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

Florian Rohart[1], Benoît Gautier[1], Amrit Singh[2,3], and Kim-Anh Lê Cao*[1]

[1]The University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, QLD 4102, Australia,
[2]UBC James Hogg Research Centre for Heart Lung Innovation, St. Paul's Hospital, University of British Columbia, Vancouver, BC, Canada.
[3]Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, BC, Canada.

Abstract

The advent of high throughput technologies has led to a wealth of publicly available biological data coming from different sources, the so-called 'omics data (transcriptomics for the study of transcripts, proteomics for proteins, metabolomics for metabolites, etc). Combining such large-scale biological data sets can lead to the discovery of important biological insights, provided that relevant information can be extracted in a holistic manner. Current statistical approaches have been focusing on identifying small subsets of molecules (a 'molecular signature') that explains or predicts biological conditions, but mainly for the analysis of a single data set. In addition, commonly used methods are univariate and consider each biological feature independently. In contrast, linear multivariate methods adopt a system biology approach by statistically integrating several data sets at once and offer an unprecedented opportunity to probe relationships between heterogeneous data sets measured at multiple functional levels.

mixOmics is an R package which provides a wide range of linear multivariate methods for data exploration, integration, dimension reduction and visualisation of biological data sets. The methods we have developed extend Projection to Latent Structure (PLS) models for discriminant analysis and data integration and include ℓ_1 penalisations to identify molecular signatures. Here we introduce the mixOmics methods specifically developed to integrate large data sets, either at the N-level, where the same individuals are profiled using different 'omics platforms (same N), or at the P-level, where independent studies including different individuals are generated under similar biological conditions using the same 'omics platform (same P). In both cases, the main challenge to face is data heterogeneity, due to inherent platform-specific artefacts (N-integration), or systematic differences arising from experiments assayed at different geographical sites or different times (P-integration). We present and illustrate those novel multivariate methods on existing 'omics data available from the package.

I. Introduction

The advent of novel 'omics technologies (*e.g.* transcriptomics, proteomics, metabololomics, etc) has enabled new opportunities for biological and medical research discoveries. Commonly, each

*Corresponding author k.lecao@uq.edu.au

feature from each technology (transcripts, proteins, metabolites, etc) is analysed independently through univariate statistical methods such as ANOVA, linear model or t-tests. Such analysis ignores relationships between the different features and may miss crucial biological information. Indeed, biological features act in concert to modulate and influence biological systems and signalling pathways. Multivariate approaches, which model features as a set, can therefore provide a more insightful picture of a biological system, and complement the results obtained from univariate methods. In mixOmics we considered multivariate projection-based methodologies for 'omics data analysis Meng et al. (2016) because of several appealing properties. Firstly, they are computationally efficient to handle large datasets, where the number of biological features (usually in the thousands) is much larger than the number of samples (usually less than 50). Secondly, they perform dimension reduction by projecting the data into a smaller subspace while capturing and highlighting the largest sources of variation from the data, resulting in powerful visualisation of of the system under study. Lastly, they are highly flexible to answer various biological questions (Boulesteix and Strimmer, 2007): mixOmics multivariate methods have been successfully applied in several recent studies to identify biomarkers in 'omics studies ranging from metabolomics, brain imaging to microbiome and statistically integrate data sets generated from difference biological sources (Labus et al., 2015; Cook et al., 2016; Guidi et al., 2016; Mahana et al., 2016; Ramanan et al., 2016; Rollero et al., 2016).

In this paper, we introduce the mixOmics multivariate methods developed for *supervised analysis*, where the aims are to classify or discriminate sample groups, to identify the most discriminant subset of biological features, and to predict the class of new samples. In particular, our two novel frameworks were implemented for the integration of multiple data sets. DIABLO enables the integration of the same biological N samples measured on different 'omics platforms with (N-integration, Singh et al. 2016), MINT enables the integration of several independent data sets or studies measured on the same P predictors (P-integration, Rohart et al. 2016a). One of the main challenges in N- and P-integration is to overcome the technical variance among 'omics platforms - either between different types of 'omics, or within the same type of 'omics but generated from several laboratories, to extract common information. To date, very few statistical methods can perform N- and P-integration in a supervised context. For instance, N-integration is often performed by concatenating all the different 'omics datasets (Liu et al., 2013), thus ignoring the heterogeneity between 'omics platforms, or by combining the molecular signatures identified from separate analyses of each 'omics platform (Günther et al., 2012), thus disregarding the relationships between the different 'omics functional levels. With P-integration, statistical methods are often sequentially combined to accommodate for technical differences among studies or platforms before classifying samples. Such sequential approach is not appropriate for the prediction of new samples as they are prone to overfitting (Rohart et al., 2016a). Our two promising frameworks have the high potential to lead to new discoveries by either modelling relationships between different types of 'omics data (N-integration) or by enabling the integrative analysis of independent 'omics studies and increasing sample size and statistical power (P-integration).

The present article introduces the main functionalities in mixOmics, presents our multivariate frameworks for the identification of molecular signatures in one and several data sets and illustrates each framework in a case study available in the package.

II. The mixOmics R package.

mixOmics is a user-friendly R package dedicated to data exploration, mining, integration and visualisation. It provides a wide range of innovative multivariate methods for the analysis and integration of large data sets in several settings (sparse PLS-DA, DIABLO and MINT, Fig. 1) with

appealing outputs such as (i) insightful visualisations of the data whose dimension has been reduced with the use of latent components, (ii) identification of molecular signatures and (iii) improved usage with common calls to all visualisation and performance assessment methods (see a list of those S3 functions in Suppl. S1).

Multivariate projection-based methods. The multivariate dimension reduction techniques implemented in mixOmics perform unsupervised analyses such as Principal Component Analysis (using NonLinear Iterative Partial Least Squares, Wold 1975), Independent Component Analysis (Yao et al., 2012), Partial Least Squares regression (PLS, also known as Projection to Latent Structures, Wold 1966), regularised Canonical Correlation Analysis (rCCA, González et al. 2008) and Generalised Canonical Correlation Analysis (rGCCA, based on a PLS algorithm Tenenhaus and Tenenhaus 2011), multi-group PLS (Eslami et al., 2013a) as well as supervised analyses such as PLS-Discriminant Analysis (PLS-DA, Nguyen and Rocke 2002b,a; Boulesteix 2004), and recently GCC-DA (Singh et al., 2016) and multi-group PLS-DA (Rohart et al., 2016a).

While each multivariate method aims at answering a specific biological question, the uniqueness of the mixOmics package is to provide novel sparse variants to enable the identification of key predictors (e.g. genes, proteins, metabolites) in large biological data sets. Feature selection is performed via ℓ^1 regularisation (LASSO, Tibshirani 1996), which is implemented directly into the optimisation of the statistical criterion specific to each method. Such criterion include the maximisation of the most important source of variation in the data, of the covariance or correlation between different 'omics sets, or of the segregation of a categorical outcome of interest. Solving the optimisation criterion enables to seek for *latent components* and *loading vectors*. Latent components are linear combinations of the original predictors, where each predictor is assigned a coefficient indicated in the loading vectors. The ourrefore, linear multivariate methods reduce the dimension of the data into a space spanned by a few components, by projecting the samples into a smaller, interpretable space.

In mixOmics methods, the parameters to choose include the total number of components, also called dimensions H, and the ℓ^1 penalty on each dimension for all sparse methods. Contrary to other R packages implementing ℓ^1 penalisation methods (*e.g.* `glmnet`, Friedman et al. 2010, `PMA`, Witten et al. 2013), and it order to improve usability of the methods, the ℓ^1 parameter is solved via soft-thresholding and equivalently replaced by the number of features to select on each dimension. In our multivariate models, the tuning of the number of features to select is performed via repeated cross-validation. The result is a selection of a subset of correlated features that best discriminate the outcome and constitute a *molecular signature*.

