# Significance

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 datasets1,2. 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, gender3) for participating individuals, which can be used by an adversary for re-identification of the identities. With the recent announcement of Precision Medicine Initiative4, a large body of datasets are to be generated and shared among researchers5. 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 individual3,8,9, these datasets have the potential to exacerbate the risk of privacy breach.

Many consortia, like GTex10, ENCODE11, 1000 Genomes12, and TCGA13, 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 genotypes14,15.

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 generated16,17. With the increase in the number of large scale genotyping and phenotyping studies, the protection of privacy of participating individuals emerged as an important issue. Homer et al18 proposed a statistical testing procedure that enables testing whether a genotyped individual is in a pool of samples, for which only the allele frequencies are known. Im et al19 showed that, given the genotypes of a large set of markers for an individual, an attacker can reliably predict whether the individual participated to a QTL study or not. These attacks, which we refer to as “detection of a genome in a mixture”, are one type of attacks on privacy (Fig S6). There is yet another important attack where the attacker links two or more datasets to pinpoint individuals in datasets and reveal sensitive information. One well-known and illustrative example of these “linking attacks”, although not in a genomic context, is the linking attack that matched the entries in Netflix Prize Database and the Internet Movie Database (IMDB)20. For research purposes, Netflix released an anonymized dataset of movie ratings of thousands of viewers, which they thought was secure as the viewers’ names were removed. However, Narayanan et al20 used IMDB database, a seemingly unrelated and very large database of movie viewers, linked the two databases, and revealed identities and personal information (movie history and choices) of many viewers in the Netflix database. The fact that Netflix and IMDB host millions of individuals in their databases renders the question of detection of an individual in these database irrelevant since any random individual is very likely to be in one or both of these databases but the focus of attacks turns to matching individuals in the databases. Consequently, as the databases grow, the attacks for detection of an individual in a database become unimportant and the linking attacks become more admissible in order to characterize individuals’ sensitive information. In the genomic privacy context, as the size and number of the genotype and phenotype datasets increase, possibility of potentially linkable datasets will increase, which may make scenarios similar to Netflix attacks a reality in genomic privacy (Fig S6, Section S2).

[[INNOVATION]] 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 link different datasets and potentially 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 of the high dimensional genomic quantitative traits and corresponding QTLs are gene expression levels (eQTLs), protein levels (pQTLs32,33), DNase hypersensitivity site signals (dsQTLs34), ribosome occupancy (rQTLs35), DNA methylation levels (meQTLs36), histone modification levels (haQTLs37–39), RNA splicing (sQTLs40), and also higher order traits like network modularity (modQTLs41). Other QTLs associated with single dimensional non-genomic phenotypes include body mass index42, basal glucose levels43, and serum cholesterol levels32,44. Each QTL can potentially cause a small amount of genotypic information leakage. As these QTLs are often identified and reported at genomic scale, when an adversary utilizes a large number of QTLs in the attack, he/she can accurately 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.

## AIM1: Development of a Statistical Formalism for Leakage from QTL Sets

We will develop a statistical framework for analysis of linking information leakage from QTL sets.

## 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 . The eQTL (gene) expression levels and eQTL (variant) genotypes are stored in and matrices and , respectively, where and denotes the number of individuals in gene expression dataset and individuals in genotype dataset. The row of , , contains the gene expression values for eQTL entry and represents the expression of the gene for individual. Similarly, row of , , contains the genotypes for eQTL variant and represents the genotype ( ϵ {0,1,2}) of variant for 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 eQTL entry are distributed randomly over the samples in accordance with random variables (RVs) which we denote with and , respectively. We denote the correlation between the RVs with . 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 indicates the strength of association between the eQTL genotype and the eQTL expression level. The sign of 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., , 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.

## 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 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 individuals in the expression dataset among 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 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 . 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 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 correctly predicted eQTL genotypes:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where is the random variable that corresponds to the genotypes for the kth eQTL, is a specific genotype (Refer to Methods Section 3.1 for more details), and denotes the genotype frequency of 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 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 . *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 . Thus, the attacker will try to select the eQTLs whose genotypes are the most correctly predictable to maximize *ICI* leakage. Although 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 individual, we compute the predictability of the eQTL genotypes as

