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.Gerstein@yale.edu
http://bioinfo.mbb.yale.edu

Today's date

Dear Editor of Nature Methods,

Please find enclosed our manuscript entitled "Information theory-based measures and privacy-preserving file formats for sensitive information leakage from raw functional genomics data", which we hope will be considered for publication in Nature Methods.

Our study relates to the balance between open data and protection of patients' sensitive information. We focused on the raw data from functional genomics experiments, which are becoming increasingly important with the rise of personalized medicine. We believe this study will be of great interest to Nature Methods readership, because

(1) Privacy, by itself, is one of the most critical topics of debate in data science that stands at the corner of many different fields, including ethics, sociology, law, political science, and forensic science.
(2) Functional genomics experiments, especially RNA-Seq, improved our understanding of important biological activities in disease and health, hence they sit at the heart of focus of many researchers and consortiums such as ENCODE, GTEx, and IHEC.
(3) A recent editorial by Nature Methods named "Sharing epigenomes globally" gave a great summary of the importance of raw epigenomic data and the hurdles for the full access to the data. This made us believe that our solutions to the raw epigenomic data sharing might be of great interest.
(4) Legislation like the Precision Medicine Initiative and NIH's new data sharing policies, which plan to increase the public sharing of research results and biomedical datasets, make it necessary to review data sharing and publishing policies and to find means to share data while protecting privacy.

In this study, we present a universal framework that can assess the private information leakage of raw functional genomics data. In light of our findings, we propose a powerful yet simple file format manipulation that allows sharing of raw functional genomics data by largely reducing the sensitive information leakage. Our file format called pBAM is based on the widely used standard file format system SAM / BAM and is compatible with many softwares and pipelines. We tested this new file format in various ENCODE data processing pipelines and observed only a small amount of utility loss. In fact, ENCODE Data Coordination Center is currently implementing pBAMs into their pipelines.

We believe our framework for quantification of sensitive data will help researchers to understand the privacy leaks in their data before release. We also strongly believe that our new file format system will be an important step towards open data in biomedical data science, and therefore will greatly increase reproducibility.

We were a bit unsure of the appropriate format of the manuscript for Nature Methods. We have submitted this as a full-length 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 the paper.

Yours sincerely,

Mark Gerstein
Albert L. Williams Professor
of Biomedical Informatics