Analysis, Quantification, and Protection of Sensitive Information Leakage 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 sensitive personal information is a vital issue to be addressed for protection of privacy. In this paper, we present a comprehensive framework for quantification and analysis of sensitive information in the gene expression datasets. We present a general scenario where an adversary can utilize gene expression datasets in conjunction with expression quantitative loci, eQTL, datasets to correctly predict the genotypes of the eQTL variants to link two datasets and re-identify individuals. In the context of this re-identification scenario, we first propose measures for studying the tradeoff between quantification of the leakage of individual identifying information and predictability of the eQTL variant genotypes. Next we present a general framework that consists of 3 steps for individual identification and utilize it on a representative dataset to show that significant fraction of individuals become vulnerable for identification. Finally, we present a simple genotype prediction method and utilize it in our framework to show in a simple practical setting that a significant fraction of the samples can be re-identified.

1 BACKGROUND

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

The decreasing cost of DNA sequencing [1] has rendered a massive increase in the annual amount of high-dimensional personalized "–omic" data being generated [2]. Many large consortia, like GTex [3], ENCODE [4], 1000 Genomes [5], and TCGA [6], are generating large amount of high dimensional –omics datasets. Coupled with the generated data, the sophisticated analysis methods are being developed to discovery complex biological correlations between the molecular signatures and phenotypes, which can contain sensitive information about individuals like disease status. It is therefore necessary for the models for sharing these datasets to keep up with the analysis methods so as to control the leakage of the predictable sensitive information in each study.

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

Several previous studies have demonstrated the possibility of individual identification in different specific scenarios by exploiting different statistical and genomic attributes of the generated datasets. A review of breaches of genomic privacy can be found [7]. In [8] authors propose a novel statistical analysis methodology for testing whether an individual is in a pool of samples, where only the allele frequencies are known. In [9], the authors identify the identities of several male participants of 1000

Genomes [5] project by exploiting that the short tandem repeats on Y-chromosome can be used as an individual identifying biomarker. In [10], the authors demonstrate that one can build a model for predicting genotypes for eQTLs using gene expression levels and use the model to identify individuals with high accuracy.

[[Previous approaches: Differential privacy, different types of attacks, model inversion attack, linking attack]] [[Genetic leakage protection: Several of these: De-identification based (removal of names), Encryption based, more complicated de-identification techniques (k-anonymization), different types of attacks, model inversion attack, linking attack]] [[Genetic leakage protection: Several of these: De-identification based (removal of names), Encryption based, more complicated de-identification techniques (k-anonymization), different types of attacks makes a very good compromise between data utility and privacy's sake). These are active fields of research.]] [[It has been shown previously that differential privacy framework is currently 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 leaks in different datasets and how much of the sensitive data affects data utility. To accomplish this, This study furthers the understanding of the presence of sensitive genetic information in gene expression datasets.]]

In addition, different formalities have been proposed for protecting sensitive information. For example differential privacy [11] establishes bounds on the leakage on sensitive information in statistical databases. The main issue with this formality is that there is a stringent tradeoff between utility and privacy. Thus, it has been shown that differential privacy mechanisms can decrease the utility of the biological information [12]. In addition, homomorphic encryption [13], which enable performing operations on encrypted data directly, are possible approaches that can offer protection of sensitive information as the real data is never seen by the information processors. These approaches require from very high computational complexity and storage requirements for encrypted data. Another well-established formality is k-anonymization [14]. In this formality, the released dataset is anonymized by different data perturbation techniques for ensuring that no combination of features in the dataset can be shared by less than k individuals. This approach, however, is computationally very high complexity with large features and usually not practical for high dimensional biomedical datasets. Several variants have been proposed that extend k-anonymity framework [15, 16]. As the size and nature of the biomedical datasets change, it is necessary to build analysis frameworks that can quantify the correlations between different data types that can lead to sensitive information leakage.

