

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

Arif Harmanci, Jieming Chen, Dov Greenbaum, Mark Gerstein

ABSTRACT

[[MG-abstract-below]]

Genomic privacy is currently a matter receiving a lot of attention with the reality that thousands if not millions of genomes will be sequenced in the near future. Most of the attention has been focused on genomic variants. However some appreciate that this is a substantial privacy risk in the large amount of quantified molecular phenotyping data being released. This data ostensibly does not have any variant information. However because of its very high dimensional nature and because a small amount of genotype-to-phenotype correlation exists it can potentially be linked to genotypes in a statistical way. Here we develop a formalism for characterizing the amount of potential information leakage in a potential attack and showing the tradeoff between the predictability of the genotype, genotype linkage and the amount of potential information leaked. Then we show how one could practically instantiate an attack focusing on the most commonly used data sets now that of RNA-seq and eQTL. We develop a three step procedure showing how an attacker would select eQTL, statistically predict a few SNPs and then identify individuals based on that.

This attack becomes particularly easy to implement when one deals with outlier gene expression levels and we developed a particular realization of this attack for these outlier cases quantifying the amount of information leakage.

[[Previous-abstract-below]]

The unprecedented increase in the breadth and depth of “-omic” datasets enforces the data sharing mechanisms to adapt to the risks associated with leakage of sensitive personal medical information. ~~As~~ The genome-wide studies on association between the genetic variants and the molecular phenotypic profiling data have identified correlations between large number of genetic loci and different phenotypes. Although these correlations are valuable for biological understanding increase about how of how phenotype and genotype interacts, they can serve to an adversary as a backdoor for predicting the genotypes from phenotypes or vice versa. When the prediction is done over very large number of genetic loci, this allows the adversary to link the entries in genotype and phenotype datasets so as to reveal sensitive phenotypes like disease status relate phenotypic information about individuals. Even though majority the genomic privacy studies has focused solely on protection of genetic variants, it is

~~necessary to the molecular signatures in the omics datasets, the risk that an individual analyze how these correlations can be characterized in terms of their sensitive information also increases. As individuals take part in different omic profiling studies, the probability that they can be characterized increases substantially. lead to a linking attack with other datasets and lead to privacy breach.~~

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 matching, or linking, the entries in ~~two phenotype and genotype~~ datasets. ~~While doing this the attacker utilizes a third dataset that contains the genotype-to-phenotype correlations.~~ We focus on ~~a scenario where one dataset contains variant the correlations between~~ genotypes and ~~identities for a set of individuals and the other dataset contains the gene expression levels and disease states for another set of individuals, whose identities are removed from the dataset. The attacker aims at exploiting the correlations between variant genotypes and expression levels, which are available reported in public eQTL datasets, to predict the genotypes of the variants using the expression levels in second dataset. Attacker then matches the predicted genotypes to the first dataset so as to link the disease states to the individual identities to characterize the disease state of the individuals in first dataset.~~

~~In this scenario, we. We~~ first perform a quantitative analysis ~~of the tradeoff between the characterizing information leakage and the correct predictability of the information genotypes.~~ We propose two quantification metrics that can be used for evaluating the amount of leakage at different levels of prediction. We then present a generalized framework for analysis of the individual characterization and evaluate the fraction of characterizable individuals in a general setting on the representative dataset. For a illustrating the practicality of these ~~results analyses,~~ we ~~finally~~ 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 ~~for to~~ analysis of ~~characterizing information leakage individual characterizability in large scale gene expression profiling datasets other genotype-to-phenotype correlation studies.~~

1 BACKGROUND

The decreasing cost of DNA sequencing [1] has rendered a massive increase in the amount of high-dimensional personalized ~~“omic” biomedical~~ data being generated [2]. Many large consortia, like GTEx [3], ENCODE [4], 1000 Genomes [5], and TCGA [6], are generating large amount of ~~omics personalized biomedical~~ datasets. Coupled with the generated data, ~~the~~ sophisticated analysis methods are being developed to discover correlations between the ~~molecular signatures genotypes~~ and phenotypes, ~~some of~~ which can contain sensitive information ~~about individuals~~ like disease status. ~~For example, the phenotype-to-genotype correlations can be utilized by an adversary in two different ways: First he can estimate a sensitive phenotype for an individual, whose genotype information is available. Secondly, the adversary can predict the genotype of an individual, whose phenotype information is available. As the genotype is a unique identifier for an individual, the individual's sensitive phenotype can be revealed as a consequence. Even when the strength of each phenotype-to-genotype correlation is not significantly high, the availability of millions of phenotype-to-genotype correlations to the adversary increases the risk of leakage of sensitive information.~~ It is therefore necessary for the ~~data~~ sharing mechanisms for

TRAV

these datasets to keep up with the analysis methods account for the types and extent of the genotype-to-phenotype correlations so as to control that the leakage of the predictable sensitive information in each study public datasets can be controlled.

