Dr. Kim-Anh Lê Cao

Senior Lecturer, Statistical Genomics

NHMRC Career Development Fellow

School of Mathematics and Statistics

Centre for Systems Genomics Bld 184

The University of Melbourne | VIC 3010

T: +61 (0)3834 43971 | kimanh.lecao@unimelb.edu.au

27th March 2018

Dear Editor of Genome Biology,

We wish to submit our manuscript “DIABLO: from multi-omics assays to biomarker discovery, 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 feel 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 facing 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 multi-steps methods that either concatenate or combine the analyses from each data set separately, 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 DIABLO 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 omic data sets 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 multi omics data currently being generated and push novel biological discoveries of 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, along with detailed tutorials on the 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. A previous submission to *Genome Biology*, GBIO-D-16-01112 was rejected from Genome Biology after revisions. We have carefully considered the reviews of the reviewers and have considerably revised the current version of the manuscript. We provide a point-by-point response to reviewers in the next section. We look forward to your reply.

Yours sincerely,

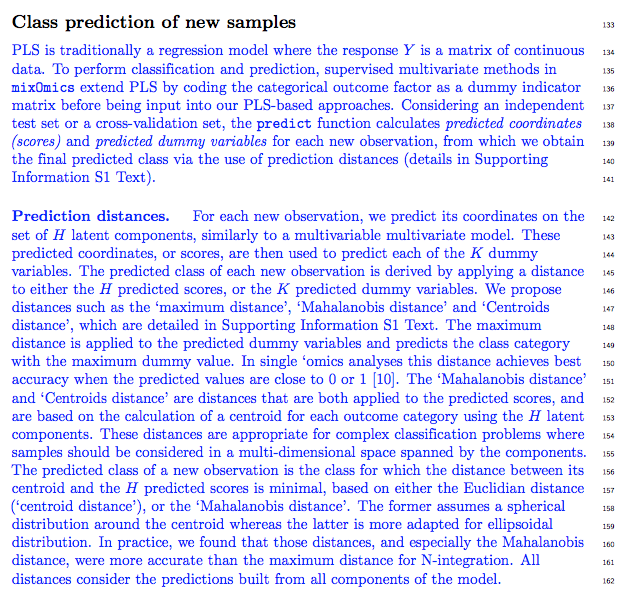
Dr. Kim-Anh LÊ CAO

**Reviewer #1: General: This is a well written manuscript addressing a very important topic. The software, which the authors have compiled, is well designed and extremely useful, and I think it is a good idea to make it familiar to a wider audience via this paper. Nonetheless, I see a number of smaller points that should be addressed before publication:**

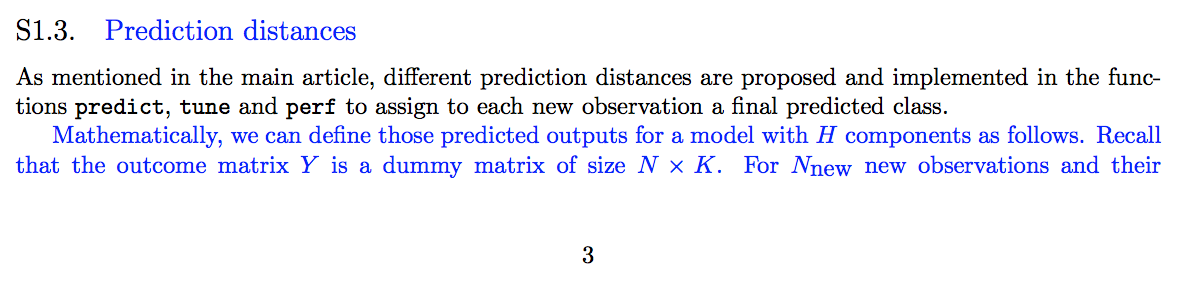
We appreciate the positive overall comments from the reviewer and the careful review. We hope to have adequately addressed the issues raised, as detailed below. The package has been updated to version 6.2.0 to reflect those changes and is available from the CRAN.

