Dear Dr. [Gregory S. Barsh](http://www.plosgenetics.org/static/eic#eic) and Dr. [Greg Copenhaver](http://www.plosgenetics.org/static/eic),

We have uploaded our manuscript entitled, “A simple model-based approach to inferring and visualizing cancer mutation signatures” by Yuichi Shiraishi, et al., which we would like to submit for publication in PLoS Genetics.

It has been known that some cancer types and even individual instances within the same cancer type have characteristic pattern of somatic mutations or ``mutation signatures’’ that are often associated with some carcinogens (e.g., frequent C > T and CC > TT mutations in skin cancers caused by ultraviolet). Thanks to a massive amount of somatic mutation data produced by recent high-throughput sequencing technologies, we can now analyze mutation signatures at an unprecedented resolution, which many researchers expect leads to identification of novel mechanisms of cancer development.

To fully utilize the massive somatic mutation data for analyzing mutation signatures, we need efficient statistical approaches. Previous studies have successfully utilized the nonnegative matrix factorizations for extracting mutation signatures from large-scale somatic mutation data. However, the current approach has several limitations, one of which is that we cannot include many contextual factors because the parameter dimension increases exponentially as the number of factors to take into account.

In this paper, to address the above problem, we propose a novel probabilistic approach for detecting mutation signatures.

Our specific contributions are:

1. We provide a novel probabilistic model for mutation signatures that can robustly incorporate many contextual factors by decomposing the elements constituting mutation signatures (e.g., substitution types, flanking bases, transcription strand) into separate ``mutation features’’ and reducing the number of parameters.
2. Also, we provide an intuitively interpretable way of visualizing mutation signatures, which is analogous to “sequence logos” widely used for representing transcription factor binding sites.
3. By analyzing the large-scale somatic mutation data from 30 cancer types, we demonstrate that the proposed approach can not only gives us highly robust estimates, but also capture novel characteristics of mutation signature such as base sequence context effects at the two base 5' to the mutated sites.
4. We highlight the close relationship between the mutation signature model and the “mixed membership model”, which has been extensively studied in other field such as machine learning and population genetics. We believe noticing this relationship will be extremely helpful for future improvement of statistical methods for mutation signature problems.
5. We provide a highly efficient R package (<https://github.com/friend1ws/pmsignature>) as well as a web application (<https://friend1ws.shinyapps.io/pmsignature_shiny/>) implementing the proposed approach.

We believe that our work will be beneficial to a community of cancer genome researchers by providing; a novel insight for mutation signature problems, a new starting point for improving statistical approaches, and a highly efficient software that facilitates detection of novel mutation signatures. I would very much appreciate if you could consider our manuscript for publication in PLoS Genetics. I look forward to hearing from you soon.

Yours Sincerely,

Yuichi Shiraishi

Human Genome Center, Institute of Medical Science, The University of Tokyo

4-6-1, Minato-ku, Shirokanedai, Tokyo 108-8639, Japan

E-mail: [yshira@hgc.jp](mailto:yshira@hgc.jp)

Matthew Stephens

Department of Human Genetics, University of Chicago, Chicago Illinois, United States of America

Department of Statistics, University of Chicago, Chicago Illinois, United States of America

E-mail: mstephens@uchicago.edu