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

9th April 2018

Dear Editor of Genome Biology,

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

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

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

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

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

The submitted manuscript has been approved by all authors and has not been submitted to any other journal. This manuscript is a substantial revised version to our previous submission to *Genome Biology*, **GBIO-D-16-01112**. We improved the method and added four more case studies to benchmark the method to address the reviewers' comments. We provide a point-by-point response to reviewers in the next section. We look forward to your reply.

Yours sincerely,
Dr. Kim-Anh LÊ CAO

Reviewer #1:

The article has several strengths:
a) The article is very well written and provides a good overview of various statistical methods for analyzing
genomic data.
b) I think it presents an honest analysis of the data. The authors resist the temptation to oversell their method.
They acknowledge that their method does not outperform existing methods when it comes to accuracy.
c) The authors have implemented the method in an R package
d) This is a multi-omic method that integrates data.
e) The authors apply their method to both empirical data and to simulated data.

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

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

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

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

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

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

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

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

Overall, I am not sure how much biology can be learnt by applying this method. Bottom line: this predictive method does not seem to improve predictive accuracy.

We have refocused our manuscript on biological insights primarily, rather than prediction performance as the former was our main motivation in driving methodological developments. By extending our analyses with six multi-omics studies including two case studies in human breast cancer and asthma we believe we have demonstrated that data integration performed using appropriate computational methods generate new biological insights and novel hypotheses to be further tested in the laboratory. The important contribution of DIABLO is its resulting molecular signatures that both explain the correlation structure across multiple biological domains and discriminate multiple phenotypic groups, with increased biological enrichment compared to other methods.

Reviewer #2:

This is a well written article which addresses an important need in the field. 1) In the introduction, the longer intro to sparse CCA should be provided. In the methods the actual method is more clearly stated "DIABLO extends sparse gCCA to a classification framework".

We thank the reviewer for their appreciative comments. In the revised version of the manuscript we provide a clearer explanation of our method DIABLO (see lines 99-116).

*"DIABLO (**D**ata **I**ntegration **A**nalysis for **B**iomarker discovery using **L**atent c**O**mponents) maximizes the common or correlated information between multiple omics (multi-omics) datasets while identifying the key omics variables (mRNA, miRNA, CpGs, proteins, metabolites, etc.) and characterizing the disease sub-groups or phenotypes of interest. DIABLO uses Projection to Latent Structure models (PLS) [1], and extends both sparse PLS-Discriminant Analysis [2] to multi-omics analyses and sparse Generalized Canonical Correlation Analysis [3] to a supervised analysis framework. In contrast to existing penalized matrix decomposition methods [4], DIABLO is a component-based method (or a dimension reduction technique) that transforms each omic dataset into latent components and maximizes the sum of pairwise correlations between latent components (user-defined) and a phenotype of interest [5]. DIABLO is, therefore, an integrative classification method that builds predictive multi-omics models that can be applied to multi-omics data from new samples to determine their phenotype. Users can specify the number of variables to select from each dataset and visualize the omics data and the multi-omics panel into a reduced data. The method is highly flexible in the type of experimental design it can handle, ranging from classical single time point to cross-over and repeated measures studies. Modular-based analysis can also be incorporated using pathway-based module matrices [6] instead of the original omics matrices, as illustrated in*

one of our case studies."

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

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

Currently, our method does not account for completely missing observations or variables.
Random missing values are allowed in the dataset matrices as local regressions are fitted in the model and missing values will be omitted when calculating the latent components and loading vectors. The prediction step however, as highlighted in the Breast Cancer case study can be performed with an entire dataset missing.
l 401: '*As the class prediction relies on individual vote from each omics set, DIABLO allows for some missing datasets X_k during the prediction step, as illustrated in the Breast Cancer case study.*'
l209: '*The training data consisted of four omics-datasets (mRNA, miRNA, CpGs and proteins) whereas the test data included all remaining samples for which the protein expression data were missing.*'
As such we do not think this is an information that should appear all throughout the document.

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

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

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

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

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

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

"Finally, DIABLO, like other methods we benchmarked, will be affected by technical artifacts of the data, such as batch effects and presence of confounding variables that may affect downstream integrative analyses. Therefore, we recommend exploratory analyses be carried out in each single omics dataset to assess the effect, if any, of technical factors and use of batch removal methods prior to the integration analysis [8–10]*."*

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

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

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

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

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

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

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

We provide a description of eigengene summarization in the revised manuscript (lines 504-509). *"**Modular analysis:** Eigengene summarization is a common approach to decompose a n x p dataset (where n is the number of samples and p is the number of variables in a module), to a component (linear combination of all p variables) that represents the summarized expression of genes in the module [6]. For the asthma study, 15,683 genes were reduced to 229 KEGG pathways and 292 metabolites were reduced to 60 metabolic pathways using eigengene summarization."*

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

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

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

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

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

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

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

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

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

13) is the analysis effective by the number of variables. For example if dataset A has several thousand variables and dataset B has less than 50, would this impact the analysis?
The difference in the number of variables in each dataset should not impact the analysis as each dataset is summarised by its own set of latent components, so that components across data sets are maximally correlated, irrespective of how many variables there are in each dataset. This makes DIABLO a much more attractive solution than a concatenation method, where datasets than include a large number of variables tend to be more 'favoured' in the molecular signature compared to smaller data sets.

