*Bioinformatics*
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

# DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays

**Amrit Singh [1], Casey P. Shannon [1], Benoît Gautier [2], Florian Rohart [3], Michaël Vacher [4], Scott J. Tebbutt [1] and Kim-Anh Lê Cao [5,*]**

[1]Prevention of Organ Failure (PROOF) Centre of Excellence, University of British Columbia, Vancouver, BC, Canada, [2]The University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, QLD 4102, Australia, [3]Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia [4]Australian eHealth Research Centre, Commonwealth Scientific and Industrial Research Organisation, Brisbane, Queensland, Australia, [5]Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** In the continuously expanding omics era, novel computational and statistical strategies are needed for data integration and identification of biomarkers and molecular signatures. We present DIABLO, a multi-omics integrative and versatile method that seeks for common information across different data types through the selection of a subset of molecular features, while discriminating between multiple phenotypic groups.

**Results:** Using simulations and benchmark multi-omics studies, we show that DIABLO identifies features with superior biological relevance compared to existing unsupervised integrative methods, while achieving predictive performance comparable to state-of-the-art supervised approaches. DIABLO is versatile, allowing for modular-based analyses and cross-over study designs. In two case studies, DIABLO identified both known and novel multi-omics biomarkers (mRNA, miRNA, CpGs and proteins).

**Availability:** DIABLO is implemented in the mixOmics R Bioconductor package with functions for visualisation and choice of parameters to assist in the interpretation of the integrative analyses, along with tutorials on `http://mixomics.org` and our Bioconductor vignette.

**Contact:** kimanh.lecao@unimelb.edu.au

**Suppl. information:** Suppl. data are available at *Bioinformatics* online.

## 1 Introduction

Technological improvements have allowed for the collection of data from different molecular compartments (*e.g.* gene expression, DNA methylation status, protein abundance) resulting in multiple omics (multi-omics) data from the same set of biospecimens or individuals (*eg.* transcriptomics, proteomics, metabolomics). Systems biology approaches, by incorporating data from multiple biological compartments, provide improved biological insights compared to traditional single omics analyses (Zhu *et al.*, 2012; Kim *et al.*, 2013; Wang *et al.*, 2014). One reason might be that interactions between omics layers is not taken into account in single omics analysis and prevents the reconstruction of accurate molecular networks. These molecular networks are dynamic, changing under perturbed conditions such as disease, response to therapy, and environmental exposures. Therefore, adopting a holistic approach by integrating multi-omics data may bridge this information gap, and uncover networks that are representative of the underlying molecular mechanisms (Ritchie *et al.*, 2015; Yugi *et al.*, 2016).

Many strategies (component-based, message-passing, Bayesian methods, network-analysis, classification schemes) have been proposed for multi-omics data integration to answer various questions, incorporating experimental data as well as curated data from biological databases (see Suppl. Fig. S1, Zeng and Lumley 2018; Ritchie *et al.* 2015; Bersanelli *et al.* 2016; Meng *et al.* 2016; Huang *et al.* 2017; Rohart *et al.* 2017b). These include data-driven methods for identifying novel phenotypic clusters such

 

as Similarity Network Fusion (Wang *et al.*, 2014), Bayesian Consensus Clustering (Kirk *et al.*, 2012), and methods for extracting common sources of variation such as joint Non-negative Matrix Factorization (Zhang *et al.*, 2012), Joint and Individual Variation Explained (Lock *et al.*, 2013), sparse MultiBlock Partial Least Squares (Li *et al.*, 2012), regularized and sparse Generalized Canonical Correlation Analysis (Tenenhaus and Tenenhaus, 2011; Tenenhaus *et al.*, 2014) and Multi-Omics Factor Analysis (Argelaguet *et al.*, 2018). Other methods such as Passing Attributes between Networks for Data Assimilation (Glass *et al.*, 2013), Sparse Network regularized Multiple Non-negative Matrix Factorization (Zhang *et al.*, 2011) and Reconstructing Integrative Molecular Bayesian NETworks (Zhu *et al.*, 2012) can be used to incorporate curated data with experimental data in order to reconstruct biological networks. All of these methods are examples of unsupervised multi-omics data integration, that is, without the need of sample labels that categorize samples based on a certain phenotype or trait. However, researchers are also interested in multi-omics biomarkers that are predictive of disease, *i.e.* supervised methods in which molecular patterns that span across biological domains explain or characterise a known phenotype.

Supervised data integration approaches for the classification of multiple phenotypes (*e.g.* PAM50 breast cancer phenotypes) include multi-step approaches that concatenate all data prior to applying a classification model, or ensemble-based in which a classification model is applied separately to each omics data and the resulting predictions are combined based on average or Majority vote (Günther *et al.*, 2012). These approaches can be biased towards certain omics data types, and do not account for interactions between omic layers (Aben *et al.*, 2016; Ma *et al.*, 2016). Recently, classification approaches such as Network smoothed t-statistics Support Vector Machines (Cun and Fröhlich, 2013), Generalized Elastic Net (Sokolov *et al.*, 2016), and adaptive Group-Regularized ridge regression (van de Wiel *et al.*, 2016) have incorporated curated biological data such as PPI data, genetic pathway data, and type of methylation probes. These methods are still limited to single omics data such that, either the concatenation or ensemble-based schemes must be applied to incorporate additional data-types. Other approaches include The Analysis Tool for Heritable and Environmental Network Associations (ATHENA) based on a Grammatical Evolution Neural Network that integrates multi-omics data for the prediction of clinical outcomes (Kim *et al.*, 2013). However, the approach requires initial filtering, feature selection and modelling independently on each omics dataset prior to integration.