In this paper, we are proposing metrics and an analysis framework for quantification of sensitive information in gene expression datasets that can be used for re-identification of individuals. The expression datasets generated by RNA-sequencing [17] can be utilized directly for identification of personal variants directly from the reads. This can, however, be easily remedied by removing the nucleotide information in the datasets [18] or by releasing the gene expression levels in the publicly accessible datasets. Another information that can be utilized for genotype identification from gene expression datasets is the expression quantitative trait loci (eQTLs) datasets. Each eQTL contains a common genetic variant and a gene expression such that the genotype of the variant is significantly correlated with the expression level of the gene. Each eQTL entry contains typically the strength of the correlation and a gradient information that tells which genotype is associated with higher or lower expression level. The eQTLs are especially useful since there are large eQTL datasets that are publicly available online. [[cite GTex project hosts approximately 30 million eQTLs whose gradient and significance information can be viewed freely through eQTL Browser [[\cite GTex Browser]].

[[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.]]

We concentrate on the linking attack scenario. In this scenario, the attacker gains access to an expression dataset where the expression levels of participants are stored with sensitive information. The attacker also gains access to a genotype dataset where the genotypes of a set of individuals are stored with their identities. The aim of the attacker is to match the individuals in a gene expression dataset to

individuals in a genotype dataset where each match enables the attacker to link the identity of an individual in the genotype dataset to the sensitive information in expression dataset.

[[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 evaluate the incorporation of auxiliary information. 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 1a illustrates the privacy breaching scenario that is considered. In the context of breach, there are two datasets. First dataset contains gene expression levels and certain sensitive information (e.g., disease status) for n_e individuals. The gene expression dataset is de-identified by removal of the names. This dataset is release for public access. The second dataset contains the genotypes and the identities for n_v individuals. We assume that this dataset is released with restricted access. It should be noted that the number of individuals in genotype dataset is assumed to be larger than the number of individuals in expression dataset. The adversary gains access to both datasets and intends to identify the identities of each of the n_e individuals in the gene expression dataset. For this, attacker predicts the genotypes of the variants for each individual in gene expression dataset and links the individuals in the expression dataset to the individuals in the genotype dataset. The linking process is basically comparison of the predicted genotypes for each individual and identifying the best matching individual. In the genotype prediction, the attacker concentrates on expression quantitative trait loci (eQTL) in the attack. The attacker aims at exploiting the correlation between the eQTL variant genotypes and eQTL gene expression levels for predicting eQTL variant genotypes with high accuracy.

[[We first present the notations]]

Figure 1b illustrates the eQTL, expression, and genotype datasets. The eQTL dataset is composed of a list of gene-variant pairs such that the gene expression levels and variant genotypes are significantly correlated. We will denote the number of eQTL entries with n_q. The eQTL (gene) expression levels and eQTL (variant) genotypes are stored in $n_q \times n_e$ and $n_q \times n_v$ matrices e and v, respectively, where n_e and n_v denotes the number of individuals in gene expression dataset and individuals in genotype dataset. k^{th} row of e, e_k, contains the gene expression values for k^{th} eQTL entry and $e_{k,j}$ represents the expression of the k^{th} gene for j^{th} individual. Similarly, k row of v, v_k, contains the genotypes for k^{th} eQTL variant and $v_{k,j}$ represents the genotype ($v_{k,j} \in \{0,1,2\}$) of k variant for j^{th} individual. We assume that the variant genotypes and gene expression levels for the k^{th} eQTL entry are distributed randomly over the samples in accordance with random variables (RV) which we denote with E_k and V_k, respectively. As explained earlier, these random variables are correlated with each other. We denote

the correlation with $\rho(E_k, V_k)$. In most of the eQTL studies, the value of the correlation is reported in the eQTL dataset. The absolute value of $\rho(E_k, V_k)$ indicates the strength of association between the eQTL genotype and the eQTL expression level. 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. This forms the basis for correct predictability of the eQTL genotypes using eQTL expression expression levels: The homozygous genotypes associate with the extremes of the gene expression levels, i.e., the highest of the lowest levels of expression and the heterozygous genotypes associate with moderate levels of expression. Most of the eQTL studies utilize complicated linear models to identify this association between the genotypes and the gene expression levels.

