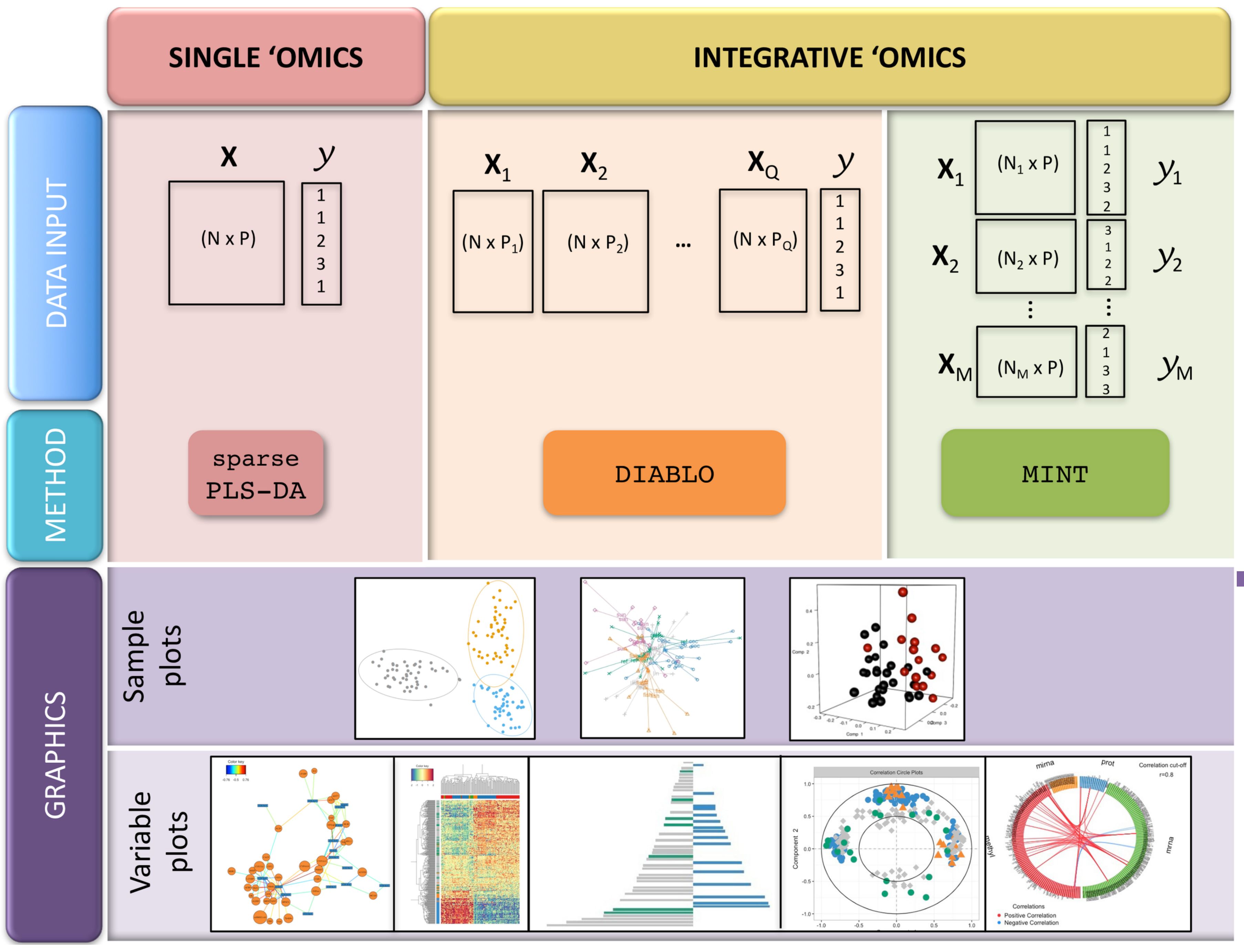
mixOmics: an R package for 'omics feature selection and multiple data integration

