Analysis of Individual Characterizing Information Leakage in Gene Expression and Genotype Datasets

Arif Harmanci, Jieming Chen, Dov Greenbaum, Mark Gerstein

# ABSTRACT

Genomic privacy is gaining much attention with the unprecedented increase in the breadth and depth of personal “-omic” datasets. While most of the studies are focused on protection of genomic variants in personal genomes, the analysis of sensitive information leakage in molecular phenotypic datasets, like functional genomics datasets, is in its inception. A significant amount of leakage can be caused by the phenotype –to-genotype correlations identified by the genome-wide studies, where associations between large number of genetic loci and different molecular phenotypes are discovered. Although these correlations are valuable for understanding how phenotype and genotype interact, an adversary can utilize the phenotype-to-genotype correlations for predicting the variant genotypes for the individuals, for which only phenotype information is available. When performed over a large number of predicted genotypes the adversary can accurately link the entries in phenotype datasets to the entries in a genotype dataset to reveal sensitive information.

In this paper, we study the characterizability of individuals in the context of linking attacks, where an adversary aims at revealing an individual’s sensitive information by linking the entries in phenotype and genotype datasets. While doing this the attacker utilizes a third dataset that contains the phenotype-to-genotype correlations. We focus on the correlations between genotypes and gene expression levels reported in eQTL datasets. We first quantitatively assess the relation between the amount of information leakage the adversary can cause by genotype prediction and how accurately the leakage can be performed. We propose two quantification metrics that can be used for evaluating the amount of leakage at different levels of correct predictability. We then present a general framework for analysis of the individual characterization and evaluate the fraction of characterizable individuals in a general setting on the representative dataset. For illustrating the practicality of these analyses, we present a simple practical genotype prediction method, which, when employed on a representative dataset, yields a significant fraction of individuals characterizable. Overall, the quantification metrics and the analysis framework can be utilized in analysis of individual characterizability in other studies where phenotype-to-genotype correlations are investigated.