~~Analysis of~~ Information Leakage in Phenotype and Genotype Datasets

, PRAC LINKING ATTACKS ON RNASEQ DATABSE

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

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

Deleted: mark.gerstein@gersteinlab.org

[[Fix \cite's]]

[[Fix Figure/Section References]]

[[Use past tense, all around]]

[[At most 10 equations in the online methods?]]

[[Methods only references]]?

[[Number of references? 30 now, too low?]]

[[Figure Captions]]

[[Read one more time]]

[[TIFF figures: http://www.jimmunol.org/site/pdf/PPTtoTIFF.pdf, RGB, 300 DPI]]

Formatted: Default Paragraph Font, Font: 11 pt

[[Clarify MAP prediction in the figures]]

ABSTRACT

MANY HIGH D

Privacy is receiving much attention with the increase in the breadth and depth of personalized biomedical datasets. Studies on genomic privacy are mainly focused on protection of variants. Molecular phenotype datasets 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. The links can then be used to characterize individuals' sensitive phenotypes. Here, we first develop a formalism for the quantification 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 for individual characterization. Then we present a general three step procedure that can be used to practically instantiate an accurate attack. We develop a particular realization of the attack for outlier cases and study different aspects of the attack.

LINKING ?

1 INTRODUCTION

Genomics has recently emerged as one of the major foci of studies on privacy. This can be attributed to high throughput biomedical data acquisition that brings about a surge of datasets[1–3]. Among these, molecular phenotype datasets, like functional genomics measurements, substantially grow the list of the quasi-identifiers[4] which may lead to re-identification and characterization[4–6]. In general, statistical analysis methods are used to discover genotype-phenotype correlations[7,8], which can be utilized by an

SPELL OUT

adversary for linking the entries in genotype and phenotype datasets, and revealing sensitive information. The availability of a large number of correlations increases the possibility of linking[9,10].

Along with the initial genotype-phenotype association studies, the protection of privacy of participating individuals emerged as an important issue. Several studies addressed the problem of detecting whether an individual, with known genotype, has participated in a study[11]. As study participants choose to remain anonymous, the detection of an individual causes privacy concern[12–15] by revealing their existence in the study cohort. We refer to these systematic breaches as "detection of a genome in a mixture" attacks (**Supplementary Fig. 1**). However, as the number and size of phenotype and genotype datasets increase, the detection of individuals in these datasets will be irrelevant since any individual will already have their genotype or phenotype information stored in a dataset, i.e., participation will already be known. This opens up a new route to breaching privacy: An adversary can now aim at pinpointing an individual among multiple, seemingly independent, genotype and phenotype datasets by linking the entries in these datasets. It is most certain that as personalized genomics gain more prominence, e.g. large genotype and phenotype datasets are used in medicine, the attackers will focus on gaining access to these datasets, then aim at linking different datasets in order to reveal sensitive information. We will refer to these attacks as "linking attacks"[4,5]. One well-known example of these attacks is the attack that matched the entries in Netflix Prize Database and the Internet Movie Database (IMDB)[16], and revealed sensitive personal information. For research purposes, Netflix released an anonymized dataset of movie ratings of thousands of viewers, which was assumed to be secure as the viewer's names were removed. However, Narayanan et al used the Internet Movie Database (IMDb), a seemingly unrelated and very large database of movie viewers, linked two databases, and revealed identities and personal information of many viewers in the Netflix database. This attack is underpinned by the fact that both Netflix and IMDb host millions of individuals and any individual who is in one dataset is very likely to be in the other dataset. As the size and number of the genotype and phenotype datasets increase, number of potentially linkable datasets will increase, which can render similar scenarios a reality in genomic privacy (Supplementary Note).

Deleted: As

Deleted: Consequently, an

Deleted: then

Deleted: As

Deleted: that can

ALREADY EXPOSED ON IMDB BUT NOT WANT VIEWING HABIT...

Deleted: Different aspects of genomic privacy, pertaining linkability of high dimensional phenotype datasets to genotypes, are yet to be explored.

CHARACTERIZING

2 RESULTS

2.1 Linking Attack Scenario

In the linking attacks, the attacker aims at characterizing sensitive information about a set of individuals in a stolen genotype dataset (**Fig. 1**). For each individual in the genotype dataset, she aims at querying the publicly available anonymized phenotype datasets in order to characterize their sensitive phenotypes. For this, she first utilizes a public quantitative trait loci (QTL) dataset that contains phenotype-genotype correlations. She statistically predicts genotypes using the phenotypes and QTLs. Then she compares the predicted genotypes to the genotype dataset and links the entries that have good genotype concordance. Consequently, the sensitive information for the linked individuals in genotype dataset is revealed to the attacker.

Among the QTL datasets, the abundance of eQTL datasets makes them most suitable for linking attacks. In an eQTL dataset, each entry contains a gene, a variant, and correlation coefficient, denoted by ρ, between the expression levels and genotypes. We assume that the attacker aims to build a genotype prediction model that utilizes the relation between expression levels and genotypes (**Fig. 2a**, **Supplementary Fig. 2**). As a representative dataset for reporting results and for performing mock linking attacks, we use the eQTLs and gene expression levels from the GEUVADIS project[17], and the genotypes from the 1000 Genomes Project[18].

2.2 Genotype Predictability and Information Leakage

We assume that the attacker will behave in a way that maximizes her chances of correctly characterizing the most number of individuals. Thus, she will try and predict the genotypes, using the phenotype measurements, for the largest set of variants that she believes 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 then predicting the genotypes for each variant (**Supplementary Fig. 3**). The attacker will encounter a tradeoff: As she goes down the list, more individuals can be characterized (more genotypes can characterize more individuals) but it also becomes more likely that she makes an error in the prediction since the correlation decreases going down the list. 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). We will propose two measures, cumulative individual characterizing information (*ICI*) and genotype predictability (π), to study this tradeoff.