Historically, our first methods were dedicated to the integration of two 'omics data sets (González et al., 2008; Lê Cao et al., 2008, 2009b,a), or the discriminant analysis of a single 'omics data set (Lê Cao et al., 2011). The integrative methods presented in this manuscript focus on the integration of multiple biological data sets to address cutting-edge biological and biomedical questions.

Implementation. mixOmics is fully implemented in the R language and exports more than 30 functions for either performing a statistical analysis, tuning its parameters or visualising its results. mixOmics mainly depends on the R base packages (parallel, methods, grDevices, graphics, stats, utils) and recommended packages (MASS, lattice), but also imports functions from a limited number of other R packages (igraph, rgl, ellipse, corpcor, RColorBrewer, plyr, dplyr, tidyr, reshape2, ggplot2). In mixOmics, we provide generic R/S3 functions to assess the performance of the methods (`predict, plot, print, perf, auroc`, etc) and visualisation the results (`plotIndiv, plotArrow, plotVar, plotLoadings`, etc) as described in the next paragraph.

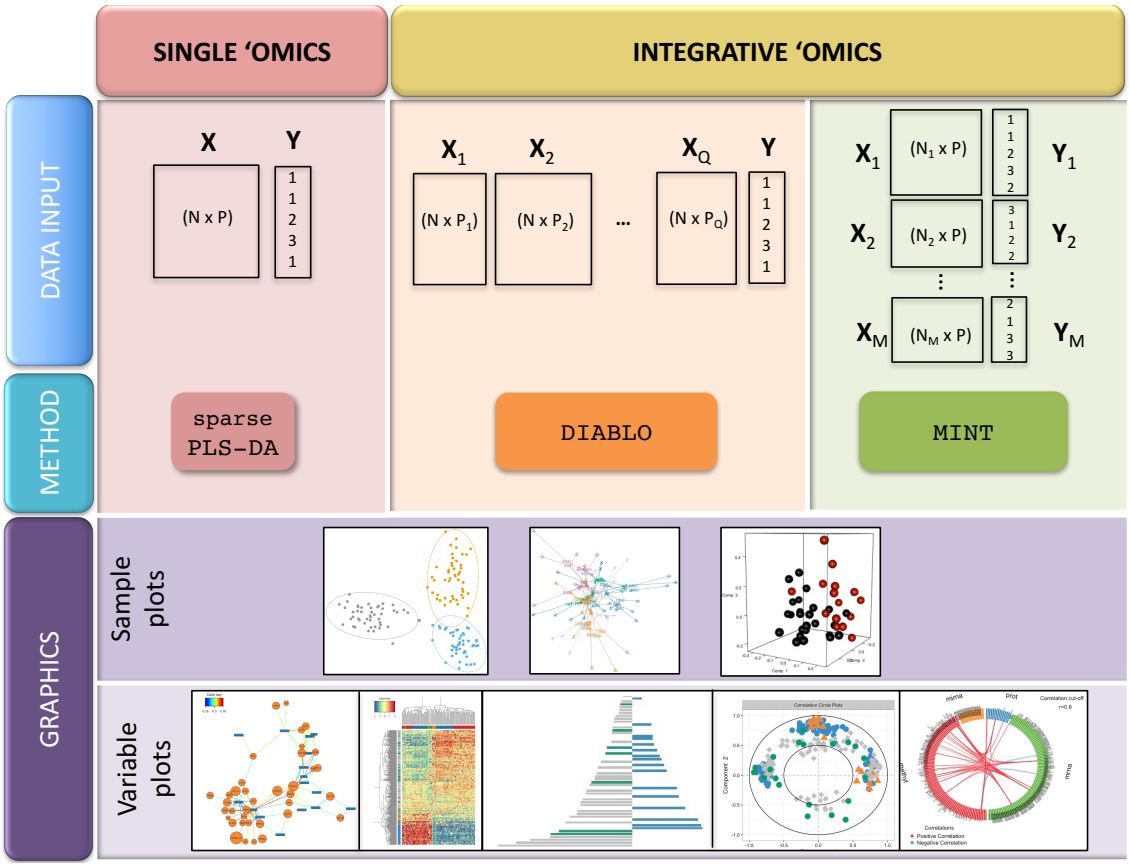


Figure 1: *Overview of the* ***mixOmics*** *multivariate methods for single (sparse PLS-DA) and integrative (DIABLO and MINT) 'omics supervised analyses. X denote a predictor dataset, and Y a categorical outcome response. Integrative analyses include N-integration (across studies generated on the same N samples and different types of predictor features), and P-integration (the same P predictors are measured on independent studies). See also Suppl. S1 for a summary of the different method call and plot functions.*

Currently, seventeen methods are implemented in `mixOmics` to integrate large biological datasets, amongst which twelve have similar names `(mint).(block).(s)pls(da)` (see Table 1) as they are wrappers of a single main hidden function of `mixOmics`. The wrapper functions check and shape the input parameters before passing them to the hidden function that extends the SGCCA algorithm (Tenenhaus et al., 2014) to perform either N- or P-integration. The remaining four statistical methods are PCA, sparse PCA, IPCA, rCCA and rGCCA. Each statistical method implemented in `mixOmics` returns a list of essential outputs which are used in our S3 visualisation functions.

Graphical outputs to visualise multivariate analysis results. `mixOmics` aims to provide insightful and user-friendly graphical outputs to interpret the statistical and biological results, some of which were introduced in González et al. 2012. Thanks to R/S3 functions as listed in S1, the function calls are identical for all multivariate methods implemented in the mixOmics package, as we illustrate in the next sections. We provide various visualisations, including sample plots and feature plots, which are based on the component scores and the loading vectors, respectively.

Framework		sparse	Function name
Single 'omics	unsupervised	-	pca
		-	ipca
		✓	spca
	supervised	-	plsda
		✓	splsda
Two 'omics	unsupervised	-	rcca
		-	pls
		✓	spls
P-integration	unsupervised	-	mint.pls
		✓	mint.spls
	supervised	-	mint.plsda
		✓	mint.splsda
N-integration	unsupervised	-	wrapper.rgcca
		-	block.pls
		✓	block.spls
	supervised	-	block.plsda
		✓	block.splsda (DIABLO)

Table 1: *Seventeen statistical methods available in* ***mixOmics***

Here we list the main important visualisation functions in `mixOmics`.

- `plotIndiv` (Sample plot): represents samples by plotting the component scores. Such plot visualises similarities between samples in the small subspace spanned by the components. For the integrative methods described in Sections V and IV, samples from each data set, or each study are represented on separate plots, allowing to visualise the agreement between the data sets at the sample level. Confidence ellipse plots for each class can be displayed.

- `plotArrow` (Arrow representation): plots the components scores associated to either $\boldsymbol{X}$ data (start of the arrow) or $\boldsymbol{Y}$ outcome (tip of the arrow). As such, short arrows indicate a good discrimination of the classes. In the case of N-integration, the start of the arrow indicates the centroid between all data sets for a given sample and the tips of the arrows the location of that sample in each data set. In that specific case, short arrows indicate a strong agreement between the matching data sets, long arrows a disagreement between the matching data sets.

- `plotVar` (Correlation circle plots): displays features selected by the multivariate method. Each feature coordinate is defined as the Pearson correlation between the original data and the loading vector for each dimension (see González et al. (2012) for a detailed description). Correlation circle plots are particularly useful to visualise the contribution of each feature to define each component (feature close to the large circle of radius 1), as well as the correlation structure between features (clusters of features). The cosine angle between any two points represent the correlation (negative positive, null) between two features.

