THE UNIVERSITY OF
MELBOURNE

FACULTY OF
SCIENCE

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

April 23, 2018

Dear Professor Dina Schneidman,

Re: Pre-submission enquiry

We wish to enquire about the submission of our manuscript ‘DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays’ as a research method article in *PLOS Computational Biology*. This methodological paper is one of the method highlighted in our recent software article ‘mixOmics: An R package for ‘omics feature selection and multiple data integration’ that was viewed > 11,000 times since publication in your journal in November 2017.

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 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 highly relevant, correlated and discriminative biomarkers in six multi-omics studies including two case studies in human breast cancer and asthma, and in a comprehensive simulation study. We integrated a wide range of omics datasets, from transcriptomics (mRNA, miRNA), epigenomics (CpGs), proteomics and cell-type frequencies. We benchmarked DIABLO against integrative methods recently proposed in this emerging field, such as MOFA (https://doi.org/10.1101/217554, bioRxiv, April 2018).

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 *PLOS Computational Biology* who wish to capitalise on fastly generated multi-omics data and push novel biological discoveries to an unprecedented level.

Our R scripts are available in R markdown format as supplementary material, the method is implemented in the open source R package mixOmics, with detailed tutorials on our companion website http://www.mixOmics.org/mixDIABLO. I attach to this pre-sbumission enquiry our manuscript in its final stage, along with Figures and Supplemental material. I look forward to your reply.

Yours sincerely,

Dr. Kim-Anh LÊ CAO