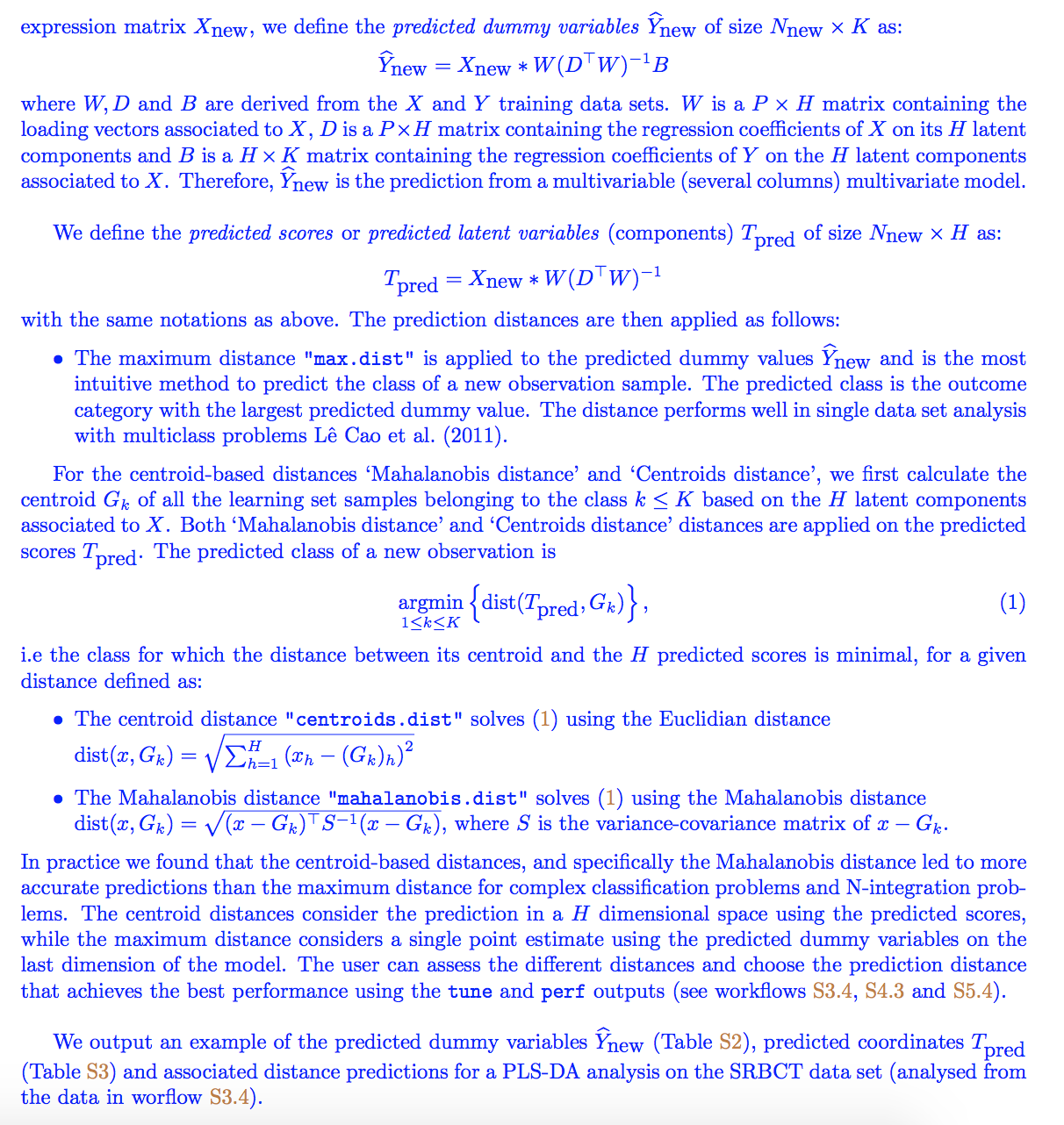
**Major**:  
**1. Prediction distances: This topic could be explained in more detail from my point of view. How is the "predicted score" defined exactly? What are the specific assumptions and motivations underlying the different prediction methods?**

Prediction is one of the key feature of our supervised methods. We have clarified the two main prediction outputs, the predicted dummy variables and the predicted latent variables or scores. In the supplemental material **S1.3**, we mathematically defined those outputs and how the prediction distances are applied. In the **main manuscript**, we expanded on the different types of distances, how they are applied to either of the predicted outputs and when those distances perform well in practice.