Deleted: his or

ICI can be interpreted as the total amount of information in a set of variant genotypes that can be used to pinpoint an individual in a linking attack. This quantity depends on the joint frequency of the variant genotypes. For example, if the set contains many common genotypes, they will not be very useful for pinpointing individuals. On the other hand, rare variant genotypes would give much information for linking. Thus, the information content of a set of genotypes is inversely proportional to the joint frequency of the variant genotypes. We utilize this property to quantify *ICI* in terms of genotype frequencies (Online Methods, **Fig 3**). In order to estimate the joint frequency of variant genotypes, we assume independence of variant genotypes (Online Methods, Supplementary Note).

For a set of variants, π measures how predictable genotypes are given the gene expression levels. Since genotypes and expression levels are correlated, knowledge of the expression enables one to predict the genotype more accurately than predicting genotype with no knowledge. In order to quantify the predictability, we use an information theoretic measure for randomness left in genotypes, given gene expression levels (Online Methods, **Fig. 3**). This has several advantages over using reported correlation coefficients for each eQTL for quantifying predictability. Although the correlation coefficient is a measure of predictability, it is computed differently in different studies and there is no easy way to combine and interpret the correlation coefficients when we would like to estimate the joint predictability of multiple eQTL genotypes. On the other hand, joint predictability of multiple eQTL genotypes given gene expression levels can be easily performed using π as it fits naturally to the information theoretic formulations (Online Methods).

Deleted: Although the reported

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. We computed, for each eQTL, average predictability and average *ICI* over all the individuals (**Fig. 4a**).

Deleted: π

ALSO BEYOND LINKING

Most of the data points are spread along the anti-diagonal: The eQTL variants with high major allele frequencies have high predictability and low *ICI* and vice versa for eQTL variants with lower major allele frequencies (**Fig. 4b**). 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 and vice versa for the eQTLs with smaller major frequency alleles. The predictability versus *ICI* leakage for the eQTLs in the shuffled eQTL dataset (Online Methods) is dominantly on the anti-diagonal. This is also expected as the gene expression levels in the shuffled eQTL dataset do not supply any information for predicting variant genotypes. Thus the predictabilities for shuffled eQTL genotypes depend mainly on how frequent they are in the population (major frequency genotypes are much easier to predict but have low *ICI* and vice versa), as explained above. On the other hand, the real eQTLs (**Fig. 4b**) deviate from the anti-diagonal, compared to shuffled eQTLs, which reflects the fact that expression supplies much information for predicting eQTL genotypes (**Fig. 4c**). The eQTLs with high correlation have substantially ~~high ICI and high predictability~~.

IS THIS THE PROOF OF CSTT

Deleted: The eQTLs with high absolute correlation (**Fig. 4b**) deviate from the anti-diagonal, compared to shuffled data (**Fig. 4c**). These eQTLs have

When multiple genotypes are utilized, the information leakage is greatly increased. To study this, we computed *ICI* (in bits) and predictability for increasing number of eQTLs (Supplementary Note, **Fig. 4d**). As expected, the predictability decreases with increasing *ICI* leakage. Inspection of mean predictability versus mean cumulative *ICI* enables us to estimate the number of vulnerable individuals at different predictability levels. For example, at 20% predictability, there is approximately 8 bits of cumulative *ICI* leakage. At this level of leakage, the adversary can pinpoint an individual, with 20% accuracy, within a sample of $2^8 = 256$ individuals. Thus, within any sample of 256 individuals, we expect the attacker to be correctly link 256x20%=51 individuals. At 5% predictability, the leakage is 11 bits and the attacker can pinpoint an individual in a sample of $2^{11} = 2048$ individuals. This corresponds to approximately 100 individuals getting correctly linked (5% of 2048). Auxiliary information can be easily added into *ICI*. For example, gender information, which can be predicted with high accuracy from many molecular phenotype datasets brings 1 bit of additional auxiliary information to *ICI* (Supplementary Note).

Formatted: Font: Italic

2.3 Framework for Instantiation of Linking Attacks

We present a three step framework for practical instantiation of linking attacks (**Fig. 2b**). This framework can be used to perform mock linking attacks on datasets for evaluating whether they will be effective for risk assessment purposes. We use this framework to simulate mock attacks in the following sections for assessing their accuracies. The input is the phenotype measurements for an individual, who is being queried for a match to individuals in the genotype dataset (**Fig. 1**). In the first step, the attacker selects the QTLs, which will be used in linking. The selection of QTLs can be based on different criteria. As discussed earlier, the genotype predictability (π) is the most suitable QTL selection criterion. Although the attacker cannot practically compute predictability using only the QTL list, any function of predictability would still be useful to the attacker for selecting QTLs. For example, the most accessible criterion is selection based on the absolute strength of association, $|\rho|$, between the phenotypes and genotypes. The second step is genotype prediction for the selected QTLs using a prediction model. The third and final step of a linking attack 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 (Online Methods).

