Analysis of Information Leakage in Phenotype and Genotype Datasets

Arif Harmanci[1,2], Mark Gerstein[1,2,3,*]

1 Program in Computational Biology and Bioinformatics, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
2 Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
3 Department of Computer Science, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
*Corresponding authors: Mark Gerstein pi@gersteinlab.org

[[FIGURE CAPTIONS]]

[[FIGURE UPDATES]]

[[RESPONSE LETTER]]

[[Figure references]]

All [[…]]'s

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 be the key consideration in designing data sharing mechanisms. Most studies on genomic privacy are focused on protection of variants in personal genomes. Molecular phenotype datasets, however, can also contain substantial amount of sensitive information. Although there is no explicit genotypic information in them, subtle genotype-phenotype correlations can be used to statistically link the phenotype and genotype datasets. The links can then be used to characterize individuals by identifying their sensitive phenotypes and breaching privacy. Here, we develop a formalism for the quantification and analysis of 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 perform linking based on the predicted genotypes. The linking 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 leaked.

1 BACKGROUND

Privacy is one of the most important topics of debate in data science that stands at the corner of many different fields, including ethics, sociology, law, political science, and forensic science. Recently, genomics has emerged as one of the major foci of studies on privacy. This can mainly be attributed to the advancement of technologies for high throughput biomedical data acquisition that bring about a surge of datasets [1,2]. Among these, high throughput molecular phenotype datasets, like functional genomic and metabolomic measurements, substantially grow the list of the *quasi-identifiers* (such as birth date, ZIP code, gender[3]) for participating individuals, which can be used by an adversary for re-identification of the identities. With the recent announcement of Precision Medicine Initiative[4], a large body of datasets are to be generated and shared among researchers[5]. The National Institutes of Health also released the plans to encourage public access to biomedical datasets from scientific studies [5–7]. Considering the fact that one does not need many identifiers to uniquely pinpoint an individual[3,8,9], these datasets have 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 are 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 genotype and phenotype datasets. For example, when a phenotype dataset is available, the adversary can utilize the genotype-phenotype 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 genotype-phenotype correlation is not high, the availability of a large number of genotype-phenotype correlations increases the scale of linking. In fact, an adversary can perform correct linking with relatively small number of genotypes[14,15].

[[Divide the genomic privacy attacks here into several different categories? Genome in a mixture? Linking? Data publishing/serving?]]

Many different aspects of privacy have been intensely studied. Recently, genomic privacy is receiving much attention as a result of the deluge of personalized genomics datasets that are being generated[16,17]. Several studies have demonstrated the possibility of individual re-identification based on analysis of genotypic information. Homer et al[18] showed that a statistical testing procedure enables testing whether a genotyped individual is in a pool of samples, for which only the allele frequencies are known.

[[Im et al reference]]

Deleted: -Cox

In another study[19], the authors identify the identities of several male participants of 1000 Genomes Project[12] by using the short tandem repeats on Y-chromosome as an individual identifying biomarker. A

detailed review can be found elsewhere[20]. In addition, different formalisms for protecting sensitive information have been proposed and applied to genomic privacy. These censor or hide information, or aim at ensuring statistical indistinguishability of individuals in the released data. For example, differential privacy[21] involves building data release mechanisms that have guaranteed bounds on the leakage of sensitive information. The release mechanisms track how much information is leaked and stops release when the estimated leakage is above a predetermined threshold. Although this approach is theoretically very appealing, studies showed that it can substantially decrease the utility of the biological data[22]. In addition, the release mechanism must keep track of all the queries, which can cause complications in data sharing[23]. Homomorphic encryption[24] enables performing analysis 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[25,26]. Before releasing the dataset, it is anonymized by data perturbation techniques for ensuring that no combination of features in the dataset are 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[27]. Several variants have been proposed for extending k-anonymity framework[28,29]. A majority of these studies aim at protecting the genomic variants and identities of individuals in databases. Different aspects of genomic privacy, pertaining linkability of high dimensional phenotype datasets to genotypes, are yet to be explored.

In this paper, we focus on characterizability of the individuals' sensitive information in the context of linking attacks, where the adversary exploits the genotype-phenotype correlations to reveal sensitive information. In general, the high dimensional phenotype datasets generated in genomic studies harbor a number of phenotypes that contain sensitive information, like disease status, and other phenotypes, while not sensitive, may have subtle correlations with genomic variant genotypes. Many quantitative phenotypes can be linked to genotypes using public quantitative trait loci (QTL) datasets. Some quantitative traits and corresponding QTLs can be body mass index[30], basal glucose levels[31], serum cholesterol levels[32,33], gene expression levels (eQTLs), protein levels (pQTLs[33,34]), DNase hypersensitivity site signals (dsQTLs[35]), and also higher order traits like network modularity (modQTLs[36]). Correlations can potentially cause a small amount of genotypic information leakage, which, when utilized by an adversary at a large number of loci, can be used to link the sensitive phenotypes to the genotype dataset. Since genotypes can almost perfectly identify an individual, this linking attack can potentially cause a breach of privacy for the individuals who participated in the studies.

Among all the datasets, the most abundant and well-studied genotype-phenotype 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 [37–39]. The eQTL datasets are especially useful in the context of linking attacks since there is a large and growing compendium of public eQTL datasets[40]. 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 [10,36]. 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, however, worth noting that most of the results and analyses can be trivially generalized to other types of genotype-phenotype correlations.

One publication[41] relates to our study, 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 [41] and generalizes the results in two ways: First we study quantification of characterizing information leakage versus risk of characterization in an information theoretic setting. Secondly, we show that the linking can be performed in a much simplified setting by just utilizing the outliers in the data. For this, we introduce a new metric, we termed extremity, and show that this metric can be utilized in genotype prediction and linking attacks with high accuracy with a model-free procedure. When a large set of eQTLs are used, linking can be done with high accuracy.

Deleted:

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 the practicality of the attack scenario, we present extremity based genotype prediction method and evaluate the fraction of characterizable individuals on the representative dataset. The analysis tools and code are available for download at http://privaseq.gersteinlab.org.

