Genome Biology

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

--Manuscript Draft--

Manuscript Number:	
Full Title:	DIABLO: identifying key molecular drivers from multi-omic assays, an integrative approach
Article Type:	Method
Funding Information:	National Institute Of Allergy And Infectious Diseases of the National Institutes of Health (U19AI118608) — Dr Casey P Shannon the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health (U19AI118608) — A/Prof Scott J Tebbutt National Health and Medical Research Council (GNT1087415) — Dr Kim-Anh Lê Cao Canadian Institutes of Health Research Doctoral Award (NA) — Dr Amrit Singh
Abstract:	Systems biology approaches, leveraging multi-omics measurements, are needed to capture the complexity of biological networks while identifying the key molecular drivers of disease mechanisms. We present DIABLO, a novel integrative method to identify multi-omics biomarker panels that can discriminate between multiple phenotypic groups. In the multi-omics analyses of simulated and real-world datasets, DIABLO resulted in superior biological enrichment compared to other integrative methods, and achieved comparable predictive performance with existing multi-step classification schemes. DIABLO is a versatile approach that will benefit a diverse range of research areas, where multiple high dimensional datasets are available for the same set of specimens. DIABLO is implemented along with tools for model selection, and validation, as well as graphical outputs to assist in the interpretation of these integrative analyses (http://mixomics.org/).
Corresponding Author:	Kim-Anh Lê Cao, Ph.D University of Melbourne Brisbane, Queensland AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Melbourne
Corresponding Author's Secondary Institution:	
First Author:	Amrit Singh
First Author Secondary Information:	
Order of Authors:	Amrit Singh Casey P Shannon Benoît Gautier Florian Rohart Michaël Vacher Scott J Tebbutt Kim-Anh Lê Cao, Ph.D

Order of Authors Secondary Information:	
Suggested Reviewers:	Marylyn D Ritchie Pennsylvania State University marylyn@pennmedicine.upenn.edu Prof Ritchie is a computational biologist, and an expert in multi omics data integration and publishes methods in this specific area of research
	Susan Holmes Stanford University susan@stat.stanford.edu Prof Holmes is a statistician internationally renowned for the type of multivariate methods we have developed in this manuscript, and their application to biological data.
	Aedin Culhane Dana-Farber Cancer Institute aedin@jimmy.harvard.edu Prof Culhane is a statistician with expertise in the type of methods we developed in this manuscript for multi omics data integration.
	Levi Waldron City University of New York - Hunter College levi.waldron@sph.cuny.edu A/Prof Waldron is a statistician expert in omics data analysis and multi omics data integration in particular.
Additional Information:	
Question	**Response**
Has this manuscript been submitted to this journal before?	Yes
Please provide the manuscript identification number from your previous submission. If you no longer have the identification number, please specify this in the text box below. as follow-up to "Has this manuscript been submitted to this journal before?"	GBIO-D-16-01112
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly	Yes

encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	

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

Amrit Singh[1,2,3], Casey P. Shannon[3], Benoît Gautier[4], Florian Rohart[5], Michaël Vacher[6,9], Scott J. Tebbutt[1,3,7], Kim-Anh Lê Cao[8]

[1]Centre for Heart Lung Innovation, St. Paul's Hospital, University of British Columbia, Vancouver, BC, Canada;
[2]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada;
[3]Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, BC, Canada.
[4]The University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, QLD 4102, Australia
[5]Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia
[6]Australian Research Council Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, Western Australia, Australia
[7]Department of Medicine (Respiratory Division), University of British Columbia, Vancouver, BC, Canada.
[8]Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia
[9]current address: Australian eHealth Research Centre, Commonwealth Scientific and Industrial Research Organisation, Brisbane, Queensland, Australia

Corresponding author:
Dr Kim-Anh Lê Cao
Melbourne Integrative Genomics and School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia
T: +61 (0)3834 43971
kimanh.lecao@unimelb.edu.au

40

Abstract

Systems biology approaches, leveraging multi-omics measurements, are needed to capture the complexity of biological networks while identifying the key molecular drivers of disease mechanisms. We present DIABLO, a novel integrative method to identify multi-omics biomarker panels that can discriminate between multiple phenotypic groups. In the multi-omics analyses of simulated and real-world datasets, DIABLO resulted in superior biological enrichment compared to other integrative methods, and achieved comparable predictive performance with existing multi-step classification schemes. DIABLO is a versatile approach that will benefit a diverse range of research areas, where multiple high dimensional datasets are available for the same set of specimens. DIABLO is implemented along with tools for model selection, and validation, as well as graphical outputs to assist in the interpretation of these integrative analyses (http://mixomics.org/).

Keywords: Systems biology, biomarkers, data integration, data visualization, asthma, classification, breast cancer, multi-omics, network analysis

Background

Technological improvements have allowed for the collection of data from different molecular compartments (*e.g.*, gene expression, methylation status, protein abundance) resulting in multiple omics (multi-omics) data from the same set of biospecimens (*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 explains phenotypic variation. Already, systems biology approaches which incorporated data from multiple biological compartments, have shown improved biological insights compared to traditional single omics analyses [1–3]. This may be because single omics analyses cannot account for the interactions between omic layers and, consequently, are unable to reconstruct 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 [4,5].

Preliminary approaches to data integration included multi-step approaches that leveraged existing single-omics methods: multi-omics data were concatenated, or ensembles of single omics models created [6]. These approaches can be biased towards certain omics data types, however, and do not account for interactions between omic layers [7,8]. Recently, more

sophisticated integrative approaches have been proposed (**Supplementary Fig. 1**) [4,9–12]. They can be broadly divided into unsupervised analyses, which identify coherent relationships across 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.

To address these knowledge gaps, we introduce DIABLO, a method that incorporates information across high dimensional multi-omics data while discriminating phenotypic groups. DIABLO uncovers robust biomarkers of dysregulated disease processes that span multiple functional layers. We demonstrate the capabilities and versatility of DIABLO both in simulated and real-world data, integrating multi-omics datasets to identify relevant biomarkers of various diseases. DIABLO is available through the mixOmics data integration toolkit (www.mixomics.org [12]) which contains a wide range of multivariate methods for the exploration and integration of high dimensional biological datasets.

Results

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) datasets while identifying the key omics variables (mRNA, miRNA, CpGs, proteins, metabolites, *etc.*) and characterizing the disease sub-groups or phenotypes of interest. DIABLO uses Projection to Latent Structure models (PLS) [13], and extends both sparse PLS-Discriminant Analysis [14] to

multi-omics analyses and sparse Generalized Canonical Correlation Analysis [15] to a supervised analysis framework. In contrast to existing penalized matrix decomposition methods [16], DIABLO is a component-based method (or a dimension reduction technique) that transforms each omic dataset into latent components and maximizes the sum of pairwise correlations between latent components (user-defined) and a phenotype of interest [17]. DIABLO is, therefore, an integrative classification method that builds predictive multi-omics models that can be applied to multi-omics data from new samples to 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. The method 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 [18] instead of the original omics matrices, as illustrated in one of our case studies.

DIABLO selects correlated and discriminatory variables

Briefly, three omic datasets consisting of 200 samples (split equally over two groups) and 260 variables were generated by modifying the degree of correlation and discrimination, resulting in 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 (**Supplementary Note, Supplementary Fig. 2**). 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 with either a 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, **see 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 (see **Supplementary Fig. 3)**. The purpose of the simulation study was to compare DIABLO models with existing multi-step integrative classifiers with respect to 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.

The concatenation, ensemble and DIABLO_null classifiers performed similarly across 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 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 1 or 2). This simulation highlights how the design (connection between datasets) affects the flexibility of the DIABLO model, resulting in a trade-off between discrimination or correlation. DIABLO_null focused on selecting discriminatory variables and disregarded most of the correlation between datasets (null design), whereas DIABLO_full selected highly correlated variables across all datasets. Since the variables selected by DIABLO_full reflect the correlation structure between biological

compartments, we hypothesized that they might provide a balance between prediction accuracy and biological insight.

DIABLO identifies molecular networks with superior biological enrichment

To assess this, we turn to real biological datasets. We applied various integrative approaches to cancer multi-omics datasets (mRNA, miRNA, and CpG) – colon, kidney, glioblastoma (gbm) and lung – and identified multi-omics biomarker panels that were predictive of high and low survival times (**Table 1**). We then compared the network properties and biological enrichment of the selected features across approaches.

Multi-omics biomarker panels were developed using component-based integrative approaches that also performed variable selection: supervised methods included concatenation and ensemble schemes using the sPLSDA classifier [14], and DIABLO with either the null or full design (DIABLO_null, and DIABLO_full); unsupervised approaches included sparse generalized canonical correlation analysis [15] (sGCCA), Multi-Omics Factor Analysis (MOFA), and Joint and Individual Variation Explained (JIVE) [23] (see **Supplementary Note** 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).

Each multi-omics biomarker panel included 180 features (60 features of each omics type across 2 components). Approaches generally identified distinct sets of features. **Fig. 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 DIABLO_full had greater feature overlap 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 identified by DIABLO_full, were more similar to those identified by unsupervised approaches, including 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 (**Supplementary Fig. 4**). For example, **Fig. 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 DIABLO_full and the unsupervised approaches as compared to the supervised ones. The corresponding component plots show a clear separation between the high and low survival groups for the panels derived using supervised approaches, whereas the unsupervised approaches could not segregate the survival groups [**Fig. 2c** (lower panel), see **Supplementary Fig. 5 and 6** for other cancer datasets].

Finally, we carried out gene set enrichment analysis on each multi-omics biomarker panel (using gene symbols of mRNAs and CpGs) against 10 gene set collections (see **Methods**) and tabulated the number of significant (FDR=5%) gene sets (**Table 2**). The 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 (5 collections)

and lung (5 collections) cancer datasets, whereas JIVE outperformed all other methods in the kidney cancer datasets (6 collections). Unlike all other approaches considered, DIABLO_full, which aimed to explain both the correlation structure between multiple omics layers and a phenotype of interest, implicated the greatest number of known biological gene sets (pathways/functions/processes *etc.*).

Case study 1: DIABLO identified 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 to characterize and predict PAM50 breast cancer subtypes (**Supplementary Fig. 7**). After preprocessing and normalization of each omics data-type, the samples were divided into training and test sets (**Methods, Table 1**). 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 repeated 5 times (**Supplementary Fig. 8**). The number of variables that resulted in the minimum balanced error rate were retained as previously described in [12]. The optimal multi-omics 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, 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) [23], miRCancer [24], Online Mendelian Inheritance in Man (OMIM) [25], and DriverDBv2 [26]. **Figure 3a** depicts the variable contributions of each omics-type indicated by their loading weight (variable importance). Variables not found in any database 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 majority of the samples were within the ellipses, suggesting a reproducible multi-omics biomarker panel from the training to the test set, that was predictive of breast cancer subtypes (balanced error rate = 22.9%). The consensus plot corresponded strongly with the mRNA component plot, depicting a strong separation of the Basal (error rate = 4.9%) and Her2 (error rate = 20%) subtypes. We observed a weak separation of Luminal A (LumA, error rate = 13.3%) and Luminal B (LumB, error rate = 53.3%) 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**). 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 (**Fig. 3d**). The largest cluster in **Fig. 3d** consisted of 72 variables; 20 mRNAs, 21 miRNAs, 15 CpGs and 16 proteins (red bubble) and was further investigated using gene set enrichment analysis. 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), demonstrating the ability of DIABLO to identify a biologically plausible multi-omics biomarker panel. This panel

generalized to new breast cancer samples and implicated previously unknown molecular features in breast cancer, which could be further validated in experimental studies.

