Yale University

**Bass Building, Rm 432A**

**260/266 Whitney Avenue**

**PO Box 208114**

**New Haven, CT 06520-8114**

**Telephone: 203 432 6105**

**Fax: 360 838 7861**

[**Mark.Gerstein@yale.edu**](mailto:Mark.Gerstein@yale.edu)

April 7, 2014

Dear Editor,

Please find enclosed our manuscript entitled ‘**A flexible framework to annotate and prioritize noncoding somatic variants from cancer whole-genome sequencing**’ by Yao Fu, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y. Yip, Ekta Khurana and Mark Gerstein. We hope you will consider our manuscript for publication in *Genome Biology*.

Cancer is a genetic disease driven by aberrant mutations. Rapid and low-cost next generation sequencing technologies have opened new avenues in cancer research. Effort to catalogue genetic alterations may lead to new discoveries that will be translated to diagnostic and therapeutic targets. Cancer genome sequencing generally identifies thousands of somatic alterations in individual genomes. A few of them, called drivers, contribute to oncogenesis, whereas the majority are passenger mutations accumulated during cancer progression. Systematic studies of human cancer genomes have discovered a wide range of cancer driver genes. However, comparatively less effort has been invested in the noncoding portions of the genome. Recent discovery of somatic mutations in telomerase reverse transcriptase (*TERT*) promoter implicates that regulatory variants may constitute driver events. International cancer consortia, such as TCGA (The Cancer Genome Altas) and ICGA (The International Cancer Genome Consortium), plan to perform large-scale cancer whole genome sequencing in the near future. Thus, there is a great demand for high-throughput computational methods analyzing those variants.

In this study, we reported a flexible framework to annotate and prioritize ‘high-impact’ somatic alterations from cancer whole genome sequencing. We briefly describe the key elements here. To explore functional impact of noncoding variants, our method analyzed evolutionary and population-level conservations and nucleotide-resolution loss-of and gain-of function events. We then systematically identified proximal and distal associations between regulatory variants and target genes. We investigated the essentiality of associated genes, by incorporating network topology analysis, gene function prioritization, differential gene expression detection and sample-specific epigenetic (and/or open-chromatin) profiles. Integrating these features, we developed a weighted scoring scheme to predict potentially deleterious driver mutations. To illustrate performance, we applied our method to cancer somatic recurrent mutations and germline pathogenic variants. We used recurrence to approximate the deleteriousness of somatic variants from two perspectives - the same-site recurrence and the recurrent regulatory elements. Our method predicted higher scores for somatic recurrent and germline pathogenic variants, compared to non-recurrent and neutral variants respectively. As a test case, we then applied our method to an individual Medulloblastoma genome with the known *TERT* promoter mutation. Our method was able to prioritize this mutation and provided clues of its potential functional impact. We also compared our scoring scheme to the two recently published germline variants prediction methods: GWAVA and CADD. In addition, our method can detect recurrence across multiple genomes and we have created a recurrence database using 570 whole-genome sequenced samples of 10 cancer types and COSMIC (Catalogue of Somatic Mutations in Cancer) data. The usage of the database will further benefit the prioritization.

This study introduces an approach to understand the functional implications of regulatory mutations in cancer genomes. We believe the method will be useful for researches to prioritize a few mutations for further investigation and may help to understand the mechanisms underlying oncogenesis. Hence, we hope that you will consider our manuscript for publication in *Genome Biology*. Below we provide a list of suggested referees.

Yours sincerely,

Mark B Gerstein

Albert L Williams Professor of Biomedical Informatics,  
 Molecular Biophysics & Biochemistry,   
 and Computer Science

co-Director of   
 the Yale Program in Computational Biology & Bioinformatics

**Suggested Referees:**

Prof. Nicholas Luscombe

The CRUK London Research Institute and UCL, UK

Email: [nick@valleyofpigs.org](mailto:nick@valleyofpigs.org)

<http://www.luscombelab.org/Contact/index.php>

Prof. Kevin White

University of Chicago

Email: [kpwhite@uchicago.edu](mailto:kpwhite@uchicago.edu)

Prof. Andrey Rzhetsky

University of Chicago

Email: [arzhetsk@medicine.bsd.uchicago.edu](mailto:arzhetsk@medicine.bsd.uchicago.edu)

Prof. Mark Rubin

 Weill Cornell Medical College, New York

Email: [rubinma@med.cornell.edu](mailto:rubinma@med.cornell.edu)

Prof. Rachel Karchin

Johns Hopkins University

Email: [karchin@jhu.edu](mailto:karchin@jhu.edu)

Prof. Zhiping Weng

UMass Medical School, USA

Email: [Zhiping.Weng@umassmed.edu](mailto:Zhiping.Weng@umassmed.edu)

Prof. Robert Grossman

University of Chicago

Email: [robert.grossman@uchicago.edu](mailto:robert.grossman@uchicago.edu)