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. The second dataset contains the genotypes and the identities for another set of individuals. The third dataset contains 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 datasets as the representative 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. The k^{th} row of e, $\boldsymbol{e_k}$, 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^{th} row of v, $\boldsymbol{v_k}$, 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. The coding of the genotypes from homozygous or heterozygous genotype categories to the numeric values are done according to the correlation dataset (See Methods Section 4.1). 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 terms of a gradient (or the regression coefficient) in addition to the significance of association (p-value) between genotypes and expression levels. 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 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. Given this knowledge, the adversary aims at reversing this operation so as to predict genotypes for each individual, using the respective gene expression levels and the genotype-phenotype correlation. For general applicability of the analysis, we assume that he/she utilizes a prediction model that estimates correctly the *a posteriori* distribution of the eQTL genotypes given the eQTL expression levels, i.e., $p(V_k|E_k)$, as illustrated in Fig S2b. 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 Correct Predictability of 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/she believes he/she can predict correctly. The most obvious way that the attacker does this is by first sorting the genotype-phenotype pairs with respect to decreasing strength of correlation as illustrated in Fig 2a. He/She will then predict the genotypes starting from the top genotype-phenotype 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 genotype-phenotype 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 (fraction of the linkings that are correct) and recall (fraction of individuals that are correctly linked). In this section we will propose two measures to quantify this tradeoff.

In the context of the linking attack, the attacker aims to correctly characterize n_e individuals in the expression dataset among n_v individuals in the genotype dataset. 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

Deleted: predictability of the genotypes

Deleted: what

Deleted: the

Deleted: can be characterized by

Deleted: predicted genotypes

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 joint frequency of the set of predicted genotypes for the selected eQTLs should be $\frac{1}{n_v}$. 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 at least $\log_2(n_v)$ bits of information, the individual is vulnerable to characterization of his/her phenotypes. Following this statement, we can quantify the leakage from a set of correctly predicted eQTL variant genotypes as the logarithm of their joint frequency. Assuming that the genotypes of different eQTLs (See Section 5) are independent from each other, we can decompose the quantity of individual characterizing information that is leaked for a set of n correctly predicted eQTL genotypes:

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

where V_k is the random variable 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 $p(V_k = g_k)$ denotes 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 $\log(n_v)$. *ICI* can also be interpreted as the number of rare SNP genotypes that an individual harbors.

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 the joint predictability of multiple eQTL genotypes. In order to uniformly quantify the joint (correct) predictability of the eQTL genotypes using the expression levels, we use the exponential of entropy of the conditional genotype distribution given gene expression levels. Given the expression levels for j^{th} individual, we compute the predictability of the k^{th} eQTL genotypes as

Deleted: at most

Deleted:

$$\pi(V_k|E_k = e_{k,j}) = \underbrace{\exp(-1 \times \overbrace{H(V_k|E_k = e_{k,j})}^{\substack{\text{Randomness left in } V_k \\ \text{given } E_k = e_{k,j}}})}_{\substack{\text{Convert the entropy to} \\ \text{average probability}}} \quad (2)$$

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

We first considered each eQTL and evaluated the genotype predictability versus the characterizing information leakage. We use the GEUVADIS dataset as a representative dataset for this computation (Refer for Section 5). For this, we computed, for each eQTL, average π and average *ICI* over all the individuals, which is plotted in Fig 2b. Most of the data points are spread along the diagonal, which indicate that there is a natural tradeoff between correct predictability and *ICI* leakage. The eQTL variants with rare (minor) allele frequencies have high predictability and low *ICI* and vice versa for common eQTL variants with common allele frequencies (Fig 2b, left). This is expected because the genotypes of the rare variants can be predicted, on average, easily (most individuals will harbor the major allele) and consequently does not deliver much characterizing information. The genotypes for the eQTLs with common alleles, however, are harder to predict as they are mostly uniformly distributed among population. On the other hand, these eQTLs contain high *ICI* on average. The eQTLs with high correlation (Fig 2b, right) deviate from the diagonal with high ICI and high predictability. It can be seen that these highly informative eQTLs are also variants with high allele frequency. In principle, the adversary will aim at identifying and using these eQTLs with high ICI and predictability. The shuffled gene-variant pairs, on the other hand, are distributed mainly along the diagonal (Fig S1a).

The risk of characterizability increases substantially when the adversary utilizes multiple genotype predictions at once. We will now use ICI and π to evaluate how predictability changes with increasing leakage when multiple genotypes are utilized. As discussed earlier, the attacker will aim at predicting the largest number of eQTL genotypes given the expression levels to maximize 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 characterizing information of the top predictable eQTLs and their predictabilities, we plotted average *ICI* versus average π for top genotype predictions. 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 as illustrated in Fig S2a. We then plotted mean π versus mean *ICI* for each *n* which is shown in Fig 2c. From the plot, we can first estimate the number of vulnerable

Deleted: rare

Deleted: eQTLs

Deleted: common

individuals at different predictability levels. For example, at 20% predictability, there is approximately 8 bits of ICI leakage. At this level of leakage, the adversary can correctly link all individuals, on average with 20% chance, in a sample of $2^8 = 256$ individuals. At 5% predictability, the leakage is 11 bits and the characterizable sample size is $2^{11} = 2048$ individuals, which can be interpreted as a higher risk of characterizability. These estimates are useful when releasing QTL datasets such that the leakage risks can be assessed besides the released list of genotype-phenotype correlations. Another view is to evaluate the risk at which a given sample of individuals can be characterized. For a dataset of n_v individuals, as explained earlier, it is necessary to predict $\log(n_v)$ bits of genotypic information correctly. The risk of characterization can be determined from the graph as the predictability level at which $\log(n_v)$ bits of *ICI* leakage is observed. The auxiliary information knowledge can also be incorporated into this analysis easily. For example, assuming that the sample set contains 10,000 individuals, it is necessary to correctly predict $\log(n_v = 10{,}000) = 13.3$ bits of information. At around 5% predictability, the adversary can gain 11 bits of information. Even though this cannot uniquely characterize all individuals, if the attacker can gain $13.3 - 11 = 2.3$ bits of auxiliary information, e.g. gender and ethnicity, he/she can characterize all individuals correctly. Since many phenotypic measurements have significant predictive power for gender, the attacker can predict it correctly, which gains the attacker 1 bit of auxiliary information.

