Analysis of Information Leakage in Phenotype and Genotype Datasets

Arif Harmanci, Jieming Chen, Dov Greenbaum, Mark Gerstein

ABSTRACT

Genomic privacy is receiving much attention with the unprecedented increase in the breadth and depth of biomedical datasets. Moreover, considering the legislative plans for encouraging public data sharing in biomedical research fields, privacy will the key factor in designing mechanisms for how these datasets will be shared. Most studies on genomic privacy are focused on protection of variants in personal genomes. Molecular phenotype datasets, like functional genomics datasets, can also contain substantial amount of sensitive information. Although there is no explicit genotypic information in them, the subtle phenotype-genotype correlations can be used to statistically predict genotypes from phenotypes. Predicted genotypes can then be used to link the phenotype datasets the genotype datasets. The linkages can potentially characterize sensitive information about individuals, e.g. disease information. In this paper, we develop a formalism for quantification and analysis of potential individual characterizing information leakage in a linking attack. We analyze the tradeoff between the predictability of the genotypes and the amount of leaked information that can be used in linking and individual characterization. Then we show how one could practically instantiate an attack focusing on the most commonly available data sets, those of RNA-seq and eQTL. We develop a three step procedure showing how an attacker would select eQTLs, statistically predict the genotypes, and then perform linking based on the predicted genotypes which can be very accurate considering the high dimensionality of phenotypes. The linking attack becomes particularly easy to perform when one deals with outlier gene expression levels. To study this, we developed a particular realization of this attack for the outlier cases and quantified the amount of information leakage.

Deleted: Individual Characterizing

Deleted: however

Deleted: entries in

Deleted: to those in

Deleted: Each linkage

Deleted: some

Deleted: an individual. This linking attack can be very accurate considering the high dimensionality of phenotypes.

Deleted: .

1 BACKGROUND

The decreasing cost of DNA sequencing [1] has rendered a massive increase in the amount of high-dimensional personalized biomedical data being generated [2]. The molecular phenotype datasets, when generated in a high-throughput manner, substantially grow the list of the *quasi-identifiers* (such as birth date, ZIP code, gender) for the respective individual, which can be used by an adversary to identify them. Moreover, with the recent announcement of Precision Medicine Initiative [3], a very large body of these datasets is to be generated and shared among researchers [4]. Following this, National Institutes of Health recently released the plans to encourage public access to biomedical datasets from scientific studies [4–6]. Considering the fact that one does not need many identifiers to uniquely pinpoint an individual [7–9], these datasets has the potential to exacerbate the risk of privacy breach.

Many consortia, like GTex [10], ENCODE [11], 1000 Genomes [12], and TCGA [13], are generating large amount of personalized biomedical datasets. Coupled with the generated data, sophisticated analysis methods are being developed to discover correlations between genotypes and phenotypes, some of which can contain sensitive information like disease status. Although these correlations could be useful for discovering how genotypes and phenotypes interact, they could also be utilized by an adversary in a linking attack for matching the entries in datasets where genotypes and phenotypes are stored. For example, when phenotype dataset is available, the adversary can utilize the phenotype-genotype correlations to statistically predict the genotypes, compare the predicted genotypes with the entries in another dataset that contains genotypes. For the entries that are correctly matching, he/she can reveal sensitive phenotypes of the individuals and characterize them. Even when the strength of each phenotype-to-genotype correlation is not high, the availability of a large number of phenotype-genotype correlations to the adversary increases the accuracy of correct linking. In fact, an adversary can perform correct linking with relatively small number of genotypes [14, 15].

Several previous studies have demonstrated the possibility of individual identification under specific scenarios. In [16], 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 [17], the authors identify the identities of several male participants of 1000 Genomes Project [12] by using the Y-chromosome short tandem repeats as an individual identifying biomarker. A more detailed review can be found [18]. In addition, different formalisms have been proposed for protecting sensitive information. For example, differential privacy [19] establishes bounds on the leakage of sensitive information in statistical databases. The solution involves building release mechanisms that track how much information is leaked and stops release when the estimated leakage privacy is above a predetermined threshold.