Several previous studies have demonstrated the possibility of leakage of sensitive information under specific scenarios by exploiting 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 the fact 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.

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 addition, different formalities have been proposed for protecting sensitive information. For example differential privacy [10] establishes bounds on the leakage on sensitive information in statistical databases. This formality imposes a stringent tradeoff between utility and privacy. Thus, it has been shown that differential privacy mechanisms can substantially decrease the utility of the biological information [11]. Another approach is homomorphic encryption [12], which enables performing operations on encrypted data directly. This framework offers complete protection of sensitive information as the processors of the data never interacts with the unencrypted sensitive information. This approach, however, is not practically applicable as they require very high computational complexity and storage requirements for encrypted data. Another well-established formality is k-anonymization [13]. In this formality, the released dataset is anonymized by 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 high in computational complexity and not practical for high dimensional biomedical datasets. Several variants have been proposed that extend k-anonymity framework [14, 15]. Much of

the previous literature focused on protection of genotype datasets. As the size and nature of the biomedical datasets change, it is necessary to build analysis frameworks that can uniformly quantify the predictability of genotypes and characterizability of individuals using the phenotype datasets exploiting the phenotype-to-genotype datasets.

In this paper, we focus on characterizability of the individuals' sensitive information in the context of linking attacks, where the adversary exploits the phenotype-to-genotype correlations to reveal sensitive information. In the linking attack, there are ~~two~~three datasets: The first dataset contains ~~identities and gene expression levels~~the measurement of a series of phenotypes for a set of individuals. ~~The second dataset contains genotypes and a piece of sensitive information.~~Examples for each individual, e.g., HIV the phenotypes can be blood sugar level, measurement of several metabolite and biomarker levels, and gene expression levels in the blood but also disease status, for states like HIV state, and cancer diagnosis and prognosis. As these phenotypes can be sensitive, the dataset is de-identified by removal of names and then it is released publicly. The second dataset contains the genotypes of another set of individuals. The attacker aims at linking the identities in gene expression dataset to the sensitive information in ~~genotype dataset by first~~Since genotype information can reliably identify individuals as shown in previous publications, this dataset is not released publicly and released by permission only. The adversary gains access to these datasets. He then aims at characterizing the individuals in the genotype ~~dataset by predicting the genotypes of variants using the gene expression levels then~~from the phenotypes and matching the predicted genotypes to the genotype dataset. For ~~the prediction, the attacker~~he utilizes publicly available eQTL datasets, which report the genes and variants whose expressions and genotypes are significantly ~~a~~third dataset, where correlations between the genotypes and phenotypes are reported. For each individual ~~in the phenotype dataset, using the value of a phenotype, the attacker computationally predicts the mostly genotype that is correlated.~~This attack enables the attacker to characterize sensitive information about the individuals, and represents a privacy breach.

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 only the quantification of gene expression levels in the publicly accessible datasets. Even when only the expression levels are released, expression quantitative trait loci (eQTLs) can be utilized for linking gene expression datasets to the genotype datasets. Each eQTL contains a common genetic variant and a gene such that the genotype of the ~~variant~~with that phenotype. The basic idea is significantly that the prediction will be of higher accuracy, compared to random guessing of genotypes, given that the genotype and phenotype are 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 is a large and growing compendium of public eQTL datasets. For example, GTex project hosts approximately 30 million eQTLs whose gradient and significance information can be viewed freely through eQTL Browser [3], ~~each other.~~It should also be noted that the attacker aims at predicting as many genotypes correctly as he can so that the most number of individuals are characterized correctly.

Among all the datasets, the most abundant and well-studied phenotype-to-genotype correlation dataset is expression quantitative trait loci (eQTL) datasets. These datasets are generated by genome-wide screening for correlations between the variant genotypes and gene expression levels usually through RNA sequencing or expression arrays [16, 17]. Several publications evaluated different aspects of RNA-seq data privacy [18, 19]. The eQTL datasets are especially useful in the context of linking attacks since there is a large and growing compendium of public eQTL datasets. For example, GTex project hosts millions of eQTLs where the users can view in detail how the genotypes and expression levels are associated [3]. In order to demonstrate our results and build the formulations in a specific context, we will focus on eQTL datasets and linking of gene expression and genotype datasets. It is worth noting that most of the results and analyses can be extended to other types of phenotype-to-genotype correlations.

