RESPONSE TO REVIEWERS FOR "ANALYSIS OF INFORMATION LEAKAGE IN PHENOTYPE AND GENOTYPE DATASETS"

RESPONSE LETTER

-- Ref1: Introduction –--

Reviewer Comment	A. Harmanci and Gerstein demonstrate a three step procedure of how to initiate an attack on group privacy, through the seemingly innocuous use of aggregate datasets - those focusing on the quantification of expression quantitative trait loci (eQTL). At risk from the Harmanci-Gerstein Attack on Individual Privacy is the suspect's participation in any number of massive studies on obesity, body mass index, cholesterol, or even other hypothetical eQTL datasets that without fail (as shown in figure 1) contain HIV status as a covariate. While Harmanci-Gerstein Attack on Individual Privacy method does not immediately reveal whether the individual being targeted by Harmanci and Gerstein attack is indeed overweight and in need of a dietary intervention - or secretly harboring their high cholesterol numbers from a loved one. As hypothesized in this article, the fact that they have participated in biomedical research studies funded could lead to any number of negative consequences, including psychological trauma and taunts from peers for participation in a study published in a low impact journal. Most importantly, the perpetuator of the Harmanci-Gerstein attack would know that just beyond the dbGap chasm of click-through's, institutional monitoring, progress reports, more progress reports, and IRB's assuring that dbGap is absolved of privacy breaches' - well lies the suspect's genetic blue print - their individual level data. Harmanci and Gerstein advocate for changes the ways laws are made as an important step - specifically, adding risks estimates of leakage within future legislative decision making as a first step, which this paper helps to provide insight into.
Author Response	We thank the reviewer for providing detailed insight into our manuscript.
Excerpt From Revised Manuscript	

-- Ref 1: The reviewer suspects that the authors are unaware that very similar work was published in 2012 --

Reviewer Comment	The reviewer suspects that the authors are unaware that very similar work was published in 2012 with a fair amount of discussion and attention showing the core principles of this work on eQTL under what the reviewer considers a more broadly applicable mathematical framework. While the author's focus on using extremes or outliers as information sources has some unique aspects, the innovative work was in the original work by Im, Cox and colleagues in the American Journal of Human Genetics. Indeed it was a complete surprise at that time to those who read and went to meetings where this work was presented. I am sure the authors of this paper are in no doubt aware that Dr. Cox leads one of the largest NIH funded efforts putting forth eQTL data. Thus its reassuring to see that her team prospectively put for the careful analytical consideration of risk for the community to vet at that time in 2012.
Author Response	We thank the reviewer for pointing us to the Im et al 2012 study, which is an important study relating to Genomic Privacy which we should have cited in our manuscript. We have carefully reviewed the Im et al paper in detail. Interestingly, the reviewer views the scenario that is presented in Im et al study as the only way that the QTLs can be used to breach privacy and views the study as the de-facto standard on the problems of privacy breaches that uses genotype-phenotype correlations as a way to breach privacy. We believe there are major conceptual and technical differences in Im et al study and our study, which we list below. In the Im et al study, the authors address "detection of a genome in a mixture" in the setting of QTL GWAS studies. It should be noted, however, that we have cited Homer et al 2008 study, which is one of the earlier "detection of a genome in a mixture" studies. In Im et al paper, the attacker gains access to the allelic dosages (from genotyping arrays or DNA sequencing) at a large number of SNP sites for an individual and the regression coefficients of the SNP genotypes to certain phenotypes, the attacker can statistically identify whether the individual has participated in the original GWAS study or not. The output is a yes/no answer for indicating whether the individual has attended the study or not. We are, however, studying a different problem with a different setup: We are undertaking the "Linking Attack" problem. In this attack, the attacker aims at characterizing the individuals by linking the genotype and phenotype datasets ***to pinpoint and match the individuals in these datasets***. In our setting, as described in Figure 1 (And new Figure S5), we assume that the attacker gets access to 2 databases where first contains (de-identified)

Deleted: and Cox

Deleted: paper

Deleted: a very

Deleted: , Cox

Deleted: -Cox

Deleted: out

Deleted: -Cox

Deleted:

Deleted: -Cox

Deleted: as many individuals as possible.

measurements of a large number of phenotypes and second database contains genotypes and individual identities. The attacker aims at linking the first dataset to the second dataset, where the attacker uses one or more of the phenotypes in the first dataset and the phenotype-genotype correlations between the one or more of the phenotypes in the first dataset and the genotypes in second dataset. This way, the attacker can link the rows in the first dataset to the second dataset. Each correct linking of rows in the datasets, links of all the phenotype information (from 1st database) to the identity in the 2nd database, even the ones that were not used in linking. In this attack, the attacker is not necessarily aiming to identify a specific individual (as in "detection of a genome in a mixture") but rather tries to characterize as many individuals as possible. The accuracy and size estimation is the main focus of our study. In Section 2.2, we are aiming to jointly quantify the correct predictability of genotypes versus the amount of characterizing information leakage. Im-Cox et al do not address the issue of "linking", which is the 3rd step in the individual characterization.

This final point is important for following reason: Let's consider that our study is redundant in comparison to Im et al's study. This would suggest that an attacker could utilize Im et al attack to perform a linking attack. However, if an attacker tried to perform the linking attack as per Im et al study, the input and outputs of the method does not support a linking attack: The attacker could certainly utilize the Im et al's attack to each individual in the genotype dataset using the regression coefficients and determine whether they are in the phenotype dataset or not. After this, however, there is no machinery that is presented in Im et al study to link each individual in genotype dataset to an individual in the phenotype dataset. Therefore, we believe the linking attacks that we are focusing on are out of the scope of Im et al's study.