This formalism imposes a stringent tradeoff between utility and privacy. It has been shown that differential privacy mechanisms can substantially decrease the utility of the biological data [20]. In addition, the release mechanism must keep track of all the queries, which can complicate the way that the data is served [21]. Another approach is homomorphic encryption [22], which enables performing operations on encrypted data directly. Complete protection of sensitive information is guaranteed as the data processors never interact with the unencrypted sensitive information. The drawback, however,

Deleted: Many consortia, like GTex [3], ENCODE [4], 1000 Genomes [5], and TCGA [6], are generating large amount of personalized biomedical datasets.

Moved down [1]: Coupled with the generated data, sophisticated analysis methods are being developed to discover correlations between genotypes and phenotypes, some of which can contain sensitive information like disease status. Although these correlations could be useful for discovering how genotypes and phenotypes interact, they could also be utilized by an adversary in a linking attack for matching the entries in datasets where genotypes and phenotypes are stored. For example, when phenotype dataset is available, the adversary can utilize the phenotype-genotype correlations to statistically predict the genotypes, compare the predicted genotypes with the entries in another dataset that contains genotypes. For the entries that are correctly matching, he/she can reveal sensitive phenotypes of the individuals and characterize them. Even when the strength of each phenotype-to-genotype correlation is not high, the availability of a large number of phenotype-genotype correlations to the adversary increases the accuracy of correct linking

Moved (insertion) [1]

Deleted: Several previous studies have demonstrated the possibility of individual identification under specific scenarios. In [7], 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 [8], the authors identify the identities of several male participants of 1000 Genomes Project [5] by using the Y-chromosome short tandem repeats as an individual identifying biomarker. A more detailed review can be found [9]. In addition, different formalisms have been proposed for protecting sensitive information. For example differential privacy [10] establishes bounds on the leakage of sensitive information in statistical databases. This formalism imposes a stringent tradeoff between utility and privacy. 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. Complete protection of sensitive information is guaranteed as the data processors never interact with the unencrypted sensitive information. The drawback, however, is high computational and storage requirements. Another well-established formalism is k-anonymization [13]. 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, has high computational complexity and is 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--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--

is high computational and storage requirements. Another well-established formalism is k-anonymization [23]. 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. In this approach the anonymization process has, however, excessive computational complexity and is not practical for high dimensional biomedical datasets [24]. Several variants have been proposed that extend k-anonymity framework [25, 26]. 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-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-genotype correlations to reveal sensitive information. In the linking attack, there are three datasets: The first dataset contains the measurement of a series of phenotypes for a set of individuals.

[[A More technical list of phenotypes here illustrates the point better.]]

Examples for the phenotypes can be blood sugar level, measurement of several metabolite and biomarker levels, and gene expression levels in the blood but also disease 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. 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 from the phenotypes and matching the predicted genotypes to the genotype dataset. For prediction, he utilizes 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 most likely genotype that is correlated with that phenotype. The basic idea is that the prediction will be of higher accuracy, compared to random guessing of genotypes, given that the genotype and phenotype are correlated with 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 [27–29]. The eQTL datasets are especially useful in the context of linking attacks since there is a large and growing compendium of public eQTL datasets [30]. For example, GTex project hosts a sizeable set of eQTL dataset from multiple studies where the users can view in detail how the genotypes and expression levels are associated [10, 31]. In order to demonstrate our results and build the formulations in a specific context, we will focus on eQTL datasets and linking of

Deleted: [16–18]. The eQTL datasets are especially useful in the context of linking attacks since there is a large and growing compendium of public eQTL datasets [19]. For example, GTex project hosts a sizeable set of eQTL dataset from multiple studies where the users can view in detail how the genotypes and expression levels are associated [3].

gene expression and genotype datasets. It is, however, worth noting that most of the results and analyses can be extended to other types of phenotype-to-genotype correlations.