Case study 2: DIABLO for 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 [27], as well as incorporating module-based analyses that incorporate prior biological knowledge [28–30]. We use a small multi-omics asthma dataset, including pre and post intervention timepoints, to compare a DIABLO model that can account for repeated measures (multilevel DIABLO) with the standard DIABLO model as described above [20,21]. An allergen inhalation challenge was performed as we previously described in [20,21] in 14 subjects and blood samples were collected before (pre) and two hours after (post) challenge; cell-type frequencies, leukocyte gene transcript expression and plasma metabolite abundances were determined for all samples (**Table 1**). We observed a net decline in lung function after allergen inhalation challenge (**Supplementary Fig. 9)**, 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 summarization [18], **see Methods**) was used to transform both the gene expression and metabolite datasets into pathway datasets. Consequently, each variable in those two datasets now represented the scaled 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 annotations provided by Metabolon Inc. (Durham, North Carolina, USA) (**Fig. 4a**). To account for the repeated measures experimental design, a

multilevel approach [27] was first used to isolate the within-sample variation from each dataset (see **Methods**), and then DIABLO was applied to identify a multi-omics biomarker panel consisting of cells, gene and metabolite modules that discriminated pre- from post-challenge samples. We contrast the resulting 'multilevel DIABLO' (mDIABLO) with a standard DIABLO model that disregards the paired nature of this study by comparing their cross-validation classification performances (**Fig. 4b**). mDIABLO outperformed DIABLO (AUC=98.5% vs. AUC=62.2%, leave-one-out cross-validation, **see Methods**), and we observed a greater degree of separation between the pre- and post-challenge samples for mDIABLO compared to DIABLO (**Fig. 4c)**. Common features (pathways) were identified across omics-types in the mDIABLO model, but not in the standard DIABLO model (**Fig. 4d**). Tryptophan metabolism and Valine, leucine and isoleucine metabolism pathways were identified in both the gene and metabolite module datasets using mDIABLO. The heatmap of pairwise associations of all features identified with mDIABLO demonstrated the ability of DIABLO to select groups of correlated features which were predictive of pre- and post-challenge samples. The Asthma pathway was also identified [even though individual gene members were not significantly altered post-challenge (**Supplementary Fig. 10**)] and was negatively associated with Butanoate metabolism and positively associated with basophils, a hallmark cell-type in asthma (**Fig. 4e**). These findings depict DIABLO's flexibility and sensitivity to detect subtle differences between repeated designs, and its ability to identify common molecular processes spanning different biological layers. The biological pathways identified suggest a mechanistic link with response to allergen challenge.

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 poor discriminative ability. In contrast, features identified by supervised methods were discriminative, but formed sparsely connected networks. This trade-off between correlation and discrimination is a fundamental challenge when trying to identify biologically relevant biomarkers that are also clinically relevant [31]. DIABLO controls this trade-off by incorporating *a priori* relationships between different omic domains to adequately model dysregulated biological mechanisms between phenotypic conditions. This may explain the superior biological enrichment of the DIABLO_full models in our benchmarking experiments where the mRNA and miRNA expression as well as methylation activity were assumed to be correlated (**Table 2**). 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 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 linear relationship between the selected omics features to explain the phenotypic response may not apply in some biological research areas, for example when integrating distance-based metagenomics studies, where kernel approaches could be further explored [32]. Selecting the optimal number of variables requires repeated cross-validation 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 [12], 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 [12]. Finally, DIABLO, like other methods we benchmarked, will be affected by technical artifacts of the data, such as batch effects and presence of confounding variables that 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 integration analysis [33–35].

To summarize, 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.

Online Methods

Code availability and software tool requirements. The DIABLO framework is implemented in the mixOmics R package [12]. mixOmics currently includes 19 multivariate methodologies, for single-omics and integrative analyses. All scripts and tutorials are provided in our companion web-page http://www.mixomics.org/mixDIABLO. All analyses were performed using the R statistical computing program (version 3.4.1) and the mixOmics package (version 6.3.0).

Statistical methods and analysis

General multivariate framework to integrate multiple datasets measured on the same samples. DIABLO extends sparse generalized canonical correlation analysis (sGCCA) [15] to a classification (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 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 normalized, centered and scaled datasets X_1 (n x p_1), …, X_K (n x p_K), measuring the expression levels of $p_1, p_2, \ldots, p_K$ omics variables on the same n samples, $k = 1, \ldots, K$, sGCCA solves the optimization function:

$$\max_{\boldsymbol{a}^1,\ldots,\boldsymbol{a}^K} \sum_{k,j=1,k\neq j}^{K} c_{jk} cov(X_k \boldsymbol{a}^k, X_j \boldsymbol{a}^j), \quad s.t. \ \left||\boldsymbol{a}^k|\right|_2 = 1 \text{ and } \left||\boldsymbol{a}^k|\right|_1 < \lambda_k \qquad (1)$$

where c_{jk} indicates whether to maximize the covariance between the datasets X_k and X_j according to the design matrix, with c_{jk} values ranging from 0 (no relationship modelled between the datasets) to 1 otherwise, $\boldsymbol{a}^k$ is the variable coefficient vector for each dataset X_k, λ_k is a non-negative parameter that controls the amount of shrinkage and thus the number of non-zero coefficients in $\boldsymbol{a}^k$. Similar to Lasso [36] and other l_1 – penalized multivariate models developed for single omics analysis [14], the l_1 penalization improves the interpretability of the component scores $X_k \boldsymbol{a}^k$ that is now only defined on a subset of variables with non-zero coefficients in X_k. The result is the identification of variables that are highly correlated between and within omics datasets.

Equation (1) describes the sGCCA model for the first dimension. Once the first set of coefficient vectors $\boldsymbol{a}_1^k$ and associated component scores $\boldsymbol{t}_1^k = X_k \boldsymbol{a}_1^k$ are obtained, residual matrices are calculated during the 'deflation' step for the second dimension, such that $X_k^2 = X_k^1 - \boldsymbol{t}_1^k \boldsymbol{a}_1^k$, 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 by X_k^2 in (1). This process is repeated until a sufficient number of dimensions (or set of components) is obtained.

The underlying assumption of the sGCCA model is that the major source of common biological variation can be extracted via the first sets of component scores $X_k \boldsymbol{a}^k$, while any unwanted variation due to heterogeneity across the datasets X_K does not impact the statistical model. The optimization problem (1) is solved using a monotonically convergent algorithm [15].

DIABLO for supervised analysis and prediction.

To extend sGCCA for a classification framework, we substitute one omics dataset X_k in (1) with a dummy indicator matrix Y of size (n x G), where G is the number of phenotype groups that indicate the class membership of each sample. In addition, and for easier use of the method, we replaced the l_1 penalty parameter λ_k by the number of variables to select in each dataset and each component, as there is a direct correspondence between both parameters.

Denote a new sample i which is measured across the different types of omics datasets $\boldsymbol{x}_k^i$, its class membership is predicted by the fitted sGCCA model with the estimated variable coefficients vectors $\widehat{\boldsymbol{a}^k}$ to obtain the predicted scores $\boldsymbol{t}^{k,i} = \boldsymbol{x}_k^i \, \widehat{\boldsymbol{a}^k}$, $k = 1, \dots, K$. Therefore, to each dataset k corresponds a predicted continuous score $\boldsymbol{t}^{k,i}$. The predicted class of sample i for each dataset is obtained from the predicted score using one of the distances Maximum, Centroids or Mahalanobis [37] as described in [12]. The consensus class membership is determined using either a majority vote, or by averaging all $\boldsymbol{t}^{k,i}$ across all K datasets before using the prediction distance of choice ('average prediction' 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. 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 [12] for details about distance measures and voting schemes that can be used).

Design matrix in DIABLO. The design matrix C is a (K x K) matrix with values ranging from 0 to 1 which specifies whether the covariance between two datasets should be maximized DIABLO (see equation (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):

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

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 below ***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 [12].

Input data in DIABLO. While DIABLO does 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 (see below).

Parameters tuning.

The first parameter to tune is the design matrix C, which can be determined using either prior biological knowledge, or a data-driven approach. The latter approach uses PLS method implemented in mixOmics that models pair-wise associations between omics datasets. 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 the DIABLO design as a 1 value.

The second parameter to tune is the total number of components. In several analyses we found that G − 1 components were sufficient to extract sufficient information to discriminate all phenotype groups [14], but this can be assessed by evaluating the model performance across all specified components (described below) as well as using graphical outputs such as sample plots to visualize the discriminatory ability of each component.

Finally, the third set of parameters to tune is the number of variables to select per dataset and per component. Such tuning can rapidly become cumbersome, as there might be numerous combinations of selection sizes to evaluate across all *K* datasets. For the breast cancer study, we used 5-fold cross-validation repeated 50 times to evaluate the performance of the model over a grid of different possible values of variables to select (**Supplementary Fig. 8**). The performance of the model for a given set of parameters (including number of component and number of variables to select) was based on the balanced classification error rate using majority vote or average prediction schemes with centroids distance. The balanced classification error rate is useful in the case of imbalanced class sizes, where the majority classes can have strong influence on the overall error rate. The balanced error rate measure calculates the weighted average of the individual class error rates with respect to their class sample size. In our experience, the number of variables to select in each dataset provided less of an improvement on the error rate compared to tuning the number of components. Therefore, even a grid composed of a small number of

variables (<50 with steps of 5 or 10) may suffice as it does not substantially change 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 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.

Visualization outputs with DIABLO. To facilitate the interpretation of the integrative analysis, several types of graphical outputs were implemented in mixOmics.

Sample plots. The consensus plot which depicts the samples is computed by calculating the average of the components from each dataset. Omics specific samples plots can also be obtained by plotting components associated to each data set. The sample plot are useful to visualize the ability of the DIABLO model to extract common information at the sample level for each dataset, and the discriminatory power of each data type to separate the phenotypic groups. The scatterplot matrix represents the correlation between components for the same dimension but across all omics datasets. This plot assesses the model's ability to maximize the correlation as indicated in the design matrix. Separation of subjects according to their phenotypic groups can be visualized.

Variable plots. To visualize selected variables, we proposed circos plot to represent correlations between and within 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 [38]. For each omics dataset, DIABLO produces a variable coefficient matrix of size (p_kx H), where H is the total number of components in the

model. The product of any two matrices approximates the association score between variables of the two omics datasets. The association between variables 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.

Clustered Image Map (CIM). A clustered image map [38] based on the Euclidean distance and the complete linkage displays an unsupervised clustering between the selected variables (centered and scaled) and the samples. Color bars represent the sample phenotypic groups (columns) and the type of omics (rows) variables.