Core team: Florian ROHART[2], Sébastien DÉJEAN[3], Kim-Anh LÊ CAO[1,2]
Key contributors: Benoît GAUTIER[2], Amrit SINGH[4], François BARTOLO[3]

[1] Integrative Genomics & School of Mathematics and Statistics, University of Melbourne, **Australia**; [2] The University of Queensland Diamantina Institute, Translational Research Institute, **Australia**
[3] Institut de Mathématiques, UMR5219 CNRS, Université Toulouse 3 Paul Sabatier, Toulouse, **France**; [4] Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, **Canada**.

www.mixomics.org | mixomics@math.univ-toulouse.fr | @mixOmics_team

mixOmics is an R package dedicated to the **multivariate** analysis of biological data sets with a specific focus on **data exploration**, **dimension reduction and visualisation**. By adopting a systems biology approach, the toolkit provides a wide range of methods that statistically **integrate** several data sets at once to probe relationships between heterogeneous **'omics data sets**. Our recent methods[1-3] extend Projection to Latent Structure (PLS) models for discriminant analysis, for data integration across multiple 'omics data or across independent studies, and for **the identification of molecular signatures**.



SINGLE 'OMICS

INTEGRATIVE 'OMICS

DATA INPUT

METHOD

GRAPHICS

sparse PLS-DA

DIABLO

MINT

Sample plots

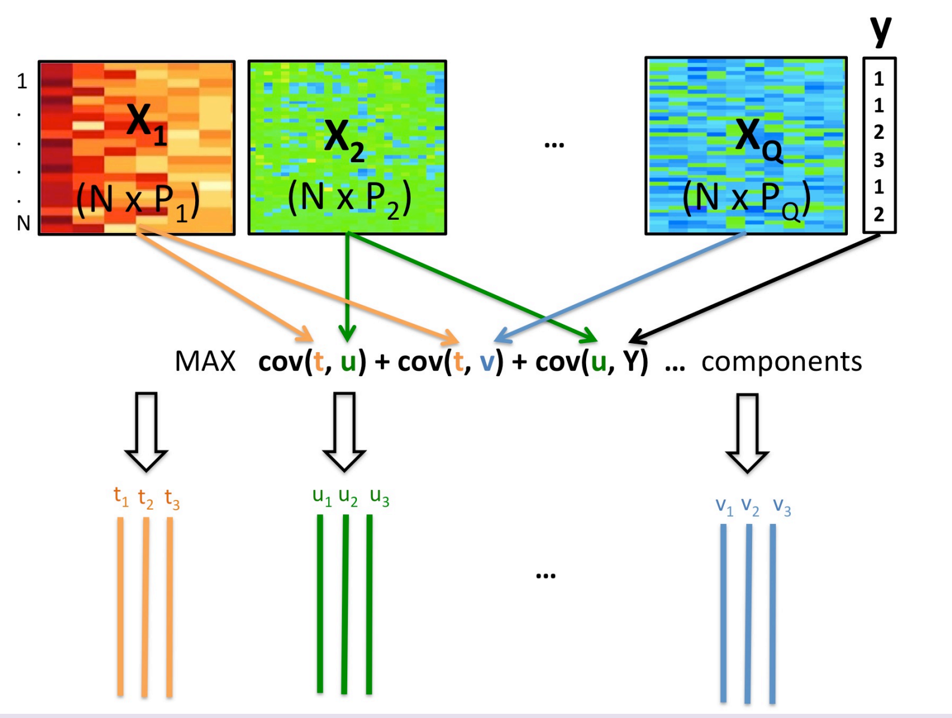
Variable plots

Methods based on Projection to Latent Structures (PLS) models[5]

Nineteen multivariate methods (13 are novel)