The section ‘**Class prediction of new samples**’ now reads:

The supplemental material section ‘**S1.3.** **Prediction distances’** was extensively detailed and Tables S2 and S3 were added for further illustration.

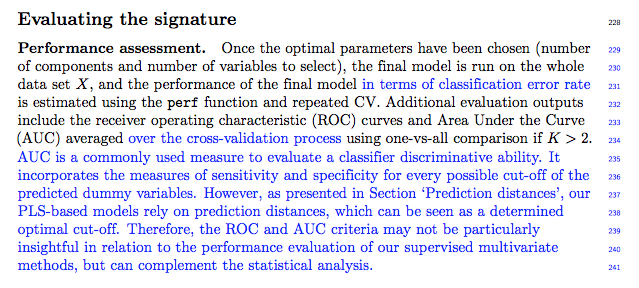
****

**  
2. Performance evaluation (p 5): Why is the area under ROC curve not considered as a popular performance measure?**

Evaluating biomarkers discriminative abilities with other criteria than statistical significance classically relies on the area under the ROC curve. The AUC incorporates the measures of sensitivity and specificity for every possible cut-off of the prediction scores. The AUC is not particularly informative in relation to our PLS-type models that rely on distances to predict classes and determine an optimal cut-off (as presented above). Therefore, the classification error rates proposed in our method and the AUC measure different aspects of the classifier performance and the AUC does not reflect the true performances of a PLS-type model.

As a side note, AUC is often calculated on the training set directly, and not on a cross-validated or independent test sets. Such use of AUROC leads to overfitting, and thus over-optimistic performances. We avoid this practice in mixOmics as the AUC is calculated on cross-validation sets in our perf function. The current AUROC outputs complement the evaluation of the discriminative ability of our models but are not used as prediction thresholds in the package.

In the manuscript we added a new section **‘Evaluating the signature’**, where we clarified:



See also our answer to issue 8.

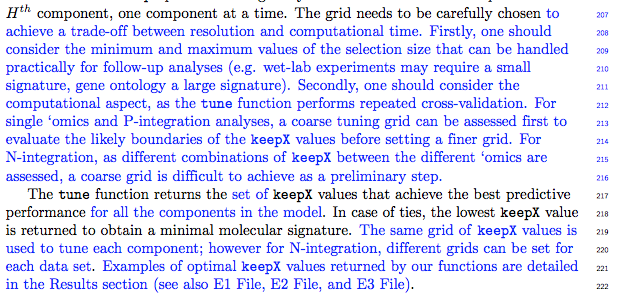
**3. l1-penality (p 6): Why is it reasonably (from the perspective of the authors) to assume the same number of relevant features for each canonical covariate?**

We assume the reviewer refers to the paragraph:

*‘[…] mixOmics uses soft-thresholding to improve usability by replacing the l1 parameter by the number keepX of features to select on each dimension. The performance of the model is assessed for each value of keepX provided as a grid by the user from the first component to the Hth component, one component at a time. […] Note that the same grid of keepX values is used to tune* ***each component***’

Although the grid of keepX values is the same on each component, the resulting number of chosen variables can be different for each component, as can be observed in the sPLS-DA analysis (Electronic Supporting file **E1.5**) where 8 genes were selected on the first component, then 290 on the second and 30 on the third.

We clarified this in the paragraph ‘**l1 penalty or the number of features to select’:**



**Furthermore: Are there any guidelines, which values of keepX to try in practice?**

We can choose the keepX tuning grid based on several criteria:

1/ user expectation: the user arbitrarily chooses to bound the grid with the maximal number of features that can be handled in the follow-up analyses, e.g. gene ontology analyses (usually a large number of features), wet-lab experimental validations (usually a small number of features) etc.

2/ practical computational aspects: for ‘simpler’ models such as sPLS-DA and MINT P-integration, it is best to try a coarse grid to first assess the boundaries of the grid, before running a much finer model. For DIABLO N-integration, since they are many combinations to assess, we often rely more on strategy 1.

