OXFORD UNIVERSITY PRESS | Bioinformatics

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

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2018-1115.R1
Category:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Singh, Amrit; University of British Columbia, Centre for Heart Lung Innovation Shannon, Casey; UBC, Centre for Heart + Lung Innovation Gautier, Benoit; University of Queensland, The University of Queensland Diamantina Institute Rohart, Florian; The University of Queensland, Diamantina Institute Vacher, Michael; The University of Western Australia, 6Australian Research Council Centre of Excellence in Plant Energy Biology Tebbutt, Scott J; UBC, PROOF Centre of Excellence; UBC, Medicine, Division of Respiratory Medicine Lê Cao, Kim-Anh; University of Melbourne, Melbourne Integrative Genomics
Keywords:	Systems biology, biomarkers, Data integration, omics, asthma, Network analysis

Dr. Kim-Anh Lê Cao
Snr Lecturer, Statistical Genomics
School of Mathematics & Statistics
Melbourne Integrative Genomics
The University of Melbourne VIC 3010
T: +61 (0)3834 43971
@: kimanh.lecao@unimelb.edu.au

Nov 30, 2018

Dear Editor,

Please find attached a revision to our manuscript 'DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays' as a research article for the Systems Biology category in Bioinformatics.

We would like to thank the reviewers for their comments and we have substantially extended our simulation analyses and improved the clarity of the text. We have made extensive changes in the main manuscript and the supplemental material to address their comments.

The method is implemented in the open source R package mixOmics (now moved to Bioconductor), and our R scripts in R markdown format, along with detailed tutorials are available on our companion website http://www.mixOmics.org/mixDIABLO. We look forward to your reply.

Yours sincerely,

Dr. Kim-Anh LÊ CAO

Reviewer: 1

Comments to the Author

Summary

Singh, et al. present a new method, DIABLO, for supervised biomarker discovery of multiple 'omics datasets. More specifically, DIABLO is designed to overcome the computational challenge of identifying molecular features in different datasets predictive of a phenotypic response (e.g. cancer subtype). This is an important and difficult challenge given the scale of 'omics data and that it is increasingly common for researchers to take multiple types of molecular measurements (e.g. mRNA, miRNA, protein expression, …) per sample. Singh, et al. use a matrix factorization approach, specifically a generalized version of canonical correlation analysis to incorporate supervision in the form of phenotypic labels. They demonstrate DIABLO on simulated and real data, including a breast cancer and asthma dataset, and compare supervised/unsupervised and integrative/non-integrative approaches. The results on both simulated and real data are somewhat mixed.

Overall, the DIABLO method is novel and interesting, as are the applications and some of the analysis. However, despite these contributions, I recommend that the authors revise their manuscript for two main reasons. First, in multiple places the manuscript reads like a draft and requires major edits. Second, the analysis of the results of DIABLO on simulated and real data is incomplete. I elaborate on these and other points below.

Major comments

(1) In many places the manuscript reads like a draft and/or is missing key details, and also includes many typos. These include:

* Limited motivation for the new supervised approach. In particular, there is no substantive review of related work. As such, the authors' claim that existing "supervised strategies are unable to capture the shared information across multiple biological domains when identifying the key molecular drivers associated with a phenotype" (page 3) is unsupported.

We agree with the reviewer that compressive review of current knowledge and gaps would benefit the motivation of our approach. We had provided a summary figure in our Supplemental Material Fig S1, but we also have better detailed one paragraph in the introduction, which reads as:

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.

Note that most of the introduction section has been rewritten.

* Confusing presentation of the DIABLO algorithm. The authors should write out the DIABLO algorithm in full.
* Confusing presentation of the sGCCA algorithm. The notation for a_h^k is inconsistent, and more importantly, all a_h^k are completely missing from the objective of the optimization problem. Further, the authors do not review how sGCCA solves the optimization problem.

* Never defining sPLSDA
* Never defining N_new
* "validatio" —> "validation" (page 2)

The method section has been entirely rewritten. Section 1.1 presents sGCCA and Section 1.2 introduces DIABLO. All notations have been checked and made consistent all throughout. Acronyms have been defined and typos fixed.

(2) The results on simulated and real data are also concerning, particularly in the comparatively worse performance of DIABLO (full) at classification (full refers to the design matrix which controls which omics datasets are "connected").
* It is concerning that DIABLO (full) has the worst phenotypic classification performance on simulated data. The authors claim that there is a tradeoff between discrimination and correlation, and that DIABLO is better at selecting interpretable variables. This makes sense, but is unexplored and incomplete. The authors should extend their simulation analysis to show settings in which DIABLO (full) is at least as good as existing methods, and whether the design matrix can be used to achieve stronger classification performance even in the current simulated data setup.

We decided to thoroughly extend our simulation analyses to address these important questions. We agree with the reviewer that our discussion about the discrimination and correlation trade-off needed further exploration. Therefore, in our new simulation scheme we have further studied the relationship between the covariance between datasets, classification performance (error rate) and number of variables selected. Changes appear in the main document **subsection 3.1** Correlation and discrimination trade-of, **Suppl Section S1 and Suppl Fig S3**.

The updated simulation scheme detailed in Suppl S1 is depicted in the figure below, where:

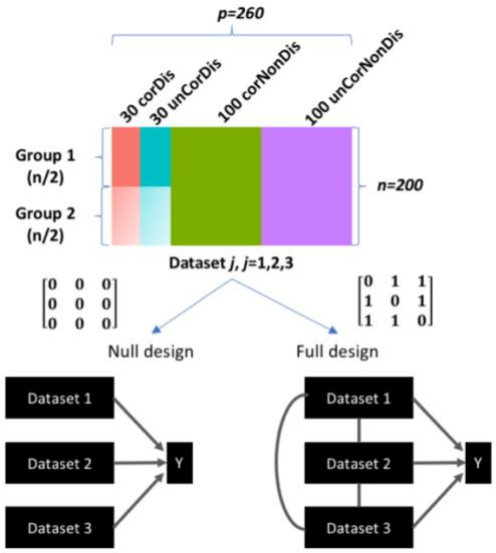


Figure. Simulated multi-omics data. *Each simulated dataset consisting of four types of variables: 30 correlated-discriminatory (corDis) variables, 30 uncorrelated-discriminatory (unCorDis) variables, 100 correlated-nondiscriminatory (corNonDis) variables and 100 uncorrelated-nondiscriminatory (unCorNonDis) variables*

Three datasets were simulated each with 200 observations (n) and 260 variables (p). The 200 observations were split equally over two groups (G1 and G2), whereas the 260 variables were generated by varying the covariance ($\sigma^2_{XY} = [0,5,10,15]$) between datasets and fold-change (δ=[0,1,2]) between G1 and G2: 30 correlated-discriminatory (corDis) variables, 30 uncorrelated-discriminatory (unCorDis) variables, 100 correlated-nondiscriminatory (corNonDis) variables, and 100 uncorrelated-nondiscriminatory (unCorNonDis) variables were simulated (see Figure 1A). The resulting dataset is of the form:

$$X_j = \left[X_j^{corDis} \mid X_j^{unCorDis} \mid X_j^{corNonDis} \mid X_j^{unCorNonDis} \right] + E_j, \text{ where j = 1, 2, 3}$$

The matrix containing correlated and discriminatory variables, X_j^{corDis} was generated using the following model:

$$X_j^{corDis} = \boldsymbol{u}_j^{corDis} \boldsymbol{w}_j^t, \text{ where } \|\boldsymbol{w}\| = 1, j = 1,2,3$$

where the loadings, $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ were 30-vectors, and the elements were drawn from a uniform distribution in the interval of [-0.3, 0.2] U [0.2, 0.3]. For G1, the outer components $\boldsymbol{u}_1^{corDis}$, $\boldsymbol{u}_2^{corDis}$, $\boldsymbol{u}_3^{corDis}$ were 3-vectors drawn from a multivariate normal distribution with a mean value of 0 and a mean value of δ=[0,1,2] for G2. The covariance

between pairs of components was set to cov($\boldsymbol{u}_i^{corDis}$, $\boldsymbol{u}_j^{corDis}$) = σ_{ij}^2 (for $i \neq j$) where σ_{ij}^2=[0,5,10,15] and cov($\boldsymbol{u}_i^{corDis}$, $\boldsymbol{u}_j^{corDis}$) =0 (for *i=j*).

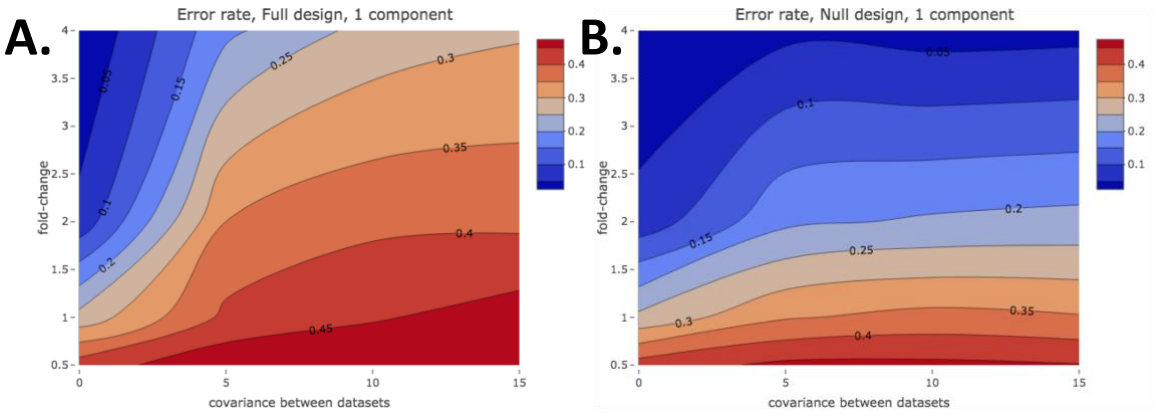
The matrix containing uncorrelated and discriminatory variables, $X_j^{unCorDis}$ was generated using the following model:

$$X_j^{unCorDis} = \boldsymbol{u}_j^{unCorDis}\boldsymbol{w}_j^t, \text{ where } \|\boldsymbol{w}\| = 1, j = 1,2,3$$

where the loadings, $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ were 30-vectors, and the elements were drawn from a uniform distribution in the interval of [-0.3, 0.2] U [0.2, 0.3]. For G1, the outer components $\boldsymbol{u}_1^{unCorDis}$, $\boldsymbol{u}_2^{unCorDis}$, $\boldsymbol{u}_3^{unCorDis}$ were 3-vectors drawn from a multivariate normal distribution with a mean value of 0 and a mean value of δ=[0,1,2] for G2. The covariance between pairs of components was set to cov($\boldsymbol{u}_i^{unCorDis}$, $\boldsymbol{u}_j^{unCorDis}$) = $\sigma_{ij} i \neq j \sigma_{ij} \boldsymbol{u}_i^{unCorDis}\boldsymbol{u}_j^{unCorDis}$0 for all *i* and *j*.