Framework		Function name	Sparse	Prediction
Single 'omics	unsupervised	pca	-	-
		ipca	-	-
		sipca	✓	-
		spca	✓	-
	supervised	plsda	-	✓
		splsda	✓	✓
N-integration	unsupervised (2 'omics)	rcca	-	-
		pls	-	✓
		spls	✓	✓
	unsupervised	wrapper.rgcca	-	-
		wrapper.sgcca	✓	✓
		block.pls	-	✓
		block.spls	✓	✓
	supervised (DIABLO)	block.plsda	-	✓
		block.splsda	✓	✓
P-integration (MINT)	unsupervised	mint.pls	-	✓
		mint.spls	✓	✓
	supervised	mint.plsda	-	✓
		mint.splsda	✓	✓

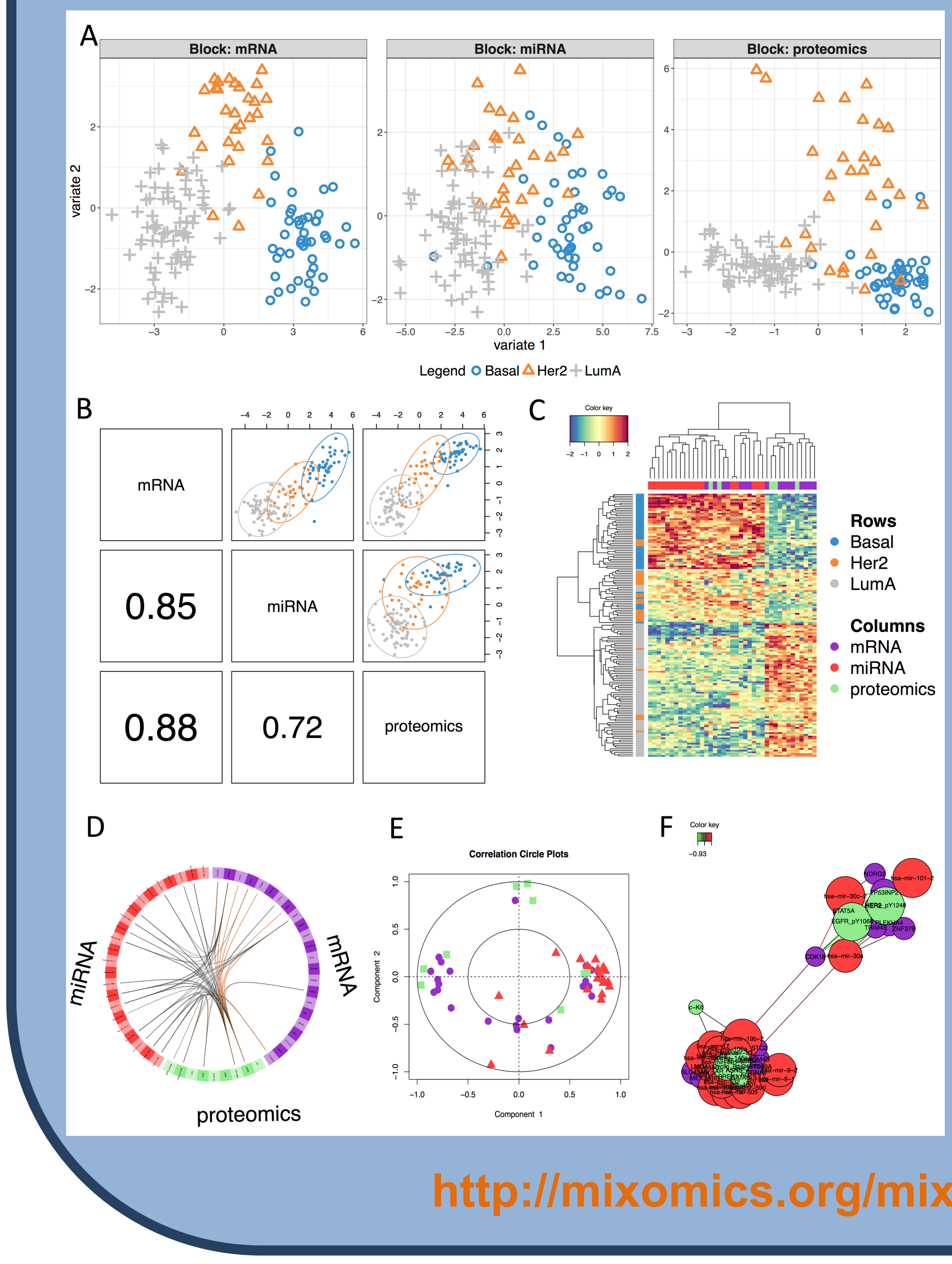
Integrative 'omics



- Matrix decomposition into **latent components**:
  - dimension reduction
  - visualisation (projection of large datasets into the components subspace)
- **Covariance** between components is maximised
- Feature selection via **LASSO (sparse methods)**

N-integration of multiple 'omics measured on the same biological samples[2]

Aims:
- identify a multi 'omics signature that explains a phenotype
- achieve maximal correlation between molecular features of different types for greater biological insights



Challenges:
- large number of highly collinear variables
- vague biological question (*'I want to integrate my data'*)

Toy example on Breast Cancer from TCGA:
- 150 samples
- 3 training datasets: mRNA, miRNA, proteomics
- 3 tumor subtypes
- 2 test datasets: mRNA and miRNA

http://mixomics.org/mixdiablo/

Other frameworks: Single 'omics[4] and *P*-integration[3]

mixMC[4]: multivariate analysis to characterise and compare microbial communities (16S, metagenomics)
Motivation: first multivariate method for beta diversity analyses to identify bacteria driving changes in microbial community.

http://mixomics.org/mixmc