Gene-set enrichment analyses

Significance of enrichment was determined using a hypergeometric test of the overlap between the selected features (mapped to official HUGO gene symbols or official miRNA symbols) and the various gene sets contained in the collections. In order to carry out the comparison, each feature set was mapped back to official HUGO gene symbols. This was done as follows across the respective data types: mRNA, CpGs and proteins. The following collections were used as gene-sets for the enrichment analysis [39]: C1 - positional gene sets for each human chromosome and cytogenetic band. C2 – curated gene sets (Pathway Interaction DB [PID], Biocarta [BIOCARTA], Kyoto Encyclopedia of Genes and Genomes [KEGG], Reactome [REACTOME], and others), C3 - motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes. C4 – computational gene sets (from the Cancer Gene Neighbourhoods [CGN] and Cancer Modules [CM] – citation available via the MolSigDB [23]. C5 - GO gene sets consist of genes annotated by the same GO terms. C6 –

ontologic gene sets (Gene sets represent signatures of cellular pathways which are often dis-regulated in cancer). C7 - immunologic gene sets defined directly from microarray gene expression data from immunologic studies. H - hallmark gene sets are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes. & A. BTM - Blood Transcriptional Modules [40]. B. TISSUES - cell-specific expression from Benita *et al.* [41].

Modular analysis: Eigengene summarization is a common approach to decompose a *n* x *p* dataset (where *n* is the number of samples and *p* is the number of variables in a module), to a component (linear combination of all *p* variables) that represents the summarized expression of genes in the module [18]. For the asthma study, 15,683 genes were reduced to 229 KEGG pathways and 292 metabolites were reduced to 60 metabolic pathways using eigengene summarization.

Multilevel transformation: 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) [42], 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.

Declarations

Acknowledgements

The authors would like to thank Mr Kevin Chang (University of Auckland) for some preliminary exploratory analyses of the breast cancer dataset. We would also like to thank Dr Chao Liu (University of Queensland) for obtaining the PAM50 phenotypic information for the TCGA datasets.

Competing interests

The authors declare no competing interests.

Funding

AS is the recipient of the Canadian Institutes of Health Research Doctoral Award – Frederick Banting and Charles Best Canada Graduate Scholarship and the Michael Smith Foreign Study Supplement award. Research reported in this publication 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 and Medical Research Council (NHMRC) Career Development fellowship (GNT1087415).

Authors' contributions

AS performed the data pre-processing, the statistical analyses and developed the DIABLO method. BG implemented the R scripts for DIABLO and graphical outputs, CPS performed the

gene enrichment analyses, MV implemented the circos plots, FR and BG implemented the R scripts in mixOmics along with the S3 functions, SJT supervised AS and participated in the design of the study. KALC supervised AS, BG, MV and FR, participated in the development of the DIABLO method and provided statistical advice. AS and KALC edited the manuscript, with editorial input from SJT and CPS.

Tables

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

Analysis	Dataset	Number of samples	Sample size in each subtype			Omics	Number of variables
Benchmark cancer datasets (Wang et al. [3])	Colon	92	High (33) Low (59)			mRNA	17,814
						miRNA	312
						CpGs	23,088
	Kidney	122	High (61) Low (61)			mRNA	17,665
						miRNA	329
						CpGs	24,960
	Glioblastoma (gbm)	213	High (105) Low (108)			mRNA	12,042
						miRNA	534
						CpGs	1,305
	Lung	106	High (53) Low (53)			mRNA	12,042
						miRNA	353
						CpGs	23,074
Case study 1 (The Cancer Genome Atlas **) [19]**	Breast cancer	989		Train	Test	mRNA	16,851
			Basal	76	102	miRNA	349
			Her2	38	40	CpGs	9,482
			LumA	188	346	Proteins	Train: 115 Test: 0
			LumB	77	122		
Case study 2 (Singh et al. [20,21])	Asthma	28	Pre (14) Post (14)			Cell-types	9
						mRNA-modules	229
						metabolite-modules	60

Table 2. Number of significant gene sets for each integrative method and benchmarking cancer dataset. Best performing method is indicated in the shaded cell. Each row represents a gene set collection (**see Methods** for details, FDR = 5%).

		Unsupervised, integrative			Supervised, non-integrative			Supervised, integrative
disease	**collection**	**JIVE**	**MOFA**	**sGCCA**	**Concatenation**	**Ensemble**	**DIABLO_null**	**DIABLO_full**
Colon	**BTM**	0	4	0	0	0	0	23
	C1	0	0	0	0	0	0	0
	C2	15	14	5	12	3	21	113
	C3	8	5	14	11	2	6	0

	C4	0	1	0	1	2	1	46
	C5	19	36	147	7	0	0	216
	C6	0	0	0	0	0	0	0
	C7	1	87	11	61	10	62	218
	H	0	0	0	0	0	2	7
	TISSUES	2	12	0	0	0	0	16
	TOTAL	45	159	177	92	17	92	639
Gbm	**BTM**	0	0	19	10	9	10	30
	C1	0	0	0	0	0	0	0
	C2	275	337	193	258	358	312	426
	C3	94	64	37	14	15	15	34
	C4	49	43	68	47	50	62	125
	C5	825	708	706	526	669	776	693
	C6	22	25	18	30	24	24	21
	C7	460	82	526	432	173	147	869
	H	12	8	8	19	23	20	19
	TISSUES	18	29	21	10	12	14	44
	TOTAL	1755	1296	1596	1346	1333	1380	2261
Kidney	**BTM**	1	0	0	0	0	0	0
	C1	0	0	1	0	0	0	1
	C2	42	33	7	10	5	15	4
	C3	8	80	1	4	35	23	1
	C4	17	6	0	7	1	3	0
	C5	157	110	1	55	27	46	0
	C6	0	0	0	0	0	0	0
	C7	0	74	15	93	13	10	18
	H	6	3	0	1	0	1	0
	TISSUES	2	0	0	0	0	0	0
	TOTAL	233	306	25	170	81	98	24
Lung	**BTM**	0	0	0	0	0	2	0
	C1	0	0	0	1	0	0	1
	C2	4	17	2	0	0	1	33
	C3	48	20	57	50	26	21	19
	C4	17	0	47	0	0	18	13
	C5	35	127	42	0	25	22	193
	C6	1	0	1	3	2	5	7
	C7	18	13	78	0	7	72	100
	H	0	2	0	0	1	0	0

	TISSUES	0	0	0	0	0	9	20
	TOTAL	123	179	227	54	61	150	386

Figure captions

Figure 1. 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 (**see Supplementary Note**). 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).

Figure 2. 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.

Figure 3. 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 4. 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 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.

References

1. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, et al. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. Levchenko A, editor. PLoS Biol [Internet]. 2012 [cited 2016 Jan 19];10:e1001301. Available from: http://dx.plos.org/10.1371/journal.pbio.1001301

2. Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData Min. 2013;6:23.

3. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods [Internet]. 2014 [cited 2016 Jan 19];11:333–7. Available from: http://www.nature.com/doifinder/10.1038/nmeth.2810

4. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet [Internet]. 2015 [cited 2015 Jul 10];16:85–97. Available from: http://www.nature.com/doifinder/10.1038/nrg3868

5. Yugi K, Kubota H, Hatano A, Kuroda S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. Trends Biotechnol [Internet]. 2016 [cited 2018 Feb 21];34:276–90. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0167779915002735

6. Günther O, Chen V, Freue GC, Balshaw R, Tebbutt S, Hollander Z, et al. A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers. 2012 [cited 2016 Jan 19];13:326. Available from: http://summit.sfu.ca/item/13303

7. Aben N, Vis DJ, Michaut M, Wessels LFA. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. Bioinformatics [Internet]. 2016 [cited 2017 Aug 2];32:i413–20. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw449

8. Ma S, Ren J, Fenyö D. Breast cancer prognostics using multi-omics data. AMIA Summits Transl Sci Proc [Internet]. 2016 [cited 2017 May 30];2016:52. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001766/

9. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics [Internet]. 2016 [cited 2016 May 8];17. Available from: http://www.biomedcentral.com/1471-2105/17/S2/15

10. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform [Internet]. 2016 [cited 2018 Feb 21];17:628–41. Available from: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv108

11. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet [Internet]. 2017 [cited 2018 Feb 21];8. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2017.00084/full

12. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Comput Biol [Internet]. 2017 [cited 2018 Jan 29];13:e1005752. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752

13. Wold H. Estimation of Principal Components and Related Models by Iterative Least squares. Multivar Anal. 1966;391–420.

14. Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics [Internet]. 2011 [cited 2015 Jul 15];12:253. Available from: http://www.biomedcentral.com/1471-2105/12/253/

15. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. Biostatistics [Internet]. 2014 [cited 2015 Jul 15];15:569–83. Available from: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxu001

16. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics [Internet]. 2009 [cited 2016 Jul 27];10:515–34. Available from: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxp008

17. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res [Internet]. 2004 [cited 2016 Mar 30];14:1085–1094. Available from: http://genome.cshlp.org/content/14/6/1085.short

18. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics [Internet]. 2008 [cited 2016 Apr 4];9:559. Available from: http://www.biomedcentral.com/1471-2105/9/559

19. The TCGA Research Network. The Cancer Genome Atlas [Internet]. Available from: http://cancergenome.nih.gov/

20. Singh A, Yamamoto M, Kam SHY, Ruan J, Gauvreau GM, O'Byrne PM, et al. Gene-metabolite expression in blood can discriminate allergen-induced isolated early from dual asthmatic responses. Hsu Y-H, editor. PLoS ONE [Internet]. 2013 [cited 2015 Jul 18];8:e67907. Available from: http://dx.plos.org/10.1371/journal.pone.0067907

21. Singh A, Yamamoto M, Ruan J, Choi JY, Gauvreau GM, Olek S, et al. Th17/Treg ratio derived using DNA methylation analysis is associated with the late phase asthmatic response. Allergy Asthma Clin Immunol [Internet]. 2014 [cited 2016 Mar 2];10:32. Available from: http://www.biomedcentral.com/content/pdf/1710-1492-10-32.pdf

22. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat [Internet]. 2013 [cited 2018 Jan 24];7:523–42. Available from: http://projecteuclid.org/euclid.aoas/1365527209

23. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst [Internet]. 2015 [cited 2018 Jan 30];1:417–25. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2405471215002185

24. Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics [Internet]. 2013 [cited 2018 Jan 30];29:638–44. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt014

25. Hamosh A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res [Internet]. 2004 [cited 2018 Jan 30];33:D514–7. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki033

26. Chung I-F, Chen C-Y, Su S-C, Li C-Y, Wu K-J, Wang H-W, et al. DriverDBv2: a database for human cancer driver gene research. Nucleic Acids Res [Internet]. 2016 [cited 2018 Jan 30];44:D975–9. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1314

27. Liquet B, Lê Cao K-A, Hocini H, Thiébaut R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. BMC Bioinformatics [Internet]. 2012 [cited 2015 Jul 18];13:325. Available from: http://www.biomedcentral.com/1471-2105/13/325/

28. Allahyar A, de Ridder J. FERAL: network-based classifier with application to breast cancer outcome prediction. Bioinformatics [Internet]. 2015 [cited 2018 Feb 1];31:i311–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv255

29. Cun Y, Fröhlich H. Network and data integration for biomarker signature discovery via network smoothed t-statistics. Boccaletti S, editor. PLoS ONE [Internet]. 2013 [cited 2017 May 30];8:e73074. Available from: http://dx.plos.org/10.1371/journal.pone.0073074

30. Sokolov A, Carlin DE, Paull EO, Baertsch R, Stuart JM. Pathway-based genomics prediction using generalized elastic net. PLoS Comput Biol [Internet]. 2016 [cited 2017 May 30];12:e1004790. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004790

31. Wang TJ. Assessing the Role of Circulating, Genetic, and Imaging Biomarkers in Cardiovascular Risk Prediction. Circulation [Internet]. 2011 [cited 2018 Feb 23];123:551–65. Available from: http://circ.ahajournals.org/cgi/doi/10.1161/CIRCULATIONAHA.109.912568

32. Mariette J, Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. Bioinformatics [Internet]. 2017 [cited 2018 Mar 6]; Available from: http://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btx682/4565592

33. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics [Internet]. 2007 [cited 2016 May 12];8:118–27. Available from: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxj037

34. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics [Internet]. 2012 [cited 2018 Mar 6];13:539–52. Available from: https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxr034

35. Parker HS, Corrada Bravo H, Leek JT. Removing batch effects for prediction problems with frozen surrogate variable analysis. PeerJ [Internet]. 2014 [cited 2016 May 12];2:e561. Available from: https://peerj.com/articles/561

36. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996;58:267–88.

