# Supplementary Note

[Supplementary Note 1](#_Toc508923060)

[Real datasets 2](#_Toc508923061)

[Benchmarking cancer datasets 2](#_Toc508923062)

[Breast cancer multi-omics study 2](#_Toc508923063)

[Asthma multi-omics study 3](#_Toc508923064)

[Simulated datasets 3](#_Toc508923065)

[Generating multi-omics data 3](#_Toc508923066)

[Simulation analysis 4](#_Toc508923067)

[Description of methods used for the benchmarking experiments 4](#_Toc508923068)

[References 7](#_Toc508923069)

## Real datasets

### Benchmarking cancer datasets

All cancer (colon, glioblastoma, kidney and lung) datasets used for the benchmarking analyses were obtained from <http://compbio.cs.toronto.edu/SNF/SNF/Software.html> (Wang *et al.*1). For the mRNA datasets, all transcripts with the same gene symbol were averaged.

### 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 1,098 subjects for 3,703 variables. 29 unannotated transcripts were removed from the mRNA dataset composed resulting in 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 platforms, 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 1,182 samples were obtained from the TCGA staff.All datasets were restricted to samples coming from the primary solid tumor (sample type code 01) and to the first vial (vial code A).

*Normalization and pre-filtering:*The count data for the mRNA dataset was normalized to log2-counts per million (logCPM), similar to limma voom 2:

**

After library size normalization, genes with counts less than 0 were removed. The PAM50 genes were also removed from the mRNA dataset prior to analyses. Similarly, the miRNA count data was normalized to logCPM and miRNA transcripts with counts less than 0 were also removed.

### Asthma multi-omics study

*Datasets accession:* Paired blood samples were obtained from 14 asthmatic individuals undergoing allergen inhalation challenge as previously described 3. 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 4,5.

*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).

## Simulated datasets

### Generating multi-omics data

Three datasets were simulated each with 200 observations (n) and 260 variables (p). The 200 observations were split equally over two groups (G1 and G2), whereas the 260 variables were generated by varying the degree of correlation and fold-change (δ) between G1 and G2: 30 correlated-discriminatory (corDis) variables, 30 uncorrelated-discriminatory (unCorDis) variables, 100 correlated-nondiscriminatory (corNonDis) variables, and 100 uncorrelated-nondiscriminatory (unCorNonDis) variables. The resulting dataset was of the form:



The discriminatory variables (corDis and unCorDis) were generated using the following model:



where the loadings, **w**1, **w**2, and **w**3 were 30-vectors, and the elements were drawn from a uniform distribution in the interval of [-0.3, 0.2] U [0.2, 0.3]. For G1, the outer components **u**1, **u**2, **u**3 were 3-vectors drawn from a multivariate normal distribution with a mean value of -δ/2 and a mean value of δ/2 for G2. For corDis variables, cor(**u**1, **u**2) = 1, cor(**u**1, **u**3) = 1, cor(**u**2, **u**3) = 1, whereas for unCorDis variables, cor(**u**1, **u**2) = 0, cor(**u**1, **u**3) = 0, cor(**u**2, **u**3) = 0.

The nondiscriminatory variables (corNonDis and unCorNonDis) were generated by drawing 100-vectors each with 200 elements, from a multivariate normal distribution with a mean of 0. For corNonDis variables, cor(**u**1, **u**2) = 1, cor(**u**1, **u**3) = 1, cor(**u**2, **u**3) = 1, whereas for unCorNonDis variables, cor(**u**1, **u**2) = 0, cor(**u**1, **u**3) = 0, cor(**u**2, **u**3) = 0.

***E****j* is a 200 x 260 residual matrix where each element is drawn from a normal distribution with zero mean and variance according to the grid [0.1, 0.2, 0.6, 1]. The following grid of values were used for the fold-change: [0.1, 0.5, 1, 2].

### Simulation analysis

Using fold-change values of [0.5, 1, 2] and noise values of [0.2, 0.5, 1, 2], 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 (across all datasets). In addition, other integrative schemes such as concatenation and ensemble-based classifiers were also tested using the sPLSDA classifier. For the concatenation-based scheme, all datasets were concatenated into one matrix containing 3x260=880 variables and sPLSDA was applied, retaining 1 component and 90 variables. 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 10-fold cross-validation averaged over 50 simulations was used to evaluate the performance of each method/scheme and the number of each type of variable selected in each model was recorded.

