Exploration and Integration with mixOmics: our latest improvements

F BARTOLO¹,  S DÉJEAN¹, B GAUTIER², I GONZÁLEZ², F ROHART³ and K-A Lê CAC²

¹ Institut de Mathématiques, UMR5219 CNRS, Université Toulouse 3 Paul Sabatier; 118, route de Narbonne, 31062, Toulouse, Cedex 9, France, ² The University of Queensland Diamantina Institute, Translational Research Institute and ³ Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, QLD 4072, Brisbane, Australia

http://mixomics.org/  mixomics@math.univ-toulouse.fr

mixOmics is an R package dedicated to the exploration and integration of 'omics datasets (transcriptomics, proteomics, metabolomics etc.). Its first release to the CRAN in 2009 proposed statistical methodologies to integrate two 'omics data sets. Since then numerous methodologies and sparse variants have been implemented such as Generalized Canonical Correlation Analysis (GCCA) to integrate more than two datasets. These latest developments require effective computational optimisation and memory management.

### BACKGROUND

#### Statistical methodologies implemented in mixOmics to analyse high throughput data

- **One data set**
  - PCA (Principal Component Analysis) and PLS-DA (Partial Least Squares-Discriminant Analysis)

- **Two data sets**
  - PLS and CCA (Canonical Correlation Analysis)

- **More than two data sets**
  - GCCA (Generalized Canonical Correlation Analysis)

And associated sparse and regularized variants for feature selection and dealing with the high dimension.

#### Several types of graphical outputs

- **Sample plots**
  - a) 2D scatterplots
  - b) 3D scatterplots

- **Variable plots**
  - c) Correlation circle plots
  - d) Networks
  - e) Clustered Image Map

#### Exemplar datasets

- breast.tumors
- liver.toxicity
- multidrug
- nutrimouse
- prostate
- srbc
- vac18

### I. EFFECTIVE COMPUTATIONAL OPTIMISATION

- Sequential optimisation (pre-compilation of functions)
- Parallel computation (with package parallel)

#### Before optimisation

- Time [seconds]

#### After optimisation

- Time [seconds]

- 3 CPU

- 1 CPU

### II. GRAPHICAL IMPROVEMENTS

#### Sample plots

- with ggplot2
- with lattice
- ellipse plot

#### Variable plots

- Contribution plot

- Clustered Image Map

### LATEST PUBLICATIONS


- crGCA: Generalized Canonical Correlation Analysis Methods (2012), R package version 1.0.3.