Analysis and Protection of 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.

[[Previous approaches: Differential privacy, different types of attacks, inversion attack, linking attack]]

[[It has been shown previously that differential privacy formality, which is theoretically the most complete data protection scheme, for releasing genomic information may lead to very poor utility~\cite{XX,XX}. It is therefore necessary to analyze where the sensitive information exists in different datasets and how protection of the sensitive data affects data utility. To accomplish this, This study furthers the understanding of the predictable sensitive genetic information from gene expression datasets.]]

In this paper, we present a comprehensive framework for analysis of sensitive 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 re-identify individuals.

[[1st act::First we present quantification of the predictability of eQTL genotypes from gene expression and the leakage of individual identifying information in the predicted genotypes. Using the quantifications, we study the tradeoff between the predictability of the genotypes and the amount of identifying information leakage.]]

[[2nd act::We next present a 3-step general framework for individual identification. This analysis can bring important insight into the extent of vulnerability of individuals and what can be predicted, which is important for designing differentially private release algorithms for analysis of gene expression datasets. In addition, the framework that we are presenting can be utilized for the analysis of vulnerability in the future eQTL studies.]]

[[We also analyze the cross-population analysis of leakage]] ?

[[We need to say somewhere that we are concentrating on linking type of attacks]]

[[We also do a cross-dataset analysis of the reproducibility of genetic leakage attack using different datasets.]]

[[3rd act::We finally propose a practical linking attack method. (Extremity attack) An example for the practicality of all the analysis.]]

[[This is also the first time a systematic analysis of genetic leakage is analyzed with respect to prediction from gene expression.]]

1 BACKGROUND

[[Define sensitive information: Anything that the individuals do not want leaked]]

[[Previous work: Homer, Schadt, Erlich, …]]

[[Genetic leakage protection: Several of these: De-identification based (removal of names), Encryption based, more complicated de-identification techniques (k-anonymization), differential privacy based (makes a very high compromise of utility for privacy's sake). Last two are active field of research.]]

[[In this paper, we analyze identifiability of SNP genotypes and identifiability of individuals in the context of linking attacks. These are the most prevalent attacks that can affect the currently generated genomics datasets.]]

[[First, we present an analysis framework that formalizes and decomposes the analysis of genetic leakage in the context of linking attacks. Our framework decomposes the linking attack into 3 steps that we study in detail.

-- We make the assumption that the attacker recovers the conditional probabilities perfectly, which enables us to be as stringent about what the attacker can predict as possible.

-- We evaluate the incorporation of auxiliary information.

-- Simulate suboptimal conditional probabilities: Relax the assumption that the attacker can build the posterior probabilities of genotypes given expressions. How much vulnerability exist?

This framework can be used for leakage analysis in the future studies.

We finally present a practical attack for prediction of genotypes from gene expression levels.]]

[[The paper is organized as follows: We first analyze the predictability of the SNPs and evaluate the tradeoff between the amount of identifying information recovered versus the predictability of the eQTLs using expression datasets. Next we present the 3 step individual identification framework and study different aspects of vulnerability using the framework. In the last section, we present a novel and simple but effective genotype prediction method, which can be employed in most scenarios, and use it in our framework.]]

2 RESULTS

2.1 Overview of the Privacy Breaching Scenario by Linking Attacks

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.

COULD DO IN GETEX

[[Present GEUVADIS dataset]] MEAN?

2.

[[We first present the notations]] The gene expression and genotype datasets are stored in $N_q \times N_i^e$ and $N_q \times N_i^v$ matrices e and v, respectively, where N_i^e and N_i^v denotes the number of individuals in gene expression and genotype datasets, respectively; N_q denote the number of variants and genes, respectively, in the eQTL dataset. k^{th} row of e, e_k, contains the expression values for k^{th} gene and $e_{k,j}$ represents the expression of the k^{th} gene for j^{th} individual. Similarly, l^{th} row of v, g_l, contains the genotypes for l^{th} variant and $v_{l,j}$ represents the genotype ($v_{l,j}$ ϵ {0,1,2}) of l^{th} variant for j^{th} individual. We will denote the random variables (RVs) whose values represent that the gene expression of k^{th} gene and the variant genotypes for l^{th} variant with $\{E_k\}$ and $\{V_l\}$, respectively. The rows of the expression and genotype dataset matrices are matched to each other such that the gene and genotype RV pairs $\{(E_k, V_k)\}, k < N_q$, are highly correlated. We will denote the correlation with $\rho(E_k, V_k)$. In many of the eQTL studies, this correlation is reported with the statistical significance and several other information (for example, population of individuals for which the correlation is observed) in a table. The sign of $\rho(E_k, V_k)$ represents the direction of association, i.e., which genotype corresponds to higher expression and the magnitude represents the strength of the association.

