Quantification of Private Information Leakage from Phenotype-Genotype Data: Linking Attacks

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

General comments:

Word count should ideally be 3000 words and not exceed 3500 – I have made some suggestions, but further streamlining is required to hit 3500.

Editorial summary:

Linkage attacks can identify individuals by the presence of their seemingly independent data, such as molecular phenotypes and genotypes, in different databases and are a threat to privacy. The authors statistically quantify the extent of this risk and propose means to reduce it.

ABSTRACT

Studies on genomic privacy have traditionally focused on analysis of individual identification using DNA variants in personal genomes. Molecular phenotype data, such as gene expression levels, are generally assumed as free of sensitive information. Although there is no explicit genotypic information in them, an adversary can utilize the publicly available genotype-phenotype correlation databases to statistically link phenotypes to genotypes. This can be very accurate when many phenotypes are utilized and the resulting links can be used to reveal sensitive personal information. Here, we develop frameworks and tools for quantification of the individual characterizing information leakage from phenotype datasets. These can be used for estimating leakage from large datasets before being released. We also present a general three-step procedure for practically initiating linking attacks. We showcase an attack using outlier gene-expression levels that is simple yet accurate. We then study effectiveness of this attack under different scenarios.

Commented [A1]: We are currently counting less than 3,000 words.

Deleted:

Commented [NR2]: I realize that this is a complex topic but I ask you to present the question in a more generally accessible way in the abstract. As written it will not be easy to understand bythe non geneticists among our readers. Since this is a topic of concern to everybody it is important to present the abstract in a way that engages a wide range of readers. What input data are needed and what privacy breaches can be done with them?

Deleted: Privacy is receiving much attention with the increase ...

Deleted: protection of

Commented [NR3]: Not clear what you mean here, DNA ...

Commented [NR4]: You are con trasting DNA and RNA, ...

Deleted: However, m

Deleted: sets

Deleted:

Deleted: (e.g.

Deleted: , such as RNA-sequencing,

Deleted:)

Deleted: can also contain substantial amounts of sensitive ...

Commented [NR5]: Where does this information come from ...

Commented [NR6]: Where does this information come from ...

Deleted: subtle

Deleted: correlations

Deleted: high-dimensional data

Deleted: , and

Deleted: the

Deleted: then

Deleted: characterize sensitive phenotypes

Deleted: formalism

Commented [NR10]: In RNA-seq datasets? Where does this ...

Deleted: the

Deleted: leakage of

Deleted: and the tradeoff connection between the total amou ...

Deleted: Finally, w

Deleted: instantiating

Deleted: a particular realization of

Deleted: for

Deleted: ?

Commented [NR12]: It may be worth also mentioning that t ...

Commented [NR13]: Not sure what you mean by "applying ...

Deleted: present

Deleted: applicability

1 INTRODUCTION

Genomic privacy has recently emerged as an important issue, particularly in light of a surge in biomedical data acquisition [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 of individuals[4–6]. In general, statistical analysis methods are used to discover genotype-phenotype correlations[7,8], which can be utilized by an adversary for linking the entries in genotype and phenotype datasets, thereby revealing sensitive information. The availability of a large number of correlations increases the possibility of linking[9,10].

Protecting the privacy of participating individuals emerged as an important issue in genotype-phenotype association studies. Several studies addressed the problem of detecting whether an individual, with known genotype, has participated in a study[11] which raises privacy concerns[12–15]. 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 them 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 cross-referencing multiple, seemingly independent, genotype and phenotype datasets and pinpointing an individual to characterize her sensitive phenotypes. It is most certain that as personal genomics gains more prominence, the attackers will 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 is the attack that matched the entries in Netflix Prize Database and the Internet Movie Database[16]. For research purposes, Netflix released an anonymized dataset of movie ratings of thousands of viewers. This dataset was assumed to be secure as the viewer's names were removed. However, Narayanan et al used the Internet Movie Database, in which the identities of many users are public but only some of their movie choices are available, and linked it to the Netflix dataset. This revealed the identities and personal movie preference information of many users in the Netflix dataset. This attack is underpinned by the fact that both Netflix and the Internet Movie Database 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, the number of potentially linkable datasets will increase (Supplementary Note).