One publication that relates to our study is [32], 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 [32] and generalizes the results 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 linking can be performed in a much simplified genotype prediction approach by just utilizing the outliers in the data. For this, we introduce a new simple metric extremity and show that this metric can be utilized in genotype prediction. When large set of eQTLs are utilizes, linking can be done with high accuracy.

The paper is organized as follows: We first analyze the genotype predictability and evaluate the tradeoff between the amount of information leakage and correct predictability of the genotypes. Next we present the 3 step individual characterization framework and study different aspects of vulnerability using the framework. 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.

[[Can we make the genotype-phenotype analysis code available for download?]]

2 RESULTS

2.1 Overview of the Individual Characterization Scenario by Linking Attacks

Figure 1a illustrates the general privacy breaching scenario that is considered. There are three datasets in the context of the breach. First dataset contains the phenotype information for a set of individuals. The 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 another set of individuals. The third dataset contains a correlations between one or more of the phenotypes in the phenotype dataset and the genotypes. 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 (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 q. The eQTL (gene) expression levels and eQTL (variant) genotypes are stored in $q \times n_e$ and $q \times n_v$ 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, $\boldsymbol{e_k}$, contains the gene expression values for k^{th} eQTL entry and $e_{k,j}$ represents the

Deleted: One publication that relates to our study is [20]

Deleted: [20] and generalizes the results of

Deleted: n_q.

Deleted: $n_q \times n_e$

Deleted: $n_q \times n_v$

expression of the k^{th} gene for j^{th} individual. Similarly, k row of v, $\boldsymbol{v_k}$, contains the genotypes for k^{th} eQTL variant and $v_{k,j}$ represents the genotype ($v_{k,j}$ ϵ {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. We denote the correlation between the RVs 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 homozygous genotype corresponds to higher expression levels. 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 and the heterozygous genotypes associate with moderate levels of expression. The eQTL studies utilize linear models to identify the gene and variant pairs whose expressions and genotypes that are significantly correlated.

Deleted: complicated

Given this knowledge, the adversary aims at reversing this operation so as to predict genotypes given the gene expression levels. For generalization of our analysis, we assume that the he/she utilizes a prediction model that estimates correctly the *a posteriori* distribution of the eQTL genotypes given the eQTL expression levels, i.e., $p(V_k|E_k)$. This enables us to perform the analysis independent of the prediction methodology that 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/her chances of characterizing the most number of individuals. Thus, he/she will try and predict the genotypes, using the phenotype measurements, 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 phenotype-genotype pairs with respect to decreasing strength of correlation as illustrated in Fig 2a. He will then predict the genotypes starting from the top phenotype-genotype pair. As he/she predicts more genotypes, he/she increases his/her chances of characterizing more individuals. As the attacker goes down the list, however, the correct predictability of the genotypes diminish, i.e., the strength of phenotype-genotype correlation decreases. Thus, each time he/she predicts a new genotype, he/she will encounter a tradeoff between the number of genotypes that can be predicted correctly versus the cumulative correctness of the all the predicted genotypes. This tradeoff can also be viewed as the tradeoff between precision (correct predictability of the genotypes) and recall (what fraction of the individuals can be characterized by correctly predicted genotypes). In this section we will propose two measures to quantify this tradeoff.

Deleted: genotype-to-
Deleted: correlations
Deleted: .
Deleted: genotype-
Deleted: genotype-

[[Following is moved here from below, must blend it with the text]]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

Moved (insertion) [2]

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.