We introduce DIABLO, a multi-omics method that simultaneously identifies key omics variables (mRNA, miRNA, CpGs, proteins, metabolites, *etc.*) during the integration process and discriminates phenotypic groups. DIABLO (**D**ata **I**ntegration **A**nalysis for **Bi**omarker discovery using **L**atent c**O**mponents) maximizes the common or correlated information between multiple omics datasets. It is the first multivariate integrative classification method of its kind that builds a predictive model for prediction on new samples. The method is based on Projection to Latent Structure, allowing for powerful visualizations. DIABLO is highly flexible in the type of experimental design it can handle, ranging from classical single time point to cross-over and repeated measures studies. Modular-based analysis can also be incorporated using pathway-based module matrices (Langfelder and Horvath, 2008) instead of the original omics matrices. We demonstrate the capabilities and versatility of DIABLO below, both in simulated and real multi-omics studies to identify relevant biomarkers of various diseases.

## 2 Methods

### 2.1 General multivariate integrative framework.

DIABLO extends sparse generalized canonical correlation analysis (sGCCA, Tenenhaus *et al.* 2014) to a classification or supervised framework. sGCCA is a multivariate dimension reduction technique that uses singular value decomposition and selects co-expressed (correlated) variables from several omics datasets. sGCCA maximizes the covariance between linear combinations of variables (latent component scores) and projects the data into the smaller dimensional subspace spanned by the components. The selection of the correlated molecules across omics levels is performed internally with $\ell_1$ penalization on the variable coefficient vector defining the linear combinations. Since all latent components are scaled in the algorithm, sGCCA maximizes the correlation between components. However, we will retain the term 'covariance' instead of 'correlation' throughout this section to present the general sGCCA framework.

Denote $Q$ normalized, centered and scaled datasets $X^{(1)}(N \times P_1)$, $X^{(2)}(N \times P_2)$, ..., $X^{(Q)}(N \times P_Q)$ measuring the expression levels of $P_1, \ldots, P_Q$ 'omics variables on the same $N$ samples. sGCCA solves the optimization function for each component $h = 1, \ldots, H$:

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

where $a_h^{(q)}$ is the variable coefficient or loading vector on component $h$ associated to the residual matrix $X_h^{(q)}$ of the dataset $X^{(q)}$, and $C = \{c_{i,j}\}_{i,j}$ is the design matrix. $C$ is a $Q \times Q$ matrix that specifies whether datasets should be connected. Elements in $C$ can be set to zeros when datasets are not connected and ones where datasets are fully connected, as we further describe in section 2.2. In addition in (1), $\lambda^{(q)}$ is a non-negative parameter that controls the amount of shrinkage and thus the number of non-zero coefficients in $a_h^{(q)}$. Similar to Lasso (Tibshirani, 1996) and other $\ell_1$ penalized multivariate models developed for single omics analysis (Lê Cao *et al.*, 2011), the penalization enables the selection of a subset of variables with non-zero coefficients that define each component score $t_h^{(q)} = X_h^{(q)} a_h^{(q)}$. The result is the identification of variables that are highly correlated between and within omics datasets.

The sGCCA model (1) is iterative; a first set of coefficient vectors $(a_1^{(1)}, \ldots a_1^{(Q)})$ is obtained by maximizing (1) for $h = 1$ with $X_1^{(q)} = X^{(q)}$, before maximizing (1) for $h = 2$ using residual matrices $X_2^{(q)} = X_1^{(q)} - t_1^{(q)} a_1^{(q)}, 1 \leq q \leq Q$. This process is repeated until a sufficient number of dimensions (or set of components) is obtained. The underlying assumption of sGCCA is that the major source of common biological variation can be extracted via the first sets of component scores $t_1^{(q)}, \ldots, t_h^{(Q)}$, while any unwanted variation due to heterogeneity across the datasets $X^{(q)}$ does not impact the statistical model. The optimization problem (1) is solved using a monotonically convergent algorithm (Tenenhaus *et al.*, 2014).

### 2.2 DIABLO: supervised analysis and prediction

To extend sGCCA for a classification framework, we substitute one omics dataset $X^{(q)}$ in (1) with a dummy indicator matrix $Y$ $(N \times G)$ to indicate the class membership of each sample, where G is the number of phenotype groups. For easier use of DIABLO, we replaced the $\ell_1$ penalty parameter $\lambda^{(q)}$ by the number of variables to select in each dataset and each component, as there is a direct correspondence between both parameters.

*Input data.* While DIABLO does not assume particular data distributions, all datasets should be normalized appropriately according to each omics

platform and preprocessed if necessary (see normalisation steps described in Suppl. Section S2 for each case study). Samples should be represented in rows in the data matrices and match the same samples across omics datasets. The phenotype outcome $y$ is a factor indicating the class membership of each sample and is internally transformed into a dummy matrix $Y$ in `mixOmics`. In addition, each variable is centered and scaled internally, as is conventionally performed in PLS-based models. A multilevel variance decomposition option is available for repeated measures and cross-over study designs, as illustrated in the Asthma study section 3.