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, she aims at querying the publicly available anonymized phenotype datasets in order to characterize their HIV status. For this, she first utilizes a public quantitative trait loci (QTL) dataset that contains genotype-phenotype 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 is revealed to the attacker.

Deleted: Genomics

Deleted: one of the major foci of studies on privacy

Deleted: . This can be attributed to high throughput

Deleted: that brings about a surge of datasets

Deleted: Along with the initial genotype-phenotype association studies, p

Deleted: . As study participants choose to remain anonymous, the detection of an individual

Deleted: by revealing their existence in the study cohort

Commented [NR14]: It is likely that any person will have participated in a study, but this still does not say which study. Revealing the identity of a particular person in a particular study is still a threat, isn't it? Don't these two threats , genome in a mixture and linking attacks stand in parallel?

Commented [A15]: Detection attacks do not reveal identity of a person, they just reveal whether they attended the study or not. So, revealing identity of a particular individual in a study is performed via a linking attack.

Our point here is that as the size and number of the datasets grow, the detection attacks becomes less of an issue. We are pointing out the paradigm shift in the way attacks will happen.

Deleted: ,

Deleted: e.g. large genotype and phenotype data are used in medicine

Deleted: focus on gaining access to these data, then

Commented [NR16]: Very good illustration!

Deleted: ,

Deleted: which can render similar scenarios a reality in genomic privacy

Commented [A17]: Done!

Commented [NR18]: Please reconfigure this figure to have the box after step 4 align next to the box after step 3. The arrow going form 1 to 4 can be shortened and the Genotype comparison and matching moved above box 4. As presented there is too much white space in the center of the figure which does not look good in print and takes up too much space.

Commented [NR19]: Can you give an example for a sensitive phenotype?

Deleted: sensitive phenotypes

Among the QTL datasets, the abundance of expression QTL (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 relationship between expression levels and genotypes (**Fig. 2a**, **Supplementary Fig. 2**). 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] as representative datasets.

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 greatest 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 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 genotype-phenotype correlations decrease. 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.

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 give more information for linking. Thus, the information content of a set of genotypes is inversely proportional to the joint frequency of the 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 that the variant genotypes are distributed independently (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 without such information. 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 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, given gene expression levels, joint predictability can be easily quantified using π as it fits naturally to the information theoretic formulations (Online Methods). Furthermore, the predictability estimated via π can accommodate the non-linear relationships between genotype and phenotype –unlike the correlation coefficient, which generally measures linear relationships.

Deleted: that the attacker does this

Formatted: Highlight

Commented [NR21]: Fig. 2b needs to be called out before Fig. 3.
Fig. 3 is large and very complex

Formatted: Highlight

Formatted: Highlight

We first considered each eQTL and evaluated the genotype predictability versus the characterizing information leakage. We computed, for each eQTL in the GEUVADIS dataset, average predictability and average *ICI* over all the individuals (**Fig. 4a**). Most of the data points are spread along the anti-diagonal: eQTL variants with high major allele frequencies have high predictability and low *ICI;* and vice versa for variants with lower 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 do not deliver much characterizing information and vice versa for the eQTLs with lower frequency alleles. In order to evaluate how much gene expression levels contribute to predictability of genotypes, we use a shuffled eQTL dataset. The predictability versus *ICI* leakage for the eQTLs in the shuffled eQTL dataset (Online Methods) is dominantly on the anti-diagonal (**Fig. 4c**). This is also expected as the predictabilities for shuffled eQTL genotypes depend mainly on how frequently they occur in the population (major frequency genotypes are much easier to predict but have low *ICI*), as explained above. On the other hand, the real eQTLs (**Fig. 4b**) deviate from the anti-diagonal, compared to shuffled eQTLs, which shows that expression supplies much information for predicting eQTL genotypes (**Fig. 4c**). The eQTLs with high correlation have substantially higher ICI and greater predictability. These results illustrate the fact that π measures the total effect of genotype frequencies and expression levels on the predictability of genotypes.