  Both `plotIndiv` and `plotVar` offer usual plot arguments to display symbols, colours and legend. Graphic styles include default `ggplot2`, `graphics`, `lattice` and 3D plots.

- `cim` (Clustered Image Maps): heatmap plots to visualise the distances between features with respect to each sample. By default we use Euclidian distance and complete linkage method.

For the specific case of N-integration, the function `cimDiablo` represents the selected features from the different data sets.

- `network` (Relevance networks): represents the correlation structure between features of different types. A similarity matrix representing the association between pairs of features across all components is calculated as the sum of the correlations between the original features and the loading vector across all dimensions of interest $h = 1, \ldots, H$ (see González et al. (2012) for more details). Those networks are bipartite and thus only a link between two features of different types is represented.

- `plotLoadings`: represents the loading coefficient of each feature selected on each dimension of the multivariate model. Features are ranked according to their contribution to the component (bottom to top), colors indicate the class for which the mean (or median) expression value is the highest (or the lowest) for each feature. Such graphical output enables more insight into the molecular signature, especially when interpreted in conjunction with the sample plot.

Other graphical outputs are available in `mixOmics` to represent classification performance of multivariate models using the generic function `plot`. The listing of the functions for each framework presented here are summarised in Suppl. S1.

General notations. We assume each data set has been normalised using appropriate techniques specific for the type of 'omics platform. Let X denote a data matrix of size N observations (rows) $\times$ P predictors (e.g. expression levels of P genes, in columns). The categorical outcome y is expressed as a dummy matrix Y in which each column represents one outcome category and each row indicates the class membership of each sample. Y is of size N observations (rows) $\times$ K categories outcome (columns). We denote for all $a \in \mathbb{R}^n$ its ℓ^1 norm $||a||_1 = \sum_1^p |a_j|$ and its ℓ^2 norm $||a||_2 = (\sum_1^p a_j^2)^{1/2}$. For any matrix we denote by $^\top$ its transpose.

III. Multivariate analysis of one data set

Linear Discriminant Analysis (LDA) and Projection to Latent Structure (PLS, Wold 1966) are popular multivariate methods for supervised analyses. In `mixOmics` we mainly focus on PLS methods for their flexibility to solve a variety of analytical problems (Boulesteix and Strimmer, 2007). PLS regression (Wold, 1966) was originally developed for unsupervised analysis to integrate two continuous data sets measured on the same observations. We introduce here a supervised version, called PLS-Discriminant Analysis (PLS-DA, (Nguyen and Rocke, 2002a; Barker and Rayens, 2003), a natural extension that substitutes one of the data set for a dummy matrix Y. PLS-DA fits a classifier multivariate model that assigns samples into known classes, with the ultimate aim to predict the classes of external test samples where the outcome is often unknown.

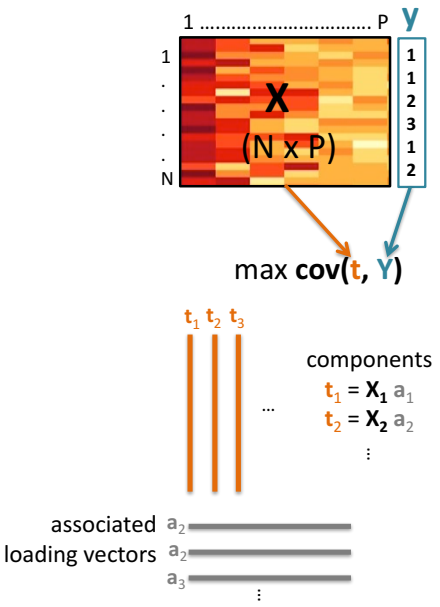


Figure 2: *Example of data matrix decomposition for single 'omics analysis. The predictor matrix X is decomposed into a set of components $(t_1, \ldots, t_H)$ and associated loading vectors $(a_1, \ldots, a_H)$, $\boldsymbol{Y}$ is the outcome coded as a dummy matrix and combined linearly (see exact formula in Equation (1)). $\boldsymbol{X_h}$ is the deflated (residual) matrix starting with $\boldsymbol{X_1} = \boldsymbol{X}$, for $h = 1 \ldots H$ the dimension of the model (number of components).*

PLS-DA. Briefly, PLS-DA is an iterative method that constructs H successive artificial (latent) components $t_h = X_h a_h$ and $u_h = Y_h b_h$ for $h = 1, .., H$, where the h^{th} component t_h (respectively u_h) is a linear combination of the X (Y) features. H denotes the dimension of the PLS-DA model. The weight coefficient vector a_h (b_h) is the loading vector that indicates the *importance* of each feature to define the component. For each dimension $h = 1, \ldots, H$ PLS-DA seeks to maximize

$$\max_{(a_h, b_h)} \text{cov}(X_h a_h, Y_h b_h), \qquad \text{s.t.} \quad ||a_h||_2 = ||b_h||_2 = 1 \tag{1}$$

where X_h, Y_h are the residual (deflated) matrices extracted from each iterative linear regression (see Lê Cao et al. 2011 for more details). The PLS-DA model assigns to each sample i a pair of H scores (t_h^i, u_h^i) which effectively represents the projection of that sample into the X- or Y-space spanned by those PLS components. As $H << P$, the projection space is small, allowing for dimension reduction as well as insightful sample plot representation. Note that the projection into the Y-space is of no use for a Discriminant Analysis as PLS-DA.

Feature selection with sparse PLS-DA. We developed a *sparse* version of PLS-DA (Lê Cao et al., 2011) which includes an ℓ^1 penalisation (Tibshirani, 1996) on the loading vector a_h to shrink some coefficients to zero. Thus, for each dimension $h = 1, .., H$, sPLS-DA solves:

$$\max_{(a_h, b_h)} \text{cov}(X_h a_h, Y_h b_h), \qquad \text{s.t.} \quad ||a_h||_2 = ||b_h||_2 = 1 \text{ and } ||a_h||_1 \leq \lambda_h \tag{2}$$

where λ_h is a non negative parameter that controls the amount of shrinkage in a_h. The component scores $t_h = X_h a_h$ are now defined on a small subset of features with non-zero coefficients, leading

to feature selection that aims to optimally maximise the discrimination between the K outcome classes in Y.

Prediction. Once fitted, the (sparse) PLS-DA model can be applied on an external test set $\tilde{X}$ of size $(N_{test} \times P)$ to predict the class of new samples (see Lê Cao et al. 2011 for more details). The `predict` function outputs the predicted scores for each test sample. Since the predicted scores are expressed as continuous values, a prediction distance must be applied to obtain the final predicted class membership. Distances such as 'maximum distance', 'Malhanobis distance' and 'Centroid distance' are provided (see Figure 3B1).

Choice of parameters. One important parameter to choose in PLS-DA and sPLS-DA is the number of components, or dimension of the model, called `ncomp` in `mixOmics`. While this parameter has mostly been ignored from the PLS-DA literature it plays a crucial role to ensure maximal prediction accuracy. Our experience when analysing a large number of 'omics data sets has shown that `ncomp = K-1` was usually sufficient to summarise most of the discriminatory information from the data (Lê Cao et al., 2011; Shah et al., 2016). The second parameter to choose pertains to the λ_h penalisation parameters for sPLS-DA, which was replaced by the number of features to select on each component with the argument `keepX`.

