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

Florian Rohart [1], Benoît Gautier [1], Amrit Singh [2,3], K-A. Lê Cao [1,4*]

[1]The University of Queensland Diamantina Institute,
The University of Queensland, Translational Research Institute, QLD 4102, Australia,
[2]UBC James Hogg Research Centre for Heart Lung Innovation, St. Paul's Hospital and
[3]Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, BC, Canada
[4]Centre for Systems Genomics and School of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia

* kimanh.lecao@unimelb.edu.au

Abstract

The advent of high throughput technologies has led to a wealth of publicly available 'omics data coming from different sources, such as transcriptomics, proteomics, metabolomics. 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') to explain or predict biological conditions, but mainly for a single type of 'omics. In addition, commonly used methods are univariate and consider each biological feature independently.

We introduce `mixOmics`, 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 system 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 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. We illustrate our latest `mixOmics` integrative frameworks for the multivariate analyses of 'omics data available from the package.

Introduction

The advent of novel 'omics technologies (*e.g.* transcriptomics for the study of transcripts, proteomics for proteins, metabolomics for metabolites, etc) has enabled new opportunities for biological and medical research discoveries. Commonly, each feature from each technology (transcripts, proteins, metabolites, etc) is analysed independently through univariate statistical methods including ANOVA, linear models or t-tests. However, 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. Our package `mixOmics` proposes multivariate projection-based methodologies for 'omics data analysis as those provide several attractive properties to the data analyst (Lê Cao et al., 2017). Firstly, they are computationally efficient to handle large datasets, where the number of biological features (usually 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 the biological system under study. Lastly, their relaxed assumptions about data distribution make them highly flexible to answer topical questions across numerous

biology-related fields (Boulesteix and Strimmer, 2007; Meng et al., 2016). `mixOmics` multivariate methods have been successfully applied to statistically integrate data sets generated from difference biological sources, and to identify biomarkers in 'omics studies such as metabolomics, brain imaging and microbiome (Labus et al., 2015; Cook et al., 2016; Guidi et al., 2016; Mahana et al., 2016; Ramanan et al., 2016; Rollero et al., 2016).

We introduce `mixOmics` in the context of *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. We further extended our core method sparse Partial Least Square - Discriminant Analysis (sPLS-DA Lê Cao et al. 2011) that was originally developed for the supervised analysis of one dataset. Our two novel frameworks `DIABLO` and `MINT` focus on the integration of multiple data sets for different biological questions (Fig 1). `DIABLO` enables the integration of the same biological N samples measured on different 'omics platforms (N-integration, Singh et al. 2016), while `MINT` enables the integration of several independent data sets or studies measured on the same P predictors (P-integration, Rohart et al. 2017). 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), which ignores the heterogeneity between 'omics platforms and mainly highlights one single type of 'omics. The other common type of N-integration is to combine the molecular signatures identified from separate analyses of each 'omics (Günther et al., 2012), which disregards the relationships between the different 'omics functional levels. With P-integration, statistical methods are often sequentially combined to accommodate or correct for technical differences ('batch effects') among studies before classifying samples with a suitable classification method. Such sequential approaches are not appropriate for the prediction of new samples as they are prone to overfitting (Rohart et al., 2017). Our two frameworks model relationships between different types of 'omics data (N-integration) or integrate independent 'omics studies to increase sample size and statistical power (P-integration). Both frameworks aim at identifying biologically relevant and robust molecular signatures to suggest novel biological hypotheses.

The present article first introduces the main functionalities of `mixOmics`, then 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 from the package. Reproducible Sweave code is provided for all analyses.

Design and Implementation

`mixOmics` is a user-friendly R package dedicated to the exploration, mining, integration and visualisation of large data sets (Lê Cao et al., 2017). It provides attractive functionalities such as (i) insightful visualisations with dimension reduction (Fig 1), (ii) identification of molecular signatures and (iii) improved usage with common calls to all visualisation and performance assessment methods (Fig S2).

Multivariate projection-based methods

`mixOmics` offers a wide range of multivariate dimension reduction techniques designed to each answer specific biological questions, via unsupervised or supervised analyses. The `mixOmics` functions are listed in Table 1. Unsupervised analyses methods includes Principal Component Analysis (also based on NonLinear Iterative Partial Least Squares for missing values Wold 1975), Independent Component Analysis (Yao et al., 2012), Partial Least Squares regression (PLS, also known as Projection to Latent Structures, Wold 1966), multi-group PLS (Eslami et al., 2013), regularised Canonical Correlation Analysis (rCCA, González et al. 2008) and regularised Generalised Canonical Correlation Analysis (RGCCA, based on a PLS algorithm Tenenhaus and Tenenhaus 2011). Supervised analyses methods includes PLS - Discriminant Analysis (PLS-DA, Nguyen and Rocke 2002b,a; Boulesteix 2004), GCC-DA (Singh et al., 2016) and multi-group PLS-DA (Rohart et al., 2017). In addition, `mixOmics` provides novel sparse variants that enable *feature selection*, the identification of key predictors (e.g. genes, proteins, metabolites) that constitute a *molecular signature*. Feature selection is performed via ℓ^1 regularisation (LASSO, Tibshirani 1996), which is implemented into each method's statistical criterion to be optimised. For supervised analyses, `mixOmics` provides functions to assist users