2.4 Individual Characterization by Linking Attacks

Using the three step approach, we first evaluated the accuracy of linking when the attacker utilizes genotype prediction where the attacker builds the posterior distribution of genotypes given the expression levels (Supplementary Note) where the attacker knows the exact joint distribution of genotypes and expression. This sets a baseline reference for comparison of linking accuracy, where we assume the attacker knows the exact joint distribution. The attacker builds the posterior distribution of genotypes given expression levels from the joint distribution. Finally, she predicts each genotype by selecting the genotype with maximum *a posteriori* probability given gene expression level (Supplementary Note, **Supplementary Fig. 4**). For several eQTL selections with changing correlation threshold, the linking accuracy is above 95% and gets close to 100% when auxiliary information is available (**Fig. 5a**).

awk -

In general, knowledge or correct reconstruction of the exact joint genotype expression distribution may not be possible because the the genotype-phenotype correlation coefficient alone is not sufficient to perfectly reconstruct the genotype distribution 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 (Online Methods, Supplementary Note, **Supplementary Fig. 5**). We focus on a highly simplified model where the attacker exploits the knowledge that the eQTL genotypes and expression levels are correlated such that the extremes of the gene expression levels (highest and smallest expression levels) are observed with extremes of the genotypes (homozygous genotypes). We use a measure, termed extremity, to quantify the outlierness of expression levels (Online Methods, Supplementary Note, **Supplementary Fig. 6, 7**). Based on the extremity of expression level and the gradient of association, the attacker first builds an estimate of the joint genotype-expression distribution, then constructs the posterior distribution of genotypes and finally chooses the genotypes with maximum *a posterior* probability (Online Methods, Supplementary Note, **Fig. 2a, b**).

The extremity based prediction methodology assigns zero probability to heterozygous genotype, and assigns only homozygous genotypes to variants, for which the associated gene's expression level has absolute extremity higher than a threshold. We performed linking attack using this prediction method (in 2nd step of linking). In the 1st step of the attack, we used absolute correlation and extremity thresholds for eQTL selection. The linking accuracy is higher than 95% for much of the eQTL selections (**Fig 2a**, **Supplementary Fig. 6d**). We also observed that changing extremity threshold does not affect the linking accuracy substantially compared to changing absolute correlation threshold. We thus focus on attack scenarios where the absolute extremity threshold is set to zero. This also simplifies the attack scenario by removing one parameter from genotype prediction. With this approach, the genotype prediction accuracy increases with increasing absolute correlation threshold (**Supplementary Fig. 6c**). We performed linking attack with this model where we used the correlation based eQTL selection in step 1, then extremity based genotype prediction in step 2. In the step 3, we evaluated two distance measures for linking the predicted genotypes to the individuals in genotype dataset (Online Methods,

Deleted: , as expected

Deleted: next

Supplementary Fig. 8). More than 95% of the individuals (**Fig. 5c, d**) are vulnerable for most of the parameter selections. When the auxiliary information is present, the fraction of vulnerable individuals increases to 100% for most of the eQTL selections. We also observed that the extremity attack may link close relatives to each other, which can create potential privacy concerns for the family (**Supplementary Fig. 9d**). These results show that linking attack with extremity based genotype prediction, although technically simple, can be extremely effective in characterizing individuals.

CMPTO
MAP
BASELINE

We evaluated whether the attacker can estimate the reliability of the linkings so as to focus on highly reliable linkings. We observed that the measure we termed, *first distance gap*, denoted by $d_{1,2}$ (Online Methods), serves as a good reliability estimate for each linking. We computed the positive predictive value (PPV) versus sensitivity of the linkings in the testing set with changing $d_{1,2}$ threshold (Online Methods). Compared to random sortings, the attacker can link a large fraction (79%) of the individuals at a PPV higher than 95% (**Fig. 5d**, **Supplementary Fig. 9** CX). We also studied several biases that can affect linking accuracy. First when the eQTLs are discovered on a sample set that and the linking attack is performed on another sample set, the accuracies are still very high (Supplementary Note, **Supplementary Fig. 9a**). Moreover, attacks are accurate when there is mismatch between the tissue or population of eQTL discovery sample set and tissue or population of linking attack sample set (Supplementary Note, **Supplementary Table 1a, b**). In addition, we observed that the extremity attack is still effective when genotype sample size is very large (Supplementary Note, **Supplementary Fig. 9b, c**), which points out the applicability on large sample sizes.

3 DISCUSSION

In genomic privacy, it is necessary to consider the basic premise of sharing any type of information: There is always an amount of sensitive information leakage in every released dataset[19]. It is therefore essential for the genomic data sharing and publishing mechanisms to incorporate statistical quantification methods to objectively quantify risk estimates before the datasets are released. The quantification methodology and the analysis frameworks presented here and in future studies can be used for analysis of the information leakage where the correlative relations between datasets can be exploited for performing linking attacks (Supplementary Note, **Supplementary Fig. 10**).

In the context of linking attacks, an individual's existence in two seemingly independent databases (e.g., phenotype and the genotype) can cause a privacy concern when the attacker statistically links the databases using the a priori information about correlation of different entries in the phenotype and genotype databases. The methods that we proposed can be integrated directly into the existing risk assessment and risk management strategies. One such approach is k-anonymization and its extensions[20–22]. This technique performs anonymization of the datasets by ensuring that no combination of the features (e.g., predicted genotypes) can be used to pinpoint an individual to less than *k* individuals. This is done by censoring entries in the dataset or noise addition into the dataset. The estimates of genotype predictability and *ICI* leakages can be used to select which entries in the phenotype dataset should be anonymized so as to achieve anonymity. Another approach is to serve phenotypic data from a statistical database. In this context, differential privacy has been proposed as an

EFF of ABSOLUTE ?

Moved (insertion) [1]

Deleted: 9c, d

Moved up [1]: We also observed that the extremity attack may link close relatives to each other, which can create potential privacy concerns for the family (**Supplementary Fig.**

Deleted: 10).