The nondiscriminatory variables (corNonDis and unCorNonDis) were generated by drawing 100-vectors each with 200 elements, from a multivariate normal distribution with a mean of 0. For correlated variables, the covariance between pairs of components was set to cov($\boldsymbol{u}_i^{corDis}$, $\boldsymbol{u}_j^{corDis}$) = σ_{ij}^2 (for $i \neq j$) where σ_{ij}=[0,5,10,15] and cov($\boldsymbol{u}_i^{corDis}$, $\boldsymbol{u}_j^{corDis}$) =0 (for *i=j*). For uncorrelated variables the covariance between pairs of components was set to cov($\boldsymbol{u}_i^{unCorDis}$, $\boldsymbol{u}_j^{unCorDis}$) = 0 for all *i* and *j*. $\boldsymbol{E}_j$ is a 200 x 260 residual matrix where each element is drawn from a normal distribution with zero mean and variance equal to 0.5.

For each simulated set of datasets with a given covariance and fold-change level, DIABLO models were constructed either with the null or full design and their performance was evaluated using 10-fold cross-validation. This procedure was repeated 20 times and the classification error rates were averaged.

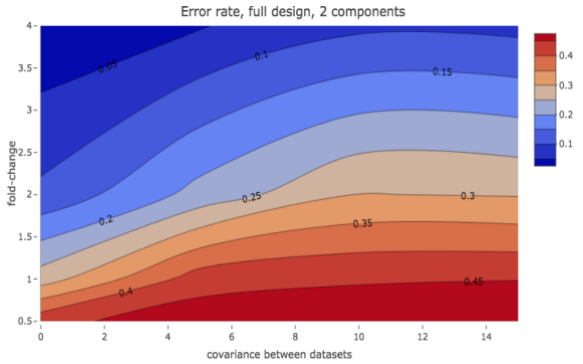


The figure above (**Suppl. Fig. S3**) depicts contour plots of the error rate for different degrees of covariance and fold-change (signal) either using the full or null design in the diablo models (selecting 60 variables on 1 comp). As can be observed in **A**, increasing the covariance between datasets leads to an increase in the classification error rate (blue to red) for a given fold-change. For the Null design in **B** however, the error rate is similar for given effect size, irrespective of the covariance between datasets.

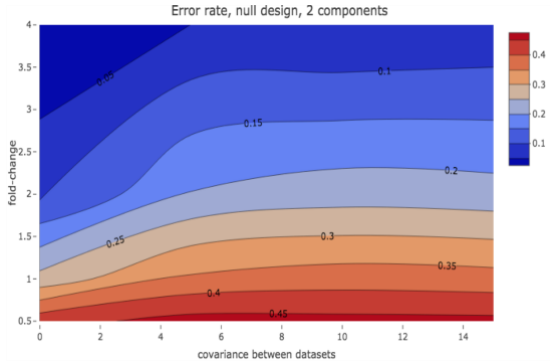
Our second step was then to study whether the addition of components improves the classification performance of DIABLO_full. The underlying assumption of DIABLO_full is that a common source of variance exists between datasets. We hypothesized that the high classification error rate was due to the fact that the model seeks for highly correlated components across datasets, however a good classification tasks requires uncorrelated information to be extracted. One way to add uncorrelated information in DIABLO_full is to consider the other dimensions of the model (i.e. the components built on the residual data after the deflation step).

The figure below panels **C** and **D** (**Suppl. Fig. S3**) show the contour plots of the classification error rate for different degrees of covariance and fold-change (signal) either using the full or null design as in **A** and **B** but when the components of the second dimension of DIABLO have been added. Since the sets of components in dimension 1 are orthogonal to the set of components in dimension 2, the variables selected on the second set of components are uncorrelated with those from the first component but still predictive of the outcome, resulting in a lower error rate.

C.



D.



To conclude on these simulation analyses, we have shown that DIABLO_full's performance can be affected by the covariance structure between datasets, and that the performance can be improved when adding orthogonal information (components) in the model.

* On simulated data, the authors only perform limited benchmarking against existing approaches, only comparing to sPLSDA, and do not provide an explanation for this missing analysis. There is more extensive benchmarking on real data.

The aim of the simulation study was mainly to evaluate the classification performance of the DIABLO designs and whether the methods were able to identify the correct variables. Therefore, we could only include the supervised methods (i.e. Concatenation and Ensemble scheme with sPLS-Discriminant Analysis).

* Some of the results on real data are not well-explained and/or do not have sufficient context.

Unfortunately we are lacking space in the main manuscript to fully describe the data. The information is presented in **Suppl Section S2**, along with the pre-processing steps.

* For example, there are quite substantial error rates for predicting PAM50 breast cancer subtypes (ranging from ~5-50%). How are these results to be interpreted? The PAM50 subtypes have known clinical implications, so if the authors find subtypes that are refined or different from PAM50, they should provide some sort of validation (e.g. with clinical data such as survival). Otherwise, what is the point of using supervision?

The purpose of the Breast Cancer analysis was to focus on the application of DIABLO to multiclass phenotypes and the biological relevance of the variables selected. We describe that some phenotypes are easier to separate (Basal, Her2) as compared to other subtypes (LumA, and LumB), as we depict using various illustration such as component plots, and heatmaps. It was not our intent to identify new subtypes, but rather to investigate whether there were multi-omics panels that could predict PAM50 subtypes. Therefore downstream validation with survival such as through analysis has not been considered. However, and to answer Reviewer 2 comments, we have added a section on classification performance as we have access to an independent test set in this study:

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. 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.

* The description of the "multilevel DIABLO" approach is confusing, and does not seem to be discussed in the Methods (though the authors say it is in the Results on page 23).

Due to the word limit restrictions, this section was moved to the supplementary materials and was not corrected in the main text. We apologize for this over sight. Additional details for this approach can be in **Suppl Section S7** and is stated as follow:

For multivariate analyses, A multilevel approach separates the within subject variation matrix (X_w) and the between subject variation (X_b) for a given dataset (X) (Westerhuis *et al.*, 2010; Liquet *et al.*, 2012), ie. $X = X_w + X_b$. In the case of a two-repeated measured problem (e.g. pre vs post challenge), the within subject variation matrix is similar to calculating the net difference for each individual between the data obtained for pre and post challenge. For each omics dataset, the within-subject variation matrix was extracted prior to applying DIABLO. In the asthma study, the multilevel approach (called variance decomposition step) was applied to the cell-type, gene and metabolite module datasets.

Minor comments

* The authors have integrated DIABLO into their mixOmics R package, and it seems well-documented.

Thank you!

* What is the runtime and memory footprint of DIABLO?

Computational footprint of DIABLO has been largely reduced since the first submission of this manuscript. Runtime and memory usage are reported in the following table on simulated data, comparing a single omics analysis using sPLS-DA and an integrative analysis using DIABLO in mixOmics, with a macbook pro 2013, 2.6GHz, 16Go Ram. The `tune` function is used to identify the number of variables to select from each dataset using cross-validation and grid of values for the number of components and number of features to select per component (the tune method is currently not implemented for NA values). The `model` is the final model run based on the tuned parameter. V6.3.2 is the current version in mixOmics in Bioconductor.

	Single 'omics sPLS-DA				*N*-integration DIABLO			
N	1000				1000			
P	20,000				10,000; 10,000			
NA	no		yes		no		yes	
	`model`	`tune`	`model`	`tune`	`model`	`tune`	`model`	`tune`
v6.1.1 (sec)	13.7	160	370	42 min	172	40 min	-	-
v6.3.2 (sec)	1.0	12	3.2	20 sec	2	18 sec	3.4	29
× faster	14	13	115	126	86	133	-	-
v6.1.1 (Go)	6.1	78	16.7	200	9.4	202	-	-
v6.3.2 (Go)	0.8	4.6	2.4	14.1	0.8	5.1	2.4	18.8
× better	8	17	7	14	12	40	-	-

Reviewer: 2

Comments to the Author

In their manuscript the authors present a method (DIABLO) to integrate data from multiple omics in a semi-supervised manner providing a balance between unsupervised methods that do not take into account known labels and supervised methods that do not take into account correspondence structures between omics. The method is based upon sGCCA (Tenenhaus et al al 2014) by including the labels as an additional block and implemented as part of the mixOmics package.

Overall, the method and the results are presented in a clear manner and the method seems to provide a good balance between supervised and unsupervised approaches. The authors demonstrate the ability of the method to find correlated discriminative features in simulations and convincingly show that the method is able to infer discriminative and biological meaningful components in several applications. More care could be taken when discussing the relationship of the proposed methods to existing approaches and in evaluating its predictive performance.

Major comments:

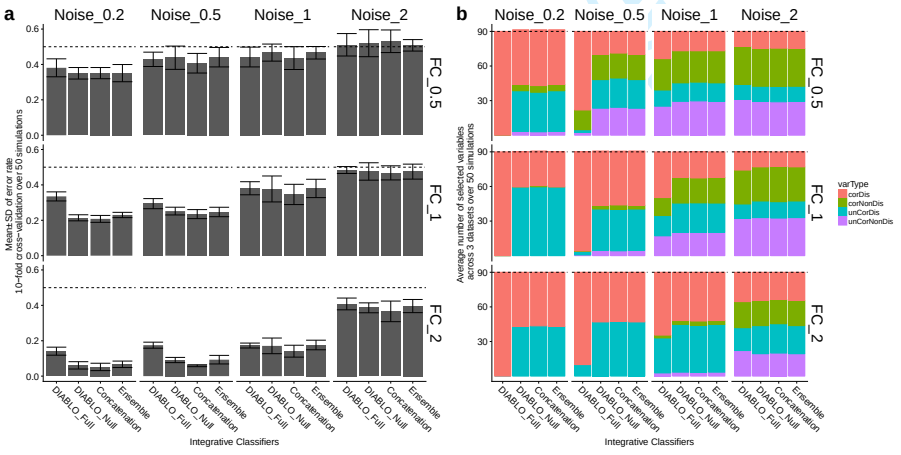
1. In the introduction the authors comment on supervised and unsupervised methods, however they do not relate their method to existing methods that aim at partly supervised integration of multiple data types such as for example sparse Multi-Block Partial Least Squares or sparse supervised CCA. This relationship and the contributions should be discussed more carefully.