[[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.]]

For generalization of the analysis, we assume that the attacker can utilize a prediction model that can estimate the *a posteriori* distribution of the eQTL genotypes given the eQTL expression levels, i.e., $p(V_k|E_k)$. This allows us to quantify the individual identifying information and also analyze the fraction of individuals that are vulnerable to linking attack in different settings, without making any assumptions on the prediction model that is utilized by the attacker.

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 context of the linking attack introduced in Section 2.1, the attacker aims to correctly identify n_e individuals in the expression dataset among n_v individuals in the genotype dataset. In order to identify an individual, the attacker should select a set of eQTLs that he believes he can predict correctly. Next, given the individual's expression levels, the attacker should predict the genotypes for the selected eQTLs correctly such that the predicted set of genotypes are not shared by more than 1 individual, i.e., the predicted genotypes identify the individual. In other words, the frequency of the set of predicted genotypes for the selected eQTLs should be at most $\frac{1}{n_v}$. We can rephrase this condition as following in information theoretic terms: If the attacker can reliably predict $\log_2(n_v)$ bits of information using the genotypes predicted from expression data for an individual, the individual is vulnerable. It should be noted that, assuming the independence of the genotypes for different eQTLs, we can decompose the quantity of individual identifying information that is leaked for a set of n correctly predicted eQTL genotypes:

$$III(\{V_1 = g_1, V_2 = g_2, \dots, V_n = g_n\}) = -\sum_{k=1}^{n} \log\big(p(V_k = g_k)\big)$$

where V_k is the k^{th} eQTL and g_k is a specific genotype for the eQTL (Refer to Methods Section 3.1 for more details), $p(V_k = g_k)$ denote the genotype frequency of g_k within the population, and *III* denotes the total individual identifying information. Evaluating the above formula, *III* increases as the frequency of the variant's genotype g_k decreases. In other words, the more rare genotypes contribute higher to *III* compared to the more common ones. Thus, individual identifying information can be interpreted as a

quantification of how rare the predicted genotypes are. The attacker aims to predict as many eQTLs as possible such that *III* for the predicted genotypes is at least $\log(n_v)$.

In order to maximize the amount of *III*, the attacker will aim at predicting as many eQTL genotypes correctly as possible. The (correct) predictability of the eQTL genotypes from expression levels, however, varies over the eQTL dataset as some of the eQTL genotypes are more highly correlated with the expression levels compared to others, given in $|\rho(E_k, V_k)|$. Thus, the attacker will try to select the eQTLs whose genotypes are the most correctly predictable to maximize *III* leakage. Although $\rho(E_k, V_k)$ is a measure of predictability, it is computed differently in different studies. In addition, there is no easy way to combine these correlation values when we would like to estimate the predictability of multiple eQTL genotypes. In order to uniformly quantify predictability of the eQTL genotypes from expression levels, we use an information theoretic measure. We use the exponential of the entropy of the conditional distribution of genotype given gene expression level as a measure of predictability. Given the expression levels for j^{th} individual, we compute the predictability of the k^{th} eQTL genotypes as

$$\pi(V_k|E_k = e_{k,j}) = \underbrace{\exp(-1 \times \overbrace{H(V_k|E_k = e_{k,j})}^{\text{Randomness left in } V_k \text{ given } E_k = e_{k,j}})}_{\text{Convert the entropy to average probability}}$$

where π denotes the predictability of V_k given the gene expression level $e_{k,j}$. π can be interpreted as the average probability that the attacker can correctly predict the eQTL genotype given the expression level. In the equation for π, the conditional entropy of the genotypes given the gene expression level is a measure for the randomness that is left in genotype distribution when the expression level is known. In the case of high predictability, the conditional entropy is close to 0, and there is little randomness left in the genotype distribution. Taking the exponential of negative of the entropy converts the entropy to average probability of correct prediction of the genotype. In the most predictable case (conditional entropy close to 0), π is close to 1, indicating very high predictability. In order to extend the predictability measure to multiple of eQTLs, we use exponential of the negative of joint conditional entropy. (Refer to Methods Section 4.1 for more details).