As we generate and gather larger and more inclusive genotype-phenotype databases, the linking attacks will become more relevant to privacy in comparison to the genome in a mixture identification, as many people will most definitely be in one or more of these databases. Consider following situation, which should clarify the differences even better: Attacker gets access to a genotype dataset of 100,000 individuals and that the attacker most definitely knows that the individuals in his/her phenotype dataset are already in this genotype dataset; i.e., no need to predict participation. The logical question that the attacker would ask is: Can I identify these people in the phenotype dataset within the genotype dataset? He/she would perform this using our

CLARIFY

Deleted: "

Deleted: "

Deleted: because of

Deleted: -Cox

Deleted: -Cox

Deleted: -Cox

Deleted: -Cox

Deleted: -Cox

Deleted: -Cox

Deleted: [[Within a million individuals, we most likely already know that the person attended the study, but who is he in that database?]]¶
¶
Another

manuscript's main focus, the linking attack. Im et al attack is not useful to the attacker at all as the participation is already known.

An important technical difference between the two approaches is that the statistical test in Im et al 2012 exploits the phenotype to genotype correlations of the specific phenotype and genotype datasets, and not the actual biological correlation:

> note that our method relies on "over fitting" of the data that occurs for individuals in the sample and not on any real relationship between genotype and phenotype. As previously mentioned, we found that the method worked equally well when a simulated phenotype was used.

On the other hand, in our study, we assume that the attacker utilizes a third party phenotype-genotype correlation dataset, which is utilized for linking. In our study, the information leakage happens through this "biological channel" (using genotype predictions via inversion of genotype-to-phenotype correlations), unlike the Im et al study, where the leakage happens through a "statistical channel".

One other technical difference is that Im et al perform classification of class membership (Participated/Not participated) using a statistical test that uses a statistic defined as following:

Let $\hat{Y}$ be defined as

$$\hat{Y}_I = \frac{n}{M} \sum_{j=1}^{M} \hat{\beta}_j (X_{Ij} - \hat{X}_j), \qquad \text{(Equation 1)}$$

where X_{Ij} is the allelic dosage of individual I at SNP j, $\hat{\beta}_j$ is the estimated coefficient from fitting the model $Y_I = \alpha_j + \beta_j X_{Ij} + e_I$, and $\hat{X}_j$ is the estimated mean of allelic dosage (twice the allele frequency) for SNP j computed with the reference group.

This statistic is genotype based, i.e. it takes the genotype based information, e.g., the authors utilize the DNA genotyping array based allelic dosage information in the results section. The authors propose two additional statistics, which are also genotype based. Our methodology, however, is based the genotype prediction, using the phenotypes. The extremity statistic, for example, is based on the phenotypic information.

Another important technical difference is that the class membership classification in Im et al attack works well (in terms of power, See Section name "Power of the Method" in 2012 paper)

Deleted: -Cox

Deleted: -Cox

Deleted: -Cox

Deleted: This is one of the main methodological differences between the two studies:

Deleted: -Cox

	when M>>n>>1, where M is the number of independent SNPs to be used in the classification and n is the number of individuals. Authors use M/n=300 in their experimental validations. Translating this to our test scenario, M/n=300 means, for GEUVADIS dataset where n=421, that one requires ***126,300 expression-genotype regression coefficients for each gene***. From the available files, the largest M for any gene goes upto at most several thousands of regression coefficients, where most of the correlations are against variants that are in LD (i.e. regression coefficients are not independent), which do not give much information. (It is worth mentioning also that, in the case of simulated dataset, we used n=100,211). Moreover, the attacker also needs to ensure M>>n*>>1; which indicates that the same criteria has to be satisfied with respect to the reference population. Considering the attacker uses 1000 Genomes as reference, i.e., n*=1092, the required number of regression coefficients are even much higher. Although for some eQTL studies all gene to all SNP pairwise correlations are made publicly available, they are, to our knowledge, not available in GEUVADIS project. These issues render the attack almost non-applicable on the GEUVADIS dataset. On the contrary, we evaluate our method's performance using one marker per phenotype, i.e.,one gene-one SNP, and using much less number of QTLs in the individual characterization, which highlights the applicability of the linking attack. We believe that above points clarify our study's differences from the Im et al study and other "genome in a mixture identification" studies, too. We believe this confusion is caused on our part as we may not have clarified well the attack setting. We have added a citation to Im et al paper in the background section and made updates to the introduction and methods section to ensure that our manuscript is clearer. We added Figures S5 and S6 to make linking attack scenario and differences with genome in a mixture identification attack scenario clearer.
Excerpt From Revised Manuscript	

Formatted: Font: Bold, Italic

Deleted: .

Deleted: n*=1092 as in

Deleted: (It is worth mentioning also that, in the case of simulated dataset, we used n*=100,211).

Deleted: -Cox

Deleted: -Cox

-- Ref1: The review views the incremental advancements over the 2012 paper do not support the far-reaching conclusions that the work by Harmanci and Gerstein for changing legistlative

decision making process in a way that the Im et al paper did not.