We added some guidelines in the main text, see screenshot in point 3 above.

**Finally: Is it really necessary to re-calculate the entire model, if another value of keepX is tested or could it be possible to update a previously calculated model with a new keepX (i.e. using a 'warm start')?**

We thank the reviewer for the suggestion. Our PLS-based models sPLS-DA and MINT start with Singular Value Decomposition and converge quickly, which reduces the need for a warm start. We are unsure about the advantage of a warm start for our sparse approaches as the solution to a model with 10 selected variables might not be closer to the solution to a model with, say, 20 selected variables (those 10 variables might not be included in the 20). Nonetheless, we will investigate this option in future versions of mixOmics that will focus on improving computational time.

**4. Tuning with constraint method: This paragraph is not clear to me at all. Please explain better the difference to the method in the previous paragraph.**

We apologize for the lack of clarity. As presented in section ‘l1 penalty or the number of features to select’, the tuning is performed based on the *number* of features keepX to retain in the model, as specified in the tuning grid, rather than the actual selected features. The selected features are returned from the final multivariate model run on the chosen parameters (see example in in Electronic files **E1.4 & E1.5**).

Originally, the constraint model allowed to perform the tuning based on a list of features selected from the previous components, while assessing the keepX value on the next component. However, after additional analyses for this revision we assessed a risk for overffiting. We hence decided to remove this option in the updated version of the package 6.2.0 and updated all results, figures and graphical outputs accordingly.

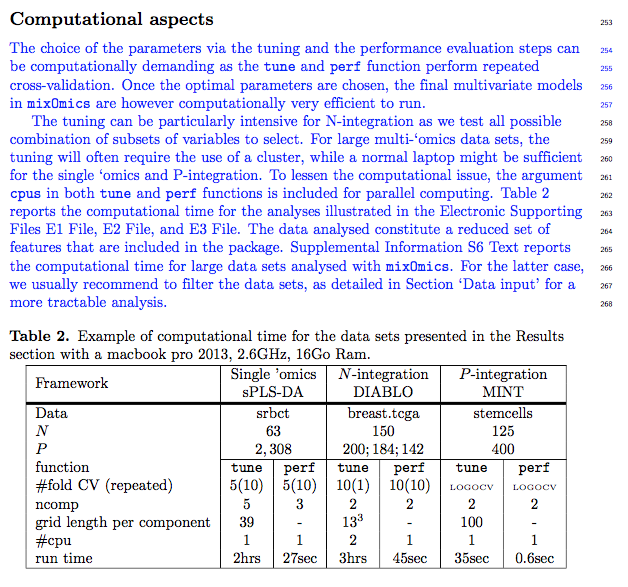
**5. The authors should elaborate a bit more about parallel computing support given the high computational demand of their approach.**

The methods implemented in mixOmics can be computationally demanding but for two specific cases only:

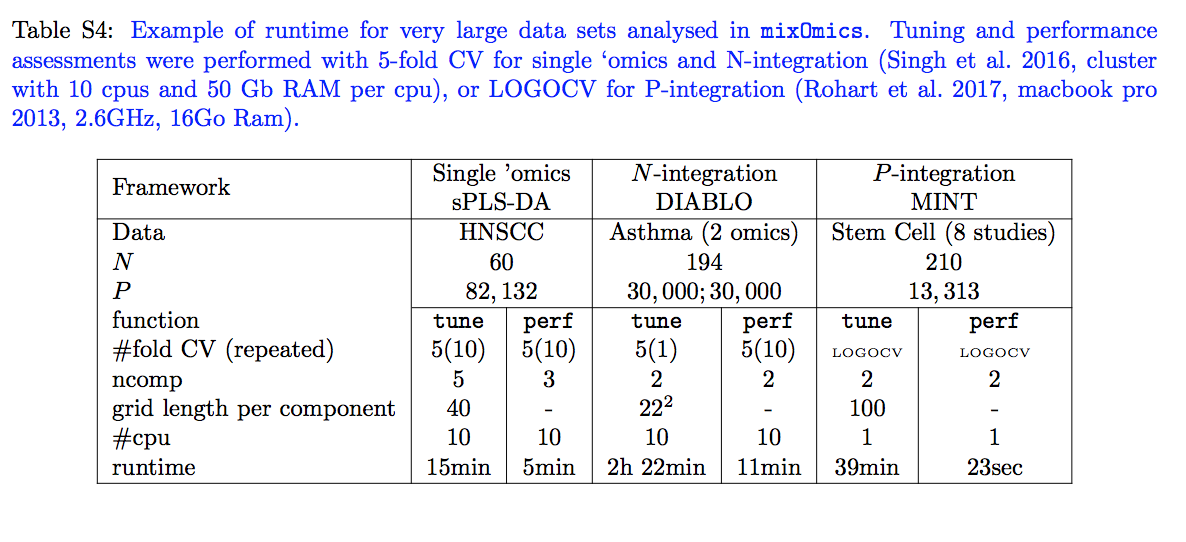
1/ to tune the parameters of the model as optimally as possible with the tune function.