[[Say sth about the allele frequency, predictability, and information content: The eQTLs are common variants, thus are not very informative]]

At this point, it is useful to note that there is a natural tradeoff between the correct predictability of eQTLs and the leaking individual identifying information. For example, the eQTLs that have the highest individual identifying information, i.e., high $-\log\big(p(V_k = g_k)\big)$, must have small genotype frequency in the population. The low frequency genotypes, however, are most likely not highly correlated with the gene expression levels, i.e., π is smaller for those variants.

As discussed earlier, the attacker will aim at predicting the largest number of eQTL genotypes given the expression levels. For this, we assume the attacker will sort the eQTLs with respect to absolute correlation then predict the eQTL genotypes starting from the first eQTL. Within this scenario, 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. For this, we first sorted the eQTLs with

respect to the reported $|\rho(E_k, V_k)|$. Then for top *n=1,2,3,…,20* eQTLs, we estimated mean π and mean *III* over all the samples. We then plotted mean π versus mean *III* for each *n* which is shown in Fig 2a. 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. These results show the extent of leakage of identifying information from the gene expression datasets.

2.3 A Generalized Framework for Analysis of Individual Identification

[[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 framework for individual identification. Figure 3a summarizes the steps in the individual identification for each individual. The input is the gene expression levels for j$^{\text{th}}$ individual in the expression dataset, e_j. The aim of the attacker is correctly identifying each of the n_e individuals in the expression dataset in genotype dataset. In the first step, the attacker selects the eQTLs (among n_q eQTLs) that will be used in linking j$^{\text{th}}$ individual. The selection of eQTLs can be based on different criteria. As described in the previous section, the most accessible criterion is selecting the eQTLs for which absolute value of the reported correlation coefficient, $|\rho(E_k, V_k)|$, is greater than a predefined threshold. In our analysis, we evaluate the effect of changing correlation coefficient. 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. The second step is genotype prediction for the selected eQTLs using a prediction model. For general applicability of our analysis we are assuming that the attacker's prediction model can reliably construct the posterior probability distribution of the genotypes given the gene expression levels. The attacker then uses the posterior probabilities of the genotypes to identify the maximum *a posteriori* (MAP) genotype for each eQTL. In this prediction, the attacker assigns the genotype that has the highest *a posteriori* probability given the expression level (Refer to Methods Section 4.3). The third and final step of individual identification is comparison of the predicted genotypes to the genotypes of the n_v individuals in genotype dataset to identify the individual that matches best to the predicted genotypes. In this step, the attacker links the predicted genotypes to the individual in the genotype dataset with the smallest number of mismatches compared to the predicted genotypes (Refer to Methods Section 4.4).

2.3.1 Individual 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 number of selected eQTLs and the fraction correctly predicted MAP genotypes with changing absolute correlation thresholds.

[[Fraction of vulnerable individuals]]

Using the list of predicted eQTL genotypes 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 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. [[This illustrates the tradeoff having more correct eQTLs versus the accuracy of predictions]]

[[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. These results show that a significant fraction of individuals are vulnerable for most of the correlation thresholds that the attacker can utilize.

2.4 Anonymization

[[How many eQTL associations should be removed to make vulnerability small?]]