– --

Reviewer Comment	Again, a major aspect of this 2012 work was indeed privacy risk via eQTL, and indeed at that time it was a major shock to myself and other colleagues how powerful eQTL data really can be. In comparison of the two papers, the 2012 seems focused on a broader problem building from eQTL in line with Nature Methods as premier journal to publish methodological firsts. The review views the incremental advancements over the 2012 paper do not support the far-reaching conclusions that the work by Harmanci and Gerstein for changing legistlative decision making process in a way that the Im et al paper did not. I remain more impressed to see how Cox and colleagues in 2012 provider a broader framework and a bit stunned that p-values and odds ratios from enough SNPs limit absolute privacy. This generalizable framework intuitively makes sense - when asking one question about a person's membership in a cohort can we use thousands and thousands of correlated measurements to infer correctly the answer. The privacy risk management issue covered elsewhere then is towards what is the probability of this impacting a specific person's privacy.
Author Response	The reviewer finds our study's contributions not very impressive compared to Im et al study. As we outlined above, our study addresses a different aspect of genomic privacy compared to Im et al study. Our study's main aim is to first bring into public view the potential risks behind releasing seemingly unrelated phenotyping datasets. The linking attacks attacks underpins these risks. We concentrate on quantification of the leakage in these attacks and show how extremity based genotype prediction can be utilized to perform a very effective linking attack. Extremity is a fairly central theme in privacy analysis: Any time an individual is outlier in any feature, they can be distinguished easily from other individuals. Although fairly simple to implement, our results demonstrate the usage of extremity in the context of genotype prediction and linking attacks. The reviewer puts forward Im et al and the "genome in a mixture identification" as a meaningful and generalizable framework. Although we agree with the reviewer that a meaningful risk management should be defined in studies on privacy, we believe that the Linking Attacks should be analyzed in a different scenario compared to the studies on "genome in a mixture identification". We thank the reviewer for articulating on our suggestions for changing the legislative decision making processes. We are not aiming to create a panic environment. In the contrary, our aim is to

Deleted: -Cox

Deleted: we believe

Deleted: -Cox

Deleted: -Cox

	build analysis frameworks against the linking attacks. Our main goal with these suggestions is that the approaches for bioinformatics analysis of genomic privacy proposed by our study and many others before our study should be used more extensively while data sharing mechanisms are designed. For this, we also made our tools available. [[We have reworded the legislative clauses to ensure that this study advances on all the previous studies]]
Excerpt From Revised Manuscript	

-- Ref1: the paper doesn't consider a hallmark of risk management of also considering the probability of a 'meaningful' privacy breach –--

Reviewer Comment	This brings the second major critique of the paper, that the paper doesn't consider a hallmark of risk management of also considering the probability of a 'meaningful' privacy breach to an individual and damages incurred under proper analysis of risk management. The paper brings up the legislature goals, and thus that lack of utilization of standard approaches for managing and quantifying risk management is a fair area of critique and a deficiency. Of course, a major premise of legislative privacy is the impact or damage to an individual by a privacy breach. The question can be framed: "What is the probability that a person with information they wished to remain protected from other individuals is compromised, and what is the tort damages if so? " The authors frame privacy risk through an anecdotal example that seems unfounded in individual privacy - in contrary to the example the authors used, privacy risk is not only about speculating that a person exists who wants to expose as many people as possible, as is hypothesized in this paper. Pragmatically, it's more probable that a person would search for a specific person, such as a child of a sperm-donor father.
Author Response	We understand that the reviewer finds our scenario anecdotal and unrealistic. We agree that the attack scenarios should provide a reasonable argument showing a real risk on individual privacy. We, however, do not agree with the reviewer's view that our scenario, privacy breach via linking attacks, is not founded in individual privacy. Firstly, Schadt et al's 2012 study (Cited in the Background Section) takes on the linking attacks in a scenario that is practically the same as ours. Apart from this, linking attacks have a very rich literature in the field of privacy research. One very well-known example is Latanya Sweeney's[1] demonstration of a linking which characterized the

Deleted: background section

Deleted: ¶
Apart from this, linking attacks have a very rich literature in the field of privacy research. One very well known example is Latanya Sweeney's demonstration which characterized the governor of Massacusetts, in addition to many other individuals, by linking the voter registration list to the Group Information Comission using several common columns in these databases.¶
¶

	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. In addition, 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 performed by Narayanan et al[2] 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. The story can be found here: https://en.wikipedia.org/wiki/Netflix_Prize#Privacy_concerns To relate this further to our study; any movie enjoying person can be expected to be in one of these datasets, which renders the prediction of participation problem (Im et al study) somewhat useless. Actually, Netflix is enormously popular and includes millions of individuals in their databases. There is a very good chance that any person in a group of intellectual individuals that we randomly pick will be in one of these databases. The question that an attacker would be, can I identify who these people are and what their preferences are? In addition, the literature on linking attacks (and on any privacy aware data publishing/serving mechanism, for that matter) consider any type of sensitive information leakage will lead to a privacy breach and must be protected. Formalisms that try to limit the leakage are: k-anonymization and differential privacy, l-diversity, t-closeness, etc. Following this, we would like to argue that the risk management (via anonymization) that these formalisms provide do not conform with the reviewer's view of a reasonable risk of privacy breach. In these studies, for example k-anonymization, any individual that can be characterized/identified is considered a serious risk, and thus must be protected. In other words, characterization of even one individual is as serious a risk as characterization of many. [[We have added a discussion that explains above points about risk management]]
Excerpt From Revised Manuscript	

Deleted:

Deleted: researchers who linked the IMDB records and Netflix Prize competition database to reveal identities of Netflix users.