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, he/she should select a set of eQTLs that he/she believes he/she 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 correctly predict a subset of genotypes that contain $\log_2(n_v)$ bits of information, 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:

$$ICI(\{V_1 = g_1, V_2 = g_2, \dots, V_n = g_n\}) = \overbrace{\sum_{k=1}^{n} \underbrace{-\log\big(p(V_k = g_k)\big)}_{\substack{\text{Convert the genotype} \\ \text{frequency to number of bits} \\ \text{that can be used to characterize} \\ \text{individual}}}}^{\substack{\text{Sum individual characterizing} \\ \text{information for all variants}}}$$

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

Formatted: Font color: Text 1

Deleted: the attacker

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 (over sampling of individuals from the 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. (Refer to Methods Section 4.1 for more details).

In general, ICI represents the recall rate for the attacker's predictions and π represents the precision of his/her predictions. We will now use ICI and π to evaluate how predictability changes with increasing leakage. We will use GEUVADIS dataset as the 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 correlation, $|\rho(E_k, V_k)|$. Then for top *n=1,2,3,…,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 2b. 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, 2048 individuals, which on average can characterize all samples in the dataset. 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 General 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

Deleted: general

Deleted: In order to extend the predictability measure to multiple of eQTLs, we use exponential of the negative of joint conditional entropy.

Moved up [2]: 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.¶

Formatted: Font color: Text 1

Deleted: The relation between

Deleted: 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.

Deleted: in the individual characterizing information on the

Deleted: , which we use

Deleted: a

Deleted: 2a

Deleted: all the

Deleted: sample

Deleted: (At 12.4% predictability, the leakage is approximately 9 bits for 6 top eQTLs.)

Deleted: Generalized

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).

[[Talk a bit more about linking step: Perfect matches versus imperfect?]]

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).

2.3.1 Fraction of Vulnerable Individuals with MAP Genotype Prediction

To illustrate the results of linking attack, we evaluate the fraction of individuals that are vulnerable to characterization using gene expression and genotype data in GEUVADIS Project. 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 for characterization. The genotypes for the selected eQTLs are predicted using MAP prediction (Refer to Methods Section 4.3). Figure 4a shows, for each correlation threshold, the number of selected eQTLs and the fraction correctly predicted genotypes.

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). Each individual in expression dataset, who is linked to the right individual are flagged as vulnerable. 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 behavior 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

Deleted: Vulnerable to Characterization

Formatted: Heading 3

Deleted: In this section

Deleted: utilize the general setting we presented in Section 2.3 and

Deleted: characterizable

Deleted: in the representative dataset

Deleted: . Fig SXX shows the distribution of the absolute correlation levels for the eQTL dataset.

Deleted: MAP

Deleted: with changing absolute correlation thresholds

Formatted: Font: 11 pt

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 individuals 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.

2.4 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 only the phenotype-genotype 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 k^{th} eQTL, $\boldsymbol{e_k}$, $extremity$ of the jth individual with expression level $e_{k,j}$ is defined as

Deleted: kth

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

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

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

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

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

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

With the current pace of data generation coupled with the policies to encourage genomic data sharing, genomic privacy will be one of the a topic of hot debate. In the analysis of genomic privacy, however, it is necessary to consider the basic premise of sharing any type of personal information: There is always an amount of leakage in the sensitive information [33]. We believe this makes it necessary for the genomic data sharing and publishing mechanisms to incorporate quantification methods before the datasets are released. The quantification methodology and the analysis frameworks presented in this study can be applied for analysis of the information leakage in the datasets where the correlative relations between datasets can be exploited for performing linking attacks.

Deleted: 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 [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.

We introduced a simple yet effective genotype prediction method that utilizes the simple extremity statistic. This approach capitalizes on the fact that an individual who is an outlier for a phenotype will most likely harbor a homozygous genotype. When employed in the individual identification framework, this simple approach renders a very significant number of individuals vulnerable. This illustrates the viability of individual characterization from gene expression datasets.

