DIABLO – an integrative, multi-omics, multivariate method for multi-group classification

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

[1]UBC James Hogg Research Centre for Heart Lung Innovation, St. Paul's Hospital, University of British Columbia, Vancouver, BC, Canada.
[2]Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, BC, Canada.
[3]The University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, QLD 4102, Australia
[4]Australian Research Council Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, Western Australia, Australia
[5]Department of Medicine (Respiratory Division), University of British Columbia, Vancouver, BC, Canada.

Corresponding author:
Kim-Anh Lê Cao
The University of Queensland
Diamantina Institute, Translational Research Institute
37 Kent Street, Woolloongabba, QLD 4102, Australia.
Tel.: +61 7 3443 7069
Fax: +61 7 3443 6966.

Abstract

Rapid advances in technology have led to a wealth of large-scale molecular omics datasets. Integrating such data offers an unprecedented opportunity to assess molecular interactions at multiple functional levels and provide a more comprehensive understanding of the biological pathways involved in different diseases subgroups. However, multiple omics data integration is a challenging task due to the heterogeneity in the different platforms used. There is a need to address the complex and correlated nature of different data-types, in order to identify a robust and reliable multi-omics signature that can predict a phenotype of interest.

We introduce a novel multivariate dimension reduction method for multiple omics integration, classification and identification of a multi-omics molecular signature. DIABLO - Data Integration Analysis for Biomarker discovery using a Latent component method for Omics studies, models the correlation structure between omics datasets, resulting in an improved ability to associate biomarkers across multiple functional levels to phenotypes of interest. We demonstrate the capabilities of DIABLO using simulated data and studies of breast cancer and asthma, integrating up to four types of omics datasets to identify relevant biomarkers, while still retaining competitive classification and predictive performance compared to existing methods.

Our statistical integrative framework can benefit a diverse range of research areas with varying types of study designs, as well as enabling module-based analyses. Importantly, graphical outputs of our method assist in the interpretation of such complex analyses and provide significant biological insights.

Keywords: Systems biology, biomarker discovery, data integration, data visualization, PAM50, asthma, classification, breast cancer

Background

Systems biology approaches combine information from different biological components in order to unravel complex processes involved in health and disease [1,2]. Such biological processes are comprised of interactions between different biological layers such as the genome, methylome, transcriptome, proteome and metabolome. These interactions are often missed when each omics level is studied in isolation, leading to an increased number of false positives, loss of information (false negatives) and irreproducible findings [3]. The advent of technological advances coupled with decreasing experimental costs have however made it possible to obtain multiple high dimensional omics datasets, such as transcriptomics, proteomics, and metabolomics, for the same group of individuals or biological samples. Systems approaches including multivariate approaches [4], Bayesian methods [5], and network analyses [6] have been used to combine data originating from different biological layers with the aim to provide a holistic and accurate depiction of molecular processes within biological systems. For example, multivariate methods such as multiple co-inertia analysis (MCIA), used to integrate gene and protein expression of the NCI-60 cell line, revealed pathways not uncovered by single-omics analyses [7]. Bayesian network algorithms have been used to integrate datasets consisting of expression, variation and interaction data from yeast, and identified causal regulators of networks and novel biological mechanisms of expression quantitative trait loci (eQTL) hot spots [8]; condition-specific regulatory networks produced using multiple datasets have been shown to be more accurate than using individual datasets [9]. Others have used modular approaches that reduce high dimensional data to modules (clusters) representing distinct functional processes [10,11]. The next logical step was to determine whether the identified interactions, modules, and networks differed between different disease conditions [12,13]. To that end, differential modular or network analyses were proposed to determine whether modules or networks were statistically different between groups. However, such inferential methods are not predictive models and, as such, cannot be used to classify new subjects into different phenotypic groups.

Machine learning algorithms construct predictive models by “learning” from the data a classification rule that is then used to predict or assign the class membership of new individuals. Common classification methods include discriminant analysis, neural networks, decision trees, support vector machine (SVM) and random forest (RF) to name a few, where performance is assessed using indices such as the classification error rate, prediction accuracy, and the area under the receiver operating curve. A study comparing 176 classifiers showed that RF and SVM led to superior performance accuracy [14]. Linear penalized regression models such as elastic net have been also proposed to simultaneously perform shrinkage and variable selection, thereby resulting in a parsimonious linear model with improved predictive performance [15]. These methods are suited to single dataset analyses, whereas methods that can construct predictive models from multiple high dimensional omics datasets are required as multiple sources of information captured via different data-types becomes available for the same individuals.

Omics data integration as defined by Ritchie *et al.* [16] refers to combining multiple omics datasets in order to develop multivariate models that are predictive of complex traits or phenotypes. However, the task is non-trivial given the numerous analytical challenges. The high dimensionality of each dataset requires not only efficient computational techniques, but also the development of novel methods which are able to identify relevant information; a ‘multi-omics molecular signature’ from the tens of thousands of predictors that are measured [17]. Additionally, the small number of samples compared to the large number of predictors limits statistical power and accuracy of current methods, which may lead to non-reproducible molecular signatures. Finally, while the same samples are profiled within a single study, the different omics platforms employed have their own inherent platform-specific artifacts such as variation between manufacturers and omics technologies. Data heterogeneity is therefore a major obstacle to combining multiple omics studies [17].

Current data integration frameworks enabling the identification of multi-omics molecular signatures in a data-driven analysis include concatenation-based [18] and model-based integration (*e.g.* ensemble classifiers) [19] (**Figure 1A-B**). Concatenation-based integration combines multiple datasets into a single large dataset, with the aim to predict a phenotype of interest. Model-based integration approaches such as ensemble classification construct a predictive model on each individual dataset before combining the model predictions. None of these approaches however account or model relationships between datasets and thus limit our understanding of molecular interactions at multiple functional levels. Therefore, there is a crucial need for novel integrative modeling methods, that can identify a multi-omics molecular signature by borrowing discriminatory strength from complementary information, across multiple functional levels while providing greater insight into disease mechanisms.

We introduce a multivariate dimension reduction discriminant analysis method, DIABLO (Data Integration Analysis for Biomarker discovery using a Latent component method for Omics studies, **Figure 1C**) as part of the mixOmics Data Integration Project (http://mixomics.org/) [20,21]. DIABLO aims to maximize the common or correlated information between multiple datasets whilst identifying in an optimal manner the key omics variables (mRNA, miRNA, CpGs, proteins, metabolites, etc.) that explain and reliably classify disease sub-groups or phenotypes of interest. DIABLO builds on Projection to Latent Structure models (PLS) [22], substantially extends both sparse PLS-Discriminant Analysis [23] to multi-omics analyses and sparse generalized canonical correlation analysis [24] to a discriminant analysis framework. In contrary to existing penalized matrix decomposition methods [25] DIABLO models and maximizes the correlation between pairs of pre-specified omics datasets to unravel similar functional relationships between those omics data [26]. In addition, DIABLO provides appealing features by 1) allowing the user to specify the number of variables to select from each dataset 2) constructing a predictive multi-omics model that can be applied to classify new samples even if some datasets are missing, and by 3) allowing for the assessment of the classification performance of the predictive model. The dimension reduction process enables visualization of the samples, as well as biologically relevant variables. DIABLO is a highly flexible method that can handle classical single time point experimental designs, as well as cross-over or repeated measures study designs. Modular-based analysis can also be used in conjunction with DIABLO by inputting pathway-based module matrices [11] instead of omics matrices.

We demonstrate the ability of DIABLO to select relevant, correlated and discriminatory biomarkers, using synthetic data as well as multi-omics datasets from human breast cancer and asthma case studies. In those studies, we integrate up to four omics datasets and show that DIABLO has competitive classification performance with existing single-omics methods and multi-omics integrative frameworks. Importantly, DIABLO yields improved biological insights of multi-omics signatures as we demonstrate in both case studies.

Results

The mixDIABLO integrative framework

We describe the mixDIABLO pipeline in **Figure 2** to integrate multiple omics datasets and identify a multi-omics biomarker panel, assess the predictive performance of the model, and generate visualizations to aid in the interpretation of the results. The first step inputs multiple omics datasets measured on the same individuals, that were previously normalized and filtered with optional preprocessing steps such multilevel transformation (for repeated measures study designs) and module-transformations (for Pathway analyses, as described in **Methods**). Prior to multivariate data integration, exploratory and unsupervised data analyses of each omics dataset with Principal Component Analysis (PCA), or sparse PCA built only on a smaller subset of variables [27] can be useful to visualize and understand the major sources of variation in each dataset to be integrated.. Then, the omics datasets, along with the phenotype information indicating the class membership of each sample (two or more groups) are input in the multivariate integrative method DIABLO. DIABLO is a multivariate dimension reduction method that seeks for latent components – linear combinations of variables from each omics dataset, that are maximally correlated as specified in a design matrix. The design matrix indicates which datasets should be connected such that their pair-wise correlations are maximized (**Figure 1C**). The design can be determined according to *prior* knowledge (*e.g.* mRNA and miRNA datasets can be assumed to be connected since miRNAs regulate mRNA expression), or using our proposed data-driven approach that indicates when to connect pairs of datasets (**see Methods**). The identification of a multi-omics panel is performed via l_1 penalties that shrink the variable coefficients defining the latent components to zero (**see Methods**). The performance of the DIABLO model and associated multi-omics panel is then assessed using cross-validation repeated several times to ensure reliable evaluation and the balanced error rate (BER) or area under the receiver operating curve (AUC, for two groups) are reported. Lastly, numerous visualizations are proposed to provide insights into the multi-omics panel and guide the interpretation of the selected omics variables, including sample and variable plots (**see Methods**).

Validation of the DIABLO method on synthetic data

An extensive simulation study was performed to numerically validate the DIABLO classification performance method and its ability to identify a highly correlated and discriminatory signature, as well as investigate the impact of the design matrix on the selected variables (**Additional file 1**). Briefly, two design matrices were tested; the full design (all datasets were connected) and the null design (where no datasets were connected, **see Methods**). Three datasets were generated with equal numbers of observations (n=100, divided into two sample groups) and 150 variables. Amongst the 150 variables, 100 variables were deemed 'irrelevant', i.e. not correlated between datasets and not discriminatory, whereas the 50 remaining variables were either 1) correlated across all three datasets but not discriminatory between groups (called *CorNonDis*), 2) correlated and discriminatory (*CorDis*) or 3) not correlated but discriminatory (*NonCorDis*). Several scenarios were considered to simulate these three datasets, such as different fold-change values between the two sample groups (varied from 0 to 1.5) as well as levels of noise. Furthermore, the strength of correlation was varied based on specified variance-covariance matrices (see details in **Additional file 1**). Simulated datasets were generated 20 times for each type of variable and DIABLO was applied with the full and null design selecting 50 variables with one component. We used 10x5-fold cross-validation to evaluate the performance of the integrative model on the different simulated scenarios.

As expected DIABLO with a full design correctly identified a greater proportion of *CorNonDis* variables compared to a null design (**Figure 3A**). The difference increased further with the correlation strength between the variables. DIABLO with a full design also selected a greater proportion of *CorDis* variables compared with a null design, however this difference decreased as the fold-change increased, while no difference was found between the full and null designs when *NonCorDis* variables were simulated. Interestingly, we observed very similar classification error rates between the full and null design for the *CorDis* and *NonCorDis* variables (**Figure 3B**). The error rate was lower when the datasets contained *NonCorDis* instead of *CorDis* variables. As expected, when DIABLO was applied to datasets including only *CorNonDis* and irrelevant variables we observed a random prediction of the model (error rate ~ 50%).

In summary, the design matrix is an important parameter in the model, as it affected the types of variables selected by DIABLO. The purpose of the full design is to select highly correlated variables regardless of their discriminatory power whereas the null design selects discriminatory variables regardless of the correlation structure between variables. The error rate was similar in both designs, although the presence of highly correlated and discriminatory variables led to a slight increase in the error rate.