[[Add a discussion of how this is useful: Exact matching and querying a database.]]

Deleted: It is worth noting that since

Deleted: , most of the time,

Deleted: bits

Formatted: Highlight

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 3 summarizes the steps in the individual characterization for each individual. The input is the phenotype measurements for j^{th} individual. 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 QTLs, which will be used in linking j^{th} individual. The selection of QTLs can be based on different criteria. As described in the previous section, the most accessible criterion is selection based on the absolute gradient or the absolute strength of association between the phenotypes and genotypes. In the case of eQTLs, this is the reported correlation coefficient, $|\rho(E_k, V_k)|$. In our analysis, we evaluate the effect of changing correlation coefficient. It is worth noting that the adversary can use other measures of correct predictability to select the set of QTLs that he/she will utilize in the linking process. The second step is genotype prediction for the selected QTLs 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 phenotypes. The attacker then uses the posterior probabilities of the genotypes to identify the maximum *a posteriori* (MAP) genotype. 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).

2.3.1 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 4b 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 (Fig S3). 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. It has been previously shown that gene expression levels show widespread differences with respect to gender[42]. In addition, it has been shown that the ethnicity and population differences can be observed in the gene expression levels[43,44]. These indicate that gender and ethnicity can be inferred from gene expression levels. 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 4b 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.

Deleted:

Deleted:

2.4 Individual Characterization using Extremity based Genotype Prediction

In the previous section, we presented a general framework for analysis of vulnerability. For the 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 genotype-phenotype correlation coefficient is not enough to regenerate the *a posteriori* distribution of genotypes given the expression levels.

The attacker can, however, utilize a priori knowledge about relation between gene expression levels and genotypes to estimate roughly the *a posteriori* distribution of genotypes. Even though the genotype prediction may not be very accurate, the attacker can utilize a large number of eQTLs to maximize the

Deleted: In this section, we present a simple approach for estimating the *a posteriori* distribution of eQTL genotypes given the expression levels.

linking accuracy. For this, the attacker exploits the knowledge that the eQTL genotypes and expression levels are correlated such that the allele effects on expression are additive and 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 estimate coarsely the joint distribution of the genotypes and expression levels. This idea is illustrated in Fig 5a. Using an estimate of 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 j^{th} individual with expression level $e_{k,j}$ is defined as

$$extremity(e_{k,j}) = \frac{\text{rank of } e_{k,j} \text{ in } \{e_{k,1}, e_{k,2}, \dots, e_{k,n_e}\}}{n_e} - 0.5. \quad (3)$$

Extremity can be interpreted as a normalized rank, which is bounded between -0.5 and 0.5. Figure S4 shows the mean absolute extremity distribution of all the gene expression levels for all the individuals. The average median extremity is uniformly distributed among individuals.

Following from the above discussion, the adversary builds the posterior distribution for k^{th} eQTL genotypes as

$$P(V_k = 0 \mid E_k = e_{k,j}) = \begin{cases} 0 \text{ if } extremity(e_{k,j}) \times \rho(E_k, V_k) > 0 \\ 1 \text{ otherwise} \end{cases} \quad (4)$$

$$P(V_k = 2 \mid E_k = e_{k,j}) = \begin{cases} 1 \text{ if } extremity(e_{k,j}) \times \rho(E_k, V_k) > 0 \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

$$P(V_k = 1 \mid E_k = e_{k,j}) = 0. \quad (6)$$

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. This is expected since we are focusing on the extremes and heterozygous genotype is expected to co-incide with medium levels of gene expression. For ease of interpretation, the genotype prediction can be interpreted as a rank correlation between the genotypes and expression levels and choosing the homozygous genotypes that maximize the rank correlation. Thus, this process can be generalized as a rank correlation based prediction. Using the probabilities, we utilized extremity based prediction and assessed the genotype prediction accuracy. Figure 5b shows the accuracy of genotype predictions with changing correlation threshold. As expected, the accuracy of genotype predictions increases with increasing correlation threshold.

It is worth noting that extremity is a fairly central concept in privacy. Extreme phenotypes and features represent potential identifying information. Any time an individual harbors an extreme (or outlier) feature, he/she can be distinguished from a large number of individuals in the whole population. Thus

Deleted: therefore

Deleted: [[How to interpret extremity and refer to extremity-like measures in the literature]]¶

Deleted: these

this feature can be used to identify him/her. Much of the data anonymization techniques aim at identifying and protecting the individuals that are outliers with respect to certain features. Here, we utilize the extremity to infer the genotypes and link the individuals based on the predicted genotypes.

We next utilized extremity based genotype prediction in the 2nd step of the individual characterization framework (Fig 3) 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 order to demonstrate the utility of the 3-step analysis framework; we evaluated two different distance measures for linking the predicted genotypes to the individuals in genotype dataset in the 3rd step of the attack. First is based on comparison of the predicted genotypes to all the genotypes in genotype dataset. Second is based on comparison of the predicted genotypes to only the homozygous genotypes in the genotype dataset (See Methods Section for details). The motivation for using this distance measure is that the extremity based genotype prediction never assigns heterozygous genotypes. Thus the heterozygous genotypes are excluded from distance computation.

For each measure, the attacker links the predicted genotypes to the individual whose genotypes minimize the selected distance measure. Figure 5c shows the fraction of vulnerable individuals for both distance measures. More than 95% of the individuals are vulnerable for most of the parameter selections for both distance measures. The homozygous genotype matching distance measure has slightly higher linking accuracy. 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 show that linking attack with extremity based genotype prediction, although technically simple, can be extremely effective in characterizing individuals. We will focus on homozygous genotype matching based distance computation in the rest of the paper for simplicity of presentation.