When multiple genotypes are utilized, the information leakage is greatly increased. To study this, we computed *ICI* and predictability for increasing numbers 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 correctly link 51 (20% of 256) individuals. Although the attacker would not know which individuals are correctly linked, she can estimate reliability of linkings by statistical methods presented in following Section and focus on the most reliable ones. 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).

2.3 Framework for Linking Attacks

We present a three-step framework for practical analysis of linking attacks (**Fig. 2b**). This framework can be used to perform mock linking attacks on datasets to assess their privacy risks. 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

Commented [NR22]: But the attacker will not know which individuals are correctly linked, right?

Deleted: be

Deleted: Instantiation of

Commented [NR24]: Note that panels need to be cited in order

Deleted: instantiation

Deleted: examples?

Commented [NR25]: As meant?

Deleted: .

Deleted: evaluate whether they will be effective for risk assessment purposes

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 best matches 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 using a genotype prediction model where the attacker knows exact joint genotype-expression distribution (Supplementary Note). Although not very realistic, this scenario is useful as a baseline reference for comparison of linking accuracy. 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 (Supplementary Note, **Supplementary Fig. 4**) and links the predicted genotypes to the individual whose genotypes match best. For several eQTL selections with changing correlation threshold, the linking accuracy is above 95% and approaches 100% when auxiliary information is available (**Fig. 5a**).

In general, knowledge or correct reconstruction of the exact joint genotype-expression distribution may not be possible because the genotype-phenotype correlation coefficient alone is not sufficient to reconstruct the genotype distribution given the expression levels. The attacker can, however, utilize a priori knowledge about the genotype-expression relation and build the joint 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 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. 6a,b and 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 genotype with the maximum *a posterior* probability (Online Methods, Supplementary Note, **Fig. 2a**).

The extremity based prediction methodology assigns zero probability to the heterozygous genotype. Thus, it 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 ($|\rho|$) and extremity thresholds ($|\delta|$) for eQTL selection. The linking accuracy is higher than 95% for most 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, **Supplementary Fig. 8**). More than 95% of the individuals (**Fig. 5b,c**) are vulnerable for most of the parameter selections, which is more accurate compared to the baseline linking attack (**Fig 5a**). When the auxiliary information is used, the fraction of vulnerable individuals is 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 Note, **Supplementary Fig. 9d**). These results show that linking attack with extremity-based genotype prediction, although technically simple, can be extremely effective in characterizing individuals.

We evaluated whether the attacker can estimate the reliability of the linkings. We observed that the measure we termed, *first distance gap*, denoted by $d_{1,2}$, serves as a good reliability estimate for each linking. We computed the positive predictive value (PPV) versus sensitivity of the linkings with varying $d_{1,2}$ thresholds. For the eQTL selection where overall linking accuracy is 84%, the attacker can link a large fraction (79%) of the individuals at a PPV higher than 95% (Online Methods, **Fig. 5d**, **Supplementary Fig. 9a**).

We also studied several biases that can affect linking accuracy. First, when the eQTL discovery sample set is different from the samples set on which linking attack is performed, 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 those 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**).

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 framework presented here can be used for analysis of the information leakage when the correlative relations between datasets can be exploited for performing linking attacks (Supplementary Note, **Supplementary Fig. 10**).

Commented [NR27]: This goal should be mentioned in the abstract.

Commented [A28]: Please check updated abstract.