## Description of methods used for the benchmarking experiments

For the purposes of this study, only component-based methods that integrated multiple datasets and perform variable selection were considered. Since tuning the number of variables to retain in each model would result in biomarker panels with different numbers of variables, for the purposes of this study all variables were retained in each model. The features were instead ranked based on their absolute value of their loadings (importance) and 60 variables were selected from each omic type, resulting in multi-omic biomarker panels with 180 variables (60 mRNAs, 60 miRNAs and 60 CpGs). Equal numbers of variables allowed for a fair comparison in the gene set enrichment analysis.

|  |  |
| --- | --- |
|  | **Parameter settings** |
| **Supervised** | |
| DIABLO\_null | ncomp = 2 (# of components)  keepX = all variables were retained from each omics dataset    default parameters were used for the other arguments:  scheme="horst",  mode="regression",  scale = TRUE,  init = "svd",  tol = 1e-06,  max.iter = 100 |
| DIABLO\_full | ncomp = 2 (# of components)  keepX = all variables were retained from each omics dataset    default parameters were used for the other arguments:  scheme="horst",  mode="regression",  scale = TRUE,  init = "svd",  tol = 1e-06,  max.iter = 100 |
| Concatenation-sPLSDA | ncomp = 2 (# of components)  keepX = all variables were retained from each omics dataset  default parameters were used for the other arguments:  mode = "regression"  scale = TRUE,  tol = 1e-06,  max.iter = 100 |
| Ensemble\_sPLSDA | ncomp = 2 (# of components)  keepX = all variables were retained from each omics dataset  default parameters were used for the other arguments:  mode = "regression"  scale = TRUE,  tol = 1e-06,  max.iter = 100 |
| **Unsupervised** | |
| sGCCA6 | ncomp = 2 (# of components)  keepX = all variables were retained from each omics dataset    default parameters were used for the other arguments:  scheme = "horst",  mode="canonical",  scale = TRUE,  init = "svd.single",  tol = .Machine$double.eps,  max.iter=1000, |
| JIVE\*7 | default parameter settings from the jive() from the r.jive R-package were used:   1. scale = TRUE, center = TRUE 2. method = “perm”   sPCA parameters:  ncomp = 2 (# of components)  keepX = rep(ncol(X),ncomp)(all variables were retained from each omics dataset  default parameters were used for the other arguments:  center = TRUE  scale = TRUE,  max.iter = 500,  tol = 1e-06 |
| MOFA8 | factors=2 (# of components)  default parameter settings recommended by MOFA were used:   1. likelihoods=( gaussian gaussian gaussian ) 2. Convergence criterion (tolerance=0.01, nostop=0) 3. Training components (startDrop=1 # initial iteration to start shutting down factors, freqDrop=1 # frequency of checking for shutting down factors, dropR2=0.00 # threshold on fraction of variance explained) 4. hyperparameters for the feature-wise spike-and-slab sparsity prior [learnTheta=( 1 1 1 ) # 1 means that sparsity is active whereas 0 means the sparsity is inactivated; each element of the vector corresponds to a view, initTheta=( 1 1 1 ) # initial value of sparsity levels (1 corresponds to a dense model, 0.5 corresponds to factors ); each element of the vector corresponds to a view, startSparsity=250 # initial iteration to activate the spike and slab, we recommend this to be significantly larger than 1]   Intercept was set to TRUE (learnIntercept=1) |

\*since the variable selection functionality has not been added to JIVE R-function, sparse Principal Component Analysis (sPCA) from the mixOmics R-package was applied to the joint variation matrix obtained after applied JIVE to the multi-omics cancer datasets.

## References

1. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11,** 333–337 (2014).

2. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15,** R29 (2014).

3. Singh, A. *et al.* Plasma proteomics can discriminate isolated early from dual responses in asthmatic individuals undergoing an allergen inhalation challenge. *PROTEOMICS - Clin. Appl.* **6,** 476–485 (2012).

4. Singh, A. *et al.* Gene-metabolite expression in blood can discriminate allergen-induced isolated early from dual asthmatic responses. *PLoS ONE* **8,** e67907 (2013).

5. Singh, A. *et al.* Th17/Treg ratio derived using DNA methylation analysis is associated with the late phase asthmatic response. *Allergy Asthma Clin. Immunol.* **10,** 32 (2014).

6. Tenenhaus, A. *et al.* Variable selection for generalized canonical correlation analysis. *Biostatistics* **15,** 569–583 (2014).

7. Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7,** 523–542 (2013).

8. Argelaguet, R. *et al.* Multi-Omics factor analysis disentangles heterogeneity in blood cancer. *bioRxiv* 217554 (2017).