The previous results show that extremity based linking attacks are highly effective when the eQTLs are identified and linking attack is performed using the same expression and genotype datasets. In order to assess the accuracy when the eQTLs are computed and tested on different datasets, we divided the dataset into a training and a testing dataset. The training dataset, of 210 individuals, is used to discover the eQTLs, using Matrix eQTL[45] method (See Methods Section for details). The testing dataset, of 211 individuals, is utilized for assessing the accuracy of linking. Figure 6a shows the linking accuracy for individuals in testing dataset. The accuracy is very high, around 95%, which suggests that extremity based linking attacks are potentially effective when the datasets where eQTLs are identified do not match the data being tested. This is an important aspect of genotype prediction based linking attacks, as they exploit the generalizability of the correlations between phenotypes and genotypes.

We evaluated whether the attacker can estimate the reliability of the linkings. This may potentially increase the effectiveness of the linking and increase the risk associated with linking attacks because the attacker can estimate reliability of the linkings and choose the ones that are more likely to be correct. This increases the risk associated with the linking attacks because although he/she may not have a high overall accuracy of linkings, the high ranking linkings may be much higher in accuracy. We observed that the measure we termed, *first distance gap*, denoted by d_{2-1} (See Methods), serves as a

good reliability estimate for each linking. For a given linking, d_{2-1} is the difference between the genotype distances of the 1st closest and 2nd closest individuals to the predicted genotypes. When the linking is incorrect, we observed that d_{2-1} is very likely to be smaller than the distance difference when the linking is correct.

To evaluate this measure further, we computed the positive predictive value (PPV) versus sensitivity of the linkings of individuals in the testing set with changing d_{2-1} threshold. For this, we first computed d_{2-1} for each linking, then filtered the linkings that did not satisfy the threshold. Then we computed PPV and sensitivity of the linkings (See Methods), which is plotted in Fig 6b. It can be seen that the PPV of linkings can get very high at the same time with high sensitivity. For example, the attacker can link around 79% of the individuals at a PPV higher than 95%. The random sorting of the linkings, on the other hand, have significantly lower PPV (cyan in the plots) at the same sensitivity levels. These results suggest that the attacker can increase the potential risk (accuracy of linkings) of the attack by focusing on a slightly smaller set of linkings with high reliability.

We compared the accuracy of extremity based linking attack with the model based linking approach by Schadt et al[41] (See Section SXX). The results with different number of top eQTLs are shown in Table SXX. This shows that the model-free extremity based linking attack has comparable linking accuracy to the model based method, even when small number of markers are used in linking.

An important practical question is how well the linking accuracy changes with increasing genotype data size. In order to evaluate this, we simulated the genotypes of the eQTLs (discovered in the training set) for 100,000 individuals. The 100,000 simulated individuals are then merged with the testing dataset of 211 individuals to build the large testing dataset. We then performed the extremity attack using the expression levels of the testing dataset and linked them to the merged testing dataset of size 100,211 individuals. The linking accuracy is plotted in Fig 7a with changing eQTL selection criteria. The linking accuracy is very high (Around 96%). This result suggests that the extremity attack can be extended to a large testing sample set. Figure 7b shows the sensitivity versus PPV (with changing first gap distance) for the eQTLs for which the overall linking accuracy is 70% (Yellow dashed lines on Fig. 7b). It can be seen that the attacker can link around 55% of the individuals with PPV higher than 95%.

We also studied how the linking accuracy changes when the training and testing datasets are measured in different populations. For this, we used the 1000 Genomes Project sample information and divided the GEUVADIS samples into 5 populations. Then we used each population's samples to discover the population specific eQTLs, then used the other populations to test the linking accuracy. Table S1 shows the accuracies in each case. It can be seen that when the eQTLs are disovered in European populations (CEU, GBR, TSI, FIN), the linking accuracies are very high (higher than 95%). When the eQTLs are discovered in YRI (African) population, the linking accuracies are significantly smaller in European populations. Similarly, when eQTLs are discovered on European populations, the linking accuracy in YRI sample is relatively smaller. These results illustrate that extremity attack can still be effective when eQTLs are identified in populations that are genetically close to the population(s) of testing sample and decrease when the populations are diversified. [[Make sure this makes sense]]

We next studied scenario where the eQTLs are identified in tissues that are different from the tissues on which the expression data is generated. For this, we used the eQTLs that are identified by GTex Project [[cite]]. We downloaded the eQTLs for 5 tissues and performed the linking attack using each eQTL dataset to link the individuals in the GEUVADIS gene expression dataset to the individuals in 1000 Genomes Phase 1 genotype dataset [[cite]]. The results are shown in Table S2. The accuracy is highest for Whole Blood eQTLs, which is around 88%. This is expected since the expression levels in GEUVADIS project are measured in blood cell lines. The accuracy is smallest for Muscle Skeletal eQTLs, which is 76%.

We also studied whether having close relatives in the genotype dataset affects the accuracy. To test this, we used the expression and genotype data from 30 CEU trios (mother-father-child) from available from HAPMAP project [[cite]]. We first identified the eQTLs from the 90 individuals and performed linking over the same individuals. We then computed the average rank of the (non-self) close relatives in each linking. For example, when the tested individual is a father or mother, we computed the rank of the individual child and if the tested individual is a child, we computed the rank of his/her mother and father. We also selected, for each tested individual, random individuals and computed their ranks in the linking. The distribution of the ranks are shown in Fig XX. It can be seen that the ranks of the related individuals are significantly shifted to smaller values compared to random individuals. This result shows that the close relatives can get linked to each other. This result indicates that when the individuals that are close relatives may potentially be confused with each other. While the correct person may not get characterized, the attacker can still reveal sensitive information about the individual's family, which can cause privacy breaches for the family of the individual. [[These cases of must be handled appropriately while evaluating the sensitive information leakage.]]