Deleted: personal

Deleted: leakage in the

Deleted: 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

Deleted: applied

Deleted: in the datasets

Deleted: 11

Deleted: Our study focuses on the individual privacy breaches in the context of linking attacks, where an individual's existence in two seemingly independent databases (e.g., phenotype and the genotype) can cause a privacy concern when an attacker links statistically the databases using the a priori information about correlation of different entries in the databases. The obvious risk management strategy against these attacks is restricting access to the phenotype datasets. This approach has, however, high cost in terms of lost research opportunities. Another approach is serving encrypted data, where data analysis is performed directly on the encrypted data, for example, using homomorphic encryption[20]. This approach has very high compute requirements and not practical yet. One other approach is to utilize statistical techniques like k-anonymization[21]. These can be employed on the phenotype datasets before being published. For this, it is necessary to develop new approaches and heuristics that can effectively circumvent high computational requirements[22]. In addition, several other approaches have addressed scenarios where k-anonymization may fail to protect data[23,24]. These scenarios must be properly handled in the risk management strategies. The anonymization strategies can use the estimates of leakage and predictability from our study to determine the QTLs that cause most leakage and anonymize the phenotype data accordingly. Another approach is to serve phenotypic data from a statistical database. In this context, differential privacy has been proposed as an optimal way for privacy aware data serving from statistical databases. The data release mechanisms in a differentially private scenario can benefit from the estimates of *ICI* leakage in each QTL. Differentially private data serving may, however, decrease the biological utility of the data significantly[25]. We believe new studies should address protection and risk management strategies for serving utility maximized and privacy aware high dimensional phenotype datasets.¶
DATASETS¶

optimal way for privacy aware data serving from statistical databases[23]. The data release mechanisms, which are used in differentially private databases for serving data, can make use of the estimates of *ICI* leakage in each QTL so that the total leakage can be tracked. It is worth noting that anonymized data publishing and serving mechanisms may decrease the biological utility of the data significantly[24]. Thus, it is necessary to integrate the measures of biological utility of the anonymized datasets as another quantity in addition to predictability and *ICI* leakage in risk assessment.

Moved down [2]: The normalized gene expression levels for 462 individuals and the eQTL dataset are obtained from GEUVADIS mRNA sequencing project[17]. The eQTL dataset contains all the significant (Identified at most 5% false discovery rate) 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 (Section S4). The

Moved down [3]: The genotype, gender, and population information datasets for 1092 individuals are obtained from 1000 Genomes Project[18]. For 421 individuals, both the genotype data and gene expression levels are available. For tissue analysis, the publicly available significant eQTLs for 6 tissues that are computed by the GTex project are downloaded from the GTex Portal.¶

4 ACKNOWLEDGEMENTS

Authors would like to thank Akdes Serin Harmanci for constructive comments and discussions on study design and running of external tools. Authors also would like to thank the anonymous reviewers for constructive criticism that made the manuscript complete and comprehensive.

5 AUTHOR CONTRIBUTIONS

A.H. designed the study, gathered datasets, performed experiments, and drafted the manuscript. M.G. conceived the study, oversaw the experiments and wrote the manuscript. Both authors approved final manuscript.

Authors declare no conflict of financial interest.

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

4. 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

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

6. 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

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

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

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

10. 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).

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

12. 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).

13. Im, H. K., Gamazon, E. R., Nicolae, D. L. & Cox, N. J. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet.* **90,** 591–598 (2012).

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

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

16. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. in *Proc. - IEEE Symp. Secur. Priv.* 111–125 (2008). doi:10.1109/SP.2008.33

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

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

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

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

21. 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

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

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

24. 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>

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

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

7 FIGURE LEGENDS

Figure 1: Illustration of the linking attack. The publicly available anonymized phenotype dataset contains q phenotype measurements and the HIV Status for a list of n individuals. Genotype dataset contains the variant genotypes for m individuals. Genotype-phenotype correlation dataset contains q phenotypes, variants, and their correlations. The attacker performs genotype prediction for all the variants. The attacker then links the phenotype dataset to the genotype dataset by matching the genotypes. The linking potentially reveals the HIV status for the subjects in the genotype dataset. The IDs and HIV Status are colored to illustrate how the linking combines the entries in the two datasets. The grey-shaded columns are not used for linking.

Figure 2: Illustration of genotype-expression associations and linking attacks (a) Schematic representation of genotype and expression association and simplifications for an eQTL. The trimodal gene expression distribution and the joint genotype-expression distribution are shown. The conditional distribution of expression given each genotype is illustrated with box plots in different colors corresponding to each genotype. The genotypes and expression levels are correlated (ρ) as indicated by

Deleted: Vaikuntanathan, V. Computing Blindfolded: New Developments in Fully Homomorphic Encryption. *2011 IEEE 52nd Annu. Symp. Found. Comput. Sci.* 5–16 (2011). doi:10.1109/FOCS.2011.98

Deleted: 21. SWEENEY, L. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* **10,** 557–570 (2002).¶
¶
22. 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¶
¶
23.

Deleted: 24

Deleted: 25

Deleted: 26

Deleted: 27

Deleted: (a) Phenotype

Deleted: different

Deleted: variants

Deleted: Phenotype-

Deleted: datasets

Deleted: does

Deleted: genotypes

Deleted: non

Deleted: associations. Genotype (y-axis) and expression (x-axis) are correlated, indicated by line fit and ρ. The rectangles represent conditional distribution of expression given genotype values. (b) Illustration of extremity based genotype prediction. Expression range is divided into two equal ranges (separated by e_{mid}). The blue rectangles represent the distribution used for prediction. Given the distribution of expression (tri-model distribution on right), the positive extremity is assigned genotype 2, and negative extremity is assigned genotype 0. (c) Three the line fit. The extremity based genotype prediction models the joint genotype-expression as a simplified distribution. When the genotype value is 0, uniform probability is assigned for expression values where extremity is smaller than δ (Green rectangle). For genotype value 1, no probability is assigned. When genotype value is 2, the probability is uniformly distributed over expression values for which extremity is greater than δ (Purple rectangle). Simplified extremity based prediction utilizes the same distribution by setting δ to 0. In this case, when genotype is 0, joint probability is distributed uniformly over expression levels with negative extremity. When genotype is 2, uniform probability is assigned to expression levels with positive extremity. (b) Illustration of the three step linking process. First step is selection of phenotypes and genotypes to be used in linking. Second step is prediction of genotypes. Last step is linking of predicted genotypes to the genotype dataset.