The `tune` function performs repeated cross-validation (CV) for a user-input grid of `keepX` values to assess. The `keepX` parameter that leads to the best prediction accuracy of the model is reported for each component. Prediction accuracy is evaluated according to the overall classification error rate, or the Balanced Error Rate (BER) for unbalanced number of samples per class. Both measures are calculated on the left-out samples set during the CV procedure, and averaged across the repeated CV runs. The number of folds in CV depends on the number of samples N and can be specified in the function, with a sufficient number of run (e.g. `nrepeat = 50-100`). In the case of small N, leave-one-out validation is advised and `nrepeat` is set to 1. Additional outputs from the `tune` function include 1/the stability of the selected features across all CV runs, which represents a useful measure of reproducibility of the molecular signature and 2/receiver operating characteristic (ROC) curves and Area Under the Curve (AUC) averaged using one-vs-all comparison if $K > 2$. Note however that ROC and AUC criteria may not be particularly insightful as the prediction threshold in our methods is based on a specified distance as described earlier.

An additional option that we developed and implemented for all `tune` function in `mixOmics` is to tune and fit a *constraint* model. The process is as follows: once the optimal `keepX` value is chosen on one component, the model is fitted with the specific `keepX`, and the resulting feature selection is then fixed for the tuning of the following component. In other words, the tuning is performed on the optimal *list* of selected features (`keepX.constraint`) instead of the number of features (`keepX`). Such strategy was implemented in the sister package `bootPLS` and successfully applied in our recent integrative study Rohart et al. (2016b). Our experience has shown that the contraint tuning and models improve the performance of the methods. We illustrate an example in Suppl. V.

The tuning step must be conducted with caution to avoid overfitting results, as widely described in the literature (see for example Ambroise and McLachlan 2002). Our `tune` procedure performs repeated CV, reports the frequency of selected features across all repeated CV folds and the classification error rate for each `keepX` value. Once `ncomp` and `keepX` for each component are chosen, the final PLS-DA or sPLS-DA model is fitted on the whole data set and the final performance can be obtained with the `perf` function that also performs repeated CV (see Suppl. V).

Extensions of PLS-DA for repeated measurements and 16S microbiome data. PLS-DA and sPLS-DA were extended to account for repeated measurement designs, as described in Liquet et al. (2012) by specifying the argument `multilevel` in the `plsda` and `splsda` functions. Recent extensions in the package include sPLS-DA analysis to identify microbial communities for 16S data with an additional `logratio` argument to account for compositional data in microbiome experiment (Lê Cao et al. 2016, see also our `mixMC` framework in www.mixOmics.org/mixMC).

Usage in `mixOmics`. Figure 3 illustrates the different graphical outputs obtained when analysing a single data set from unsupervised to supervised analyses. The data set analysed is a microarray data set available from the `mixOmics` package investigating Small Round Blue Cell Tumors (SRBCT, Khan et al. 2001) of 63 tumour samples with the expression levels of 2,308 genes. Samples are classified into four classes: 8 Burkitt Lymphoma (BL), 23 Ewing Sarcoma (EWS), 12 neuroblastoma (NB), and 20 rhabdomyosarcoma (RMS). The aim of this analysis is to assess similarities between tumour types, using Principal Component Analysis (Fig. 3**A**), to classify the different tumour subtypes with PLS-DA (Fig. 3**B**) and to identify a molecular gene signature discriminating the tumour types with sPLS-DA (Fig. 3**C**). The full pipeline, results interpretation and associated R code is available in Electronic Suppl. V.

IV. Integration of heterogeneous data sets with DIABLO

The integration of multiple 'omics datasets measured on the same N biological samples (Figure 1) is based on a variant of the multivariate methodology Generalised Canonical Correlation Analysis (GCCA, Tenenhaus and Tenenhaus 2011; Tenenhaus et al. 2014), which, contrary to what its name suggests, generalises PLS for N-integration. Our recent development `DIABLO` further improved the implementation of GCCA to include feature selection in a supervised framework and in a user-friendly manner (Günther et al., 2014; Singh et al., 2016).

Method. We denote Q 'omics data sets $X^{(1)}(N \times P_1)$, $X^{(2)}(N \times P_2)$, ..., $X^{(Q)}(N \times P_Q)$ measuring the expression levels of P_q 'omics features on the same N biological samples, $q = 1, \ldots, Q$. GCCA solves for each component $h = 1, \ldots, H$:

$$\max_{a_h^{(1)},\ldots,a_h^{(Q)}} \sum_{q,j=1,q\neq j}^{Q} c_{q,j} \operatorname{cov}(X_h^{(q)} a_h^{(q)}, X_h^{(j)} a_h^{(j)}), \qquad \text{s.t.} \quad ||a_h^{(q)}||_2 = 1 \text{ and } ||a_h^{(q)}||_1 \leq \lambda^{(q)} \tag{3}$$

where $\lambda^{(q)}$ is the penalisation parameter, $a_h^{(q)}$ is the loading vector on component h associated to the residual matrix $X_h^{(q)}$ of the data set $X^{(q)}$, and $C = \{c_{q,j}\}_{q,j}$ is the design matrix. C is a $Q \times Q$ matrix of zeros and ones which specifies whether datasets should be correlated; zeros when datasets are not connected and ones where datasets are connected. Thus, it is possible to constraint the model to only take into account specific pairwise covariances by setting the `design` matrix (see Tenenhaus et al. (2014) for more details). Such design thus enables to model a particular association between pairs of 'omics data, as expected from prior biological knowledge or experimental design. `DIABLO` Discriminant Analysis in `mixOmics` extends (3) to a supervised framework by replacing one data matrix $X^{(q)}$ with the outcome dummy matrix Y.

Prediction. `DIABLO` includes several predictions strategies such as a majority vote and a weighted vote. Both are based on the predictions obtained from each 'omics dataset via the $X_h^{(q)} a_h^{(q)}$ components. The majority vote consists in assigning to a sample the class that has received the