2/ to assess the performance of the final model as accurately as possible, with the perf function.

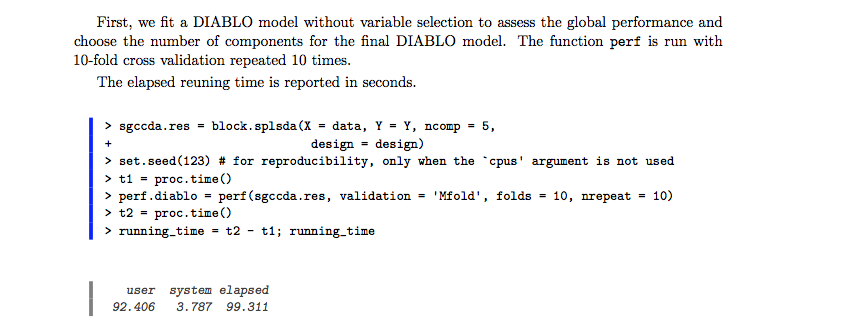
We reported the computational time for the data sets analysed in each of our Electronical Files **E1-E3** and in a new Section ‘**Computational aspects’** and **Table 2.**

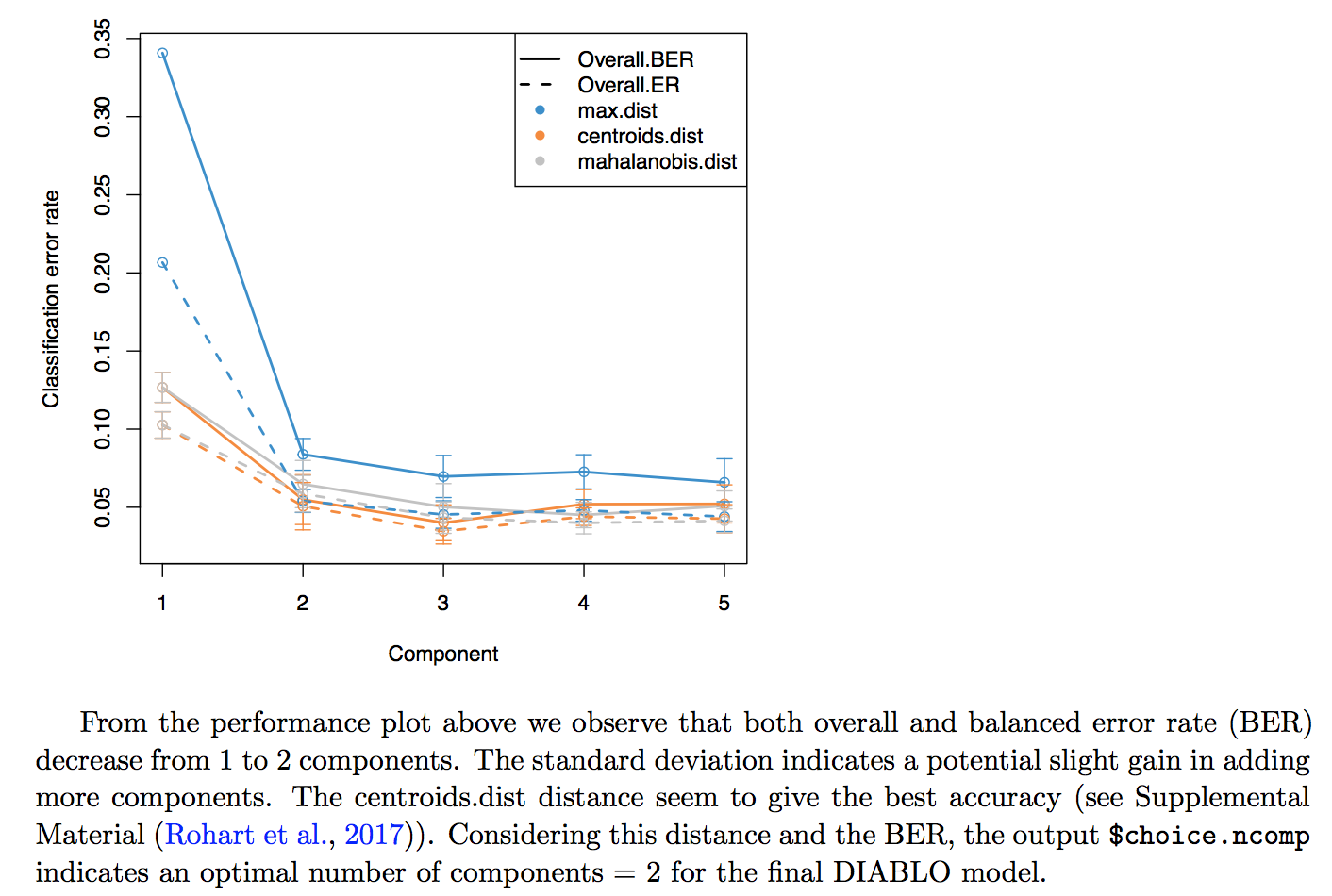


The data sets presented in the package are however quite small given the restriction of the R CRAN (< 5Mo per package). To complement this, we added a new section in Supplemental Material **S6** ‘**Computational time for large data sets’ Table S4** where we report the computational time for very large data sets (run on a cluster for all analyses except P-integration).

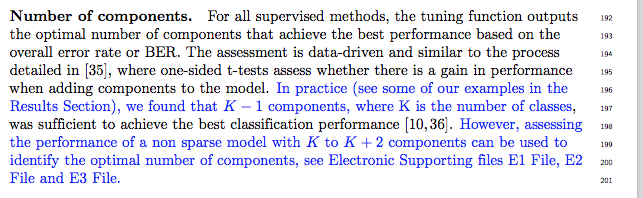
**6. p 8, Results visualisation: How was the parameter keepX tuned in that example? How many latent variables were selected and why? What were the relative contributions of the original miRNA and mRNA features to the latent variables, i.e. how did the loadings look like?**

We clarified this concern in all **Electronic files** and in the manuscript. For the example cited here p 8 for the N-integration (DIABLO) example, we first assessed the optimal number of components on a non-sparse model (section **E2.2**) using the perf function. The results indicated that 2 components were sufficient to continue on with the DIABLO analysis.

Here is a screenshot of **E2.2**:



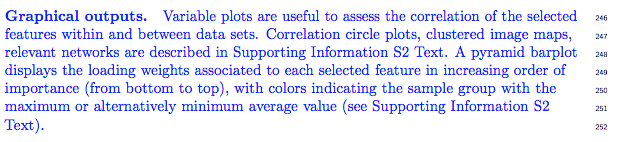
To clarify the tuning step, we added the following in the ‘**Number of components’** section in the main manuscript:



All electronic files were updated to follow the same procedure.

The loadings of all selected variables are displayed in **Supp E2.4 p11.** For the sake of conciseness, we did not show them in this particular example, but a similar loading plot is displayed in Figure 5 for P-integration with MINT.