Comparisons with existing single-omics classifiers and multi-omics integrative classifiers using human breast cancer data

The subtypes of breast cancer (Luminal A, Luminal B, Her2-enriched and Basal-like) [28] have been the most replicated subtypes of human breast cancer [29] and a risk model based on the expression levels of 50 genes (PAM50) has been shown to successfully predict breast cancer subtypes [30]. This biomarker panel has been developed using the NanoString platform, called the Prosigna^{TM} test and has been approved by the Food and Drug Administration (FDA) [31]. We integrated human breast cancer datasets (mRNA without PAM50 genes, miRNA, methylation and proteins) from The Cancer Genome Atlas (TCGA) [32] in order to achieve a systems characterization of PAM50 breast cancer subtypes with other types of omics datasets. We compared the classification performance of single-omics methods, integrative concatenation and ensemble-based approaches and DIABLO. The TCGA study was divided into a training and test set, where the training set was determined so as to include all samples in the proteomics dataset (**see Methods**). Most of the training samples were obtained in 2010, whereas the test samples were mainly obtained from 2011 to 2013 (**Table S1, Additional file 2**). The training set consisted of 379 subjects (76-Basal, 38-Her2, 188-LumA, and 77-LumB) with four omics datasets, whereas the test set consisted of 610 subjects (102-Basal, 40-Her2, 346-LumA, and 122-LumB) with only three omics datasets (mRNA, miRNA, and methylation). The omics datasets consisted of 2,000 mRNAs, 184 miRNAs, 2,000 CpG probes, and 142 proteins (**see Methods for preprocessing of datasets**). We checked that the range of expression values within each omics dataset was consistent between the training and test set (**Figure S1, Additional file 3**).

For single-omics analyses, the penalized regularization method [15] Elastic net (Enet), random forest (RF) [33] and support vector machine (SVM) [34] were used to identify biomarker panels for each omics dataset (mRNA, miRNA, CpGs and proteins), respectively. In Enet the sparsity parameter (lasso penalty) was set to 1 to determine the smallest possible biomarker panel, whereas RF and SVM do not perform variable selection and retained all variables (**Figure 4A**). Multi-omics biomarker panels (with equal number of variables from each omics dataset) were constructed using DIABLO such that the total number of variables was similar to the single-omics biomarker panels. The classification performance of single-omics biomarker panels was compared with DIABLO by using a 50x5-fold cross validation in the training set (**Figure 4B)**. Enet gave the best performance for the mRNA panel (BER = 12.8±0.9%), however, DIABLO out-performed all single-omics methods for the other types of omics panels; miRNA, CpGs, and proteins panels, with respective training BERs of 14.2±1.9%, 14.1±1.3% and 13.0±1.8%.

For the integrative methods we applied Enet, SVM and RF in the concatenation or ensemble frameworks. For similar size panels, and on the training set, Concatenation-Enet and Ensemble-Enet led to the lowest BERs (11.4±1.1% and 11.9±1.1% respectively) compared to the DIABLO panels (DIABLO9 and DIABLO11, BER = 16.4±2.1% and 13.5±1.7% respectively, **Figure 4B**). SVM and RF performed better than DIABLO using the Concatenation framework, but worse when used with the Ensemble framework. The main limitation of the Concatenation method is that all omics datasets must also be available in the test set, which in this study is missing the proteomics data. In addition, we observed that the Concatenation-based panel was heavily biased towards variables from the most discriminatory mRNA dataset (**Figure 4C**). As such, the Concatenation-Enet multi-omics signature consisted of 62% mRNAs and 4% miRNAs, 29% CpGs and 5% proteins. **Figure 4D** shows a large number of unique non-overlapping set of features between the methods. Most importantly, when examining the correlation between variables identified by each method as displayed in the circos plot (**see Methods**) we observed substantially fewer inter and intra-associations in the Concatenation and Ensemble-based approaches compared to DIABLO (**Figure 4E**). Therefore, the multi-omics biomarker panel selected by DIABLO was not only predictive of breast cancer subtypes, but also included highly correlated molecular features spanning different biological layers. Interestingly, although DIABLO did not substantially out-perform existing methods, it also did not under-perform. For example, out of all panels in **Figure 4**, the best performing panel was RF including all 2000 mRNA transcripts with a training and test performance of 13.0±1.1% and 12.0%. The second best performing panel was the DIABLO7 panel with a training and test BER of 13.0±1.8% and 12.2%, based on 60 features (15 from each omics dataset). Therefore, we conclude that DIABLO performs competitively with current methods with an enhanced focus on selecting discriminatory and correlated multi-omics variables.

Multi-omics biomarker panel predicts the PAM50 breast cancer subtypes

We demonstrate our mixDIABLO pipeline presented in **Figure 2** to identify a multi-omics biomarker signature predictive of the PAM50 human breast cancer subtypes and determine its biological significance (**Figure 5A**). The path diagram of **Figure 5A**, was determined by using a correlation cut-off 0.8 for between dataset pair-wise correlations (**Figure S2, Additional file 4**) The panel size of the Enet classifier, which is dependent on the elastic net penalty (Lasso penalty set to 1), remained quite large (hundreds of variables). DIABLO on the other hand can fit a very sparse model that contains only a few variables in each dataset. **Figure 5B** shows the tuning step to set the optimal number of variables in each dataset to be selected with a minimum BER (**see Methods**), resulting in a multi-omics panel of 9 variables selected from each dataset (BER = 13.4±1.5%). **Figure 5C** shows the sample clustering of the subjects in the training cohort, with the Basal group clearly separated from the rest of the subjects and a significant overlap between the Luminal groups. Using a correlation cut-off of 0.7, DIABLO identified 44 pair-wise associations between miRNA and other omics variables (mRNA, CpGs and proteins), amongst which 34 (77%) were negative correlations (**Figure 5D**). Similar to **Figure 5C**, the heatmap in **Figure 5E** shows tight clustering of Basal and Her2 samples, and intermixing of the Luminal A and B samples. We then performed a gene-set enrichment analysis for each set of 9 variables separately using curated gene sets and oncogenic signatures (C2 and C6 collections) (**see Methods**). The top 25 ranked pathways from the C2 and C6 collections were mainly enriched with the mRNA and proteins (**Figure 5F**). The top ranked pathways consisted of several breast cancer-related pathways such as *breast cancer ESR1 up*, *breast cancer basal down*, *breast cancer basal vs. luminal*, *breast cancer luminal vs. basal up*, and *breast cancer relapse in bone up*. Given the highly correlated nature of the variables identified through DIABLO, our analyses and identified molecular signature may suggest novel biologically plausible roles of the selected CpGs and miRNAs in breast cancer.

A holistic view of molecular processes in blood during allergen inhalation change.

Next we showed the utility of DIABLO to a repeated measures study, incorporating cell-types and pathway-based modules using molecular data from 14 asthmatic individuals undergoing allergen inhalation challenge [35,36] (**Figure 6A**). Blood samples were collected prior to (pre) and 2 hours after (post) allergen challenge and profiled for cell-type frequencies (9 cell-types), leukocyte gene transcript expression and plasma metabolite abundances. A module based approach (also known as eigengene summarization [11]) was used to transform both the gene expression and metabolite datasets into pathway datasets. Consequently, each variable in those two datasets now represented the pathway activity expression level for each sample instead of direct gene/metabolite expression. The mRNA dataset was transformed into a Kyoto Encyclopedia of Genes and Genomes (KEGG) dataset whereas the metabolite dataset was transformed into a metabolite pathway dataset (**see Methods**). We observed that metabolite modules were highly correlated with cell-counts and gene modules (Pearson correlation > 0.8) (**Figure S3, Additional file 5**). To account for the repeated measures experimental design, a variance decomposition technique was applied to all datasets (**see Methods**). **Figure 6B** shows our chosen DIABLO design to identify correlated sets of cells, gene and metabolite modules that were altered after allergen inhalation challenge. The DIABLO model identified 2 cell-types, 10 gene and metabolite modules across two components. We compared the performance of DIABLO with variance decomposition for the repeated experimental design (AUC=99%, leave-one-out cross-validation) or with no variable decomposition (AUC=85%), suggesting a high individual variability that is greater than the pre-post challenge differences (**Figure 6C**). **Figure 6D** shows that the DIABLO method is able to maximize the correlation between components from each omics dataset or module as specified in the design matrix, and the first component shows a clear separation between pre- and post-challenge samples. Interestingly, many asthma-related cell-types and molecular pathways were identified by DIABLO, as represented in **Figure 6E** The selected cell-types, eosinophils and basophils, are considered hallmarks of allergic asthma [37]. The selected gene-module pathways included *Asthma* KEGG pathway (**Figure S4, Additional file 6**) even though individual gene members were not significantly altered post-challenge (**Figure S5, Additional file 7**). DIABLO also selected the *Valine, leucine and isoleucine (branched-chain amino acids, BCAAs) biosynthesis* gene module and the *Valine, leucine and isoleucine metabolism* metabolite module, where the activity of both significantly increased post-challenge (**Figure S6, Additional file 8**). These findings depict common molecular processes that span different biological layers, and suggest the ability of our DIABLO method to identify features that suggest a mechanistic link with response to allergen challenge.

Discussion

Classification algorithms *a priori* do not focus on incorporating biological information and therefore, any derived discriminatory markers ("biomarkers") may not mechanistically link the underlying biology to the phenotype. To address this concern, we developed DIABLO, an integrative classification method which not only identifies subsets of discriminatory molecules from each omics dataset, but also aims to more plausibly model the correlation structure between them, assuming that correlation implies similar functional relationships [26].

DIABLO promotes a compromise between a performance-driven and biologically-driven multi-omics biomarker panel. For example, for the single-omics analyses, the mRNA dataset was found to be the most discriminatory and led to superior performance using Enet, SVM and RF compared to the multi-omics DIABLO panel which included equal numbers of each type of omics variables. The high performance of the mRNA dataset may be due to the fact that the PAM50 gene classifier was developed using gene expression data (even though the PAM50 genes were removed) and thus, genes correlated with the PAM50 genes may be driving the classification signal. Therefore, although DIABLO provided an enhanced set of correlated omics variables, its classification performance was hindered by variables with less discriminatory power. On the other hand, DIABLO out-performed the other single-omics panels (miRNA, CpGs and proteins), which may be explained by the integrative focus of DIABLO, where stronger discriminatory omics variables may compensate for weaker ones. While the existing integrative schemes using Enet out-performed DIABLO, they presented strong limitations, such as an over representation of selected mRNAs (Concatenation-Enet) and panels with a large number of features (Ensemble-Enet). Furthermore, the Concatenation method could not be objectively assessed in the test set that was missing proteomics data, while the Ensemble methods could only be assessed based on three omics datasets. The DIABLO classifier, however, was built by integrating all four datasets and tested using three datasets. The integration task in DIABLO models the information contained in the proteomics dataset, even though it was missing in the test set. Therefore, DIABLO gives a competitive performance compared to existing integrative methods with added user-friendly benefits of: 1) user specified number of features, 2) strong correlation between the variables identified and 3) ability to make predictions on new data even when some new datasets are missing. Lastly, although we have shown comparable performance of DIABLO with existing methods for this particular human breast cancer data, conclusions may vary with other datasets from other biological studies.

A challenge that limits the clinical translatability of multi-omics biomarker panels is the increased number of features that need to be assessed in combination. DIABLO is able to build a sparse classifier (36 variables, 9 from each omics space) with a competitive performance compared to existing methods which contained hundreds of omics-specific variables. Our analyses revealed a multi-omics characterization of the PAM50 breast cancer subtypes by combining selected mRNA, miRNA, CpGs, and proteins as validated by the gene-set enrichment analyses. Although the top ranked pathways (many related to breast cancer) were enriched with mRNA and proteins, this suggests that the identified miRNA and CpGs may also play a role in determining the PAM50 subtypes. Therefore, DIABLO helps generate novel hypotheses that arise from evidence through multiple biological layers of information and may lead to more robust signatures compared to those generated via single-omics analyses.

We also demonstrate the utility of DIABLO to study molecular processes across different omics layers by combining DIABLO with a modular approach. Often biologists are interested in identifying significant molecular pathways that are dysregulated between disease groups instead of identifying biomarker signatures that contain a limited number of features which restricts gene-set enrichment analyses. DIABLO identified known cell-types and pathwaysin the context of asthma, such as eosinophils/basophils and the *Asthma KEGG pathway*, as well as, novel pathways such as the *valine, leucine and isoleucine (branched-chain amino acids, BCAAs) pathway*. The mechanistic relationship between BCAAs and allergic responses is a novel finding and their role as potential predictors of allergic inflammation needs to be further investigated. This proof-of-concept study demonstrates the ability of DIABLO to uncover common relationships between different biological layers, resulting in novel hypotheses to be validated in the laboratory.

Despite the multi-purpose nature of DIABLO, we acknowledge some limitations of the method. The linearity assumption between the selected omics variables and the response may not be valid in some biological research areas, and the further development of kernel-based methods to model non-linear relationships between omics levels and the response may overcome this problem. The other limitation that is also encountered with other machine learning algorithms is the tuning of the parameters. The optimal number of variables to select from each dataset, can be computationally intensive, as we have used repeated cross-validation to ensure unbiased classification error rate evaluation. A grid approach was deemed reasonable and provided very good performance results, but may still be suboptimal as we had to restrict the grid space. Finally, and similar to other methods, DIABLO suffers from potential technical artifacts of the data, such as batch effects, presence of confounding variables and differences in noise levels with respect to the different technologies used for each omics dataset. Therefore, we recommend exploratory preliminary analyses for each single-omics dataset to address technical factors that may affect downstream analyses using DIABLO.