*Design matrix.* The design matrix C is a $(Q \times Q)$ matrix with values ranging from 0 to 1, which specifies whether datasets should be connected, see (1). In our simulation study, we evaluated two scenarios: a null design (DIABLO_null) when no omics datasets are connected, and a full design when all datasets are connected (DIABLO_full):

$$C_{\text{null}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad C_{\text{full}} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Every dataset is then connected to the outcome $Y$ internally. For the two case studies Breast cancer and Asthma the design matrix was chosen based on our proposed method (see Parameters tuning in 2.3). The design matrix is not restricted to 0 and 1 values only and a compromise between correlation and discrimination can also be modelled as described in Rohart *et al.* (2017b).

*Consensus class prediction for each new sample.* For a $new$ sample, a set of $H$ predicted component scores $(t^{(q)}_{1,new}, \ldots, t^{(q)}_{H,new})$ can be calculated for each type of omics $q$ by using the estimated loadings vectors $a^{(q)}$ from DIABLO. The predicted class of a new sample for each dataset is obtained from the predicted score using one of the distances Maximum, Centroids or Mahalanobis as detailed in Rohart *et al.* (2017b), which results in $Q$ class memberships for a new sample.

Since the different omics datasets may not all agree on a predicted class, a consensus class membership is determined using either a majority vote, a weighted majority vote or by averaging all $t^{(q)}_{h,new}$ for each component $h$ across all $Q$ datasets then applying a prediction distance scheme. In case of ties in the majority vote scheme, 'NA' is allocated as a prediction but is counted as a misclassification error during the performance evaluation. For the weighted majority vote, each omics dataset is weighted by the correlation between its latent components and the outcome, that is, stronger predictive datasets are up-weighted as compared to weaker omics datasets. As the class prediction relies on individual vote from each omics set, DIABLO allows for some missing datasets $X_k$ during the prediction step, as illustrated in the Breast Cancer case study. We used the Centroid distance for the weighted majority vote scheme (Breast Cancer study) and the Maximum distance for the average vote scheme (Asthma study) as those led to best performance (see Rohart *et al.* 2017b for details about distance measures and proposed voting schemes).

## 2.3 Parameters tuning

There are three types of parameters to tune in DIABLO.
- The design matrix $C$ can be determined using either prior biological knowledge, or a data-driven approach. The latter approach can use PLS that models pair-wise associations between omics datasets Lê Cao *et al.* (2008). If the correlation between the first component of each omics dataset is above a given threshold (*e.g.* 0.8) then a connection between those datasets is included in $C$ as a 1 value.
- The number of components: in several analyses we found that $G - 1$ components could extract sufficient information to discriminate all phenotype groups (Lê Cao *et al.*, 2011), but this can be assessed by evaluating the model performance across all specified components, as described below, and can be aided with graphical outputs such as sample plots to visualize the discriminatory ability of each component.
- The number of variables to select per dataset and per component. A grid composed of a small number of variables (<50 with steps of 5 or 10) may suffice as we did not observe substantial changes in the classification performance during our case study analyses. The variable selection size can also be guided according to the downstream biological interpretation. For example, a gene-set enrichment analysis may require a larger set of features than a literature-search interpretation.

## 2.4 DIABLO visualisation outputs

To facilitate the interpretation of the integrative analysis, several types of graphical outputs were proposed and implemented in `mixOmics`.
*Sample plots* include a consensus plot which depicts the samples by calculating the average of the components from each dataset (Fig 3A). Omic-specific sample plots can also be obtained by plotting components associated to each dataset (Suppl. Fig. S14).
*Variable plots* give more insights into the variables that were selected by DIABLO. Our new circos plot represents correlations between and within selected variables from each dataset. The association between variables is computed using a similarity score that is analogous to a Pearson correlation coefficient (see González *et al.* 2012); this association is displayed as a color-coded link inside the plot to represent a positive or negative correlation above a user-specified threshold. The selected variables are represented on the side of the plot, with side colors indicating each omics type, optional line plots represent the expression levels in each phenotypic group (see Suppl. Fig. S20).
*Clustered Image Maps (CIM)* based on the Euclidean distance and the complete linkage display an unsupervised clustering between the selected variables (centered and scaled) and the samples (see Suppl. Fig. S15). Color bars represent the sample phenotypic groups (columns) and the type of omics (rows) variables (see González *et al.* 2012).