In addition, a new paragraph was added in the main manuscript ‘**Evaluating the signature’** section untitled ‘**graphical outputs’:**

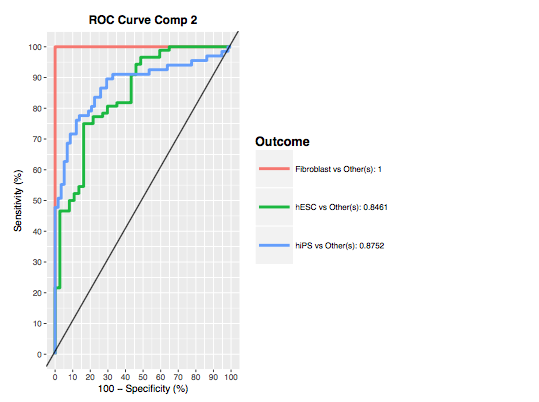
**  
7. p 9, Choice of parameters: Can the choice of 2 latent components be better explained? It looks rather ad hoc in the way it is described right now.**

We addressed this comment in the changes listed in point 6 above.

**8. p 10, Illustration of MINT analysis: Why don't the authors add the AUC as another performance measure?**

We have discussed in point 2 why AUROC is an unrelated measure to our multivariate models. However, in order to obtain consistent outputs, we added the AUC measures per study in MINT and the ROC curves can be plotted as displayed in **E3.3.6** ‘**Performance assessment and prediction’**. Those new options are now available in mixOmics 6.2.0.

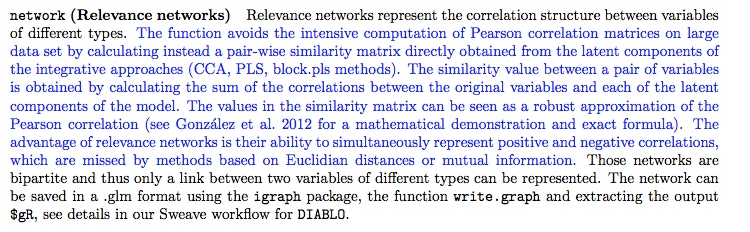
**AUROC from E3.3.6:**



**9. S 2.5 (Supplements): How does the calculation of "relevance networks" take into account that there can be more than one latent component? Please add a more detailed explanation.**

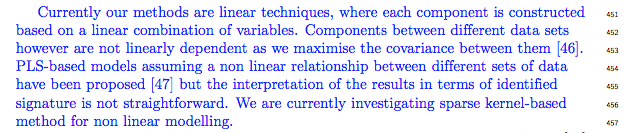
We calculate a similarity matrix where we sum the associations across all components. We originally stated this in the supplemental: ‘*A similarity matrix representing the association between pairs of variables across all components is calculated as the sum of the correlations between the original variables and the latent components across all dimensions of interest*’.

We added more details in the Supplemental, which now reads:



**9. p 10, Discussion: I think there are two important points missing in the discussion:  
a) mixOmics relies on PLS and CCA, which are linear techniques. Hence, non-linear relationships cannot be detected. This is ok, but should be made clearer.**

We thank the reviewer for the suggestion. This is indeed an important aspect that we added in the discussion.

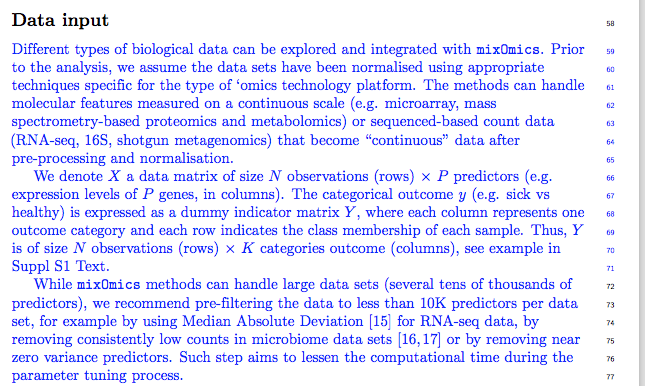


A first version of the kernel approach we mention is presented on our website as a sister package with our collaborators: <http://mixomics.org/mixkernel/> We are planning further testing and developments before the module is included in mixOmics.