An important aspect of analysis of privacy is anonymization. Here we assess how much the gene expression dataset should be anonymized for ensuring that there are no vulnerable individuals. We assume that the anonymization of a gene expression level is performed by censoring, i.e., replacing the reported value for gene expression level with 'Not Available' value in the dataset. After an eQTL gene expression level is anonymized, we assume that the attacker cannot reliably estimate the eQTL genotype, which decreases the chance that the individual is vulnerable. Given a vulnerable individual, it is useful to estimate the minimum number of genes expression levels to be anonymized to make the individual non-vulnerable. For this, we compute the genotype distance for all the individuals then sort the distances in increasing order and store it in a list. The number of expression levels to be anonymized is then given by following:

$$\text{Number expression levels to be anonymized} = \text{Genotype distance of the } 2^{\text{nd}} \text{ closest Individual} - \text{Genotype distance of } 1^{\text{st}} \text{ closest Individual}$$

[[How do we anonymize gene expression levels: Just remove the expression level]]

Figure XX shows the average number of expression values to be anonymized per individual with different absolute correlation thresholds. It can be seen that the total number of expression levels to be anonymized is maximum for $|\rho(E_k, V_k)|$, i.e., when all the eQTLs are utilized for individual identification.

[[Auxiliary information?]]

We also evaluated the number of gene expression levels to be anonymized when auxiliary information is available. [[TBA]]

[[When those eQTLs are removed, how are the correlations affected?]]

In order to evaluate how the biological utility of the eQTLs are affected after the expression levels are anonymized, we computed the spearman correlations of the eQTL genotypes and the associated gene expression levels. Fig XX shows the distribution of the absolute correlations between genotypes and gene expression levels before and after the gene expression dataset is anonymized. It can be seen that there is a significant decrease in the correlation levels of a significant number of the eQTLs.

[[This is probably just an underestimate of how much we need to remove before anonymization]]

2.5 Individual Identification with Extremity Attack

In previous sections, we presented quantification of leakage in individual identifying information and a general framework for analysis of vulnerability and studied the number of expression levels that should be anonymized to decrease vulnerability. In this section, we propose a simple genotype prediction methodology, extremity attack, and demonstrate the extent vulnerability when the attack is utilized in the individual identification framework.

Extremity attack utilizes a statistic we termed $extremity$, which quantifies how extreme an individual's gene expression level is among the expression levels of all the samples. For the gene expression level, e_k, $extremity$ is defined as:

$$extremity(e_k) = \frac{\text{rank of } e_k \text{in } \{e_1, e_2, \dots, e_{n_e}\}}{n_e} - 0.5$$

where e_k is the expression level of k^{th} individual. Extremity is bounded between -0.5 and 0.5. Figure SXX shows the mean absolute extremity distribution of all the gene expression levels for all the individuals. The average absolute extremity per individual is around 0.25.

Figure XX illustrates the extremity attack. Extremity attack utilizes the fact that the more extreme gene expression levels most likely coincide with one of the extreme genotypes, i.e., homozygous genotypes (Refer to Methods Section 4.7). For example, if the gradient of association between eQTL genotype and expression levels is positive, the individuals that have high positive extremity are most likely to have genotype value of 2 and the individuals with high negative extremity are most likely to have eQTL genotype value of 0 and vice versa when the gradient is negative. One aspect of the extremity attack is that it predicts only homozygous (i.e., most extreme) genotypes. Figure XX shows the accuracy of genotypes predictions with extremity attack. As expected the accuracy of genotype predictions increase with increasing correlation threshold.

We next used the extremity based prediction in the individual identification framework (Fig 2). Fig XX shows the fraction of vulnerable individuals. We utilized the correlation based eQTL selection in step 1,

then extremity based genotype prediction in step 2. In step 3 the individual is assigned as the individual whose genotype matches closest to the predicted genotypes. More than 95% of the individuals are vulnerable for most of the parameter selections. In addition, when the gender and/or population information is present as auxiliary information (red and green colored plots), the fraction of vulnerable individuals increases to 100% for most of the eQTL selections. These results suggest that extremity based linking attack, although technically simple, can be utilized to generate a significant amount of vulnerability.

3 CONCLUSION AND DISCUSSION

In this paper we present a framework for quantification and analysis of sensitive individual identifying information leakage from the gene expression datasets. 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 in the genomic datasets.