This comment re-joins the comment from reviewer 1, see our answer in point **(1).**

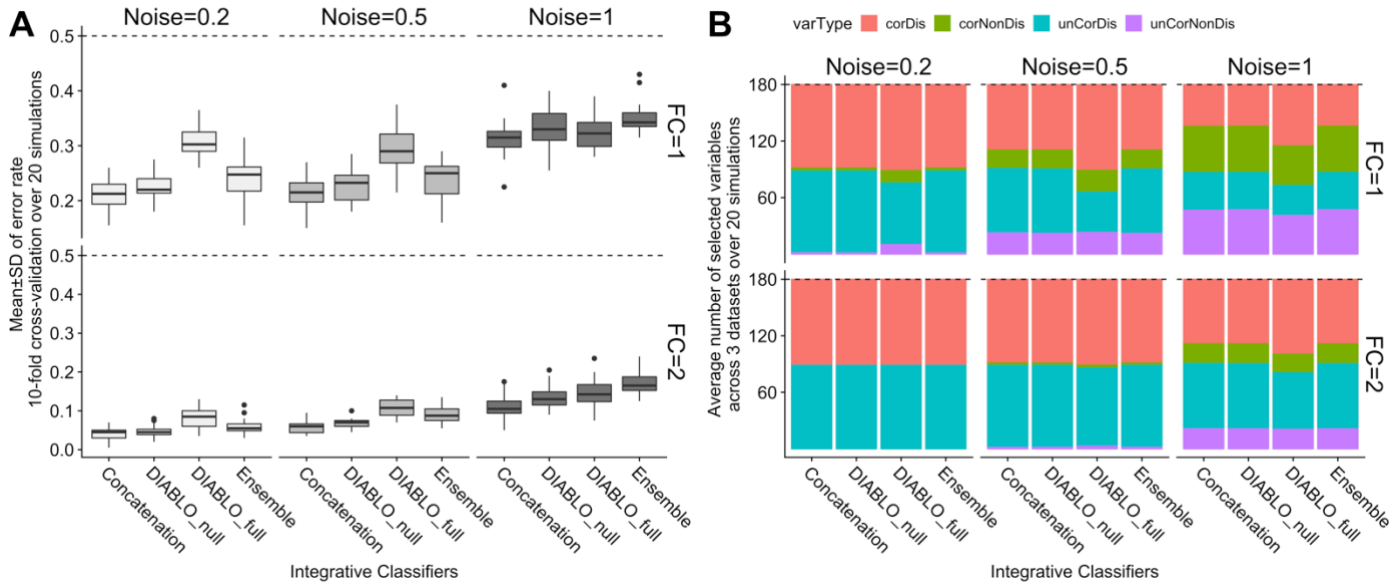
2. The authors convincingly demonstrate that the method is very good at finding biological meaningful components that well discriminate phenotypic groups. In terms of predictive performance there seems to be a risk, when concentrating on correlations between data sets, that DIABLO (with full or partly full design matrix) could overlook single strongly discriminative features in a data set when these have little correlation to other omic data sets. For example, in the simulation study DIABLO_Full mainly discovers correlated discriminative features. Would the method be able to discover all 180 discriminative features if 60 instead of 30 variables were selected from each data set?

In the previous analysis we had simulated 60 predictive variables per dataset (30 correlated and 30 uncorrelated), however as the reviewer suggests we can also allow for all methods to select 60 variables per dataset (180 in total) in order to determine whether all 180 discriminative features were selected. **Figure 1** in the manuscript has been updated and now includes the selection of 180 variables. The text in **section 3.1** has also been updated accordingly.

Before:

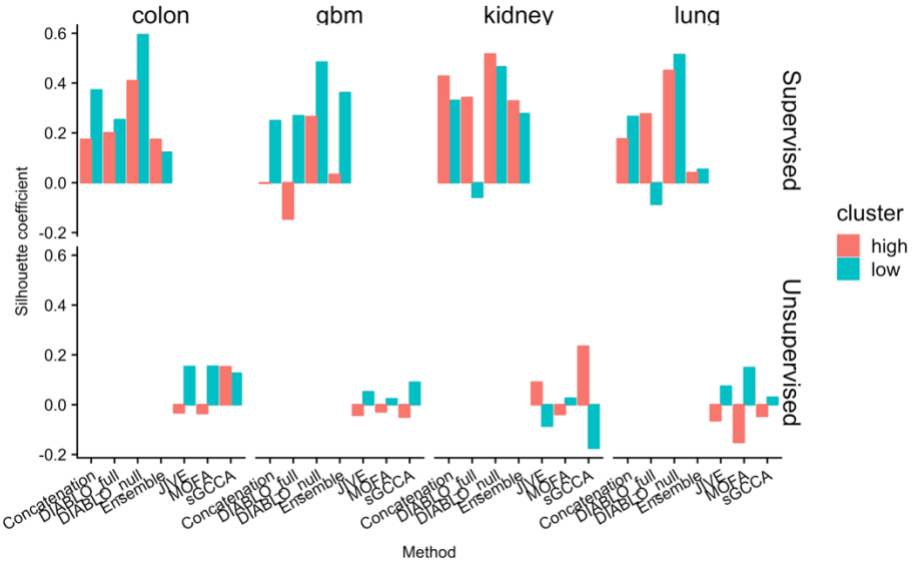


After:



In addition, it would also be good to see a method comparison in terms of classification performance on real data (e.g. on the benchmark data sets by Wang et al 2014) using independent test sets.

The purpose of the benchmarking analyses was to compare the types of variables selected across different datasets and integrative methods. However, since the methods differed with respect to variable selection, and level of supervision and cohorts lacked independent test datasets, a comparison with respect classification performance was neglected. In the revised analyses we have performed internal validation to assess for consistency within the phenotypic groups in the benchmarking experiments using the silhouette coefficient Since all integrative methods that have been included in the benchmarking experiments are component-based methods, the first two principal components were used to compute the average silhouette coefficient per dataset, per group for all methods. **Suppl. Fig S10** has been added as:



Supplementary Figure S10. Internal validation of high and low phenotypic groups for all method in the benchmarking experiments.

The silhouette for each data i, was computed as the normalized difference between two average distances (a_i and b_i), where a_i is the average distance between i and all points within its own cluster and bi is the average distance between i and all points that are not in its cluster ($s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$). The silhouette ranges from -1 to 1, 1 being a strong indicator of cluster membership and -1 being a weak indicator of cluster membership. As can be observed, the supervised methods show stronger silhouette coefficients

as compared to unsupervised methods. This is because the principal components are associated with the phenotype of interest. DIABLO_Null consistently out-performed the methods with a higher average silhouette coefficient with respect to both phenotypic groups (high and low survival). The silhouette coefficients for the other methods were variable, however, whether this translates to a lower predictive performance in independent test data remains to be observed.

In addition, we added a classification performance comparison on the Breast Cancer PAM50 subtypes, that includes independent test sets (610 samples in the test datasets: mRNA, miRNA, and CpGs). Given the limitation of the Concatenation scheme where all variables must be present in both training and test datasets, we removed the proteins dataset for this comparative analysis. The summary of the results is presented in **Suppl. Table S2** (see below) and we added a new **section 3.4**:

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. 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.

Supplementary Table S2. Classification error rates [average error (sd)] of DIABLO, Concatenation-based and Ensemble-based sPLSDA and Elastic Net (enet) classifiers on the Breast Cancer study (see Suppl. Section S5 for details).

Dataset	*p*	Train	Test
Diablo_null	mRNA: 60 miRNA: 42 CpGs: 22	0.21 (0.0091)	0.19
Diablo_full	mRNA: 55 miRNA: 17 CpGs: 17	0.22 (0.0057)	0.21
Concatenation_sPLSDA	mRNA: 60 miRNA: 0 CpGs: 0	0.15 (0.013)	0.18
Concatenation_enet	mRNA: 38 miRNA: 2 CpGs: 118	0.14 (0.0072)	0.20
Ensemble_sPLSDA	mRNA: 60 miRNA: 55 CpGs: 40	0.25 (0.014)	0.28
Ensemble_enet	mRNA: 96 miRNA: 45 CpGs: 127	0.11 (0.0016)	0.23

3. To find sparse solutions the method requires the users to choose the number active variables per dataset. However, it is unclear how users should make an informed decision on this quantity, as an exhaustive grid search can be very expensive. How sensitive is the method to this choice (which possibly could lead to strong over- or under-fitting)?

The sparse methods implemented in the mixOmics R-library, including DIABLO use soft-thresholding to replace the ℓ_1 penalty by the number of variables to select from each component. This improves the usability by the user who can determine a suitable grid for their purposes. A smaller classification model may be favored if the features will be follow-up via candidate experiments or validation studies, where large model are useful for perform gene-set enrichment analyses, as more clearly specified in the **Methods section.**

2.3 Parameters tuning

- 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.

The grid search is also mentioned in the Discussion section:

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).

We provide a tune function in mixOmics that performs parameter tuning using a repeated and stratified cross-validation (Rohart *et al.*, 2017). In regards to over and under-fitting, this can be assessed using cross-validation which provides an estimate of the generalizable test error.

Minor comments:

1. In 'parameter tuning' it is unclear what is meant by 'first component' in l.27 p. 9. Which design matrix is used to calculate this component?

To choose an appropriate design matrix, we suggest the user to first consider the pairwise correlation between components obtained from a PLS model. PLS only applies for the integration of 2 data sets and there is no design matrix needed for this method. Once the first set of components is extracted from PLS, the correlation between them will indicate the degree of correlation and will inform the design matrix for DIABLO.

2. The description of visualisation outputs on p. 10 would profit from an illustration in a supplementary figure or including pointers to a corresponding figure in subsequent analyses.

We thank the review for this suggestion, we have referred to exemplar figures in this section:

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).

3. alpha is missing in the objective of equation in p.6, l.7 and has inconsistent sub/superscripts in the equation

All notations were amended in the Methods section (we do not need an \alpha coefficient in Eq. (1)).

4. The description on p. 7 uses different notation and naming for loadings/coefficients vectors than on p.6 and differs

again from the description in “Prediction distances”. In general, it would be helpful to make this more consistent and avoid duplicated descriptions if possible on these two pages.

We have fully amended our notations and rewritten the Methods section.

5. The methods MOFA and JIVE have missing or malformatted citations in the text on p.14

This has been corrected in the revised draft.

6. Typo on p.2 l.31: validation

This has been corrected in the revised draft.

7. Why is the set size different in Fig. 2a between methods? To my understanding the same number of features were used from each method.

For each method, 2 sets of components were retained and 30 variables were selected for each dataset, resulting in 30 variables x 2 components x 3 datasets = 180 variables per method. Although the first and second components are orthogonal, there might be some overlap between variables selected on both components. The set size depicts the number of unique features and thus leads to the discrepancy observed by the reviewer. We describe this occurrence in Supplementary Figure S4 of the revised manuscript draft.

8. The message of Fig. 2c (upper panel) is unclear. Do the two large clusters correspond to the two components? What do the grey lines represent? A description thereof should be included into the caption.