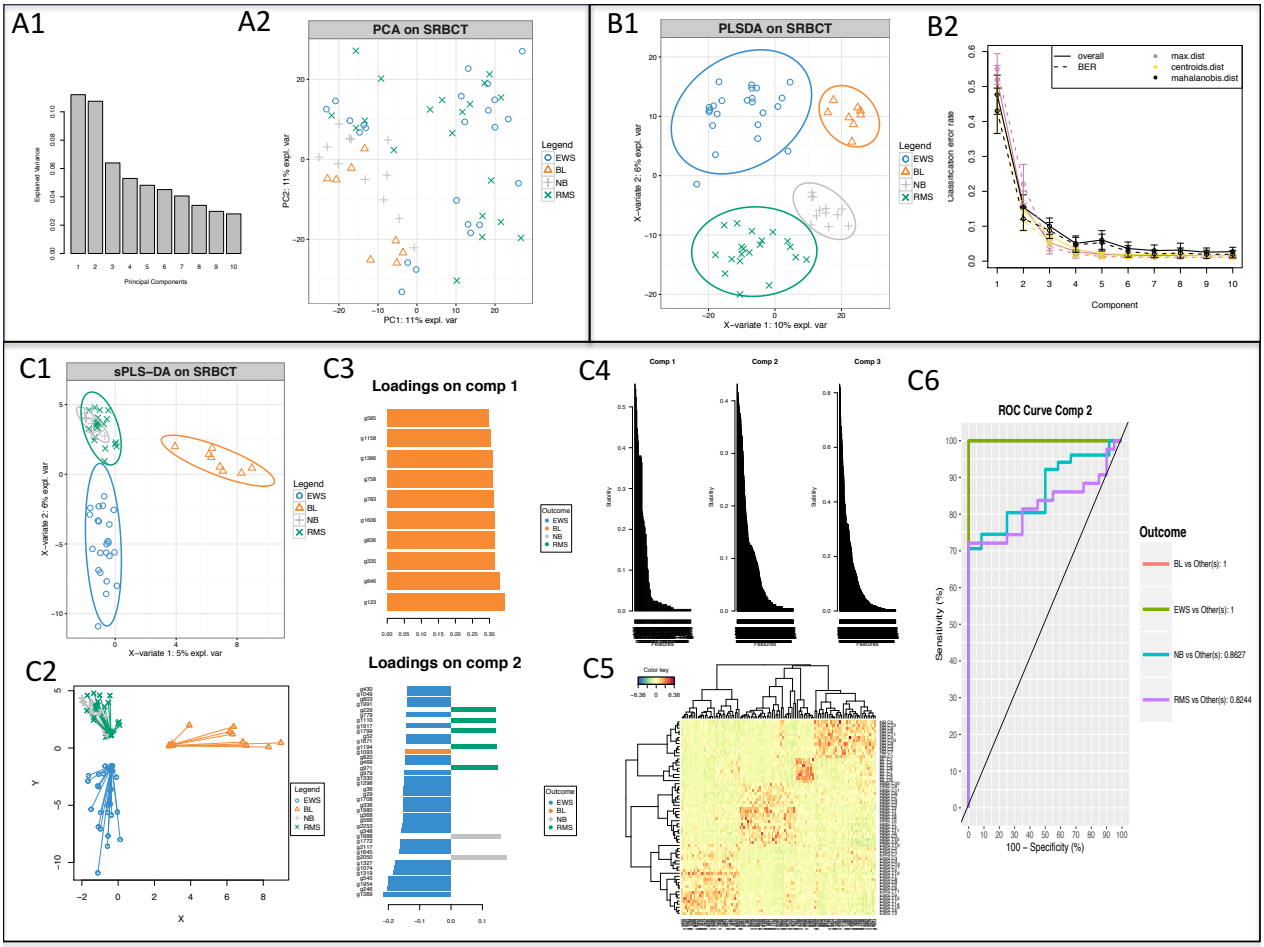


Figure 3: *Illustration of PLS-DA and sPLS-DA in mixOmics.* ***A) Unsupervised preliminary analysis with PCA****,* ***A1****: percentage of explained variance per component,* ***A2****: PCA sample plot.* ***B) Supervised analysis with PLS-DA****,* ***B1****: PLS-DA sample plot with confidence ellipse plots,* ***B2****: classification performance per component (overall or BER) for three prediction distances using 50 * 5-fold cross-validation.* ***C) Supervised analysis and feature selection with sparse PLS-DA****,* ***C1****: sPLS-DA sample plot with confidence ellipse plots,* ***C2****: arrow plot representing each sample pointing towards its outcome category,* ***C3****: coefficient weight of the features selected on component 1 and component 2, with colour indicating the class with maximal mean expression value for each feature,* ***C4****: feature stability when evaluating the performance of a sPLS-DA model with 10, 40 and 60 features on the first three components (50 * 5-fold cross-validation),* ***C5****: Clustered Image Map (Euclidian Distance, Complete linkage). Samples are represented in rows, selected features in columns (10, 40 and 60 genes selected on each component respectively),* ***C6****: receiver operating characteristic (ROC) curve and Area Under the Curve (AUC) averaged using one-vs-all comparisons.*

highest number of predictions over all 'omics data set, while the weighted vote combines the predictions of all 'omics after weighting each by the correlation between the component $X_h^{(q)} a_h^{(q)}$ and the outcome. In both strategies, a prediction distance is to be specified to obtain a predicted class, as described in Section III. Ties are indicated as 'NA' in the predicted classes.

Specific outputs to visualise multiple 'omics data sets integration. Several types of graphical outputs are available to support interpretation of the statistical results. To represent samples, `plotIndiv` displays component scores from each 'omics data set individually. Such type of plot enable to visualise the agreement between all data sets at the sample level. The `plotArrow` function also enables similar visualisation (see Section II). The function `plotDiablo` is a matrix scatterplot of the components from each data set for a given dimension; it enables to check whether the pairwise correlation between two 'omics has been modelled according to the design. The function `circosPlot` shows pairwise correlations among the selected features over all data sets. Features are represented on the side of the circos plot, with colours indicating the type of data, and external (optional) lines display expression levels for each outcome category. It is an extension of the method used in `plotVar`, `cim` and `network` (see González et al. 2012).

Parameters tuning. The first parameter to choose in `DIABLO` is the `design` matrix, which can be specified based on either prior biological knowledge, or by using a preliminary multivariate method integrating two data sets at a time (e.g. PLS) to assess the potential common information between data sets in an unsupervised analysis. In addition, the function `plotDIABLO` run on all features (non sparse model) can further confirm the suitability of the design to maximise the correlations between data sets. By default the design links each data set to the outcome Y.
Similar to PLS-DA, the number of components `ncomp` needs to be specified. We usually found that $K - 1$ components were sufficient to discriminate the sample classes but this should be further assessed with the model performance and graphical outputs (see our example 4 and Suppl. V).
Finally and most importantly, the number of features to select *per* data set and *per* component needs to be specified with the list argument `keepX`. The `tune` function evaluates the performance of the model over a grid of different `keepX` parameters using repeated cross-validation, based on the (balanced) classification error rate, with a parallelisation option (argument `cpus`). Note that this tuning step might become cumbersome as there might be numerous combinations to evaluate. A constraint tuning is also available, see Section III. Our experience shows that a minimal error rate could be attained with a rather small number of features per component and data set (<20, Singh et al. 2016). However, the user can enlarge the search grid to ensure a sufficiently large number of selected features when the focus is on the downstream biological interpretation (e.g. enrichment analyses).

Usage in `mixOmics`. Figure 4 displays some of the graphical outputs when performing N-integration. The multi-'omics breast cancer study analysed include mRNA ($P_1 = 200$), miRNA ($P_2 = 184$) and proteomics ($P_3 = 142$) data that were normalised and drastically filtered for illustrative purpose in this manuscript. The data were divided into a training set composed of $N = 150$ samples and an external test set of $N_{\text{test}} = 70$ samples where the proteomics data are missing (see details in Singh et al. 2016). The aim of N-integration is to identify a highly correlated multi-'omics signature discriminating the breast cancer subgroups Luminal A, Her2 and Basal. Figure 4**A** displays the matrix design and the sample correlation between each component from each data set, **B** the sample plots for each data set, **C** our different feature plots and **D** a clustered image map of the multi-'omics signature. The full pipeline, results interpretation and associated R code is available in Electronic Suppl. V.