**b) There should be some more details around NGS based data: How does mixOmics e.g. deal with RNASeq data? Do the authors recommend a transformation of the data (e.g. via limma-voom) and if yes, which one?**

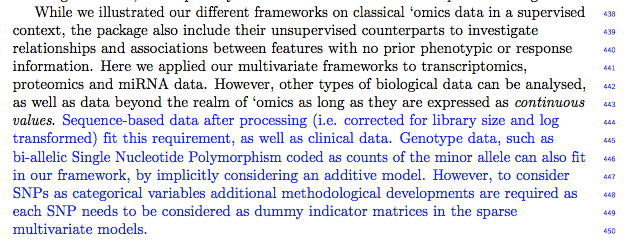
mixOmics attractive property is that it can deal with a variety of biological data sets. As each omics technology requires a specific type of processing and normalisation that evolve (sometimes too) quickly with advances in technology, the package cannot exhaustively provide all types of normalisation. We therefore request the user to choose the appropriate normalisation method and processing suitable to their own data set(s).

We have address this important point in a new section of the manuscript ‘**Data Input’**



**Which recommendations would the authors have for genetic variant calls (e.g. SNPs, CNVs) and integration with gene expression?**

So far, genetic variant calls analysis has been left out of the package as such data require specific theoretical and methodological developments. We are in the process of extending Multiple Correspondence Analysis with a sparse version for SNP data, which will eventually enable the integration with other `omics data sets. We have mentioned the limitations and future work in the Discussion section



**Reviewer #3: MixOmics is a great R package and a significant contribution to the field. This is a well-written article, with clear attractive figures. It provides an excellent tutorial on the methods available in MixOmics and provides an in-depth review of the package. This is a well-written review/tutorial and deserved to be published in a high profile journal.**

We thank the reviewer for such positive comments. The manuscript summarises many years of methods development and implementation to make the package available to the computational biology community. **Currently within the review system, this article is listed as a research article and I don't think it meets the journal criteria for publication as research article, as it provides few novel findings. The R package has been introduced in previous research papers. The author states "All methods have been published, with the exception of DIABLO (bioRxiv manuscript currently being redrafted..).". The datasets analyzed in this article are existing datasets and few novel findings are presented.**

**I would be fully-supportive of this manuscript if submitted as a software submission and I have indicated to the editor that it should be a software submission.**

**Therefore I will also provide a review of it as a software submission.**

The cover letter was a follow-up from a pre-submission enquiry letter which enquired about a software research article. We thus submitted the present article as a software submission, even though it may not be explicit from the online submission procedure from the journal. **The supplement includes the code and data to re-generate all of the figures. I ran the MINT, DIABLO, and PLSDA code and these worked without error.**

We thank the reviewer to take the time to test our R scripts that will be provided as a useful resource to our readers and users. As detailed in our answers to reviewer #1, we have updated the mixOmics package to version 6.2.0 available on the CRAN as well as all the Electronic Files, Results section and Figures. **1) For archival purposes, it would be nice to create this as a R package, which will handle software dependencies.**

All software dependencies are handled internally with the mixOmics package, and as such we do not feel the need to create an extra package. Instead we propose to include directly the three example ‘vignettes’ on our website, at the following links

<http://mixomics.org/case-studies/srbct-example/>

<http://mixomics.org/mixmint/stemcells-example/>

<http://mixomics.org/mixdiablo/tcga-example/>

Those links are in addition the ‘bundled’ Sweave and R codes available on our website <http://mixomics.org/presentations/publications/>

**2) Please provide a short discussion of related R packages (pcaMethods, RGCCA, moGSA, PTAk, OmicsPLS, etc)**

We thank the reviewer for this suggestion. The recent and excellent review from Meng et al., 2016 (Briefings in Bioinformatics 17 (4) <https://academic.oup.com/bib/article/17/4/628/2240645/Dimension-reduction-techniques-for-the-integrative> ) give an extensive list of all multivariate R packages and functions for dimension reduction and data integration. We referred to the review, and listed a few methods and packages that propose similar approaches to those implemented in mixOmics for feature selection. None of those methods, with the exception of factomineR (no feature selection) and ade4 provide visualisation of the selected features.

The discussion now reads:

