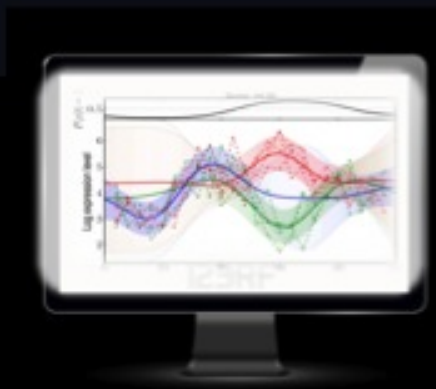
It is all about mixOmics!

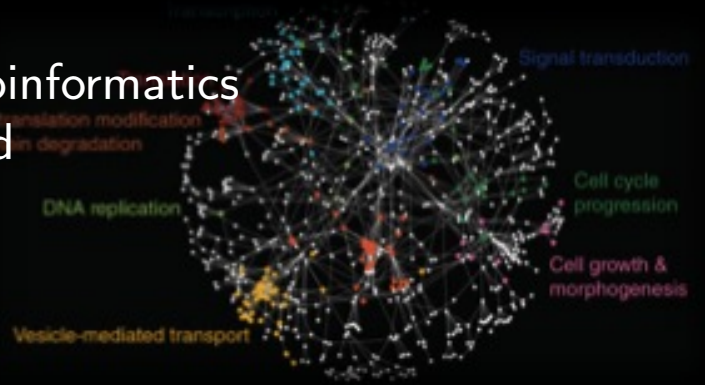
or

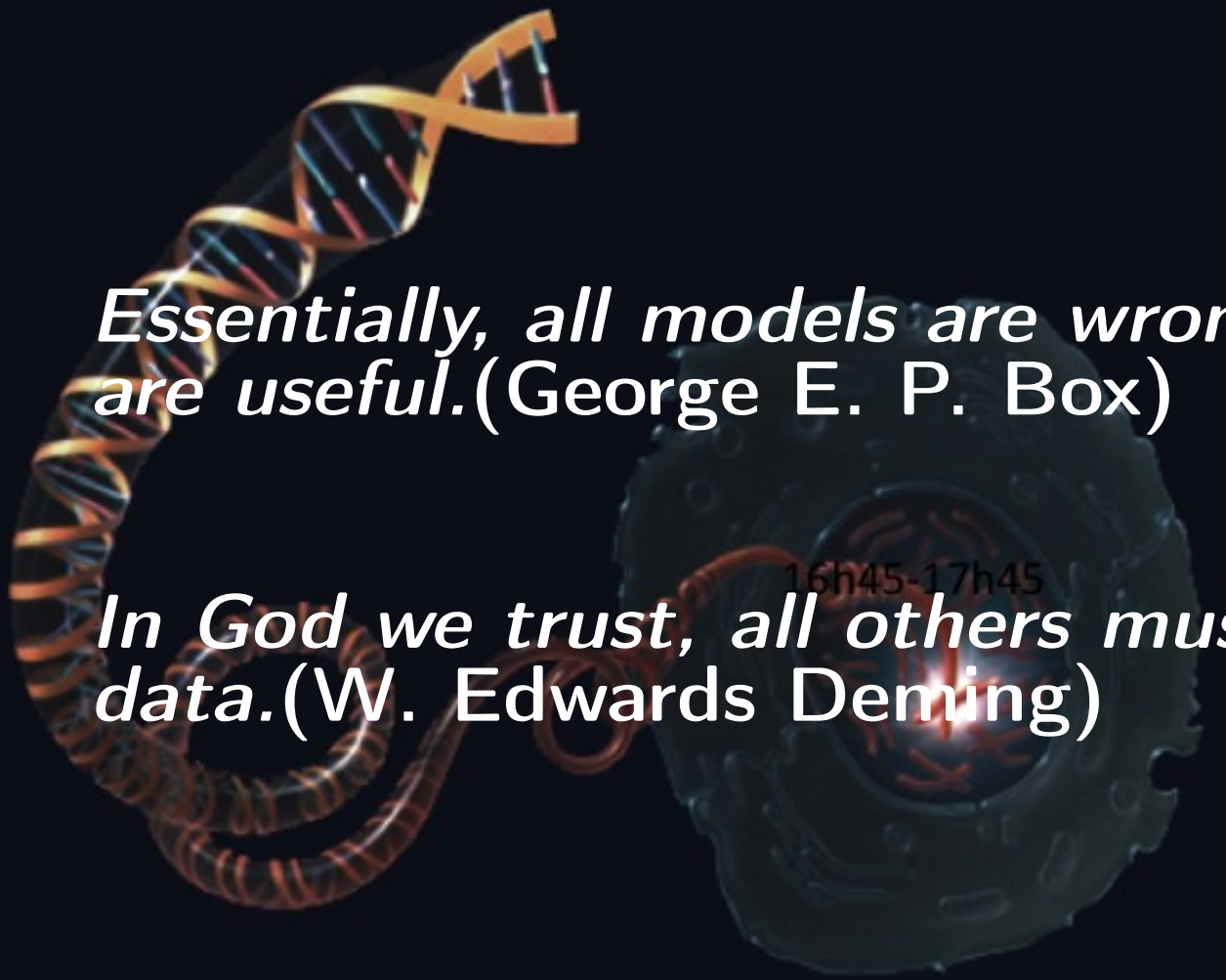
a summary of last year's research

Kim-Anh Lê Cao

Queensland Facility for Advanced Bioinformatics
The University of Queensland







Essentially, all models are wrong, but some are useful.(George E. P. Box)

In God we trust, all others must bring data.(W. Edwards Deming)

Outline of this talk and research questions

1. It is all about `mixOmics`
2. What is the common information contained in different experiments?
3. What are the correlated features across different time course experiments?
4. Can we combine similar experiments performed in different labs or/and on different platforms?

mixOmics: philosophy

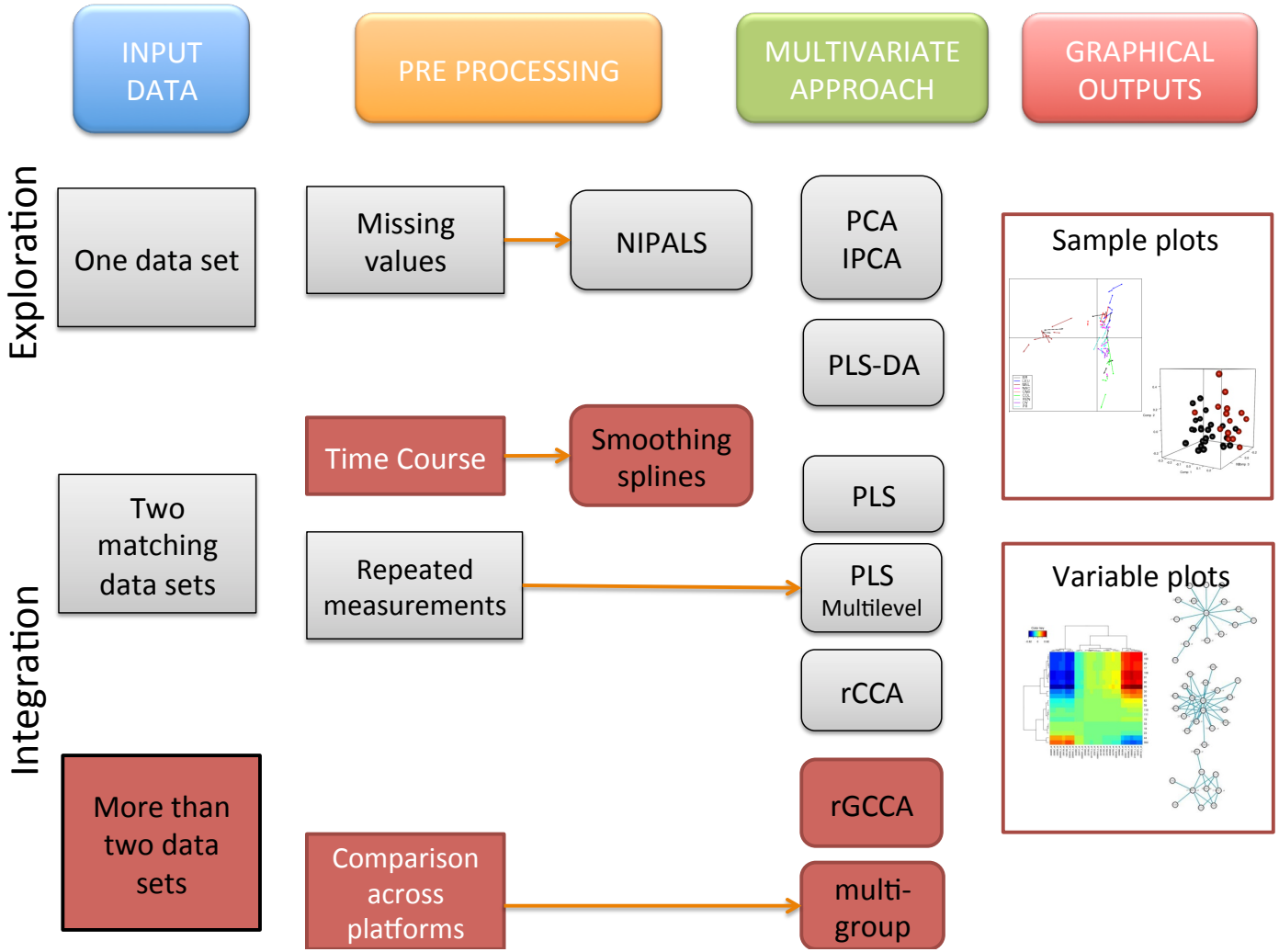
is originally an R package developed for the statistical exploration and integration of large biological data sets.

