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

RESPONSE LETTER

Formatted: Heading 3,headline,OH,3,heading 3

-- 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	

-- Ref1: The reviewer suspects that the authors are unaware that very similar work was published in 2012 --

Deleted: Ref 1

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)

	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 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: An attacker gets access to a genotype dataset of 100,000 individuals and he/she 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 link these people in the phenotype dataset to the people in the genotype dataset? He/she would perform this using our manuscript's main

Deleted: -Cox

Deleted: Attacker

Deleted: that the attacker

Deleted: identify

Deleted: within

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". In consequence, an important technical difference should be highlighted: In the linking attacks, we are utilizing the significantly correlated reproducible set of QTLs, while Im et al depends on the specific set of regression coefficients for the dataset; i.e., the attacker cannot use the QTLs from one study to perform the Im et al attack on another study.

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 uses genotypic information to compute the proposed phenotype statistic (the authors utilize the allelic dosages generated by the DNA genotyping arrays). The authors propose two additional statistics, which are also genotype based. On the other hand, our methodology is based on phenotype information; where we use the phenotypes to first perform genotype prediction, then use the predicted genotypes for linking.

ANALOGY TO NETFLIX SITUATION

Deleted: ¶

Deleted: takes the genotype based

Deleted: , e.g.,

Deleted: array based allelic dosage information in the results section.

Deleted: Our

Deleted: , however,

Deleted: using

Deleted: phenotypes

	The extremity statistic, for example, is based on the phenotypic information. In this sense, two methods use different sources of 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) 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. These differences should also be kept in mind for later as they shape and outline the differences in terms of the risk assessment and management that we delve in more detail in the following comments. 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 S6 and

Deleted: S5

	S7 to make linking attack scenario and differences with genome in a mixture identification attack scenario clearer.
Excerpt From Revised Manuscript	***Supplementary Material Section 1:*** Privacy has a multifaceted nature which can be breached under many different scenarios. In genomic privacy, the initial focus is to protect the identities of the individuals who attend genetic databases. The several first 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 study[2], or the allele frequencies[3] 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 obvious that an individual's participation in a genotyping dataset will be very highly likely within the large genotype database that is stored in the hospital of their 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.

-- 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.

[[COMBINE WITH NEXT COMMENT?]]– --

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

Deleted: S6

Formatted: Normal

Formatted: Highlight

Deleted: –

	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 and genotyping datasets. The linking attacks attacks underpin these risks. We concentrate on quantification of the leakage in these attacks and show how extremity based genotype prediction can be utilized to perform an 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 "detection of genome in a mixture" as a meaningful and generalizable framework. We believe, for the reasons we explained above, the Im et al attacks (and "detection of a genome attacks") are not generalizable to the scenario that we are presenting. 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 through a different risk management procedure than the detection attacks. It is widely accepted that the risk assessment and management methods are scarce and are in need of development. In parallel with the recently recognized needs to develop and build "measurable method for addressing privacy risk in information systems" (http://www.nist.gov/itl/201506_privacy_framework.cfm), our aim

Deleted: underpins

Deleted: a very

Deleted: identification

Deleted: in

Deleted: scenario compared

Deleted: the studies on "genome in a mixture identification".¶
¶
We thank the reviewer

Deleted: articulating on our suggestions for changing the legislative decision making processes. We are not aiming to create a panic environment. In the contrary,

	is to build analysis frameworks for the specific case of linking attacks. These frameworks help objectively quantify the risks associated with genotype-phenotype data publishing/serving. The suggestions in 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 added a section on discussion about risk management and analysis in the Supplementary Material that combines all the analysis that we presented throughout our manuscript.
Excerpt From Revised Manuscript	***Supplementary Material Section 1: 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). ***Supplemantary Material Section 5: A Basic Risk Assessment Procedure for Genotype-Phenotype 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.