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 an attacker statistically links the databases using the *a priori* information about correlation of entries in the databases. The methods that we propose can be integrated directly into the existing risk assessment and management strategies. One such strategy 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 the entries or by 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. This maximizes the utility of the anonymized dataset by focusing only on the data points that leak the most characterizing information. In addition, as the anonymization process can focus only on the sources of highest leakage, this cuts down compute requirements[23] and increase efficiency of anonymization. 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[24]. In a differentially private database, release mechanisms are used to query the database and share statistics of the underlying data. The individual records in the database are not shared. To ensure the privacy of the database, the release mechanisms keep track of the leakage in the past queries and limit access to the database. For phenotype databases, the *ICI* leakage can be incorporated into the release mechanisms so that the total leakage can be tracked. It is also worth noting that anonymized data publishing and serving mechanisms may substantially decrease the biological utility of the data[25]. Thus, it is necessary to integrate measures of biological utility of the anonymized datasets as another quantity in the risk assessment.

4 ACKNOWLEDGEMENTS

Authors would like to thank A. Serin Harmanci for constructive comments and discussions on study design and running of external tools. Authors would also like to thank D. Clarke for comments on the manuscript.

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 the 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. Erlich, Y. *et al.* Redefining genomic privacy: trust and empowerment. *PLoS Biol.* **12,** e1001983 (2014).

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., Gehrke, J., Kifer, D. & Venkitasubramaniam, M. ℓ-Diversity: Privacy beyond k-anonymity. *Proc. - Int. Conf. Data Eng.* **2006,** 24 (2006).

23. Meyerson, A. & Williams, R. On the complexity of optimal K-anonymity. in *Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Princ. database Syst.* 223–228 (2004). doi:10.1145/1055558.1055591

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

25. Fredrikson, M. *et al.* Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. in *23rd USENIX Secur. Symp.* 17–32 (2014). at <http://www.biostat.wisc.edu/~page/WarfarinUsenix2014.pdf>

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. The genotype dataset contains the variant genotypes for *m* individuals. The genotype-phenotype correlation dataset contains *q* phenotypes, variants, and their correlations. The attacker predicts genotypes of *n* individuals in phenotype dataset using the phenotype measurements. The attacker then links the phenotype dataset to the genotype dataset by matching the predicted genotypes to the genotype dataset. 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.

Formatted: Highlight

Deleted: performs

Formatted: Highlight

Commented [NR29]: Not clear what is being predicted here since you said above the genotype data contains the variant genotypes. Do you mean genotype prediction for the phenotypic variants?

Deleted: prediction

Formatted: Highlight

Deleted: for all the variants

Formatted: Highlight

Commented [NR30]: But the identity of the genotype is not know, is it? So one still cannot trace the HIV status of a particular individual?

Figure 2: Illustration of genotype-expression associations and linking attacks (a) Schematic representation of expression-genotype relationships. 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 the line fit. In the extremity-based joint distribution, when the genotype value is 0, a uniform probability is assigned for expression values where extremity is smaller than δ (Green rectangle). For a 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 model 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 (Green rectangle). When genotype is 2, uniform probability is assigned to expression levels with positive extremity (Purple Rectangle). (b) Illustration of the three step linking process; selecting phenotypes and genotypes to be used in linking (step one), predicting the genotypes (step two), linking predicted genotypes to the genotype dataset (step three). The attacker can also estimate the reliabilities of the linkings using the first distance gap metric.

Deleted: .

Deleted: The first step entails

Deleted: . The second step entails

Deleted: . The third step entails

Figure 3: Illustration of computation of the individual characterizing information (*ICI*) and correct predictability of genotypes. *ICI* for a set of n variant genotypes is computed using the genotype distributions of the variants, as illustrated by the histogram plots under each variant. 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 the genotype given expression level is *e* is computed in terms of the entropy of conditional genotype distribution, given expression level *e*. The conditional distribution is built by slicing the joint distribution at expression level *e*.