Table 1. Overview of multi-omics datasets analyzed for method benchmarking and in two case studies. The breast cancer case study includes training (test) datasets for all omics types except proteins.

| Dataset | $n$ | Omics | $p$ |
|---|---|---|---|
| Colon | 92 | mRNA | 17,814 |
| | high: 33 | miRNA | 312 |
| Wang *et al.* (2014) | low: 59 | CpGs | 23,088 |
| Kidney | 122 | mRNA | 17,665 |
| | high: 61 | miRNA | 329 |
| Wang *et al.* (2014) | low: 61 | CpGs | 24,960 |
| Glioblastoma | 213 | mRNA | 12,042 |
| | high: 105 | miRNA | 534 |
| Wang *et al.* (2014) | low: 108 | CpGs | 1,305 |
| Lung | 106 | mRNA | 12,042 |
| | high: 53 | miRNA | 353 |
| Wang *et al.* (2014) | low: 53 | CpGs | 23,074 |
| Breast Cancer | 989 | mRNA | 16,851 |
| | Basal: 76 (102) | miRNA | 349 |
| | Her2: 38 (40) | CpGs | 9,482 |
| | LumA: 188 (346) | Proteins | 115 (0) |
| TCGA Research Network (2012) | LumB: 77 (122) | | |
| Asthma | 28 | Cell-types | 9 |
| | Pre: 14 | mRNA modules | 229 |
| Singh *et al.* (2013, 2014) | Post:14 | Metabolite modules | 60 |

## 3 Results

### 3.1 Correlation and discrimination trade-off

Three omic datasets consisting of 200 samples (100 in each of the two phenotypic groups) and 260 variables were simulated (details in Suppl. Section S1). Each dataset included four types of variables: 30 correlated-discriminatory (*corDis*), 30 uncorrelated-discriminatory (*unCorDis*), 100 correlated-nondiscriminatory (*corNonDis*) and 100 uncorrelated-nondiscriminatory (*unCorNonDis*) variables. DIABLO models with either a null or full design (DIABLO_null, DIABLO_full) were compared with existing integrative classification schemes based on classification performance (10-fold cross-validation - CV, averaged over 20 simulations) and variable selection (Fig. 1). The covariance between datasets was held constant, with fold-change (FC) varying from 0 to 2, and noise (SD) between 0.2 to 1. When FC = 0, the error rate was $\sim 50\%$ for all methods regardless of noise level (Suppl. Figure S2). When FC = 1, DIABLO_full yielded a higher error rate than all other methods, for noise < 1. However, when noise = 1 and FC = 1, all methods performed similarly. Finally,

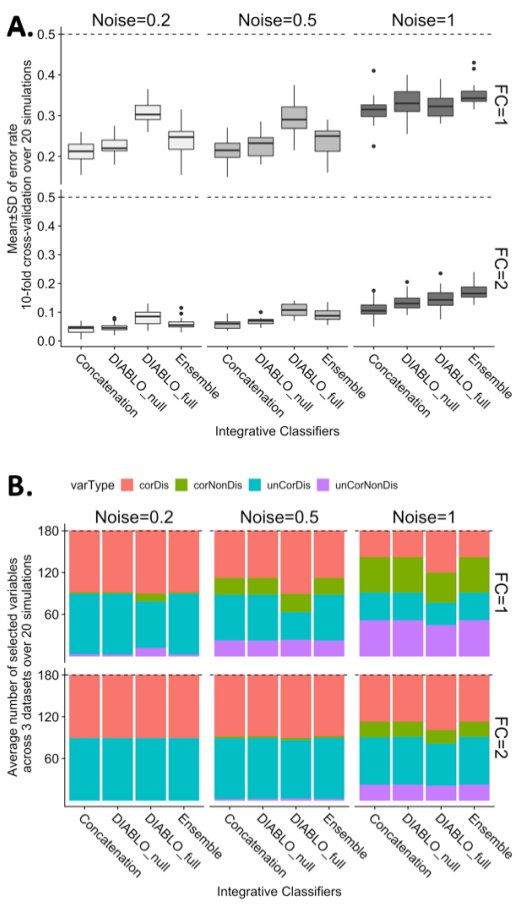


**Fig. 1.** Simulation study. A) Classification error rates (10-fold CV averaged over 20 simulations) for different fold-changes (FC) between groups and varying level of noise (sd). Dashed line indicates a random performance (error rate = 50%). B) Types of variables selected by the different classification methods amongst the 180 variables selected for each classification method.

when FC = 2 (higher than both the covariance and noise levels) the error rate of the DIABLO_full model decreased further. We hypothesized that the increased error rate between the DIABLO models was due to the covariance constraint used to extract a common source of variation across datasets instead of independent sources of variation from each dataset. Therefore, we varied the covariance value between datasets and performed