3 CONCLUSION AND DISCUSSION

[[Add linking attack's importance, Add the relevance/generality of extremity/outlierness in the privacy literature]]

Increasing pace of data generation and the policies to encourage genomic data sharing will make genomic privacy a topic of hot debate. In the analysis of genomic privacy, 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 [46]. In addition, as shown by previous studies, we often cannot propose black-and-white solutions to problems in privacy which mainly roots from the multifaceted nature of privacy. We believe these make it necessary for the genomic data sharing and publishing mechanisms to incorporate statistical quantification methods before the datasets are released. Legislative decision making processes should incorporate the quantified risk estimates of leakage as an objective factor. 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. In accordance to a utility policy, the leakage risk can be evaluated against the utility requirements so as to assess the suitability of different data release mechanisms.

Deleted: [[We also compared the performance of the extremity based attack and Schadt et al]]¶

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 at different predictability levels. Depending on the risk 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. The 3-step framework aims at representing the framework for studying specific instantiations of the linking attacks. The decomposition of the attack into steps makes the analysis of different attacks easier as each step can be separately evaluated. For example, the genotype prediction and linking steps can be replaced with different approaches so as to evaluate how the linking accuracy varies. These can reveal insight into how the datasets should be protected. We also presented a simple yet accurate linking attack that utilizes genotype prediction method based on the 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. In addition, we also showed that the attacker can estimate the reliability of the linkings using the first gap distance statistic so as to increase the risk of correct characterization. This illustrates the viability of individual characterization utilizing technically simple approaches. Even though we observed that the attacker can characterize a large fraction of individuals with high PPV, the smaller fraction of individuals that are linked at the top with high gap distance statistic are under higher risk of being characterized.

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[47]. The metrics that we presented can be used to analyze the correlative structures as the leakage sources and quantify the risk and amount of leakage associated with these sources.

4 METHODS

4.1 Quantification of Individual Characterizing Information and Predictability

The genotype RV V_k takes 3 different values, {0,1,2}, where the genotype coding is done per counting the number of alternate alleles in the genotype. Given that the genotype is $g_{k,j}$, we quantify the individual characterizing information in terms of *self-information*[48] of the event that RV takes the value $g_{k,j}$:

$$ICI(V_k = g_{k,j}) = I(V_k = g_{k,j}) = -\log(p(V_k = g_{k,j})) \quad (7)$$

where V_k is the RV that represents the k^{th} eQTL genotype, $p(V_k = g_{k,j})$ is the probability (frequency) of that V_k takes the value $g_{k,j}$, and ICI denotes the individual characterizing information. Given multiple eQTL genotypes, assuming that they are independent, the total individual characterizing information is simply summation of those:

$$\begin{aligned} ICI(\{V_1 = v_{1,j}, V_2 = v_{2,j}, \dots, V_N = v_{N,j}\}) \\ = -\sum_{k=1}^{N} \log\left(p(V_k = v_{k,j})\right). \end{aligned} \quad (8)$$

The genotype probabilities are estimated by the frequency of genotypes in the genotype dataset. As presented in the Results Section 2.2, 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,

$$\pi(V_k|E_k = e) = \exp(-H(V_k|E_k = e)) \quad (9)$$

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

$$\begin{aligned} \pi(\{V_k\}, \{E_k = e_{k,j}\}) &= \exp(-H(\{V_k\} \mid \{E_k = e_{k,j}\})) \\ &= \exp\left(-\sum_k H(V_k|E_k = e_{k,j})\right) \end{aligned} \quad (10)$$

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

We estimate the genotype entropy using the Shannon's entropy[48]:

$$H(V_k) = -\sum_{v\in\{0,1,2\}} p(V_k = v) \times \log(p(V_k = v)) \quad (11)$$

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. These probabilities are estimated from the distribution of

Deleted: ¶
Where

genotypes over all the samples. As the genotypes are discrete valued, the above formula can be computed in a straightforward way by the summation after the probabilities are estimated.

In the formulation for conditional predictability of genotypes given expression levels, we also use the conditional specific entropies[48] of the genotypes given the gene expression levels. For this, we use the following formulation:

$$H(V_k|E_k = e_{k,j}) = -\sum_{v\in\{0,1,2\}} p(V_k = v|E_k = e_{k,j}) \times \log(p(V_k = v\,|E_k = e_{k,j})) \quad (12)$$

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 [49] to choose the number of bins. This rule states that the number of bins should be selected as:

$$n_b = \lceil \log(n_e) \rceil + 1 = \lceil \log(426) \rceil + 1 = 10 \quad (13)$$

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 $n_b = 10$ bins of equal size and each expression level is mapped to a value between in $[0, n_b - 1]$. The expression level of k^{th} gene in j^{th} individual, $e_{k,j}$, is mapped to

$$\tilde{e}_{k,j} = \left\lfloor \frac{(e_{k,j} - \min(\boldsymbol{e_k})) \times n_b}{\max(\boldsymbol{e_k}) - \min(\boldsymbol{e_k})} \right\rfloor \quad (14)$$

where $\min(\boldsymbol{e_k})$ and $\max(\boldsymbol{e_k})$ represents the minimum and maximum values, respectively, 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:

$$p(V_k = v|\tilde{E}_k = \tilde{e}_{k,j}) = \frac{\sum_i I(\tilde{e}_{k,i} = \tilde{e}_{k,j}, V_{k,i} = v)}{\sum_i I(\tilde{e}_{k,i} = \tilde{e}_{k,j})} \quad (15)$$

where $I(.)$ is an indicator function for counting the number of matching mapped expression and genotype values:

$$I(\tilde{e}_{k,i} = \tilde{e}_{k,j}, V_{k,i} = v) = \begin{cases} 1; & \text{if } \tilde{e}_{k,i} = \tilde{e}_{k,j}, V_{k,i} = v \\ 0; & \text{otherwise} \end{cases} \quad (16)$$

Finally, we utilize compute the Shannon entropy of the estimated conditional distribution as the condition specific entropies.