- Development of statistical multivariate approaches
- Variable selection included in the methodologies
- Graphical outputs
- User-friendly use: website, R package, web interface

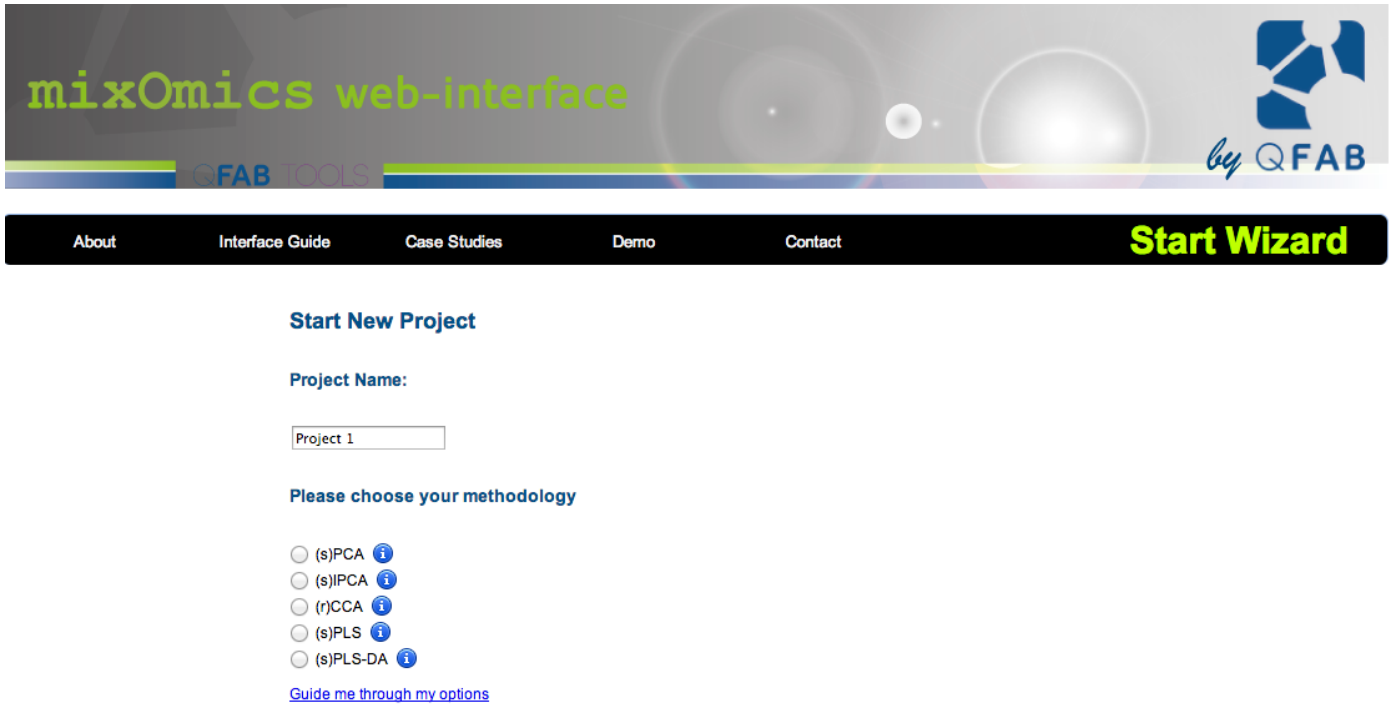
First R CRAN release in May 2009.

Lê Cao et al. mixOmics: an R package to unravel relationships between two omics data sets, *Bioinformatics*, 25 (21).

Framework


INPUT DATA
PRE PROCESSING
MULTIVARIATE APPROACH
GRAPHICAL OUTPUTS
Exploration
One data set
Missing values
NIPALS
PCA
IPCA
PLS-DA
Sample plots
Integration
Two matching data sets
Time Course
Smoothing splines
Repeated measurements
PLS
PLS Multilevel
rCCA
Variable plots
More than two data sets
Comparison across platforms
rGCCA
multi-group

Web interface



http://www.qfab.org/mixomics/

Statistical data integration

- What is the common information contained in different experiments?
  - Requires the same samples measured across different experiments
  - Large number of variables → variable selection
  - Emphasis on graphical outputs to ease results interpretation

Work in close collaboration with various **French teams** (INRA, Univ Toulouse, Supelec Paris).

Work in progress!

Linear multivariate approaches

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

PCA is an (almost) compulsory first step in exploratory data analysis to:

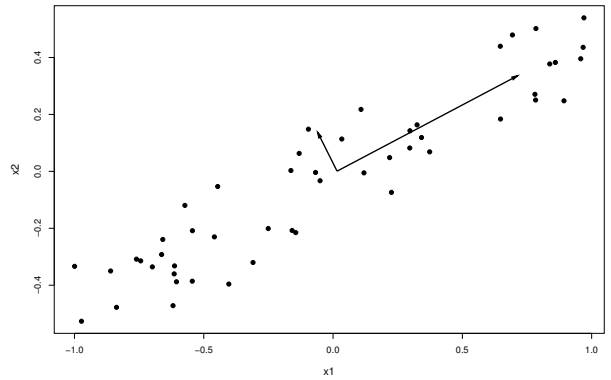
- Have a first understanding of the underlying data structure
- Identify bias, experimental errors, batch effects

Problem: interpretation can be difficult with very large number of (possibly) irrelevant variables.

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:



$$c_1 = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

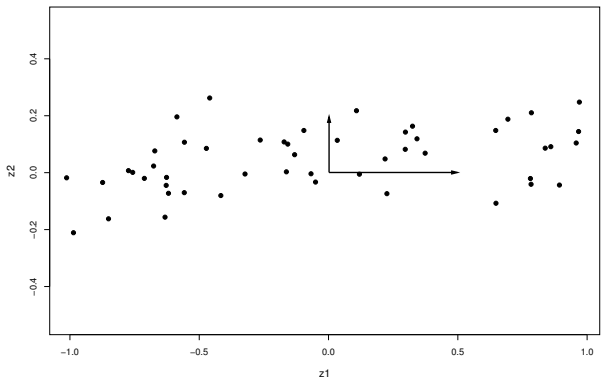
where

- where c_1 is the **first** principal component with max. variance
- $\{w_1, \ldots, w_p\}$ are the weights in the linear combination
- $\{x_1, \ldots, x_p\}$ are the gene expression profiles.

All PCs are mutually orthogonal. $(c_1, c_2, \ldots)$

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

Project the data on these new axes to summarize the information related to the variance.



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

PCA is an (almost) compulsory first step in exploratory data analysis to:

- Have a first understanding of the underlying data structure
- Identify bias, experimental errors, batch effects

Problem with PCA: interpretation can be difficult with very large number of (possibly) irrelevant variables.

Remember that the principal components are linear combinations of the original variables:

$$c = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

A clearer signal could be observed if some of the variable weights $\{w_1, \ldots, w_p\}$ could be set to 0 for the irrelevant variables:

$$c = 0 * x_1 + w_2 x_2 + \cdots + 0 * x_p$$

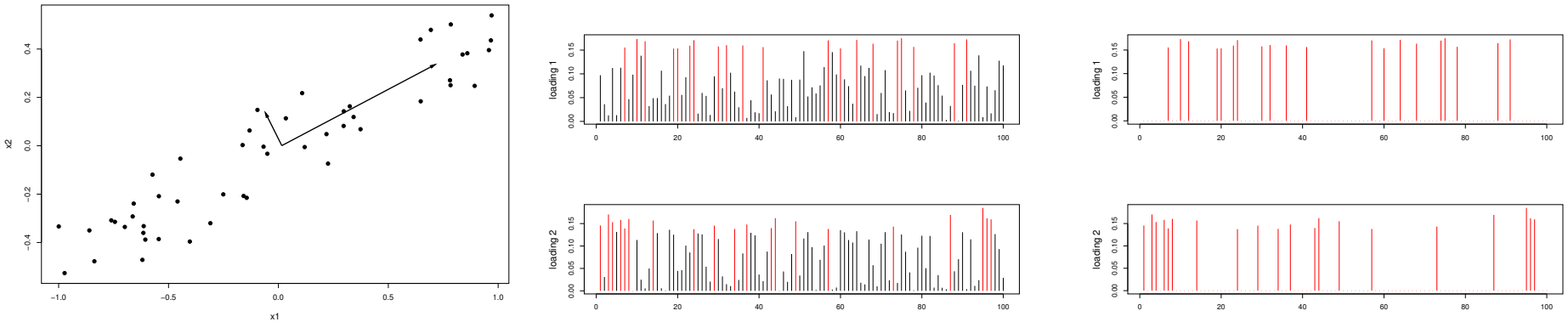
These variables weights are defined in the loading vectors.
Important weights = important contribution to the PC.
Similar weights = correlated variables.

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*).

sparse Principal Component Analysis: sPCA

sparse PCA: sparse loading vectors to remove noisy or irrelevant variables which determine the principal components.

$\rightarrow$ Solving PCA through least squares problem (SVD) allows to include regularization parameters

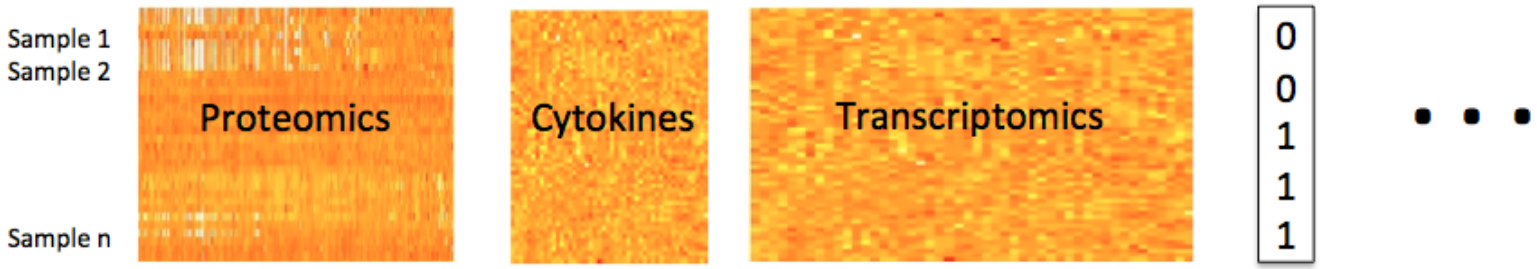
$$\min_{\mathbf{v_h}} ||X_h - \mathbf{u}_h \mathbf{v}_h^T||_F^2 + P_\lambda(\mathbf{u}_h)$$

P_λ is a penalty function with tuning regularization parameter λ

$\rightarrow$ use Lasso penalization, or soft-thresholding
$\rightarrow$ obtain sparse loading vectors, with very few non-zero elements

Shen, H., Huang, J.Z. 2008. Sparse principal component analysis via regularized low rank matrix approximation, *J. Multivariate Analysis*.

Integration of multiple data sets



- Define the relationships between the different data sets
- Select relevant biological entities which are correlated across the different data sets

Kidney transplant study: Transcriptomics and proteomics study of 40 patients with kidney transplant, rejecting ($n_1 = 20$) or not ($n_2 = 20$) the transplant, PROOF Centre, UBC.

