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

August 31, 2017

Dear Dr. Rusk,

Please find our enclosed manuscript entitled “Supervised enhancer prediction with epigenetic pattern recognition and targeted validation across organism”, which we hope will be considered for publication in Nature Methods. I personally talked about this study with you while you were visiting Yale. In this method, we aggregated the epigenetic signals from large scale of validated enhancers, extracted patterns of features in a supervised fashion, and scan the whole genome with a matched filter. To the best of our knowledge, it is the first endeavor to apply signal processing methods to analyze epigenetic signals for enhancer prediction. The dominating peak-trough-peak pattern observed within the signal of certain post-translational histone modifications at active enhancers has a natural correspondence in biological mechanisms. The model also aggregates different epigenetic signals in a cell-type dependent fashion, which allows it to be applied to many different cell lines and species. Our method is validated through multiple rounds of *in vivo* and *in vitro* assays across species.

The new method we developed is trained with the output of massively parallel reporter assays. Traditionally, enhancers were characterized using low throughput validation assays, resulting in rigorous validation of very few cell-type specific mammalian enhancers. Those enhancers were also typically selected based on certain genomic characteristics, which introduces selection bias. Both the experimental size and selection bias hindered effective training and cross-validation of enhancer prediction models. The development of massively parallel reporter assays allows for the first time the identification of thousands of enhancers. Using data from these assays, we are able to rigorously train and test statistical models for enhancer prediction. The predictions are comprehensively validated in multiples ways including *in vivo* transgenic assays in six different mouse tissues, as well as *in vitro* transduction assays in human cell lines. The large number of validation assays we performed in mammalian tissues and cell lines also provide useful resource to the community.

We believe that our framework of genome-wide enhancer prediction will be useful to researchers. Its ability to predict tissue-specific enhancers across species allows it for broad application. The source code of the software is readily available in the github repository (https://github.com/gersteinlab/MatchedFilter), along with our genome-wide prediction of regulatory regions in different mouse tissues and human cell lines. This paper is submitted as part of the mouse developmental ENCODE package [ENC 017]. We hope you would reconsider it for publication in Nature Methods.

For your reference, we have listed a number of suitable reviewers for this work:

Alexander Stark
(alexander.Stark@imp.ac.at, Research Institute of Molecular Pathology, Austria)
Duncan Odom
(duncan.odom@cruk.cam.ac.uk, Cancer Research UK Cambridge Institute)
Piero Carninci
(carninci@riken.jp, RIKEN Center for Life Science Technologies)
Ian Dunham
(dunham@ebi.ac.uk, EMBL-EBI)
Nadav Ahituv
(nadav.ahituv@ucsf.edu, UCSF)
Katie Pollard
(kpollard@gladstone.ucsf.edu, UCSF)
Brendan Frey
(frey@psi.toronto.edu, University of Toronto)

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