[[Nature of eQTL gene expression correlations: Extremity based associations (extremities in both the genotypes and in the gene expression levels associate with each other) are identified in eQTL studies. This is the main point of leakage of genetic information from gene expression datasets, which are identified generally via a linear model.]]

[[For generalization of the analysis, we assume that the attacker can predict with high certainty the posterior probabilities. Previous studies have presented different approaches for predicting a-posteriori probabilities of genotypes given gene expression levels.]]

2.2 Quantification of Tradeoff between Predictability of the SNP Genotypes and Individual Identification

[[Predictability of the eQTL genotypes, individual identification information. This is the analysis where the attacker is to match with no database at hand by just predicting all the SNPs he chooses to predict.]]

In the linking attack, the attacker aims to identify the correct individual among N_i individuals. In order to identify an individual, if the attacker can predict the genotypes for a set of eQTLs such that the joint probability of the genotypes is $1/N_i$, the attacker can distinguish the individual from other individuals with certainty. In other words, if the attacker can gain access to $\log_2(N_i)$ bits of information using the genotypes predicted from expression data, the individual is vulnerable. To quantify the identifying information, we introduce individual identifying information (III):

$$III(\{V_1 = g_1, V_2 = g_2, \ldots, V_N = g_N\}) = -\sum_{k=1}^{N_q} \log\big(p(V_k = g_k)\big).$$

where V_k is the k$^{\text{th}}$ eQTL and g_i is a specific genotype for the eQTL (Refer to Methods Section 3.1 for more details) and III denotes the individual identifying information. Practically, the individual identifying information can be interpreted as a quantification of how rare the genotypes are. If the list contains many rare genotypes, it contains significant amount of identifying information. The attacker aims to predict as many eQTLs as possible such that *III* is at least $\log(N_i)$. of pred

In order to maximize the amount of III, the attacker will aim at predicting as many eQTL genotypes as possible. The predictability of the eQTLs from gene expression, however, is not uniform as some of them are more highly correlated with the gene expression levels compared to others, given in $|\rho(E_k, V_k)|$. Thus, the attacker will try to select the most predictable eQTLs genotypes that are most correctly predictable so as to maximize the amount of leaked identifying information. To quantify predictability of eQTL genotypes from expression levels, we use exponential of the conditional distribution of genotype given gene expression level as a measure of predictability. Given the gene expression levels for j^{th} individual, the predictability of all the eQTL genotypes is computed as

$$\pi\big(\{V_k\}, \{E_k = e_{k,j}\}\big) = \exp\left(-\sum_{k=1}^{N_q} H(V_k | E_k = e_{k,j})\right)$$

MAKE INTUIT

where π denotes the predictability of V_k given the gene expression level $e_{k,j}$. Given a list of eQTLs, the joint entropy is used, which (Refer to Methods Section 3.1 for more details). This measure can be interpreted as the average probability of correct prediction of the genotypes given the expression levels.