4.3 Maximum *a posteriori* (MAP) Genotype Prediction

While assigning the genotypes using maximum *a posteriori* prediction, the attacker assigns to V_k the genotype that maximizes the estimated conditional probability:

$$\text{MAP}(V_k|\tilde{E}_k = \tilde{e}_{k,j}) = \tilde{v}_{k,j} = \underset{v}{\text{argmax}}(p(V_k = v|\tilde{E}_k = \tilde{e}_{k,j})) \quad (17)$$

where the conditional probabilities are estimated as in Methods Section 4.2 and $\tilde{v}_{k,j}$ denotes the predicted genotype for V_k, given $\tilde{E}_k = \tilde{e}_{k,j}$.

4.4 Linking of the Predicted Genotypes to Genotype Dataset

The linking is the 3rd and last step of the linking attack. The aim is to compare the predicted genotypes from the phenotype dataset to the genotypes in the genotype dataset so as to match the samples in the phenotype dataset to those in genotype dataset. We will use the linking approach that evaluates the minimal distance between the compared genotypes but different methods can be used for genotype comparison. Given a set of predicted eQTL genotypes for individual j, $\tilde{\boldsymbol{v}}_{\cdot,\boldsymbol{j}} = \{\tilde{v}_{1,j}, \tilde{v}_{2,j}, \dots, \tilde{v}_{n_q,j}\}$, the attacker links the predicted genotypes to the individual whose genotypes have the smallest distance to the predicted genotypes:

$$pred_j = \underset{a}{\text{argmin}}\{d(\tilde{\boldsymbol{v}}_{\cdot,\boldsymbol{j}}, \boldsymbol{v}_{\cdot,\boldsymbol{a}})\}. \quad (18)$$

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

$$d(\tilde{\boldsymbol{v}}_{\cdot,\boldsymbol{j}}, \boldsymbol{v}_{\cdot,\boldsymbol{a}}) = \sum_{k=1}^{n_q}(1 - I(\tilde{v}_{k,j}, v_{k,a})) \quad (19)$$

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

$$I(\tilde{v}_{k,j}, v_{k,a}) = \begin{cases} 1 \text{ if } \tilde{v}_{k,j} = v_{k,a} \\ 0 \text{ otherwise} \end{cases} \quad (20)$$

Finally, j^{th} individual is vulnerable if $pred_j = j$. When auxiliary information is available, the attacker constrains the set of individuals while computing $d(\tilde{\boldsymbol{v}}_{\cdot,\boldsymbol{j}}, \boldsymbol{v}_{\cdot,\boldsymbol{a}})$ 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 $d(\tilde{\boldsymbol{v}}_{\cdot,\boldsymbol{j}}, \boldsymbol{v}_{\cdot,\boldsymbol{a}})$. This way the auxiliary information decreases the search space of the attacker.

4.5 Homozygous Genotype Matching based Linking of the Predicted Genotypes to Genotype Dataset

The extremity based genotype prediction predicts only homozygous genotypes. Therefore heterozygous genotypes in the genotype dataset will always increase the distance in linking step. To correct for this, the attacker can focus only on the homozygous genotypes while he/she is linking the

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

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

Deleted: ()

Deleted: get around

predicted genotypes to the genotype dataset. For this, a simple modification of the distance function is sufficient:

$$d^H(\tilde{\boldsymbol{v}}_{\cdot,j}, \boldsymbol{v}_{\cdot,\boldsymbol{a}}) = \frac{\sum_{k=1}^{n_q}(1 - I^H(\tilde{v}_{k,j}, v_{k,a}))}{n_a^H} \quad (21)$$

where n_a^H represents the number of homozygous genotypes in $\boldsymbol{a}$th individual and $I^H(\tilde{v}_{k,j}, v_{k,j})$ represents the homozygous match indicator:

$$I^H(\tilde{v}_{k,j}, v_{k,j}) = \begin{cases} 1 \text{ if } v_{k,a} = 0, \tilde{v}_{k,j} = v_{k,a} \\ 1 \text{ if } v_{k,a} = 2, \tilde{v}_{k,j} = v_{k,a} \\ 1 \text{ if } v_{k,a} = 1 \\ 0 \text{ otherwise} \end{cases} \quad (22)$$

This indicator function does comparison only when the genotype being matched ($v_{k,a}$) is homozygous. When $v_{k,a}$ is heterozygous, it acts as if the genotypes are the same, thus the distance function is updated only when the genotype being matched is a homozygous genotype. The normalization is necessary to convert the distance into a fraction so that the distances can be compared among different genotype samples.

4.6 First Distance Gap Statistic For Linking Reliability Estimation

Following the previous section, the attacker computes, for each individual, the distance to all the genotypes in genotype dataset, then identifies the individual with smallest distance. Let $d_{j,(1)}$ and $d_{j,(2)}$ denote the minimum and second minimum genotype distances (among $d^H(\tilde{\boldsymbol{v}}_{\cdot,\boldsymbol{j}}, \boldsymbol{v}_{\cdot,\boldsymbol{a}})$ for all $\boldsymbol{a}$) for jth individual. We propose using the difference between these distances as a measure of reliability of linking. For this, the attacker computes following difference:

$$d_{1,2} = d_{j,(2)} - d_{j,(1)} \quad (23)$$

First distance gap can be computed without the knowledge of the true genotypes, and is immediately accessible by the attacker with no need for auxiliary information.

[[Motivation for this statistic]]

4.7 eQTL Identification on Training Sets with Matrix eQTL[45]

For identification of eQTLs, we used Matrix eQTL[45] method. We first generated the testing and training sample lists by randomly picking 210 and 211 individuals, respectively, for testing and training sets. We then separated the genotype and expression matrices into training and testing sets. In order to decrease the run time, Matrix eQTL is run in cis-eQTL identification mode. After the eQTLs are generated, we filtered out the eQTLs whose FDR was larger than 5%. We finally removed the redundancy by ensuring that each gene and each SNP is used only once in the eQTL final list.

Deleted: [[Distance function definition]]¶

Deleted: , i.e., d_(1).

Deleted: the smallest and second smallest

Deleted: .