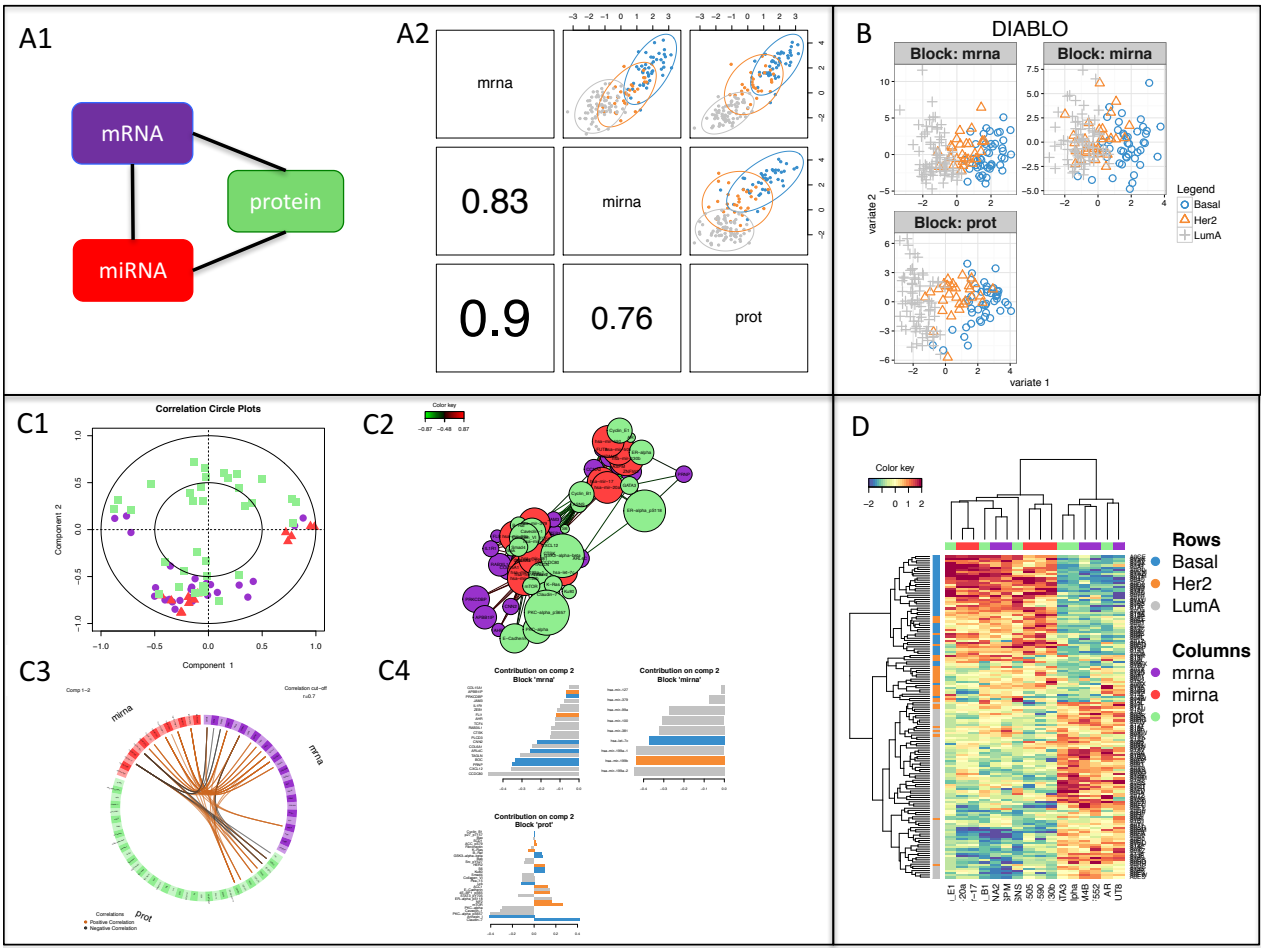


Figure 4: *Illustration of DIABLO analysis in mixOmics.* ***A1****, design and* ***A2****, sample scatterplot from* `plotDiablo` *displaying the first component in each data set (upper plot) and Pearson correlation between each component (lower plot).* ***B****: sample plot per data set (block),* ***C) feature outputs****,* ***C1****: Correlation Circle plot representing each type of selected features,* ***C2****: relevance network visualisation of the selected features,* ***C3****: Circos plot shows the positive (negative) correlation ($r > 0.7$) between selected features as indicated by the brown (black) links, feature names appear in the quadrants,* ***C4****: coefficient weight of the features selected on component 1 in each data set, with color indicating the class with a maximal mean expression value for each feature.* ***D*** *Clustered Image Map (Euclidian distance, Complete linkage) of the multi-omics signature. Samples are represented in rows, selected features on the first component in columns.*

V. P-integration across independent data sets with MINT

The integration of independent data sets measured on the same common P features under similar conditions or treatments (Figure 1) is a useful approach to increase sample size and gain statistical power. In this context, the challenge is to accommodate for systematic differences that arise due to differences between protocols, geographical sites or the use of different technological platforms to generate the same type of 'omics data (e.g. transcriptomics). The systematic unwanted variation, also called 'batch-effect', often acts as a strong confounder in the statistical analysis and may lead to spurious results and conclusions if it is not accounted for in the statistical model.

Method. MINT (Rohart et al., 2016a) is an extension of the multi-group PLS framework (mg-PLS, Eslami et al. 2013b, 2014), where 'groups' represent independent studies, to a supervised framework with feature selection. MINT seeks to identify a common projection space for all studies that is defined on a small subset of discriminative features and that display an analogous discrimination of the samples across studies.

We combine M datasets denoted $X^{(1)}(N_1 \times P)$, $X^{(2)}(N_2 \times P)$, ..., $X^{(M)}(N_M \times P)$ measured on the same P predictors but from independent studies, with $N = \sum_{m=1}^{M} N_m$. Each data set $X^{(m)}$, $m = 1, \ldots, M$, has an associated dummy outcome $Y^{(m)}$ in which all K classes are represented. We denote X $(N \times P)$ and Y $(N \times K)$ the concatenation of all $X^{(m)}$ and $Y^{(m)}$ respectively. In the MINT particular framework, each feature of the datasets $X^{(m)}$ and $Y^{(m)}$ is centered and scaled. For each component h, MINT solves :

$$\max_{a_h, b_h} \sum_{m=1}^{M} N_m \, \text{cov}(X_h^{(m)} a_h, Y_h^{(m)} b_h), \qquad \text{s.t.} \quad ||a_h||_2 = 1 \text{ and } ||a_h||_1 \leq \lambda \tag{4}$$

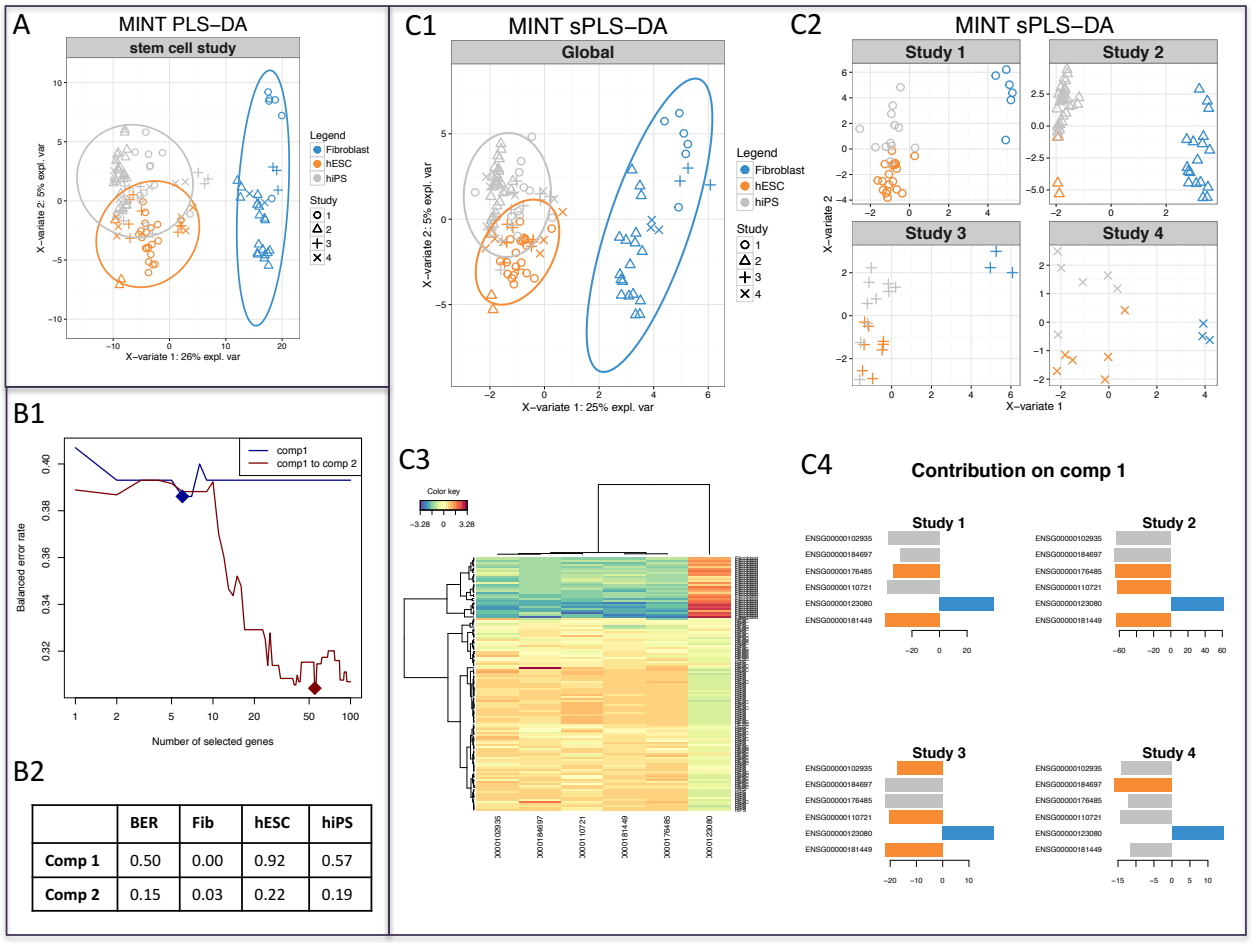
where a_h and b_h are the global loadings vectors common to all studies, $t_h^{(m)} = X_h^{(m)} a_h$ and $u_h^{(m)} = Y_h^{(m)} b_h$ are the partial PLS-components that are study specific. Residual (deflated) matrices are calculated for each iteration of the algorithm based on the global components and loading vectors (see Rohart et al. 2016a). Thus the MINT algorithm models the study structure during the integration process. The penalisation parameter λ controls the amount of shrinkage and thus the number of non zero weights in the global loading vector a. Similarly to sPLS-DA (Section III) MINT selects a combination of features on each PLS-component.