Nowadays, system biologists, computational biologists and bioinformaticians dealing with multi-omics studies face the challenge of 'missing' the biological question, thus relying on data-driven statistical approaches in the absence of specific hypotheses to be tested. Our study shows that a precise biological question is crucial to perform integrative analyses, as it will aid the choice of the design in the DIABLO model, whether to use of variance decomposition and whether to use pathway-based modules. Our proposed pipeline has strong potential to identify multi-omics signatures that discriminate multiple phenotypic groups and can be interpreted through the use of various graphical outputs. Our ultimate goal is that those identified molecular signatures will help in generating novel biological hypotheses to be tested and validated back in the laboratory, thus eventually filling the gap of the missing biological question.

Conclusions

We introduced DIABLO, a dimension reduction multivariate method to integrate several omics datasets measured on the same set of samples, while accounting for the heterogeneity between omics platforms. The aim of DIABLO is to classify samples according to known phenotypic groups, to identify a small but robust multi-omics molecular signature that can predict phenotypic groups in new test samples.

To our knowledge, DIABLO is the only integrative classification method that models the correlation structure between omics data spaces, thus improving biological insights by linking biology to phenotype. We propose a flexible framework for different data-types that can be applied to any type of datasets (not only omics), various study designs and pathway-based module analyses. The mixDIABLO framework will allow researchers to explore datasets, build multi-omic panels, assess the performance of these integrative statistical models, create visualizations to assist in the interpretation of these models in the biological context, and, ultimately, generate novel hypotheses to be validated in the laboratory.

Methods

Code availability and software tool requirements. The DIABLO framework is implemented in the mixOmics R package [20,21]. mixOmics currently includes 15 multivariate methodologies, for single-omics analysis and integration of two datasets. All scripts/tutorials can be found on the webpage (http://www.mixomics.org/mixDIABLO). All analyses were performed using the R statistical computing program [43] (version 3.3.1) and the mixOmics package (version 6.0.0).

Statistical methods and analysis

General multivariate framework to integrate multiple datasets measured on the same samples. DIABLO extends sparse generalized canonical correlation analysis (sGCCA) [24] to a classification 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 $c_{jk} = 0$ (no relationship modelled between the datasets) or $c_{jk} = 1$ 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 [44] or l_1 –penalized multivariate model for one single-omics dataset [23], 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 omics variables with a non-zero coefficient from the omics dataset 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 achieved.

The underlying assumption of the sGCCA model is that the major source of common biological variation can be extracted via the 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 [24].

DIABLO for supervised classification 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, the l_1 penalty parameter λ_k was replaced by the number of variables to select in each dataset and each component, as there is a direct correspondence between both parameters.

The class membership of a new sample i which is measured across the different types of omics datasets $\widetilde{X_k^i}$ is predicted using the fitted sGCCA model with the estimated variable coefficients vectors $\hat{\boldsymbol{a}}^k$ to estimate the predicted scores $t^{k,i} = \widetilde{X_k^i}\hat{a}^k$, k = 1, …, K. To each dataset k corresponds a predicted continuous score $t^{k,i}$ which assigns a predicted class using a distance such as the Maximum, Centroids or Mahalanobis [20], as described in Lê Cao *et al.* [23] and in the mixOmics package. Each component $t^{k,i}$ associated to each dataset k predicts the class membership of the new sample i, and the consensus class membership across all K datasets is determined using either a majority vote or by averaging all $t^{k,i}$ 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. Because the class prediction relies on individual vote from each omics set, DIABLO is highly flexible and thus allows for some missing datasets X_k during the prediction step. In our two studies we used the centroid distance for the majority vote scheme (breast cancer study) and the maximum distance for the average vote scheme (asthma study) during performance evaluation and test set prediction.

Design matrix in DIABLO. The design matrix C is a KxK matrix of zeros and ones which specifies whether the covariance between two datasets should be maximized in the DIABLO model, as presented in equation (1). In our simulation study we evaluated two different scenarios: a null design is when no datasets are connected, and a full design is when all datasets are connected:

$$\mathrm{C}_{null} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathrm{C}_{full} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Note that internal to the DIABLO method, the design always links each dataset to the outcome Y. For the two case studies (breast cancer and asthma) the design matrix was computed based on our proposed method (see below ***Parameters tuning***).

Parameters tuning.

The first parameter to tune in 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.

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 [23], 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. 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. In our experience, the number of variables to select in each dataset provides 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. Also, 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 signature than a literature-search interpretation.

Visualization outputs with DIABLO. Several types of graphical outputs were made available in mixOmics to improve the interpretation of the DIABLO results.

Sample plots. Pairs of components associated to each dataset are used to represent the samples projected in the space spanned by those components in each individual omics dataset. The sample plot enables the user to visualize the ability of the DIABLO model to extract common information at the sample level for each dataset, as well as to visualize the discriminatory power of each data type to separate the phenotypic groups. The scatterplot matrix (**Figure 5C, Figure 6D**) represents correlation between components for the same dimension but across all omics datasets to verify that the model maximizes the correlation as indicated in the design matrix. Since DIABLO is a supervised method, separation of subjects of different phenotypic groups can be seen using this type of plot.

Variable plots. To visualize selected variables, we proposed circos plot (**Figure 5D**) 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 [45]. For each omics dataset, DIABLO produces a variable coefficient matrix of size (p_k x 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. When we compared several approaches that do not output latent components (e.g. Enet) we calculated instead a Pearson correlation matrix, where each link represents a Pearson correlation coefficient.

Clustered Image Map (CIM). A clustered image map [45] 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: 1) mRNA – gene symbols used as-is. 2) DNA methylation – features were mapped to coding gene symbol manually from downloaded annotation file. 3) Protein – features mapped to coding gene symbol manually from downloaded annotation file. 4) miRNA – a previously described strategy was used [46]. Briefly, all gene sets were mapped back to a set of miRNAs associated with them, using a database of computationally predicted target genes for each miRNA (e.g. if a gene set is composed of genes A, B and C, genes A and B are targets of miRNA X, while gene C is a target of miRNA Y and Z, the new gene set will be made up of miRNA X, Y and Z. This effectively deals with deduplication issues.) Enrichment of the miRNA features was then assessed against these transformed gene sets.

The following collections were used as gene-sets for the enrichment analysis [47]: 1) **C2** is a collection of curated gene sets such as Pathway Interaction DB (PID), Biocarta (BIOCARTA), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome (REACTOME). 2) **C6** is a collection of oncogenic gene sets (signatures of cellular pathways which are often dysregulated in cancer).

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).

Data description and preprocessing

Breast cancer multi-omics study.

Datasets accession. The level 3 TCGA data (version 2015_11_01) were retrieved from firebrowse.org hosted by the Broad Institute. The clinical data file (Merge_Clinical) was downloaded from the Primary tab of the BRCA Clinical Archives. The mRNA RSEM normalized dataset (illuminahiseq_rnaseqv2-RSEM_genes_normalized) was downloaded from the Primary tab of the BRCA mRNASeq Archives. The miRNA datasets (illuminahiseq_mirnaseq-miR_gene_expression and illuminaga_mirnaseq-miR_gene_expression) were downloaded from the Primary tab of the BRCA miRSeq Archives. The reverse phase protein array dataset (mda_rppa_core-protein_normalization) was downloaded from the Primary tab of the BRCA RPPA Archives. The beta values for the methylation datasets (humanmethylation27-within_bioassay_data_set_function and humanmethylation450-within_bioassay_data_set_function MD5) were downloaded from the Primary tab of the BRCA Methylation Archives.

Data processing. Clinical data were present for 1098 subjects for 3,703 variables. Un-annotated (29) transcripts were removed from the mRNA dataset (20,502 genes x 1212 samples). Two transcripts corresponded to *SLC35E2*, therefore one of the transcripts was re-labelled *SLC35E2.rep*. The miRNA datasets (1,046 miRNA x 1190 samples) was derived using two different Illumina technologies, the Illumina Genome Analyzer (341 samples) and the Illumina HiSeq (849 samples). The read counts instead of the reads_per_million_miRNA_mapped were used. The proteomics dataset obtained using a reverse phase protein array consisted of 142 proteins for 410 samples. The methylation data was derived from two different platform, the Illumina Methylation 27 (27,578 CpG probes x 343 subjects) and the Illumina 450K (485,577 CpG probes x 885 subjects). There were 25,978 CpG probes in common between the platforms. The PAM50 labels for 1182 samples were obtained from the TCGA staff.

Since some samples were derived from the same individuals, all datasets were restricted to samples coming from the primary solid tumor (sample type code 01) and to the first vial (vial code A), resulting in the following datasets for mRNA (20,502 genes x 1080 subjects), miRNA (1,046 miRNAs x 1066 subjects), proteins (142 proteins x 403 subjects), CpGs (25,978 CpG probes x 1066 subjects) and 1049 subjects with PAM50 subtypes present.

Training and test cohorts. There were 387 subjects (Basal: 76, Her2: 38, LumA: 188, LumB: 77 and Normal: 8) common between the clinical, mRNA, miRNA, proteomics, methylation and PAM50 label datasets. The biomarker analysis was performed using 4 molecular datasets, mRNA, miRNA, CpGs and proteins. Since the proteomics dataset was the limiting dataset, the test datasets only consisted of the mRNA, miRNA and CpG data matrices. The test cohort consisted of 638 subjects; Basal: 102, Her2: 40, LumA: 346, LumB: 122 and Normal: 28. Given the limited number of normal subjects, they were not used in the biomarker analysis.

Normalization and pre-filtering. The count data for the mRNA dataset was normalized to log2-counts per million (logCPM), similar to limma voom [48]:

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

After library size normalization, genes with counts less than 0 were removed. In addition, the 3000 most variable genes based on the median absolute deviation (MAD) were retained for downstream analysis. 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 were also removed. The CpG probes containing missing data were removed from the methylation data and the 2000 most variable probes based on MAD were retained for downstream analysis.

Asthma multi-omics study

Datasets accession. Paired blood samples were obtained from 14 asthmatic individuals undergoing allergen inhalation challenge as previously described[49]. Cell counts were obtained from a hematolyzer (percentage of Neutrophils, Lymphocytes, Monocytes, Eosinophils and Basophils) and DNA methylation analysis (percentage of T regulatory cells, T cells, B cells and Th17 cells). Gene expression profiling was performed using Affymetrix Human Gene 1.0 ST (GSE40240). Metabolite profiling was performed by Metabolon Inc. (Durham, North Carolina, USA). All asthma data have been published as part of previous studies[35,36].

Normalization. Microarray data was normalized using Robust MultiArray Average (RMA), consisting of background correction, quantile normalization and probe summarization using median polish. Preprocessing of mass spectrometry data including data extraction, peak-identification and data preprocessing for quality control and compound identification was performed by Metabolon Inc. (Durham, North Carolina, USA).

Modular analysis. Eigengene summarization is a common approach to decompose a n by 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 [11]. 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 repeated measures study designs. 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) [50], 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.

List of abbreviations

DIABLO, Data Integration Analysis for Biomarker discovery using a Latent component method for Omics studies; AUC, area under the receiver operating curve; PLS, Projection to Latent Structure models; sPLS-DA, sparse PLS-Discriminant Analysis, sGCCA, sparse generalized canonical correlation analysis; PCA, Principal Component Analysis; BER, Balanced Error Rate; Enet, elastic net; RF, random forest; SVM, support vector machine; KEGG, Kyoto Encyclopedia of Genes and Genomes;

Declarations

- **Acknowledgements**

- The authors would like to thank Dr. Kevin Chang (University of Auckland) for some preliminary exploratory analyses of the breast cancer dataset. We would also like to thank Mr. 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. KALC was supported in part by the National Health and Medical Research Council (NHMRC) Career Development fellowship (APP1087415).

-

- **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.