Multiblock analysis: Regularized Generalised CCA

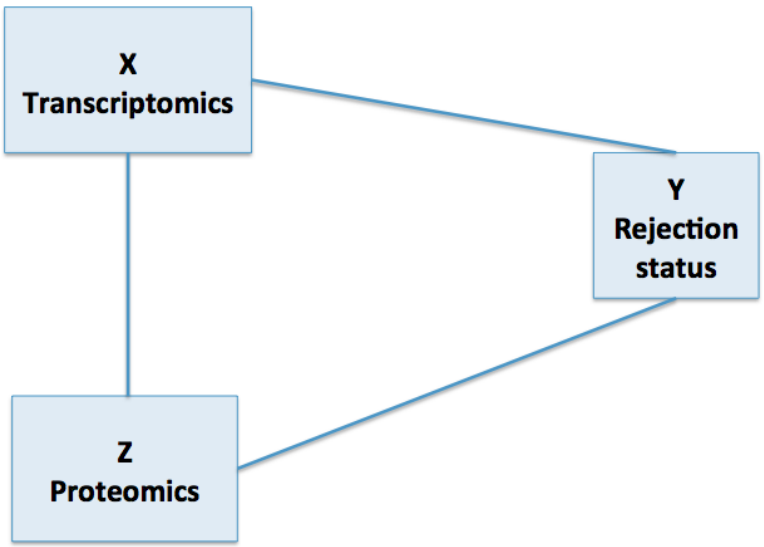
Canonical Correlation Analysis (CCA) maximises the correlation between 2 data sets, but numerical limitations when $p >> n$
$\rightarrow$ regularization needed

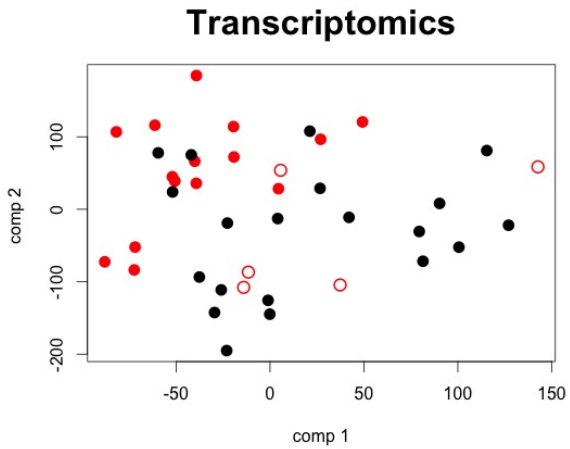
- RGCCA generalizes CCA to more than 2 data sets
- Constitutes a general framework for many multi-block data analysis methods
- Objective: seeks linear combinations of variables from different platforms (a.k.a. *block*):
  - block components explain their own block well and/or
  - block components that are assumed to be connected are highly correlated.

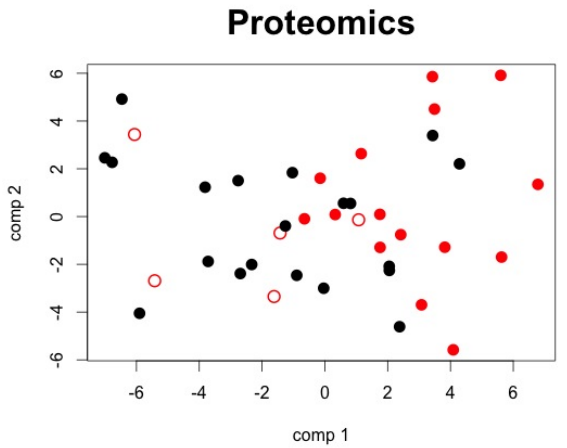
Tenenhaus, A., Tenenhaus, M (2011) Regularized Generalised Canonical Correlation Analysis, *Psychometrika*, 76 (2).

Illustration: design 1

Design	X	Y	Z
X	0	1	1
Y	1	0	1
Z	1	1	0





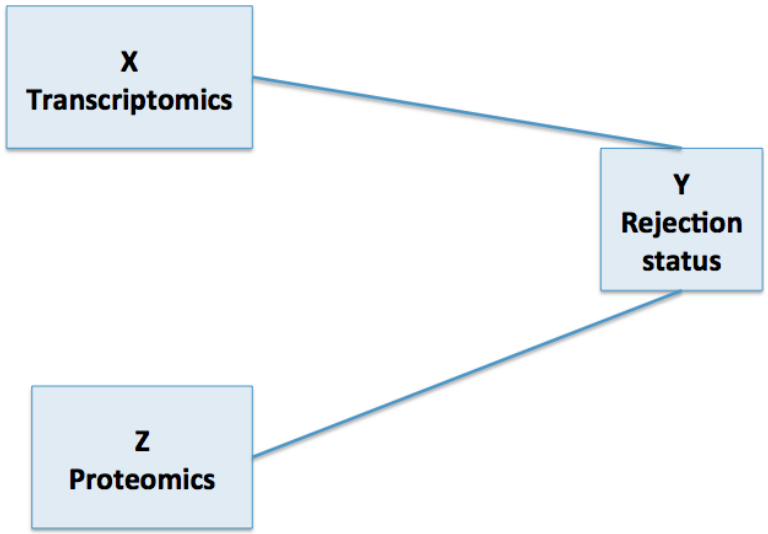


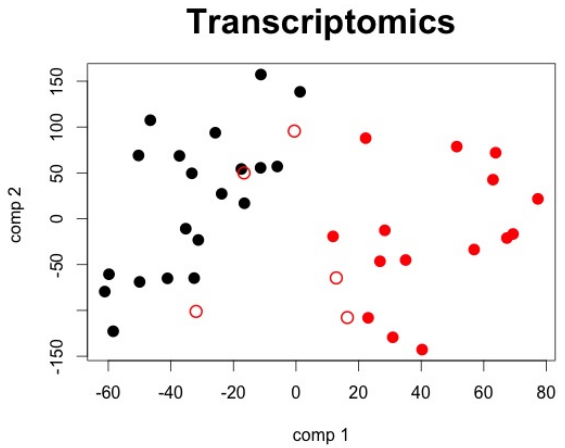
AR LR NR

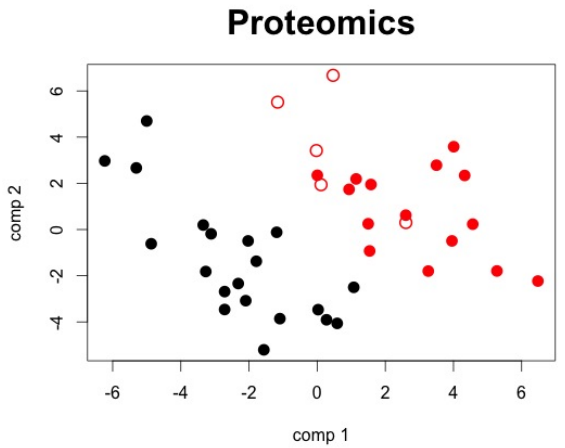
Kidney transplant study

Illustration: design 2

Design	X	Y	Z
X	0	**0**	1
Y	**0**	0	1
Z	1	1	0



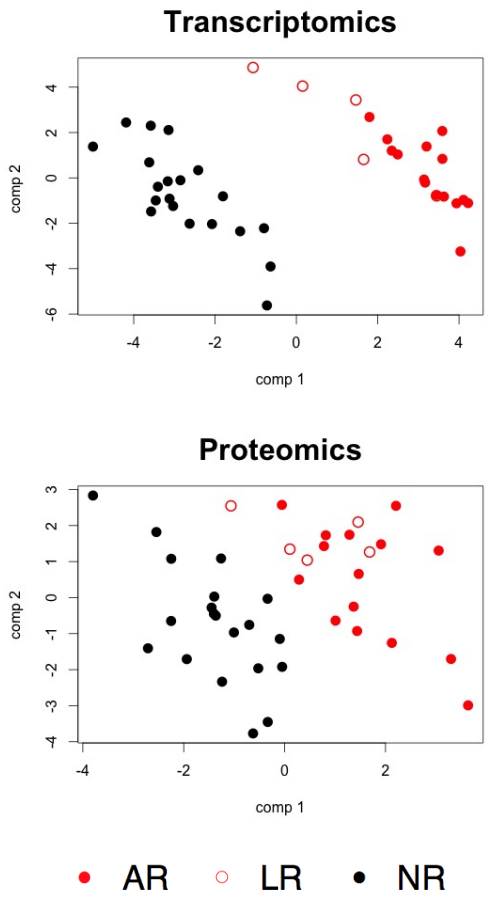


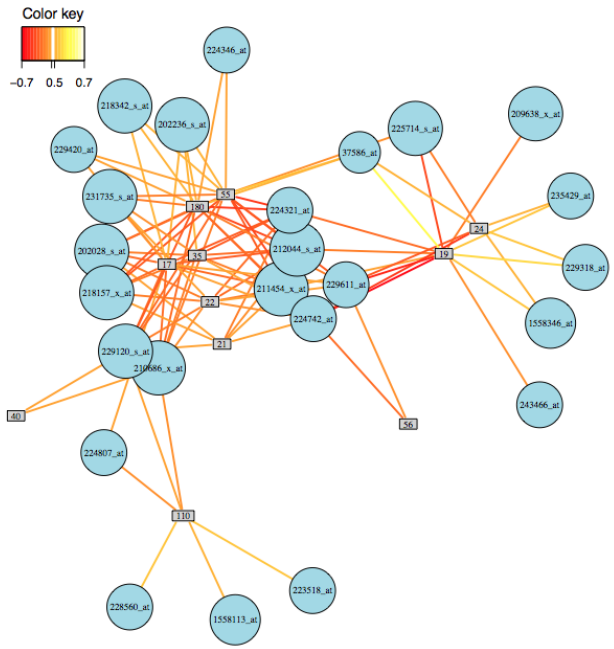


Kidney transplant study

AR LR NR

New developments include a sparse RGGCA (sGCCA) to select variables across the different platforms





Relevance networks Gonzales, Lê Cao et al. (2012), J. Data Mining

Conclusions: statistical data integration

Multivariate integrative approaches

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

Approaches implemented in our R package mixOmics:

- implements 6 different methodologies plus their sparse variants
- data integration, variable selection, graphical outputs
- includes a web-associated interface (will soon be released through Galaxy too)
- tutorials: http://www.math.univ-toulouse.fr/~biostat/mixOmics

Time course data integration

- What are the correlated features across different time course experiments?
  - Requires the same samples measured across different time course experiments
  - 'Noisy' variables $\rightarrow$ variable selection
  - Hypothesize that correlated biological entities belong to the same biological pathway
  - Time delay

Work in close collaboration with various **French teams** (INRA, Univ Toulouse, Supelec Paris) , **Dr Kathy Ruggiero** (Auckland) and **Jasmin Straube** (Ph.D student).