Deleted: [[What is the motivation for this?]]¶

Deleted: [[5% FDR, focus on only the cis-eQTLs]]¶

5 DATASETS

The normalized gene expression levels for 462 individuals and the eQTL dataset are obtained from gEUVADIS mRNA sequencing project[50]. The eQTL dataset contains all the significant gene-variant pairs with high genotype-expression correlation. To ensure that there are no dependencies between the variant genotypes and expression levels, we used the eQTL entries where gene and variants are unique. In other words, each variant and gene are found exactly once in the final eQTL dataset. The genotype, gender, and population information datasets for 1092 individuals are obtained from 1000 Genomes Project [12]. For 421 individuals, both the genotype data and gene expression levels are available.

Deleted:

6 FIGURE CAPTIONS

Figure 1: Illustration of the linking attack. (a) Phenotype dataset contains *q* different phenotype measurements and the HIV Status for a list of individuals. Genotype dataset contains the variants genotypes for *n* individuals. Phenotype-Genotype correlation datasets contains *q* phenotypes, variants, and their correlations. The attacker does genotype prediction for all the variants and links the phenotype dataset to the genotype dataset by matching the genotypes. The linking potentially reveals the HIV status for the subjects in the genotypes dataset. (b) Illustration of the expression and genotype datasets. Variant genotype dataset contains the genotypes for *q* eQTL variants for n_v individuals. j^{th} entry for k^{th} eQTL is denoted by $v_{k,j}$. Similarly, the expression dataset contains the expression levels for *q* genes. The k^{th}expression level for j^{th} individual is denoted by $e_{k,j}$. The variant genotypes for k^{th} variant is distributed over samples with distribution specified by the random variable V_k. Likewise, the expression levels for k^{th} gene is distributed per random variable E_k. These random variables are correlated with each other with correlation coefficient, denoted by $\rho(E_k, V_k)$ (bottom).

Figure 2: Quantification of ICI and correct genotype predictability (a) Adversary's genotype prediction strategy. The phenotype-genotype correlations ρ_1, ρ_2,… are sorted with respect to decreasing absolute correlation, as shown on each line. For a selected set of *n* variants, the genotypes are predicted using the phenotypes. The green and red individuals on the right represent the vulnerable and non-vulnerable individuals, respectively. (b) Plots show, for each, eQTL the information leakage (x-axis) versus correct genotype predictability (y-axis). For each eQTL, the estimated ICI leakage and genotype predictability are plotted. Each eQTL's point is colored with respect to allele frequency (top left) and with respect to absolute correlation of the eQTL (top right). (c) Average predictability versus average individual characterizing information leakage. For the top 20 eQTLs, the plot shows the distribution of average predictability and average ICI leakage for the top eQTLs. The number of eQTLs that are used for computing the values at each point are shown next the the point. Only 10 of them are numbered in the figure. The error bars show the standard deviations among the sample set. The cyan plot shows the same plot for shuffled gene-variant pairs. The error bars are left out for simplification.

Figure 3: The figure illustrates the steps of the linking attack. The first step consists of selecting the phenotypes and genotype to be used in linking. The absolute value of correlation can be used as one of the selection criteria. The second step comprises the genotype prediction using the selected set of phenotypes. Maximum *a posteriori* genotype prediction can be used for prediction. Third step in

characterization is the linking step, where the predicted genotypes are matched to the genotype dataset. The matching can be performed by comparing the distance between the predicted genotypes and individual genotypes in the dataset.

Figure 4: MAP genotype prediction accuracy and vulnerable fraction. (a) The number of eQTLs selected (blue) and the number of correctly predicted eQTL genotypes (red). At each absolute correlation threshold, the number of eQTLs passing the threshold are shown and the number of correctly predicted genotypes using MAP prediction are shown. The error bars show the distribution of accuracy over all the samples. (b) The fraction of vulnerable individuals with MAP genotype prediction. X-axis shows the absolute correlation threshold used to select eQTLs. Y-axis shows the fraction of vulnerable individuals. At correlation threshold of 0.35, the fraction is maximized, as indicated by the dashed yellow line. The red, green, and cyan lines show the fraction of vulnerable individuals when gender, population, and gender and population information, respectively, are available as auxiliary information.

Figure 5: Extremity based genotype prediction and extremity based linking attack (a) Figure illustrates the extremity based genotype prediction. The joint distribution of expression levels and genotypes is shown on left. Given the relation between expression and genotypes, the lower expression levels (Labelled with "Negative Extremity" shown in red ellipse on left) co-associate with the genotype "TT" and higher expression levels (Labelled with "Positive Extremity" shown in green ellipse on left) co-associate with the genotype "CC". (b) The extremity based genotype prediction accuracy versus the absolute correlation threshold used to select the eQTLs. (c) The fraction of vulnerable individuals versus the correlation threshold in blue. The red, green, and cyan plots show the vulnerable fraction when gender, population, and both gender and population are available, respectively, as auxiliary information.

[[New figure captions]]

Figure S1: The figure shows different properties of the eQTLs. (a) The average ICI leakage versus the genotype predictability is shown for real (red) and shuffled (blue) eQTL dataset is shown. (b) The absolute correlation versus predictability is shown.

Figure S2: Figure shows the attacker's presumed strategy for linking attack. (a) The phenotype and variant pairs are sorted with respect to decreasing absolute correlations values. For the top n pairs, joint predictability and ICI are computed. (b) Illustration of prior, joint, and posterior distribution of genotypes and expression levels. Leftmost figure shows the distribution of genotypes over the sample set, which is labelled as the prior distribution. Middle figure shows the joint distribution of genotypes and expression levels. Notice that there is a significant negative correlation between genotype values and the expression levels. Rightmost figure shows the posterior distribution of genotypes given that the gene expression level is 10. The posterior distribution has a maximum at genotype 2, which is indicated by a star.

