Supplementary Material for “Analysis of Information Leakage in Phenotype and Genotype Datasets”

Arif Harmanci1,2, Mark Gerstein1,2,3,\*

1 Program in Computational Biology and Bioinformatics, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA

2 Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA

3 Department of Computer Science, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA

\*Corresponding authors: Mark Gerstein [pi@](mailto:mark.gerstein@yale.edu)gersteinlab.org

# Motivation on Extremity Attack: Outlier Attacks in Privacy

Extremity is a central concept in privacy. This is because the individuals who are outliers in certain characteristics are statistically more distinguishable than other samples, which makes them more prone to be targeted by the privacy breaching attacks. A simple example follows: “If a person is driving a very expensive vehicle, it can be deduced with high certainty that he/she is wealthy”. It is worth noting that the reverse is not always true; i.e., a wealthy person can also drive a mid-range priced vehicle. Thus the extremity of the vehicle price enables us to estimate very roughly the economical status of a person. In formalisms like k-anonymization1, the aim is to protect published datasets by imposing statistical indistinguishability of the rare and extreme features using different methods like censoring, swapping, adding noise. In our study, the attacker uses extremity to evaluate the outlierness of the individuals’ phenotypes, then he/she predicts the genotypes and distinguishes them from other individuals. Since the extremity is simple to estimate from the data, the extremity based attack can be implemented easily, which makes it fairly applicable and realistic in most situations.

In our study, we focus on the extremities of phenotypes, expression levels, to infer genotypes then link to the genotype datasets. The extremity based prediction exploits the outliers; i.e, the outliers in the expression levels are associated with the outliers in the genotypes, i.e., the homozygous genotypes. The heterozygous genotypes, do not co-incide with the extremes of the expression levels, i.e., they co-incide with the medium expression levels. Thus, we do not assign the heterozygous genotype in the genotype prediction. Although predicting only homozygous genotypes decreases the genotype prediction accuracy, the main goal is linking the individuals correctly. Thus, in the linking step, we utilize only the homozygous genotypes to compute the distances and perform matching.

# Comparison of “Detection of a Genome in a Mixture” and “Linking Attacks” in Genomic Privacy

Privacy has a multifaceted nature which can be breached under many different scenarios. The methods that assess and manage the risks, however, are scarce and are in need of development. Along this, our study aims at building analysis frameworks that will develop “measurable method for addressing privacy risk in information systems” (<http://www.nist.gov/itl/201506_privacy_framework.cfm>). In genomic privacy, the initial focus is to protect the identities of the individuals who attend genetic databases. The initial studies on privacy, therefore, focused on the statistical methods to predict whether a certain individual with known genotypes attended a study or not. We refer to this scenario as “detection of a genome in a mixture”. These are illustrated in Fig S6b. The attacker gets access to a genotype dataset (green). The attacker acquires also the statistics for the study in which he/she is to evaluate the participation of the individuals. The statistics can be simply the regression coefficients in a QTL study2, or the allele frequencies3 in a large scale genotyping study. He/she also needs a reference population on which the allele frequencies are known. These datasets are fed into a statistical testing procedure to decide whether the individuals in the genotype dataset have attended the study or not. Among all the scenarios, these attacks will breach privacy when an individual would like to hide their participation in a study. Although this holds true for many of the datasets, it is not relevant when the individual’s participation is almost certainly true or known. For example, if DNA genotyping becomes a routine operation in hospitals in near future, it will be most likely that an individual has participated in the genotyping dataset in their hospital of choice. The privacy concern will then be whether an attacker can pinpoint the individual among all other people within the large genotype database. The linking attacks become much more relevant at this point: If the attacker gets access to the genotype database, and can link it to another database with this individual’s phenotypes, he/she can reveal sensitive information (like disease status, address, sensitive phenotypes) by the linked entries in the databases. One famous example of these attacks (although not in a genomic context) is Latanya Sweeney’s1 demonstration of a linking which characterized the governor of Massachusetts, in addition to many other individuals, by linking the voter registration list to the Group Insurance Commission’s publicly released de-identified records using shared common columns in these databases. Sweeney also demonstrated that the identities of several personal genome project (PGP) participants can be re-identified by linking the PGP database to the voters list in a similar fashion as above4. Another well-known example was the demonstration of the linking attack on the Netflix and Internet Movie Database Records (IMDB). Netflix was sued by many people over the privacy concerns that stem from the linking attack demonstrated by Narayanan et al5 who linked the IMDB records and Netflix Prize competition database (seemingly unrelated databases of a very large number of individuals) to reveal identities of Netflix users, in addition to sensitive information about them. As it can be seen, the genomic linking attacks are almost orthogonal (or independent) to the detection of a genome in a mixture attacks since the attacker most certainly knows that the individuals at hand are in the genotype dataset that he/she is trying to link to.