Work in progress!

Why PCA can 'fail' to summarize the data?

- In some time course experiments, the subject variation can be larger than the time variation
- PCA makes the assumption that samples are independent of each other
- In univariate analysis we use a paired t-test instead of a t-test
- In multivariate analysis we use a multilevel approach:
  - different sources of variation can be separated (treatment effect within subjects and differences between subjects)
  - gain in power

Multilevel approach

- The variation in the data is separated: within matrix and between matrix
- Multivariate tools can then be applied on the within matrix (Westerhuis, 2008)
- We can take into account the repeated measures design of the experiment

VEGFC Study: Human lymphatic endothelial cells were treated in vitro with recombinant VEGF-C for 16 time points in triplicates, CAGE data (FANTOM5, Riken Institute).

Liquet*, B., Lê Cao*, K-A., Hocini, H., Thiébaut, R. A novel approach for biomarker selection and the integration of repeated measures experiments from two platforms, *BMC Bioinformatics*.

VEGFC study: high individual effect

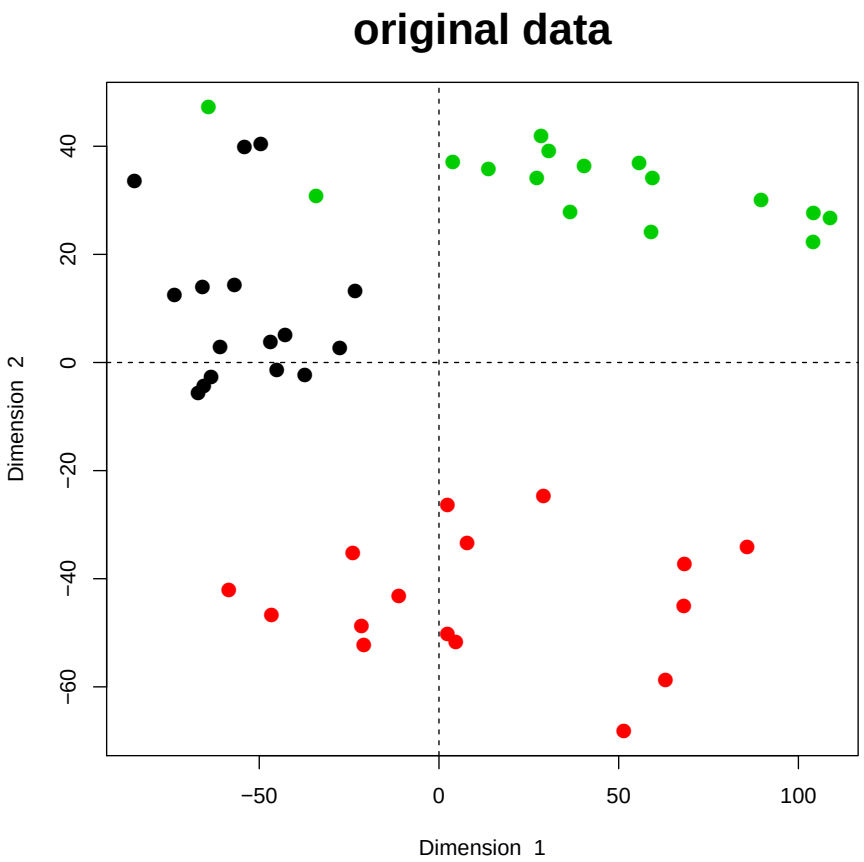


Figure: color = patient

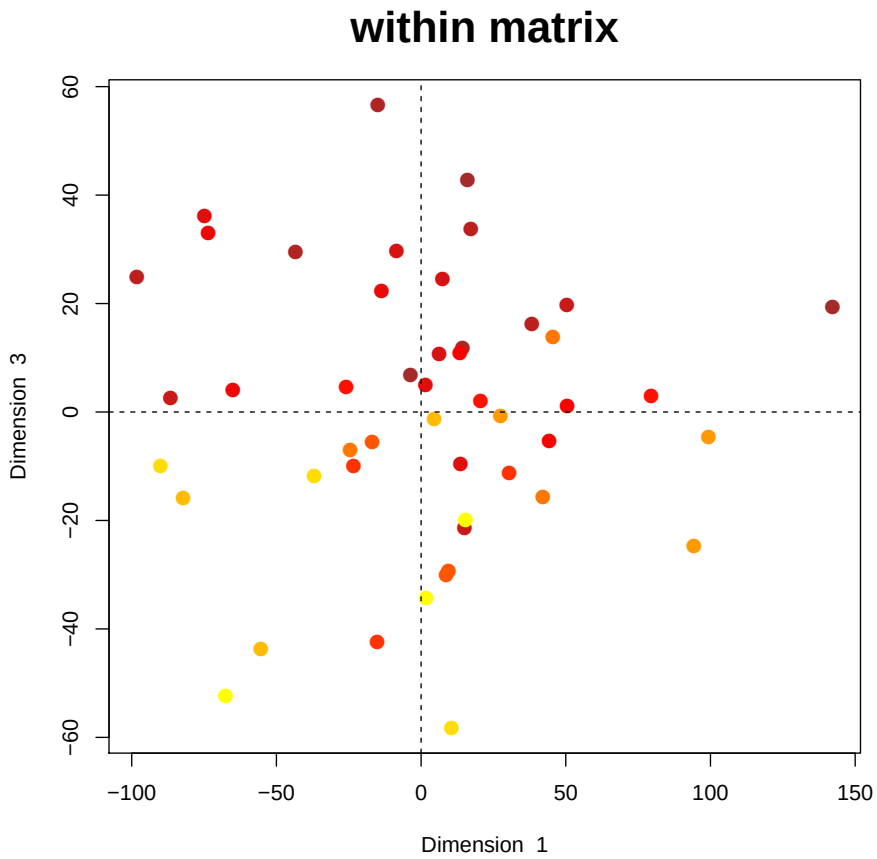
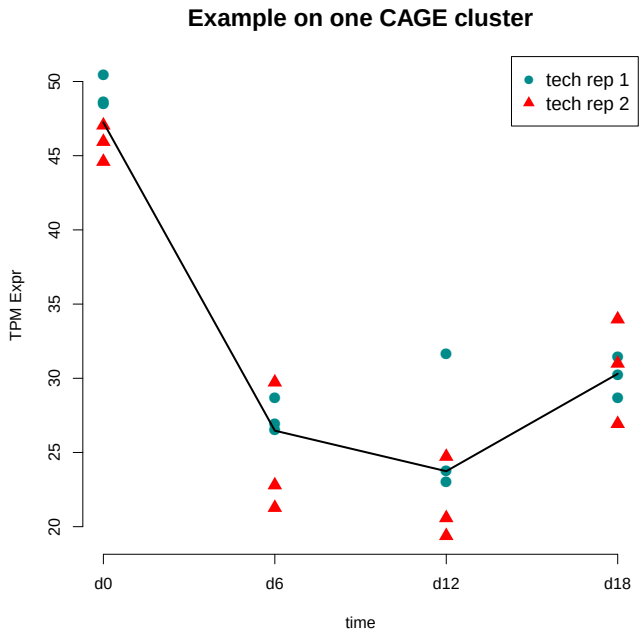


Figure: gradient color = time

Modelling trajectories: cubic smoothing splines

Aim: summarize the trajectory of each variable

- Use cubic smoothing splines to summarize each profile
- The derivative between each time point can be estimated
- Fit a non-supervised algorithm to cluster the profiles based on the derivative (k-means, SOM)



Déjean et al. (2008), Clustering Time-Series Gene Expression Data Using Smoothing Spline Derivatives *Eurasip J.*