37. Le Cao K-A, Gonzalez I, Dejean S. integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics [Internet]. 2009 [cited 2016 Apr 3];25:2855–6. Available from: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp515

38. González I, Lê Cao K-A, Davis MJ, Déjean S. Visualising associations between paired 'omics' data sets. BioData Min [Internet]. 2012 [cited 2015 Jul 15];5:1–23. Available from: http://link.springer.com/article/10.1186/1756-0381-5-19

39. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci [Internet]. 2005 [cited 2016 Jul 26];102:15545–15550. Available from: http://www.pnas.org/content/102/43/15545.short

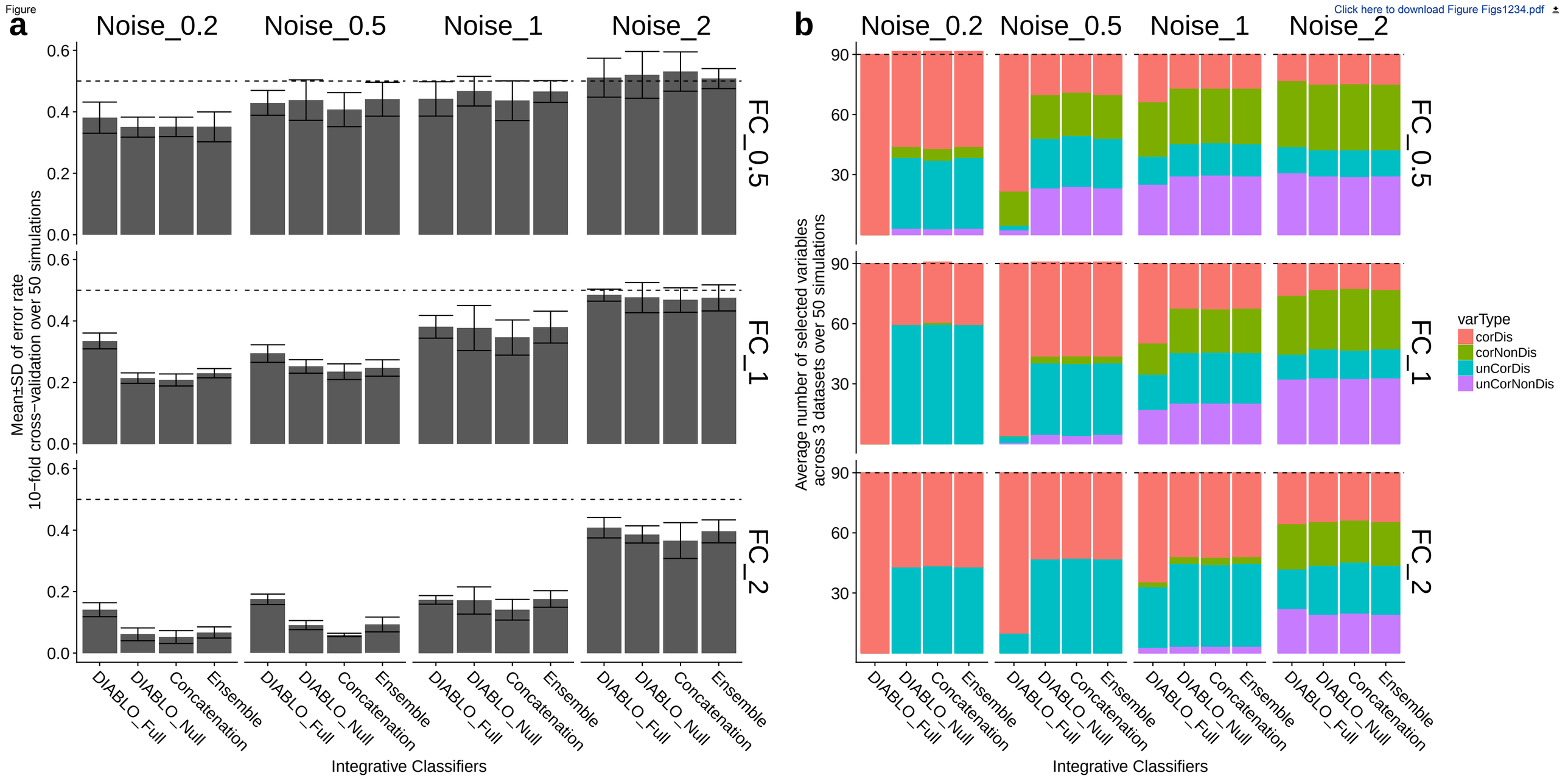
40. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. Immunity [Internet]. 2008 [cited 2016 Jul 22];29:150–64. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1074761308002835

41. Benita Y, Cao Z, Giallourakis C, Li C, Gardet A, Xavier RJ. Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT repressor. Blood [Internet]. 2010 [cited 2018 Mar 5];115:5376–84. Available from: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2010-01-263855

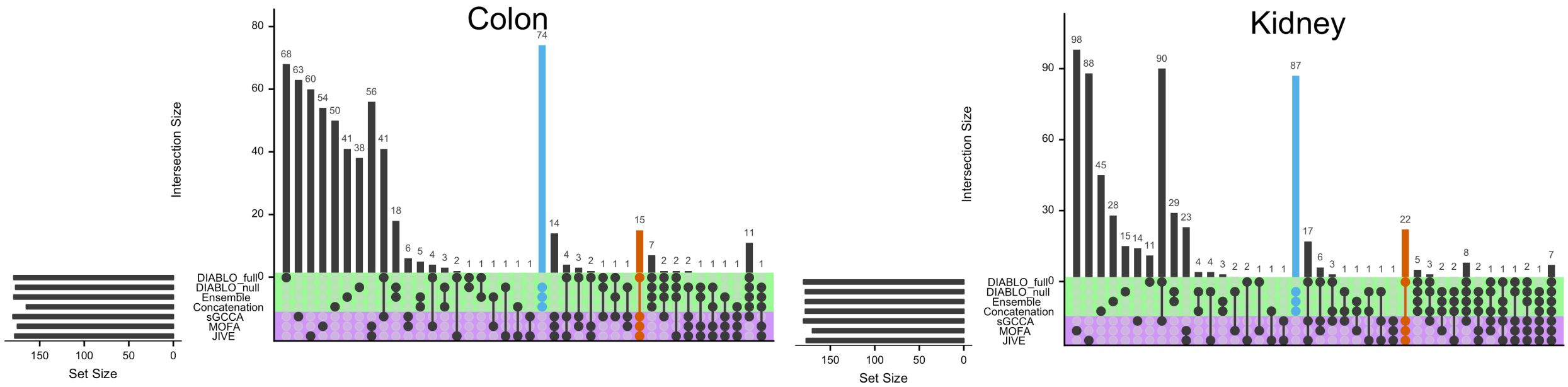
42. Westerhuis JA, van Velzen EJJ, Hoefsloot HCJ, Smilde AK. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metabolomics [Internet]. 2010 [cited 2016 Jul 27];6:119–28. Available from: http://link.springer.com/10.1007/s11306-009-0185-z

Figure

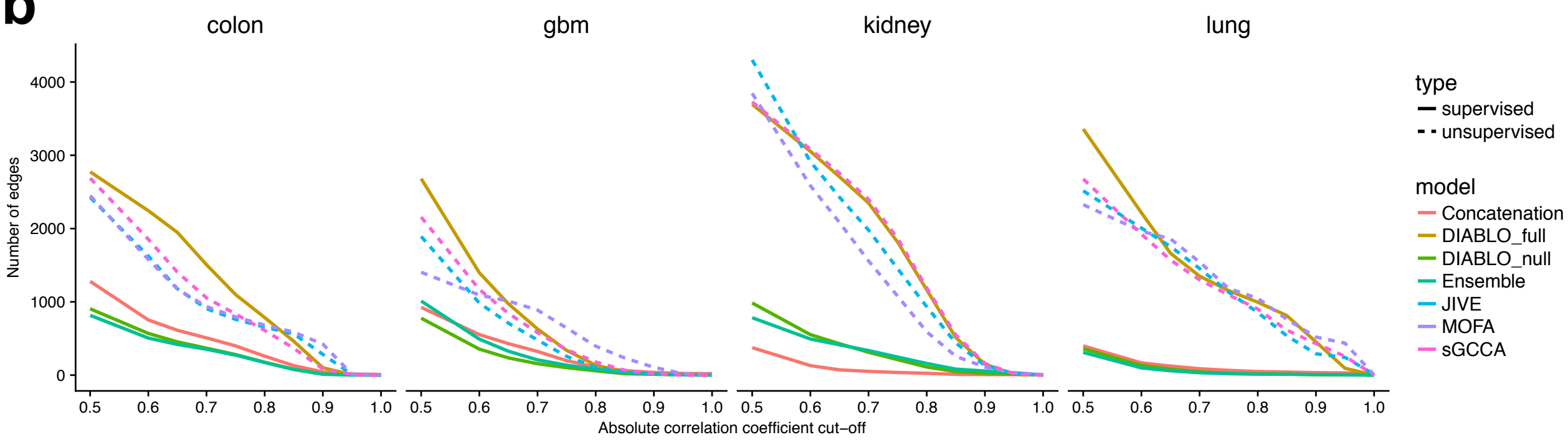
Click here to download Figure Figs1234.pdf