RNA
SEQ

One publication that relates to our study is [20], where the authors demonstrate that an adversary can build a model for predicting genotypes for eQTLs using gene expression levels. The authors show that given the model, individuals can be identified with high accuracy. Our study follows the study in [20] and generalizes the results of in two ways: First we ~~study~~ quantifying the amount of characterizing information leakage that can be generalized to other types of genotype-to-phenotype correlations. Secondly, we show that the adversary does not require a model based prediction. For this, we introduce a new simple metric extremity, which can be computed in the absence of a model show that this metric can be utilized in linking and can characterize a high fraction of individuals in the representative dataset.

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 characterization framework and study different aspects of vulnerability using the framework. We utilize a representative dataset and evaluate the fraction of characterizable individuals in a general setting. In the last section, to illustrate a practicality of the attack scenario, we present a simple and generally applicable genotype prediction method and evaluate the fraction of characterizable individuals on the representative dataset.

WAY

DISCUS

Formatted: Font: 2 pt

2 RESULTS

2.1 Overview of the Individual Characterization Scenario by Linking Attacks

Figure 1a illustrates the general privacy breaching scenario that is considered. ~~In the context of breach~~As we introduced earlier, there are ~~two~~three datasets. ~~in the context of the breach~~. First dataset contains gene expression levels and HIV disease status (HIV+/HIV-) ~~the phenotype information~~ for ~~a~~ set of individuals. The ~~gene expression dataset is de-identified by removal of the names. This dataset is released for public access.~~phenotypes can include sensitive information such as disease status in addition to several molecular phenotypes such as gene expression levels, blood cholesterol levels, and other metabolite levels. The second dataset contains the genotypes and the identities for ~~n_p~~ 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 disease state (i.e., HIV+ or HIV-) of each of the n_g individuals. The third dataset contains a correlation between one or more of the phenotypes in the gene expression phenotype dataset. For this, the attacker predicts and 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. In this dataset, each entry contains a phenotype, a variant, and the degree to which these values are correlated. In order to formulate and demonstrate the results, we will focus on the gene expression dataset as the phenotype dataset. As explained earlier, the abundance of gene expression-genotype correlation between the eQTL variant genotypes and eQTL gene expression levels for predicting eQTL variant genotypes with high accuracy (eQTL) datasets makes these datasets most suitable for linking attacks.

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 (RVs) which we denote with V_k and E_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 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 our 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 analysis independent of the fraction of individuals prediction methodology that are vulnerable to

linking attack in different settings, the attacker utilizes 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 Leakage of Individual Characterizing Information

We assume that the attacker will behave in a way that maximizes his chances of characterizing the most number of individuals. Thus, he will try and predict the genotypes for the largest set of variants that he believes are he can predict correctly. The most obvious way that the attacker does this is by first sorting the genotype-to-phenotype correlations with respect to decreasing strength of correlation. He will then predict the genotypes starting from the top genotype-phenotype pair. As he predicts more genotypes, he increases his chances of characterizing more individuals. As the attacker goes down the list, though, the correct predictability of the genotypes diminish, i.e., the strength of genotype-phenotype correlation decreases. Thus, each time the attacker predicts a new genotype, he will encounter a tradeoff between the number of genotypes that the attacker can predict correctly versus the cumulative correctness of the all the predicted genotypes. This tradeoff can also be viewed as the tradeoff between sensitivity (what fraction of the individuals can be characterized by correctly predicted genotypes) and positive predictive value (correct predictability of the genotypes) of the characterization. In this section we will propose two measures to quantify this tradeoff.

In the context of the linking attack introduced in Section 2.1, the attacker aims to correctly characterize n_e individuals in the expression dataset among n_v individuals in the genotype dataset whose disease states are known. In order to correctly characterize 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 can be matched to the correct 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:

Given the genotypes of an individual, if the attacker can reliably correctly predict a subset of genotypes that contain $\log_2(n_v)$ bits of information using the genotypes predicted from expression data for an individual, the individual is vulnerable to characterization of their disease state. It should be noted that, assuming the independence of the genotypes for different eQTLs, we can decompose the quantity of individual characterizing information that is leaked for a set of n correctly predicted eQTL genotypes:

$$ICI(\{V_1 = g_1, V_2 = g_2, \dots, V_n = g_n\}) = \sum_{k=1}^n \frac{-\log(p(V_k = g_k))}{\text{Convert the genotype frequency to number of bits that can be used to characterize individual}}$$

Sum individual characterizing information for all variants

where V_k is the RV that corresponds to the genotypes for the k^{th} eQTL, g_k is a specific genotype (Refer to Methods Section 3.1 for more details), and $p(V_k = g_k)$ denote the genotype frequency of g_k within the population, and ICI denotes the total individual characterizing information. Evaluating the above

formula, ICI increases as the frequency of the variant's genotype g_k decreases. In other words, the more rare genotypes contribute higher to ICI compared to the more common ones. Thus, individual linking 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 ICI for the predicted genotypes is at least $\log(n_v)$.