Example with K-means

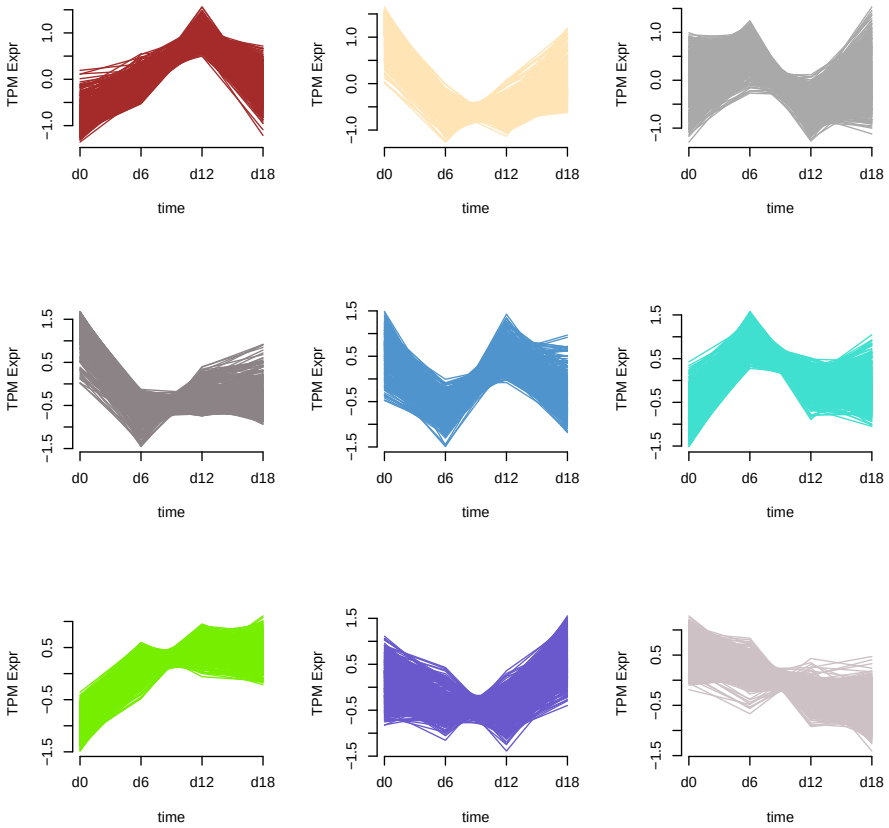


Figure: K-means on derivatives

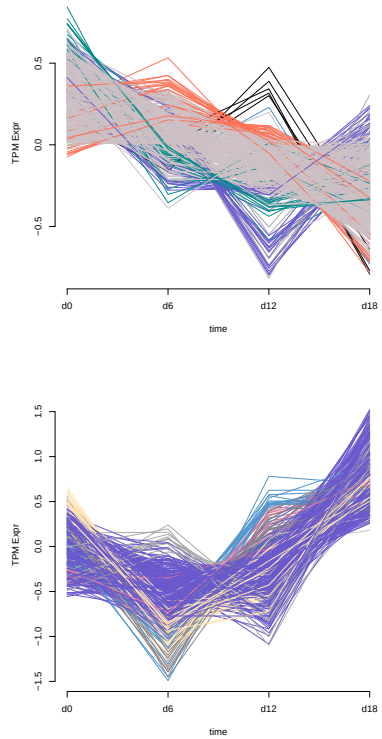


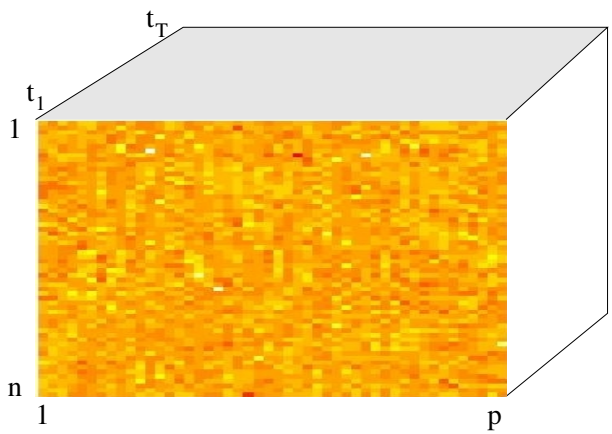
Figure: K-means on original data

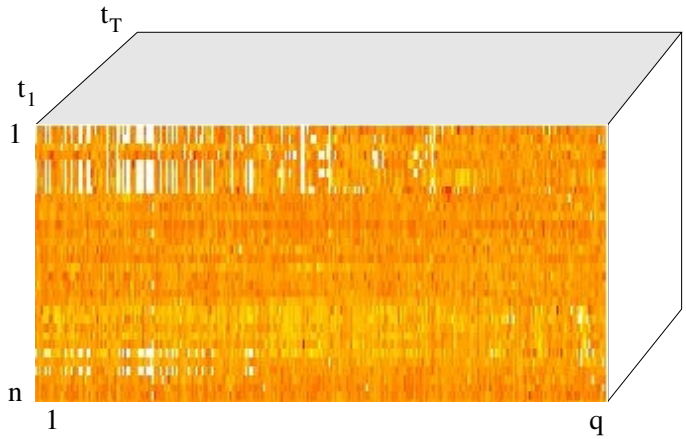
There is a link between smoothing splines and LMM

- Appropriate for correlated repeated measures
- Enables interpolation of missing values

- Fits into a linear mixed model framework (Verbyla et al. 1999)
  - no parameters to tune,
  - flexibility of the model
  - model the shape of the trajectories
  - assesses the variability for each feature (technical , biological variability)
    $\rightarrow$ filters 'noisy' variables

Integration of two longitudinal studies





- Select correlated profiles across time, between and within each data set.
- But difficult to deal with 3D data sets!
- Projection-based multivariate methods can integrate data sets of 2 dimensions (sparse Partiel Least Squares (PLS) regression, RGCCA ...).

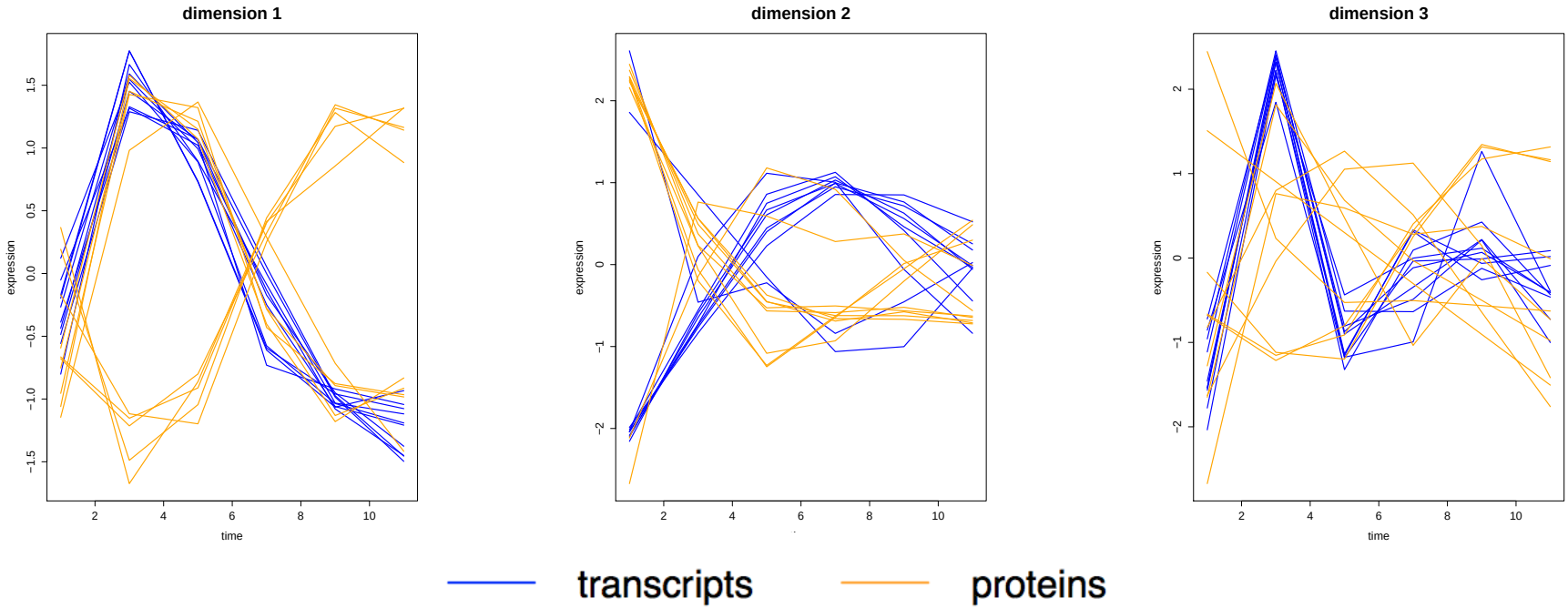
Step 1: use cubic smoothing splines to reduce one dimension (samples dimension)

Step 2: apply sGCCA on the estimated splines to identify correlated profiles both within and between the two data sets

→ modelisation of the trajectories
→ filtering of the profiles
→ integration of two types of data
→ selection and clustering of the correlated time profiles

Kidney transplant study: Transcriptomics and proteomics study of 40 patients with kidney transplant, rejecting ($n_1 = 20$) or not ($n_2 = 20$) the transplant. Follow up on 5 time points (weeks), PROOF Centre, UBC.

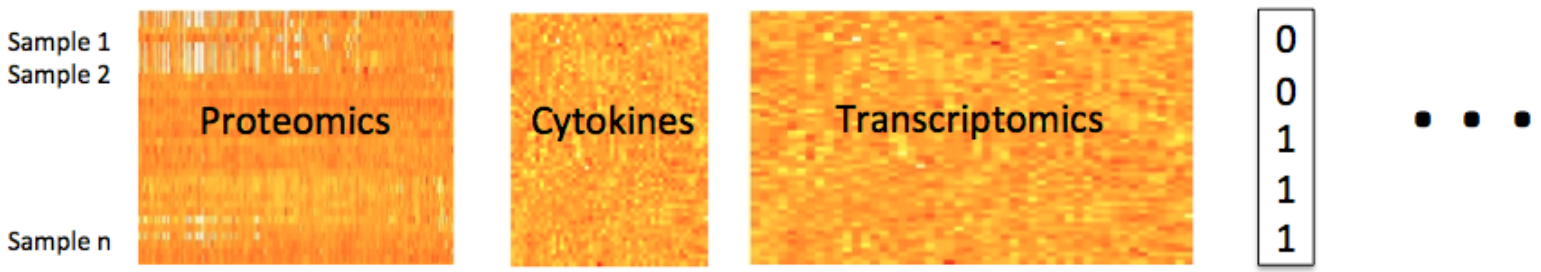
Profile clusters on kidney transplant study



- sPLS selects both transcripts and proteins which are positively or negatively correlated across time
- Quality of clusters decreases with the number of PLS components (dimensions) as obvious patterns cannot be extracted anymore