Deleted: extermity

Compared to other formalisms, our study aims to develop and build on other studies for quantifying the information leakage and help setup a framework for analysis of the leakage of individual characterizing information. Differential privacy, 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, however, it is necessary to analyze the sources of sensitive information leakage so that one can design the utility maximizing release mechanisms [34]. The metrics that we presented can be used to analyze the correlative structures as the sources and quantify the risk and amount of leakage associated with these sources.

Deleted: Compared to other formalisms, our study aims more to characterize the leakage of individual characterizing information. Differential privacy, 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, , however, it is necessary to analyze the sources of sensitive information leakage so that one can design the utility maximizing release mechanisms [22]. Our study contributes to quantifying the individual characterizing information leakage.¶

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:

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

where V_k is the RV that represents the k^th eQTL genotype 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 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:

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

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_k 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

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

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

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

Deleted: $V_{(l_i)}$

Deleted: j^th

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

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

Deleted: for Quantification of Predictability

Formatted: Font: 2 pt

Deleted: [[We bin the expression values to log_2(N_i) different bins \cite{...}]]¶

We estimate the genotype entropy using the Shannon's entropy [35]:

$$H(V_k) = -\sum_{v\in\{0,1,2\}} p(V_k = v) \times \log(p(V_k = v))$$

Where V_k represents the RV for k^th eQTL variant genotypes and $p(V_k = v)$ represents the probability that V_k takes the value v. This probability can be also interpreted as the population frequency of the genotype v at the k^{th} eQTL's variant locus. As the genotypes are discrete valued, the above formula can be computed in a straightforward way by the summation.

[[Rewrite above; add histogram building.]]

In the formulations, we also use the conditional specific entropies [35] of the genotypes given the gene expression levels. For this, we use the following formulation:

$$H(V_k|E_k = e_{k,j}) = -\sum_{v\in\{0,1,2\}} p(V_k = v|E_k = e_{k,j}) \times \log(p(V_k = v\,|E_k = e_{k,j}))$$

where $p(V_k = v|E_k = e_{k,j})$ represents the conditional probability that V_k takes the value v under the condition that the RV representing gene expression level for k^{th} eQTLs (E_k) is $e_{k,j}$. Since the gene expression levels are continuous, to estimate the conditional probabilities of genotypes given expression levels; we start with the joint distribution of E_k and V_k, then bin the gene expression levels. For this, we use Sturges' rule [36] to choose the number of bins. This rule states that the number of bins should be selected as:

$$n_{bins} = \lceil \log(n_e) \rceil + 1 = \lceil \log(426) \rceil + 1 = 10$$

The binning is done for each gene by first sorting the expression levels for all the individuals, then the range of gene expression levels are divided into 10 bins of equal size and each expression level is mapped to a value between 1 and 10. The expression level of k^{th} gene in j^{th} individual, $e_{k,j}$, is mapped to

$$\tilde{e}_{k,j} = \left\lceil \frac{(e_{k,j} - \min(\boldsymbol{e_k})) \times 10}{\max(\boldsymbol{e_k}) - \min(\boldsymbol{e_k})} \right\rceil$$

Where $\min(\boldsymbol{e_k})$ represents the minimum and maximum values for the k^{th} expression level over all the samples and $\tilde{e}_{k,j}$ represents the binned expression level. After the gene expression levels are binned, we

use the binned expression levels and compute the conditional distribution of the variant genotypes at each binned gene expression level using the histograms:

$$p(V_k = v|\tilde{E}_k = \tilde{e}_{k,j}) = \frac{\sum_i I(\tilde{e}_{k,i} = \tilde{e}_{k,j}, V_{k,i} = v)}{\sum_i I(\tilde{e}_{k,i} = \tilde{e}_{k,j})}$$

where

$$I(\tilde{e}_{k,i} = \tilde{e}_{k,j}, V_{k,i} = v) = \begin{cases} 1; & \text{if } \tilde{e}_{k,i} = \tilde{e}_{k,j}, V_{k,i} = v \\ 0; & \text{otherwise} \end{cases}$$

Finally, we utilize the probabilities estimated from histograms to compute the condition specific genotype entropies.