Figure 3: Illustration of individual characterizing information (*ICI*) and correct predictability of genotypes. *ICI* for a set of n variant genotypes is computed in terms of population genotype frequencies. Each genotype contributes to *ICI* additively with the logarithm of reciprocal of the genotype frequency (illustrated by the genotype distributions). Given an eQTL where genotype of variant V_1 is correlated to expression of gene 1 (E_1), the predictability of genotype given expression level is e is computed in terms of exponential of the entropy of conditional genotype distribution, given expression level *e*. While computing the entropy of genotypes, the conditional distribution is built by slicing the joint distribution at expression level *e*. The entropy of the conditional distribution is then used for predictability. The genotype frequencies for *ICI* computation can also be computed from the joint genotype-expression distribution by marginalizing over expression levels.

Figure 4: Estimates of *ICI* leakage versus predictability. 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. The dots are colored with respect to the major allele frequency (a) and with respect to absolute correlation of the eQTL (b), and real versus shuffled eQTL dataset (c). The average cumulative *ICI* leakage versus joint genotype predictability is shown (d) when multiple eQTLs are utilized with shuffled eQTL dataset. The arrows on the plot indicates the increasing number of eQTLs used in estimated joint predictability and cumulative *ICI* leakage.

Figure 5: Accuracy of linking attacks. (a) Accuracy of linking with exact joint distribution based genotype predictions. Absolute correlation threshold (x-axis) versus fraction of vulnerable individuals (y-axis) is plotted. The yellow arrow indicates the maximized position of linking accuracy. Red, green, and cyan plots show linking accuracy with gender, population, and gender and population as auxiliary information, respectively. (b) Linking accuracy with extremity based linking with all genotypes. (c) Linking accuracy with extremity based linking with homozygous genotypes. (d) Sensitivity versus positive predictive value of linkings chosen with changing d_1,2 threshold in comparison to the random selections of linkings.

Supplementary Figure 1: Schematic comparison of linking attacks (a) and detection of a genome in a mixture attacks (b). Each box in the figure represents a dataset in the form of a matrix. Multiple boxes

Formatted: Font: Italic

Deleted: (a) Graphical representation of ICI formulation.

Formatted: Font: Italic

Formatted: Font: Italic

Deleted: (b) Graphical representation of π.

Deleted: the joint distribution of

Deleted: and

Deleted: (shown below), the conditional distribution of genotypes

Deleted: e

Formatted: Font: Italic

Deleted: . The

Deleted: conditional distribution

Deleted: is used for

Deleted: predictability

Deleted: π for each eQTL

Formatted: Font: Italic

Deleted:). ICI versus π for shuffled data (red) is compared to the real dataset (blue) in (c).

Deleted: measures for

Deleted: Linking accuracy

Deleted: MAP

Deleted:).

Deleted: +

Deleted: .

Deleted: The genotype prediction

Deleted: . The genotype prediction accuracy (y-axis) of

Deleted: changing absolute correlation threshold (x-axis).

Deleted: all genotypes. (d) Linking accuracy with extremity based linking with

Deleted: ***[[FOLLOWING ARE NOT UPDATED, YET]]¶***
Table S1

Moved down [4]: ***:*** Linking accuracy of extremity based linking attack using the eQTLs are identified in different populations and different tissues. (a) The table shows the linking accuracies (for populations shown in the rows) when the eQTLs that are identified using data (indicated in each column) from different populations. (b) The linking accuracy of individuals in GEUVADIS project when eQTLs identified from different tissues are used in linking.

Deleted: ¶
Table S2:

Moved down [5]: Linking attack accuracy comparison. The table shows linking accuracy for Schadt et al and extremity based linking attack methods. Each row corresponds (for Schadt et al Method) to a different number of data points in the training datasets that is input to Schadt et al method.¶

Deleted: Left

Deleted: Right

next to each other correspond to concatenation of matrices. Linking attacks aim at linking genotype and phenotype datasets. The phenotype datasets contain both “predicting” phenotypes and other phenotypes, some of which can be sensitive. The attacker first predict genotypes for each of the predicting phenotype. The predicted genotypes are then compared with the genotypes in the genotype dataset. After the linking, all the datasets are concatenated where the identifiers can be matched to the sensitive phenotypes. Different colors indicate how the linking merges different information. The detection of a genome in a mixture attacks start with a genotype dataset. The attacker gets access to the statistics of a GWAS or genotyping dataset (for example, regression coefficients or allele frequencies). Then the attacker generates a statistic and tests it against that of a reference population. The testing result can be converted into the study membership indicator (attended/not attended) which shows whether the tested individual was in the study cohort or not.

Supplementary Figure 2: Illustration of the expression and genotype datasets. Variant genotype dataset contains the genotypes for *q* eQTL variants for n_v individuals. jth entry for kth eQTL is denoted by $v_{k,j}$. Similarly, the expression dataset contains the expression levels for *q* genes. The kth expression level for jth individual is denoted by $e_{k,j}$. The variant genotypes for kth variant is distributed over samples with distribution specified by the random variable V_k. Likewise, the expression levels for kth gene is distributed per random variable E_k. These random variables are correlated with each other with correlation coefficient, denoted by $\rho(E_k, V_k)$ (right).

Deleted: ***Supplementary Figure 2:*** Representation 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.¶
¶
¶

Supplementary Figure 3: Figure shows the attacker’s presumed strategy for linking attack. (a),(b) 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. (c) The average joint predictability of genotypes versus the average cumulative *ICI* leakage for multiple eQTLs. The error bars (one standard deviation) for *ICI* and predictability are shown on the real eQTLs.