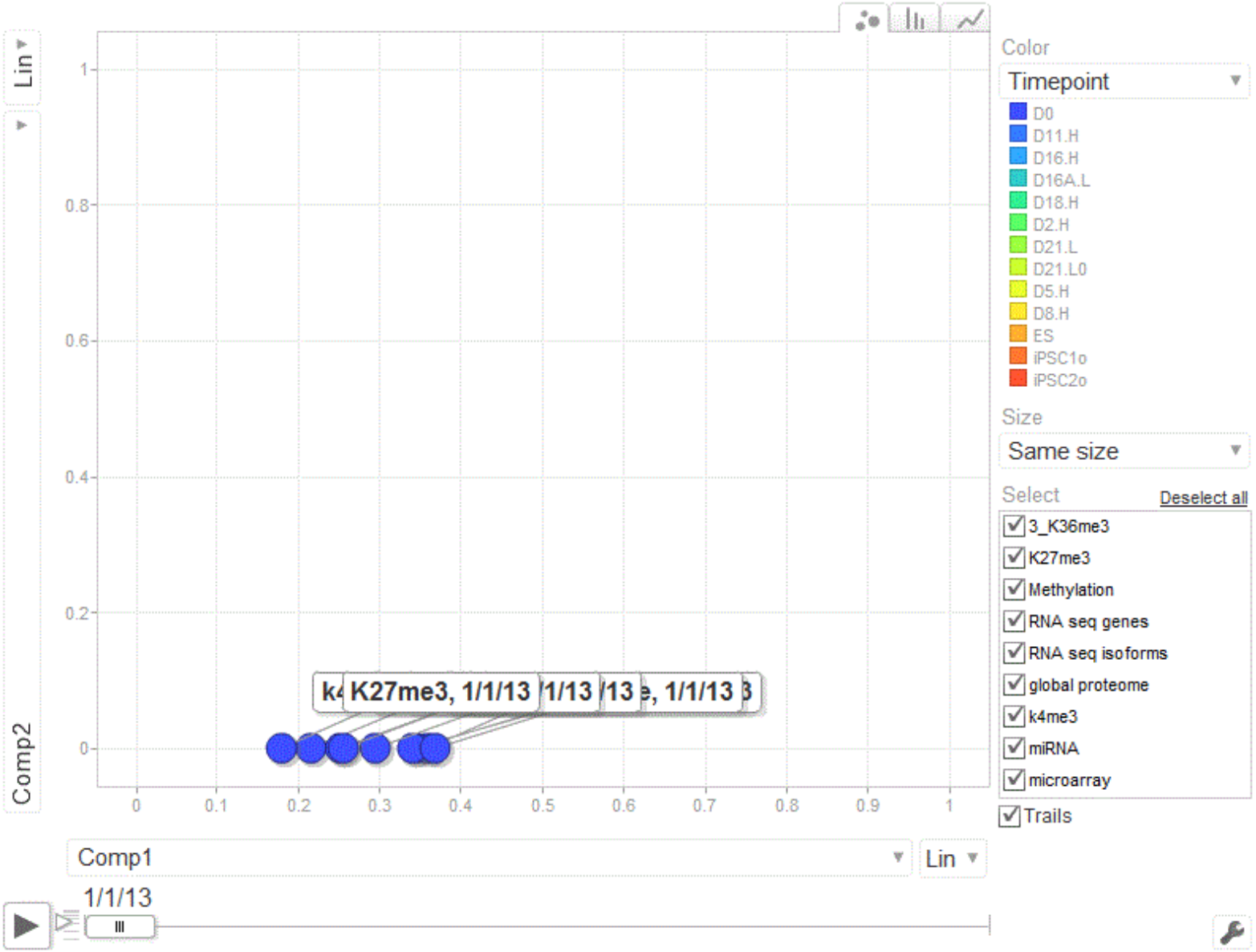
Integration of multiple data sets

Integrate heterogeneous data sets



Grandiose project: Longitudinal study of cell reprogramming across 8 time points. Multi platforms: 10 platforms: microarray, cell surface proteome, total proteome, RNA-seq isoform, RNA-seq genes, miRNA...

Results on Grandiose project



Conclusions: integrating longitudinal data sets

- Statistical exploratory and integrative tools to model and extract patterns in time course/longitudinal data
- Help generating new hypotheses, further statistical tests can then be applied
- Future directions: variable selection in the multi block case, biological interpretation of the gene lists, time delay, identifying discordant clusters across data sets for the same genes ...

Cross-platform comparison

- Can we combine similar experiments performed in different labs and/or on different platforms?
  - Not the same samples measured across different experiments
  - 'Noisy' variables
  - Experimental biases (batch effects)
  - Well known fact: microarray experiments across studies bring different results!
    $\rightarrow$ "But I would very much like to compare my gene signature from **my** experiment to my colleagues' gene signatures from **their** experiments!"

Aim: Identify genes diagnostic of the key characteristics of stem cells derived from independent studies

Work in close collaboration with **Stéphanie Bougeard**, **Aida Eslami** (ANSES) and **Florian Rohart** (AIBN, UQ). Work in progress!

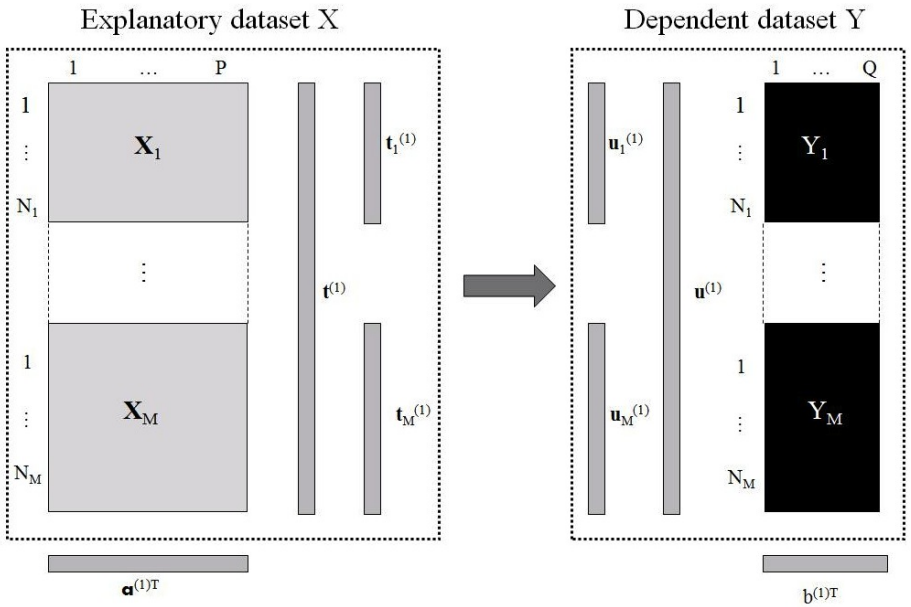
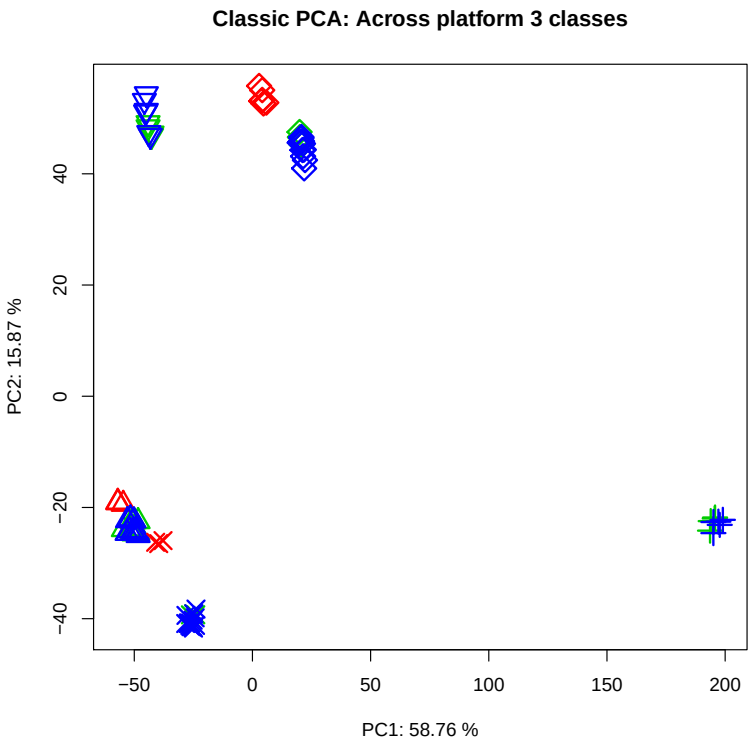


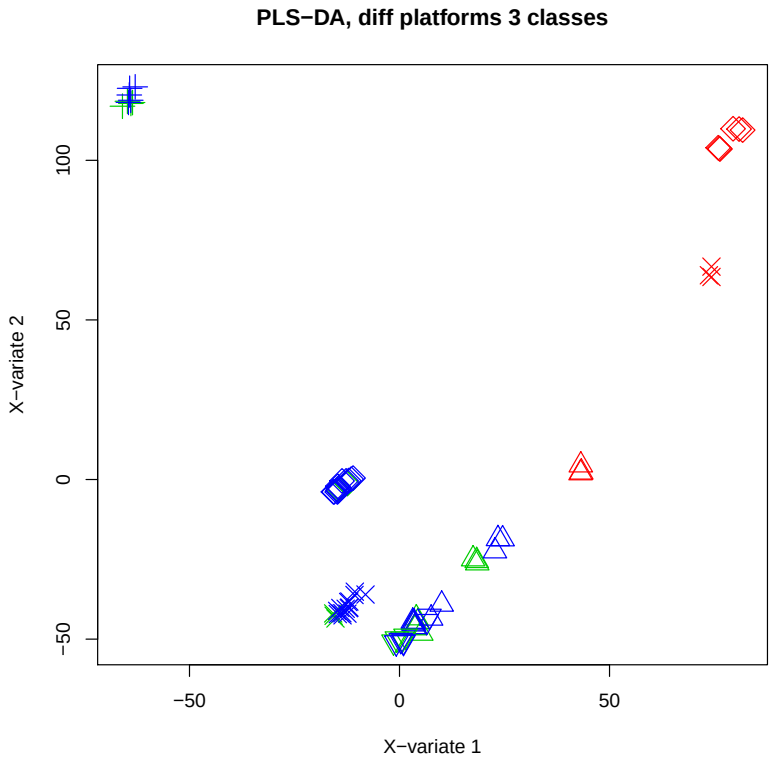
Fig. from Aida Eslami

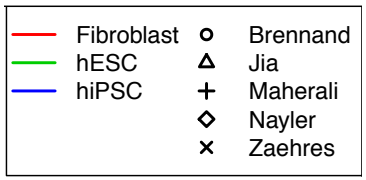
- X = several microarray experiments, performed in different labs but studying the same biological conditions
  $X = (|X_1|X_2|...|X_g|...X_G|)^T$, $p >> n$ ($p \sim 20K$, $n_g \sim 10 - 20$)
- Group = experiment/platform/lab Y = biological conditions
  $\rightarrow$ Supervised problem

We propose: multi-group analysis

What is a 'batch' effect?



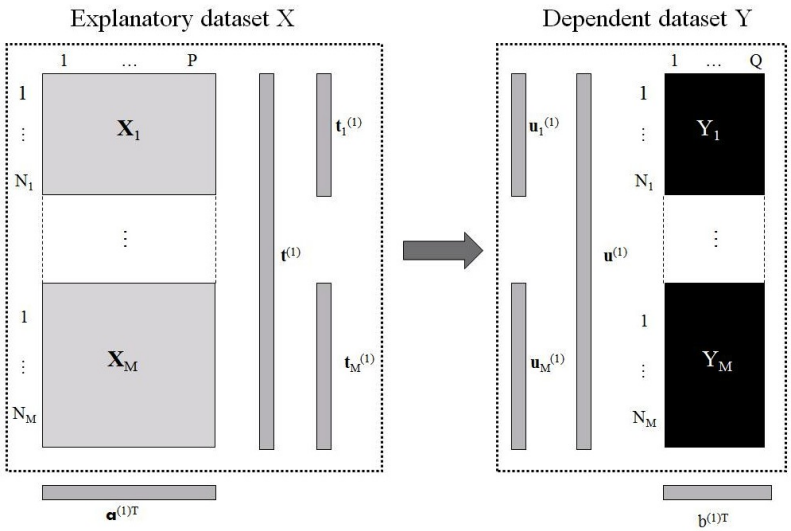




Classical PCA and PLS-DA both highlight a 'lab effect'.