Formatted: Hyperlink

Deleted: ¶
¶

Deleted: these

Deleted: also

Deleted: breacy

Deleted: even

YES

-- Ref1: The review views the incremental advancements over the 2012 paper do not support the far-reaching conclusions that the work by Harmanci and Gerstein for changing legistlative decision making process in a way that the Im et al paper did not. – --

Reviewer Comment	As such, and as has been generally modeled in other frameworks, the focus should be on positive predictive value. Given a person is trying to keep information private that would be damaging (legislative tort is framed in damages both punitive and otherwise as such as HiV stat), what is the probability that a person would correctly identify something about their privacy. Thus this metric considers - well most people don't participate in studies and that too many false positives makes an approach unreliable at detecting a rare event. It also reflects that a privacy breach for a random person visually obese would not be meaningful for many people who have pride in participating in a biomedical study. Thus the reviewer provides a specific suggestion that is to frame improvements of their methods in comparison to the proposed methods as either PPV or AUC, given the overall prevalence of people participating in eQTL databases that could expose potentially damaging information. The review concern is that they rare 'outlier information' would lower the prevalence and thus not increase diagnostic accuracy.
Author Response	We understand that the reviewer's suggestion about comparison of our proposed method in terms of positive predictive value. We have made two changes to the manuscript to address these concerns. First, in order evaluate the risks that are incurred by the extremity based attack, we evaluated the positive predictive value of the linkings. For this, we propose the *first distance gap*, $d_{1,2,}$ which the attacker can compute for each linking to estimate reliabilities of the linkings. The attacker can use this measure to sort the linkings and evaluate whether to use the linkings or not. We have included sensitivity versus PPV plots (Figs 5, 6) for the different linking scenarios. It can be seen that when the attacker utilizes this measure, among all the test scenarios, more than 50% of the linkings (sensitivity) can be performed with PPV greater than 95%. In some cases the sensitivity goes up to 80% or more while PPV is greater than 95%. These results show that our method does not link only the obvious outlier individuals but a much larger fraction. Among the methods that are mentioned, the most relevant to our method is Schadt et al 2012's methodology. In order to compare the two methods, we use the testing and training datasets and selected different number of eQTLs with highest correlation to

Deleted: use a metric, d_1,2,

Deleted: XX, XX

Deleted: (at PPV>95%)

Deleted: upto

Deleted: .

Deleted: distance measures that we use against that of Schadt et al's distance measure based on posterior probability computation

Deleted: used

	evaluate the effect of changing SNP numbers on the linking accuracy of the Methods. The results can be seen in Table S1, which show that both methods perform very similarly and identify very high fraction of individuals. These show that the extremity based linking can characterize individuals very similarly in terms of linking accuracy as the model based approach proposed by Schadt et al. It should be noted that Schadt et al's method requires, in addition to the list of eQTLs, a training dataset to build a model for genotype prediction, while our method requires only the list of eQTLs to be used in linking. In order to make a comparison of accuracy versus input size, we evaluated how the accuracy of Schadt et al method changes with changing training data size. For this, we evaluated the linking accuracy of Schadt et al with changing training data size. The results are tabulated in Table S1b. These show that the accuracy of Schadt et al's method decreases as the training data size decreases and requires at least 60 data points (30 expression and genotype values) per eQTL to train the model robustly and accurately. Our method requires roughly 60 times less data (only 1 parameter per eQTL is necessary), which illustrates the difference in terms of the required input size to each method. This also reflects the applicability of each method by an attacker: Extremity based linking requires much less information and thus is much easier to implement compared to Schadt et al's methodology. In more simple terms, our method can bring a very high and comparable linking accuracy as the Schadt et al's method, while requiring much less input information. We also want to emphasize that the results of a comparison of privacy breaching methods should be treated with caution. Our aim is to evaluate whether using a model-free approach (extremity based) decreases the linking accuracy of the attacker significantly compared to the model based attack. Since all the attacks represent a different routes to a privacy breach, the data publishing/sharing mechanisms must consider and protect against all of these attacks, rather than considering just the "best" one.
Excerpt From Revised Manuscript	

Deleted: distance measures..

Deleted: are shown in Table SXX. It

Deleted: for most part

Deleted: measures

Deleted: at

Deleted: accuracy. At very small SNP numbers (at 50 SNPs), our distance measure has slightly lower accuracy (9

Deleted: less than Schadt et al's distance measure out of 211

Deleted:). These results show that our model-free approach can capture much

Deleted: the

Deleted: compared to

Deleted: of

Deleted: al's approach.

Deleted: ta

Deleted: It is not to evaluate which method works better, as is done, for example, in protein structure prediction literature.

-- Ref1: the reviewer profusely thanks the authors for putting forth a paper that breaks the monotony of boring and dry introductions/discussions –--

Reviewer Comment	Finally, the reviewer profusely thanks the authors for putting forth a paper that breaks the monotony of boring and dry introductions/discussions, for one that confidently suggests the legislature should carefully utilize this framework for their deliberation to protect our privacy. Enjoying both the tone of the discussion and introduction, I was only disappointed to see no references to the NSA, Edward Snow, or Jennifer Lawrence woven into sections on privacy breaches. The reviewer suspects the authors were unaware of prior similar work and similarly appreciates a periodically 'tongue and cheek' and playful review critique.
Author Response	[[Closing statements, not to be included]] We thank the reviewer for constructive suggestions, which we believe made our manuscript much more complete. After consideration, we did not find the suggested individuals to be sufficiently related to biomedical data privacy.
Excerpt From Revised Manuscript	

