Dear Drs. Barsh and Copenhaver,

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

Our manuscript provides several important advances in the analysis of ``cancer mutation signatures”, which are characteristic patterns of somatic mutations that can vary both among and within tumor types. Such patterns are often associated with a specific mechanism - for example, C > T and CC > TT mutations in skin cancers are associated with damage due to ultraviolet light. The massive amounts of somatic mutation data produced by recent high-throughput sequencing technologies have the potential to identify mutation signatures at an unprecedented resolution, and many researchers expect this type of analysis to lead to identification of novel mechanisms of cancer development. However, to make the most of this opportunity requires effective statistical tools. Our manuscript addresses this important need.

Our contributions can be summarized by considering two (apparently unrelated) areas　where statistical methods have made a strong impact on genetic applications: the use of admixture analysis (also known as the “Structure” model) for identifying population structure, and the use of sequencing logos and position weight matrices (PWMs) to represent transcription factor binding sites. In essence, we introduce methods and software analogous to these methods — which have become established as methods of choice in their original domains of application— to the problem of identifying mutation signatures. To expand on these analogies: Whereas the admixture model assumes each sample has some proportion of their genome arising from each ancestral population, our model here assumes that each tumor sample has some proportion of their mutations arising from each signature. And whereas PWM-based methods represent each binding motif by a product of simple probability distributions, our approach represents each mutation signature by a similar product. One important consequence is that mutation signatures become much easier to visualize: see Figure 1 in the paper for example.

The end result is a set of new methods that are statistically more efficient, more flexible, and more interpretable than existing methods for this problem, which are based on``nonnegative matrix factorization”. One particular limitation of the current methods, addressed by our approach, is the ability to include many contextual factors into a signature. Existing methods suffer in this regard, because the number of parameters increases exponentially with the number of factors, but our method does not. For example, the current approach requires 1535 parameters per signature just to include ±2bp　flanking bases into the signature, whereas our new method requires just 17. For this reason we believe that our method not only makes several important immediate contributions, but has great potential for further impact in the future, as it will allow researchers to incorporate additional epigenetic and genetic features into mutation signature profiles, in a way that is currently unimaginable.

We believe that our work can become widely used for cancer mutation detection, just as Structure and PWMs have become widely used in the analysis of population structure and binding motifs. 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 hope that these tools will be beneficial to the community of cancer genome researchers, providing novel insight into mutation signature problems, and a new starting point for improved statistical approaches to this problem. We very much hope you will consider our manuscript for publication in PLoS Genetics.

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](mailto:mstephens@uchicago.edu)