Additional files

Additional file 1: Simulation Study. (PDF 382KB)

Additional file 2: Table S1. Year of collection for TCGA breast cancer samples. (PDF 34KB)

Additional file 3: Figure S1. Overlap of expression between train and test sets. (PDF 54KB)

Additional file 4: Figure S2. Design matrix used for Figure 5. (PDF 71KB)

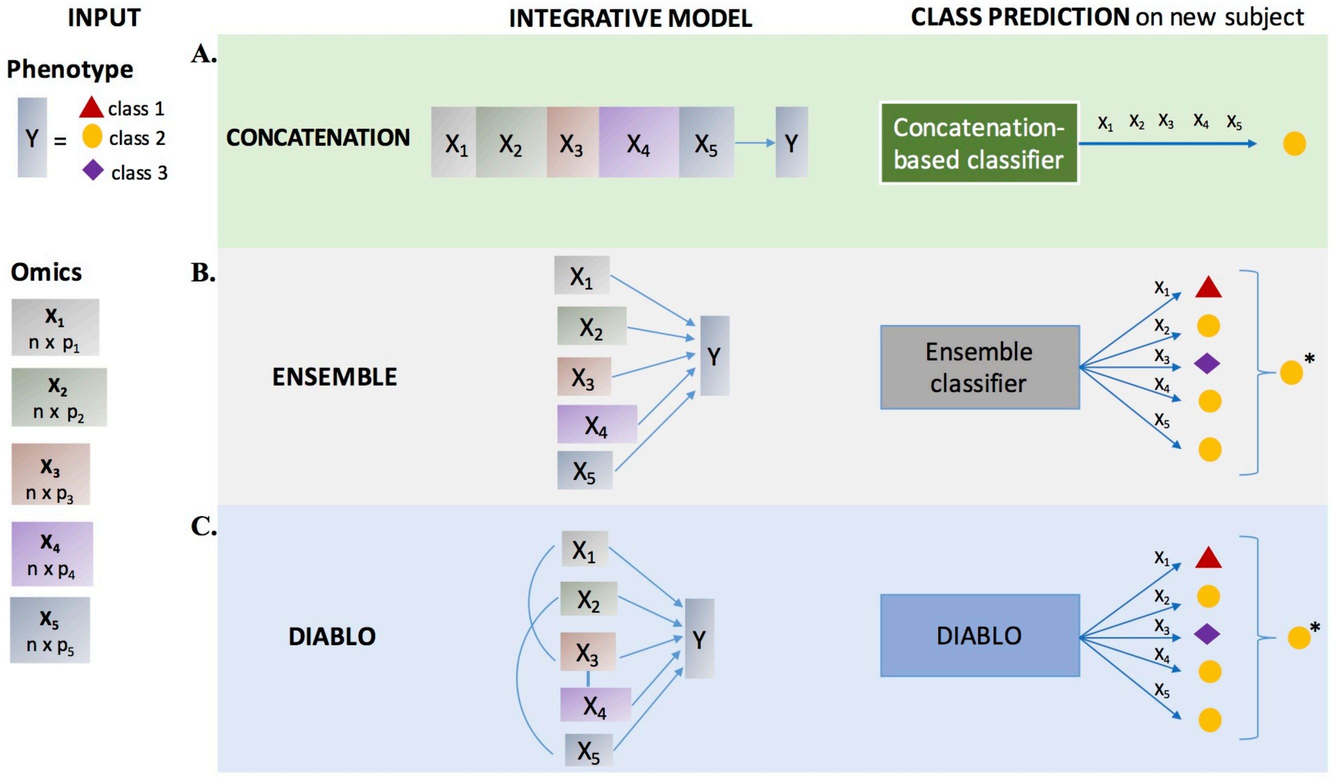
Additional file 5: Figure S3. Design matrix used for Figure 6. (PDF 64KB)

Additional file 6: Figure S4. Asthma KEGG pathway. (PDF 100 KB)

Additional file 7: Figure S5. Volcano plot of genes in the Asthma KEGG pathway. (PDF 65KB)

Additional file 8: Figure S6. Valine, leucine and isoleucine gene and metabolite pathway. (PDF 98KB)

Figures


INPUT
INTEGRATIVE MODEL
CLASS PREDICTION on new subject
Phenotype
Y = ▲ class 1 ● class 2 ◆ class 3
A.
CONCATENATION
X_1 X_2 X_3 X_4 X_5 → Y
Concatenation-based classifier
X_1 X_2 X_3 X_4 X_5
Omics
X_1 n x p_1
X_2 n x p_2
X_3 n x p_3
X_4 n x p_4
X_5 n x p_5
B.
ENSEMBLE
X_1 X_2 X_3 X_4 X_5
Y
Ensemble classifier
X_1 X_2 X_3 X_4 X_5
*
C.
DIABLO
X_1 X_2 X_3 X_4 X_5
Y
DIABLO
X_1 X_2 X_3 X_4 X_5
*

* Majority Vote or Average prediction

Figure 1. Data integrative frameworks and class prediction of new samples. We consider the case where the integration of 5 omics datasets measured on the same samples is required to predict three phenotype groups. **A)** The concatenation framework first concatenates all omics datasets into one joint matrix before fitting a classification model. Prediction is based on a single statistical model. **B)** The ensemble framework fits a classification model for each omics dataset

independently. Prediction is performed by combining all individual predictions. **C)** The proposed DIABLO method models relationships between omics datasets based on a given design determined according to a data-driven or knowledge-driven approach. The multivariate method then maximizes the correlation between latent components of each omics when specified in the design. Prediction is based by combining all omics latent components predictions.

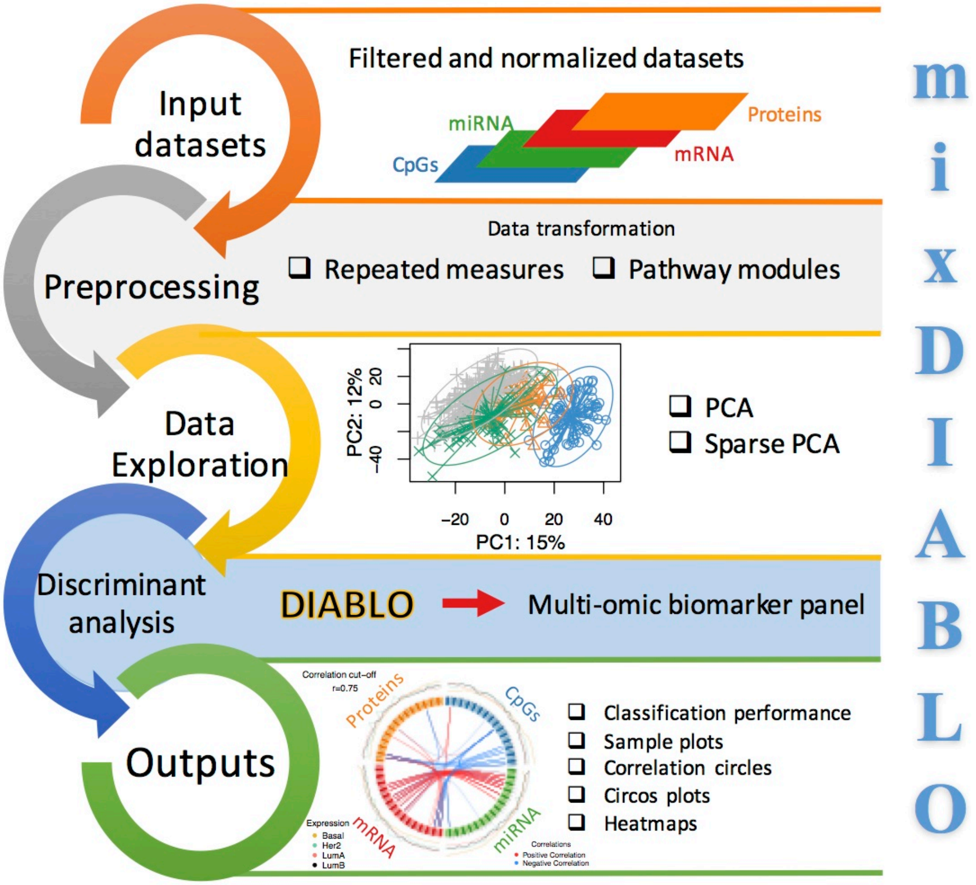


Figure 2. mixDIABLO - a framework for multi-omics data integration and identification of multi-omic panels. Multiple datasets measured on the same samples assigned to two or more phenotype groups are integrated. Once the data are normalized and filtered additional transformation may be applied to account for repeated measurement designs or to summarize gene modules. Preliminary unsupervised analysis is performed to determine the design which

will be modeled in the DIABLO method to identify a multi-omics biomarker panel. Classification performance is assessed using repeated cross-validation and interpretation of the results is enabled through various sample and variable plots, as well as pathway enrichment analysis.

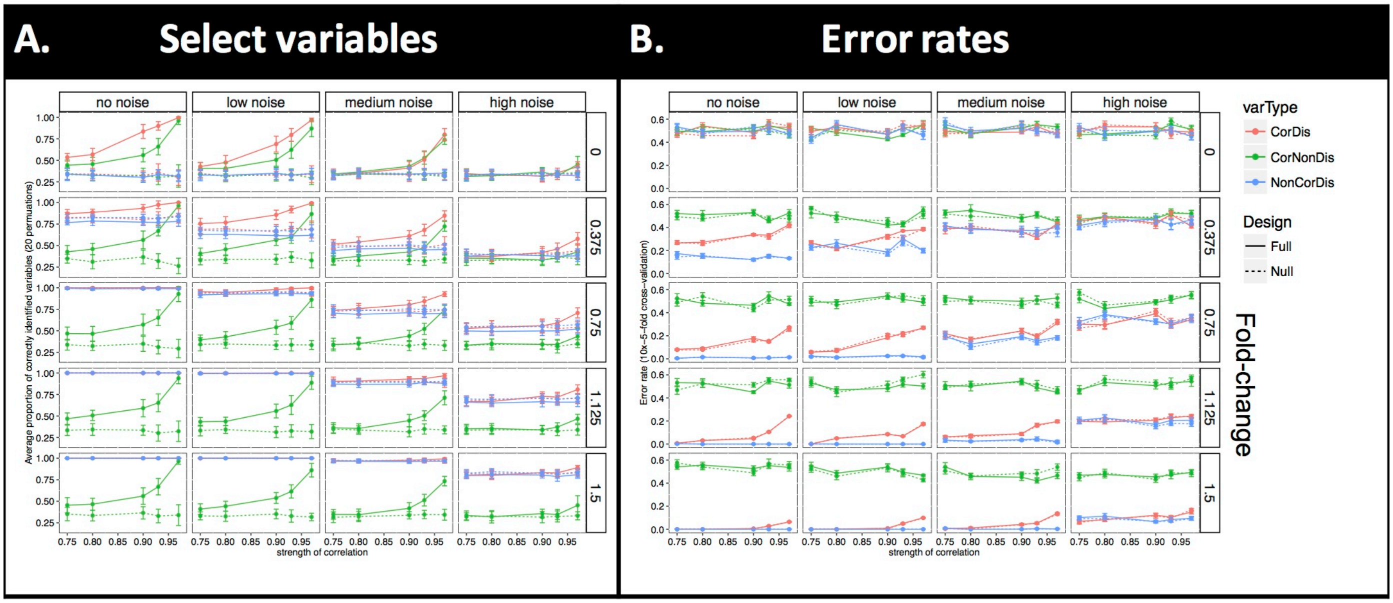


Figure 3. Simulation study: proportion of relevant selected variables and classification error rates. We assessed the impact of the full and null DIABLO designs on the different types of variables that were selected, namely CorDis: correlated and discriminatory variables across the two phenotype groups, CorNonDis: correlated but non-discriminatory, NonCorDis: not correlated but discriminatory variables, when varying the strength of correlation, fold-change and noise, as described in **Additional file 1**. **A.** Proportion of variables correctly identified by DIABLO. DIABLO identified a greater proportion of CorNonDis and CorDis variables in the full design compared to the null design. The proporition of CorDis is higher with a high fold-change. No difference was identified for NonCorDis variables between the two designs. **B.** Averaged DIABLO classification error rates. When fold-change = 0, all models had an average error rate of 0.50 corresponding to a random prediction. When the fold-change increases,

NonCorDis variables lead to a better performance than CorDis variables regardless of the design while CorNonDis variables and irrelevant variables are (by definition) not useful for classification.

