# Response to reviewer 3 for “Analysis of Information Leakage in Phenotype and Genotype Datasets”

# Response Letter

### -- Ref3: The authors make a distinction between linking and genome-in-a-mixture attacks. This is not a tight distinction –--

|  |  |
| --- | --- |
| ReviewerComment | 1. The authors make a distinction between linking and genome-in-a-mixture attacks. This is not a tight distinction in the sense, that identifying if a genome is in a mixture can lead to linking of genotype to phenotype. Consider the scenario where the genome-in-a-mixture is looking to see if a genome falls in cases vs controls (this linkage would not occur in the case of quantitative traits as in Im et al.). As the authors point out, I think the two use different types of information (large numbers of phenotypes vs large numbers of SNPs). |
| AuthorResponse | The reviewer brings out an interesting scenario that can be considered almost as a hybrid of genome-in-a-mixture attacks and linking attacks, which, as the reviewer suggests, is not the main focus of Im et al and our study. We also agree with the reviewer that different variations of genome-in-a-mixture attacks may lead to linking attacks. The studies designs based on case vs control comparisons, for example GWAS studies, might present new dimensions to consider in the analysis of sensitive information leakage. We also would like to emphasize that this scenario illustrates our point of the multifaceted nature of the genomic privacy and how slight modifications of the scenarios can lead to breaches. We, however, believe that this scenario does not correspond, as the reviewer points out, strictly to the cases that our manuscript is focused on. The main difference is following: In the linking attacks that we are studying, there are phenotypes (potentially not sensitive) that are used for linking, for example gene expression levels. As we explain in Results Section, the linking attack links the sensitive phenotypes (disease status), which are not used for linking, to the individual’s identity. ***Exact pinpointing of the individuals*** makes possible the linkage of sensitive phenotypes, which are not used in linking, to the individuals (This is also the privacy concern in Netflix Attack). In the scenario that reviewer raises, as there is no pinpointing of individuals, the attacker can only reveal the case versus control status of the individuals. Thus, no other information can be revealed. In addition, regarding the approaches that are proposed for detection-of-a-genome attacks; the statistical problem that is being addressed is detecting whether a genome is in a mixture, which, in our opinion, may not be directly applied to detecting whether an individual is in case versus control. We believe the latter problem requires a different statistical approach.We added discussion of the alternative route of privacy breach that the reviewer pointed out to the Supplementary Note. |
| Excerpt FromRevised Manuscript | ***Supplementary Note: “Comparison of “Detection of a Genome in a Mixture” and “Linking Attacks” in Genomic Privacy”***There are however, scenarios that can be thought of hybrid of detection of a genome and linking attacks. For example, in the case versus control studies, for example GWAS studies, if an attacker can predict whether an individual has participated in the case group or a control group, he or she can reveal the associated phenotype of an individual. |

### -- Ref3: You would alway choose the point that dominates the others to get a monotonic curve –

#### [[MAKE SURE WHAT I AM SAYING HERE IS MEANINGFUL. I AM NOT SURE IF I AM GETTING THE REFFREE CORRECTLY]]

|  |  |
| --- | --- |
| ReviewerComment | In figs 6b and 7b, the curves for the random experiments are non-monotonic but you would alway choose the point that dominates the others to get a monotonic curve (see Davis and Goadrich ICML 2006). |
| AuthorResponse | We agree with the point that reviewer is raising but we also believe that this result has not much practical importance for attacks:The sensitivity versus positive predictive value plots for random sortings of linkings in the Figure 6b and 7b show 10 random sortings of the dataset, so that we can compare how well sorting with respect to first distance gap statistic performs against random sortings. We reviewed the reference that is provided by the reviewer. Although the top performing random curves would generate a monotonic curve, this result does not have any practical use in an attack scenario because the attacker has no way of knowing which curve is going to perform best, unless he or she has access to real identities of the individuals, which is not possible by design of the attack. In other words, the attacker could generate each curve independently, however, he or she would have no way of choosing the “dominating one” among the random sortings unless he uses a measure like first distance gap. We added a note in the Supplementary Note to convey this result. |
| Excerpt FromRevised Manuscript | ***Supplementary Note: “Accuracy Measures for Linking”***It is worth noting that the sensitivity versus PPV curves with dominating PPV for the random sortings will generate a monotonically decreasing curve44. |