MINT[3]: multivariate P-integration of independent studies on the same variables (genes)
Motivation: Combining several independent transcriptomics datasets increases sample size, avoids data obsolescence, and identifies a platform agnostic signature.

http://mixomics.org/mixmint

Example of computational time cluster with 10 cpus and 50 Gb RAM

Framework	Single 'omics sPLS-DA		*N*-integration DIABLO		*P*-integration MINT	
Data	HNSCC		Asthma (2 omics)		Stem Cell (8 studies)	
N	60		194		210	
P	82, 132		30, 000; 30, 000		13, 313	
function	tune	perf	tune	perf	tune	perf
#fold CV (repeated)	5(10)	5(10)	5(1)	5(10)	LOGOCV	LOGOCV
ncomp	5	3	2	2	2	2
grid length per component	40	-	22^2	-	100	-
#cpu	10	10	10	10	1	1
runtime	15min	6min	19min	3min	17min	12sec

Latest publications

1. Rohart F, Gautier B, Singh A, Lê Cao K-A (2017). **mixOmics: an R package for 'omics feature selection and multiple data integration.** *PLoS Comp Biol*, *in press*
2. Singh A, Gautier B, Shannon C, Vacher M, Rohart F, Tebbutt S, Lê Cao K-A. **DIABLO – an integrative, multi-omics, multivariate method for multi-group classification.** *bioRxiv* 067611
3. Rohart F, Eslami A, Matigian, N, Bougeard S, Lê Cao K-A (2017). **MINT: A multivariate integrative approach to identify a reproducible biomarker signature across multiple experiments and platforms.** *BMC Bioinformatics* **18**:128.
4. Lê Cao K-A&, Costello ME&, Lakis VA, Bartolo F, Chua XY, Brazeilles R and P Rondeau P (2016). **mixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities.** *PLoS ONE*, 11(8):
5. Tenenhaus A, Phillipe C, Guillemot V, Lê Cao K-A, Grill J, Frouin V (2014). **Variable selection for generalized canonical correlation analysis**, *Biostatistics,* 15(3):569-83