

Dr. Kim-Anh Lê Cao
Snr Lecturer, Statistical Genomics
School of Mathematics & Statistics
Melbourne Integrative Genomics
The University of Melbourne VIC 3010
T: +61 (0)3834 43971
@: kimanh.lecao@unimelb.edu.au

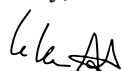
Dec 19, 2018

Dear Editor,

Please find attached our second revision to our manuscript 'DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays' as a research article for the Systems Biology category in Bioinformatics. You can find below our answers to the minor comments raised by the reviewers.

We look forward to your reply.

Yours sincerely,



Dr. Kim-Anh LÊ CAO

12-Dec-2018

Manuscript ID: BIOINF-2018-1115.R1

Title: DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays

Here are the Associate Editor's comments:

Please address the remaining minor issues in full before we make our final decision.

Here are the comments of the reviewers:

Reviewer: 1

Comments to the Author

The authors present a strongly revised manuscript that provides a clear and focused description and sound validation of the proposed method. This constitutes a valuable addition to the toolset available for multi-omics integration. All my previous comments have been well addressed by the authors.

Minor comments:

The following suggestions could be incorporated to further improve the structure and appearance of the text and the supplement:

- Consider merging Sections 3.3 and 3.4 with a joint description of evaluation on the test data. Currently, the correspondence of the numbers in Table S2 (classification error of 0.21) and the

text of Section 3.3 (BER of 22.9%) as well as the numbers of selected variables per omic (with/without proteins) is unclear.

We agree with the reviewer that the current presentation of the results can be confusing to readers given the discrepancy in the results and Table S2. This is because those error rates do not refer to the same analyses.

3.3 DIABLO identifies known and novel multi-omics biomarkers of breast cancer

The optimal DIABLO model is tuned using **mRNA, miRNA, CpGs and proteins** in the training dataset and applied to the test data (BER = 21%)

3.4 Competitive classification performance of DIABLO

Unlike DIABLO and ensemble classifiers, Concatenation-based classifiers require all data to be present for the training and test cohorts. Therefore all classifiers were tuned using only **mRNA, miRNA and CpGs**. The DIABLO model with the full design using $\frac{3}{4}$ available datasets results in a BER of 22.9%.

We have merged the two sections into 3.3 and swap the order of the two paragraphs. The manuscript now reads (see our track changes in *Diablo_diff2.pdf*):

3.3 Competitive performance and identification of known and novel multi-omics biomarkers of breast cancer subtypes

On the TCGA breast cancer study we focused our analyses on characterizing and predicting PAM50 breast cancer subtypes. Processing and normalisation is described in Suppl. Section S2 and Fig. S11.

Classification performance benchmark. First, we compared the classification error rates of DIABLO models (DIABLO_null and DIABLO_full) with existing classification schemes (Concatenation and Ensemble) using sPLSDA and Elastic Net (enet) classifiers. For the purposes of this comparative performance analysis, the proteomics dataset which was only available for the training set was excluded to address the limitation of the Concatenation-based scheme. Hyperparameters for all six classifiers were tuned on the training set (mRNA, miRNA, CpGs) using 5-fold CV repeated 5 times and a variable selection size grid approach on three components. The performance of the methods was assessed on the independent test set (see Suppl. Section S5 for details). DIABLO_null and DIABLO_full led to a classification error rate of 19% and 21% respectively, while Concatenation and Ensemble-based methods error rate ranged from 11 to 28% (Suppl. Table S2, all methods included three components). We noted that Concatenation-based classifiers tended to be biased towards the more predictive variables (mRNA or CpGs), whereas DIABLO selected variables evenly across datasets and had similar error rates between training and test datasets.

Identification of multi-omics biomarkers. We then applied DIABLO_full for variable selection and evaluated its prediction performance on all omics available (mRNA, miRNA, CpGs and proteins). The optimal multi-omics biomarker panel size was identified as described above and detailed in Suppl. Fig. S12. Our panel consisted of 45 mRNA, 45 miRNAs, 25 CpGs and 55 proteins selected across three components with a balanced error rate (BER, see Rohart *et al.* 2017) of $17.9 \pm 1.9\%$. This panel identified many variables with previously known associations with breast cancer, according to MolSigDB (Liberzon *et al.*, 2015), miRCancer (Xie *et al.*, 2013), Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2005), and DriverDBv2 (Chung *et al.*, 2015). In addition, we identified several variables that were not found in any database and that may represent novel biomarkers of breast cancer (Suppl. Fig. S13). Fig. 3A shows that the majority of the test samples were located within the ellipses built on the training set, suggesting a reproducible multi-omics biomarker panel from the training to the test set (see Suppl. Fig. S14 for omic-specific component plots). On the independent test set, a BER of 22.9% indicated a relatively good prediction accuracy of breast cancer subtypes. The consensus plot corresponded strongly with the mRNA component plot, with a strong separation of the Basal (error rate = 4.9%) and Her2 (20%) subtypes, and a weak separation of Luminal A and Luminal B (error rates of 13.3% and 53.3% respectively) subtypes. A heatmap of the biomarker panel showed similar results (Suppl. Fig. S15). Overall, the features of the multi-omics biomarker panel formed a network of four densely connected clusters of variables (Fig. 3B). The largest cluster of 72 variables (20 mRNAs, 21 miRNAs, 15 CpGs and 16 proteins) was further investigated using gene set enrichment analysis as described in Section 3.2 and presented in Suppl. Fig. S16. We identified many cancer-associated pathways (*e.g.* FOXM1 pathway, p53 signaling pathway), DNA damage and repair pathways (*e.g.* E2F mediated regulation of DNA replication, G2M DNA damage checkpoint) and various cell-cycle pathways (*e.g.* G1S transition, mitotic G1/G1S phases). Therefore, DIABLO was able to identify a biologically plausible multi-omics biomarker panel that generalized to test samples. The panel also included unknown molecular features in breast cancer suggesting novel molecular features whose importance would require further experimental validations.