4.3 MAP (Maximum *a posteriori*) Genotype Prediction

While assigning the genotypes, the attacker assigns to V_k the genotype that maximizes the estimated conditional probability:

$$\text{MAP}(V_k\,|\tilde{E}_k = \tilde{e}_{k,j}) = \underset{v}{\text{argmax}}(p(V_k = v|\tilde{E}_k = \tilde{e}_{k,j}))$$

where the conditional probabilities are estimated as in Methods Section 4.2.

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:

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

$pred_j$ 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.

Deleted: -
Deleted: [[Describe
Deleted: genotypes are not assigned any
Formatted: Font: 11 pt
Deleted: binning and MAP selection of
Deleted: bc of the selection]]
Formatted: Font: 11 pt
Deleted:]]¶
[[Must include SNP selection such that some of
Formatted: Font: 11 pt
Formatted: Font: 11 pt
Formatted: Font: 11 pt
Formatted: Font: 11 pt

[[Any other ways to do match?]]

Deleted: <#>Extremity Attack¶
[[Define the extremity attack: Correlation and extremity parameters]]¶

5 DATASETS

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

6 FIGURE CAPTIONS

[[Add the figure captions]]

7 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. **jnfographic-printable.pdf** [http://www.nih.gov/precisionmedicine/infographic-printable.pdf]

4. Collins FS: **A New Initiative on Precision Medicine**. *N Engl J Med* 2015, **372**:793–795.

5. **Plan for Increasing Access to Scientific Publications - NIH-Public-Access-Plan.pdf** [https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf]

6. **GENOMIC DATA SHARING (GDS) Home** [http://gds.nih.gov/index.html]

7. Sweeney L: *Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4*. 2000.

8. Sweeney L, Abu A, Winn J: **Identifying Participants in the Personal Genome Project by Name**. *SSRN Electron J* 2013:1–4.

9. Golle P: **Revisiting the uniqueness of simple demographics in the US population**. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*; 2006:77–80.

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

11. 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.

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

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

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

Deleted: 4.

Deleted: 5

Deleted: 6

14. Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, Kidd KK: **SNPs for a universal individual identification panel**. *Hum Genet* 2010, **127**:315–324.

15. Wei YL, Li CX, Jia J, Hu L, Liu Y: **Forensic Identification Using a Multiplex Assay of 47 SNPs**. *J Forensic Sci* 2012, **57**:1448–1456.

16. 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**.

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

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

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

20. 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.

21. Adam NR, Worthmann JC: **Security-control methods for statistical databases: a comparative study**. *ACM Computing Surveys* 1989:515–556.

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

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

24. Meyerson A, Williams R: **On the complexity of optimal K-anonymity**. In *Proceedings of the twentythird ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems PODS 04*; 2004:223–228.

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

26. 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.

27. 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.

28. 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**.

Deleted: 7

Deleted: 8

Deleted: 9

Deleted: 10

Deleted: 11

Deleted: 12

Deleted: 13

Deleted: 14

Deleted: 15

Deleted: 16

Deleted: 17

29. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464**:773–777.

30. Xia K, Shabalin AA, Huang S, Madar V, Zhou YH, Wang W, Zou F, Sun W, Sullivan PF, Wright FA: **SeeQTL: A searchable database for human eQTLs**. *Bioinformatics* 2012, **28**:451–452.

31. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, et al.: **The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans**. *Science (80-)* 2015, **348**:648–660.

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

33. 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.

34. 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.

35. Cover TM, Thomas JA: *Elements of Information Theory*. 2005.

36. Herbert A. Sturges: **The Choice of a Class Interval**. *J Am Stat Assoc* 1926, **21**:65–66.

Deleted: 18

Deleted: 19

Deleted: 20

Deleted: 21

Deleted: 22

Deleted: