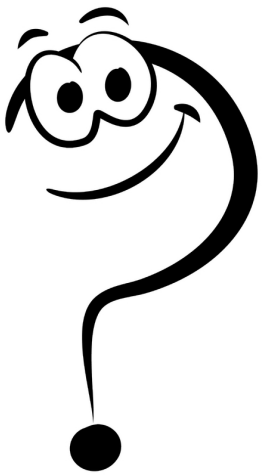
mixOmics team

Sébastien Déjean	Univ. Tlse
Ignacio González	Univ. Tlse
Xin Yi Chua	QFAB

Other contributors to mixOmics

Jeff Coquery	QFAB, Sup'Biotech
Benoit Liquet	Univ. Bordeaux
Eric Fangzhou Yao	QFAB, Shanghai Univ of Finance and Economics
Pierre Monget	QFAB, CESI

Questions?



http://mixomics.qfab.org

mixomics@math.univ-toulouse.fr

k.lecao@uq.edu.au