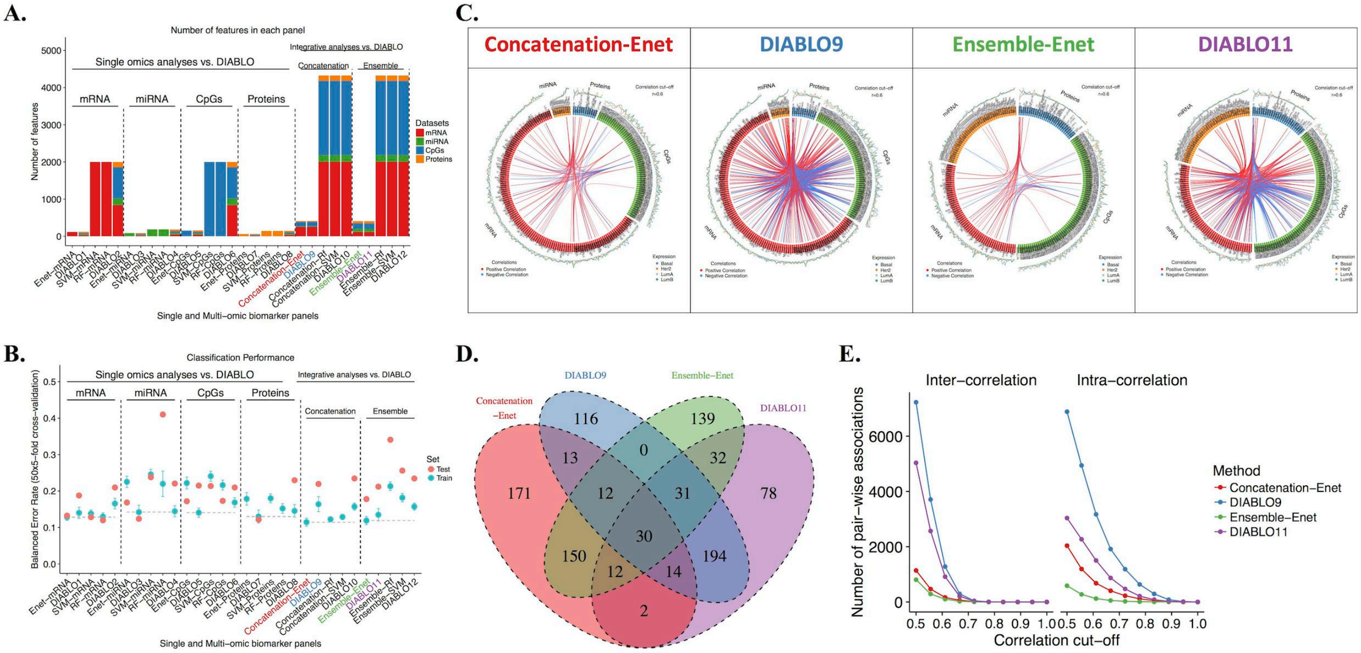


Figure 4. Comparison of single-omics and integrative frameworks on the breast cancer multi-omics study. We compared the classification methods Elastic Net (Enet), Support Vector Machine (SVM) and Random Forests (RF) applied on single-omics and integrative frameworks (concatenation and ensemble). SVM and RF do not perform variable selection. Different DIABLO models with a number of selected variable matching with Enet selection were fitted (DIABLO1-12). **A)** panel size for each method, **B)** Classification error rate based on 50x10-fold cross validation, balanced error rate is averaged for the training set. **C)** The circos plots depict the multi-omics panel identified by the methods and show the unbalance of omics variables selected in the concatenation method and the ability of DIABLO to identified highly connected multi-omics signatures**.** **D)** Overlap of commonly selected variables across the different

integrative methods for similar panel sizes. **E)** Description of the nature of the correlation in the different multi-omics panel with omics (intra) and between omics (inter) selected variables.

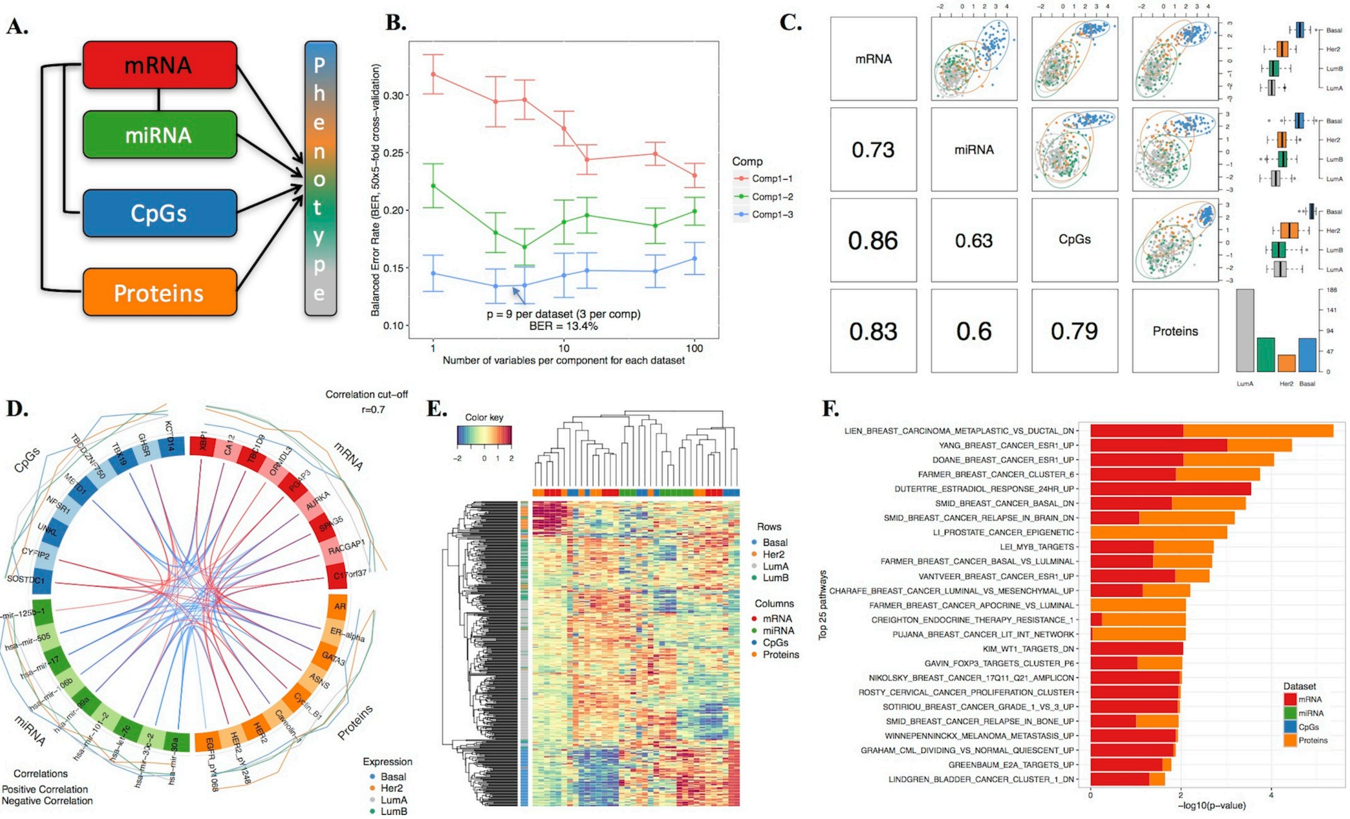


Figure 5. DIABLO graphical and numerical outputs on the breast cancer multi-omics study. A) Input design in DIABLO determined according to a data-driven approach. **B)** Classification performance using 50x5-fold cross validation to tune the number of variables to select on each component in DIABLO. **C)** Matrix scatterplot to verify that the first components related to each omics dataset (upper matrix) are maximally correlated (lower matrix, Pearson correlation) in DIABLO according to the input design in A. **D)** Circos plot of the final multi-omics signature identified in the training set. **E)** Clustered Image Map representing the multi-omics signature in relation with the samples and **F)** Pathway enrichment analysis based on the multi-omics signature.

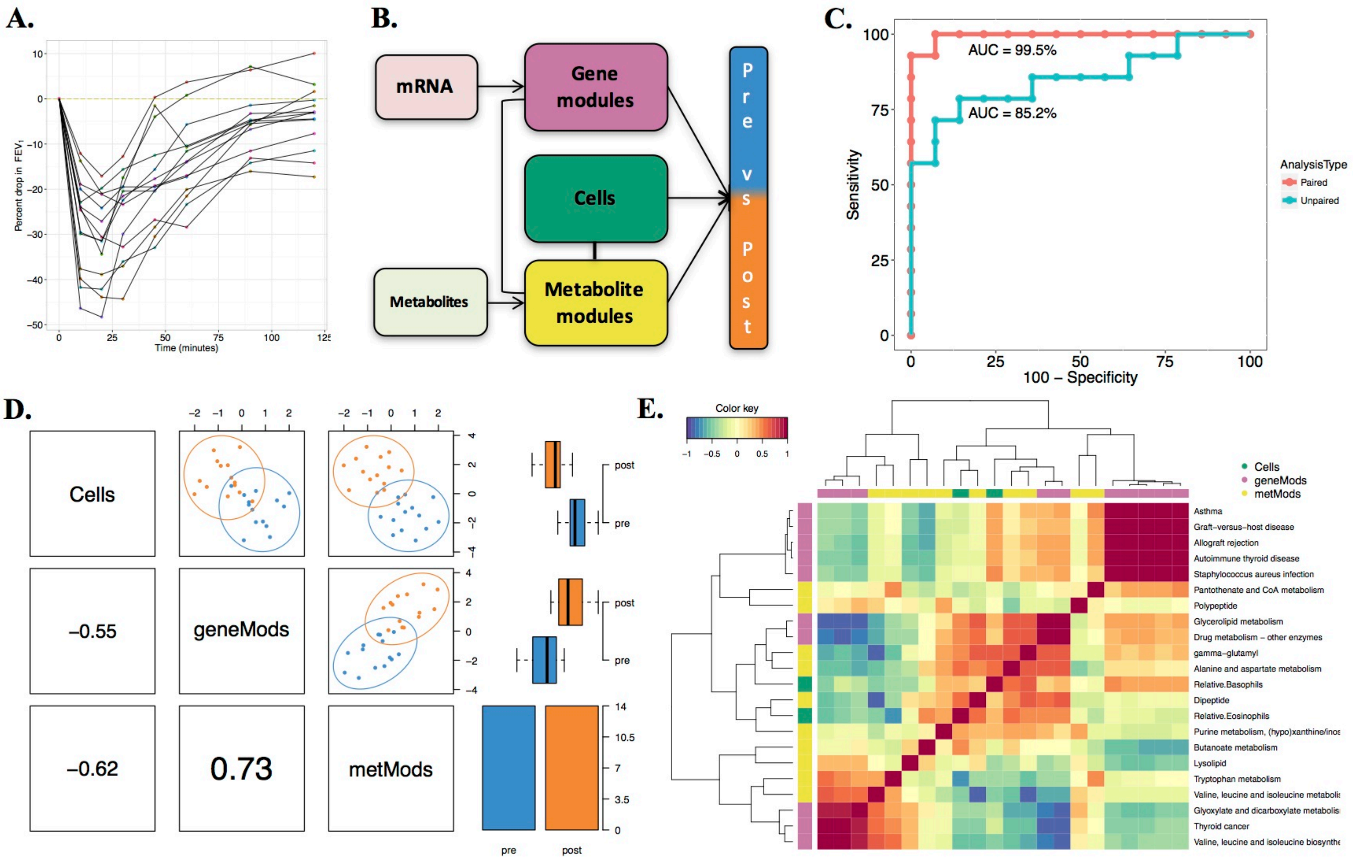


Figure 6. Systems approach to molecular changes in blood after allergen inhalation challenge. A. FEV_1 response profiles 0-2 hours after allergen inhalation. **B.** Path diagram of the connection between datasets, all predicting the time point variable. The mRNA and metabolite datasets were transformed into module datasets. **C.** ROC comparing two DIABLO models that include / exclude the repeated measures experimental design. **D.** Sample plots depicting the clustering of subjects based on the first component of each dataset from the DIABLO model. **E.** Correlation between variables selected in the DIABLO model.