Specific graphical outputs. The set of partial components $t_h^{(m)}, h = 1, ..., H$ provides study-specific outputs in `plotIndiv`. These graphics can act as a quality control step to detect studies that cluster outcome classes differently to other studies (i.e. 'outlier' studies). The function `plotLoadings` displays the coefficients weights of the features globally selected by the model but represented individually in each study. Visualisation of the global loading vectors is also available. Note the projection into the Y-space is of not useful in MINT.

Parameters tuning. We take advantage of the independence between studies to evaluate the performance based on a novel CV technique called 'Leave-One-Group-Out Cross-Validation' (Rohart et al., 2016a). LOGOCV performs CV where each group m is left out once. The aim is to reflect a realistic prediction of independent external studies. The `tune` function implements LOGOCV to choose the optimal number of features `keepX` or the optimal set of features `keepX.constraint` to select in X, as described in the earlier Sections. Note that LOGOCV cannot be repeated (no `nrepeat` argument) as this type of cross-validation is not random.

Usage in mixOmics. Figure 5 displays some of the graphical outputs when performing P-integration with `mixOmics`. We combined four independent transcriptomics stem cell studies that measure the expression levels of 400 genes across 125 samples (cells). The data were normalised and drastically filtered for illustrative purpose in this manuscript. The cells were classified into Fibroblasts, hESC and hiPSC. The aim of this $P-$ integration analysis is to identify a robust molecular signature across all studies to discriminate the three different cell types. After applying MINT via the `mint.plsda` and `mint.splsda` functions, generic visualisations functions of the `mixOmics` R-package like `plotIndiv`, `cim` and `plotLoadings` can be used. Figure 5 displays some

outputs easily obtained by calls to those functions. Figure 5**A** displays a MINT PLS-DA sample plot, Figure 5**B** the tuning and performance evaluation of the MINT sPLS-DA analysis and Figure 5**C** the different sample and feature with MINT sPLS-DA. The full pipeline, results interpretation and associated R code is available in Electronic Suppl. V.



	BER	Fib	hESC	hiPS
Comp 1	0.50	0.00	0.92	0.57
Comp 2	0.15	0.03	0.22	0.19

Figure 5: *Illustration of MINT analysis in mixOmics.* ***A: Preliminary analysis with MINT PLS-DA (no feature selection)****, sample plot displays the sample cell types.* ***B Parameter tuning and performance with MINT sPLS-DA****,* ***B1****: BER (y-axis) with respect to number of selected features (x-axis) when 1 and 2 components are successively added in the model. Full diamond represents the optimal number of features to select on each component using Leave-One-Group-Out cross-validation and the maximum distance,* ***B2****: Final performance of the MINT sPLS-DA model for a selection of 6 and 55 transcripts on each component: overall BER and error rate per cell type with the maximum distance,* ***C) MINT sPLS-DA graphical outputs using*** `plotIndiv`***,*** `cim` ***and*** `plotLoadings`*.* ***C1****: Global sample plot with confidence ellipse plots.* ***C2****: Study specific sample plot.* ***C3****: Clustered Image Map (Euclidian Distance, Complete linkage). Samples are represented in rows, selected features on the first component in columns.* ***C4****: Coefficient weight of the features selected on component 1 in each study, with color indicating the class with a maximal mean expression value for each transcript.*

Conclusions

The technological race in high-throughput biology lead to increasingly complex biological problems which require innovative statistical and analytical tools. Our package mixOmics focuses on data exploration and data mining, that are crucial steps for a first understanding of large data sets. In this article we presented our latest methods to answer cutting-edge integrative and multivariate questions in biology. In particular, our supervised frameworks DIABLO and MINT substantially extend the key PLS-DA method to perform *N*- and *P*- integration of multiple data sets, classification and class prediction of external studies. Combined together, those two framework bear the promise of NP-integration (combine multiple studies that each have several type of data). The sparse version of our methods are particularly insightful to identify molecular signatures across those multiple data sets.

Feature selection resulting from our methods help to refine biological hypotheses, suggest downstream analyses including statistical inference analyses, and may propose biological experimental validations. Indeed, multivariate methods include appealing properties to mine and analyse large and complex biological data, as they allow more relaxed assumptions about data distribution, data size ($n << p$) and data range than univariate methods, and provide insightful visualisations. In addition, the identification of a *combination* of discriminative features meet biological assumptions that cannot be addressed with univariate methods. Nonetheless, we believe that combining different types of statistical methods (univariate, multivariate, machine learning) is the key to answer complex biological questions. However, such questions must be well stated, in order for those exploratory integrative methods to provide meaningful results, and especially for the non trivial case of multiple data integration.

We illustrated our different frameworks on classical 'omics data, however, mixOmics methods can also be applied to data beyond the realm of 'omics as long as they are expressed as continuous values. Our future work will include extensive development for other types of data, such as genotypic as well as time course biological data. Finally, while our manuscript focused mainly on supervised methodologies, the package also include their unsupervised counterparts to investigate relationships and associations between features with no prior phenotypic or response information.

Availability and requirements

The R package mixOmics is available from the CRAN (R Core Team, 2016), with tutorials and newsletter updates available from our website www.mixOmics.org.

Conflict of Interest

The authors declare that they have no competing interests.

Availability of supporting data

The data sets supporting the results of this article are available from the mixOmics R package in a processed format. R scripts, full tutorials and reports to reproduce the results from the proposed framework are available as Sweave code from our website www.mixOmics.org.

Author's contributions

FR implemented the MINT method, FR, BG and AS implemented the DIABLO method, FR was the main developer of the *mixOmics* package from v6.0.0. KALC manages and supervises the mixOmics project. FR and KALC edited the manuscript.