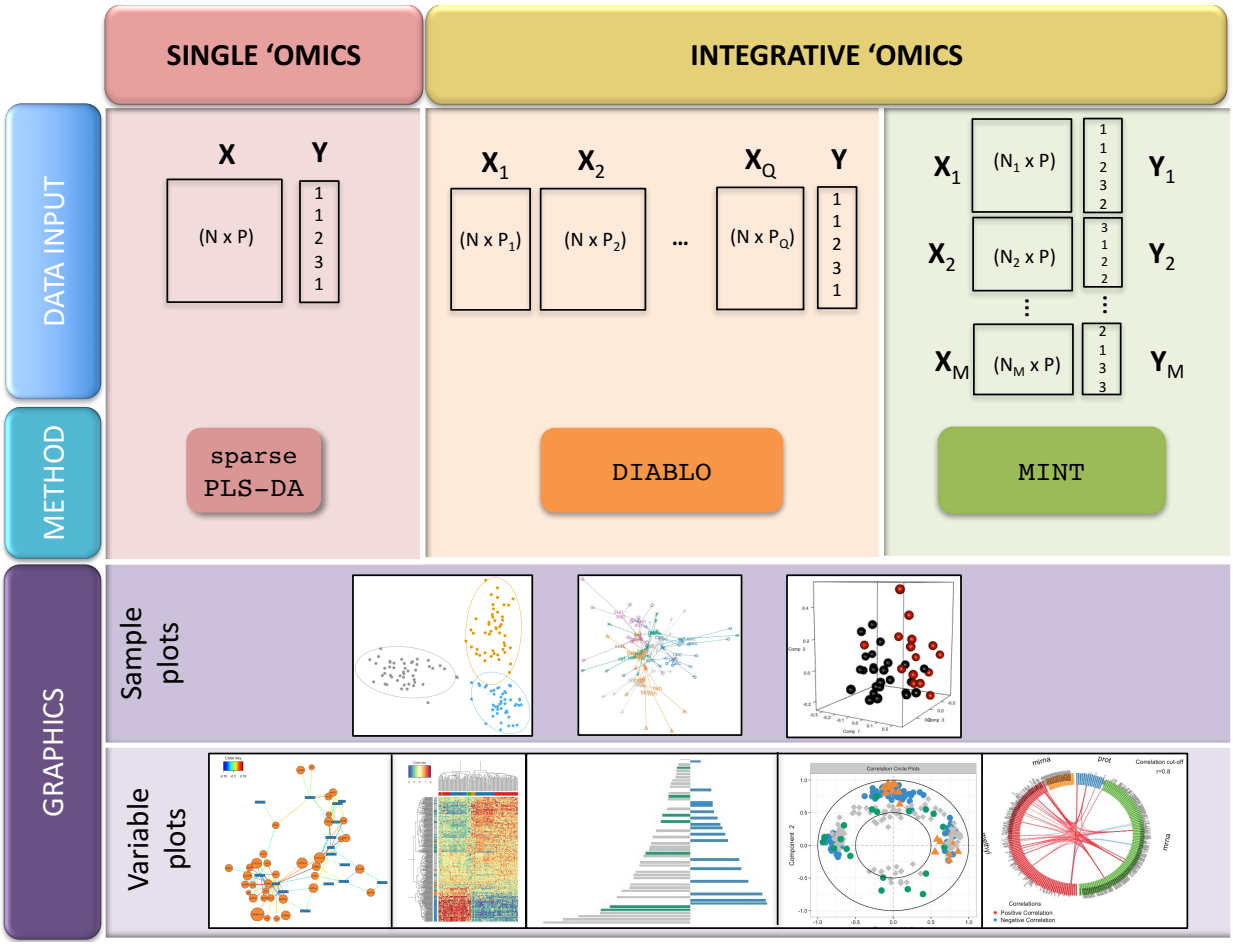


Figure 1: **Overview of the `mixOmics` multivariate methods for single and integrative 'omics supervised analyses.** X denote a predictor 'omics dataset, and Y a categorical outcome response (*e.g.* healthy *vs.* sick). Integrative analyses include N-integration with `DIABLO` (the same N samples are measured on different 'omics platforms), and P-integration with `MINT` (the same P 'omics predictors are measured in several independent studies). Sample plots depicted here use the `mixOmics` functions (from left to right) `plotIndiv`, `plotArrow` and `plotIndiv` in 3D; variable plots use the `mixOmics` functions `network`, `cim`, `plotLoadings`, `plotVar` and `circosPlot`. The graphical output functions are detailed in Supporting Information S2.

with the choice of parameters necessary for the feature selection process (see 'Choice of parameters' Section) to discriminate the outcome of interest (*e.g.* healthy *vs.* sick, or tumour subtypes, etc.).

All multivariate approaches listed in Table 1 are projection-based methods whereby samples are summarised by two sets of H *latent components* $(t_1, \ldots, t_H)$ that are defined as linear combinations of the original predictors. In the combinations $(t_1, \ldots, t_H)$, the weights of each of the predictors are indicated in the *loading vectors* $a_1, \ldots, a_H$. For instance, if X denotes the data matrix of P predictors, $X = (X^1, \ldots, X^P)$, then we define the first latent component $t_1 = Xa_1 = X^1 a_1^1 + \cdots + X^P a_1^P$. Therefore, to each loading vector a_h corresponds a latent component t_h, and there are as many pairs (t_h, a_h) as the chosen dimension H in the multivariate model, $h = 1, \ldots, H$, where $H << P$. The samples are thus projected into a smaller interpretable space spanned by the H latent components.