We assume that the attacker will sort the eQTLs in terms of their predictability. For this, we assume the attacker uses the absolute value of the correlation between the genotype and the expression, i.e., $|\rho(E_k, V_k)|$. In order to evaluate the tradeoff between the identifying information of the top predictable eQTLs and their predictabilities, we plotted average *III* versus average π in Fig 2. We first sorted the eQTLs with respect to the reported $\rho(E_k, V_k)$ then for the top 20 eQTLs, we estimated mean \pi and mean *III* for all the samples. Figure 2a shows that there is significant leakage of *III* at 20% average predictability, there is approximately 7 bits of leakage and at 5% predictability, there is around 11 bits of leakage, which is enough to identify, on average, all the individuals in the dataset. (At 12.4% predictability, the leakage is approximately 9 bits for 6 top eQTLs.) Figure 2b and 2c also shows the average leakage for the randomized eQTL dataset where the genes and eQTLs are shuffled to generate a background model. The leakage is significantly smaller compared to the original eQTL dataset (At an average predictability of 12.4%, the average leakage is approximately 3.5 bits.)

[illegible] SORT tradeoff

CORR 2.

2.3 A Generalized Individual Identification Model

[[We decompose the linking attack into 3-steps to study different variations and parameterizations of the linking attack.]]

Following the results in the previous section, we present a 3 step model for individual identification. Figure 3a summarizes the steps in the individual identification. In the first step, the attacker selects the eQTLs that will be used in the linking attack. The selection of eQTLs can be based on different criteria. As described in the previous section, the most accessible criterion is filtering the eQTLs based on absolute value of the reported correlation coefficient with a predefined threshold. Another criterion is to use the estimated conditional entropy of the genotype given the gene expression level, which is a measure of the predictability of the eQTL genotype (See Fig 3b). The second step is the prediction of the selected eQTLs. The attacker uses a predefined prediction model. In this step we are assuming that the attacker can reliably predict the posterior probabilities of the genotypes given the gene expression levels as illustrated in Fig 3b. The attacker uses the posterior probabilities of the genotypes to predict the maximum *a-posteriori* (MAP) genotype. In this prediction, the attacker assigns the genotype that has the highest *a-posteriori* probability (Refer to Methods Section 4.3). The third and final step of individual identification is comparison of the predicted genotypes to the genotypes database to identify the individual that matches the predicted genotypes. We assume that the attacker links the predicted genotypes to the individual in the genotype dataset with the smallest number of mismatches compared to the predicted genotypes.

2.

OF THE INDIV

2.3.1 Individual and Population Identification Accuracy

[[We assume that the attacker selects the eQTLs using 2 different criteria: (1) Absolute value of the gradient of correlation reported in the eQTL resource, (2) Estimated predictability of the genotype: Entropy of the conditional distribution of genotypes for each individual]]

We assume that the attacker uses the absolute value of the reported correlation between the variant genotypes and gene expression levels to select the eQTLs. Fig SXX shows the distribution of the absolute correlation levels for the eQTL dataset. The genotypes for the selected eQTLs are predicted using MAP prediction (Refer to Methods Section 4.3). Figure 4a shows the the number of selected eQTLs and the fraction correctly predicted MAP genotypes with changing absolute correlation thresholds.

[[Fraction of vulnerable individuals]]

Using the predicted eQTL genotype selected at each absolute correlation cutoff, the attacker performs the 3rd step in the attack and links the predicted genotypes to the genotype dataset to identify individuals (Refer to Methods Section 4.4). Figure 5a shows the fraction of vulnerable individuals. The fraction of vulnerable individuals increase as the absolute correlation threshold increases and fraction is maximized at around 0.35. At this value, 95% of the individuals are vulnerable. This illustrates that the power of vulnerability is maximized at absolute correlation threshold of 0.35. This can be explained by the increase in identifying information leakage as the accuracy of the predicted genotypes increase while there is a balancing decrease in the identifying information leakage with decreasing number of eQTL genotypes predicted.

[[Auxiliary Information: Gender and/or Population]]

We also evaluate the case when the attacker gains access to auxiliary information. As the sources of auxiliary information, we use the gender and population information that is available for all the participants of 1000 Genomes Project on the project web site. We assume that the attacker either gains access to or predicts the gender and/or the population of the individuals and uses the information in the 3rd step of the attack (Refer to Methods Section 4.4). Figure 5a shows the fraction of vulnerable when the auxiliary information is available. When the auxiliary information is available, more than 95% of the individuals are vulnerable to identification for all the eQTL selections up to when the absolute correlation threshold is 0.6.