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where denotes the predictability of given the gene expression level . 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 , 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), 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). We computed, for each eQTL, average 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 high frequency major allele frequencies have high predictability and low *ICI* and vice versa for eQTL variants with lower major allele frequency (Fig 2b, left). This is expected because the genotypes of the high frequency variants can be predicted, on average, easily (most individuals will harbor one dominant genotype) and consequently does not deliver much characterizing information. The genotypes for the eQTLs with smaller major frequency 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. In principle, the adversary will aim at identifying and using these highly informative eQTLs. 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 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 for top genotype predictions. For this, we first sorted the eQTLs with respect to the reported correlation, . Then for top *n=1,2,3,…,20* eQTLs, we estimated mean and mean *ICI* over all the samples as illustrated in Fig S2a. We then plotted mean versus mean *ICI* for each *n* which is shown in Fig 2c. From the plot, we can first estimate the number of vulnerable 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 individuals. At 5% predictability, the leakage is 11 bits and the characterizable sample size is 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 individuals, as explained earlier, it is necessary to predict bits of genotypic information correctly. The risk of characterization can be determined from the graph as the predictability level at which 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 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 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. The presented quantification procedure can be utilized for evaluating the risk of information leakage while releasing QTL datasets. For example, the QTLs to be released can be assessed in terms of the characterizing information leakage versus the predictability so as to estimate the size and risk of a linking attack (Fig S8) that would be mediated by these QTLs.

## 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 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 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, . 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 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).

# AIM2: Substantiating an Outlier Attack

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 via eQTLs is not enough to regenerate the *a posteriori* distribution of genotypes given the expression levels.

The attacker can, however, utilize a priori knowledge about the relation between gene expression levels and genotypes and build the joint genotype-expression distributions using models with varying complexities and parameters (See Methods Section 4.8). Even though the genotype prediction with these models may not be very accurate, the attacker can utilize a large number of eQTLs to maximize the accuracy of linking (Detailed in the Background Section). We focus on a highly simplified model here. We will assume the attacker exploits the knowledge that the eQTL genotypes and expression levels are correlated such that the allelic effects on expression are additive and extremes of the gene expression levels (highest and smallest expression levels) are observed 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 . For the gene expression levels for eQTL, , of the individual’s expression level, , is defined as

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Extremity can be interpreted as a normalized rank, which is bounded between -0.5 and 0.5. Figure S4a shows the median absolute extremity distribution of all the gene expression levels among the individuals. The average median extremity is uniformly distributed among individuals. Figure S4b shows the median number of genes with minimum extremity. Almost half of the genes in each individual have higher than 0.3 extremity in the population. Also, around 1000 genes have higher than 0.45 absolute extremity. In other words, each individual harbors substantial number of genes whose expressions are at the extremes within the population. These can potentially serve as quasi-identifiers. It is worth noting, however, that not all of these extreme genes are associated with eQTLs (See Sections S1 and S6, Figure S7).

Following from the above discussion, the adversary builds the posterior distribution for eQTL genotypes as

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | | | (4) |
|  |  | | |  |
|  | |  | | (5) |
|  | |  | |  |
|  |  | | (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 observed at medium levels of expression. The posterior distribution of genotypes in equations (4-6) can be derived from a simplified model of the genotype-expression distribution that utilizes just one parameter (See Methods Section 4.8, Fig S9). As yet another way 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 absolute values of the rank correlation. Thus, this process can be generalized as a rank correlation based prediction. We used the posterior genotype probabilities in 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. The slight decrease of genotype accuracy at correlation thresholds higher than 0.7 is caused by the fact that the accuracy (fraction of correct genotype predictions within all genotypes) is not robust at very small number of SNPs. Although we expect very high accuracy, even one wrong prediction among small number of total genotypes decreases the accuracy significantly.

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 4.5 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 and 5d show 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 eQTL52 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 also evaluated the accuracy of linking attack in comparison to the linking attack proposed by Schadt et al (See Section S3, Table S2). We observed that the two methods have comparable and very high accuracy, while extremity based linking attack uses much less input information compared to Schadt et al linking attack.

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 (See Methods Section 4.6), serves as a good reliability estimate for each linking. For a given linking, 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 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 threshold. For this, we first computed for each linking, then filtered the linkings that did not satisfy the threshold. Then we computed PPV and sensitivity of the linkings (See Methods Section 4.9), 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.

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 genotype dataset of 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 distance gap) 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%. Only the remaining 15% are predicted with accuracy lower 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 S1a 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 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 discovery and testing populations are diversified. 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 Project41. We downloaded the eQTLs for 6 tissues and performed the linking attack on the whole GEUVADIS samples as test samples. The results are shown in Table S1b. The accuracy is general high (>80%) and is highest for Whole Blood eQTLs, which is 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%. It is worth noting that the decrease in the accuracies stem also from the differences in data handling and processing between GEUVADIS and GTex projects.

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) available from HAPMAP project53,54. We first identified the eQTLs from the 90 individuals and performed linking over the same individuals. We then computed the average rank of the close relatives in each linking. For example, when the tested individual is a father or mother, we computed the rank of his/her 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, a random individual and computed his/her rank in the linking. The distribution of the ranks are shown in Fig 8. 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 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 might extend the reach of privacy breach and cause privacy concerns for the family.

# AIM3: Building Privacy Reducing File Formats