These also point to the differences in the risks incurred by linking attacks and “detection of a genome in a mixture attack” and how these risks should be managed in different contexts. The main risk in detection attacks is founded on the detectability of participation of an individual in a dataset. Since the risks are incurred by the same datasets, they can be managed by evaluating which individuals can be targeted to detection attacks and restricting access to these individuals’ genotype and phenotype data. In linking attacks, the risks are founded on the linkability of an individual in a phenotype dataset to other datasets. Specifically, the risks are based on the fact that the linked datasets reveal sensitive information about the individual. The fact that these datasets are independently published/served will grossly complicate the risk management for linking attacks. The most secure risk management is restricting access to the genotype and phenotype datasets, or the QTL datasets. Another risk management strategy that can be useful data publishing is k-anonymization utilizing data perturbation techniques1. In these techniques, the phenotype data is anonymized in a way such that no combination of quasi-identifiers (i.e., predicted genotypes) are shared among less than *k* individuals. This is ensured by different techniques such as data censoring or noise addition. *k* can be chosen as a tradeoff between utility versus the risk of a privacy breach. Higher *k* implies a stronger anonymization of the data at the expense of lower utility of the data.

# Comparison of Extremity based Linking Attack Accuracy with Linking Attack in Schadt et al6

It is worth comparing the accuracies of extremity attack and the attack proposed in Schadt et al6. This attack takes as input a training set comprising the expression and genotype dataset and the list of eQTLs. Using the training set and eQTLs, it trains a genotype prediction model, which is then used for in the linking attack. On the other hand, extremity attack takes only the list of eQTLs. In order to compare the linking accuracies, we first divided the GEUVADIS dataset into 3 sets: First set is used for identifying eQTLs (85 individuals). Second set is used for training Schadt et al method (85 individuals) and the final set is used (174 individuals) for performing the linking attack and comparing the accuracies. We utilized the 1000 top eQTLs identified on the training dataset, as used in Schadt et al study6. Extremity based linking takes as input the eQTLs and the testing expression dataset. Schadt et al method takes as input the training set (expression and genotypes) and the testing expression dataset. The linking accuracies are shown in Table S2. It can be seen that both methods perform with very high accuracy. These results show that our approach performs comparably at high accuracy as the approach proposed by Schadt et al.

As the amount of data that is required is not the same while testing two methods, we also compared the amount of input that each method requires to gain the reported linking accuracies. Our method takes, for each eQTL only 1 parameter, which is the correlation coefficient. Schadt et al method, on the other hand, takes as input a training dataset (expressions and genotypes) to build the prediction model. We changed the training data size and evaluated the linking accuracy (Results in Table S2). It can be seen when the training data size is at 30 data points per eQTL, the accuracy of Schadt et al is almost comparable to extremity based attack. This result illustrates the difference in the required data size for both methods. Extremity attack requires 20 to 30 times less data compared to Schadt et al method, which highlight the practical applicability of the extremity attack on a dataset.

# Imputed Genotypes and Linking Attacks

Many studies use imputed genotypes in building genotype datasets. One practical question is how the imputed genotypes effect the linking accuracy. In principle, the SNP genotypes that are identified via imputation are not any different from genotyped SNPs in terms of characterizing information content they provide, so the presented methods should be able to handle them properly. One important point, however, is that the SNPs that are in linkage disequilibrium blocks tend to be very highly correlated and not give any additional information. In fact addition of these may increase redundancy and add noise to linking process and decrease linking accuracy. This is why we remove all the redundancies in genes and SNPs, i.e., each SNP and gene are used once in the linking attack. One could, however, evaluate the dependencies between genotypes and build a more complicated model of genotype prediction (step 2) and also include this information in linking (step 3) so as to reach a higher linking accuracy.

# A Basic Risk Assessment Procedure for Genotype-Phenotype and QTL Datasets

Figure S8 illustrates a risk assessment procedure that puts together different parts of our study. The analysis of tradeoff between ICI leakage and predictability (Section 2.2, top path in Fig S8) can be utilized for evaluating the risks associated with releasing QTL datasets. For a newly identified set of QTLs, the data releasers can compute the average information leakage and the corresponding levels of predictability to estimate the number of individuals that are potentially vulnerable at different levels of predictability. The predictabilities can be estimated using the conditional entropies in the QTL detection datasets, and the ICI leakage can be estimated using the genotype frequencies from the population panels. Secondly, the risks associated with releasing matching genotype and phenotype datasets can be evaluated using the 3-step linking attack frameworks. For this, the vulnerable individuals are identified. Finally a risk assessment can be performed to ensure that the vulnerable individuals are protected.

# An Example of Linking by Phenotype Extremity

Figure S7 shows an example of a linking attack that utilizes phenotype extremity. The basic idea is to use the extreme phenotypes (the gene expression levels) to estimate the genotypes then match them to the genotype dataset and reveal the disease status. In the example, we are focusing on 3 individuals; Bob, Alice, and John in the genotype dataset. The attacker makes use of 6 genes and variants in this attack. The gene expression levels are represented in terms of their extremity levels and some are shown as not extreme for illustrative purposes. The extreme ones are used in genotype prediction using the eQTL dataset for 6 genes. Given the predicted genotypes (note that some are predicted wrongly), Bob and John are correctly linked to their entries in the expression dataset and their disease status are revealed as positive. In this prediction, the 3 out of 4 predicted genotypes are the same for Bob and John (rs6052708, rs12479581, rs6077023). The 4th predicted genotype (rs7274244) enables pinpointing the exact entries for Bob and John. For Alice, however, there are two entries that are equally matching to the correctly predicted genotypes. The attacker, thus, cannot characterize the disease status for Alice.

# References

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

2. 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).

3. 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).

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. 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

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