Figure 4: Estimates of *ICI* leakage versus predictability. The plots show, for each eQTL, the information leakage (x-axis) versus correct genotype predictability (y-axis). The dots are colored with respect to: (a) the major allele frequency (b) absolute correlation of the eQTL (c) real versus shuffled eQTL datasets. (d) The average cumulative *ICI* leakage versus joint genotype predictability is shown when multiple eQTLs are utilized with shuffled eQTL dataset. The arrows on the plot indicate the increasing numbers of eQTLs used in estimated joint predictability and cumulative *ICI* leakage.

Figure 5: Accuracy of linking attacks. (a) Accuracy of linking with genotype predictions where exact genotype-expression distributions are known (baseline attack). The absolute correlation threshold (x-axis) versus the fraction of vulnerable individuals (y-axis) is plotted. 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, for the eQTL selection where overall linking accuracy is 84%, in comparison to the random selections of linkings.

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 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 ($v_{k,j} \in \{0,1,2\}$) of k variant for jth individual. The coding of the genotypes from homozygous or heterozygous genotype categories to the numeric values is 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 by 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*[26] of the event that RV takes the value $g_{k,j}$:

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

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:

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

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 the logarithm that is used determines the units in which ICI is reported. When the 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,

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

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

$$\pi(\{V_k\}, \{E_k = e_{k,j}\}) = \exp(-H(\{V_k\} \mid \{E_k = e_{k,j}\})) = \exp\left(-\sum_k H(V_k|E_k = e_{k,j})\right) \quad (4)$$

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

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

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 extremity value exceeding 0.3. Also, around 1000 genes have an absolute extremity exceeding 0.45 (**Supplementary Fig. 6b**). In other words, each individual harbors a 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

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

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

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

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. The posterior distribution of genotypes in equations (6-8) 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 predictions and assessed the genotype prediction accuracy. As expected, the accuracy of genotype predictions increases with increasing correlation thresholds (**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 a small number of total genotypes decreases the accuracy significantly.

8.4 First Distance Gap Statistic Computation

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

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

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

To identify eQTLs, we used the Matrix eQTL[27] 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 final eQTL 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 the strength of association between expression levels and genotypes. A similar procedure is applied when eQTLs for 30 trios are identified.

8.6 Modeling the Genotype-Phenotype Distribution

In the second step of the linking attack, the genotype predictions are performed. As intermediary information, the genotype predictions are used 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 context of risk management strategies, to study complexities of genotype prediction methods and to 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.

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, she must first model the joint genotype-phenotype distribution and then build the posterior distribution of genotypes (**Supplementary Fig. 5a**). The first level of the model can be built by decomposing the conditional distribution of expression (given genotypes) with independent variances and means (**Supplementary Fig. 5b**). Assuming that the 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 it 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 can be introduced by assuming that the variances of the conditional expression distributions are the 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 that the conditional distributions are uniform at non-overlapping ranges of expression (**Supplementary Fig. 5d**). This model requires 4 parameters (e_1, e_2, e_3, e_4) to be trained. This model can be further simplified into a model which requires only one parameter (**Supplementary Fig. 5e**). In this model, uniform probability is assigned when homozygous genotypes is observed and expression level is higher (or lower) than e_{mid}. In addition, zero probability is assigned when heterozygous genotypes are observed. Depending on the direction of genotype-expression gradient, the expression levels higher than e_{mid} associate with one of the homozygous genotypes and expression levels lower than e_{mid} associate with the other homozygous genotype. 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 the GEUVADIS mRNA Sequencing Project[17]. The eQTL dataset contains all the significant (Identified with a false discovery rate of at most 5%) gene-variant pairs with high genotype-expression correlation. To ensure that there are no dependencies between the variant genotypes and expression levels, we used the eQTL entries where gene and variants are unique. In other words, each variant and gene are found exactly once in the final eQTL dataset. The 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 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 the tissue analysis, the publicly available significant eQTLs for 6 tissues that are computed by the GTex project are downloaded from the GTex Portal. The HAPMAP CEU trio expression and genotype datasets are obtained from the HAPMAP project web site.

8.8 Code Availability

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

9 METHODS ONLY REFERENCES

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

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