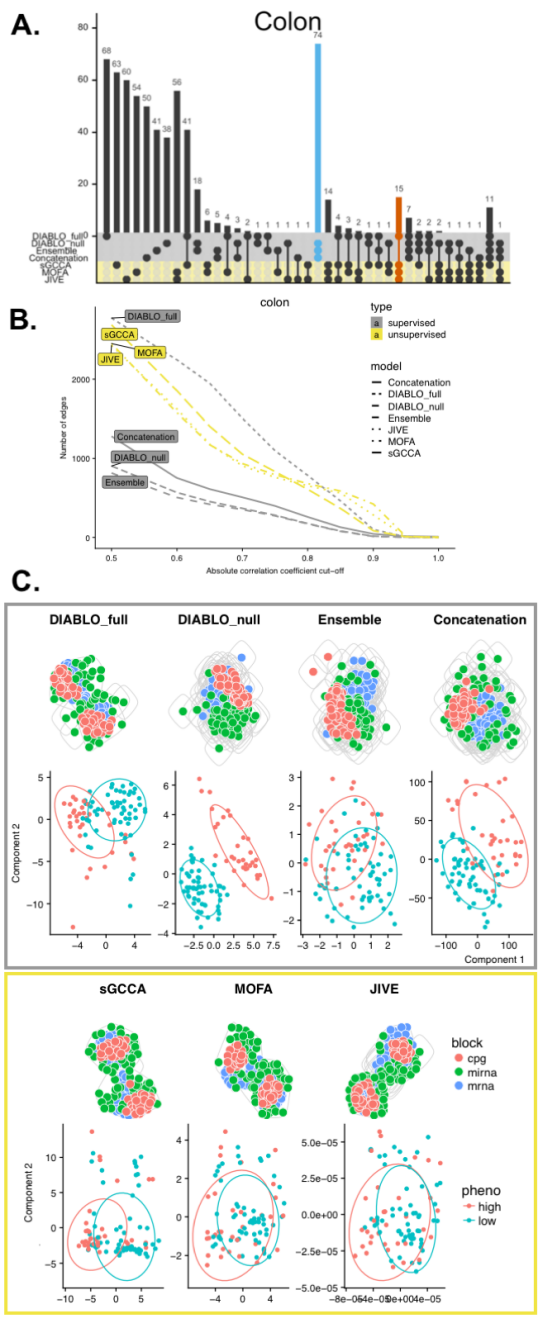


**Fig. 2.** Benchmark for colon cancer. A) Overlap of features selected by supervised or unsupervised methods. B) Number of correlated variables in the biomarker panels for various Pearson correlation cut-offs. C) Top: network modularity of each multi-omic biomarker panel. Gray circles depict modules based on the edge betweenness index from the igraph R-library. Bottom: consensus component plots depicting the separation of subjects in the high and low survival groups. Similar patterns were observed for kidney, gbm and lung cancer datasets, see Suppl. Figs S5-S9.

similar comparisons as described in Suppl. Figure S3. We found that increasing the covariance between datasets significantly increased the error rate for DIABLO_full, but not for DIABLO_null. When we added more components in DIABLO, allowing for additional independent information to be included, the classification performance improved and yielded similar results in both DIABLO designs. We hence concluded from this simulation study that the design in DIABLO achieves a trade-off between correlation and discrimination. DIABLO_null focuses on selecting discriminatory

variables and disregards most of the correlation between datasets, whereas DIABLO_full selects highly correlated and discriminatory variables across all datasets. Variables selected by DIABLO_full reflect the correlation structure between biological datasets, and may provide a balance between prediction accuracy and biological insight, as described in the next sections.

### 3.2 Benchmark: DIABLO identifies highly interconnected networks with superior biological enrichment

We applied various integrative approaches to cancer multi-omics datasets (mRNA, miRNA, and CpG): colon, kidney, glioblastoma (gbm) and lung, to identify multi-omics biomarker panels predictive of high and low survival times (see Table 1, Suppl. Section S2) and studied the network properties and biological enrichment of the selected features. Component-based integrative approaches were compared: supervised methods included concatenation and ensemble-based schemes using sparse Partial Least Squares Discriminant Analysis (sPLSDA, Lê Cao *et al.* 2011), DIABLO_null and DIABLO_full, and unsupervised approaches included sGCCA, Multi-Omics Factor Analysis (MOFA, Argelaguet *et al.* 2018), and Joint and Individual Variation Explained (JIVE, Lock *et al.* 2013) (see Suppl. Section S3 for parameter settings). Each biomarker panel consisted of 180 features (a number of variables arbitrarily chosen with the largest weights on the first two components in order to compare all methods). Across all cancer datasets, the largest overlap between biomarker panels was observed between all supervised methods with the exception of DIABLO_full whose selection was more similar to those identified with unsupervised methods (Fig. 2A and Suppl. Fig. S5 for the other studies). Interestingly, we observed similarities between the features identified by DIABLO_full and the unsupervised integrative approaches based on the following characteristics: 1) correlation between features - a large number of connections or edges regardless of the correlation cut-off was observed (Fig. 2B, Suppl. Fig. S6), 2) network attributes such as high graph density, low number of communities and large number of triads (Suppl. Fig. S7) and 3) small number of densely connected modules (Fig. 2C and Suppl. Fig. S8). The trade-off in selecting correlated features by DIABLO_full was at a slight expense of discrimination, as can be observed in the component plots which depict the separation of the high and low survival groups (Fig. 2C and Suppl. Fig. S9). DIABLO_null also achieved a good separation of the survival groups, but with biomarker panel characteristics similar to those of other supervised methods. Internal validation on the benchmark datasets showed that DIABLO_null led to better cluster consistency according to phenotypic groups compared to all other methods (Suppl. Figure S10).

Gene set enrichment analysis on each biomarker panel using gene symbols of mRNAs and CpGs against 10 gene set collections showed that DIABLO_full identified the greatest number of significant gene sets across the gene set collections and generally ranked higher than the other methods in colon (7 collections), gbm (5) and lung (5) cancer datasets (Suppl. Section S4 and Table S1). JIVE outperformed all methods in the kidney cancer datasets (6 collections). In conclusion for this benchmark study, DIABLO_full aims at explaining the correlation structure between multiple omics layers and the phenotype of interest, leading to the greatest number of known biological gene sets such as pathway, functions and processes.

