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 –--

Reviewer Comment	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).
Author Response	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. [[But this would still be different from our scenario because the attacker then only identifies whether the individual is in cases or in controls. The actual linkage in our scenario reveals the set of “carried” phenotypes with the “linking” phenotypes]] 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 the fact 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 added discussion of the alternative route of privacy breach that the reviewer pointed out to the discussion.
Excerpt From Revised Manuscript	

-- Ref3: The reviewer suspects that the authors are unaware that very similar work was published in 2012 --

Reviewer Comment	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).
Author Response	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 ones 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. In other words, the attacker could generate each curve independently, however, he/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 discussion of this point in the Supplementary Material Section XX to convey this interesting result.
Excerpt From Revised Manuscript	