-- Ref2: Introduction –--

Reviewer Comment	In this article, Harmanci and Gerstein investigated an intriguing question regarding genomic privacy: given a person 's phenotype (specifically eQTL), whether an intruder can stake advantages of known genotype-phenotype correlations existing in the public domain and reversely predict the genotype of the person. The authors showed that ... As stated by the authors, this work can be considered as an extension of an earlier work by Schadt and colleagues (Nat Gen 2012), in which they showed that given a set of high-quality mRNA expression data of a given tissue for a human cohort (and SNPs) as training data, one can predict the genotypes of another independent cohort with high accuracy. One of the major innovations of this work in comparison with the earlier work is that they showed that, inclusion of additional phenotypic data (gender and ethnicity) gives the intruder more power in predicting genotypes. The second breakthrough of this work is that, instead of using Bayesian probabilistic approach, the authors showed that the potential privacy intruder can use the extreme outliers existed in the phenotypic data as a guidance to identify the corresponding individual.
Author Response	[[Just the introduction. This is here to be complete. Probably going to remove this]]

Excerpt From Revised Manuscript	

-- Ref2: I think the work itself is interesting, however the presentation can be further clarified in places. –--

Reviewer Comment	I think the work itself is interesting, however the presentation can be further clarified in places. For starters, the equations in the manuscript need to be numbered so that it helps the readers (and reviewers) to reference the mathematical work (there are no page numbers either). The foundation of the "extremity" is described in Section 2.4, I am a little surprised that the authors did not provide any reference in this part, has the concept of Extreme Statistic not ever described in other field? I would like to see more elaboration and motivation on this part. Is the "extremity statistic" just a transformation of rank correlation? Also please clarify why genotype value 1 is never assigned to 1.
Author Response	We agree with the reviewer's rightful concern that the mathematical work is clearly labeled, which may make it harder to follow. We added numbers to all the equations and also added page numbers. These should make it much easier to follow and refer to the mathematical work in the manuscript. Extremity statistic is very much related to normalized rank, which we referred to in the manuscript. The genotype prediction by extremity statistic utilizes the fact that the extremes of gene expression levels associate with the extremes of the genotypes, i.e., homozygous genotypes. The attacker uses this to build a simplified estimate of the posterior distribution of genotypes given expression levels and utilizes this for genotype prediction. The genotype prediction for each SNP (given the expression levels) can also be conceptually interpreted as performing a rank correlation between the homozygous genotypes and the gene expression levels and selecting the genotypes that maximize the correlation. We understand that the reviewer finds extremity based genotype prediction not well motivated. In fact, using extreme phenotypes of an individual is a general route to a privacy breach. This is because, any outlier phenotype of a person is an identifying feature that can be used by an attacker to characterize/identify the person. In our study, we focus on the extremities of phenotypes 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. Finally, to address reviewer's last question: The heterozygous genotypes, do not co-incide with the

Deleted: up

Deleted: ¶
As the reviewer suggests, the extremity is related very much to a normalized rank. The genotype prediction, can thus be thought of as a rank correlation between the genotypes and the gene expression levels. We actually found that this …¶
¶

Deleted: find

Deleted: unclear

Deleted: however,

	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. Finally, in the linking step, we utilize only the homozygous genotypes in the matching, since we predict only those. We clarified the explanation of genotype prediction by extremity attack in the Results Section.
Excerpt From Revised Manuscript	

Deleted: mean level

-- Ref2: some concrete examples would be very helpful to demonstrate the power of the approach described by the authors –--

Reviewer Comment	Also, I think some concrete examples would be very helpful to demonstrate the power of the approach described by the authors, i.e. identities of individuals that would not have been discovered if only gene expression data was used or if extremity approach was not used.
Author Response	We added Figure S6 to illustrate a specific example of the linking attack by the extremity based genotype prediction. The example first illustrates the specific details of the extremity based linking attack by showing how the extremities translate to the predicted genotypes. It also shows how the extremities in gene expression levels can help the attacker can distinguish between two individuals. We believe this figure helps illustrate better the idea that gene expression extremity can lead to privacy breaches in linking attacks.
Excerpt From Revised Manuscript	

-- Ref3: Introduction –--

Reviewer Comment	Genomic privacy is an increasingly important direction of research. One of the aspects of work on genomic privacy has focused on ways to breach privacy by linking different kinds of data. This paper presents an attack that can be used to link a phenotype (in their specific case, gene expression) to a genotype and possibly to other identifying information. The study presents simulations to show the feasibility of this attack. The authors consider the following setup: an attacker has access to an individual genotype (this could be part of a larger dataset), a dataset of individual-level gene expression (but no genotypes) and a list of variants that are known to affect expression of specific genes. The attack consists of predicting the genotypes at

	the list of expression SNPs corresponding to the the gene expression data and then testing if the target individual genotype matches any of the predicted genotypes. They consider two variants. In the first (2.3), the attacker needs a prediction model to predict genotypes from expression. This, in turn, implies that the attacker would need access to data where individuals have genotypes as well as gene expression. In the second (2.4), termed Extremity-based genotype prediction, the attacker only has access to the correlation between genotype and gene expression. The authors show that for both variants, a large fraction of individuals (>=95%) are vulnerable as assessed by simulation experiments on the GEUVADIS dataset.
Author Response	[[Just the introduction]]
Excerpt From Revised Manuscript	

-- Ref3: The authors need to do a better job of clarifying their contribution and motivating the reason why variant 2 is realistic. --