### 3.3 DIABLO identifies known and novel multi-omics biomarkers of breast cancer subtypes

We applied DIABLO with a full design to the TCGA breast cancer study (TCGA Research Network, 2012) to characterize and predict PAM50 breast cancer subtypes (Table 1, Suppl. Fig. S11). Processing and normalisation is described in Suppl. Section S2. The optimal multi-omics biomarker panel size was identified using a grid approach where for any given combination of variables, we assessed the classification performance using a 5-fold CV repeated 5 times (Suppl. Fig. S12) and chose the number of variables that resulted in the minimum balanced error rate (BER, see details in Rohart *et al.* 2017b). Our panel consisted of 45 mRNA, 45 miRNAs, 25 CpGs and 55 proteins selected across three components with a balanced error rate of 17.9$\pm$1.9%. This panel identified many variables with previously known associations with breast cancer, according to MolSigDB (Liberzon *et al.*, 2015), miRCancer (Xie *et al.*, 2013), Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2005), and DriverDBv2 (Chung *et al.*, 2015). In addition, we identified several variables that were not found in any database and that may represent novel biomarkers of breast cancer (Suppl. Fig. S13). Fig.

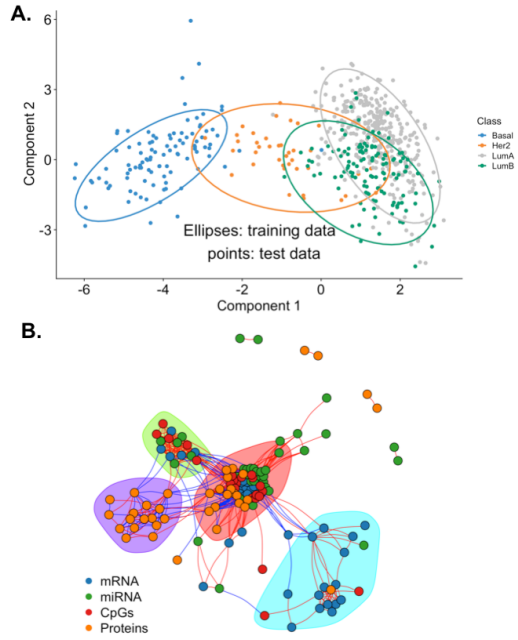


**Fig. 3.** A Multi-omics biomarker panel predictive of breast cancer subtypes. A) DIABLO consensus component plot based on the identified multi-omics biomarker panel: test samples are overlaid with 95% confidence ellipses calculated from the training data. B) Network visualization of the biomarker panel highlighting correlated variables (Pearson correlation > |0.4|) and four communities based on the edge betweeness index.

3A shows that the majority of the test samples were located within the ellipses built on the training set, suggesting a reproducible multi-omics biomarker panel from the training to the test set (see Suppl. Fig. S14 for omic-specific component plots). On the test set, a BER of 22.9% indicated a relatively good prediction accuracy of breast cancer subtypes. The consensus plot corresponded strongly with the mRNA component plot, with a strong separation of the Basal (error rate = 4.9%) and Her2 (20%) subtypes, and a weak separation of Luminal A and Luminal B (error rates of 13.3% and 53.3% respectively) subtypes. A heatmap of the biomarker panel showed similar results (Suppl. Fig. S15). Overall, the features of the multi-omics biomarker panel formed a network of four densely connected clusters of variables (Fig. 3B). The largest cluster of 72 variables (20 mRNAs, 21 miRNAs, 15 CpGs and 16 proteins) was further investigated using gene set enrichment analysis (Suppl. Fig. S16). We identified many cancer-associated pathways (*e.g.* FOXM1 pathway, p53 signaling pathway), DNA damage and repair pathways (*e.g.* E2F mediated regulation of DNA replication, G2M DNA damage checkpoint) and various cell-cycle pathways (*e.g.* G1S transition, mitotic G1/G1S phases).

Therefore, DIABLO was able to identify a biologically plausible multi-omics biomarker panel that generalized to test samples. The panel also included unknown molecular features in breast cancer suggesting novel molecular features whose importance would require further experimental validations.

### 3.4 Competitive classification performance of DIABLO

In the breast cancer study we used independent test data to compare DIABLO, Elastic Net classifiers and both Concatenation-based and ensemble-based schemes based on the sPLSDA (see Suppl. Section S5 for details). Parameters of each integrative method were tuned using 5x5-fold CV on the training datasets to identify the optimal model, before assessing the methods' prediction performance on the test data. We found that DIABLO models performed similarly to Concatenation-based classifiers and out-performed Ensemble-based classifiers (Suppl. Table S2). Concatenation-based classifiers were biased towards the more predictive variables (mRNA or CpGs), whereas DIABLO selected variables evenly across datasets and had similar error rates between training and test datasets.