14) p25, The de-duplication effort in GSEA is important and should be clear to users, If a more stringent assignment were used, would this impact results?
GSEA is impacted by the number of features that are input into the analysis and the types of gene sets that are used to determine biological enrichment. In the revised manuscript we include a benchmarking experiment where we constructed multi-omic biomarker panels of equivalent number of features with a total of 180 features. Further we tested 10 different gene set databases, from Molecular signature database[12], blood transcriptional modules[13] and cell-specific expression from Benita *et al*. [14].

15) p 27. Data Processing. Were 3,073 BRCA clinical variables used in this study? The PAM50 assignments for tumors (obtained from TCGA staff) should be made available together with the filtered TCGA data, such that others can reproduce this work.
The 3,073 variables listed describe the data that were obtained from TGCA. From the clinical data, only the PAM50 labels and sample-type variables were used. The complete code and data files can be found with the github repository (https://github.com/singha53).

16) p28. Terms in the Voom equation are not fully defined. Filtering removed "genes with counts less than 0". Does this mean the sum of the gene across all tumors was zero, or that any gene which has a zero tumor in any 1 tumor was excluded?
We have clarified the following in the revised manuscript (**see Supplementary Data file**). The count data for the mRNA dataset, X_{counts} was normalized to log2-counts per million (logCPM), X_{norm}, similar to limma voom [15]:

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

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

17) p28 Asthma study. Genes were reduced 229 KEGG pathways and metabolites were reduced to 60 pathways. Why were variables reduced to GeneSet. The rational and need for this is not *explained. Was it simply to aid biological interpretation of the data or was it for computational reasons?*

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

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

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

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

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

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

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

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

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

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

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

We decided to focus our comparisons mainly on the correlation and discrimination structure between datasets (simulation study), the effect of the design matrix on the types of variables selected and whether this led to superior biological enrichment compared to other integrative strategies (benchmark study, four new real multi-omics datasets). Computational times are provided for different scenarios in our article (Rohart et al, 2017, Plos Computational Biology 13 [11] in the main and supplemental material) for various numbers of variables and data sets, see the screenshots below.

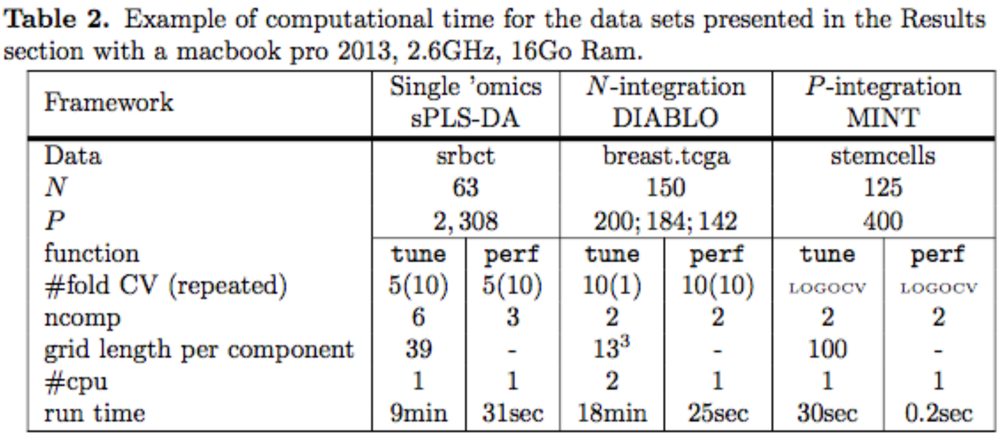

Table 2. Example of computational time for the data sets presented in the Results section with a macbook pro 2013, 2.6GHz, 16Go Ram.

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

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

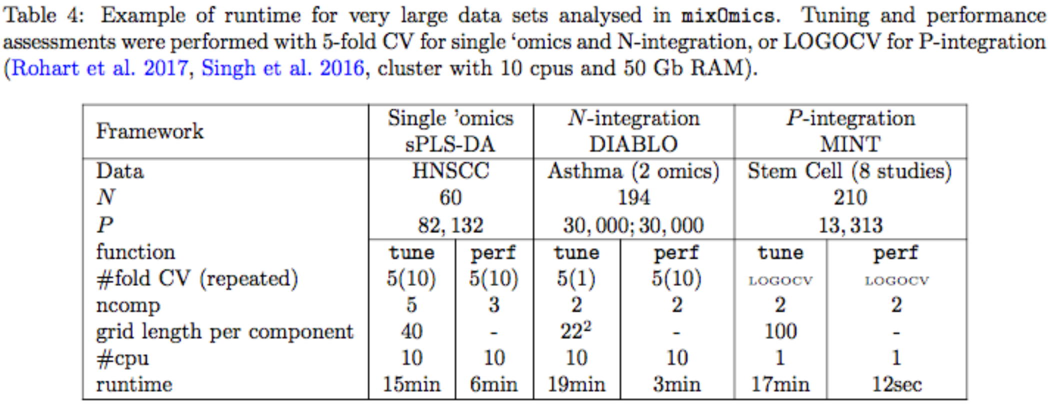

Table 4: Example of runtime for very large data sets analysed in `mixOmics`. Tuning and performance assessments were performed with 5-fold CV for single 'omics and N-integration, or LOGOCV for P-integration (Rohart et al. 2017, Singh et al. 2016, cluster with 10 cpus and 50 Gb RAM).

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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