Reviewer Comment	1. Variant 1 of the attack is very similar to the attack described in Schadt et al. (Nature Genetics 2012) which the authors cite. The only difference is that here the authors explore the number of eQTLs to use while Schadt uses 1000 top cis eQTLs. Variant 2 is novel as it relaxes the requirement that the attacker has access to joint genotype-gene expression data to learn the prediction model. The authors need to do a better job of clarifying their contribution and motivating the reason why variant 2 is realistic.
Author Response	We agree that we may have not clearly stated our contributions. We are listing them below for clarification: In Section 2.2, we are proposing quantification metrics that measure the tradeoff between predictability of the genotypes and the information leakage in the predicted genotypes. These metrics that we proposed can be utilized for evaluating the extent of leakage and the corresponding risk (predictability) of individual characterization while new phenotype-genotype correlation datasets are being released. Attack Variant 1 (Section 2.3) is a generalized analysis of the linking attack, where the attacker knows perfectly the joint expression-genotype distribution. Although seems similar to Schadt et al study, we do not assume a specific model of prediction. In Schadt et al, the authors utilize a Gaussian approximation for genotype predictions. This enables a more

Deleted: [[Clarify contribution]]¶

THIS

	generalized analysis of the linking attacks in the 3-step analysis framework that we proposed. Attack Variant 2 (Section 2.4) is the extremity attack. This attack is an instantiation of the 3-step decomposition, and also illustration of that a model-free approach can reach very high linking accuracy. As explained by the reviewers, we are investigating whether the attacker can just use a measure of "outlierness" in the gene expression levels for genotype prediction. We then evaluate under different situations the viability of this novel attack. We understand that the motivation for extremity attack may not be well-stated in our manuscript. Extremity is a fairly 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. For example, k-anonymization aim to protect published datasets at statistical indistinguishability of the rare and extreme features by different methods (e.g. censoring, swapping data, adding noise) so as to protect it. In our study, the attacker uses extremity to evaluate the outlierness of the individuals' phenotypes, predicting genotypes distinguishing them from other individuals. Since the extremity is simple to estimate from the data, which can be combined with the proposed model-free estimation procedure, the extremity based attack can be implemented easily, which makes it fairly accessible and realistic in most situations.
Excerpt From Revised Manuscript	

-- Ref3: The experimental validation needs to be improved. [[Training/Testing based eQTL selection]]–--

Reviewer Comment	a. The experimental validation needs to be improved. The authors tested their attacks on the GEUVADIS dataset. However this setting would produce optimistic results as the model was learned and the tested was done on the same data. It would be more appropriate to split the data into a training and test set where the training set is used to pick eQTLs and the test set is used for identification.
Author Response	We agree with the reviewer that matching of eQTLs and testing dataset can create a bias. To address this issue, we have divided the GEUVADIS samples randomly in two sets (210, 211 individuals, respectively). One of the sets is used for identifying the eQTLs, using Matrix eQTL tools. The generated set of eQTLs are used in the second set for computing the characterization

	accuracy. It can be seen that the characterization accuracy is slightly lower than the matching test/training sets but still very high. We have updated the …
Excerpt From Revised Manuscript	

-- Ref3: there are a number of biases that can reduce accuracy. – [[Population stratification]]--

Reviewer Comment	b.In addition, there are a number of biases that can reduce accuracy. For example, if gene expression in the training and test sets were measured in different tissues, platforms, populations. The manuscript currently does not address complications that are likely to arise in practice. I would have liked to see such a discussion as well as empirical results that document the effects of these biases.
Author Response	We agree with the reviewer that different biases can be introduced when the eQTLs are computed using datasets from different sources and technologies. To evaluate this, we focused on the population stratification, specified by the 1000 Genomes Project. We have selected 3 populations: GBR, CEU, and YRI. For each population, we identified the eQTLs (using Matrix eQTL) then tested the matching accuracy on the expression values of other populations. We observed that for GBR and CEU populations, the eQTLs provide high matching accuracy (>95%) accuracy, while the YRI eQTLs provide slightly lower accuracy (??%). These results indicate that when the eQTL dataset is generated over individuals of different background that is not close to the tested individuals, the matching accuracy can be rather low. This result can be attributed to the fact that the different genetic backgrounds can change the eQTL compositions in different populations, which decrease the power of extremity based genotype prediction, and decrease the individual matching accuracy. When the eQTL identification and testing data populations are close, however, the matching accuracy is significantly higher. These results are in accordance with the Schadt et al study. It should, however, must be noted that Schadt et al assumes that in the matching, the attacker has access to the population knowledge and genotype frequencies of the individuals being matched, while our approach has no a-priori knowledge and only depend on the eQTL knowledge. We also studied how the accuracy gets affected when eQTLs are identified from different tissues. For this, we used the eQTL

	database of GTex Project and downloaded tissues for 5 tissues. We also performed the matching against the genotypes of 1000 Genomes phase1 individuals of 1092 genomes. It can be seen that the linking accuracy is still fairly high (>80% for all tissues except Skeleton eQTLs). As expected, we observed the highest accuracy for Whole Blood eQTLs. The decreased accuracy (compared to the matching tissues) can be attributed in part to the data processing and handling differences between the studies. These results show that the linking accuracy can still be fairly accurate when the eQTLs are identified in tissues that are not matching the tissues in which expression levels are measured.
Excerpt From Revised Manuscript	

-- Ref3: It would also be interesting to understand how these attacks scale with data set size. [[100k size genotype dataset vs performance, close relatives?]]–--

