Yale University

MB&B
260/266 Whitney Avenue
PO Box 208114
New Haven, CT 06520-8114

Telephone:
203 432 6105
360 838 7861 (fax)
mark@gersteinlab.org
www.gersteinlab.org

June 3, 2018

Dear editor of *Journal Awesome*,

We would like to submit our manuscript titled "SigLASSO: a LASSO approach for identifying mutational signatures in cancer genomics" as a software article to your journal. SigLASSO is a computationally efficient framework we developed to determine active signatures in single cancer sample. Our approach jointly optimizes the mutation sampling process and uses L1 regularization to achieve sparsity. SigLASSO also allows biomedical researchers to incorporate their expert knowledge into the model.

In the recent years, research on recognizing mutational processes by their nucleotide signatures gained considerable attention. As a biomedical data science and informatics lab, many cancer researchers inquired us on how to identify mutational signatures in small, new sample cohorts. At the first glance, it looks like an easy linear system problem. However, due to the highly interdisciplinary nature of this problem, great care must be taken to derive meaningful solution that makes sense to cancer biologist. At this moment, the only off-shelf tool that we were aware of is deconstructSigs (Rosenthal et al., *Genome Biology*, 2016). This tool received remarkable attention and has been cited more than100 times. However, we, along with many labs, are not fully satisfied by its performance. What we are looking for is a tool that gives sparse, biologically interpretable solution.

Being motivated by the imperative need, we carefully designed a LASSO based framework: sigLASSO. 1) Our method jointly learns a mutation sampling process and becomes aware of the sampling variance. This helps to achieve better performance when the total mutation count is low, which is common in WES datasets. 2) SigLASSO uses mathematically well-characterized and justified LASSO to ensure sparse signature assignment, which is in line with our current knowledge of signatures. 3) sigLASSO is able to seamlessly make use of biological prior knowledge through fine-tuning the regularizer, enabling researchers to use their invaluable biology insights and clinical information to guide the signature assignment process. 4) sigLASSO is empirically parameterized by test performance to let data complexity inform model complexity. Our performance characterization shows sigLASSO is robust, efficient and produces more biologically reasonable solutions. We make sigLASSO available for the community as open software through GitHub and an R package. We believe this work is of great interest for the readers of *J. Awesome*.

This research is original, has not been submitted previously. None of the authors have any conflict of interest. We would suggest the following referees:

Alberto?
Zhaolei Zhang?
Ekta?
Maybe authors in PCAWG who use signature data but not affiliated with that group

We would like to exclude the following referees:

Ludmil B. Alexandrov (?)

Senior authors of PCAWG signature group?
Gady Gatz

Senior authors of deconstructSigs:
Charles Swanton
Barry S. Taylor
Javier Herrero

Yours sincerely,
Mark Gerstein
Albert L. Williams Professor
of Biomedical Informatics