References

1. Zhang W, Li F, Nie L. Integrating multiple “omics” analysis for microbial biology: application and methodologies. Microbiology [Internet]. 2010 [cited 2016 Jul 22];156:287–301. Available from: http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.034793-0

2. Bunyavanich S, Schadt EE. Systems biology of asthma and allergic diseases: A multiscale approach. J. Allergy Clin. Immunol. [Internet]. 2015 [cited 2015 Nov 29];135:31–42. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0091674914014869

3. Angione C, Conway M, Lió P. Multiplex methods provide effective integration of multi-omics data in genome-scale models. BMC Bioinformatics [Internet]. 2016 [cited 2016 Mar 11];17. Available from: http://www.biomedcentral.com/1471-2105/17/S4/83

4. Tenenhaus A, Tenenhaus M. Regularized Generalized Canonical Correlation Analysis. Psychometrika [Internet]. 2011 [cited 2015 Jul 15];76:257–84. Available from: http://link.springer.com/10.1007/s11336-011-9206-8

5. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics [Internet]. 2012 [cited 2016 Jan 19];28:3290–7. Available from: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts595

6. Schadt EE, Björkegren JL. NEW: network-enabled wisdom in biology, medicine, and health care. Sci. Transl. Med. [Internet]. 2012 [cited 2015 Nov 29];4:115rv1–115rv1. Available from: http://stm.sciencemag.org/content/4/115/115rv1.short

7. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics [Internet]. 2014 [cited 2016 Jan 19];15:162. Available from: http://www.biomedcentral.com/1471-2105/15/162?utm_source=dlvr.it&utm_medium=tumblr

8. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, et al. Stitching together Multiple Data Dimensions Reveals Interacting Metabolomics and Transcriptomics 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

9. Glass K, Huttenhower C, Quackenbush J, Yuan G-C. Passing Messages between Biological Networks to Refine Predicted Interactions. Semsey S, editor. PLoS ONE [Internet]. 2013 [cited 2015 Dec 1];8:e64832. Available from: http://dx.plos.org/10.1371/journal.pone.0064832

10. 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

11. 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

12. 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

13. Ha MJ, Baladandayuthapani V, Do K-A. DINGO: differential network analysis in genomics. Bioinformatics [Internet]. 2015 [cited 2016 May 3];31:3413–20. Available from: http://bioinformatics.oxfordjournals.org/content/31/21/3413.short

14. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems. J Mach Learn Res [Internet]. 2014 [cited 2016 Jul 23];15:3133–81. Available from: http://www.jmlr.org/papers/volume15/delgado14a/source/delgado14a.pdf

15. Zou H, Hastie T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. [Internet]. 2005 [cited 2015 Jul 15];67:301–20. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/pdf

16. 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

17. Fan J, Han F, Liu H. Challenges of Big Data analysis. Natl. Sci. Rev. [Internet]. 2014 [cited 2016 Jul 23];1:293–314. Available from: http://nsr.oxfordjournals.org/cgi/doi/10.1093/nsr/nwt032

18. Liu Y, Devescovi V, Chen S, Nardini C. Multilevel omics data integration in cancer cell lines: advanced annotation and emergent properties. BMC Syst. Biol. [Internet]. 2013 [cited 2016 Jan 19];7:14. Available from: http://www.biomedcentral.com/1752-0509/7/14

19. 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

20. 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

21. Lˆe Cao K-A, Rohart F, Gautier B, Bartolo F, Gonz´alez I, D´ejean S. mixOmics: Omics Data Integration Project. 2016.

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

23. 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/

24. 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

25. 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

26. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray datasets. Genome Res. [Internet]. 2004 [cited 2016 Mar 30];14:1085–94. Available from: http://genome.cshlp.org/content/14/6/1085.short

27. Shen H, Huang J. Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation. J. Multivar. Anal. 2007;99:1015–34.

28. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature [Internet]. 2000;406:747–52. Available from: http://dx.doi.org/10.1038/35021093

29. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression datasets. Proc. Natl. Acad. Sci. [Internet]. 2003 [cited 2016 Jul 25];100:8418–23. Available from: http://www.pnas.org/content/100/14/8418.short

30. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J. Clin. Oncol. [Internet]. 2009 [cited 2016 Jul 18];27:1160–7. Available from: http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2008.18.1370

31. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med. Genomics [Internet]. 2015 [cited 2016 Jul 18];8. Available from: http://www.biomedcentral.com/1755-8794/8/54

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

33. Breiman L. Random forests. Mach. Learn. [Internet]. 2001 [cited 2016 Mar 3];45:5–32. Available from: http://link.springer.com/article/10.1023/A:1010933404324

34. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. [Internet]. 2011 [cited 2016 Jul 24];2:1–27. Available from: http://dl.acm.org/citation.cfm?doid=1961189.1961199

35. 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

36. 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

37. Stone KD, Prussin C, Metcalfe DD. IgE, mast cells, basophils, and eosinophils. J. Allergy Clin. Immunol. [Internet]. 2010 [cited 2016 Apr 5];125:S73–80. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0091674909017345

38. GAUVREAU GM, LEE JM, WATSON RM, IRANI A-MA, SCHWARTZ LB, O'BYRNE PM. Increased numbers of both airway basophils and mast cells in sputum after allergen inhalation challenge of atopic asthmatics. Am. J. Respir. Crit. Care Med. [Internet]. 2000 [cited 2016 Mar 22];161:1473–8. Available from: http://www.atsjournals.org/doi/abs/10.1164/ajrccm.161.5.9908090

39. Comhair SAA, McDunn J, Bennett C, Fettig J, Erzurum SC, Kalhan SC. Metabolomics Endotype of Asthma. J. Immunol. [Internet]. 2015 [cited 2016 Apr 5];195:643–50. Available from: http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.1500736

40. Dweik RA, Sorkness RL, Wenzel S, Hammel J, Curran-Everett D, Comhair SAA, et al. Use of Exhaled Nitric Oxide Measurement to Identify a Reactive, at-Risk Phenotype among Patients with Asthma. Am. J. Respir. Crit. Care Med. [Internet]. 2010 [cited 2016 Apr 5];181:1033–41. Available from: http://www.atsjournals.org/doi/abs/10.1164/rccm.200905-0695OC

41. An Official ATS. American Thoracic Society Documents. Am J Respir Crit Care Med [Internet]. 2011 [cited 2016 Apr 5];184:602–15. Available from: http://www.thoracic.org/newsroom/press-releases/resources/ats-document-final.pdf

42. Fitzpatrick AM. Biomarkers of asthma and allergic airway diseases. Ann. Allergy. Asthma. Immunol. [Internet]. 2015 [cited 2016 Mar 29];115:335–40. Available from: http://linkinghub.elsevier.com/retrieve/pii/S108112061500589X

43. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available from: https://www.R-project.org/

44. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B Methodol. 1996;58:267–88.

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

46. Godard P, van Eyll J. Pathway analysis from lists of microRNAs: common pitfalls and alternative strategy. Nucleic Acids Res. [Internet]. 2015 [cited 2016 May 25];43:3490–7. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv249

47. 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];102:15545–50. Available from: http://www.pnas.org/content/102/43/15545.short

48. 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

49. Singh A, Cohen Freue GV, Oosthuizen JL, Kam SHY, Ruan J, Takhar MK, et al. Plasma proteomics can discriminate isolated early from dual responses in asthmatic individuals undergoing an allergen inhalation challenge. PROTEOMICS - Clin. Appl. [Internet]. 2012 [cited 2016 Mar 2];6:476–85. Available from: http://doi.wiley.com/10.1002/prca.201200013

50. 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

Additional Files

Additional File 1

Simulation study

The purpose of this simulation was to study the effects of the design matrix on the variables selected by DIABLO and their corresponding error rates. Two designs were tested: the null design where no datasets were connected and the full design where all datasets were connected. Three datasets (X, Y, Z) for two phenotypic groups (Y) were generated of equal sizes 100 observations by 150 variables. 100 out of the 150 variables were noisy irrelevant variables and we assessed the ability of DIABLO to identify the true 50 discriminative and/or correlated variables and evaluated the classification error rate for different simulation scenarios where both correlation and noise levels were varied.

To that end we simulated four types of variables, namely non-discriminatory but correlated variables, discriminatory (i.e. explaining the phenotype of interest) but correlated variables, discriminatory non-correlated variables and irrelevant (noisy) variables, as described below:

1. **Non-discriminatory correlated variables**

   We generated covariance matrices Σ of size p^1 x p^1 with different correlation strength λ [∈(0,1)], where $p^1 = 50$:

$$\Sigma_{p^1xp^1} = \lambda^{\begin{bmatrix} 0 & 1 & 2 & \cdots & p^1-1 \\ 1 & 0 & 1 & 2 & \vdots \\ 2 & 1 & 0 & 1 & 2 \\ \vdots & 2 & 1 & 0 & 1 \\ p^1-1 & \cdots & 2 & 1 & 0 \end{bmatrix}}$$

$$\Sigma_{p^1xp^1} = \begin{bmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{p^1-1} \\ \lambda & 1 & \lambda & \lambda^2 & \vdots \\ \lambda^2 & \lambda & 1 & \lambda & \lambda^2 \\ \vdots & \lambda^2 & \lambda & 1 & \lambda \\ \lambda^{p^1-1} & \cdots & \lambda^2 & \lambda & 1 \end{bmatrix}$$

   We denote $\mathbf{X_{nonDisCor}}$ a p^1-dimensional random vector, generated from the multivariate normal distribution $\mathbf{X_{nonDisCor}} \sim N(\mu, \Sigma)$ where $\mu = (0, \ldots, 0)$ as those variables are not discriminatory. We then randomly allocate $X_{\mathbf{nonDisCor}}$ into $X^1_{\text{nonDisCor}}$, $Y^1_{\text{nonDisCor}}$, and $Z^1_{\text{nonDisCor}}$, each of size 100 observations and p^1 variables for group 1, and similarly for group 2 with $\mathbf{X_{nonDisCor2}}$ into $X^2_{\text{nonDisCor}}$, $Y^2_{\text{nonDisCor}}$, and $Z^2_{\text{nonDisCor}}$. Note since the mean of the two groups are equivalent these variables are not discriminatory but are correlated.

2. **Discriminatory non-correlated variables**

   Generate a kxk covariance matrix: $\Sigma = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$, where k = 50

Using the kxk covariance matrix generate $\mathbf{x}_{\mathbf{DisNonCor1}}$, and $\mathbf{x}_{\mathbf{DisNonCor2}}$, k-dimensional random vectors; $[X_1, X_2, \ldots, X_k]$ with a multivariate normal distribution:

$\mathbf{x}_{\mathbf{DisNonCor1}} \sim N(\mu_1, \Sigma)$ & $\mathbf{x}_{\mathbf{DisNonCor2}} \sim N(\mu_2, \Sigma)$ where $\begin{matrix}\mu_1 = (0,0,..,0) \\ \mu_2 = (c,c,..,c)\end{matrix}$, where c is the fold-change ranging from 0, 1, 2, 3

Randomly divide $\mathbf{x}_{\mathbf{DisNonCor1}}$ into X_{dis1}, Y_{dis1}, and Z_{dis1} of size 100 observations, and 50 variables for group 1 and $\mathbf{x}_{\mathbf{DisNonCor2}}$ into X_{dis2}, Y_{dis2}, and Z_{dis2} of size 100 observations, and 50 variables for group 2.

3. Generate irrelevant variables

Generate a kxk covariance matrix: $\Sigma = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$, where k = 100

Using the kxk covariance matrix generate $\mathbf{x}_{\mathbf{irv}}$ be a k-dimensional random vector; $\mathbf{x}_{\mathbf{irv}} = [X_1, X_2, \ldots, X_k]$ with a multivariate normal distribution:

$\mathbf{x}_{\mathbf{irv}} \sim N(\mu, \Sigma)$ where $\mu = (0, 0, \ldots, 0)$

Randomly divide $\mathbf{x}_{\mathbf{irv}}$ into X_{irv1}, Y_{irv1}, and Z_{irv1} of size 100 observations, and 100 variables for group 1 and X_{irv2}, Y_{irv2}, and Z_{irv2} of size 100 observations, and 100 variables for group 2.

4. Generate noise

Range variance from low noise to high noise, change diagonal values to 0, to 1.5:

e.g. $\Sigma = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$

Using the pxp covariance matrix generate $\mathbf{x}_{\mathbf{noise}}$ be a k-dimensional random vector; $\mathbf{x}_{\mathbf{noise}} = [X_1, X_2, \ldots, X_k]$ with a multivariate normal distribution:

$\mathbf{x}_{\mathbf{noise}} \sim N(\mu, \Sigma)$ where $\mu = (0, 0, \ldots, 0)$

Divide $\mathbf{x}_{\mathbf{noise}}$ into X_{noise1}, Y_{noise1}, and Z_{noise1} of size 100 observations (group 1 and group2), and 150 variables (given type of variable (50) + 100 irrelevant variables)).