Acknowledgements

FR was supported, in part, by the Australian Cancer Research Foundation (ACRF) for the Diamantina Individualised Oncology Care Centre at The University of Queensland Diamantina Institute. KALC was supported, in part, by and the National Health and Medical Research Council (NHMRC) Career Development fellowship (APP1087415). The authors would like to thank the numerous mixOmics users to continuously helping us improving the package.

References

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, 99(10):6562–6566.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of chemometrics*, 17(3):166–173.

Boulesteix, A.-L. (2004). Pls dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology*, 3(1):1–30.

Boulesteix, A.-L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.*, 8(1):32–44.

Cook, J. A., Chandramouli, G. V., Anver, M. R., Sowers, A. L., Thetford, A., Krausz, K. W., Gonzalez, F. J., Mitchell, J. B., and Patterson, A. D. (2016). Mass spectrometry–based metabolomics identifies longitudinal urinary metabolite profiles predictive of radiation-induced cancer. *Cancer research*, 76(6):1569–1577.

Eslami, A., Qannari, E. M., Kohler, A., and Bougeard, S. (2013a). Multi-group pls regression: Application to epidemiology. In *New Perspectives in Partial Least Squares and Related Methods*, pages 243–255. Springer.

Eslami, A., Qannari, E. M., Kohler, A., and Bougeard, S. (2013b). Multi-group PLS Regression: Application to Epidemiology. In *New Perspectives in Partial Least Squares and Related Methods*, pages 243–255. Springer.

Eslami, A., Qannari, E. M., Kohler, A., and Bougeard, S. (2014). Algorithms for multi-group PLS. *J. Chemometrics*, 28(3):192–201.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

González, I., Déjean, S., Martin, P. G., Baccini, A., et al. (2008). CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12):1–14.

González, I., Lê Cao, K.-A., Davis, M. J., Déjean, S., et al. (2012). Visualising associations between paired 'omics' data sets. *BioData mining*, 5(1):19.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J. R., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*.

Günther, O. P., Chen, V., Freue, G. C., Balshaw, R. F., Tebbutt, S. J., Hollander, Z., Takhar, M., McMaster, W. R., McManus, B. M., Keown, P. A., et al. (2012). A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers. *BMC bioinformatics*, 13(1):326.

Günther, O. P., Shin, H., Ng, R. T., McMaster, W. R., McManus, B. M., Keown, P. A., Tebbutt, S. J., and Lê Cao, K.-A. (2014). Novel multivariate methods for integration of genomics and proteomics data: applications in a kidney transplant rejection study. *Omics: a journal of integrative biology*, 18(11):682–695.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–679.

Labus, J. S., Van Horn, J. D., Gupta, A., Alaverdyan, M., Torgerson, C., Ashe-McNalley, C., Irimia, A., Hong, J.-Y., Naliboff, B., Tillisch, K., et al. (2015). Multivariate morphological brain signatures predict patients with chronic abdominal pain from healthy control subjects. *Pain*, 156(8):1545–1554.

Lê Cao, K., Rossouw, D., Robert-Granié, C., Besse, P., et al. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7:Article–35.

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(1):253.

Lê Cao, K.-A., González, I., and Déjean, S. (2009a). integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21):2855–2856.

Lê Cao, K.-A., Lakis, V. A., Bartolo, F., Costello, M.-E., Chua, X.-Y., Brazeilles, R., and Rondeau, P. (2016). Mixmc: Multivariate insights into microbial communities. *PloS one*, 11(8):e0160169.

Lê Cao, K.-A., Martin, P. G., Robert-Granié, C., and Besse, P. (2009b). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics*, 10(1):34.

Liquet, B., Lê Cao, K.-A., Hocini, H., and Thiébaut, R. (2012). A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC bioinformatics*, 13:325.

Liu, Y., Devescovi, V., Chen, S., and Nardini, C. (2013). Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC systems biology*, 7(1):14.

Mahana, D., Trent, C. M., Kurtz, Z. D., Bokulich, N. A., Battaglia, T., Chung, J., Müller, C. L., Li, H., Bonneau, R. A., and Blaser, M. J. (2016). Antibiotic perturbation of the murine gut microbiome enhances the adiposity, insulin resistance, and liver disease associated with high-fat diet. *Genome medicine*, 8(1):1.

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*, page bbv108.

Nguyen, D. V. and Rocke, D. M. (2002a). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18(9):1216–1226.

Nguyen, D. V. and Rocke, D. M. (2002b). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramanan, D., Bowcutt, R., Lee, S. C., San Tang, M., Kurtz, Z. D., Ding, Y., Honda, K., Gause, W. C., Blaser, M. J., Bonneau, R. A., et al. (2016). Helminth infection promotes colonization resistance via type 2 immunity. *Science*, 352(6285):608–612.

Rohart, F., Eslami, A., Matigian, N., Bougeard, S., and Le Cao, K.-A. (2016a). Mint: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *bioRxiv*, page 070813.

Rohart, F., Mason, E. A., Matigian, N., Mosbergen, R., Korn, O., Chen, T., Butcher, S., Patel, J., Atkinson, K., Khosrotehrani, K., Fisk, N. M., Lê Cao, K., and Wells, C. A. (2016b). A molecular classification of human mesenchymal stromal cells. *PeerJ*, 4:e1845.

Rollero, S., Mouret, J.-R., Sanchez, I., Camarasa, C., Ortiz-Julien, A., Sablayrolles, J.-M., and Dequin, S. (2016). Key role of lipid management in nitrogen and aroma metabolism in an evolved wine yeast strain. *Microbial cell factories*, 15(1):1.

Shah, A. K., Lê Cao, K.-A., Choi, E., Chen, D., Gautier, B., Nancarrow, D., Whiteman, D. C., Baker, P. R., Clauser, K. R., Chalkley, R. J., et al. (2016). Glyco-centric lectin magnetic bead array (lemba)- proteomics dataset of human serum samples from healthy, barrettŒş s esophagus and esophageal adenocarcinoma individuals. *Data in Brief*, 7:1058–1062.

Singh, A., Gautier, B., Shannon, C. P., Vacher, M., Rohart, F., Tebutt, S. J., and Le Cao, K.-A. (2016). Diablo-an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv*, page 067611.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, page kxu001.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Witten, D., Tibshirani, R., Gross, S., and Narasimhan, B. (2013). *PMA: Penalized Multivariate Analysis*. R package version 1.0.9.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *J. Multivar. Anal.*, pages 391–420.

Wold, H. (1975). *Path models with latent variables: The NIPALS approach*. Acad. Press.

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*, 13(1):24.

Supplementary Material

	functions	PLS-DA	sPLS-DA	DIABLO	sparse DIABLO	MINT	sparse MINT
function call		plsda	splsda	block.plsda	block.splsda	mint.plsda	mint.splsda
parameters		ncomp	ncomp keepX	design ncomp	design ncomp keepX	ncomp	ncomp keepX
performance	tune, plot.tune		✓		✓		✓
	perf, plot.perf	✓	✓	✓	✓	✓	✓
	auroc	✓	✓	✓	✓	✓	✓
sample plot	plotIndiv	✓	✓	✓	✓	✓	✓
	plotArrow	✓	✓	✓	✓	✓	✓
	plotDiablo			✓	✓		
variable plot	plotVar	✓	✓	✓	✓	✓	✓
	plotLoadings	✓	✓	✓	✓	✓	✓
	circosPlot			✓	✓		
	cim	✓	✓	✓	✓	✓	✓
	network	✓	✓	✓	✓	✓	✓
variable list	selectVar	✓	✓	✓	✓	✓	✓

Figure S1: *List of the main mixOmics functions for supervised analyses.*

Electronic Supporting Information

Sweave and R code for PLS-DA analysis are available on our website at this link.

Sweave and R code for DIABLO analysis are available on our website at this link.
Sweave and R code for MINT analysis are available on our website at this link.