### 3.5 Repeated measures and module-based analysis

The asthma study investigated blood biosignatures in response to allergen inhalation challenge (AIC) in 14 subjects. Blood was collected pre- and two hours post-AIC (Singh *et al.*, 2013, 2014). Cell-type frequencies, leukocyte gene transcript expression and plasma metabolite abundances were measured (Table 1). A module-based approach (*a.k.a* eigengene summarization, Langfelder and Horvath 2008) was used to transform both gene expression and metabolite datasets into pathway datasets to include prior biological knowledge in DIABLO (Suppl. Section S6) (Allahyar and De Ridder, 2015; Cun and Fröhlich, 2013; Sokolov *et al.*, 2016). Consequently, each variable represented the pathway activity expression level for each sample rather than gene or metabolite expression in these datasets. We used KEGG for mRNA pathways and annotations provided by Metabolon Inc. (Durham, North Carolina, USA) for the metabolites pathways (Fig. 4A).

We compared the standard DIABLO with a multilevel model (mDIABLO) that accounts for the repeated measures (pre/post) experimental design by isolating the within-sample variation from each dataset (Liquet *et al.*, 2012) (Suppl. Section S7). Both DIABLO approaches were applied to identify a multi-omics biomarker panel consisting of cells, gene and metabolite modules that discriminated pre- from post-AIC samples. mDIABLO outperformed DIABLO (AUC=98.5% vs. AUC=62.2%) with greater separation between the pre- and post-AIC samples (Fig. 4B and C). Common features (pathways) were identified across omics-types in mDIABLO but not in standard DIABLO (Suppl. Fig. S17). For example, Tryptophan metabolism and Valine, leucine and isoleucine metabolism pathways were identified in both the gene and metabolite module datasets. Groups of correlated features characterizing pre- and post-AIC samples were identified with mDIABLO (Suppl. Fig. S18). Interestingly, the Asthma pathway was identified despite individual gene members not being significantly altered post-AIC (Suppl. Fig. S19) and was negatively associated with Butanoate metabolism and positively associated with basophils, a hallmark cell-type in asthma (Suppl. Fig. S20).

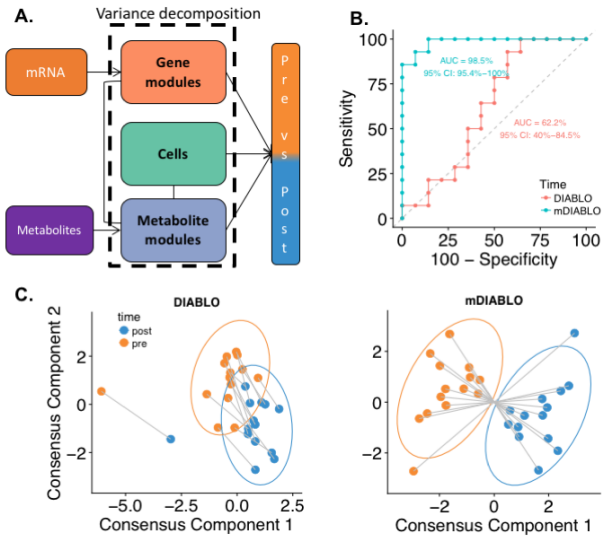


**Fig. 4.** Asthma study: cross-over design and module-based analysis. A) DIABLO design includes module-based decomposition to discriminate pre-and post-allergen challenge samples. B) Receiver operating characteristic curves comparing standard DIABLO and 'multilevel DIABLO' for repeated measures (mDIABLO) using leave-one-out CV. C) Component plots of the pre- and post-challenge samples (DIABLO and mDIABLO).

## 4 Discussion

DIABLO aims to identify coherent patterns between datasets that change with respect different phenotypes. This data-driven, holistic, and hypothesis-free tool can be used to derive robust biomarkers and, ultimately, improve our understanding of the molecular mechanisms that drive disease. We found that unsupervised methods identified features that formed strong interconnected multi-omics networks, but led to poor discriminative ability. In contrast, features identified by supervised methods were discriminative, but formed sparsely connected networks. The trade-off between correlation and discrimination is a fundamental challenge when trying to identify biologically relevant biomarkers that are also clinically relevant (Wang, 2011). DIABLO achieves this trade-off by incorporating a priori relationships between different omic datatypes to adequately model potential dysregulated processes between phenotypic groups. This may explain the superior biological enrichment of the DIABLO_full models in our benchmarking experiments. In contrast, biomarkers were different when we assumed no association between datasets with DIABLO_null and existing multi-step integrative strategies. Therefore, by controlling the trade-off between correlation and discrimination, DIABLO uncovered novel multi-omics biomarkers that have not previously been identified using existing integrative strategies. These novel biomarkers were part of densely connected clusters which have prior known biological associations, further suggesting their potential biological plausibility.

DIABLO assumes a linear relationship between the selected omics features to explain the phenotypic response, an assumption that may not apply in some biological research areas, for example when integrating distance-based metagenomics studies, where kernel approaches could be further explored (Mariette and Villa-Vialaneix, 2017). Selecting the optimal number of variables requires repeated CV to ensure unbiased classification error rate evaluation. A grid approach was deemed reasonable and provided very good performance results, but several iterations to refine the grid may be required depending on the complexity of the classification problem. The grid search algorithm is efficient (Rohart *et al.*, 2017a), but we advise using a broad filtering strategy to alleviate computational time when dealing with extremely large datasets (> 50,000 features each). DIABLO was primarily developed for omics-measurements on a continuous scale after normalization, and further developments are needed for categorical data types, such as genotype data. Finally, DIABLO, like other methods we benchmarked is likely to be affected by batch effects and presence of confounding variables. Therefore, we recommend exploratory analyses

be carried out in each single omics dataset to assess these effects prior to integration.