Each network depicted the multi-omics biomarker panel (mRNA, miRNA and CpGs) identified using the different methods. The gray circles depict modules based on the edge betweenness index from the igraph R-library. For the colon cancer dataset we observed modules (clusters circled by gray lines) that included features that selected on the two components. However, this was not true for all the other cancers datasets. The caption for Figure 2 has been amended as:

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

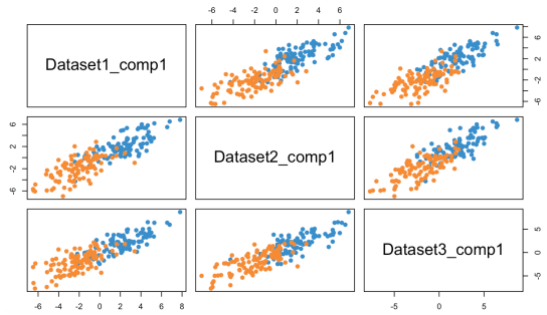
9. In the supplement the grid parameters for simulation are inconsistent within the text and with the Figure 1a.

This has been corrected in the revised draft of the manuscript.

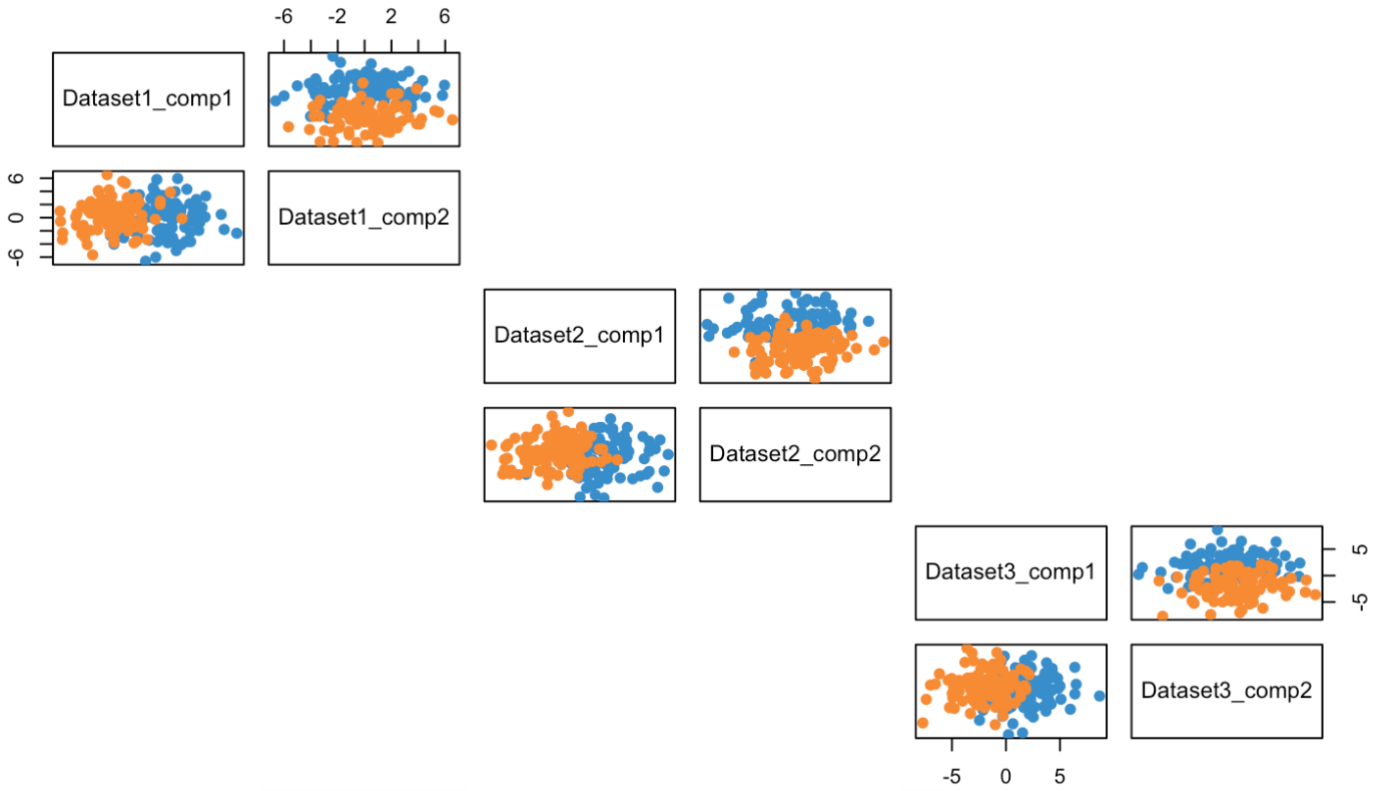
10. The correlations in Fig. S2 for the uncorrelated simulation setting seem to me still rather high. Could the authors comment on this?

Th figure was depicting an extreme case where the level of discrimination (i.e. fold-change) and correlation was very high. The fold-change is simulated as the difference between the centroids of the two groups, resulting in a high level of discrimination between groups and therefore a high degree of correlation between components of different datasets. Our scatterplot matrices created confusion as they seem to depict that the classification occurs using the components of different blocks (see below) when instead it occurs separately for each omic-type and the predictions are combined using a voting scheme (average, majority, or weighted majority, as described in the Methods section 2.2).

Previous figure: association between component 1 of different datasets. However, this plot may be misinterpreted by the reader as a depiction of the classification boundary used to discriminate the two phenotypic groups.



Correct figure (below) component plots for each dataset and each component. Given the orthogonality of components, each added component brings uncorrelated information that explains the variation in the response.



Given this confusion we have decided to remove this Supplemental figure entirely for the revised manuscript draft to avoid confusion.

11. Fig. 3a (names of proteins) and Fig. 4f are illegible

We have revised all figures to improve readability and have remove such difficult to read panels to the Supplementary Materials.

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

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 ℓ_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 ℓ_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 ℓ_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,

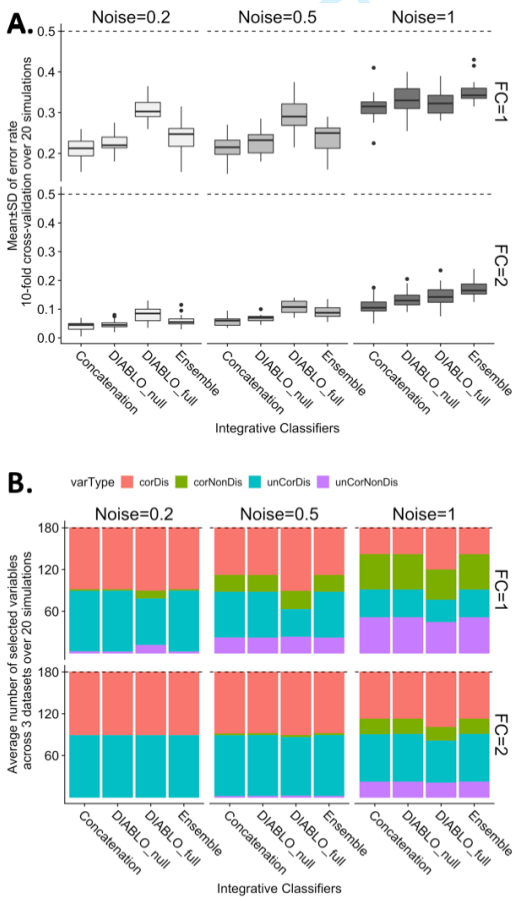


Fig. 1. Simulation study. A) Classification error rates (10-fold CV averaged over 20 simulations) for different fold-changes (FC) between groups and varying level of noise (sd). Dashed line indicates a random performance (error rate = 50%). B) Types of variables selected by the different classification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

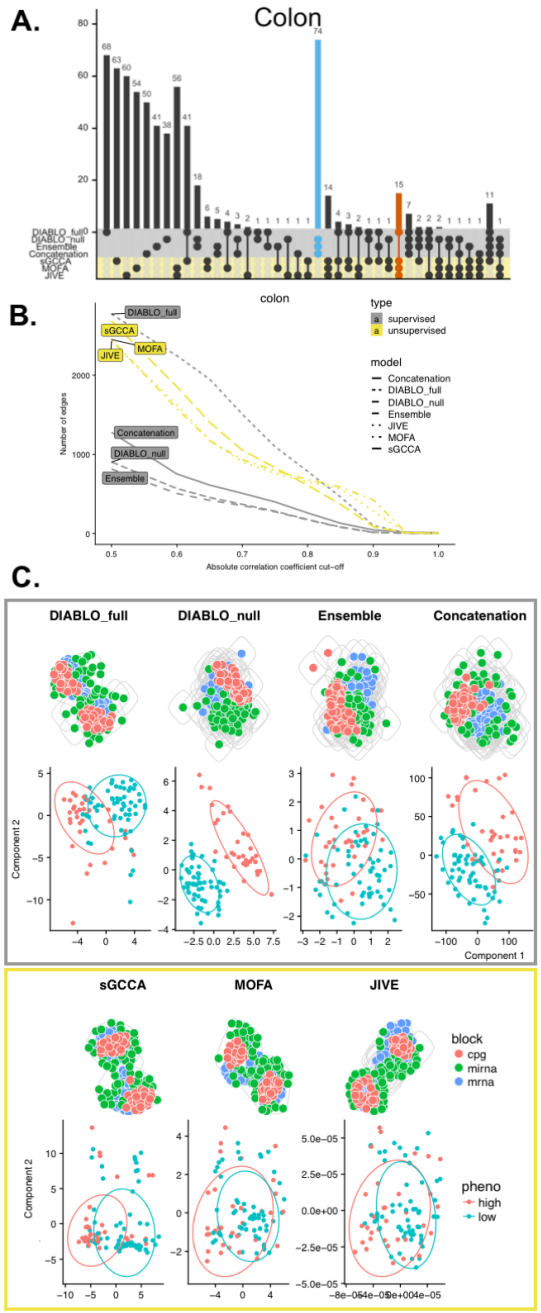


Fig. 2. Benchmark for colon cancer. A) Overlap of features selected by supervised or unsupervised methods. B) Number of correlated variables in the biomarker panels for various Pearson correlation cut-offs. C) Top: network modularity of each multi-omic biomarker panel. Gray circles depict modules based 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±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.

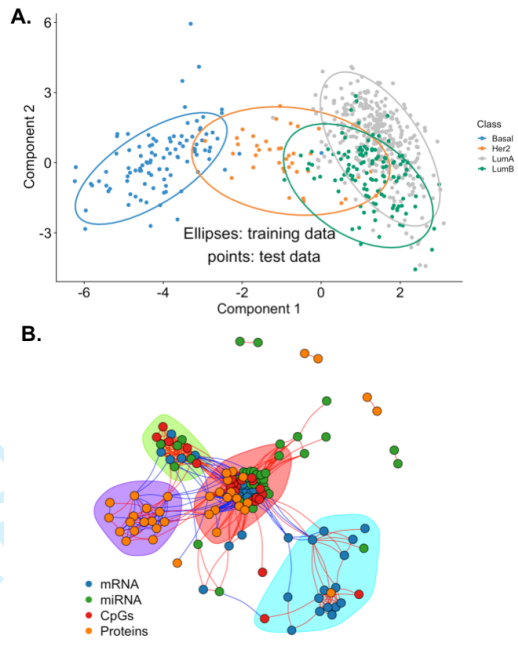


Fig. 3. A Multi-omics biomarker panel predictive of breast cancer subtypes. A) DIABLO consensus component plot based on the identified multi-omics biomarker panel: test samples are overlaid with 95% confidence ellipses calculated from the training data. B) Network visualization of the biomarker panel 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).