→ Combining different experiments is a real challenge

Multi-group Analysis

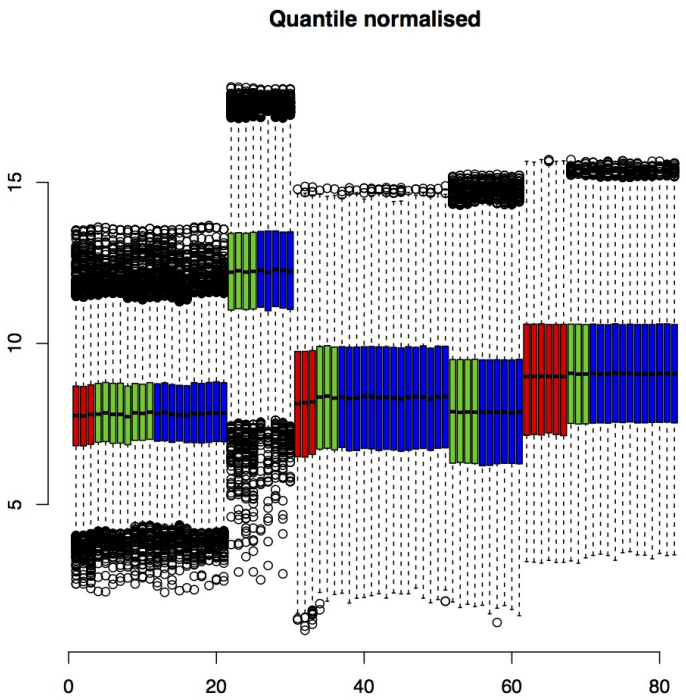


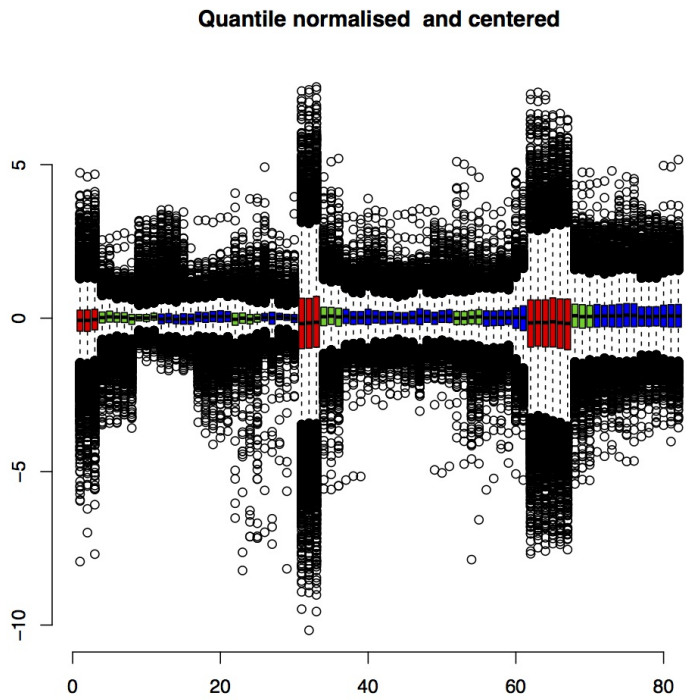
- Individuals are *a priori* structured into several groups
- within group part: group structure effect is removed
  between-group part: group structure effect is taken into account
- Partial and global components / loading vectors

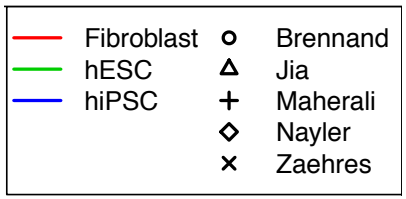
→ Understand the common stucture between the groups and within each group

→ Two block multi-group analysis by the means of PLS

The effect of centering the data







→ We cannot directly concatenate the normalised data (lab effect, LHS). By centering the data, we seem to keep the inner properties of the biological conditions (RHS).

The optimization problem to solve in multi-group PLS is:

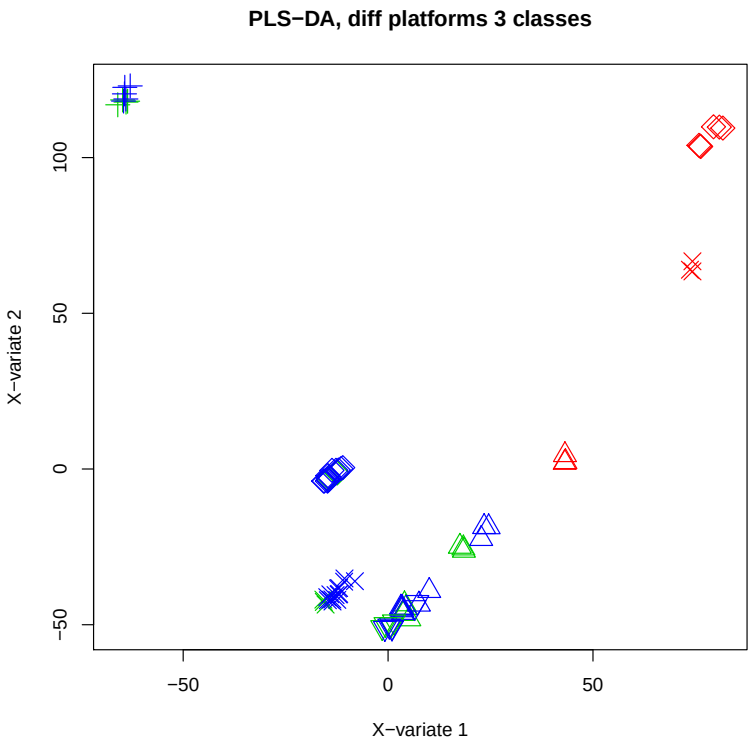
$$\max_{t_g, u_g} \sum_{1}^{G} n_g cov(t_g, u_g)$$

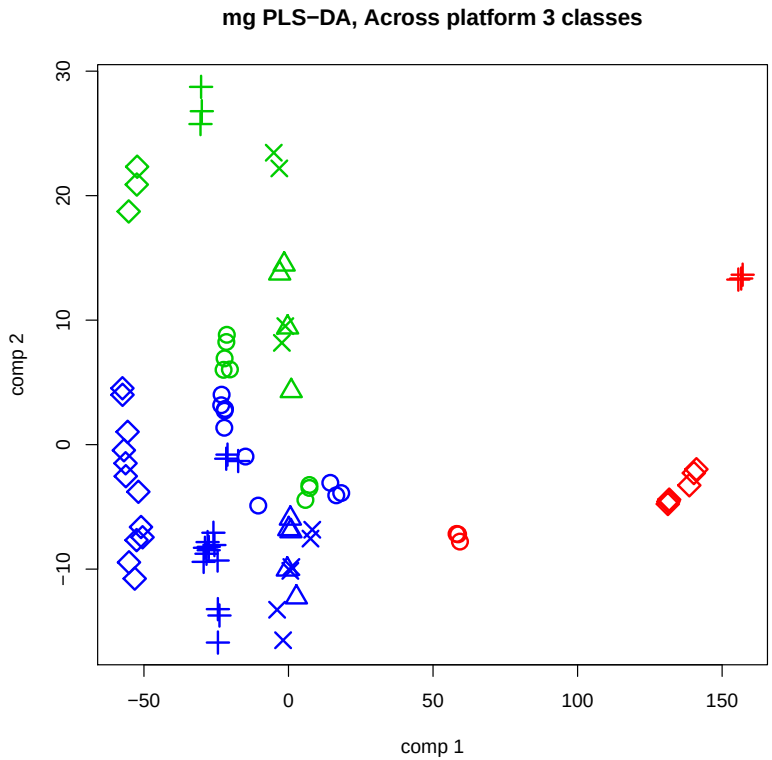
with $t_g = X_g a$, $u_g = Y_g b$ and $||a|| = ||b|| = 1$.
This is equivalent to

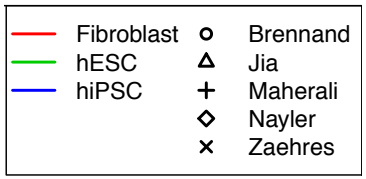
$$\max_{a,b} \sum_{1}^{G} b^T Y_g^T X_g a = a^T X^T Y b \quad \text{under the constraint } ||a|| = ||b|| = 1$$

We can compute the group components $t_g = X_g a$, $u_g = Y_g b$ as well as the global components $t = Xa$ and $u = Yb$ with X and Y centered per group. We also have the group loading vectors $a_g = X_g t_g$ and $b_g = Y_g u_g$.

Multi-group PLS-DA (in progress)