DIABLO analysis

We then ran DIABLO integrating the three simulated data sets X, Y, Z with equal numbers of observations (n=100) and including Y as the categorical outcome(class vector indicating that 100 samples belong to group1 and 100 samples to group2). Those data sets include 150 variables

each, among which 100 variables were deemed irrelevant variables (not correlated between datasets, and not discriminatory), and the remaining 50 variables were either 1) correlated across all three datasets but not discriminatory between groups (CorNonDis), or 2) correlated and discriminatory (CorDis) or 3) not correlated but discriminatory (NonCorDis).
Separate scenarios of three datasets were constructed by varying the fold-change between the two groups (from 0 to 1.5) and the levels of noise (0-no noise, 1-low noise, 5-medium noise and 15-high noise). In addition, the strength of correlation varied for the parameter λ used to generate the covariance matrices. The figure below depicts three covariance matrices for low (λ=0.75), medium (λ=0.91) and highly (λ=0.98) correlated variables.

The DIABLO analysis included 1 component and 50 variables to be selected and we compared the performance in terms of classification error rate and identification of the true relevant variables when using different matrix designs, namely a full design or a null design.

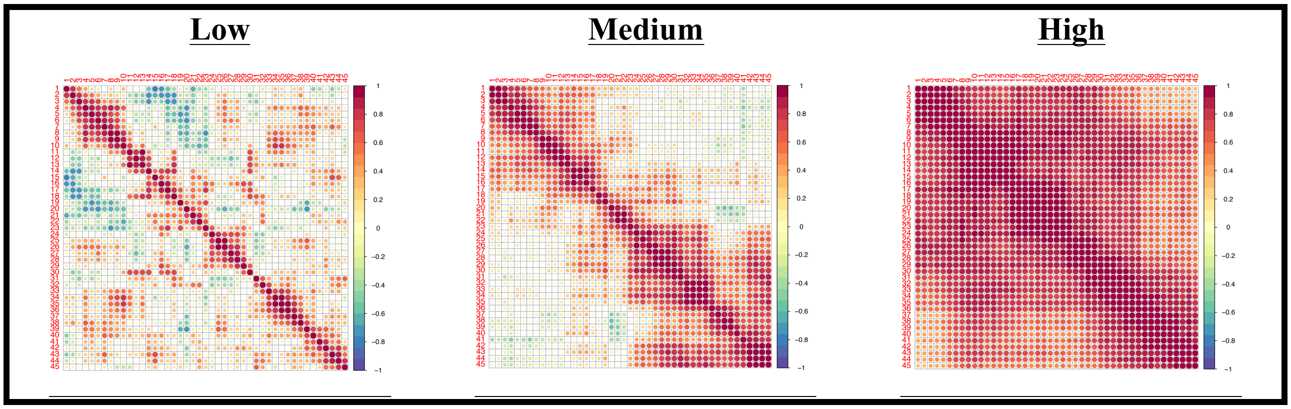


Figure. Covariance matrices varying the strength of correlation between variables.

The strength of the correlation between variables was varied to determine its effects on the number of correctly identified variables by the DIABLO classifier with respect to the design matrix specification.

Additional File 2

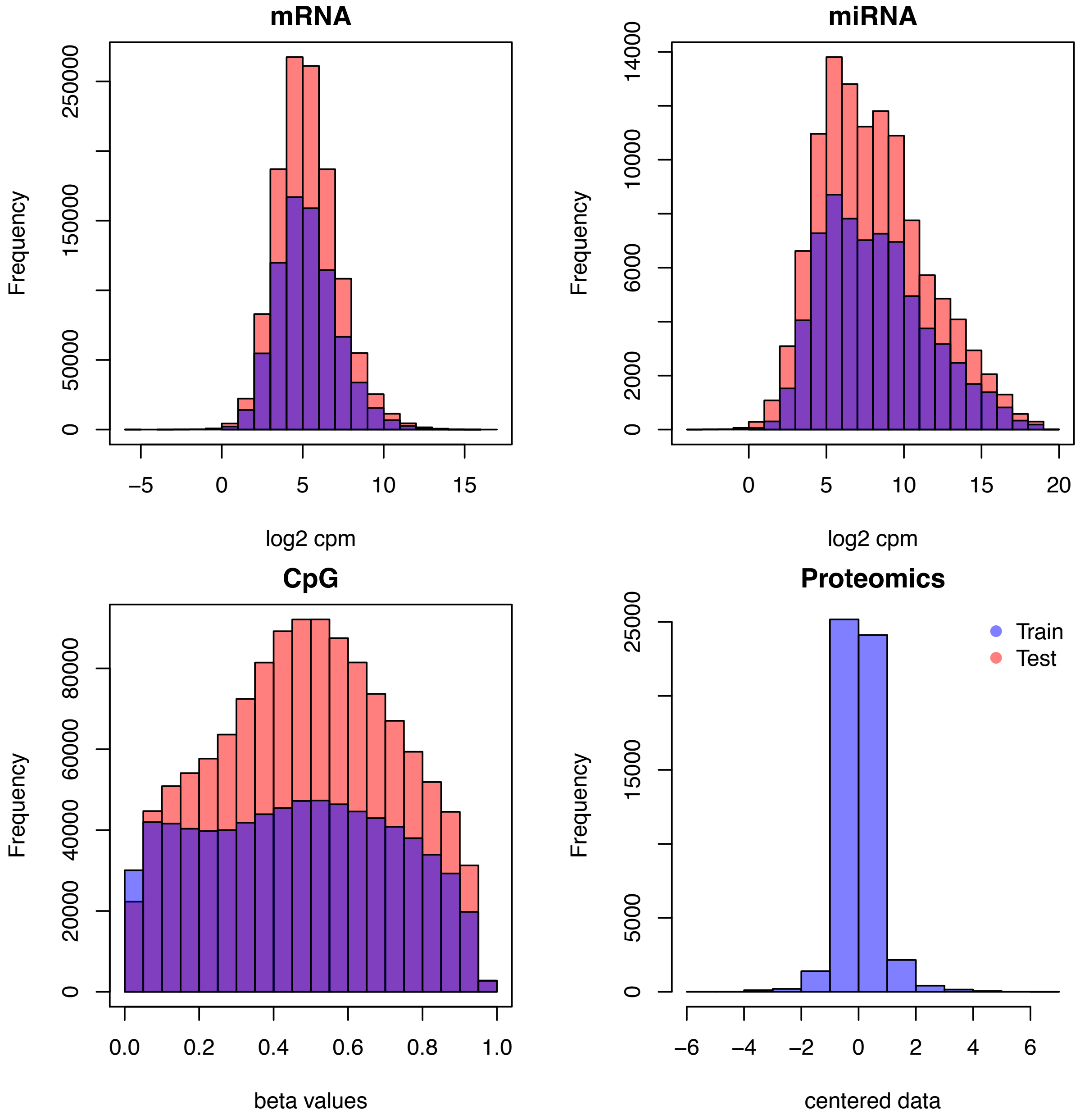
Table: Year of collection for TCGA breast cancer samples

		Year				
Set		2010	2011	2012	2013	Total number of samples
	Train	344 (91%)	35 (9%)	0	0	379
	Test	106 (17%)	325 (53%)	84 (14%)	95 (16%)	610

The majority of samples in the training set were collected in 2010, whereas the majority of samples in the test set were collected in 2011.

Additional File 3

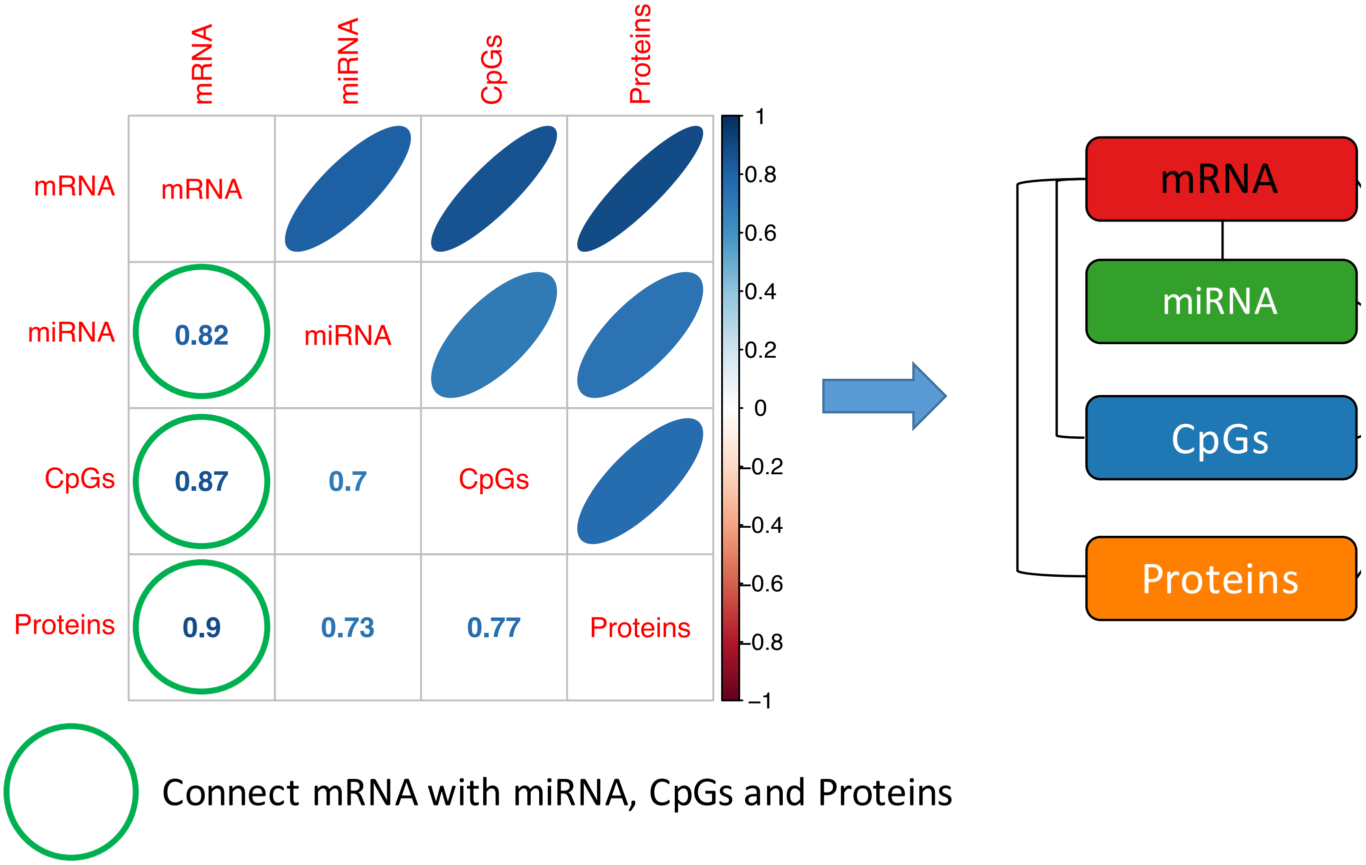
Figure S1. Overlap of expression between train and test sets.



The range of expression for each omic dataset was similar between the training and test set. Note, there was no protein expression data present for subjects in the test set.