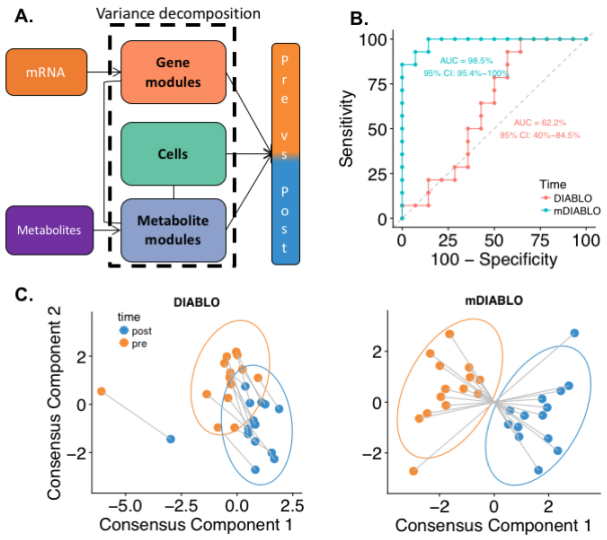


Fig. 4. Asthma study: cross-over design and module-based analysis. A) DIABLO design includes module-based decomposition to discriminate pre-and post-allergen challenge samples. B) Receiver operating characteristic curves comparing standard DIABLO and 'multilevel DIABLO' for repeated measures (mDIABLO) 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.

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

Subject Section

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

Amrit Singh [1,2,3] [1], Casey P. Shannon [3] [1], Benoît Gautier [4] [2], Florian Rohart [5] [3], Michaël Vacher [6,9] [4], Scott J. Tebbutt [1,3,7] [1] and Kim-Anh Lê Cao [8,*] [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.,~~ *e.g.* gene expression, DNA methylation status, protein abundance) resulting in multiple omics (multi-omics) data from the same set of biospecimens ~~(eg.,~~ or individuals (*eg.* transcriptomics, proteomics, metabolomics). ~~The large number of omic variables compared to the limited number of available biological samples presents a computational challenge when identifying the key drivers of disease. Further, technological limitations differ with respect to different omic platforms (e.g., sequencing vs. mass spectrometry), and biological effect sizes differ with respect to different omic variable-types (e.g., methylation status vs. protein expression). Effective integrative strategies are needed, to extract common biological information spanning multiple molecular compartments that explain phenotypic variation. Already, systems biology approacheswhich incorporated~~ Systems biology approaches, by incorporating data from multiple biological compartments, ~~have shown~~ provide improved biological insights compared to traditional single omics analyses ~~(Zhu *et al.*, 2012; Kim *et al.*, 2013; Wang *et al.*, 2014). This may be because single omics analyses cannot account for the interactions between omic layers and, consequently, are unable to reconstruct~~ (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).

~~Preliminary approaches to data integration included~~

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 ~~leveraged existing single-omics methods: multi-omics data were concatenated, or ensembles of single omics models created~~ 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, ~~however,~~ and do not account for interactions between omic layers ~~(Aben *et al.*, 2016)~~(Aben *et al.*, 2016; Ma *et al.*, 2016). Recently, ~~more sophisticated integrative approaches have been proposed (Suppl. Figure S1) (Ritchie *et al.*, 2015; Bersanelli *et al.*, 2016; Meng *et al.*, 2016; Huang *et al.*, 2017 . They can be broadly divided into unsupervised analyses, which identify coherent relationships across~~ 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 ~~datasets when samples are unlabeled, and supervised analyses, which identify multi-omics patterns that discriminate between known phenotypic sample groups. However, these supervised strategies are unable to capture the shared information across multiple biological domains when identifying the key molecular drivers associated with a phenotype. Such methods are needed to capture the dynamic nature of molecular networks under various disease conditions and ultimately provide robust biomarkers that are both biologically and clinically relevant~~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.

~~To address these knowledge gaps, we~~

We introduce DIABLO, a ~~method that incorporates information across high dimensional multi-omics data while discriminating~~ multi-omics datasets (mRNA, miRNA, CpGs, proteins, metabolites, *etc.*) during the integration process and discriminates phenotypic groups. DIABLO (**D**ata **I**ntegration **A**nalysis for **B**iomarker discovery using **L**atent c**O**mponents) maximizes the common or correlated information between multiple omics ~~(multi-omics) datasetswhile identifying the key omics variables (mRNA, miRNA, CpGs, proteins, metabolites, etc.) and characterizing the disease sub-groups or phenotypes of interest. DIABLO is , therefore, an~~ datasets. It is the first multivariate integrative classification method ~~that builds predictive multi-omics models that can be applied to multi-omics data from new samplesto determine their phenotype. Users can specify the number of variables to select from each dataset and visualize the omics data and the multi-omics panel into a reduced data space.~~ 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~~, as illustrated in one of our case studies.~~ . We demonstrate the capabilities and versatility of DIABLO below, both in simulated and ~~real-world data, integrating~~ real multi-omics ~~datasets~~ studies to identify relevant biomarkers of various diseases. ~~DIABLO is available through the mixOmics data integration toolkit, (Rohart *et al.*, 2017a) which contains a wide range of multivariate methods for the exploration and integration of high dimensional biological datasets.~~

2 Methods

2.1 ~~Statistical methods and analysis~~General multivariate integrative framework.

~~*General multivariate framework to integrate multiple datasets measured on the same samples.*~~ DIABLO extends sparse generalized canonical correlation analysis (sGCCA) ~~(Tenenhaus *et al.*, 2014)~~ , Tenenhaus *et al.* 2014) to a classification ~~(supervised)~~ 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~~in a computationally and statistically efficient manner~~. 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 ~~in sGCCA~~ with ℓ_1 penalization on the variable coefficient vector defining the linear combinations. ~~Note that since~~ 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 ~~K~~ Q normalized, centered and scaled datasets ~~X_1 $(n \times p_1), \ldots, X_K$ $(n \times p_K)$,~~ $X^{(1)}(N \times P_1)$, $X^{(2)}(N \times P_2)$, ..., $X^{(Q)}(N \times P_Q)$ measuring the expression levels of ~~$p_1, p_2, \ldots, p_K$~~

$P_1, \ldots, P_Q$ 'omics variables on the same ~~n samples, $k = 1, \ldots, K$~~ N samples. sGCCA solves the optimization function for each component $h = 1, \ldots, H$:

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

~~where $c_{k,j}$ indicates whether to maximize the covariance between the datasets X_j^h and X_k^h according to~~ where $a_k^{(q)}$ is the variable coefficient or loading vector on component h associated to the residual matrix $X_k^{(q)}$ 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), ~~with $c_{k,j}$ values ranging from 0 (no relationship modelled between the datasets)to 1 otherwise, a_k^h is the variable coefficient vector for each dataset X_k^h, λ_k is~~ $\lambda^{(q)}$ is a non-negative parameter that controls the amount of shrinkage and thus the number of non-zero coefficients in ~~a_k^h~~ $a_k^{(q)}$. Similar to Lasso (Tibshirani, 1996) and other ℓ_1 penalized multivariate models developed for single omics analysis ~~(~~(Lê Cao *et al.*, 2011), the ~~ℓ_1 penalization improves the interpretability of the component scores $X_k^h a_k^h$ that is now only defined on~~ penalization enables the selection of a subset of variables with non-zero coefficients ~~in X_k^h~~ that define each component score $t_k^{(q)} = X_k^{(q)} a_k^{(q)}$. The result is the identification of variables that are highly correlated between and within omics datasets. ~~Equation~~

The sGCCA model (1) ~~describes the sGCCA model for the first dimension. Once the first~~ is iterative; a first set of coefficient vectors ~~a_k^1 and associated component scores $t_k^1 = X_k^1 a_k^1$ are obtained, residual matrices are calculated during the 'deflation' step for the second dimension, such that $X_k^2 = X_k^1 - t_k^1 a_k^1$, where X_k^1 is the original centered and scaled data matrix. The subsequent set of components scores and coefficient vectors are then obtained by substituting X_k $(a_1^{(1)}, \ldots a_1^{(Q)})$ is obtained by X_k^2~~ ~~in~~ 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 ~~the sGCCA model~~ sGCCA is that the major source of common biological variation can be extracted via the first sets of component scores ~~$t_k^1, \ldots, t_k^h$~~ $t_1^{(q)}, \ldots, t_k^{(Q)}$, while any unwanted variation due to heterogeneity across the datasets ~~XK~~ $X^{(q)}$ does not impact the statistical model. The optimization problem (1) is solved using a monotonically convergent algorithm (Tenenhaus *et al.*, 2014).

2.2 DIABLO~~for~~ : supervised analysis and prediction

~~*Supervised Analyses.*~~ To extend sGCCA for a classification framework, we substitute one omics dataset ~~X_k in 1)~~ $X^{(q)}$ in (1) with a dummy indicator matrix ~~Y of size (n x G)~~ Y $(N \times G)$ to indicate the class membership of each sample, where G is the number of phenotype groups~~that indicate the class membership of each sample. In addition, and for~~ . For easier use of ~~the method~~DIABLO, we replaced the ~~l1 penalty parameter λ_k~~ ℓ_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. ~~A separate classification model can then be built for each omic dataset, k:~~

$$Y_k^{\text{new}} = X_k^{\text{new}} W_k (D_k^T W_k)^{(-1)} B_k = T_{\text{pred}} B_k$$

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):

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

~~The columns of W_k are the loadings vectors (computed using sGCCA), whereas D_k and B_k consist of regression coefficients computed by regressing X_k and Y on the H latent components of X_k (also computed using SGCCA) separately. Each matrix Y_k^{new} is of size $N_{\text{new}} \times G$, and consists of the predictions of each new sample for each class g. The matrix T_{pred} consists of the predicted scores or predicted latent components of the new samples and is of size $N_{\text{new}} \times H$~~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 2.4). 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 $(t_{1,new}^{(q)}, \ldots, t_{H,new}^{(q)})$ can be calculated for each type of omics q by using the estimated loadings vectors $a^{(q)}$ from DIABLO.