In order to maximize the amount of ICI , the attacker will aim at correctly predicting as many eQTL genotypes 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 (i.e., more correctly predictable) 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 ICI 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 jointly the predictability of multiple eQTL genotypes. In order to uniformly quantify the joint predictability of the eQTL genotypes using the 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

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

where π denotes the predictability of V_k given the gene expression level $e_{k,j}$. π can be interpreted as the average probability (over sampling of individuals from the general population) that the attacker can correctly predict the eQTL genotype given the expression level. In the above 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).

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 characterizing information, i.e., high $-\log(p(V_k = g_k))$, 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.

The relation between ICI and π is important as the ICI quantifies the amount of leakage in characterizability that the predicted eQTL genotypes and π quantifies how likely that characterization can occur. We will now use ICI and π to evaluate how predictability changes with increasing leakage in the individual characterizing information on the GEUVADIS dataset, which we use as a representative dataset. As discussed earlier, the attacker will aim at predicting the largest number of eQTL genotypes given the expression levels to maximize his characterization power. For this, we assume the attacker will sort the eQTLs with respect to the absolute value of correlation then predict the eQTL genotypes starting from the first eQTL. In order to evaluate the tradeoff between the identifying information of the top predictable eQTLs and their predictabilities, we plotted average ICI 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 $n=1,2,3,\dots,20$ eQTLs, we estimated mean π and mean ICI over all the samples. We then plotted mean π versus mean ICI for each n which is shown in Fig 2a. There is significant leakage of ICI 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 sample 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. On the representative dataset, these results illustrate that there is substantial amount of leakage at significant levels of predictability.

2.3 A Generalized Framework for Analysis of Individual Characterization

In this section, we present a 3 step framework for individual characterization in the context of linking attacks. Figure 3a summarizes the steps in the individual characterization for each individual. The input is the gene expression levels for j^{th} individual in the expression dataset, e_j . The aim of the attacker is to correctly link the disease state of the individual to the correct identity in the genotype dataset. In the first step, the attacker selects the eQTLs (among n_q eQTLs) that will be used in linking j^{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 characterization 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).

2.4 Fraction of Individuals Vulnerable to Characterization

In this section, we utilize the general setting we presented in Section 2.3 and evaluate the fraction of characterizable individuals in the representative dataset. 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.

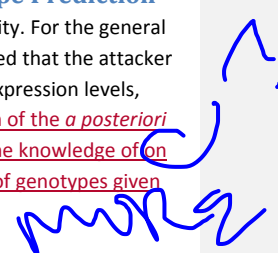
Using the list of predicted eQTL genotypes selected at each absolute correlation cutoff, the attacker performs the 3rd step in the attack and links the predicted genotypes to the genotype dataset to identify individuals (Refer to Methods Section 4.4). Figure 5a shows the fraction of vulnerable individuals. 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 characterizing information leakage as the accuracy of the predicted genotypes increase while there is a balancing decrease in the characterizing information leakage with decreasing number of eQTL genotypes predicted.

We also evaluate the scenario 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 characterization 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 choose.

2.5 Individual Characterization using Extremity based Genotype Prediction

In the previous section, we presented a general framework for analysis of vulnerability. For the general applicability of the framework in different genotype prediction scenarios, we assumed that the attacker can correctly reconstruct the *a posteriori* distribution of genotypes given the gene expression levels, which is then used to estimate the MAP genotype. In general, correct reconstruction of the *a posteriori* distribution of the genotypes given expression levels may not be possible because the knowledge of on the correlation coefficient is not enough to regenerate the *a-posteriori* distribution of genotypes given the expression levels.

