January 10, 2020

Dear Dr. Doerr and Members of the Nature Methods Editorial Board,

Please find enclosed our manuscript entitled "Improving Deconvolution Methods in Biology through Open Innovation Competitions: an Application to the Connectivity Map" for your consideration for publication as a Brief Communication in Nature Methods.

This paper deals with a recurring problem in biomedical research: how to isolate signals of distinct populations (cell types, tissues, and genes) from composite measures obtained by a single analyte?

The paper adds to the existing literature by describing new deconvolution methods for analyzing gene-expression data. These methods were obtained through an open competition that used a novel dataset of 2,200 transcriptional profiles of 1000 genes (with 122 different perturbagens at several replicates) and attracted about 300 competitors from 20 countries.

Based on the competition outcomes, we show that (1) the top-nine performing methods, that ranged from machine learning approaches (Convolutional Neural Networks and Random Forests) to more traditional ones (Gaussian Mixtures and k-means), were faster and more accurate than the benchmark; (2) the winning solution, which was based on a popular machine learning method (a random forest algorithm), achieved the highest accuracy, the lowest inter-replicate variation, and, compared to the benchmark, was able to detect more than a thousand additional extremely modulated genes, while reducing the false positives at the same time; thus showing the great potential of random forest algorithms compared to the other methods.

We believe this work will bring much-needed insights into the field on how to use new machine learning methods, such as random forest algorithms, especially in settings where more traditional deconvolution approaches are the standard and can potentially represent "local-minima" solutions to the problem.

We also believe that researchers will find the generated dataset a powerful research tool for benchmarking their methods, as well as a useful resource for multiple applications.

Thank you for your time and consideration; we look forward to your comments and feedback.

Yours Sincerely,

Andrea Blasco and **Karim R. Lakhani**

Harvard Business School

Aravind Subramanian

Broad Institute