Deleted: (b

Supplementary Figure 4: (a) Illustration of prior, joint, and posterior distributions 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 (MAP prediction) at genotype 2, which is indicated by a star. (b) The number of selected and average correctly predicted eQTL genotypes with changing absolute correlation threshold. The error bars (one standard deviation) are shown for correctly predicted eQTL genotypes.

Supplementary Figure 5: Models of joint genotype-expression distribution with varying numbers of parameters for a positively correlated eQTL. (a) The true genotype-expression distribution. Grey boxes represent the expression distributions given different genotypes. Red line show the gradient of correlation between genotype and expression. (b) First simplification of the joint distribution. The expression distribution can be modeled with Gaussians with different means and variances with total of 6 parameters. (c) Simplification of joint distribution with equal variances. The variances can be assumed same for different genotypes, where 4 parameters are required. (d) A representation of the uniform

Deleted: ***Supplementary Figure 3:*** 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.¶
Supplementary Figure 4

Moved (insertion) [6]

Moved (insertion) [7]

Moved (insertion) [8]

expression distribution given genotypes, where 4 parameters are required. The conditional distribution of expression is uniform (cross shaded rectangles) over the ranges (e_1, e_2), (e_2, e_3), and (e_3, e_4) given genotypes 0, 1, and 2, respectively. The transparent grey rectangles show the original distributions. (e) A simplification of (d) where no conditional probability of expression is assigned given genotype is 1. In this model, only one parameter (e_{mid}) is necessary. The conditional probability of expression given genotypes 0 and 2 are uniform for expression levels below e_{mid} and above e_{mid}, respectively (shown with shaded rectangles). The original distribution is shown with grey rectangles for comparison. Extremity based prediction is an instantiation of the model in (e).

Supplementary Figure 6; The median absolute gene expression extremity statistics over 462 individuals in GEUVADIS dataset. (a) For each individual, the extremity is computed over all the genes (23,662 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 midpoint between minimum and maximum values of absolute extremity. (b) For each individual, we counted the number of genes above the extremity threshold. The plot shows the extremity threshold versus the median number of genes (over 462 individuals) above the extremity threshold. Around half of the genes (indicated by dashed yellow lines) have higher than almost 0.3 extremity on average over all the individuals. Also, around median number of 1000 genes over the samples have higher than 0.45 extremity (indicated by dashed red lines). (c) Accuracy of extremity based genotype prediction with changing absolute correlation threshold. (d) The linking accuracy with changing absolute extremity (x-axis) and absolute correlation thresholds (y-axis).

Supplementary Figure 7: A representative example of extremity based linking. The phenotype dataset (Consisting of gene expression levels for 6 genes) is shown above. Each phenotype measurement is represented by blue (negative extreme), yellow (positive extreme), or grey (non-extreme) dots. Based on the extremity of phenotypes, the attacker performs prediction of genotypes, which are shown below in (2). She uses the eQTL dataset (with genes and SNPs) for prediction. Blue and brown triangles correspond to the correct genotype predictions. The grey crosses correspond to the incorrect or unavailable genotype predictions. The attacker compares the predicted genotypes to the genotype dataset in (3), where triangles show the genotypes, and performs linking. The attacker links the predicted genotypes to the genotype dataset. 3 individuals (Bob, Alice, and John) are highlighted. The attacker can link Bob and John by matching them to their genotypes. The correct prediction of rs7274244 (in yellow dashed rectangle) enables the attacker to distinguish between correct entries and reveal both of their disease status as positive. For Alice, the predicted genotypes are equally matching at two entries both of which match at 2 genotypes; PID-b and PID-k (with negative and positive disease status) thus the attacker cannot exactly reveal Alice's disease status.

Supplementary Figure 8; Illustration of linking for jth individual. The attacker first predicts the genotypes $(\tilde{\boldsymbol{v}}_{\cdot,j})$ which are then used to compute the distance to all the individuals in the genotype dataset. The computed distances are then sorted in decreasing. The top matching individual (in the

Moved (insertion) [9]

Moved (insertion) [10]

Moved (insertion) [11]

Formatted: Font: Bold, Italic

Moved (insertion) [12]

Moved up [11]: ***:*** The median absolute gene expression extremity statistics over 462 individuals in GEUVADIS dataset. (a) For each individual, the extremity is computed over all the genes (23,662 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 midpoint between minimum and maximum values of absolute extremity.

Deleted: (b) For each individual, we count

Moved up [12]: the number of genes above the extremity threshold. The plot shows the extremity threshold versus the median number of genes (over 462 individuals) above the extremity threshold. Around half of the genes (indicated by dashed yellow lines) have higher than almost 0.3 extremity on average over all the individuals. Also, around median number of 1000 genes over the samples have higher than 0.45 extremity (indicated by dashed red lines).

Deleted: ¶
Supplementary Figure 5

Moved down [13]: ***:*** Illustration of linking for jth individual. The attacker first predicts the genotypes $(\tilde{\boldsymbol{v}}_{\cdot,j})$ which are then used to compute the distance to all the individuals in the genotype dataset. The computed distances are then sorted in decreasing. The top matching individual (in the example, individual a) is assigned as the linked individual. The first distance gap, $d_{1,2}$, is computed as the difference between the second $(d_{j,(2)})$ and the first $(d_{j,(1)})$ distances in the sorted list.¶

Formatted: Font: Bold, Italic

Deleted: ***Supplementary Figure 6:***

Formatted: Font: Not Bold, Not Italic

Deleted: ***Supplementary Figure 8***

Moved (insertion) [13]

example, individual a) is assigned as the linked individual. The first distance gap, $d_{1,2}$, is computed as the difference between the second ($d_{j,(2)}$) and the first ($d_{j,(1)}$) distances in the sorted list.