In this section, we present a simple approach for estimating the *a posteriori* distribution of eQTL genotypes given the expression levels. For this, the attacker exploits the knowledge that the eQTL genotypes and expression levels are linearly correlated with each other and therefore extremes of the gene expression levels (highest and smallest expression levels) coincide with extremes of the genotypes (homozygous genotypes). Therefore, given the gradient of association, the attacker can very roughly estimate the joint distribution of the eQTL genotypes and expression levels. This idea is illustrated Fig



XX. Using the joint distribution, the attacker can compute the a posteriori distribution of genotypes given gene expression levels. To quantify the extremeness of expression levels, we use a statistic we termed *extremity*. For the gene expression levels for kth eQTL, e_k , *extremity* of the jth individual with expression level $e_{k,j}$ is defined as

$$extremity(e_{k,j}) = \frac{\text{rank of } e_{k,j} \text{ in } \{e_{k,1}, e_{k,2}, \dots, e_{k,n_e}\}}{n_e} - 0.5.$$

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 posterior distribution of kth eQTL genotypes can be formulated as

$$P(V_k = 0 | E_k = e_{k,j}) = \begin{cases} 0 & \text{if } extremity(e_{k,j}) \times \rho(E_k, V_k) > 0 \\ 1 & \text{otherwise} \end{cases}$$

$$P(V_k = 2 | E_k = e_{k,j}) = \begin{cases} 1 & \text{if } extremity(e_{k,j}) \times \rho(E_k, V_k) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$P(V_k = 1 | E_k = e_{k,j}) = 0.$$

From the *a posteriori* probabilities, when the sign of the extremity and the reported correlation are the same, the attacker assigns the genotype value 2, and otherwise, genotype value 0. Finally, the genotype value 1 is never assigned in this prediction method, i.e., the a posteriori probability is zero. Using these probabilities, we utilized extremity based prediction and assessed the accuracy. Figure XX shows the accuracy of genotypes predictions changing correlation threshold on the selected set of eQTLs. As expected, the accuracy of genotype predictions increases with increasing correlation threshold.

We next utilized the extremity based prediction in the 2nd step of the individual characterization framework (Fig 2) and evaluated the fraction of characterizable individuals in the GEUVADIS dataset. 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. Fig XX shows the fraction of vulnerable individuals. 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 plots), the fraction of vulnerable individuals increases to 100% for most of the eQTL selections. These results suggest that linking attack with extremity based genotype prediction, although technically simple, can be extremely effective in characterizing individuals.

3 CONCLUSION AND DISCUSSION

In this paper we first analyzed the leakage of individual characterizing information and its predictability. We also proposed a framework for analysis of sensitive individual characterizing information leakage in the context of linking attacks. The premise of sharing genomic information is that there is always an amount of leakage in the sensitive information [19],[21]. We believe that the quantification

methodology and the analysis framework can be applied for analysis of the *ICI* leakage in the genomic datasets where the correlative relations between datasets can be exploited for performing linking attacks.

The analysis of tradeoff between predictability and leakage of *ICI* can be generalized in two ways in future studies: First, the information theoretic measures that we proposed for measuring predictability versus the *ICI* 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 *ICI* leakage, which can identify on average 2000 individuals. At 1% predictability, there is around 18 bits of *ICI*, which can identify on average approximately 64000 individuals. Depending on the probability of leakage that can be tolerated, the predictability versus *ICI* leakage can be utilized to assess whether the dataset can be released to public access or not.

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 [20],[22]. Our study contributes to quantifying the individual identifying information leakage.

We also finally, we introduced a simple yet effective approach for characterization of individuals. The approach utilizes extremity based genotype prediction method using the extremity statistic. 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 characterization 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:

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

where V_k is the RV that represents the k^{th} eQTL genotype and g ($g \in \{0,1,2\}$) is a specific genotype for G , $p(G = g)$ is the probability (frequency) of the genotype in the sample set and *ICI* 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:

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

[[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 $V_{(i)}$ 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

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

[[How do we estimate the genotype entropy? The interpretation of the measure is that the prediction process is a uniform random guessing where the average correct prediction probability is measured by π .]]

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 , $\tilde{v}_{l,j} = \{\tilde{v}_{l,j}\}$, the attacker links the predicted genotypes to the individual whose genotypes have the smallest distance to the predicted genotypes:

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

$pred_j$ 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]]

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, Venkatasubramanian 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. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464**:768–772.~~
- ~~18. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, Price A, Raj T, Nisbett J, Nica AC, Beazley C, Durbin R, Deloukas P, Dermitzakis ET: **Patterns of Cis regulatory variation in diverse human populations.** *PLoS Genet* 2012, **8**.~~
- ~~19. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57–63.~~
- ~~20. 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.~~
- ~~21. Schadt EE, Woo S, Hao K: **Bayesian method to predict individual SNP genotypes from gene expression data.** *Nature Genetics* 2012:603–608.~~

| [1921](#). Narayanan A, Yocum K, Glazer D, Farahany N, Olson M, Stein LD, Williams JB, Witkowski JA, Kain RC, Erlich Y: *Redefining Genomic Privacy: Trust and Empowerment*. 2014.

| [2022](#). 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.