## Acknowledgements

We would like to thank Dr Kevin Chang (University of Auckland) and Dr Chao Liu (University of Queensland) for their help in the preliminary explorations of the TCGA datasets, and the reviewers for their constructive comments.

## Funding

This research was supported in part by the National Institute of Allergy and Infectious Diseases (U19AI118608: CPS/SJT) and the National Health and Medical Research Council (NHMRC) Career Development fellowship GNT1087415 (KALC).

## References

Aben, N., Vis, D. J., Michaut, M., and Wessels, L. F. (2016). Tandem: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, **32**(17), i413–20.

Allahyar, A. and De Ridder, J. (2015). Feral: network-based classifier with application to breast cancer outcome prediction. *Bioinformatics*, **31**(12), i311–9.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). MultiâŁomics factor analysisâŁ"a framework for unsupervised integration of multiâŁomics data sets. *Molecular Systems Biology*, **14**(6), e8124.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., and Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, **17**(2), S15.

Chung, I.-F., Chen, C.-Y., Su, S.-C., Li, C.-Y., Wu, K.-J., Wang, H.-W., and Cheng, W.-C. (2015). Driverdbv2: a database for human cancer driver gene research. *Nucleic acids research*, **44**(D1), D975–9.

Cun, Y. and Fröhlich, H. (2013). Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PLoS ONE*, **8**(9), e73074.

Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G.-C. (2013). Passing messages between biological networks to refine predicted interactions. *PLoS ONE*, **8**(5), e64832.

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.

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.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, **33**(suppl_1), D514–D517.

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, **8**, 84.

Kim, D., Li, R., Dudek, S. M., and Ritchie, M. D. (2013). Athena: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData mining*, **6**(1), 23.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**(24), 3290–3297.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, **9**(1), 559.

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.

Li, W., Zhang, S., Liu, C.-C., and Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**(19), 2458–2466.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell systems*, **1**(6), 417–425.

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.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, **7**(1), 523.

Ma, S., Ren, J., and Fenyö, D. (2016). Breast cancer prognostics using multi-omics data. *AMIA Summits on Translational Science Proceedings*, **2016**, 52.

Mariette, J. and Villa-Vialaneix, N. (2017). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, **34**(6), 1009–1015.

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*, **17**(4), 628–641.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, **16**(2), 85.

Rohart, F., Matigian, N., Eslami, A., S, B., and Lê Cao, K.-A. (2017a). Mint: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms.

Rohart, F., Gautier, B., Singh, A., and Le Cao, K.-A. (2017b). mixomics: an r package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, **13**(11).

Singh, A., Yamamoto, M., Kam, S. H., Ruan, J., Gauvreau, G. M., O'Byrne, P. M., FitzGerald, J. M., Schellenberg, R., Boulet, L.-P., Wojewodka, G., *et al.* (2013). Gene-metabolite expression in blood can discriminate allergen-induced isolated early from dual asthmatic responses. *PLoS ONE*, **8**(7), e67907.

Singh, A., Yamamoto, M., Ruan, J., Choi, J. Y., Gauvreau, G. M., Olek, S., Hoffmueller, U., Carlsten, C., FitzGerald, J. M., Boulet, L.-P., *et al.* (2014). Th17/treg ratio derived using dna methylation analysis is associated with the late phase asthmatic response. *Allergy, Asthma & Clinical Immunology*, **10**(1), 32.

Sokolov, A., Carlin, D. E., Paull, E. O., Baertsch, R., and Stuart, J. M. (2016). Pathway-based genomics prediction using generalized elastic net. *PLoS computational biology*, **12**(3), e1004790.

TCGA Research Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70.

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

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

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

van de Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N., and Wilting, S. M. (2016). Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine*, **35**(3), 368–381.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, **11**(3), 333.

Wang, T. J. (2011). Assessing the role of circulating, genetic, and imaging biomarkers in cardiovascular risk prediction. *Circulation*, **123**(5), 551–565.

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics*, **29**(5), 638–644.

Yugi, K., Kubota, H., Hatano, A., and Kuroda, S. (2016). Trans-omics: how to reconstruct biochemical networks across multiple âŁ˜omicâŁ™layers. *Trends in biotechnology*, **34**(4), 276–290.

Zeng, I. S. L. and Lumley, T. (2018). Review of statistical learning methods in integrated omics studies (an integrated information science). *Bioinformatics and Biology Insights*, **12**, 117793221875929.

Zhang, S., Li, Q., Liu, J., and Zhou, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, **27**(13), i401–i409.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, **40**(19), 9379–9391.

Zhu, J., Sova, P., Xu, Q., Dombek, K. M., Xu, E. Y., Vu, H., Tu, Z., Brem, R. B., Bumgarner, R. E., and Schadt, E. E. (2012). Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS biology*, **10**(4), e1001301.