Supplementary Figure 9: The linking accuracy with different setups. (a) The accuracy of linking attack when the eQTLs are discovered on the training set and linking is performed on testing set. (b) The accuracy of linking when the simulated set of 100,000 individuals are used in the genotype dataset. (c) The positive predictive value (PPV) versus sensitivity with changing d_1,2 threshold for the eQTL selection in (b) where linking accuracy is around 70%, indicated by dashed yellow line. The grey dashed line marks the 95% PPV. (d) The distribution of ranks for close relatives (blue) and for random individuals (red) in the linking in 30 HAPMAP CEU trio dataset. Assigned rank is shown in x-axis and frequency is shown on y-axis.

Supplementary Figure 10: Illustration of risk assessment procedure for joint genotyping/phenotyping data generation. There are two paths of risk assessment to be performed. The first path evaluates the risks associated with release of the QTL datasets. The genotype and phenotype data (on the left) is first used for quantitative trait loci identification (QTL identification box). This generates the significant QTLs. These are then utilized, in addition to the list of external QTL databases, in quantification of leakage versus predictability, as presented in Section 2.2. These results are then relayed to the risk assessment procedures. The second risk assessment procedure evaluates the release of genotype and phenotype datasets. For this, the datasets are input to application of a list of linking attacks (Presented in Sections 2.3, and 2.4, and other linking attacks in the literature) for evaluation of characterization risks. The results are then relayed to risk assessment procedures.

Supplementary Table 1: Linking accuracy of extremity based linking attack using the eQTLs are identified in different populations and different tissues. (a) The table shows the linking accuracies (for populations shown in the rows) when the eQTLs that are identified using data (indicated in each column) from different populations. (b) The linking accuracy of individuals in GEUVADIS project when eQTLs identified from different tissues are used in linking. (c) Linking attack accuracy comparison. The table shows linking accuracy for Schadt et al and extremity based linking attack methods. Each row corresponds (for Schadt et al Method) to a different number of data points in the training datasets that is input to Schadt et al method.

8 ONLINE METHODS

8.1 Genotype, Expression, and eQTL Datasets

The eQTL, expression, and genotype datasets contain the information for linking attack (**Supplementary Fig. 2**). 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

Deleted: ***Supplementary Figure 9:*** Models of joint genotype-expression distribution with varying numbers of parameters for a positively correlated eQTL. (a) shows the true distribution where grey

Moved (insertion) [4]

Moved (insertion) [5]

Moved up [6]: boxes represent the expression distributions given different genotypes. Red line show the gradient of correlation between genotype and expression.

Moved up [8]: representation of the uniform expression distribution given genotypes, where 4 parameters are required. The conditional distribution of expression is uniform (cross shaded rectangles) over the ranges (e_1, e_2), (e_2, e_3), and (e_3, e_4) given genotypes 0, 1, and 2, respectively. The transparent grey rectangles

Moved up [9]: simplification of (d) where no conditional probability of expression is assigned given genotype is 1. In this model, only one parameter (e_{mid}) is necessary. The conditional probability of expression given genotypes 0 and 2 are uniform for expression levels below e_{mid} and above e_{mid}, respectively (shown with

Deleted: First simplification of the model is shown in (b).

Moved up [7]: The expression distribution can be modeled with Gaussians with different means and variances with total of 6 parameters.

Moved up [10]: with grey rectangles for comparison. Extremity based prediction is an instantiation of the model in (e).¶

Deleted: The variances can be assumed same for different genotypes (c), where 4 parameters are required. (d) illustrates a

Deleted: shows the original distributions. (e) is a

Deleted: cross shaded rectangles). The original distribution is included dataset and individuals in genotype dataset. The kth row of e, $\boldsymbol{e_k}$, contains the gene expression values for kth eQTL entry and $e_{k,j}$ represents the expression of the kth gene for jth individual. Similarly, kth row of v, $\boldsymbol{v_k}$, contains the genotypes for kth eQTL variant and $v_{k,j}$ represents the genotype ($v_{k,j}$ ϵ {0,1,2}) of k variant for jth individual. The coding of the genotypes from homozygous or heterozygous genotype categories to the numeric values are done according to the correlation dataset (Online Methods). We assume that the variant genotypes and gene expression levels for the kth 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.

8.2 Quantification of 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*[25] of the event that RV takes the value $g_{k,j}$:

Field Code Changed

Deleted: [26]

$$ICI(V_k = g_{k,j}) = I(V_k = g_{k,j}) = -\log_2(p(V_k = g_{k,j})) \quad (1)$$

where V_k is the RV that represents the kth 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:

$$\begin{aligned} ICI(\{V_1 = v_{1,j}, V_2 = v_{2,j}, \dots, V_N = v_{N,j}\}) \\ = -\sum_{k=1}^{N} \log_2\left(p(V_k = v_{k,j})\right). \end{aligned} \quad (2)$$

The genotype probabilities are estimated by the frequency of genotypes in the genotype dataset. We measure the predictability of eQTL genotypes using an entropy based measure. Finally, the base of logarithm that is used determines the units in which ICI is reported. When base two logarithm is used as above, the unit of *ICI* is bits.

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

$$\pi(V_k | E_k = e) = \exp(-H(V_k | E_k = e)) \quad (3)$$

Formatted Table

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 kth individual, given the expression levels $e_{k,j}$ for all the eQTLs, the total predictability is computed as

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

$$= \exp\left(-\sum_{k} H(V_k|E_k = e_{k,j})\right) \tag{4}$$

Formatted Table

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.

