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

December 18, 2017

Dear Editor of Nature Communications,

Please find enclosed the revision of our manuscript entitled "Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions", which we hope will be considered for publication in Nature Communications. Our manuscript relates to privacy of functional genomics data and how we can protect the data against breaches.

In our revisions, we observed that much clarification was needed regarding the setup of the problem and the scenario that we are studying. There are several critical points that we clarified: Firstly, there is great incentive in the research community to share functional genomics data without much consideration around the privacy risks. The functional genomics data, such as RNA-sequencing data are generated to reveal changes in gene expression activity and regulation under different conditions, treatments, and diseases. These data are extremely useful for discovering the epigenetic changes in different diseases and there is great interest within research community to share these openly. Unlike the variants, which are by nature sensitive information, the functional genomics data have more complicated aspect with respect to privacy: They are generated for the purpose of discovering the dynamic changes in gene activity. However, the functional genomics data also leak certain amount of variant information, which may create privacy concerns. Our manuscript analyzes these leakages.

Secondly, functional genomics data are served in different representations and they leak certain amount of sensitive information. For example the raw reads from sequencing experiments leak all the variant information and as such they are not shared publicly. To enable sharing these data without privacy concerns, the researchers use aggregated data types such as read depth signal profiles and gene expression quantifications. These data types do not explicitly contain variant information so they are presumed safe to share. However, previous studies have shown that the gene expression levels do leak significant amount of variant information through eQTLs. The read depth signal profiles are, however, not studied yet. Our manuscript tackles this issue and we show that the signal profiles do leak significant amount of information. As we discuss in the updated manuscript (As per Reviewer 2's comments), other higher order and complex interactions between gene expression levels and variants can leak variant information. Our manuscript does not tackle these.

To address one of the main concerns of the Reviewer 1 (Also raised by the Editor) regarding RNA-seq not getting affected by deletions, we added a figure, Supplementary Figure 3, which shows example of RNA-seq signal profiles for 6 individuals in the GTEx project. This figure shows the signal profile around a 2 base pair deletion (Ref SNP ID: rs34043625). The dip that is caused by the deletion can be seen in 3 of the individuals. We believe this figure should clear any doubts whether the RNA-seq signal profiles could be affected by small deletions or not. Very importantly, this figure was created from a screenshot of The UCSC Genome Browser's GTEx signal profile hub that is publicly available, which means anyone can access this data.

Another major criticism from Reviewer 2 was that our anonymization strategy does not close all types of leakages. We agree with the reviewer and we have clarified our presentation of the anonymization strategy. We list a list of possible sources of leakage that are not addressed in our manuscript and point out that our manuscript deals with the leakage from the signal profiles. However, we believe that the leakage from the signal profiles is one of the most urgent leakage source that must be closed since signal profiles are extensively and publicly shared. We believe that we have clarified both reviewers' concerns and misconceptions about our study. We believe their comments have made the paper much clearer and stronger.

We are submitting our paper as a companion to the other manuscripts written in the ENCODE Consortium's 3rd phase, i.e., ENCODE3. Our study is important to the ENCODE message because it is the only study that uses ENCODE data to analyze the sensitive information leakage from functional genomics datasets. We believe that our efforts provide resources and methods that are useful for privacy community.

An important manuscript from ENCODE3 paper rollout is the paper on Encyclopedia of elements. As we mentioned earlier, our manuscript does not directly relate to it. If needed, the encyclopedia manuscript can be obtained from Orli Bahcall.

We also include a number of suitable reviewers for our manuscript.

Yours sincerely,

Mark Gerstein
Albert L. Williams Professor
of Biomedical Informatics