Figure S3: The distribution of ranks of the individuals in the linking step. At each gradient threshold, the box plots show, for each individual, their ranks in the genotype comparison in the 3rd step of linking

attack with MAP genotype prediction. Notice that at around 0.35 correlation threshold, the assigned ranks are minimized, i.e., most of the individual are linked correctly.

Figure S4: The median absolute extremity over 462 individuals in GEUVADIS dataset. For each individual, the extremity is computed over all the genes reported in the expression dataset. The median of the absolute value of the extremity is plotted. X-axis shows the sample index and y-axis shows the extremity. The absolute median extremity fluctuates around 0.25, which is exactly the mid point between minimum and maximum values of absolute extremity.

[[New supplementary figure captions]]

7 REFERENCES

1. Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* **12,** 125 (2011).

2. Rodriguez, L. L., Brooks, L. D., Greenberg, J. H. & Green, E. D. The Complexities of Genomic Identifi ability. *Science (80-.).* **339,** 275–276 (2013).

3. Sweeney, L., Abu, A. & Winn, J. Identifying Participants in the Personal Genome Project by Name. *SSRN Electron. J.* 1–4 (2013). doi:10.2139/ssrn.2257732

4. infographic-printable.pdf. at <http://www.nih.gov/precisionmedicine/infographic-printable.pdf>

5. Collins, F. S. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372,** 793–795 (2015).

6. Plan for Increasing Access to Scientific Publications - NIH-Public-Access-Plan.pdf. at <https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>

7. GENOMIC DATA SHARING (GDS) Home. at <http://gds.nih.gov/index.html>

8. Sweeney, L. *Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4*. *Forthcom. B. entitled, Identifiability Data.* (2000).

9. Golle, P. Revisiting the uniqueness of simple demographics in the US population. in *Proc. 5th ACM Work. Priv. Electron. Soc.* 77–80 (2006). doi:http://doi.acm.org/10.1145/1179601.1179615

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

11. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

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

13. Collins, F. S. The Cancer Genome Atlas (TCGA). *Online* 1–17 (2007).

14. Pakstis, A. J. *et al.* SNPs for a universal individual identification panel. *Hum. Genet.* **127,** 315–324 (2010).

15. Wei, Y. L., Li, C. X., Jia, J., Hu, L. & Liu, Y. Forensic Identification Using a Multiplex Assay of 47 SNPs. *J. Forensic Sci.* **57,** 1448–1456 (2012).

16. Church, G. *et al.* Public access to genome-wide data: Five views on balancing research with privacy and protection. *PLoS Genet.* **5,** (2009).

17. Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. From genetic privacy to open consent. *Nat. Rev. Genet.* **9,** 406–411 (2008).

18. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4,** (2008).

19. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339,** 321–4 (2013).

20. Erlich, Y. & Narayanan, A. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15,** 409–21 (2014).

21. Dwork, C. Differential privacy. *Int. Colloq. Autom. Lang. Program.* **4052,** 1–12 (2006).

22. Fredrikson, M., Lantz, E., Jha, S. & Lin, S. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. in *23rd USENIX Secur. Symp.* (2014). at <http://www.biostat.wisc.edu/~page/WarfarinUsenix2014.pdf>

23. Adam, N. R. & Worthmann, J. C. Security-control methods for statistical databases: a comparative study. *ACM Comput. Surv.* **21,** 515–556 (1989).

24. Gentry, C. A FULLY HOMOMORPHIC ENCRYPTION SCHEME. *PhD Thesis* 1–209 (2009). doi:10.1145/1536414.1536440

25. SWEENEY, L. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* **10,** 557–570 (2002).

26. Loukides, G., Gkoulalas-Divanis, A. & Malin, B. Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 7898–7903 (2010).

27. Meyerson, A. & Williams, R. On the complexity of optimal K-anonymity. in *Proc. twentythird ACM SIGMOD-SIGACT-SIGART Symp. Princ. database Syst. Pod. 04* 223–228 (2004). doi:10.1145/1055558.1055591

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

29. Ninghui, L., Tiancheng, L. & Venkatasubramanian, S. t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. in *Proc. - Int. Conf. Data Eng.* 106–115 (2007). doi:10.1109/ICDE.2007.367856

30. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42,** 937–948 (2010).

31. Cheverud, J. M. *et al.* Quantitative trait loci for obesity- and diabetes-related traits and their dietary responses to high-fat feeding in LGXSM recombinant inbred mouse strains. *Diabetes* **53,** 3328–3336 (2004).

32. Beekman, M. *et al.* Evidence for a QTL on chromosome 19 influencing LDL cholesterol levels in the general population. *Eur. J. Hum. Genet.* **11,** 845–850 (2003).

33. Holdt, L. M. *et al.* Quantitative trait loci mapping of the mouse plasma proteome (pQTL). *Genetics* **193,** 601–608 (2013).

34. Stark, A. L. *et al.* Protein Quantitative Trait Loci Identify Novel Candidates Modulating Cellular Response to Chemotherapy. *PLoS Genet.* **10,** (2014).

35. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).

36. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348,** 648–660 (2015).

37. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).

38. Stranger, B. E. *et al.* Patterns of Cis regulatory variation in diverse human populations. *PLoS Genet.* **8,** (2012).

39. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).

40. Xia, K. *et al.* SeeQTL: A searchable database for human eQTLs. *Bioinformatics* **28,** 451–452 (2012).

41. Schadt, E. E., Woo, S. & Hao, K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat. Genet.* **44,** 603–608 (2012).

42. Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* **4,** 2771 (2013).

43. Spielman, R. S. *et al.* Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* **39,** 226–231 (2007).

44. Storey, J. D. *et al.* Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* **80,** 502–509 (2007).

45. Shabalin, A. A. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358 (2012).

46. Narayanan, A. *et al. Redefining Genomic Privacy: Trust and Empowerment*. *bioRxiv* (2014). doi:10.1101/006601

47. Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P. & Palamidessi, C. Differential privacy: On the trade-off between utility and information leakage. in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **7140 LNCS,** 39–54 (2012).

48. Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. *Elem. Inf. Theory* (2005). doi:10.1002/047174882X

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

50. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–11 (2013).