Table 1: Summary of multivariate projection-based methods available in `mixOmics` version 6.0.0 or above for different types of analysis frameworks.

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

Implementation

`mixOmics` is currently fully implemented in the R language and exports more than 30 functions to perform statistical analyses, tune the methods parameters and plot insightful visualisations. `mixOmics` mainly depends on the R base packages (parallel, methods, grDevices, graphics, stats, utils) and recommended packages (MASS, lattice), but also imports functions from 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 to visualise the results as depicted in Fig 1 (`plotIndiv, plotArrow, plotVar, plotLoadings`, etc), see Fig S2 and Supporting Information S2 for an exhaustive list.

Currently, seventeen multivariate projection-based methods are implemented in `mixOmics` to integrate large biological datasets, amongst which twelve have similar names `(mint).(block).(s)pls(da)`, see Table 1. To perform either N- or P-integration, we efficiently coded the functions as wrappers of a single main hidden and generic function that is based on our extension of the SGCCA algorithm (Tenenhaus et al., 2014) . The remaining five 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 (Fig S2).

`mixOmics` aims to provide insightful and user-friendly graphical outputs to interpret statistical and biological results, some of which (correlation circle plots, relevance networks, clustered image maps) were presented in details in (González et al., 2012). The function calls are identical for all multivariate methods via the use of R/S3 functions, as we illustrate in the Results Section. `mixOmics` offers various visualisations, including sample plots and variable plots, which are based on latent component scores and loading vectors, respectively (Fig 1). Additional graphical outputs are available in `mixOmics` to illustrate classification performance of multivariate models using the generic function `plot` (see Fig S2).

Class prediction of new samples

The supervised multivariate methods in `mixOmics` can be applied on an external test set to predict the outcome of new samples with the `predict` function (Table 1), or to assess the performance of the statistical model. The `predict` function calculates *prediction scores* for each sample, or predicted coordinates, which

are equivalent to the latent component scores in the training set.

Prediction distances. Our supervised models work with dummy indicator matrices to indicate the class membership of each sample (Supplemental Information S1.1 and Fig S3), and result in a prediction score for each outcome category k, $k = 1, \ldots, K$. Therefore, the scores across all classes K need to be combined to obtain the final prediction of a given test sample using a prediction distance. We propose distances such as 'maximum distance', 'Mahalanobis distance' and 'Centroids distance', as detailed in Supplemental Information S1.3.

Visualisation of prediction area. To visualise the effect of the prediction distance, we propose a graphical output of the prediction area that overlays the sample plot (example in Fig 2 and more details in Supplemental Information S2.1).

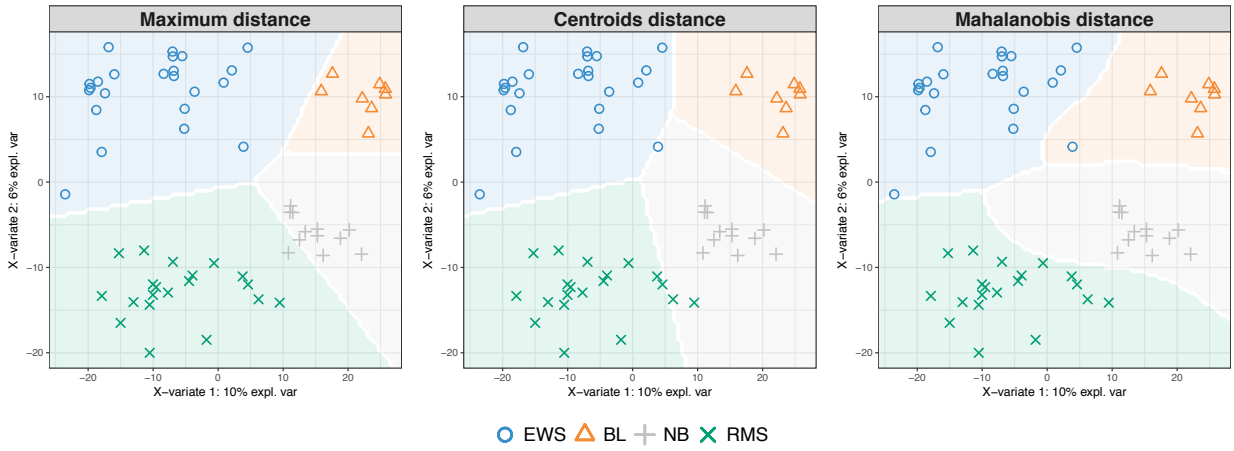


Figure 2: **Prediction area visualisation** on the Small Round Blue Cell Tumors data (SRBCT Khan et al. 2001) data, described in the Results Section, with respect to the prediction distance. From left to right: 'maximum distance', 'centroids distance' and 'Mahalanobis distance'. Sample prediction area plots from a PLS-DA model applied on a microarray data set with the expression levels of 2,308 genes on 63 samples. Samples are classified into four classes: Burkitt Lymphoma (BL), Ewing Sarcoma (EWS), Neuroblastoma (NB), and Rhabdomyosarcoma (RMS).

Prediction for N-integration. For N-integration, we obtain a predicted class *per* 'omics data set. The predictions are combined by majority vote (the class that has been predicted the most often across all data sets) or by weighted vote, where each 'omics dataset weight is defined as the correlation between latent components associated to that particular dataset and the outcome, from the training set. The final prediction is the class that obtains the highest weight across all 'omics datasets. Therefore the weighted vote gives more importance to the 'omics dataset that is best correlated to the outcome and reduces the number of ties when an even number of data sets are discordant in the case of majority vote. Ties are indicated as NA in our outputs.

Prediction for P-integration. In that specific case, the external test set can include samples from one of the independent studies used to fit the model, or samples from external studies, see (Rohart et al., 2017) for more details.

Choice of parameters for supervised analyses

For supervised analysis, `mixOmics` provides tools to help choosing the number of components H and the ℓ^1 penalty on each component for all sparse methods. These parameters are based on the performance evaluation of the statistical model.

Performance evaluation. For all supervised models, the tuning function `tune` implements repeated and stratified cross-validation (CV, see details in Supplemental Information S1.2) to obtain the predicted class of each sample. Performance is measured via overall misclassification error rate and Balanced Error Rate (BER). BER is appropriate in case of an unbalanced number of samples per class as it calculates the average proportion of wrongly classified samples in each class, weighted by the number of samples in each class. Therefore, BER is less biased towards majority classes during the performance assessment.

The choice of the parameters (described below) is made according to the best prediction accuracy - the lowest overall error rate or BER. Once the tuning process is completed, the performance of final model can be estimated using the `perf` function based on the chosen parameters and repeated CV. Additional evaluation outputs include the stability of the selected features across all CV runs, which represents a useful measure of reproducibility of the molecular signature (see example inElectronic Supplemental E1) and 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.

Number of components. For each supervised method the tuning function outputs the optimal number of components that achieve the best performance based on the overall error rate or BER. The assessment is data-driven and similar to the process detailed in (Rohart et al., 2016), where one-sided t-tests assess whether there is a gain in performance when adding a component to the model. Note that in practice (see some of our examples in the Results Section), we found that setting the number of components to $K - 1$, where K is the number of classes was sufficient to achieve the best classification performance (Lê Cao et al., 2011; Shah et al., 2016).

ℓ^1 penalty. Contrary to other R packages implementing ℓ^1 penalisation methods (*e.g.* `glmnet`, Friedman et al. 2010, `PMA`, Witten et al. 2013), `mixOmics` uses soft-thresholding to improve usability by replacing the ℓ^1 parameter by the number `keepX` of features to select on each dimension. The performance of the model is assessed for each value of `keepX` provided as a grid by the user from the first component to the H^{th} component, one component at a time. The grid needs to be carefully chosen as it is a trade-off between resolution and computational time. The `tune` function returns the `keepX` value that achieves the best predictive performance. In case of ties, the lowest `keepX` value is returned to obtain a minimal molecular signature. Note that the same grid of `keepX` values is used to tune each component.

Tuning with a constraint model. To improve performance, a new variant in all tuning functions was recently added to fit a *constraint* model (`constraint=TRUE`). The tuning is performed on the optimal *list* of selected features `keepX.constraint` from the previous components, as opposed to considering only 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., 2016). Our experience has shown that the constraint tuning and models can substantially improve the performance of the methods. We illustrate some examples in the Results section.

Special cases with N- and P- integration. For N-integration a grid of `keepX` needs to be provided for each type of 'omics data. Our example (Results section) illustrates the integration of three types of 'omics, for which three grids of `keepX` of length 13 were provided, resulting in a collection of $13 \times 13 \times 13 = 2197$ models to be assessed. For P-integration a Leave-One-Group-Out Cross Validation is performed where each study defines a subset that is left out once, as described in (Rohart et al., 2017) and in Supporting Information S5.

Results

Single 'omics supervised analyses with PLS-DA and sPLS-DA

We illustrate single 'omics multivariate methods PCA, PLS-DA and sPLS-DA on a microarray data set available from the package. The PLS-DA and sPLS-DA methods are described in the Supporting Information S3.

Data description. The study investigates 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).

Unsupervised and supervised analyses. Principal Component Analysis was first applied to assess similarities between tumour types (Fig. 3**A1**). This preliminary unsupervised analysis showed no separation between tumour types, but allows to visualise the more important sources of variation, which are summarised in the first two principal components (Fig. 3**A2**). A supervised analysis with PLS-DA focuses on the discrimination of the four tumour types (Fig. 3**B1**), and led to a good performance (Fig. 3**B2**, performance assessed when adding one component at a time in the model). We then applied sPLS-DA to identify specific discriminant genes for the four tumour types. The tuning process with the constraint model (see 'Choice of parameters' Section and Electronic Supplemental E1) led to a sPLS-DA model with 3 components and a molecular signature composed of $10, 40$ and 60 genes selected on the three component, respectively.

Results visualisation. The first sPLS-DA component discriminated BL vs the other tumour types (Fig. 3**C1**). The 10 genes selected on this component all had positive weight in the linear combination, and were highly expressed in BL (see Electronic Supplemental E1). The second component further discriminated EWS based on 40 selected genes. The genes with a negative weight were highly expressed in EWS while the genes with a positive weight were highly expressed in either NB or RMS. Finally, the third component discriminated both NB and RMS (not shown). The arrow plot displays the relationship between the samples summarised as a combination of selected genes (start of the arrow) and the categorical outcome (end of the arrow, Fig. 3**C2**).

A clustering analysis using a heatmap based on the genes selected on the first three components highlighted clusters corresponding to the four tumour types (Fig 3**C3**). ROC curve and AUC of the final model were also calculated using one-vs-all comparisons and led to satisfactory results on the first two components (Fig 3**C4**). The AUC for the first three components was 1 for all groups. Note that ROC and AUC are additional measures that may not reflect the performance of a `mixOmics` multivariate approaches since our prediction strategy is based on distances (see 'Performance' Section).

Summary. We illustrated the `mixOmics` framework for the supervised analysis of a single 'omics data set - here a microarray experiment. The full pipeline, results interpretation, associated R and Sweave codes are available in Electronic Supplemental E1. Such an analysis suggests novel biological hypotheses to be further validated in the laboratory, when one is seeking for a *signature* of a subset of features to explain, discriminate or predict a categorical outcome. The methods has been applied and validated in several biological and biomedical studies, including ours in proteomics and microbiome (Shah et al., 2016; Lê Cao et al., 2016).

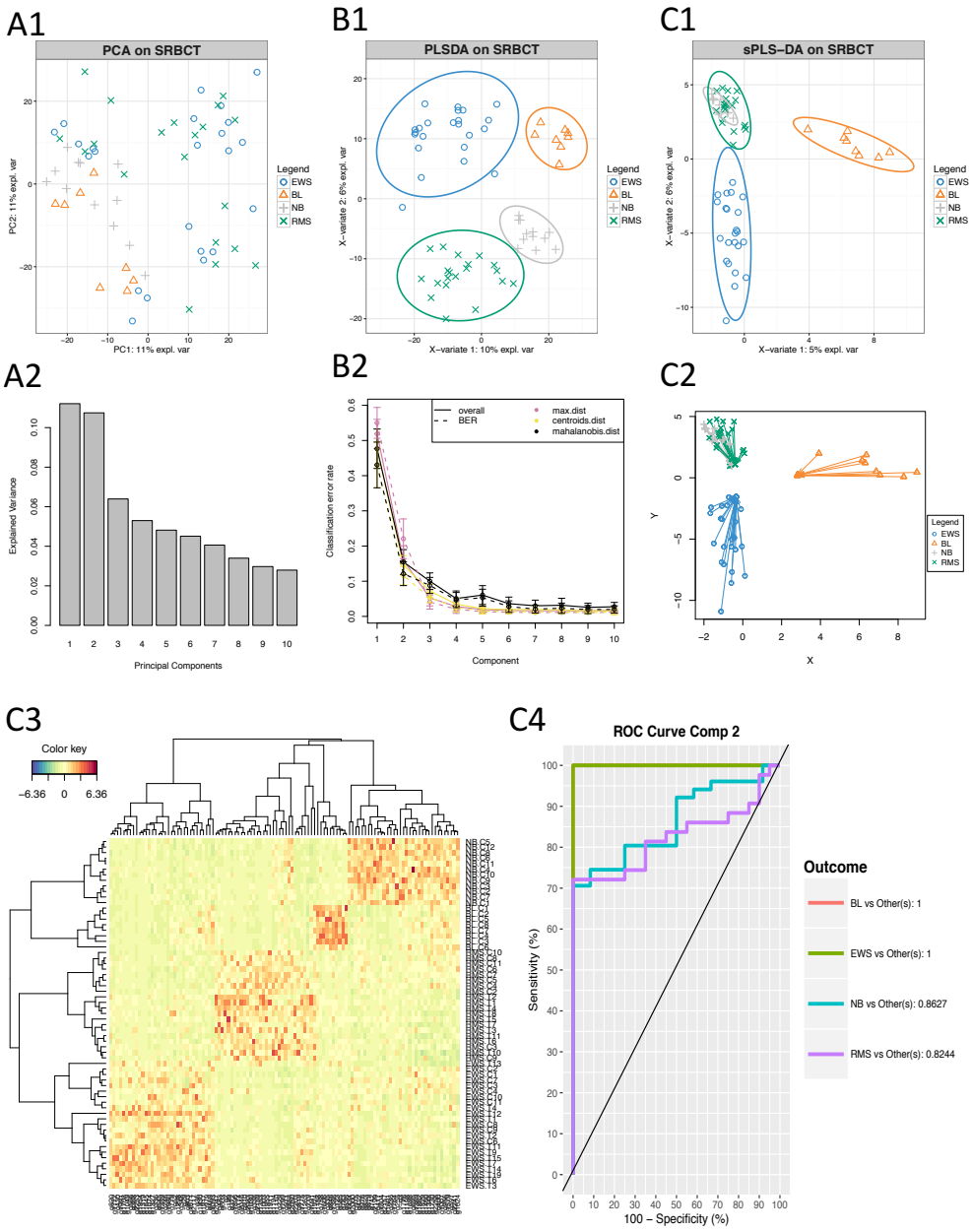


Figure 3: **Illustration of a single 'omics analysis with `mixOmics`. A) Unsupervised preliminary analysis with PCA**, **A1**: PCA sample plot, **A2**: percentage of explained variance per component. **B) Supervised analysis with PLS-DA**, **B1**: PLS-DA sample plot with confidence ellipse plots, **B2**: classification performance per component (overall and BER) for three prediction distances using repeated stratified cross-validation (50 * 5-fold CV). **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, see more details in Supplemental Information S2. **C3**: Clustered Image Map (Euclidian Distance, Complete linkage) where samples are represented in rows and selected features in columns (10, 40 and 60 genes selected on each component respectively), **C4**: ROC curve and AUC averaged using one-vs-all comparisons.