8.3 Extremity Based MAP Genotype Prediction

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's expression level, $e_{k,j}$, is defined as

$$ext(e_{k,j}) = \frac{\text{rank of } e_{k,j} \text{in } \{e_{k,1}, e_{k,2}, \dots, e_{k,n_e}\}}{n_e} - 0.5. \tag{5}$$

Extremity can be interpreted as a normalized rank, which is bounded between -0.5 and 0.5. The average median extremity is uniformly distributed among individuals (**Supplementary Fig. 6a**). In addition, around half of the genes (10,000) in each individual have higher than extremity value of 0.3. Also, around 1000 genes have higher than 0.45 absolute extremity (**Supplementary Fig. 6b**). 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.

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

$$P(V_k = 0 \mid E_k = e_{k,j}) = \begin{cases} 1 \text{ if } |ext(e_{k,j})| > \delta, ext(e_{k,j}) \times \rho(E_k, V_k) < 0 \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

$$P(V_k = 2 \mid E_k = e_{k,j}) = \begin{cases} 1 \text{ if} |ext(e_{k,j})| > \delta, ext(e_{k,j}) \times \rho(E_k, V_k) > 0 \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

$$P(V_k = 1 \mid E_k = e_{k,j}) = 0. \tag{8}$$

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. 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 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 (Online Methods). We used the posterior genotype probabilities in extremity based prediction and assessed the genotype prediction accuracy. As expected, the accuracy of genotype predictions increases with increasing correlation threshold (**Fig. 5c**). 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.

Deleted: 5b

8.4 First Distance Gap Statistic Computation

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, termed *first distance gap statistic*, as a measure of reliability of linking. For this, the attacker computes following difference:

$$d_{1,2}(j) = d_{j,(2)} - d_{j,(1)} \qquad (9)$$

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 (**Supplementary Fig. 8**). The basic motivation for this statistic comes from the observation that the first distance gap for correctly linked individuals are much higher compared to the incorrectly linked individuals.

8.5 eQTL Identification with Matrix eQTL

For identification of eQTLs, we used Matrix eQTL[26] 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. Matrix eQTL is run to identify the eQTLs using the training dataset. 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 (as reported by Matrix eQTL) 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. To accomplish this, we selected the eQTL that is correlated with highest association with each gene. The association statistic reported by Matrix eQTL was used as the measure of strength of association between expression levels and genotypes. Similar procedure is applied when eQTLs for 30 trios are identified.

Deleted: [27] method.

8.6 Modeling of Genotype-Phenotype Distribution

In the second step of the linking attack, the genotype predictions are performed. The genotype predictions are used, as an intermediate information, as input to the third step (**Fig. 2c**), where linking is performed. The main aim of attacker is to maximize the linking accuracy (not the genotype prediction accuracy), which depends jointly on the genotype prediction accuracy and the accuracy of the genotype matching in the 3rd step. Other than the accuracy of linking, another important consideration, for risk management purposes, is the amount of auxiliary input data (like training data for prediction model) that the genotype prediction takes. The prediction methods that require high amount of auxiliary data would decrease the applicability of the linking attack as the attacker would need to gather extra information before performing the attack. On the other hand, the prediction methods that require little or no auxiliary data makes the linking attack much more realistic and prevalent. It is therefore useful, in the risk management strategies, to study complexities of genotype prediction methods and evaluate how these translate into assessing the accuracy and applicability of the linking attack. We study different simplifications of genotype prediction, and illustrate different levels of complexity for genotype prediction.

In MAP based genotype prediction and linking attack, we assume that the attacker estimates the posterior distribution of genotypes and utilizes the maximum *a posteriori* estimate of the genotype as the general prediction method. For this, attacker must first model the joint genotype-phenotype distribution and then build the posterior genotype distribution (**Supplementary Fig. 5a**). The first level level of model can be built by decomposing the conditional distribution of expression with independent variances and means (**Supplementary Fig. 5b**). Assuming that mean and variance are sufficient statistics for the conditional distributions (e.g., normally distributed), the joint distributions can be modeled when the 6 parameters (3 means and 3 variances) are trained. The training can be performed using unsupervised methods like expectation maximization or can be performed using training data. This would, however, increase the required auxiliary data and decrease the applicability of the linking attack. A simplification of the model by assuming the variances of the conditional expression distributions are same for each genotype (**Supplementary Fig. 5c**). This decreases the number of parameters to be trained to 4 (3 means and 1 variance). An equally complex model with 4 parameters can be built assuming the conditional distributions are uniform at non-overlapping ranges of expression for each genotype (**Supplementary Fig. 5d**). This model requires 4 parameters to be trained corresponding to the expression range limits. Another simplification of the genotype prediction can be performed (**Supplementary Fig. 5e**), which requires only one parameter to be trained. In this model, the prediction only assigns uniform probability for homozygous genotypes when expression levels higher or lower than e_{mid} and assigns 0 conditional probability to the heterozygous genotypes, which brings up an important point: This simplified model is exactly the distribution that is utilized in the extremity based genotype prediction. In the extremity based prediction, we estimate e_{mid} simply as the mid-point of the range of gene expression levels within the expression dataset (Supplementary Note).

8.7 Datasets

The normalized gene expression levels for 462 individuals and the eQTL dataset are obtained from GEUVADIS mRNA sequencing project[17]. The eQTL dataset contains all the significant (Identified at most 5% false discovery rate) 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 (Section S4). The shuffled (randomized) eQTL datasets in comparisons are generated by shuffling the gene names in the gene-variant pairs in eQTL dataset. This way the gene and

Moved (insertion) [2]

variant matchings are randomized. The genotype, gender, and population information datasets for 1092 individuals are obtained from 1000 Genomes Project[18]. For 421 individuals, both the genotype data and gene expression levels are available. For tissue analysis, the publicly available significant eQTLs for 6 tissues that are computed by the GTex project are downloaded from the GTex Portal.

Moved (insertion) [3]

8.8 Code Availability

All the analysis code that is used to generate results can be obtained from http://privaseq.gersteinlab.org

9 METHODS ONLY REFERENCES