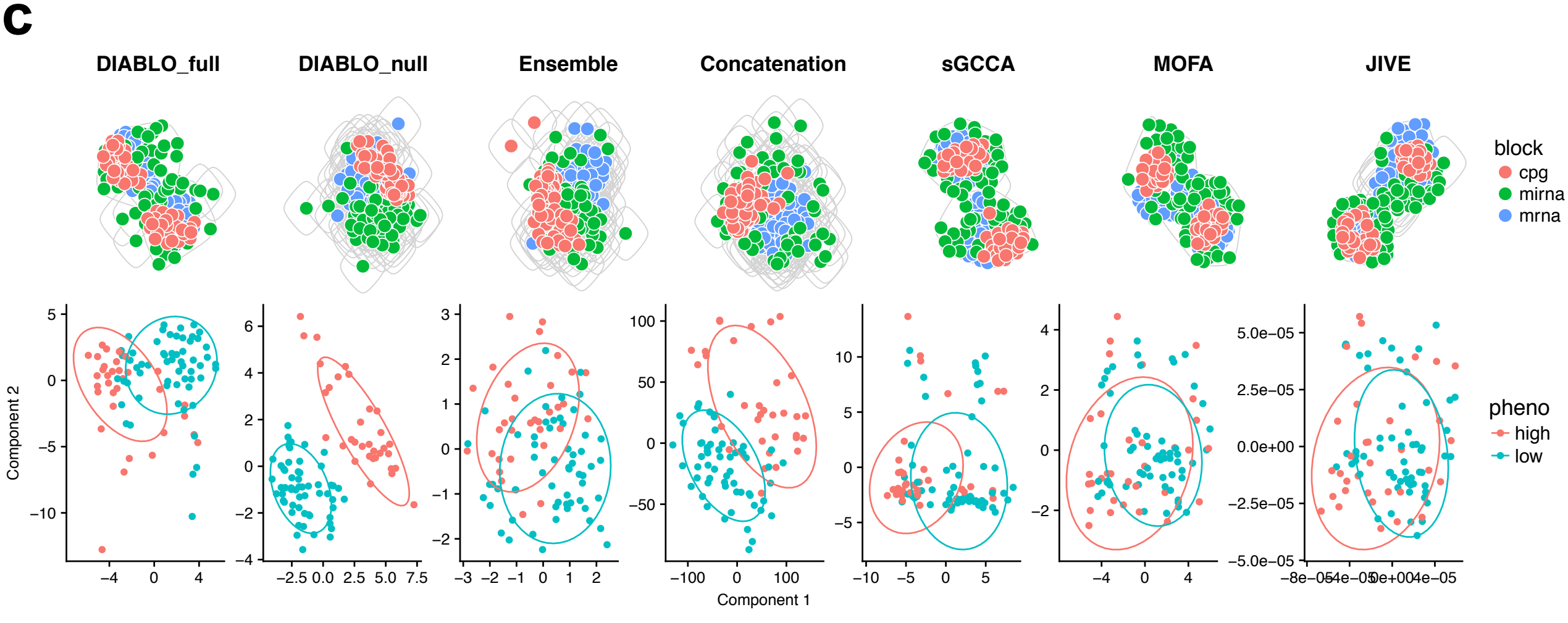
a

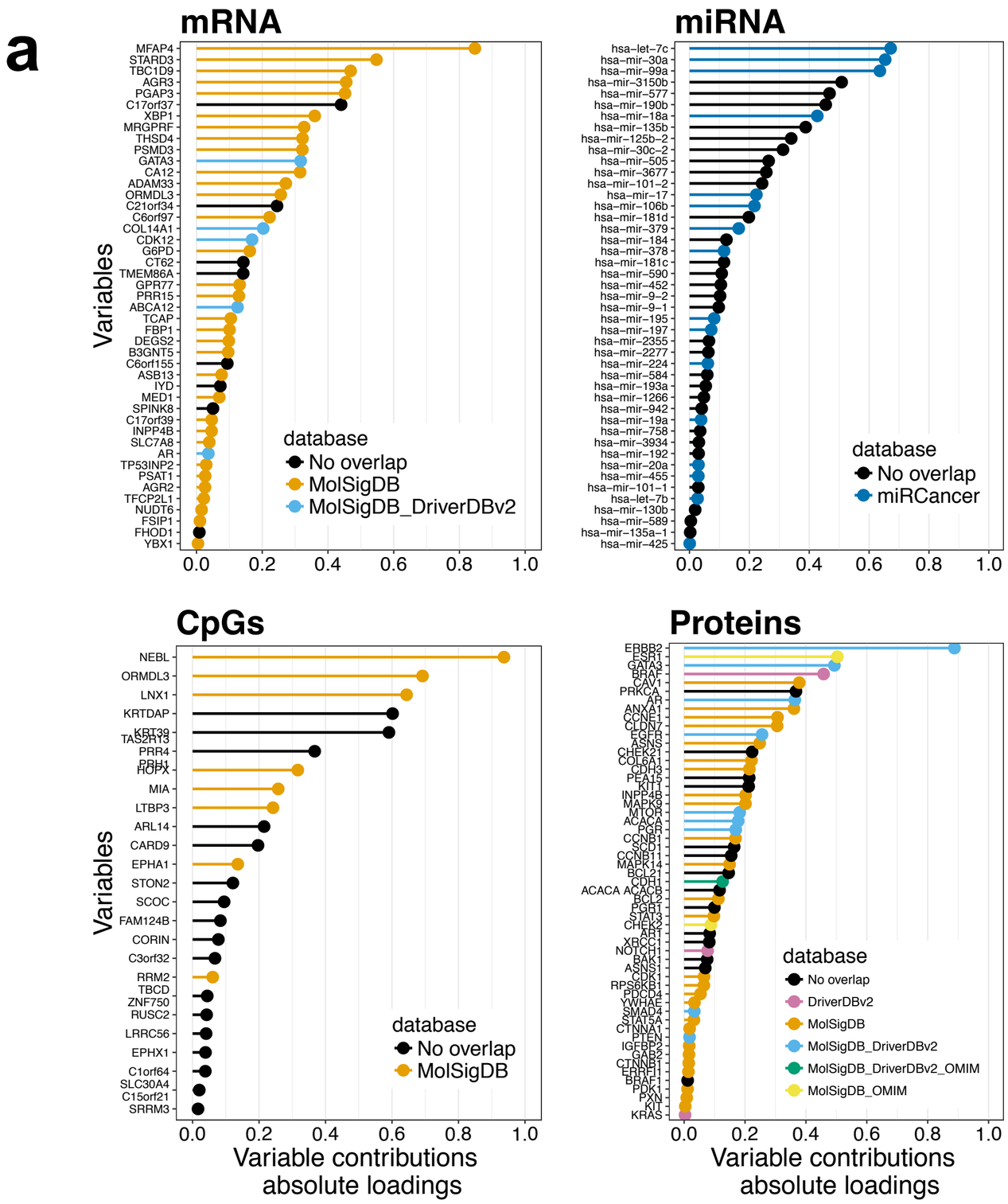

Colon
Kidney
GBM
Lung
Intersection Size
Set Size
DIABLO_full
DIABLO_null
Ensemble
Concatenation
sGCCA
MOFA
JIVE

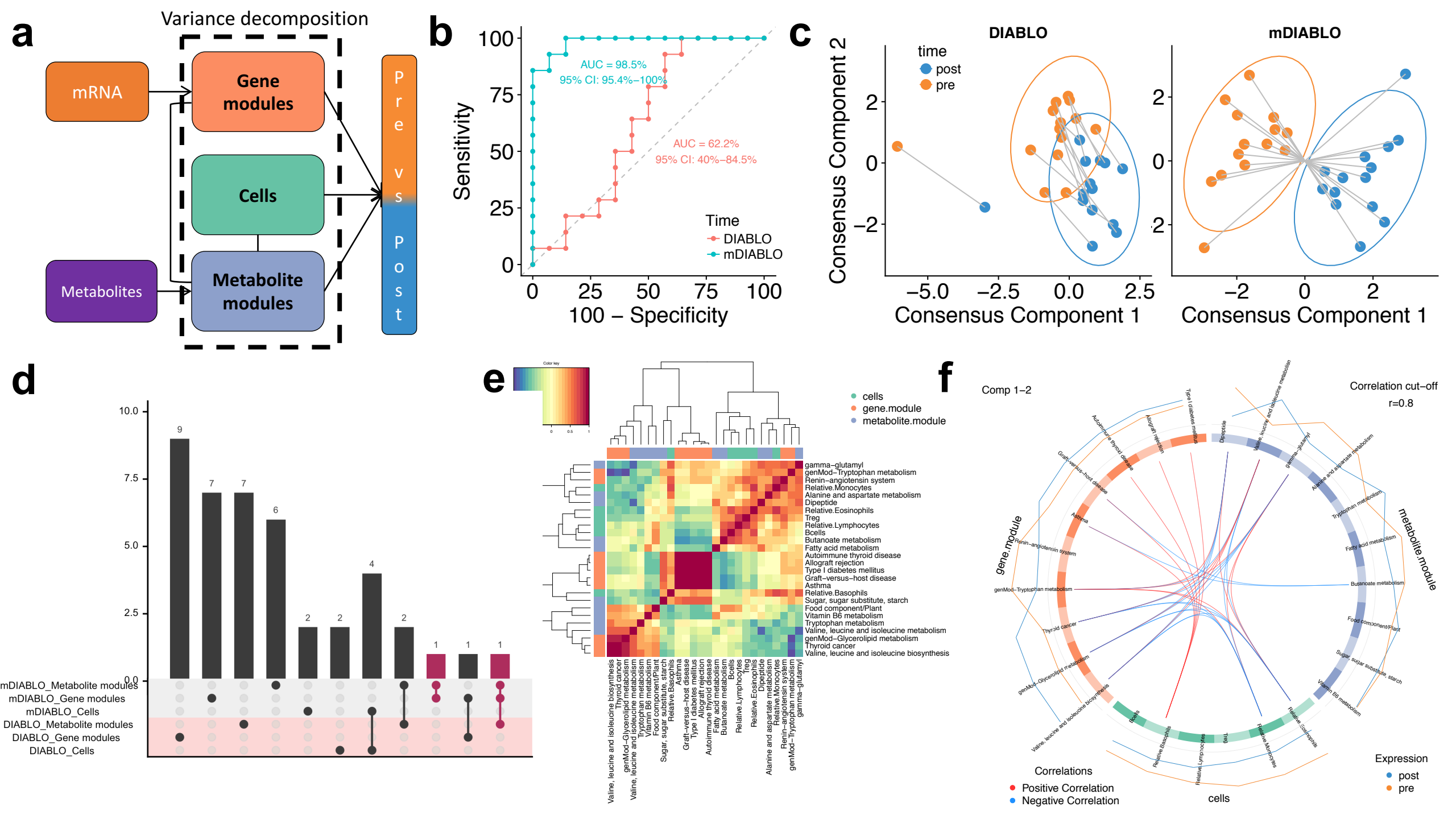
b


colon
gbm
kidney
lung
Number of edges
Absolute correlation coefficient cut−off
type
supervised
unsupervised
model
Concatenation
DIABLO_full
DIABLO_null
Ensemble
JIVE
MOFA
sGCCA