N-integration across multiple 'omics data sets with DIABLO

N-integration consists in integrating different types of 'omics data measured on the same N biological samples. In a supervised context, `DIABLO` performs N-integration by identifying a multi-'omics signature that discriminates the outcome of interest. Contrary to the concatenation and the ensemble approaches that also perform N-integration, `DIABLO` identifies a signature composed of highly correlated features across the different types of 'omics, by modelling relationships between the 'omics data sets(Singh et al., 2016). The `DIABLO` method is fully described in the Supporting Information S4. We illustrate one analysis on a multi-'omics breast cancer study available from the package.

Data description. The multi-'omics breast cancer study includes 150 samples from three types of 'omics: mRNA ($P_1 = 200$), miRNA ($P_2 = 184$) and proteomics ($P_3 = 142$) data. Prior to the analysis with `mixOmics`, the data were normalised and filtered for illustrative purpose. Samples are classified into three subgroups: 75 Luminal A, 30 Her2 and 45 Basal.

Choice of parameters and analysis. As we aim to discriminate three breast cancer subtypes we chose a model with 2 components. The tuning process with the constraint model (see 'Choice of parameters' Section and Electronic Supplemental E2) identified a multi-'omics signature of 5 and 6 mRNA features, 6 and 5 miRNA features and 6 and 18 proteomics features on the two components, respectively. Sample plots of the final `DIABLO` model in Figure 4**A** displayed a better discrimination of breast cancer subgroups with the mRNA and proteomics data than with the miRNA data. Fig 4**B** showed that the latent components of each 'omics were highly correlated between each others, highlighting the ability of `DIABLO` to model a good agreement between the data sets. The breast subtypes colors show that the components are also able to discriminate the outcome of interest.

Results visualisation. Several visualisation tools are available to help interpretation of the `DIABLO` results and to assess relationships between the selected multi-'omics features (see Supplemental Information S2 and Electronic Supplemental E2). The clustered image map (CIM) displayed a good classification of the three subtypes of breast cancer based on the 17 multi-'omics signature identified on the first component (Fig 4**C**). The CIM output can be complemented with a `circosPlot` which displays the different types of selected features on a circle, with links between or within 'omics indicating strong positive or negative correlations (Fig 4**D**). Those correlation are estimated using the latent components as a proxy, see more methodological details in (González et al., 2012). We observed strong correlations between miRNA and mRNA, but only a few correlations between proteomics and the other 'omics types. Correlation circle plots (Fig 4**E**) further highlight correlations between each selected feature and its associated latent component (see details in González et al. 2012). The 6 miRNA features selected on the first component were highly negatively correlated with the first component (red triangles close to the (-1,0) coordinates) while 4 of the 5 miRNA features selected on the second component were highly positively correlated to the second component (red triangles close to the (0,1) coordinates). Contrarily, most of the 24 proteomics features selected on the first two components were close to the inner circle, which implies a weak contribution of those features to both components. Finally, a relevance network output highlighted two clusters, both including features from the three types of 'omics (Fig 4**F**). Interactive view and .glm format are also available, see Supplemental Information S2.

Summary. We illustrated the `mixOmics` framework for the supervised analysis of a multiple 'omics study. The full pipeline, results interpretation and associated R and Sweave codes are available in Electronic Supplemental E2. Our `DIABLO` method identifies a discriminant and highly correlated multi-'omics signature. Predictive ability of the identified signature can be assessed (see E2) while the graphical visualisation tools enable a better understanding of the correlation structure of signature. Such method is the first of its kind to perform multivariate integration and discriminant analysis. `DIABLO` is useful to pinpoint a subset of different types of 'omics features in those large study, posit novel hypotheses, and can be applied as a first filtering step prior to refined knowledge- and/or data-driven pathway analyses.