Reviewer Comment	c. It would also be interesting to understand how these attacks scale with data set size. For example, how feasible is this attack within a dataset of 100,000 genotypes that are now being generated. Another interesting question is whether the method can discriminate close relatives that are likely to be present in large datasets.
Author Response	We agree that these are important points for illustrating the general applicability of the extremity attack. To evaluate how the matching genotype sample size affects the accuracy, we simulated 100,000 individuals using the 1000 Genomes genotype frequencies for the eQTL SNPs. The eQTLs are identified from the training set of 210 individuals. The 100k simulated individual genotypes are then merged with the 211 testing sample set to generate the 100,211 individual sample set. We then used the expression levels (from GEUVADIS dataset) for the test sample and performed the extremity based attack on this larger dataset to check the characterizability of individuals in testing set. We observed that the matching accuracy is very high, around 99%. This result indicates that extremity attack can potentially be effective in very large sample sizes. In order to evaluate how the existence of close relatives affect linking accuracy, we focused on the genotype and expression data for 30 CEU trios (father, mother, child) in the HAPMAP project. We identified the eQTLs using all the individuals and then performed extremity based linking attack. Although the linking accuracy is very high, we wanted to evaluate how the close relatives were scored in the linkings. Thus, we computed the ranks of close

Formatted: Not Highlight

Deleted: In order to evaluate the family … CEU trios in HAPMAP …

	relatives (child-mother, child-father linkings) in the linking process (excluding self ranks) and compared those to the ranks of randomly selected individuals in the dataset. The distriburion of ranks are plotted in Fig. 8. It can be seen that the rank distribution of the close relatives is significantly shifted towards smaller ranks; which indicates that the linking assigns smaller ranks to the close relatives. This indicates that the individuals that are close relatives This result has a significant consequence: Even when the individual that the attacker aiming to link is not in the genotype dataset, the attacker may still be able to link him/her to a close relatives that may be in the dataset, which would identify the family of the individual and cause a privacy concern.
Excerpt From Revised Manuscript	

-- Ref3: For a realistic attack, the attacker would need some threshold on the distance function to decide if a test individual is linked to a given predicted genotype. How should this threshold be chosen ? [[Rejection threshold selection]] –--

Reviewer Comment	d. The authors declare an individual to be vulnerable if pred_j = j. This is only a first step in documenting its utility. For a realistic attack, the attacker would need some threshold on the distance function to decide if a test individual is linked to a given predicted genotype. How should this threshold be chosen ? Does it give adequate power at a low false positive rate i.e. very few unrelated individuals fall below the threshold while the correct individual does ?
Author Response	The reviewer raises an important point. If the attacker can find a way to measure the reliability of the matchings he/she performed, he/she can focus on those individuals for which the linking has high reliability and increase his/her chance of a breach at the cost of a decrease in the sensitivity of matching. For this, the attacker also has to use only the information that is available to him/her, i.e., he/she cannot use the correct genotypes. We found that, for each linking, "genotype distance difference between best and second best matching individuals" (*first distance gap*) serves as a good measure, that the attacker can compute for each linking, to estimate the accuracy of the linkings. (See Methods Section) This measure stems from the observation that when the linking is incorrect, sorted distances at top are much closer to each other compared to the ones when the linking is correct.

Deleted: 1st-to-2nd

Formatted: Font: Italic

Deleted: difference

	In order to evaluate this measure's effectiveness, we evaluated the matchings when the whole eQTL list from training sample is considered. Among the 86% that is correctly identified, we are evaluating whether the ranking with respect to distance difference places the correct matchings to the top. We computed the distance difference for all the matchings that the attacker does, and sorted the matchings with respect to the difference. Finally, we computed the positive predictive value and the sensitivity over increasing distance difference cutoff values, which is plotted in Fig. 6b. Compared to random rankings of the matchings (which uniformly have 86% PPV), this sorting provides much higher PPV. In addition, upto 79% of the individuals can be linked correctly with more than 95% PPV. These results illustrate that the attacker can rank the matchings using the proposed *first distance gap* difference and select the ones that have high genotype distance to focus the attack on highly reliable linkings.
Excerpt From Revised Manuscript	

Deleted: 1st-to-2nd

Formatted: Font: Italic

Formatted: Font: Italic

Deleted: ¶
¶
[[Also, assign a notation for this distance measure]]¶
¶
[[We need a supplementary figure to illustrate this: Each linking is basically computing distance to all the individuals in the genotype dataset.]]¶

-- Ref3: The presentation could be clarified to highlight the main contributions. –--

Reviewer Comment	3. The presentation could be clarified to highlight the main contributions. a. For example, it is unclear how section 2.2 relates to the rest of the paper. While it is interesting to see the relationship between predictability and leakage, this result does not seem to be used later. The choice of eQTLs is done simply using the correlation. b. Similarly, I would have liked to see a better motivation of extremity-based prediction (which I consider to be the most interesting part of the paper) and a better experimental validation.
Author Response	We agree with the reviewer's concern. As we explained above, we have updates the conclusion section to clarify how Section 2.2 relates to the other sections. In addition, we have added Supplementary Figure S8 that illustrates how the different sections in the manuscript can be utilized in general in a risk assessment procedure. We believe these updatese clarify how different Sections fit with each other in the manuscript.
Excerpt From Revised Manuscript	

Deleted: [[Rephrase, move, clarify]]

-- Ref3: Typos –--

Reviewer Comment	Typos: Page 2: "GTex project hosts a sizable set of eQTL dataset" Page 4: "the all the predicted genotypes"