c


DIABLO_full
DIABLO_null
Ensemble
Concatenation
sGCCA
MOFA
JIVE
block
cpg
mirna
mrna
pheno
high
low
Component 1
Component 2


a
mRNA
Variables
MFAP4
STARD3
TBC1D9
AGR3
PGAP3
C17orf37
XBP1
MRGPRF
THSD4
PSMD3
GATA3
CA12
ADAM33
ORMDL3
C21orf34
C6orf97
COL14A1
CDK12
G6PD
CT62
TMEM86A
GPR77
PRR15
ABCA12
TCAP
FBP1
DEGS2
B3GNT5
C6orf155
ASB13
IYD
MED1
SPINK8
C17orf39
INPP4B
SLC7A8
AR
TP53INP2
PSAT1
AGR2
TFCP2L1
NUDT6
FSIP1
FHOD1
YBX1
database
No overlap
MolSigDB
MolSigDB_DriverDBv2
0.0
0.2
0.4
0.6
0.8
1.0
miRNA
hsa−let−7c
hsa−mir−30a
hsa−mir−99a
hsa−mir−3150b
hsa−mir−577
hsa−mir−190b
hsa−mir−18a
hsa−mir−135b
hsa−mir−125b−2
hsa−mir−30c−2
hsa−mir−505
hsa−mir−3677
hsa−mir−101−2
hsa−mir−17
hsa−mir−106b
hsa−mir−181d
hsa−mir−379
hsa−mir−184
hsa−mir−378
hsa−mir−181c
hsa−mir−590
hsa−mir−452
hsa−mir−9−2
hsa−mir−9−1
hsa−mir−195
hsa−mir−197
hsa−mir−2355
hsa−mir−2277
hsa−mir−224
hsa−mir−584
hsa−mir−193a
hsa−mir−1266
hsa−mir−942
hsa−mir−19a
hsa−mir−758
hsa−mir−3934
hsa−mir−192
hsa−mir−20a
hsa−mir−455
hsa−mir−101−1
hsa−let−7b
hsa−mir−130b
hsa−mir−589
hsa−mir−135a−1
hsa−mir−425
database
No overlap
miRCancer
CpGs
NEBL
ORMDL3
LNX1
KRTDAP
KRT39
TAS2R13
PRR4
PRH1
HOPX
MIA
LTBP3
ARL14
CARD9
EPHA1
STON2
SCOC
FAM124B
CORIN
C3orf32
RRM2
TBCD
ZNF750
RUSC2
LRRC56
EPHX1
C1orf64
SLC30A4
C15orf21
SRRM3
database
No overlap
MolSigDB
Variable contributions
absolute loadings
Proteins
EBBR2
ESR1
GATA3
BRAF
CAV1
PRKCA
AR
ANXA1
CCNE1
CLDN7
EGFR
ANS
CHEK2
COL6A1
CDH3
PEA15
KIT
INPP4B
MAPK9
MTOR
ACACA
PGR
CCNB1
SCD1
CCNB1
MAPK14
BCL2L1
CDH1
ACACA ACACB
BCL2
PGR1
STAT3
CHEK2
AR1
XRCC1
NOTCH1
BAK1
ASNS1
CDK1
RPS6KB1
PDCD4
YWHAE
SMAD4
STAT5A
CTNNA1
PTEN
IGFBP2
GAB2
CTNNB1
ERRFI1
BRAF1
PDK1
PXN
KIT
KRAS
database
No overlap
DriverDBv2
MolSigDB
MolSigDB_DriverDBv2
MolSigDB_DriverDBv2_OMIM
MolSigDB_OMIM
Variable contributions
absolute loadings
b
Consensus
Ellipses: training data
Points: test data
Component 2
Component 1
Class
Basal
Her2
LumA
LumB
mRNA
miRNA
CpGs
Proteins
c
Color key
Rows
Basal
Her2
LumA
LumB
Columns
mRNA
miRNA
CpGs
Proteins
d
mRNA
miRNA
CpGs
Proteins
e
Significance (−log10 FDR)
foxm1 pathway
p53 signaling pathway
e2f mediated regulation of dna replication
g2 m dna damage checkpoint
oocyte meiosis
g1 s specific transcription
cellcycle pathway
g2 pathway
cmyb pathway
progesterone mediated oocyte maturation
mpr pathway
g1 s transition
cell cycle checkpoints
hnf3a pathway
g2 m checkpoints
nuclear receptor transcription pathway
mitotic g1 g1 s phases
recruitment of numa to mitotic centrosomes
e2f enabled inhibition of pre replication complex formation
atf2 pathway
srcrptp pathway
ptc1 pathway
era genomic pathway
akap95 pathway
pathways in cancer
e2f pathway
cyclin a b1 associated events during g2 m transition
p53 pathway
smad2 3nuclear pathway
amino acid synthesis and interconversion transamination
signaling by erbb4
prostate cancer
stathmin pathway
cell cycle
phosphorylation of the apc c
nitrogen metabolism
g0 and early g1
apc c cdc20 mediated degradation of cyclin b
bad pathway
g1 pathway
innate immune system
nuclear signaling by erbb4
cell cycle mitotic
subcollection
CP
CP:BIOCARTA
CP:KEGG
CP:REACTOME
Gene sets


a
Variance decomposition
mRNA
Metabolites
Gene modules
Cells
Metabolite modules
Pre vs Post
b
Sensitivity
100 – Specificity
AUC = 98.5%
95% CI: 95.4%–100%
AUC = 62.2%
95% CI: 40%–84.5%
Time
DIABLO
mDIABLO
c
DIABLO
mDIABLO
time
post
pre
Consensus Component 1
Consensus Component 2
d
mDIABLO_Metabolite modules
mDIABLO_Gene modules
mDIABLO_Cells
DIABLO_Metabolite modules
DIABLO_Gene modules
DIABLO_Cells
e
Color key
cells
gene.module
metabolite.module
gamma–glutamyl
genMod–Tryptophan metabolism
Renin–angiotensin system
Relative.Monocytes
Alanine and aspartate metabolism
Dipeptide
Relative.Eosinophils
Treg
Relative.Lymphocytes
Bcells
Butanoate metabolism
Fatty acid metabolism
Autoimmune thyroid disease
Allograft rejection
Type I diabetes mellitus
Graft–versus–host disease
Asthma
Relative.Basophils
Sugar, sugar substitute, starch
Food component/Plant
Vitamin B6 metabolism
Tryptophan metabolism
Valine, leucine and isoleucine metabolism
genMod–Glycerolipid metabolism
Thyroid cancer
Valine, leucine and isoleucine biosynthesis
f
Comp 1–2
Correlation cut–off
r=0.8
gene.module
metabolite.module
cells
Correlations
Positive Correlation
Negative Correlation
Expression
post
pre

Supplementary Notes

Dr. Kim-Anh Lê Cao
Senior Lecturer, Statistical Genomics
NHMRC Career Development Fellow
School of Mathematics and Statistics
Melbourne Integrative Genomics
The University of Melbourne

THE UNIVERSITY OF MELBOURNE

FACULTY OF SCIENCE

9th April 2018

Dear Editor of Genome Biology,

We wish to submit our manuscript "**DIABLO: identifying key molecular drivers from multi-omic assays, an integrative approach**" for consideration as a research article in your journal.

In the omics era, computational solutions to integrate different types of biological data measured on the same specimens or samples are trailing behind data generation. Our manuscript aims to fill this gap by proposing an efficient, flexible and easy-to-use computational framework to integrate multiple omics data generated from emerging high-throughput technologies.

The main challenge we face in multi-omics data integration is the large heterogeneity and difference in scales between omics platforms. Statistical integrative methods for biomarker discovery are still at their infancy and provide limited insight into complex biological processes. They are built on existing methods that either concatenate or combine the independent analyses from each data set, and do not model the correlation structure between the different molecular levels. This is highly problematic as important information can be missed, leading to incorrect conclusions. DIABLO maximises the correlation between data sets whilst identifying the key molecular features that explain and reliably classify a phenotype of interest. The dimension reduction process enables intuitive visualisations of the samples and selected multi-omics signatures. We benchmarked and demonstrated the ability of our method to select relevant correlated and discriminative biomarkers in a comprehensive simulation studies and in six multi-omics studies including two case studies in human breast cancer and asthma. In each of those studies we integrated various omics datasets ranging from transcriptomics (mRNA, miRNA), epigenomics (CpGs), proteomics and cell-type frequencies.

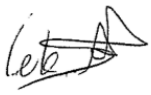
DIABLO facilitates the integration of large and heterogeneous data sets to identify relevant biomarker candidates in a wide range of biological settings. The method will be of significant interest to the scientifically diverse readership of *Genome Biology* to capitalise on fastly generated multi-omics data and push novel biological discoveries to an unprecedented level.

We are fervent advocates of open data and open science. All analyses are available in R markdown format as supplementary material, and the method is implemented in the open source R package mixOmics, with detailed tutorials on our companion website http://www.mixOmics.org/mixDIABLO.

The submitted manuscript has been approved by all authors and has not been submitted to any other journal. This manuscript is a substantial revised version to our previous submission to *Genome Biology*, **GBIO-D-16-01112**. We improved the method and added

four more case studies to benchmark the method to address the reviewers' comments. We provide a point-by-point response to reviewers in the next section. We look forward to your reply.

Yours sincerely,
Dr. Kim-Anh LÊ CAO



Reviewer #1:

The article has several strengths:

a) The article is very well written and provides a good overview of various statistical methods for analyzing

genomic data.

b) I think it presents an honest analysis of the data. The authors resist the temptation to oversell their method.

They acknowledge that their method does not outperform existing methods when it comes to accuracy.

c) The authors have implemented the method in an R package

d) This is a multi-omic method that integrates data.

e) The authors apply their method to both empirical data and to simulated data.

We appreciate the positive comments from the reviewer and the careful review. In the previous iteration of the manuscript our main focus was on the classification performance using a single breast cancer case study. However, one clear benefit of our approach is that the molecular signatures identified bring **superior biological enrichment** compared to other methods that we benchmarked on an **additional four multi-omics cancer datasets** (lung, kidney, colon and glioblastoma), each with three types of omics data (mRNA, miRNA and CpGs).

There are a few weaknesses.

The method is quite complicated and involves several parameter choices surrounding the underlying correlation structure. Why use a complicated method when simpler methods have similar predictive accuracy?

We believe our revision had addressed these weaknesses. We agree that the depiction of the method was lacking important details that led to misleading interpretations. In the revised manuscript, we have extensively benchmarked other multi-omics integration methods and demonstrated that DIABLO does not require as many parameters settings (**see Supplement**) as compared to existing methodologies. Briefly, our method requires 3 parameters, 1) number of variables to select from each omic dataset, 2) number of components to select from each omics dataset and 3) whether the correlation between certain omics datasets should be maximised (*e.g.* mRNA and miRNA). We provide a tuning function to choose parameters 1 and 2. For parameter 3 we provide guidelines that either rely on biological assumptions or a data-driven approach. Our method not only focuses on extracting the correlation structure across omic datasets but also discriminates between phenotypic groups. Such integrative approach is the first of its kind to identify molecular signatures with biological relevance and led to superior biological enrichment across various collections of gene set databases.