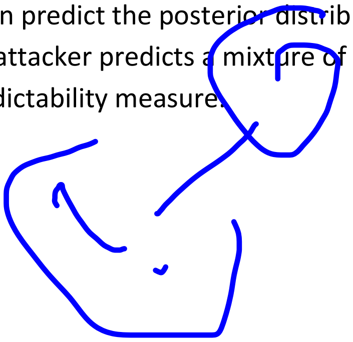
[[Population stratification of vulnerable individuals to populations: It is important to predict the probabilities from the correct population for diverse populations.]]

~~[[Population confusion matrix and accuracy statistics; YRIs get confused with other populations; this is most probably caused by the fact that at the expression levels of YRI's, the MAP estimates push them to the wrong genotype. This illustrates that matching of the target population and the training population is vital in individual identification.]]~~

[[To follow this up, we generate population specific posteriors learned on different populations. How does this change population prediction accuracy?]]

2.3.2 Effect of Posterior Probability Distributions based Identification Accuracy

In this section, we relax the assumption that the attacker can predict the posterior distribution of eQTL genotype given the expression values. We assume that the attacker predicts a mixture of the posterior and prior distributions of eQTL genotypes based on the predictability measure.



[[The bottomline is that when the posterior is not well built, the vulnerability decreases significantly, but there is still substantial vulnerability.]]

2.3.3 [[OPTIONAL]] Cross Study Individual Identification Accuracy

[[Stranger et al eQTLs on GEUVADIS expression data results: ~35% with Stranger et al eQTLs]]

2.4 Genotype Prediction by Extremity Attack

better

This analysis is useful for getting quantification of leaked genetic information from gene expression datasets. To predict the eQTL genotypes from gene expression levels, we propose using a method that we name "extremity attack". In this attack, given one gene whose expression level correlates with a variant. The prediction utilizes a statistic we termed $extremity$ of gene expression level which quantifies how extreme an individual's gene expression level is away from the mean of the distribution. Given the gene expression level, e, for a $extremity$ is defined as following:

$$extremity(e) = \frac{rank\ of\ e}{N} - 0.5.$$

Extremity is bounded between -0.5 and 0.5. Figure 8a illustrates the extremity attack. The attacker utilizes the extremity and the gradient of association between the gene expression level to assign a genotype to the associated variant.

[[Figure 3bc shows the accuracy of extremity attack with different extremity and correlation thresholds.]]

2.5 Individual Identification with Extremity Attack

[[Fig. 5a; Distribution of the maximum of absolute extremity over all the samples. How well does expression extremity identify individuals? It is mostly uniform except for some samples.]]

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 low frequency multiple SNP genotypes such that all the SNP genotypes are highly predictable using the expression dataset.

[[External information: 1 bits of gender information can be easily predicted from ; how does this change vulnerability; this justifies the fact that we need "buffering" in anonymization to protect against unaccounted external information that may cause increased vulnerability.]]

2.6 Anonymization

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

[[k-anonymization formality for guaranteeing anonymity]]

3 CONCLUSION AND DISCUSSION

In this paper we present a simple framework for quantification of the sensitive information leakage in the linking attack scenarios. The premise of sharing genomic information is that there is always an amount of leakage in the sensitive information. We believe that this quantification methodology can be utilized for more extensive analysis of the leakage in sensitive information for high level correlations in the genomic datasets. The quantification can be further developed for guaranteeing bounds on anonymized datasets.

[[How does this framework compare to other formalities? For example differential privacy? Differential privacy is about release mechanisms in statistical databases. Our analysis is about release of datasets. It is similar but differential privacy does not enable quantification of the leakage.]]

[[As the eQTL studies are done on larger and larger datasets, new (probably population specific) eQTLs are going to be identified which will increase leaking identifying information.]]

We also presented a simple attack that is based on using extremity statistic to predict genotypes that can implicate the sensitive information. Compared to previous approaches, this statistic is very easy to compute.