- Please reformat the supplement (there are still many formatting error in equations, different font sizes or line spacing, empty bullet points etc., which makes it slightly confusing to read; also, there seem to be some problems in the display of Figures S4 and S1 (in the uploaded pdf version) with ‘?’ appearing in the figures instead of spaces).

We have proof-read and carefully edited the supplement material, see our track changes on the document.

- missing word at the beginning of the Discussion: “with respect TO different phenotypes”

We have corrected this, thank you very much for a careful review.

Reviewer: 2

Comments to the Author

The authors have responded to each of my comments and addressed most of my concerns. There are a few remaining minor issues from my earlier comments as well as some typos/issues with the text that should be addressed:

1. How are the significant gene sets identified? DIABLO_full identifies many more significant gene sets than other methods, but it is unclear to me if these sets are identified based on a correlation cutoff or the edge-betweenness approach in the correlation network. How is the FDR computed?

A hypergeometric test was used to determine whether the list of features in each biomarker panel was enriched with gene-transcripts from various gene sets across 10 collections (see updated Suppl. Section S4 and Table S1). The false discovery rate was computed for each collection separately using the Benjamini Hochberg False Discovery Rate (Benjamini and Hochberg, 1995). The number of gene sets with an FDR less than 5% were determined and used as a metric to compare different multi-omics integrative methods. In the main manuscript, we have amended the associated paragraph in Section 3.2 (see also our track changes in `Diablo_diff2.pdf`):

Gene set enrichment analysis [based on hypergeometric tests](#) were conducted on each biomarker panel. [Briefly, we used gene symbols of mRNAs and CpGs of each biomarker panel and gene sets from 10 collections such as positional, curated, motif, computational, Gene Ontology, ontologic, immunologic, and hallmark gene sets as well as blood transcriptional modules and cell-specific gene sets](#) (Suppl. Section S4). DIABLO_full identified the greatest number of significant gene sets ([FDR=5%](#)) across the gene set collections that generally ranked higher than the other methods in colon (7 collections), gbm (5) and lung (5) cancer datasets (Table S1). JIVE outperformed all methods in the kidney cancer datasets (6 collections). In conclusion for this benchmark study, DIABLO_full aims at explaining the correlation structure between multiple omics layers and the phenotype of interest, leading to the greatest number of known biological gene sets such as pathway, functions and processes.

2. How many components are used in the simulated vs. real data studies? This should be made specific when describing the results. From the updated simulations, it sounds like only one component was being used initially?

We have checked and amended some of the main text to clearly include the number of components in each of the DIABLO analyses. Note that for DIABLO we refer to ‘component set’ as each dimension outputs a set of components associated to each dataset:

Section 3.1 (Simulation)

Three omic datasets consisting of 200 samples (100 in each of the two phenotypic groups) and 260 variables were simulated (details in Suppl. Section S1). Each dataset included four types of variables: 30 correlated-discriminatory (*corDis*), 30 uncorrelated-discriminatory (*unCorDis*), 100 correlated-nondiscriminatory (*corNonDis*) and 100 uncorrelated-nondiscriminatory (*unCorNonDis*) variables. DIABLO models with either a null or full design (DIABLO_null, DIABLO_full), [each with one component set](#), were compared with existing integrative classification schemes based on classification performance ... When we added [a second component set](#) in DIABLO, allowing for additional independent information to be included, the classification performance improved ...

Section 3.2 (Benchmark)

Each biomarker panel consisted of 180 features [across two components \(based on 90 variables\)](#) with the largest weights on [each of the two components](#)).

Section 3.3 (Breast Cancer)

Our panel consisted of 45 mRNA, 45 miRNAs, 25 CpGs and 55 proteins selected across three [component sets](#) with a balanced error rate ([BER, see Rohart et al. 2017b](#)) of $17.9 \pm 1.9\%$.

Section 3.4 (Asthma)

Both DIABLO approaches were applied to identify a multi-omics biomarker panel consisting of cells, gene and metabolite modules that discriminated pre- from post-AIC samples [on two component sets](#).

3. The authors have clarified the “multilevel” approach, but they do not make clear which matrix they are applying DIABLO to. Is it just X_w ?

In the asthma study, the multilevel approach was applied each of the datasets, namely the cell-type, gene and metabolite module datasets. We have amended the main text as follows:

We compared the standard DIABLO with a multilevel model (mDIABLO) that accounts for the repeated measures (pre/post) experimental design by isolating the within-sample variation from each [of the three datasets](#) (Fig. 4A, Suppl. Section S7, [Liquet et al. 2012](#)).

We also amended the Supplement Section S7 as:

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) ([Westerhuis et al., 2010](#); [Liquet et al., 2012](#)), 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 (X_w) was extracted and used to construct the multilevel DIABLO (mDIABLO) models. In the asthma study, the multilevel approach (called variance decomposition step) was applied to the cell-type, gene and metabolite module datasets.

4. What is the y-axis label in Figure 2A?

The y-axis indicates the number of overlapping features. We have amended Figure 2A to address this comment, thank you.

5. Typos: Incorrectly formatted citation of Lê Cao et al. (2008) (page 3)

This has been amended, thank you for a careful review!

References

- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, 289–300.
- Liquet, B. et al. (2012) A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC Bioinformatics*, **13**, 325.
- Westerhuis, J.A. et al. (2010) Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics*, **6**, 119–128.