***MBB***

Yale University

***266 Whitney Avenue***

***PO Box 208114***

***New Haven, CT 06520-8114***

***Telephone:***

***(203)432-6105  
(360)838-7861 (fax)***

***mark@gersteinlab.org***

***GersteinLab.org***

Today’s date

Dear Editor,

Please find enclosed our manuscript entitled “Sensitive information leakage from functional genomics data: Theoretical quantifications & practical file formats for privacy preservation”, which we hope will be considered for publication in *Nature Biotechnology*.

Our study relates to genomic privacy, which is becoming an increasingly important topic with the rise of personalized medicine and direct-to-consumer genetic testing. The advancement of technologies for high-throughput biomedical data acquisition at an ever increasing pace is bringing a surge of datasets. Consequently, one of the biggest limitations in biotechnology is how we will deal with this large-scale human data that contains private information.

In this study, we present a universal framework that can assess the private information leakage of raw functional genomics data. By instantiating simple linking attacks, we show that a small number of DNA/RNA sequencing reads can result in large privacy breaches that makes it possible to identify individuals. In light of our findings, we propose a powerful yet simple file format manipulation that allows sharing of raw functional genomics data while largely reducing sensitive information leakage. Our file format, called pBAM, is based on the widely used standard file format system SAM/BAM and is compatible with a range of software and pipelines. We tested this new file format in various ENCODE data processing pipelines and observed only a small amount of utility loss. This technology will soon have large-scale practical use as the ENCODE Data Coordination Center is currently implementing pBAMs into its pipelines. We are also working with GA4GH on standardizing the pBAM file format for worldwide use and sharing of epigenomics data. We intend to share standards and software support in a user-friendly online environment (privaseq3.gersteinlab.org).

We strongly believe this study will be of great interest to the *Nature Biotechnology* readership, because it provides a new biotechnological advancement that will impact a broad audience from the bioinformatics community to experimental biomedical data producers. We have submitted our study as a full-length article. However, if re-structuring of the manuscript is necessary to consider it for review, we would be happy to revise it.

We list a number of suitable reviewers for the paper.

Yours sincerely,

Mark Gerstein

Albert L. Williams Professor

of Biomedical Informatics