Author Response	We sincerely thank the reviewer for very careful reading of our manuscript. We have fixed the typos pointed out by the reviewer.
Excerpt From Revised Manuscript	

Deleted: the

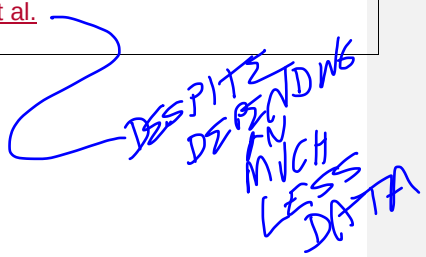
-- Ref4: Remarks to the Author –--

Reviewer Comment	The authors present a rigorous and important analysis of how predictive are genotype-phenotype correlations, using an expression quantitative trait loci (eQTL) dataset as an example. Their method predicts genotypes from eQTL gene expression with high accuracy, addressing privacy concerns related to genetic data identifiability. Despite their important contribution to addressing this problematic issue, I have some concerns and questions about this manuscript that preclude me from giving it my strongest support.
Author Response	[[This is the introduction, here for completeness, to be removed.]]
Excerpt From Revised Manuscript	

-- Ref4: Major Critique: the authors do not compare the performance of their method with this previous one. This should be done [[Schadt Comparison]] –--

Reviewer Comment	The authors rightfully cite a previous publication (Schadt et al, Nature Genetics 2012) that relates to their study, as they also developed a method to predict genotypes from eQTL gene expression. Nevertheless, the authors do not compare the performance of their method with this previous one. This should be done, as to assess the importance of this new method with the current state-of-the-art tools addressing the same issue.
Author Response	We understand that the comparison between the mathods is necessary. For this, we first requested the source code of model based method from Schadt et al. For comparison, we utilized the eQTLs identified on the training dataset. For training Schadt et al's method, we used the training set, and evaluated the accuracy of linking on the testing set. We utilized different number of eQTLs to compare the accuracy of methods with different markers. The results are shown in Table SXX. It can be seen that both methods perform with very high accuracy even at relatively smaller number of markers. These results show that our model-free approach performs comparably (at high accuracy) as the model based approach proposed by Schadt et al.

Deleted: [[The problem here is that Schadt et al does not provide source code. I can try and do my best to change the first part of the paper to match Schadt et al's model based prediction, using part of the data for "model" building, and other parts for testing. This is also useful since Ref3 also asked something similar. On the other hand, this may not be a fair comparison since it may not capture all the details of Schadt et al. We can thus just spin it by saying that we the model based method (as an alternative to Schadt et al's method) and the extremity based prediction and model based prediction are very similar in performance.]]



	[[In order to compare the amount of input to each algorithm, we ran Schadt et al algorithm with different input sizes]]
Excerpt From Revised Manuscript	

-- Ref4: the authors do not mention which was their p-value threshold. At least FDR<5% should be used. –--

Reviewer Comment	The authors use the reported eQTL correlation coefficient as the criteria for strength of the eQTL association. Nevertheless, the authors do not mention which was their p-value threshold. At least FDR<5% should be used. One of the problems of using only the correlation coefficient is that for instance for rare SNPs, the correlation coefficient might be extremely high but the p-value can be borderline significant.
Author Response	We agree with the reviewer's rightful concern. There are several eQTL datasets that we used: For eQTLs obtained from GEUVADIS project, we made sure to use FDR<5% eQTLs, which are located under project data files. For the eQTL datasets that are identified via training datasets using Matrix eQTL method, we used only the expression-genotype pairs for which Matrix eQTL reports at most 5% FDR, which is computed via Benjamini-Hochberg methodology. We have updated the Methods Section in detail to explain how eQTL selection was performed.
Excerpt From Revised Manuscript	

Deleted: data section

-- Ref4: why does the genotype accuracy decreases when the absolute correlation threshold is bigger than ~ 0.7? –--

Reviewer Comment	In Figure 5b, why does the genotype accuracy decreases when the absolute correlation threshold is bigger than ~ 0.7?
Author Response	[[This is actually a good question, the problem is with the accuracy computation: Very small number of SNPs make the genotype accuracy (the fraction) very unstable, although we expect very high accuracy, 1 wrong prediction out of a small number in the fraction makes it go down. I will look into this a little more and make sure my explanation is correct. Should be just clarification and update.]]
Excerpt From Revised Manuscript	

-- Ref4: It is not clear if your tool available at http://privaseq.gersteinlab.org can use the "Extremity based Genotype Prediction" –--

Reviewer Comment	It is not clear if your tool available at http://privaseq.gersteinlab.org can use the "Extremity based Genotype Prediction". Please clarify in a README file.
Author Response	[[Will update the README file.]]
Excerpt From Revised Manuscript	

-- Ref4: can your tool address this by being able to use imputed genotypes? [[Will we get the same privacy issue when the array studies use imputed genotypes?]]–--

Reviewer Comment	Since a lot of new studies have published eQTL datasets based on imputed genotypes, can your tool address this by being able to use imputed genotypes?
Author Response	The reviewer is raising an important point. In principle, the SNP genotypes that are identified via imputation are not any different from other SNPs in terms of characterizing information content they provide, our tool should be able to handle them properly. One important point is, however, that the SNPs that are in linkage disequilibrium blocks tend to be very highly correlated and not give any information. In fact addition of these may increase redundancy and add noise to linking process and decrease accuracy. This is why we remove all 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 accuracy. We have added a paragraph of these points in the Discussion Section.
Excerpt From Revised Manuscript	

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

2. 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