~~*Prediction distances.* Denote a new sample i which is measured across the different types of omics datasets x_k^i, its class membership is predicted by the fitted sGCCA model with the estimated variable coefficients vectors $\hat{a}^k$ to obtain the predicted scores $\hat{t}^{(k,i)} = x_k^i \hat{a}^k$, $k = 1, \ldots, K$. Therefore, to each dataset k corresponds a predicted continuous score $t^{(k,i)}$.~~ The predicted class of ~~sample i~~ a new sample for each dataset is obtained from the predicted score using one of the distances Maximum, Centroids or Mahalanobis ~~(Lê Cao *et al.*, 2009) as described in Rohart *et al.* 2017b~~as detailed in Rohart *et al.* (2017b), which results in Q class memberships for a new sample.

~~*Consensus class prediction for each new sample.* The~~ 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 ~~$t^{(k,i)}$ across all K datasets before using the prediction distance of choice ('average prediction' scheme)~~$t_{k,new}^{(q)}$ 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~~ Centroid distance for the weighted majority vote scheme (~~breast cancer~~ Breast Cancer study) and the ~~maximum~~ Maximum distance for the average vote scheme (~~asthma~~ Asthma study) as those led to best performance (see Rohart *et al.* 2017b for details about distance measures and ~~voting schemes that can be used~~proposed voting schemes).

2.3 ~~Design matrix~~Parameters tuning

~~The design matrix C is a $(K \times K)$ matrix with values ranging from 0 to 1 which specifies whether the covariance between two datasets should be maximized DIABLO, 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):~~

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

~~However, every dataset is connected to the outcome Y internally in the method. For the two case studies (breast cancer and asthma) the design matrix was chosen based on our proposed method (see Parameters tuning). Note that 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).~~

2.4 ~~Input data~~

~~While DIABLOdoes not assume particular data distributions, all datasets should be normalized appropriately according to each omics platform and preprocessed if necessary (see normalization steps described below for each case study). Samples should be represented in rows in the data matrices and match the same sample across omics datasets. The phenotype outcome Y is a factor indicating the class membership of each sample. The R function in mixOmics will internally center and scale each variable as is conventionally performed in PLS-based models and will create the dummy matrix outcome from Y. A multilevel variance decomposition option is available for repeated measures study designs.~~

2.4 ~~Parameters tuning~~

~~The first parameter to tune is the design matrix C, which~~ 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 ~~uses PLS method implemented in mixOmics~~ can use PLS that models pairwise 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.~~ *e.g.* 0.8) then a connection between those datasets is included in ~~the DIABLO design~~ C as a 1 value. ~~The second parameter to tune is the total~~
- The number of components~~. In~~ : in several analyses we found that ~~G â'' 1 components were sufficient to~~ $G - 1$ 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~~(described below) as well as using~~ , as described below, and can be aided with graphical outputs such as sample plots to visualize the discriminatory ability of each component. ~~Finally, the third set of parameters to tune is the~~
- 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 ~~it does not substantially change~~ we did not observe substantial changes in the classification performance ~~. This is because of the use of regularization constraints which reduces the variability in the variable coefficients and thus maintains the predictive ability of the model. Further, the variable~~ during our case study analyses. The variable selection size can also be guided according to the downstream biological interpretation~~to be performed~~. For example, a gene-set enrichment analysis may require a larger set of features than a literature-search interpretation.

2.4 ~~Visualization~~ DIABLO visualisation outputs

To facilitate the interpretation of the integrative analysis, several types of graphical outputs were proposed and implemented in `mixOmics`.

~~*Sample plots.* The~~
Sample plots include a consensus plot which depicts the samples ~~is computed~~ by calculating the average of the components from each dataset ~~. Omics specific samples~~ (Fig 3A). Omic-specific sample plots can also be obtained by plotting components associated to each ~~data set. The scatterplot matrix represents the correlation between components for the same dimension but across all omics datasets.~~

~~*Variable plots.* We proposed a circos plot to represent~~ 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~~at the variable level~~. The association between variables is computed using a similarity score that is analogous to a Pearson correlation coefficient ~~, as previously described in (González *et al.*, 2012). The association between variables~~ (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 ~~circos~~ plot, with side colors indicating each omics type, optional line plots represent the expression levels in each phenotypic group ~~.~~

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

~~*Clustered Image Map (CIM).* A clustered image map (González *et al.*, 2012)~~ (see Suppl. Fig. S20).
Clustered Image Maps (CIM) based on the Euclidean distance and the complete linkage ~~displays~~ 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).

3 Results

3.1 ~~DIABLO selects correlated~~ Correlation and ~~discriminatory variables~~discrimination trade-off

~~Briefly, three~~ Three omic datasets consisting of 200 samples (~~split equally over two~~ 100 in each of the two phenotypic groups) and 260 variables were ~~generated by modifying the degree of correlation and discrimination, resulting in~~ simulated (details in Suppl. Section S1). Each dataset included four types of variables: 30 correlated-discriminatory (~~corDis) variables~~*corDis*), 30 uncorrelated-discriminatory (~~unCorDis) variables~~*unCorDis*), 100 correlated-nondiscriminatory (~~corNonDis) variables,~~ *corNonDis*) and 100 uncorrelated-nondiscriminatory (~~unCorNonDis) variables (Suppl. Section S1 and Suppl. Figure S2). Three integrative classification methods were applied to generate multi-omic biomarkers panels of 90 variables each (30 variables from each omic dataset): a DIABLO model~~ *unCorNonDis*) variables. DIABLO models with either a null or full design (~~where the correlation between all pairwise combinations of datasets, as well as between each dataset and the phenotypic outcome, were maximised) or the null design (where only the correlation between each dataset and the phenotypic outcome was maximised, Methods) , a concatenation-based sPLSDA classifier which consists of naively combining all datasets into one, and an ensemble of sPLSDA classifiers where a separate sPLSDA classifier was fitted for each omics dataset and the consensus predictions were combined using a majority vote scheme (Suppl. Figure S3~~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 ~~purpose of the simulation study was to compare DIABLO models with existing multi-step integrative classifiers with respect to~~ 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 ~~and types of variables selected as part of the multi-omic biomarker panels. A secondary aim was to determine the effect of design matrix on the resulting multi-omic biomarker panels identified using DIABLO.~~

~~**Simulation study: performance assessment and benchmarking**. Simulated datasets included different types of variables: correlated & discriminatory (corDis); uncorrelated & discriminatory (unCorDis); correlated & nondiscriminatory (corNonDis) and uncorrelated & nondiscriminatory (unCorNonDis) for different fold-changes between sample groups and different noise levels (Suppl. Section S1). Integrative classifiers included DIABLO with either the full or null design, concatenation and ensemble-based sPLSDA classifiers and were all trained to select 90 variables across three multi-omics datasets. a) Classification error rates (10-fold cross-validation averaged over 50 simulations). Dashed line indicates a random performance (error rate = 50%). All methods perform similarly with the exception of DIABLO_full which has a higher error rate. b) Number of variables selected according to their type. DIABLO_full selected mainly variables that were correlated & discriminatory (corDis, red), whereas the other methods selected an equal number of correlated or uncorrelated discriminatory variables (corDis and unCorDis, red and blue).~~

~~The concatenation, ensemble and DIABLO_null classifiers performed similarlyacross the various noise and fold-change thresholds. At lower noise levels(simulated using a multivariate normal distribution with mean of zero and standard deviation of 0.2 or 0.5) the~~ was ~ 50% for all methods regardless of noise level (Suppl. Figure S2). When FC = 1, DIABLO_~~full classifier had a slightly higher error rate compared to the other approaches (Fig. 1a), but consistently selected mostly correlated and discriminatory (corDis) variables, unlike the other integrative classifiers (Fig. 1b). All methods behaved similarly with respect to the error rate and types of variables selected at higher noise thresholds (simulated using a multivariate normal distribution with mean of zero and standard deviation of~~ full yielded a higher error rate than all other methods, for noise < 1. However, when noise = 1 ~~or~~ and FC = 1, all methods performed similarly. Finally, when FC = 2 ~~). This simulation highlights how the~~

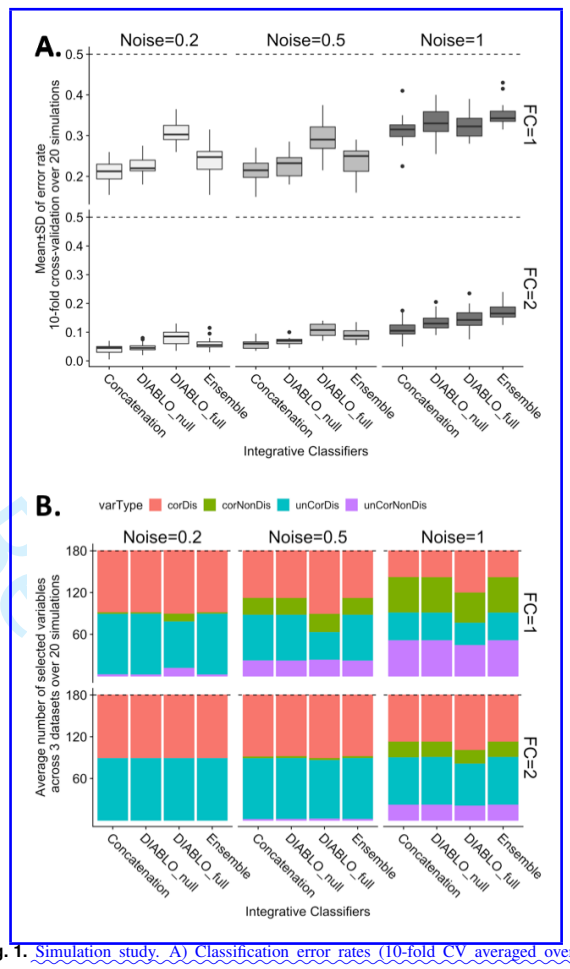


Fig. 1. Simulation study. A) Classification error rates (10-fold CV averaged over 20 simulations) for different fold-changes (FC) between groups and varying level of noise (sd). Dashed line indicates a random performance (error rate = 50%). B) Types of variables selected by the different classification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

~~design (connection between datasets) affects the flexibility of the DIABLO model, resulting in~~ (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 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

