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

17 August 2015

Dear Editor of Nature Methods,

We are submitting a revised version of our manuscript entitled "Analysis of Information Leakage in Phenotype and Genotype Datasets". We also include a revised set of figures, a supplementary material document, and a document with itemized responses to all of the reviewer comments. Briefly, in the revision we have added a number of new experiments that show the general applicability of the extremity attack under different scenarios. These results show how easily the extremity attack can be implemented. As per 1st and 3rd reviewer's suggestions, we performed analysis on whether the attacker can evaluate the reliability of the linkings that he/she makes. This enables the attacker to focus on the more reliable linkings, which can escalate the privacy risks and concerns.

In response to the specific concern of Referee 1 with regard to Im et al 2012 study, we have included, in the response letter and also in the Manuscript (Background Section) and in the Supplementary Material (Section 1) a detailed comparison of our manuscript with Im et al study. We also added new figures to clarify our attack scenario and compare it with the scenario in Im et al study. We are afraid that the Referee 1 is not knowledgeable about the conceptual setup and technical details of neither our study nor Im et al 2012 study: Im et al 2012 study focuses on "detection of a genome in a mixture" attacks, whereas we are studying the linking attacks. First and foremost, the machinery that is presented in Im et al study is not applicable and not suitable for an attacker to perform the attacks that we are studying in our manuscript. The reviewer, however, has the interesting view that Im et al is the definitive study that encompasses all the potential privacy breaches related to QTL studies, which we find rather incomplete. In response to the reviewer's comments, we explained how the two studies address significantly different scenarios in genomic privacy. We believe that as the number and the size of genotype and phenotype datasets grow, these datasets will get stolen and hacked. The detection of a genome in a mixture attacks will become needless as the participation of individuals in these datasets will be most certainly known. The attacks that are exemplified in our manuscript, however, will become much more prevalent because, as we show in our study, the individuals can be pinpointed by the linking attacks and their sensitive information be compromised. In other words, the attack scenario that we are presenting is almost orthogonal to that presented in Im et al study. We also listed a number of technical differences. These show the novelty of our study compared to Im et al study.

In addition, we made a comparison to the Schadt et al study and show that our approach can utilize much less information and obtain very high and comparable individual

characterization accuracy, which makes our attack scenario much more realistic and applicable. We made several schematic figures, added a supplementary materials document, and updates to the manuscript so as to clarify our contributions and different aspects of the linking attacks that we study in our manuscript.

As per reviewer's request, we also added a paragraph to the conclusions and added supplementary discussions for summarizing how an individual's privacy get compromised as the result of the linking attacks and discussed the possible risk management strategies, which should provide guidance in protecting phenotype datasets.

We also wanted to emphasize that our study contributes to a very important and topical subject. As we pointed in our conclusion, the dilemma that we cannot share data that is both perfectly useful and private is unavoidable. These are highlighted recently in *The Economist* and in a special issue of *Science Magazine*[1]. Genomic privacy will, with no doubt, become the center of these discussions. The objective measures and associated tools that we are presenting in our study will add to the currently limited arsenal of methods that will be used to evaluate privacy risks in biomedical data publishing.

We do realize that the manuscript is above the word limit that you indicated and we can most certainly reorganize the manuscript to have it fit to the word limits.

Yours sincerely,

Mark Gerstein
Albert L. Williams Professor
of Biomedical Informatics

---

[1] 1. We'll see you, anon | The Economist. at <http://www.economist.com/news/science-and-technology/21660966-can-big-databases-be-kept-both-anonymous-and-useful-well-see-you-anon>

2. Science/AAAS | Special Issue: The end of privacy. at <http://www.sciencemag.org/site/special/privacy/index.xhtml>