Why measures different types of data when a single data source (e.g. mRNA) already leads to good accuracy?

We agree with the reviewer that if the focus is on biomarker discovery, why not use the simplest, and cheapest strategy to identify biomarkers? But the focus of this manuscript is rather to capitalise on multi-omics studies and extract complementary information across omics data. Therefore, our focus is not only on identifying strong biomarkers, but also markers correlated across functional levels to give more insight into disease mechanisms. Therefore, we have changed the title to "**DIABLO: identifying key molecular drivers from multi-omic assays, an integrative approach**", to reflect the focus on key molecular drivers rather than biomarkers only.

I am not convinced that the method helps to elucidate the underlying biology. I understand that the latent structure might uncover interesting biology but I would never use this method to learn biology. Rather, I would use cluster analysis or unsupervised learning methods.

In the previous version of the manuscript we did not compare the biological enrichment of the various methods that were used. However, based on the reviewers' comments, we extensively explored this area, using multiple cancer multi-omics datasets, multiple gene-set databases with both unsupervised and supervised integrative methods that can perform variable selection. We demonstrate that our method outperforms unsupervised methods with respect to biological enrichment thus elucidating more known biology. In the human breast cancer study, we show that DIABLO can also detect novel biomarkers that have not been previously associated with breast cancer.

We also researched the literature to give an overview of the current state in integrative methods either supervised or unsupervised, and with or without variable selection, to highlight where the gaps are in terms of methods development (**see Supplementary Fig. 1**). In the revised version of the manuscript we have included unsupervised methods used for multi-omics data integration as well as supervised multi-step approaches.

Overall, I am not sure how much biology can be learnt by applying this method. Bottom line: this predictive method does not seem to improve predictive accuracy.
We have refocused our manuscript on biological insights primarily, rather than prediction performance as the former was our main motivation in driving methodological developments. By extending our analyses with six multi-omics studies including two case studies in human breast cancer and asthma we believe we have demonstrated that data integration performed using appropriate computational methods generate new biological insights and novel hypotheses to be further tested in the laboratory. The important contribution of DIABLO is its resulting molecular signatures that both explain the correlation structure across multiple biological domains and discriminate multiple phenotypic groups, with increased biological enrichment compared to other methods.

Reviewer #2:
This is a well written article which addresses an important need in the field. 1) In the introduction, the longer intro to sparse CCA should be provided. In the methods the actual method is more clearly stated "DIABLO extends sparse gCCA to a classification framework".
We thank the reviewer for their appreciative comments. In the revised version of the manuscript we provide a clearer explanation of our method DIABLO (see lines 99-116).

*"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) datasets while identifying the key omics variables (mRNA, miRNA, CpGs, proteins, metabolites, etc.) and characterizing the disease sub-groups or phenotypes of interest. DIABLO uses Projection to Latent Structure models (PLS) [1], and extends both sparse PLS-Discriminant Analysis [2] to multi-omics analyses and sparse Generalized Canonical Correlation Analysis [3] to a supervised analysis framework. In contrast to existing penalized matrix decomposition methods [4], DIABLO is a component-based method (or a dimension reduction technique) that transforms each omic dataset into latent components and maximizes the sum of pairwise correlations between latent components (user-defined) and a phenotype of interest [5]. DIABLO is, therefore, an integrative classification method that builds predictive multi-omics models that can be applied to multi-omics data from new samples to 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. The method 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 [6] instead of the original omics matrices, as illustrated in*

one of our case studies."

The mathematical formulas such as the sGCCA algorithm, and its extension to a discriminant framework is detailed in the Methods section.

2) Can the approach can handle missing data, that is missing row or column observations or is it only missing datasets. I presume, the later, as the intersection of tumors with complete data was used in training real data. This is important and should be made clear in the intro, abstract and discussion.

Currently, our method does not account for completely missing observations or variables.
Random missing values are allowed in the dataset matrices as local regressions are fitted in the model and missing values will be omitted when calculating the latent components and loading vectors. The prediction step however, as highlighted in the Breast Cancer case study can be performed with an entire dataset missing.
l 401: '*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.*'
l209: '*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.*'
As such we do not think this is an information that should appear all throughout the document.

3) Can PLS DA be applied to multi class classification. Was this tested?

Yes, the revised version of the manuscript uses sparse Partial Least Squares Discriminant Analysis (sPLS-DA), in various multi-step classification schemes such as concatenation and ensemble-based schemes. Generally speaking sPLS-DA can handle multiple classes (Lê Cao et al., 2011, BMC Bioinformatics 22:253). However, this is not highlighted in this study as we used sPLS-DA for the cancer benchmark data sets that only include 2 classes.

4) The order of pair comparisons appears important. (discussion page 17, 18 and methods).
Those unfamiliar with their data may specify a suboptimal Design Matrix. Could there be some tools that provide guidance? For example, multiple factorial analysis or one of many tensor decompositions could be used to compute an RV coefficient. Alternative, can datasets be weighted in the analyses? In multi dataset approaches, data are often weighted by quality/size, the first eignenvector etc (reviewed by Meng et al., Brief Bioinform (2016) doi: 10.1093/bib/bbv108). If data has a batch effect, and this data were used to seed the analysis (aka in the first pair of data analyzed) , would that skew the results ? Could this please be tested.

We thank the reviewer for their suggestions In fact, there is no order to the pairwise comparisons in the DIABLO framework: it is the **sum of pairwise correlations** that is maximized, **see Methods**. Therefore the pairwise correlations are considered simultaneously in the SGCCA algorithm[3].

In the revised manuscript we got inspiration from the multiblock literature such as multiblock partial least squares (MBPLS[7]) whose datasets (also called blocks) are weighted based on their correlation with the response variable. In the new DIABLO implementation, we have used a weighted majority vote scheme based on the correlation between the latent component of each omics dataset with the latent component from the response matrix. This has significantly improved our classification error rates, as the strongest discriminatory datasets is given a higher weight in the overall class prediction for a new sample. Further, the weighted majority vote option in our function overcome the case where an equal number of voting classifiers and no consensus can be achieved.

In the discussion, we underlined the influence of batch effects on the multivariate modelling performed by our method (see lines 322-327), as this is outside the focus of this manuscript (but developments are in progress for other types of data).

"Finally, DIABLO, like other methods we benchmarked, will be affected by technical artifacts of the data, such as batch effects and presence of confounding variables that 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 integration analysis [8–10]*."*

5) On page 8, "validation of the Diablo methods on synthetic data" . Three different criteria are explored 1) CorNonDis 2) CorDis 3) NonCorDis. Please explain the rational behind nonCorDis should be explained. In a 2 class system, methods such as CCA or PLS extract eigenvectors of correlated variables. Therefore a discriminate eigenvector will represent a set of correlated variables. Gene expression and 'omics data, measure genes which work in pathways, and therefore data has considerable correlation structure. Discriminatory non-correlated vectors, may reflect system noise.

The rationale for including four types of variables was to determine the influence of the correlation structure between datasets as well as discrimination between phenotypic groups. This is why the simulation framework includes different combinations of discrimination (discriminatory, non-discriminatory) and correlation (correlated, uncorrelated) variables. Unlike CCA and PLS which can only maximize the correlation between at most two data matrices, DIABLO can simultaneously maximize the correlation between any number of data matrices (**see Methods**). In our previous simulation study, we generated four types variables by controlling the correlation between variables or discrimination between groups. In our revised version, we have instead generated the correlation structure first by controlling the different relationships between latent components of different datasets (**see Supplementary Fig. 2**). The latent components are than used to compute the four-types of variables based on different correlation structures (**see Supplement for complete details**). The relevant variables include 30 corDis (correlated and discriminatory) and 30 unCorDis (uncorrelated and discriminatory) variables, in order to determine the effect of the design matrix on the types of variables selected. We also simulated 100 corNonDis (correlated and non-discriminatory) and 100 unCorNonDis (uncorrelated and non-discriminatory) variables. Therefore it is the corNonDis and unCorNonDis variables that represent noise and irrelevant variables, although by chance some of these variables might be correlated with the response. The purpose of the simulation was to determine whether any of multi-step classification schemes and DIABLO model happen to (wrongly) select these irrelevant variables (corNonDis, unCorNonDis) and relevant variables (corDis, unCorDis).

6) please provide a discussion on filtering data. In each case, data were filtered and reduced. Is this to reduce "noise" or for computational efficiency. Please discuss and comments on the computational cost of larger datasets.

We provide a discussion about filtering data in the revised version of the manuscript (lines 319-320). *"...we advise using a broad filtering strategy to alleviate computational time when dealing with extremely large datasets (e.g. > 50,000 features each)."* However, for this revised manuscript, we have not performed any filtering for the benchmark datasets and retained all the variables that were downloaded from their respective websites. For the breast cancer case study, some filtering was involved to remove low abundance variables, mostly to reduce some amount of noise rather than saving on computational time. We also provide additional guidance on filtering in the mixOmics article [11], which we refer to in this manuscript. **Table 1** lists the size of the datasets we analysed.

7) On page 11 the acronym BER (balanced error rate) is used before it is defined.

Since we do not use the acronym BER many times in the revised manuscript, we have removed it altogether and explicitly stated 'balanced error rate'.

8) on Page 13, please describe "eigengene summarizaton" in more detail. Please describe how to interpret the results, saying it is "common approach" on page 29 is insufficient;

We provide a description of eigengene summarization in the revised manuscript (lines 504-509).

*"**Modular analysis:** Eigengene summarization is a common approach to decompose a n x p dataset (where n is the number of samples and p is the number of variables in a module), to a component (linear combination of all p variables) that represents the summarized expression of genes in the module [6]. For the asthma study, 15,683 genes were reduced to 229 KEGG pathways and 292 metabolites were reduced to 60 metabolic pathways using eigengene summarization."*

9) page 16. Why were 9 variables (36 in total) selected in analysis of the BRCA data? Is there any guidance as to how many variables should be selected. For example, mRNA and protein were more informative in gsea, therefore it might be better to select more variables from these datasets?

Yes, we provide a tuning function which is implemented along with DIABLO in the mixOmics R package [11]. The function uses a grid approach to select an optimal number of variables to select from each omics dataset. A section on parameter tuning discusses the grid approach to identifying the optimal number of variables and components to select (lines 438-456).

*"Finally, the third set of parameters to tune is the number of variables to select per dataset and per component. Such tuning can rapidly become cumbersome, as there might be numerous combinations of selection sizes to evaluate across all K datasets. For the breast cancer study, we used 5-fold cross-validation repeated 50 times to evaluate the performance of the model over a grid of different possible values of variables to select (**Supplementary Fig. 8**). The performance of the model for a given set of parameters (including number of component and number of variables to select) was based on the balanced classification error rate using majority vote or average prediction schemes with centroids distance. The balanced classification error rate is useful in the case of imbalanced class sizes, where the majority classes can have strong influence on the overall error rate. The balanced error rate measure calculates the weighted average of the individual class error rates with respect to their class sample size. In our experience, the number of variables to select in each dataset provided less of an improvement on the error rate compared to tuning the number of components. Therefore, even a grid composed of a small number of variables (<50 with steps of 5 or 10) may suffice as it does not substantially change 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 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."*

10) page 17, 3rd line from top. "known cell-types and pathwaysin the" typo. Insert space

Thank you, the typo has been corrected in the revised version of the manuscript.