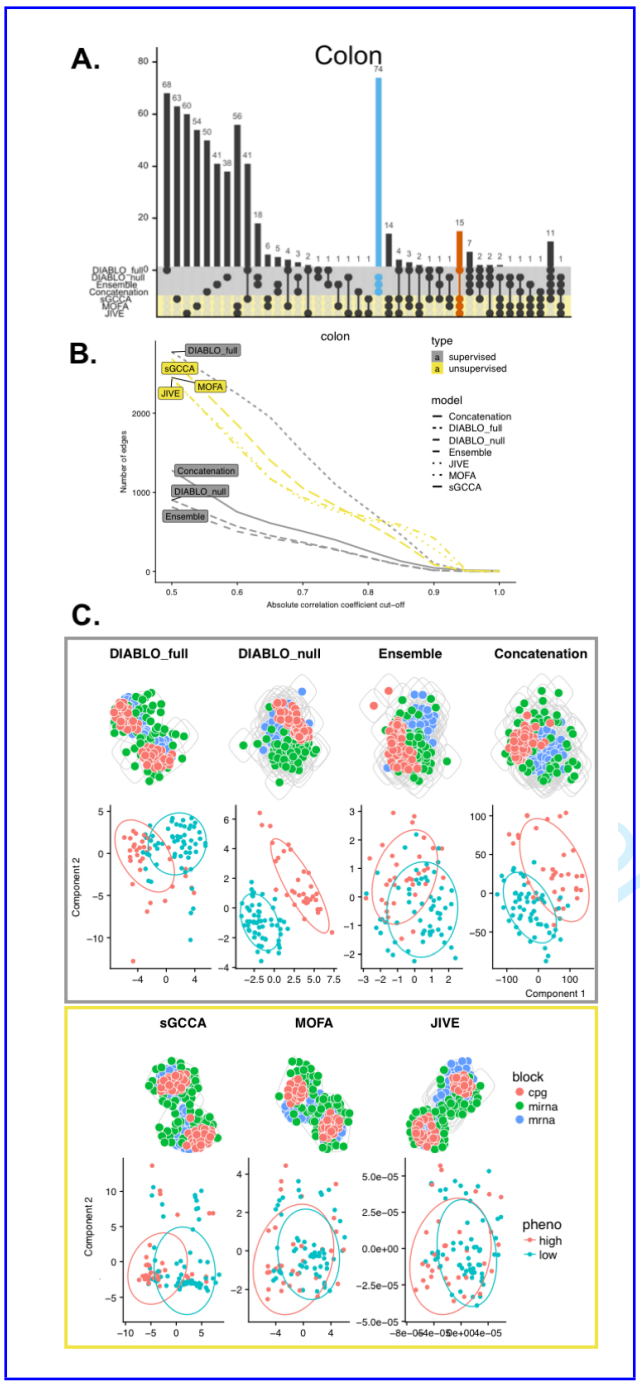


Fig. 2. Benchmark for colon cancer. A) Overlap of features selected by supervised or unsupervised methods. B) Number of correlated variables in the biomarker panels for various Pearson correlation cut-offs. C) Top: network modularity of each multi-omic biomarker panel. Gray circles depict modules based 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.

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 ~~discrimination or correlation~~ correlation and discrimination. DIABLO_null ~~focused~~ focuses on selecting discriminatory variables and ~~disregarded~~ disregards most of the correlation between datasets~~(null design)~~, whereas DIABLO_full ~~selected highly correlated~~ selects highly correlated and discriminatory variables across all datasets. ~~Since the variables~~ Variables selected by DIABLO_full reflect the correlation structure between biological ~~compartments, we hypothesized that they~~

might datasets, and may provide a balance between prediction accuracy and biological insight.

Dataset *n* Omics *p* 92 mRNA 17, 814 high: 33 miRNA 312 low: 59 CpGs 23, 088

122 mRNA 17,665 high: 61 miRNA 329 low: 61 CpGs 24,960

213 mRNA 12,042 high: 105 miRNA 534 low: 108 CpGs 1,305

106 mRNA 12,042 high: 53 miRNA 353 low: 53 CpGs 23,074

989 mRNA 16,851 Basal: 76 (102) miRNA 349 Her2: 38 (40) CpGs 9,482 LumA: 188 (346) Proteins 115 (0)LumB: 77 (122)

28 Cell-types 9 Pre: 14 mRNA modules 229 Post:14 Metabolite modules 60

as described in the next sections.

3.2 DIABLO identifies molecular networks with superior biological enrichment

To assess this, we turn to real biological datasets (Suppl. Section S2).

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 and identified , to identify multi-omics biomarker panels that were predictive of high and low survival times (Table 1). We then compared see Table 1, Suppl. Section S2) and studied the network properties and biological enrichment of the selected featuresacross approaches. Multi-omics biomarker panels were developed using component-based integrative approaches that also performed variable selection. Component-based integrative approaches were compared: supervised methods included concatenation and ensemble schemes using the sPLSDAclassifier (Lê Cao *et al.*, 2011), and DIABLOwith either the null or full design (DIABLO_null 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 sparse generalized canonical correlation analysis (Tenenhaus *et al.*, 2014) (sGCCA)sGCCA, Multi-Omics Factor Analysis (MOFA, Argelaguet *et al.* 2018), and Joint and Individual Variation Explained (JIVE) 23(Suppl. , Lock *et al.* 2013) (see Suppl. Section S3 for parameter settings). Both supervised and unsupervised approaches were considered in order to compare and contrast the types of omics variables selected, network properties and biological enrichment results. A distinction was made between DIABLO models in which the correlation between omics datasets was not maximized (DIABLO_null) and those when the correlation between omics datasets was maximized (DIABLO_full).

Benchmarking integrative methods using multi-omics biomarker panels for different cancers. a) Overlap of selected features using both supervised (green) and unsupervised approaches (purple): a strong overlap was observed between the supervised approaches with the exception of DIABLO_full (blue bars) which showed more similarity to unsupervised methods (dark orange bars). b) Number of edges within each panel network at various Pearson correlation cut-offs: unsupervised approaches panels were more connected than those from supervised approaches, with the exception of DIABLO_full which led to a highly-connected panel. An edge is present if the association between two omic variables is greater than a given correlation cut-off. c) Upper panel: network modularity of each multi-omic biomarker panel for colon cancer showed that unsupervised approaches and DIABLO_full resulted in a few groups of highly connected features, whereas supervised approaches identified networks with many groups of sparsely connected features. Lower panel: component plots depicting the clear separation of subjects in the high and low survival groups for supervised methods as opposed to the unsupervised methods.

Each multi-omics biomarker panel included Each biomarker panel consisted of 180 features (60 features of each omics type across 2 components). Approaches generally identified distinct sets of features. Figure 2a depicts the distinct and shared features between the seven multi-omics panels obtained from the unsupervised (purple, sGCCA, MOFA and JIVE) and supervised (green, Concatenation, Ensemble, DIABLO_null and DIABLO_full) methods . Supervised methods selected many of the same features (blue), but 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 had greater feature overlap whose selection was more similar to those identified with unsupervised methods (orange). The level of connectivity of each of the seven multi-omics panels was assessed by generating networks from the feature adjacency matrix at various Pearson correlation coefficient cut-offs (Fig. 2b). At all cut-offs, unsupervised approaches produced networks with greater connectivity (number of edges)compared to supervised approaches. In addition, biomarker panels A and Suppl. Fig. S5 for the other studies). Interestingly, we observed similarities between the features identified by DIABLO_full , were more similar to those identified by unsupervised approaches, including 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 , indicating that DIABLO_full identified discriminative sets of features that were tightly correlated across biological compartments (Suppl. Figure S4). For example, Figure 2c (upper panel) depicts the networks of all multi-omics biomarker panels for the colon cancer dataset, which show higher modularity (a limited number of large clusters of variables; circled)for the (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 and the unsupervised approaches as compared to the supervised ones. The corresponding component plots show a clear separation between 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 for the panels derived using supervised approaches, whereas the unsupervised approaches could not segregate the survival groups (Figure 2c lower panel, Suppl. Figure S5 and Suppl. Figure S6 for other cancer datasets.

Total number of significant gene set for different types of collections for each integrative method and benchmark dataset (see details in Supplementary Material, FDR = 5%).. Dataset JIVE MOFA sGCCA (Fig. 2C and Suppl. Fig. S9). DIABLO_null DIABLO_full 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).

Colon 45 159 177 92 **639** GBM 1755 1296 1596 1380 **2261** Kidney 233 **306** 25 98 24 Lung 123 179 227 150 **386**

Finally, we carried out gene Gene set enrichment analysis on each multi-omics biomarker panel (biomarker panel using gene symbols of mRNAs and CpGs) against 10 gene set collections (Suppl. Section S4) and tabulated the number of significant (FDR=5%) gene sets (Table **??**). The showed that DIABLO_full model identified the greatest number of significant gene sets across the 10 gene set collections and generally ranked higher than the other methods in the colon (7 collections), gbm (5collections) and lung (5collections) cancer datasets , whereas (Suppl.

Section S4 and Table S1). JIVE outperformed all ~~other~~ methods in the kidney cancer datasets (6 collections). ~~Unlike all other approaches considered~~In conclusion for this benchmark study, DIABLO_full ~~, which aimed to explain both~~ aims at explaining the correlation structure between multiple omics layers and ~~a~~ the phenotype of interest, ~~implicated~~ leading to the greatest number of known biological gene sets ~~(pathways/functions /processesete.).~~

such as pathway, functions and processes.

3.3 ~~Case study 1:~~ DIABLO ~~identified~~ identifies known and novel multi-omics biomarkers of breast cancer subtypes

We ~~next demonstrate that DIABLO can identify novel biomarkers in addition to biomarkers with known biological associations using a case study of human breast cancer. We applied our biomarker analysis workflow to breast cancer datasets~~ applied DIABLO with a full design to the TCGA breast cancer study (TCGA Research Network, 2012) to characterize and predict PAM50 breast cancer subtypes (~~Suppl. Figure S7). After preprocessing and normalization of each omics data-type, the samples were divided into training and test sets (Table 1)~~Table 1, Suppl. Fig. S11). Processing and normalisation is described in Suppl. Section S2. ~~The training data consisted of four omics-datasets (mRNA, miRNA, CpGs and proteins) whereas the test data included all remaining samples for which the protein expression data were missing.~~ 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 ~~cross-validation~~ CV repeated 5 times (~~Suppl. Figure S8). The~~ Suppl. Fig. S12) and chose the number of variables that resulted in the minimum balanced error rate ~~were retained as previously described in (Rohart *et al.*, 2017b). The optimal multi-omics~~ (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 17.9Â±±1.9%. This panel identified many variables with previously known associations with breast cancer, ~~as assessed by looking at the overlap between the panel features and gene sets related to breast cancer based on the Molecular Signature database (MolSigDB)~~ 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).