The predictability vs III-leakage tradeoff analysis that we performed can be generalized in two ways in future studies: First, the information theoretic measures that we proposed for measuring predictability versus the *III* leakage can be utilized for analyzing the tradeoff in other biomedical datasets where correlations can be exploited in linking attacks. Second, the analysis that we performed can be used to extrapolate the number of vulnerable individuals in a large dataset at different predictability levels. For example, in Figure XX, at 5% predictability level there is 11 bits of *III* leakage, which can identify on average 2000 individuals. At 1% predictability, there is around 18 bits of *III*, which can identify on average approximately 64000 individuals. Depending on the privacy leakage that can be tolerated, the predictability versus *III* leakage can be utilized to propose new metrics for quantifying the risk of individual identification.

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

[[There is also utility satisfying differential privacy: Our study is useful for understanding which utility to hide and which to reveal.]]

Compared to other formalities, our study aims more to characterize the leakage of individual identifying information. Differential privacy formality, for example, aims at proposing release mechanisms for statistical databases where the mechanism guarantees that queries return results such that the probability of identifying a specific individual's contribution to the result is vanishingly small. In order to maximize the utility of the biological data, it is, however, necessary to analyze the points of sensitive information leakage so that one can design the utility maximizing release mechanisms [19]. Our study contributes to quantifying the individual identifying information 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.]] [[The leakage of individual identification from gene expression datasets is rather complicated to analyze. The quantification method that we presented here is an underestimation of the leakage since it utilizes perfect matching of the predicted genotypes to the individual genotypes. With increasing efforts to identify the correlation of the genetic variation to quantifiable phenotypes.]] [[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.]]

We also introduced a simple yet effective approach for identification of individuals. The approach utilizes extremity based genotype prediction. When employed in the individual identification framework, this simple approach renders a very significant number of individuals vulnerable. This illustrates the amount the viability of individual identification from gene expression datasets.

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:

$$III(V_k = g_{k,j}) = I(V_k = g_{k,j}) = -\log(p(V_k = g_{k,j}))$$

where V_k is an eQTL genotype RV 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:

$$III(\{V_1 = g_{1,j}, V_2 = g_{2,j}, \dots, V_N = g_{N,j}\}) = -\sum_{k=1}^{N} \log\left(p(V_k = g_{k,j})\right).$$

[[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 genotype RV, V_k, and the correlated gene expression RV, E_k,

$$\pi(V_k|E_k = e) = \exp(-H(V_k|E_k = e))$$

where π denotes the predictability of $V_{(l_i)}$ given the gene expression level e, and H denotes the entropy of V_k given gene expression level e for E_k. 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

$$\pi(\{V_k\}, \{E_k = e_{k,j}\}) = \exp(H(-\{V_k\} \mid \{E_k = e_{k,j}\}))$$

$$= \exp\left(-\sum_k H(V_k|E_k = e_{k,j})\right)$$

[[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. This is the reciprocal of Shannon diversity; the average number of genotype predictions that you can randomly equally likely choose from.]]

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]]

[[Must include SNP selection such that some of the genotypes are not assigned any genotype bc of the selection]]

4.4 Linking of the Predicted Genotypes to Genotype Dataset

Given a set of predicted eQTL genotypes for individual j, $\tilde{v}_{\cdot,j} = \{\tilde{v}_{l,j}\}$, the attacker links the predicted genotypes to the individual whose genotypes have the smallest distance to the predicted genotypes:

$$pred_j = \underset{a}{\operatorname{argmin}}\{d(\tilde{v}_{\cdot,j}, v_{\cdot,a})\}\,.$$

$pred_j$ denotes the index for the linked individual and $d(\tilde{v}_{\cdot,j}, v_{\cdot,a})$ represents the distance between the predicted eQTL genotypes and the genotypes of the a^th individual:

$$d(\tilde{v}_{\cdot,j}, v_{\cdot,a}) = \sum_{k=1}^{n_q} (1 - I(\tilde{v}_{k,j}, v_{k,j}))$$

where $I(\tilde{v}_{k,j}, v_{k,j})$ is the match indicator:

$$I(\tilde{v}_{k,j}, v_{k,j}) = \begin{cases} 1 \text{ if } \tilde{v}_{k,j} = v_{k,j} \\ 0 \text{ otherwise} \end{cases}$$

