**Simulation study**

Three datasets are considered, each with 200 observations and four types of variables, 30 correlated-discriminatory variables, 30 noncorrelated-discriminatory variables, 100 correlated-nondiscriminatory variables, and 100 noncorrelated-nondiscriminatory variables.

The different types of variables were simulated as follows:

* 1. *correlated-discriminatory (corDis)*

Each n x pj dataset X(corDis)j, j = 1, 2, 3 is generated using the following model:

, where ||w||=1

For group1, the outer components u1, u2, u3 are 3-vectors draw from a multivariate normal distribution with a mean value of **–fc/2**, and the following correlation structure: cor(u1, u2) = 1, cor(u1, u3) = 1, cor(u2, u3) = 1.

For group2, the outer components u1, u2, u3 are 3-vectors draw from a multivariate normal distribution with a mean value of **fc/2** (real number), and the following correlation structure: cor(u1, u2) = 1, cor(u1, u3) = 1, cor(u2, u3) = 1.

The loadings, w1, w2, and w3 are 30-vectors, where the elements are drawn from a uniform distribution in the interval of [-0.3, 0.2] U [0.2, 0.3].

* 1. *noncorrelated-discriminatory (nonCorDis)*

Each n x pj dataset X(nonCorDis)j, j = 1, 2, 3 is generated using the following model:

, where ||w||=1

For group1, the outer components u1, u2, u3 are 3-vectors draw from a multivariate normal distribution with a mean value of **–fc/2** (real number), and the following correlation structure: cor(u1, u2) = 0, cor(u1, u3) = 0, cor(u2, u3) = 0.

For group2, the outer components u1, u2, u3 are 3-vectors draw from a multivariate normal distribution with a mean value of **fc/2** (real number), and the following correlation structure: cor(u1, u2) = 0, cor(u1, u3) = 0, cor(u2, u3) = 0.

The loadings, w1, w2, and w3 are 30-vectors, where the elements are drawn from a uniform distribution in the interval of [-0.3, 0.2] U [0.2, 0.3].

* 1. *correlated-nondiscriminatory (corNonDis)*

Each n x pj dataset X(corNonDis)j, j = 1, 2, 3 is generated by drawing 100-vectors each with 200 elements, from a multivariate normal distribution with a mean of 0 and the same correlation as the corDis variables:



* 1. *noncorrelated-nondiscriminatory (nonCorNonDis)*

Each n x pj dataset X(nonCorNonDis)j, j = 1, 2, 3 is generated by drawing 100-vectors each with 200 elements, from a multivariate normal distribution with a mean of 0 and the same correlation as the corDis variables:



The resulting list of three datasets are as follows:



Ej is a 200 x pj residual matrix where each element is drawn from a normal distribution with zero mean and variance equal to **noise**.

Using fold-change values of [0.1, 0.5, 1, 2] and noise values of [0.1, 0.2, 0.6, 1], 16 (4x4) sets of three datasets were generated, and DIABLO was applied, either with the full or null design (DIABLO\_Full and DIABLO\_Null). The full design, connects all blocks in the design matrix (C), such that cij=1, i=1,2,3 and j=1,2,3, whereas the null design does not connect any datasets in the design matrix (C), such that cij=0, i=1,2,3 and j=1,2,3. One component was retained in the DIABLO model, selecting 30 variables from each dataset for a total of 90 variables. In addition, other integrative schemes such as Concatenation and Ensembles were also tested using the sPLSDA classifier. All datasets were concatenated into one matrix containing 3x260 variables and sPLSDA was applied, retaining 1 component and 90 variables for the Concatenation-based method. For the ensemble-based scheme, a sPLSDA classifier was applied to each dataset separately retaining one component and 30 variables per dataset. The consensus predictions were determined using a majority vote scheme. A 5-fold cross-validation repeated 5 times (5x5-fold cross-validation) was used to evaluate the performance of each method/scheme and the proportion of each type of variable selected in each model was recorded.

|  |  |
| --- | --- |
| **Correlated-discriminatory scores** | **Noncorrelated-discriminatory scores** |
|  |  |
| **Correlated-nondiscriminatory scores** | **Noncorrelated-nondiscriminatory scores** |
|  |  |

Error rates

Figures/errorRates.pdf

**Figure 1. Cross-validation error rates of multi-omic integrative classifiers.** The top row depicts an average error of 50%, when the level of noise is equal or greater than the average difference between the two groups. As the fold-change increases the error rates decreases, and as the noise increases, the error rate increases across all classifiers. The performance of the DIABLO model with the full design (DIABLO\_Full) consistently has a high error rate compared to the DIABLO model with a null design (DIABLO\_Null), and the other classifiers. However, this difference becomes negligible as the fold-change increases.

Figures/selectedVariables.pdf

Figure 2. Types of variables selected in each integrative classifier. When the noise level is greater than or equal to the fold-change (top row), all classifiers mostly select nondiscrminatory variables (green and purple). As the noise level increases more nondiscriminatory variables are selected. When the fold-change is greater than the level of noise, all classifiers with the exception of DIABLO\_Full selected noncorrelated and discriminatory variables. The DIABLO model with the full design selected only variables that were discriminatory but also explained the correlation structure between blocks.

The dual functionality of DIABLO with its ability to explain the correlation structure between blocks (albeit with a compromise in the classifier performance), may identify novel biological mechanisms that are dysregulated between phenotypic groups. In order to explore this idea further, we applied these integrative classifiers to real world datasets and evaluated their biological enrichment and connectivity between variables of different omic layers.