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

May 17, 2016

Dear Editor of Nature Genetics,

Please find our enclosed manuscript entitled "Enhancer Prediction Using Pattern Recognition of Epigenetic Signals", which we hope will be considered for publication in your journal. Our study develops a new enhancers prediction method that 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 and these enhancers were typically selected based on certain genomic characteristics. As such, the tiny number and selection bias within these enhancers implied that they could not be used for training and cross-validating enhancer prediction models. Recently, a large number of massively parallel reporter assays were developed and these assays have identified thousands of putative enhancers. Using the data from these assays, we are able to rigorously train and test statistical models for enhancer prediction based upon these assays.

These assays have established that a peak-trough-peak pattern is observed within the signal of certain post-translational histone modifications at active enhancers. In this work, we create shape-matching filters to identify the occurrence of promoter and enhancer-associated patterns in different epigenetic signals. In addition, we use simple linear models to combine different epigenetic features and predict active enhancers and promoters in a cell-type dependent fashion. We also test these models using data from multiple assays and show that our models are transferable across cell-types as well as species without change. This allows us to apply our models to many different cell-lines and species. By applying it to the H1-human embryonic stem cell, a highly studied ENCODE cell-line, we were able to construct a secondary model that differentiates between enhancers and promoters based on the transcription factors that bind to these regions.

We have submitted this as an Analysis article. But if re-structuring of the manuscript seems necessary for consideration of review, we would be happy to revise it.

We list a number of suitable reviewers for this work.

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