Additional File 4

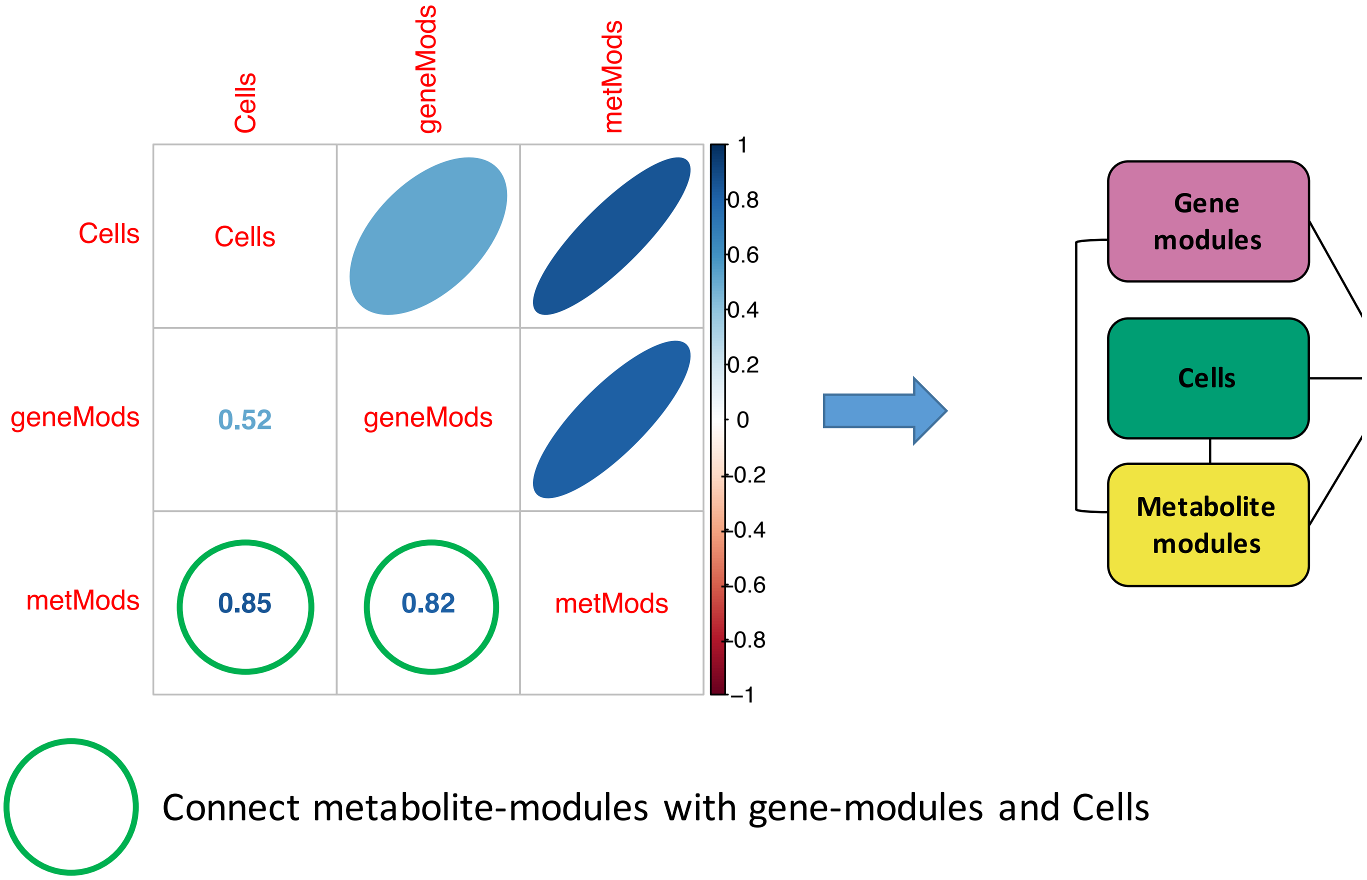
Figure S2. Design matrix used for Figure 5



A correlation cut-off of 0.8 was used to determine the connectivity for the design matrix.

Additional File 5

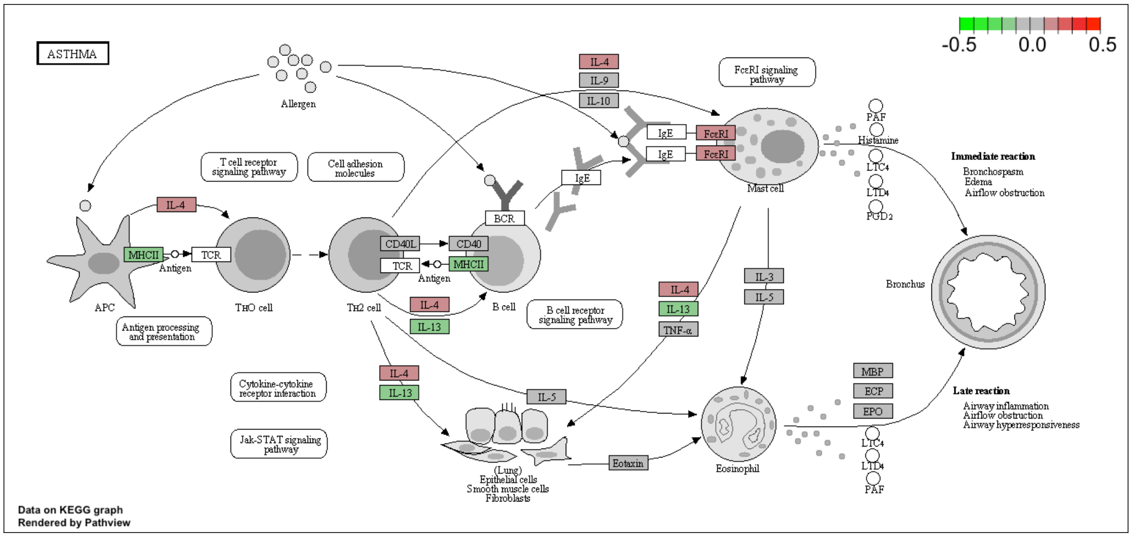
Figure S3. Design matrix used for Figure 6.



Connect metabolite-modules with gene-modules and Cells

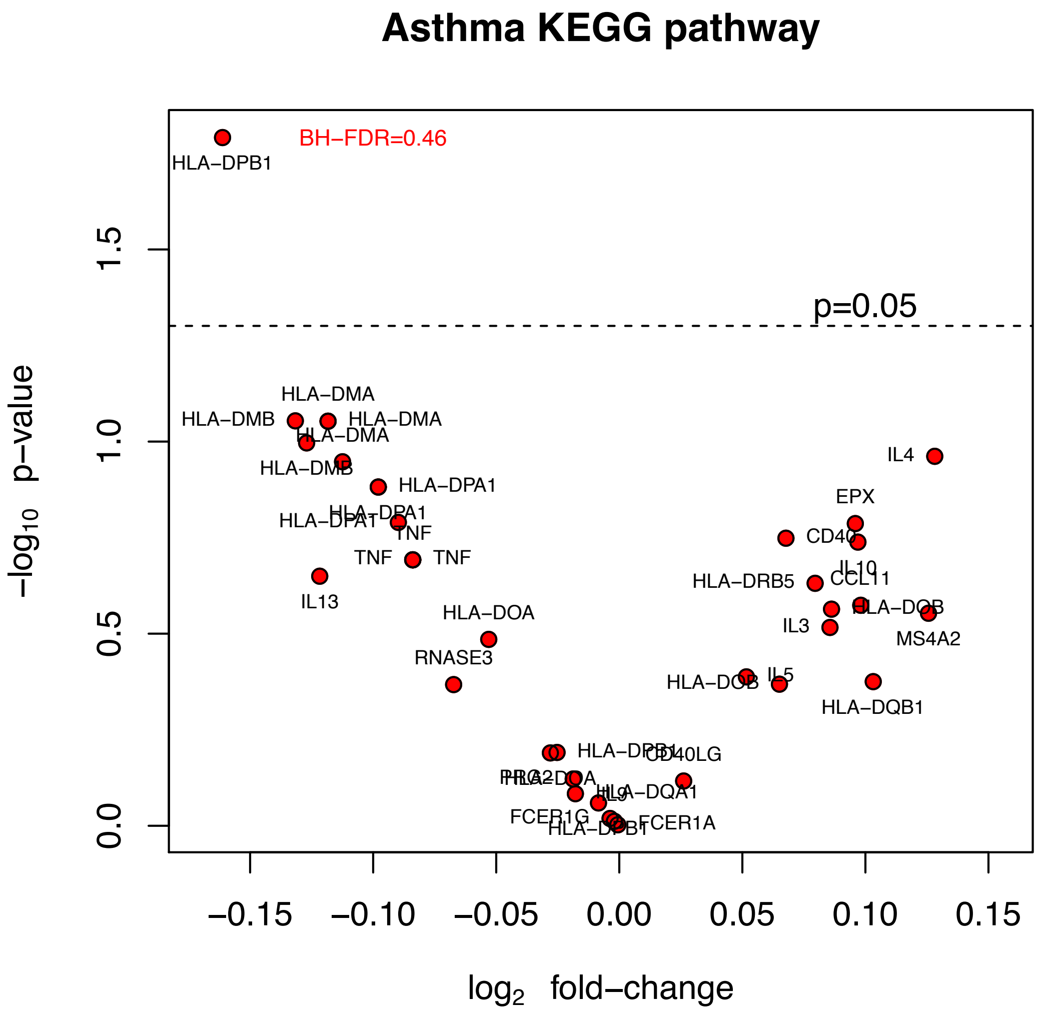
Additional File 6

Asthma KEGG pathway.



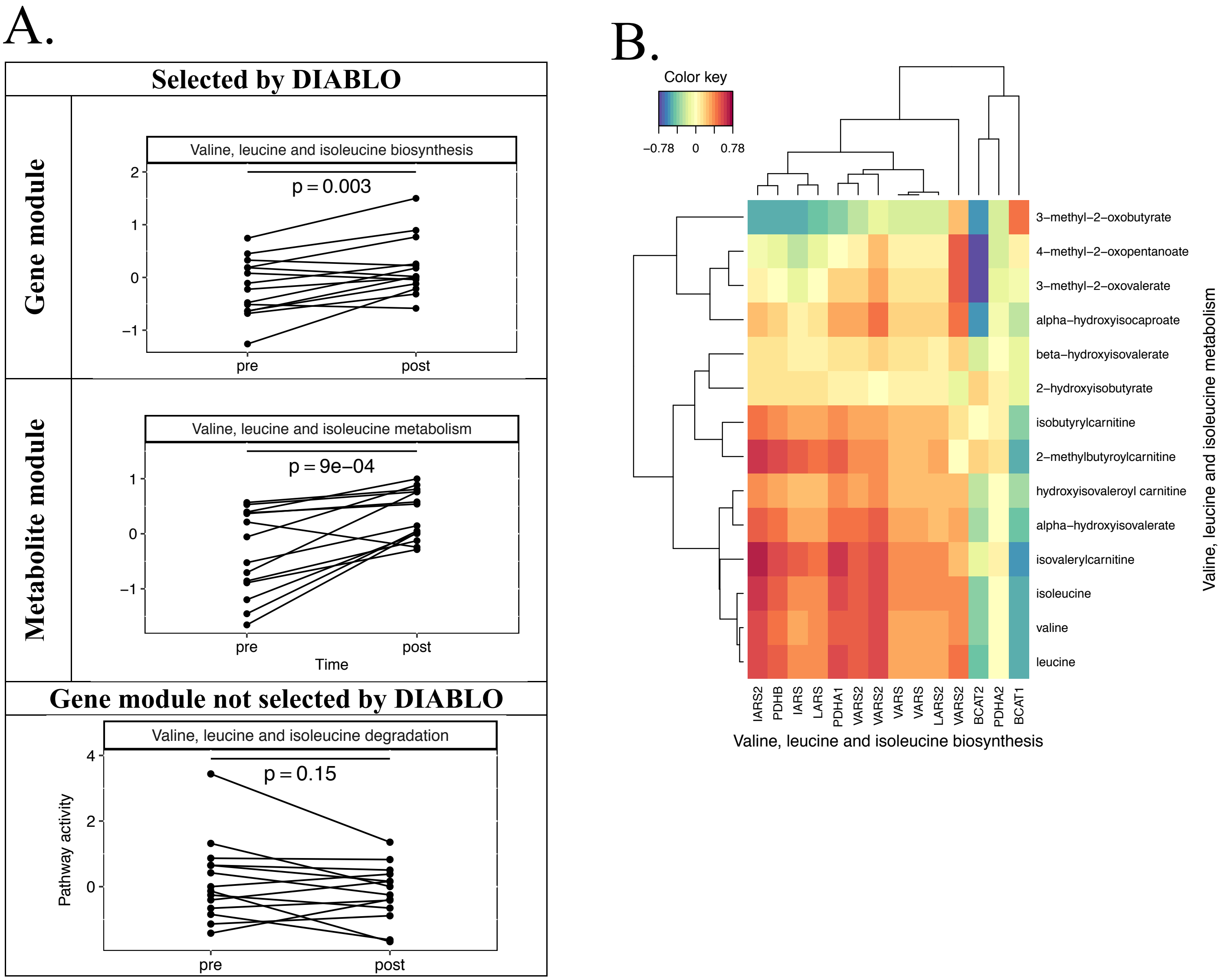
Red depicts up-regulated genes whereas green depicts down-regulate genes after allergen inhalation challenge.

Additional File 7
Volcano plot of genes in the Asthma KEGG pathway.



All genes in the asthma pathway (many not represented on the KEGG pathway diagram in Supplementary Figure 6A. The volcano plot shows that apart from HLA-DPB1 no other genes within the Asthma pathway was significant at the nominal p-value cut-off of 0.05. After correcting for multiple testing, HLA-DPB1 corresponding to an Benjamini Hochberg False Discovery Rate (BH-FDR) of 0.46.

Additional File 8
Figure S6: Valine, leucine and isoleucine gene and metabolite pathway.



A. The valine, leucine and isoleucine biosynthesis (gene-module) and valine, leucine and isoleucine metabolism (metabolite-module) were selected by DIABLO, but not the valine, leucine and isoleucine degradation pathway. B. The features genes and metabolites with the selected modules were positively correlated.