11) The statement on p 18, "To our knowledge DIABLO is the only integrative classification methods that models the correlation structure between omics data" is inaccurate. See the Meng et al., review. There are many many methods that use dimensions reduction to extract correlated structure in multiple 'omics data. Also Jeffrey et al., 2007 Bioinformatics. 2007 Feb 1;23(3):298P305. Epub 2006 Nov 24. used discriminative analysis with coinertia analysis and described a supervised integrative latent variable approach, which is related to this work but did not employ sparse methods.

We have removed this erroneous statement from the revised manuscript, thank you.

12) In the methods, p21. In the abstract/intro, the method is describes as a supervised PLS/DA approach, but on p21 it appears to be cluster to partitioning around centroids, with majority voting. Please describe the approach clearly and consistently.

In the revised version of the manuscript we have added additional details regarding the development of DIABLO, which extends sparse Generalised Canonical Correlation Analysis. We believe the confusion from the reviewer may come from the fact that multiple-types of prediction distances can be used in DIABLO such as centroids, max distance, and Mahalanobis distance. Please see lines 347-383 for a general description of the sGCCA algorithm, lines 386-397 for the classification implementation and lines 399-417 for the implementation for the different types of error rate that can be computed.

13) is the analysis effective by the number of variables. For example if dataset A has several thousand variables and dataset B has less than 50, would this impact the analysis?

The difference in the number of variables in each dataset should not impact the analysis as each dataset is summarised by its own set of latent components, so that components across data sets are maximally correlated, irrespective of how many variables there are in each dataset. This makes DIABLO a much more attractive solution than a concatenation method, where datasets than include a large number of variables tend to be more 'favoured' in the molecular signature compared to smaller data sets.

14) p25, The de-duplication effort in GSEA is important and should be clear to users, If a more stringent assignment were used, would this impact results?

GSEA is impacted by the number of features that are input into the analysis and the types of gene sets that are used to determine biological enrichment. In the revised manuscript we include a benchmarking experiment where we constructed multi-omic biomarker panels of equivalent number of features with a total of 180 features. Further we tested 10 different gene set databases, from Molecular signature database[12], blood transcriptional modules[13] and cell-specific expression from Benita *et al*. [14].

15) p 27. Data Processing. Were 3,073 BRCA clinical variables used in this study? The PAM50 assignments for tumors (obtained from TCGA staff) should be made available together with the filtered TCGA data, such that others can reproduce this work.

The 3,073 variables listed describe the data that were obtained from TGCA. From the clinical data, only the PAM50 labels and sample-type variables were used. The complete code and data files can be found with the github repository (https://github.com/singha53).

16) p28. Terms in the Voom equation are not fully defined. Filtering removed "genes with counts less than 0". Does this mean the sum of the gene across all tumors was zero, or that any gene which has a zero tumor in any 1 tumor was excluded?

We have clarified the following in the revised manuscript (**see Supplementary Data file**). The count data for the mRNA dataset, X_{counts} was normalized to log2-counts per million (logCPM), X_{norm}, similar to limma voom [15]:

$$X_{norm} = \log_2\left(\frac{\left(X_{counts} + 0.5\right)^T}{\left(lib.size + 1\right) * 10^6}\right)$$

After library size (lib.size = total number of reads per sample) normalization, genes with counts less than 0 in more than 70% of samples were removed. The PAM50 genes were also removed from the mRNA dataset prior to analyses. Similarly, the miRNA count data was normalized to logCPM and miRNA transcripts with counts less than 0 in more than 70% of the samples were also removed.

17) p28 Asthma study. Genes were reduced 229 KEGG pathways and metabolites were reduced to 60 pathways. Why were variables reduced to GeneSet. The rational and need for this is not

explained. Was it simply to aid biological interpretation of the data or was it for computational reasons?

For this specific case study we wished to incorporate modular-based analyses within the DIABLO framework to focus on pathways spanning common biological mechanisms that significantly changed in response to allergen inhalation challenge. The purpose of this analysis was only to aid in the biological interpretation and the reduction to gene sets was not performed for computational reasons. A secondary reason for including this approach was to demonstrate to potential users the benefits of combining modular-based analyses with the DIABLO framework. Other types of approaches that identify modules such as data-driven techniques like WGCNA (weighted gene co-expression networks) may also be incorporated with the DIABLO framework, since each cluster of variables can be reduced to a single variable that explains the entire cluster of features.

18) Figure 1 A) not clear if concatenation is performed on genes or tumors (rows/cols). C) The DIABLO diagram is confusing. it is not clear that DIABLO is a pairPwise approach.

The figure has been updated to clarify the integration and classification aspects of the DIABLO framework (**see Supplementary Figure 3**). Each dataset is a n x p_j matrix, where p_j is the number of variables (columns) for the j^{th} dataset. For the concatenation-based analysis, the datasets are combined row-wise since the number of samples are the same for each omics dataset, that is, the multi-omics data is obtained for the same set of samples.

Although DIABLO computes the pairwise correlation between latent components of pairs of omic datasets, similar to PLS and CCA, its objective is to maximize the sum of pairwise correlation between different omics datasets (see objective function in **Methods**), see our earlier answer to Question 4.

19) Figure 4 legend. DIABLO 1P12 are not defined, What was the difference between these models.

In the previous version of the manuscript, the concatenation and ensemble biomarker panels were tuned such that each panel consisted of a specific number of variables and DIABLO panels matched the same number of variables to keep the comparisons consistent. However, given this extra confusion, we have added 4 benchmark datasets where each method selects the same number of variables of each omic-type.

20) Figure 5. There is no scale on the ciros plot (gene level) which makes its interpretation difficult. Also please add Gene Names to the heatmap (E)

The purpose of the circosplot is to depict the inter-correlations between omic datasets. These can be observed from the red (positive) and blue (negative) lines between omics datasets (different colors). The scale for the lines surrounded the ideogram is not depicted as it is centered at zero, therefore the line height represent the average expression levels of a given variable in a given phenotypic groups compared to others. Gene names have not been added to the heatmap, due to size limitation of the figure. However, the feature plot in Figure 3a lists all the features selected by the multi-omic biomarker panel.

21). Reference 38 Gauvreau et al., is in upper case

All references have been checked and the capitalization has now been fixed.

22) Please provide more details on the computational complexity of the method as 1) the number of variables increases 2) the number of datasets increased 3) the impact of correlated datasets (eg microarray and RNAseq)

We decided to focus our comparisons mainly on the correlation and discrimination structure between datasets (simulation study), the effect of the design matrix on the types of variables selected and whether this led to superior biological enrichment compared to other integrative strategies

(benchmark study, four new real multi-omics datasets). Computational times are provided for different scenarios in our article (Rohart et al, 2017, Plos Computational Biology 13 [11] in the main and supplemental material) for various numbers of variables and data sets, see the screenshots below.

Table 2. Example of computational time for the data sets presented in the Results section with a macbook pro 2013, 2.6GHz, 16Go Ram.

Framework	Single 'omics sPLS-DA		N-integration DIABLO		P-integration MINT	
Data	srbct		breast.tcga		stemcells	
N	63		150		125	
P	2,308		200; 184; 142		400	
function	`tune`	`perf`	`tune`	`perf`	`tune`	`perf`
#fold CV (repeated)	5(10)	5(10)	10(1)	10(10)	LOGOCV	LOGOCV
ncomp	6	3	2	2	2	2
grid length per component	39	-	13^3	-	100	-
#cpu	1	1	2	1	1	1
run time	9min	31sec	18min	25sec	30sec	0.2sec

Figure 1. Computational time for various mixOmics methods, including DIABLO, as published in [11] in the main article.

Table 4: Example of runtime for very large data sets analysed in `mixOmics`. Tuning and performance assessments were performed with 5-fold CV for single 'omics and N-integration, or LOGOCV for P-integration (Rohart et al. 2017, Singh et al. 2016, cluster with 10 cpus and 50 Gb RAM).

Framework	Single 'omics sPLS-DA		N-integration DIABLO		P-integration MINT	
Data	HNSCC		Asthma (2 omics)		Stem Cell (8 studies)	
N	60		194		210	
P	82, 132		30, 000; 30, 000		13, 313	
function	`tune`	`perf`	`tune`	`perf`	`tune`	`perf`
#fold CV (repeated)	5(10)	5(10)	5(1)	5(10)	LOGOCV	LOGOCV
ncomp	5	3	2	2	2	2
grid length per component	40	-	22^2	-	100	-
#cpu	10	10	10	10	1	1
runtime	15min	6min	19min	3min	17min	12sec

Figure 2Computational time for various mixOmics methods, including DIABLO, as published in [11] in the supplement material.

References

1. Wold H. Estimation of principal components and related models by iterative least squares. Multivar Anal. 1966;391–420.

2. Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics [Internet]. 2011 [cited 2015 Jul 15];12:253. Available from: http://www.biomedcentral.com/1471-2105/12/253/

3. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. Biostatistics [Internet]. 2014 [cited 2015 Jul 15];15:569–83. Available from: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxu001

4. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics [Internet]. 2009 [cited 2016 Jul 27];10:515–34. Available from: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxp008

5. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res [Internet]. 2004 [cited 2016 Mar 30];14:1085–1094. Available from: http://genome.cshlp.org/content/14/6/1085.short

6. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics [Internet]. 2008 [cited 2016 Apr 4];9:559. Available from: http://www.biomedcentral.com/1471-2105/9/559

7. BOUGEARD S, QANNARI EM, LUPO C, HANAFI M. From Multiblock Partial Least Squares to Multiblock Redundancy Analysis. A Continuum Approach. :16.

8. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics [Internet]. 2007 [cited 2016 May 12];8:118–27. Available from: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxj037

9. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics [Internet]. 2012 [cited 2018 Mar 6];13:539–52. Available from: https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxr034

10. Parker HS, Corrada Bravo H, Leek JT. Removing batch effects for prediction problems with frozen surrogate variable analysis. PeerJ [Internet]. 2014 [cited 2016 May 12];2:e561. Available from: https://peerj.com/articles/561

11. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Comput Biol [Internet]. 2017 [cited 2018 Jan 29];13:e1005752. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752

12. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst [Internet]. 2015 [cited 2018 Jan 30];1:417–25. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2405471215002185

13. Chaussabel D, Baldwin N. Democratizing systems immunology with modular transcriptional repertoire analyses. Nat Rev Immunol [Internet]. 2014 [cited 2016 Jul 22];14:271–80. Available from: http://www.nature.com/doifinder/10.1038/nri3642

14. Benita Y, Cao Z, Giallourakis C, Li C, Gardet A, Xavier RJ. Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT repressor. Blood [Internet]. 2010 [cited 2018 Mar 5];115:5376–84. Available from: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2010-01-263855

15. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol [Internet]. 2014 [cited 2016 Mar 2];15:R29. Available from: http://www.biomedcentral.com/content/pdf/gb-2014-15-2-r29.pdf