Finally, j^{th} individual is vulnerable if $pred_j = j$. When auxiliary information is available, the attacker constrains the set of individuals while computing $d(\tilde{v}_{\cdot,j}, v_{\cdot,a})$ to the individuals with matching auxiliary information. For example, if the gender of the individual is known, the attacker excludes the individuals whose gender does not match while computing $d(\tilde{v}_{\cdot,j}, v_{\cdot,a})$. This way the auxiliary information decreases the search space of the attacker.

4.5 Extremity Attack

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

4.6 Anonymization

[[How many gene expression values should be anonymized on average so that closest match to the predicted genotypes is not the correct individual.]]

Given that j^{th} individual is vulnerable; we would like to estimate (Results Section 2.4) the number of genes expression levels to be anonymized to make the individual non-vulnerable. For this, we compute the distances $d(\tilde{v}_{\cdot,j}, v_{-,a})$ for all the individuals then sort the distances in increasing order and store it in a list. Let $d_{(k)}(\tilde{v}_{\cdot,j})$ denote the number of mismatching genotypes for the k^th individual in the sorted list. The number of expression levels to be anonymized is then given by following:

$$\#\text{ genes to anonymize} = d_{(1)}(\tilde{v}_{\cdot,j}) - d_{(2)}(\tilde{v}_{\cdot,j})$$

but which one?

[[How do we anonymize gene expression levels: Just remove the expression level]]

5 DATASETS

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

6 REFERENCES

1. Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB: **The real cost of sequencing: higher than you think!** *Genome Biology* 2011:125.

2. Rodriguez LL, Brooks LD, Greenberg JH, Green ED: **The Complexities of Genomic Identifi ability**. *Science (80-)* 2013, **339**(January):275–276.

3. Consortium TG: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45**:580–5.

4. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.

5. The 1000 Genomes Project Consortium: **An integrated map of genetic variation**. *Nature* 2012, **135**:0–9.

6. Collins FS: **The Cancer Genome Atlas (TCGA)**. *Online* 2007:1–17.

7. Erlich Y, Narayanan A: **Routes for breaching and protecting genetic privacy.** *Nat Rev Genet* 2014, **15**:409–21.

8. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson J V., Stephan DA, Nelson SF, Craig DW: **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays**. *PLoS Genet* 2008, **4**.

9. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: **Identifying personal genomes by surname inference.** *Science* 2013, **339**:321–4.

10. Schadt EE, Woo S, Hao K: **Bayesian method to predict individual SNP genotypes from gene expression data**. *Nature Genetics* 2012:603–608.

11. Dwork C: **Differential privacy**. *Int Colloq Autom Lang Program* 2006, **4052**:1–12.

12. Fredrikson M, Lantz E, Jha S, Lin S: **Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing**. In *23rd USENIX Security Symposium*; 2014.

13. Gentry C: **A FULLY HOMOMORPHIC ENCRYPTION SCHEME**. *PhD Thesis* 2009:1–209.

14. SWEENEY L: **k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY**. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002:557–570.

15. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M: **L -diversity**. *ACM Trans Knowl Discov Data* 2007, **1**:3–es.

16. Ninghui L, Tiancheng L, Venkatasubramanian S: **t-Closeness: Privacy beyond k-anonymity and ℓ-diversity**. In *Proceedings - International Conference on Data Engineering*; 2007:106–115.

17. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57–63.

18. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M: **RSEQtools: A modular framework to analyze RNA-Seq data using compact, anonymized data summaries**. *Bioinformatics* 2011, **27**:281–283.

19. Alvim MS, Andrés ME, Chatzikokolakis K, Degano P, Palamidessi C: **Differential privacy: On the trade-off between utility and information leakage**. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. *Volume 7140 LNCS*; 2012:39–54.