~~**Identification of a multi-omics biomarker panel predictive of breast cancer subtypes.** a) Variable contributions of each omics-type biomarker that are important to discriminate breast cancer subtypes. b) DIABLO component plots and the derived biomarker panel: 95% confidence ellipses were calculated from the training data set and points depict samples from the test set. c) Heatmap of the scaled expression of variable from the biomarker panel. d) Network visualization of the biomarker panel highlights correlated variables (Pearson correlation > |0.4|) and four communities based on edge betweeness scores. e) A gene set enrichment analysis was conducted on the largest community from d (red cluster) where many cancer related pathways were identified.~~

~~Figure 3a depicts the variable contributions of each omics-type indicated by their loading weight (variable importance). Variables~~ In addition, we identified several variables that were not found in any database and that may represent novel biomarkers of breast cancer ~~. Figure 3b shows the consensus and individual omics component plots based on this biomarker panel, along with 95% confidence ellipses obtained from the training data and superimposed with the samples from the test data. The~~ (Suppl. Fig. S13). Fig. 3A shows that the majority of the ~~samples were~~ 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 ~~, that was predictive of~~ (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~~(balanced error rate = 22.9%)~~. The consensus plot corresponded strongly with the mRNA component plot, ~~depicting~~ with a strong separation of the Basal (error rate = 4.9%) and Her2 (~~error rate =~~ 20%) subtypes~~. We observed~~ , and a weak separation of Luminal A ~~(LumA, error rate = 13.3%)~~ and Luminal B (~~LumB, error rate =~~ error rates of 13.3% and 53.3% respectively) subtypes. ~~Similarly, the heatmap showing the scaled expression of all features of the multi-omics biomarker panel , depicted a strong clustering of the Basal and Her2 samples whereas the Luminal A and B were mixed (Fig. 3c~~A heatmap of the biomarker panel showed similar results (Suppl. Fig. S15). Overall, the features of the multi-omics biomarker panel formed a ~~densely connected network comprising of four communities where variables in each community (cluster) were densely connected with themselves and sparsely connected with other clusters~~ network of four densely connected clusters of variables (Fig. 3~~d~~B). The largest cluster ~~consisted~~ of 72 variables ~~;~~ (20 mRNAs, 21 miRNAs, 15 CpGs and 16 proteins~~(red bubble) and~~) was further investigated using gene set enrichment analysis ~~.~~ (Suppl. Fig. S16). We identified many cancer-associated pathways (~~e.g.~~ *e.g.* FOXM1 pathway, p53 signaling pathway), DNA damage and repair pathways (~~e.g.~~ *e.g.* E2F mediated regulation of DNA replication, G2M DNA damage checkpoint) and various cell-cycle pathways (~~e.g.~~ *e.g.* G1S transition, mitotic G1/G1S phases)~~, demonstrating the ability of DIABLO~~ . Therefore, DIABLO was able to identify a biologically plausible multi-omics biomarker panel ~~. This panel generalized to new breast cancer samplesand implicated previously~~ that generalized to test samples. The panel also included unknown molecular features in breast cancer ~~, which could be further validated in experimental studies.~~

~~**Asthma study: cross-over design and module-based analysis with DIABLO.** a) DIABLO design includes a module-based decomposition approach to discriminate pre-and post-inhalation challenge samples. b) Receiver operating characteristic curves comparing the performance of~~

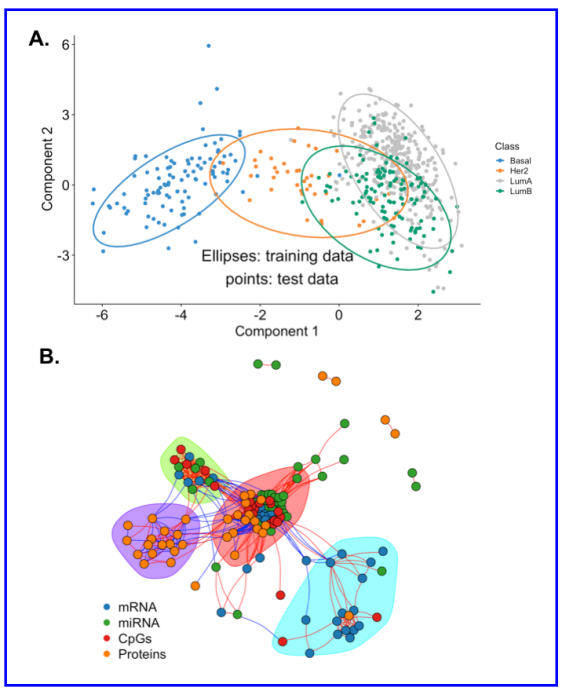


Fig. 3. A Multi-omics biomarker panel predictive of breast cancer subtypes. A) DIABLO consensus component plot based on the identified multi-omics biomarker panel: test samples are overlaid with 95% confidence ellipses calculated from the training data. B) Network visualization of the biomarker panel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

the standard DIABLO and 'multilevel DIABLO' for repeated measures (mDIABLO) using leave-one-out cross-validation. c) Component plots depicting the separation of the pre- and post-challenge samples based on DIABLO and mDIABLO. d) Overlapping features selected from either DIABLO or mDIABLO. e) Heatmap of the Pearson correlation values between the features selected with mDIABLO. f) Circos plot depicting the strongest correlations between different omics features from the mDIABLO panel.

suggesting novel molecular features whose importance would require further experimental validations.

3.4 Case study 2: Competitive classification performance of DIABLOfor repeated measures designs and module-based analyses

Next, we demonstrate the flexibility of DIABLO by extending its use to a repeated measures cross-over study , as well as incorporating module-based analyses that incorporate prior biological knowledge (Suppl. 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) (Allahyar and De Ridder, 2015; Cun and Fröhlich, 2013; Sokolov *et al.*, 2016) . We use a small multi-omics asthma dataset, including pre and post intervention timepoints, to compare a DIABLO modelthat can account for repeated measures (multilevel DIABLO) with the standard DIABLO model as described above (Suppl. Section S6)for details). An 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 was performed (AIC) in 14 subjectsand blood samples were collected before (pre) . Blood was collected pre- and two hours after (post) challenge (Singh *et al.*, 2013, 2014) post-AIC (Singh *et al.*, 2013, 2014). Cell-type frequencies, leukocyte gene transcript expression and plasma metabolite abundances were determined for all samples measured (Table 1). We observed a net decline in lung function after allergen inhalation challenge (Suppl. Figure S9), and the goal of this study was to identify perturbed molecular mechanisms in the blood in response to allergen inhalation challenge. A module based approach (also known as eigengene summarizationA module-based approach (*a.k.a* eigengene summarization, Langfelder and Horvath 2008) was used to transform both the gene expression and metabolite datasets into pathway datasets (Langfelder and Horvath, 2008). 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 in those two datasets now represented the scaled represented the pathway activity expression level for each sample instead of direct gene /metabolite expression . The mRNA dataset was transformed into a dataset of metabolic pathways (based on the Kyoto Encyclopedia of Genes and Genomes, KEGG) whereas the metabolite dataset was transformed into a metabolite pathway dataset based on 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). A).

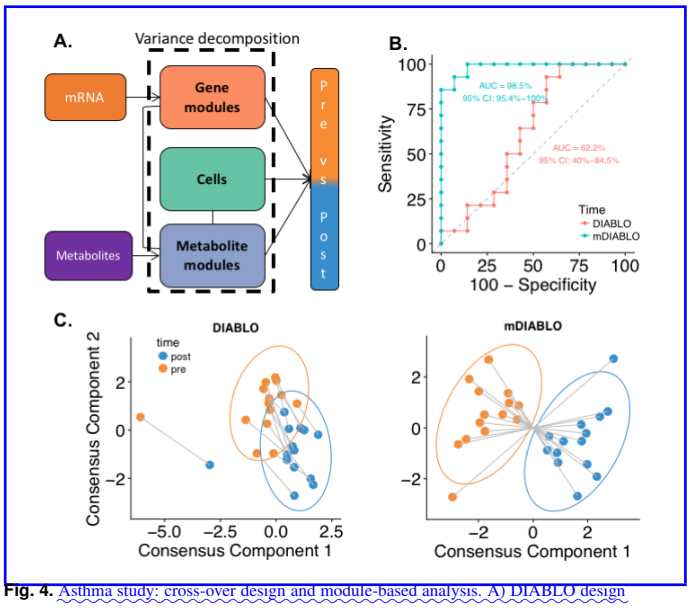


Fig. 4. Asthma study: cross-over design and module-based analysis. A) DIABLO design includes module-based decomposition to discriminate pre-and post-allergen challenge samples. B) Receiver operating characteristic curves comparing standard DIABLO and 'multilevel DIABLO' for repeated measures (mDIABLO)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

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).

4 Discussion

DIABLO aims to identify coherent patterns between datasets that change with respect different phenotypes. This purely 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 had led to poor discriminative ability. In contrast, features identified by supervised methods were discriminative, but formed sparsely connected networks. This 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 controls (Wang, 2011). DIABLO achieves this trade-off by incorporating a priori relationships between different omic domains datatypes to adequately model dysregulated biological mechanisms

between phenotypic conditionspotential dysregulated processes between phenotypic groups. This may explain the superior biological enrichment of the DIABLO_full models in our benchmarking experimentswhere the mRNA and miRNA expression as well as methylation activity were assumed to be correlated (Table ??). Since these omic domains are known to form real regulatory relationships in order to control complex biological processes, these multi-omic biomarker panels may be capturing this biological complexity. . In contrast, these biomarkers were not uncovered when no association was assumed between omic datasets , as in the case of the biomarkers were different when we assumed no association between datasets with DIABLO_null models 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 of omic variables which have prior known biological associations, further suggesting their potential biological plausibility.

There are areas of improvement that DIABLO will benefit from in the near future. The assumption of 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 cross-validation 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 was recently improved (Rohart et al., 2017), is efficient (Rohart *et al.*, 2017a), but we advise using a broad filtering strategy to alleviate computational time when dealing with extremely large datasets (e.g. > 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, as mentioned in (Rohart et al. , 2017). . Finally, DIABLO, like other methods we benchmarked , will is likely to be affected by technical artifacts of the data, such as batch effects and presence of confounding variablesthat may affect downstream integrative analyses. Therefore, we recommend exploratory analyses be carried out in each single omics dataset to assess the effect, if any, of technical factors and use of batch removal methods prior to the integrationanalysisthese effects prior to integration.

5 Conclusion

DIABLO is a versatile, component-based method that can integrate multiple high dimensional datasets and identify key variables that discriminate between phenotypic groups. DIABLO identified more biologically relevant and tightly correlated features across datasets when compared to existing multi-step classification schemes and integrative methods. The framework is highly flexible, suitable for single point or repeated measures study designs, and can accommodate various data transformations, such as feature summarization at the pathway level to enhance biological interpretability. DIABLO's implementation includes intuitive graphical outputs to facilitate the interpretation of integrative analyses.

Acknowledgements

The authors We would like to thank Dr Kevin Chang (University of Auckland) for some preliminary exploratory analyses of the breast cancer dataset. We would also like to thank and Dr Chao Liu (University of Queensland) for obtaining the PAM50 phenotypic information for their help in the preliminary explorations of the TCGA datasets. , and the reviewers for their constructive comments.

Funding

AS is the recipient of the Canadian Institutes of Health Research Doctoral Award âĽ" Frederick Banting and Charles Best Canada Graduate Scholarship and the Michael Smith Foreign Study Supplement award. Research reported in this publication This research was supported in part by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number U19AI118608 (CPS and SJT) . The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health . KALC is supported in part by the National Health 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 analysisâĽ"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., Martin, P. G., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics*, **10**(1), 34.

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.