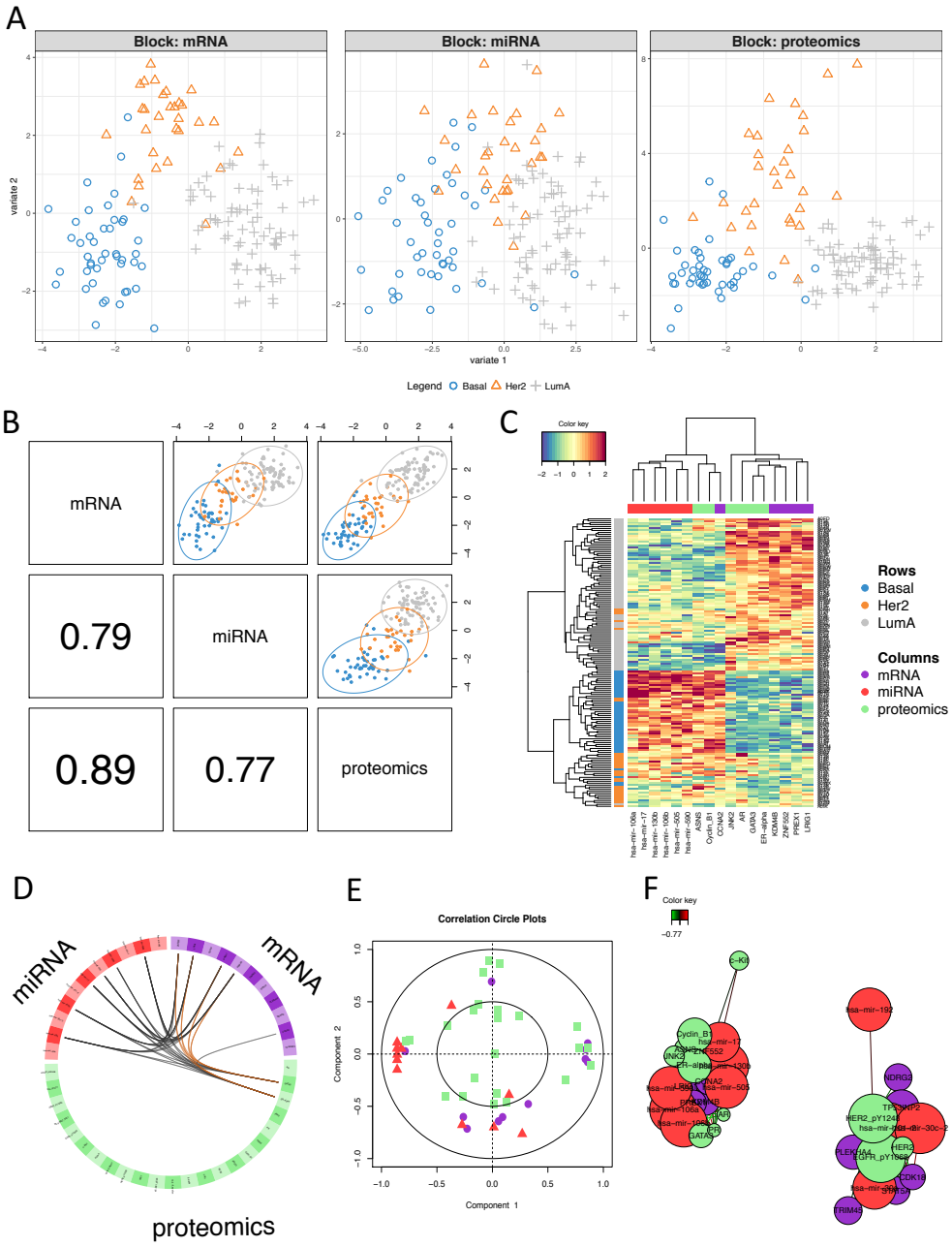


Figure 4: **Illustration of N-integrative supervised analysis with** `DIABLO`. **A**: sample plot per data set, **B**: sample scatterplot from `plotDiablo` displaying the first component in each data set (upper plot) and Pearson correlation between each component (lower plot). **C**: Clustered Image Map (Euclidian distance, Complete linkage) of the multi-omics signature. Samples are represented in rows, selected features on the first component in columns. **D**: 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, **E**: Correlation Circle plot representing each type of selected features, **F**: relevance network visualisation of the selected features.

P-integration across independent data sets with MINT

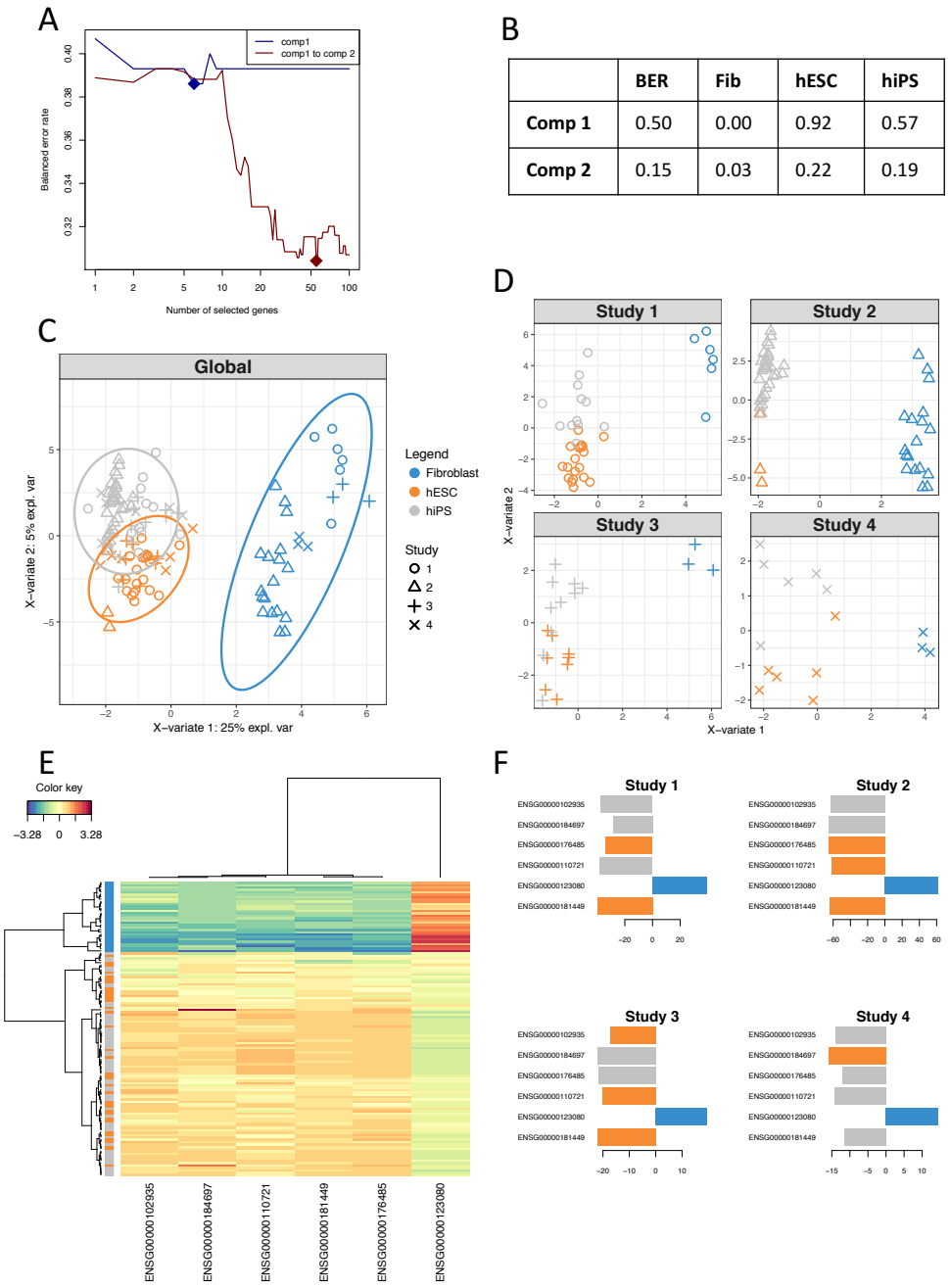
P-integration consists in integrating several independent studies measuring the same P predictors, and, in a supervised context, in identifying a robust molecular signature across multiple studies to discriminate biological conditions. The advantages of *P*- integration is to increase sample size while allowing to benchmark or compare similar studies. Contrary to usual approaches that sequentially accommodate for technical differences among the studies before classifying samples, `MINT` is a single step method that reduces overfitting and that predicts the class of new samples (Rohart et al., 2017). The `MINT` method is described in Supporting Information S5. We illustrate the `MINT` analysis on a stem cell study available from the package.