4 METHODS

4.1 Quantification of Individual Identifying Information and Predictability

To quantify the individual identifying information, we use surprisal, measured in terms of self-information of the genotypes:

ref

$$III\left(V_{(l_i)}=g\right)=I\left(V_{(l_i)}=g\right)=-\log(p\left(V_{(l_i)}=g\right))$$

where $V_{(l_i)}$ is an eQTL variant and g $(g\epsilon\{0,1,2\})$ is a specific genotype for G, $p(G=g)$ is the probability (frequency) of the genotype in the sample set and III denotes the individual identifying information. Assessing this relation, the genotypes that have low frequencies have high identifying information, as expected. Given multiple eQTL genotypes, assuming that they are independent, the total individual identifying information is simply summation of those:

$$III\left(\{V_{(l_1)} = g_1, V_{(l_2)} = g_2, \ldots, V_{(l_N)} = g_N\}\right) = -\sum_{i=1}^{N} \log\left(p\left(V_{(l_i)} = g_i\right)\right).$$

The individual identifying information after the gene expression levels are revealed is basically the conditional III given the gene expression levels:

$$III_{remaining}(\{G_1 = g_1, G_2 = g_2, \ldots\} \mid \{E_1 = e_1, E_2 = e_2, \ldots\}) = -\sum_{i=1}^{N} \log(p(G_i = g_i \mid E_i = e_i))$$

where E_i represents the gene expression level for the i^th gene, which is associated with the genotype of G_i. The leakage in III is the remaining III after expression levels are revealed:

$$III_{leaked} = III - III_{remaining}.$$

[[Predictability: Exponential of the conditional distribution given the gene expression levels]]

We measure the predictability of eQTL genotypes using an entropy based measure. Given the eQTL, $V_{(l_i)}$, and the correlated gene expression $E_{(k_i)}$

$$\pi\left(V_{(l_i)} \middle| E_{(k_i)} = e\right) = \exp(-H(V_{(l_i)}|E_{(k_i)} = e)$$

DISC

where π denotes the predictability of $V_{(l_i)}$ given the gene expression level e, and H denotes the entropy of $V_{(l_i)}$ given gene expression level e for $E_{(k_i)}$. The extension to multiple eQTLs is straightforward. For the j^th individual, given the expression levels $e_{k,j}$ for all the eQTLs, the total predictability is computed as

$$\pi\left(\{V_{(l_i)}\}, \{E_{(k_i)} = e_{k_i,j}\}\right) = \exp\left(H\left(-\{V_{(l_i)}\} \mid \{E_{(k_i)} = e_{k_i,j}\}\right)\right)$$

$$= \exp\left(-\sum_{i} H(V_{(l_i)}|E_{(k_i)} = e\right)$$

[[Cite and show that this measure is in [1/3,1] for one genotype. The interpretation of this measure is that the prediction process is converted to random guessing with uniform probability distribution where average correct prediction probability is \pi.]]

In addition, this measure is guaranteed to be between 0 and 1 such that 0 represents no predictability and 1 representing perfect predictability. The measure can be thought as mapping the prediction process to a uniform random guessing where the average correct prediction probability is measured by π.

4.2 Estimation of Genotype Entropy for Quantification of Predictability

[[How did we estimate the genotype entropy and conditional specific entropies?]]

[[We bin the expression values to log_2(N_i) different bins \cite{…}]]

4.3 MAP (Maximum *a-posteriori*) Genotype Prediction

[[Describe the binning and MAP selection of genotypes]]

4.4 Linking of the Predicted Genotypes to Genotype Dataset

Given set of predicted genotypes for individual j, $\{v'_{l,j}\}$;

$$pred_j = \underset{a}{\operatorname{argmax}}\{\sum_{b} I(v'_{b,j}, v_{b,a})\}$$

If $pred_j =$ j; j is vulnerable

[[Formulate when the auxiliary information is available?]]

4.5 Population Specific Posterior Distributions

[[Describe the binning and population based selection of genotypes.]]

4.6 Suboptimal Posterior Generation

[[Describe a-priori mixed posterior generation]]

4.7 Extremity Attack

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

4.8 K-Anonymization

[[Define k-anonymization]]

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

5 Datasets

[[GEUVADIS dataset, and eQTLs, 1000 genomes dataset]]

[[Other eQTL datasets?]]