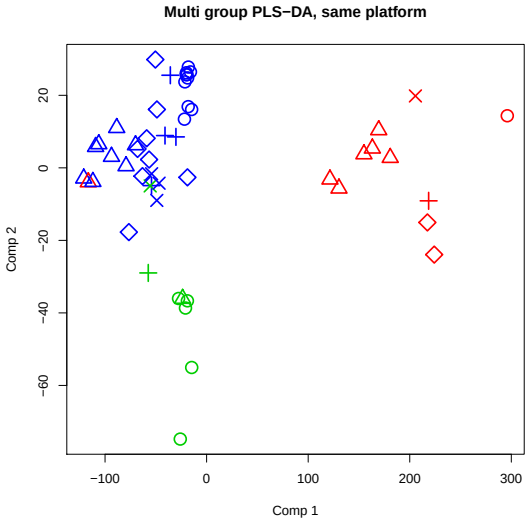


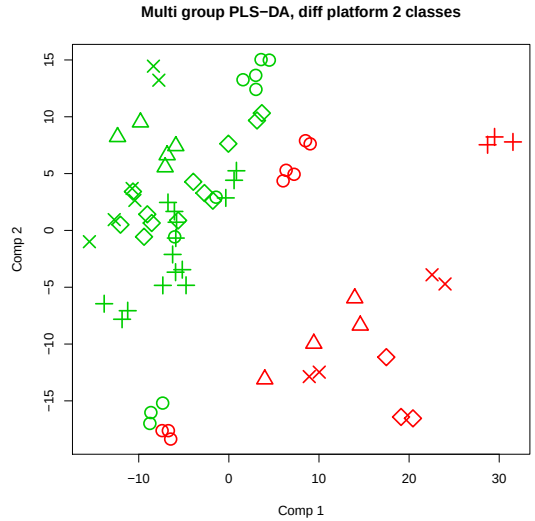


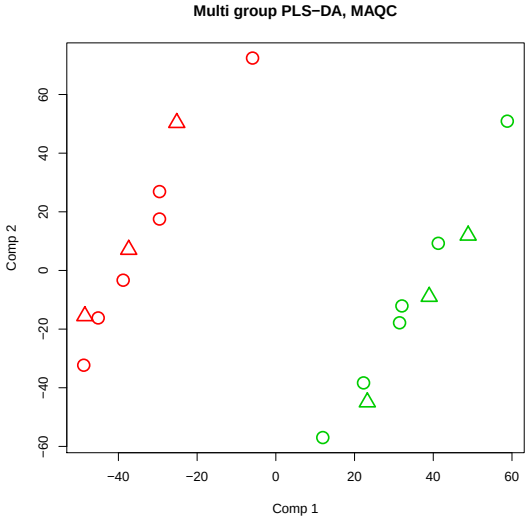


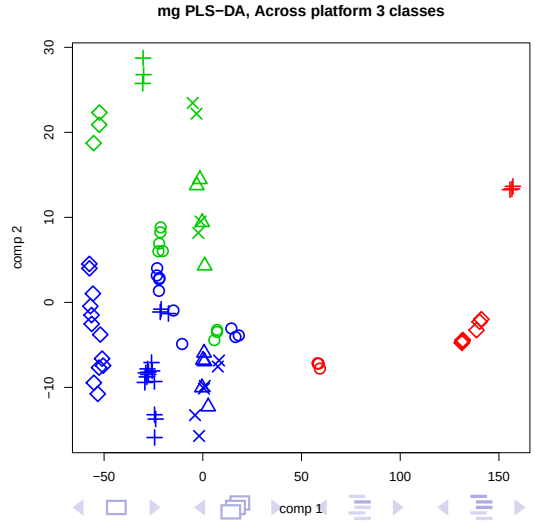
Classical PLSDA (LHS) vs multi-group PLSDA (RHS), the latter gives better (visual) results.

Multi-group PLS-DA, more data sets

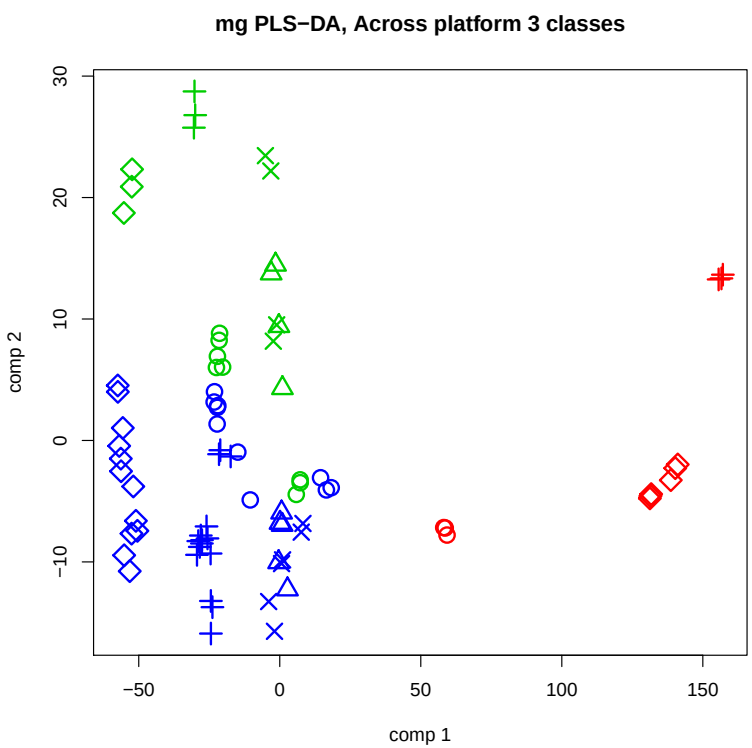


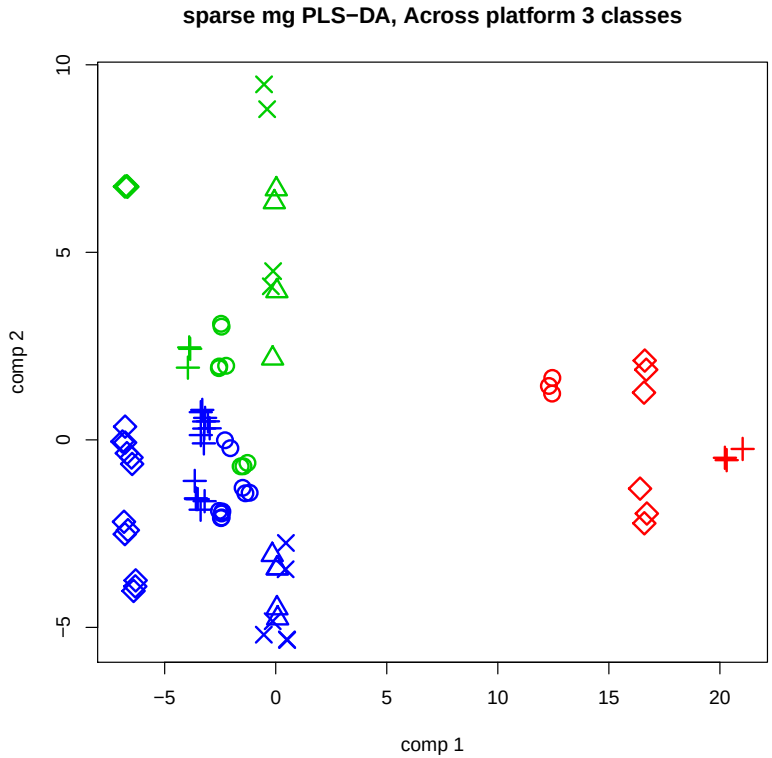






Multi-group sparse PLS-DA (in progress)





sparse multi group PLSDA (RHS) has been implemented to perform variable selection, here a selection of 20 genes.

Performance

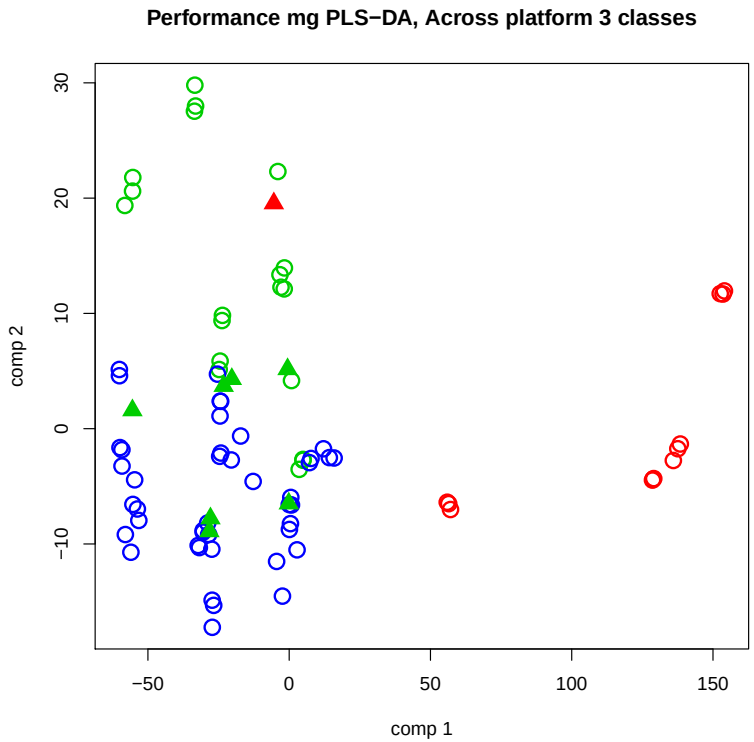


Figure: training set: circles, testing set: triangles with associated colors as predicted

	same platform 3 classes	diff. platform 2 classes	diff. platform 3 classes
# samples	53	70	82
# variables	22K	9K	9K
# groups	5	5	5
mg PLS-DA	8.9 %	15.3 %	16.1 %
sparse mg PLS-DA	7.7 %	4.5 %	10.6 %
var. selected	40	30	20

Table: Classif. error rate (in %) based on 10-fold cross validation

Conclusion to date: cross-platform comparison

- Multi group PLS-DA seems to get rid of the 'batch' effect.
- Multi group PLS-DA seems to be able to address a challenging current problem in molecular biology.
- The sparse version could help identifying a universal gene signature.
- Need more numerical and experimental validations
- Implementation in `mixOmics` is in progress

Kidney transplant study

Oliver Günther UBC

Grandiose project

Andras Nagy Univ. Toronto

mixOmics team

Sébastien Déjean Univ. Tlse

Ignacio González Univ. Tlse

Xin Yi Chua QFAB

Multi group

Stéphanie Bougeard ANSES

Aida Eslami ANSES, ONIRIS

VAC18 Project and multilevel

Benoît Liquet Univ. Bdx2

Multiple data integration

Arthur Tenenhaus Supelec Paris

Time course developments

Jasmin Straube QFAB

Kathy Ruggiero Univ. Auckland

Sébastien Gadat Univ. Toulouse

Christèle Robert INRA Toulouse

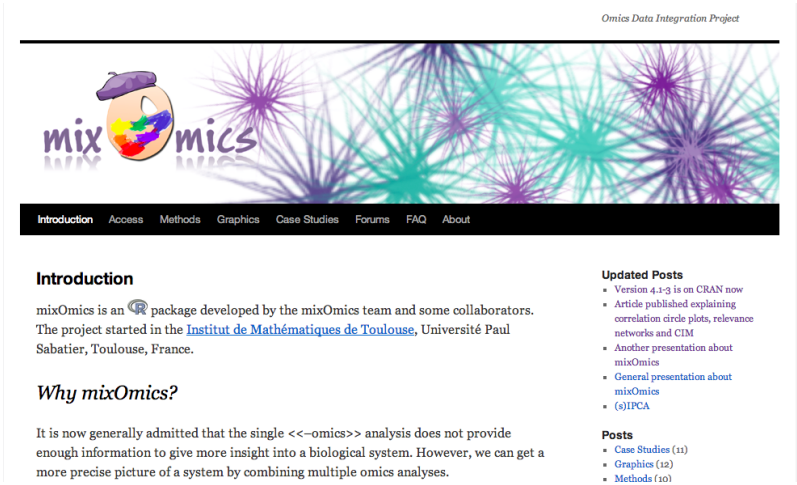
Stenformatics

Christine Wells AIBN, UQ

Florian Rohart AIBN, UQ

Questions?

Workshop `mixOmics` ??



Express your interest

mixomics@math.univ-toulouse.fr

http://www.math.univ-toulouse.fr/~biostat/mixOmics