Data description. We combined four independent transcriptomics stem cell studies measuring the expression levels of 400 genes across 125 samples (cells). Prior to the analysis with `mixOmics`, the data were normalised and filtered for illustrative purpose. Cells were classified into 30 Fibroblasts, 37 hESC and 58 hiPSC.

Choice of parameters and analysis. As we aim to discriminate three classes, we chose a model with 2 components. The tuning process of a MINT sPLS-DA with a constraint model identified a molecular signature of 6 and 55 genes on the first two components, respectively (Fig 5**A**). A `MINT` model based on these parameters led to a BER of 0.15 (Fig 5**B**), which was substantially less than the BER of 0.52 from MINT PLS-DA when no feature selection was performed (see details in Electronic Supplemental E3).

Results visualisation. Global sample plot (Fig 5**C**) and study-specific sample plots highlighted a good agreement between the four studies (Fig 5**D**). The first component segregated fibroblasts vs. hiPSC and hESC, and the second component hiPSC vs. hESC. Such observation was confirmed with a Clustered Image Map based on the 6 genes selected on the first component (Fig 5**E**). Importantly, the loading plots depicted in Fig 5**F** showed consistent weights assigned by the MINT model to each selected genes across each independent study.

Summary. We illustrated the `MINT` analysis for the supervised integrative analysis of multiple independent 'omics studies. The full pipeline, results interpretation and associated R and Sweave codes are available in Electronic Supplemental E3. Our framework proposes graphical visualisation tools to understand the identified molecular signature across all independent studies. Our own applications of the method have showed strong potential of the method to identify reliable and robust biomarkers across independent transcriptomics studies (Rohart et al., 2016, 2017).



B

	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**: Parameter tuning of a MINT sPLS-DA model with two components, 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 maximum distance, **B**: Performance of the final MINT sPLS-DA model including selected features based on BER and classification error rate per class, **C**: Global sample plot with confidence ellipse plots, **D**: Study specific sample plot, **E**: Clustered Image Map (Euclidian Distance, Complete linkage). Samples are represented in rows, selected features on the first component in columns. **F**: Loading plot of each feature selected on the first component in each study, with color indicating the class with a maximal mean expression value for each transcript.

Conclusions and Future Directions

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, which 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.

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 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 for more relaxed assumptions about data distribution, data size 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.

While we illustrated our different frameworks on classical 'omics data in a supervised context, the package also include their unsupervised counterparts to investigate relationships and associations between features with no prior phenotypic or response information. Here we applied our multivariate frameworks to transcriptomics, proteomics and miRNA data. However, other types of biological data can be analysed, as well as data beyond the realm of 'omics as long as they are expressed as continuous values. The sPLS-DA framework was recently extended for microbiome 16S data (Lê Cao et al., 2016), and we will further extend `DIABLO` and `MINT` for microbiome - 'omics integration, as well as for genomic data and time-course experiments. These two promising integrative frameworks can also be combined for NP-integration, to combine multiple studies that each include several types of 'omics data and open new avenues for large scale multiple data integration.

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 version 6.0.0. KALC supervised the whole 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 who continuously help in improving the usability of the package.

Electronic Supporting Information

Supp E1: **Sweave and R codes for PLS-DA analysis** are available on our website at this link.

Supp E2: **Sweave and R codes for DIABLO analysis** are available on our website at this link.

Supp E3: **Sweave and R codes for MINT analysis** are available on our website at this link.

References

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. (2013). Multi-group pls regression: Application to epidemiology. In *New Perspectives in Partial Least Squares and Related Methods*, pages 243–255. Springer.

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.

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.-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., 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., Rohart, F., Gonzalez, I., Déjean, S., Gautier, B., Bartolo, F., Monget, P., Coquery, J., Yao, F., and Liquet., B. (2017). *mixOmics: Omics Data Integration Project*. R package version 6.1.3.

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 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).

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. (2016). 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, barretts 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 Lê Cao, K.-A. (2016). Diablo-an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv*, 067611.

Tenenhaus, A., Philippe, C., Guillemot, V., Lê Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569 – 83.

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.