Analysis and Protection of Privacy-Sensitive Information in Gene Expression Datasets

Arif Harmanci, Jieming Chen, Dov Greenbaum, Mark Gerstein

ABSTRACT

With the unprecedented increase in the size of genomic datasets, the quantification and protection of privacy-sensitive information is a vital issue to be addressed for protection of anonymity of the participants of the scientific studies.

[[Define sensitive information]]

[[Privacy can also be defined as a certain type of information that individuals do not want to leak: For example GWAS, HGMD SNPs]]

In this paper, we present an framework for analysis and protection of private information in the gene expression datasets. We present a general scenario where the gene expression datasets can be exploited to predict eQTL genotypes to link independently distributed anonymized datasets by an adversary to de-identify individuals. We first analyze the amount of leakage of genetic information for each eQTL SNP when predicted using the gene expression datasets. We propose a simple method to predict eQTL genotypes from gene expression datasets. We then utilize the prediction method for low frequency multiple SNP genotype prediction, which can be used to de-identify individuals. Using publicly available gene expression dataset we illustrate that a significant fraction of the samples are vulnerable to de-identification. As a remedy for the privacy loss, we focus on anonymization of the gene expression datasets and present a method for anonymizing the gene expression dataset. We illustrate that the datasets can be anonymized with very small amount of loss in the biological information.

1 BACKGROUND

[[Introduction goes here]]

2 RESULTS

2.1 Overview of the Privacy Breaching Scenario

Figure 1 illustrates the privacy breaching scenario that is considered. The breach occurs by linking two datasets such that one of the datasets contains the individual identities and corresponding genotypes and the second dataset contains the gene expression levels and sensitive information (e.g. disease

status) about each individual. The second dataset is assumed to be anonymized by removal of the individual identities to protect the individuals. The adversary gains access to both datasets and links the datasets to associate the sensitive information to individuals. While performing the linking "attack" the adversary utilizes publicly available databases. In the considered scenario, the eQTL databases are utilized which enable linking the expression levels to the genotypes.

[[We also consider the inclusion of extra information, in this case, gender and race]]

[[Our novelty lies in showing the linking attack in the context of expression-genotype association with a simple attack]]

2.2 Quantification of Loss of Privacy: Identifiability of Individual SNPs

We first analyze in a general setting the amount of identifiable genetic information using the linking of gene expressions with genotypes. For this, we utilize the mutual information based metric that has been applied for quantifying privacy loss. Figure 2a and 2b show the distribution of privacy loss for all the eQTLs given the gene expression levels. It can be seen that, given the expression levels, there is significant loss of genetic information compared to a random background in the genotype information, which is as high as 20% for some of the eQTLs.

This analysis does not give us a way to predict the genotype information from the expression dataset. We propose using a method that we termed as extremity attack to generate a genotype given the gene expression value in a gene expression dataset.

[[How do we predict the genotypes? Extremity attack?]]

2.3 Identifiability of Individuals by Low Frequency Multiple SNP Genotypes in k-Anonymization Framework

The individual SNP identification is useful for a conceptual analysis of the leakage of genetic information. The eQTLs, however, are not suitable for identifying individuals since they are most common variants. We therefore utilize multiple SNP genotypes that have low frequency in the database. This is the basis of utilizing the common variants in the linking attack.

To formalize the analysis using the low frequency multi-SNP genotypes, we utilize the k-anonymization framework. K-anonymization formalizes a way to identify the number of vulnerable individuals and also to ensure the anonymization, which is presented in Section 2.5. Briefly, in order to identify the individuals that are vulnerable to the linking attack, we identify the individuals that have the multi-SNP genotypes, which are highly predictable using the expression dataset.

We first analyze the information content and information leakage in each eQTL. Figure 2c shows the

2.4 Linking Attack by Prediction of Discriminating SNPs

[[eQTL based prediction of the SNPs]]

[[Extremity of the gene expression levels: Extremity as a statistic]]

2.5 Anonymization of the Dataset

[[Do anonymization for all possible parametrizations to decrease the privacy loss to minimum]]

[[k-anonymization formality for guaranteeing anonymity]]

3 METHODS

3.1 Quantification of Genotype Information Content and Loss of Privacy

[[MI and entropy based definition of IC and Loss of Privacy]]

3.2 Extremity Attack

[[Define the extremity attack: Correlation and extremity parameters]]

3.3 K-Anonymization

[[Define k-anonymization]]

[[Present in detail the anonymization procedure that we propose]]

4 CONCLUSION AND DISCUSSION

We pretty much hacked the geuvadis dataset and will send this to E. Snowden.