Deleted: against

Deleted: Our main goal

Deleted: these

Deleted: is that the approaches for bioinformatics analysis of genomic privacy proposed by

Deleted: [[

Deleted: reworded

Deleted: legislative clauses to ensure

Deleted: this study advances on

Deleted: previous studies]]

Formatted: Normal

-- 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 well-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 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. Latanya Sweeney also demonstrated the identities of several personal genome project participants can be re-identified by linking the PGP database to the voters list in a similar fashion as above. 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 asking is: Can I characterize these people are and reveal 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, without regard to whether they would like to be protected. A dataset is k-anonymous when all the individuals that satisfy k-anonymity condition, not just a selected set of individuals. In other words, characterization of even one individual is as serious a risk as characterization of many (any person who is not a sperm donor still has the right to stay private). In our study, we are showing that the linking attacks can target and characterize a large fraction of the individuals (supported by the PPV analysis), which indicates that the linking attack has realistic levels of associated risk and can target an individual with high probability.

We have updated the manuscript and added a discussion in the Supplementary Material that summarizes the above points.

Deleted: , can

Deleted: identify who

Deleted: .

Deleted: [[

Deleted: explains

Deleted: about risk management]]

Excerpt From Revised Manuscript	***Supplemantary Material Section 2: Comparison of "Detection of a Genome in a Mixture" and "Linking Attacks" in Genomic Privacy*** One famous example of these attacks (although not in genomic context) is Latanya Sweeney's[1] 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 above[4]. 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 al[5] 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 attacks 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 techniques.

Formatted: Normal

-- 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 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

Formatted: Not Highlight

	very high fraction of individuals. These show that the extremity based linking can characterize individuals very similarly in terms of linking accuracy as the approach proposed by Schadt. 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 30 data points (15 expression and genotype values) per eQTL to train the model robustly and accurately. Our method requires roughly 30 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 extremity based genotype prediction approach decreases the linking accuracy of the attacker significantly compared to the attack by Schadt et al. 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	***Section 2.4:*** We evaluated whether the attacker can estimate the reliability of the linkings. This may potentially increase the effectiveness of the linking and increase the risk associated with linking attacks because the attacker can estimate reliability of the linkings and choose the ones that are more likely to be correct. This increases the risk associated with the linking attacks because although he/she may not have a high overall accuracy of linkings, the high ranking linkings may be much higher in accuracy. We observed that the measure we termed, *first distance gap*, denoted by d_{2-1} (See Methods), serves as a good reliability estimate for each

Deleted: model based

Deleted: et al.

Deleted: 60

Deleted: 30

Deleted: 60

Deleted: a model-free approach (

Deleted:)

Deleted: model based

linking. For a given linking, d_{2-1} is the difference between the genotype distances of the 1st closest and 2nd closest individuals to the predicted genotypes. When the linking is incorrect, we observed that d_{2-1} is very likely to be smaller than the distance difference when the linking is correct.

To evaluate this measure further, we computed the positive predictive value (PPV) versus sensitivity of the linkings of individuals in the testing set with changing d_{2-1} threshold. For this, we first computed d_{2-1} for each linking, then filtered the linkings that did not satisfy the threshold. Then we computed PPV and sensitivity of the linkings (See Methods), which is plotted in Fig 6b. It can be seen that the PPV of linkings can get very high at the same time with high sensitivity. For example, the attacker can link around 79% of the individuals at a PPV higher than 95%. The random sorting of the linkings, on the other hand, have significantly lower PPV (cyan in the plots) at the same sensitivity levels. These results suggest that the attacker can increase the potential risk (accuracy of linkings) of the attack by focusing on a slightly smaller set of linkings with high reliability.

Supplementary Material Section 3:

It is worth comparing the accuracies of extremity attack and the model based attack proposed in Schadt et al[6]. The model based 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. 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 model based 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.

Formatted: Normal

-- 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	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	

Deleted: [[Closing statements, not to be included]]¶

-- 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.

Formatted: Not Highlight

Formatted: Not Highlight

	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 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	***Section 2.4:*** For ease 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 rank correlation. Thus, this process can be generalized as a rank correlation based prediction. ***Supplementary Section 1: Motivation on Extremity Attack: Outlier Attacks in Privacy*** 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. 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-anonymization[1], 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 predicting the genotypes and distinguishing 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 accessible 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.

Formatted: Normal

-- 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, while the third individual does not get resolved correctly because there are not enough extreme identifiers to pinpoint that individual. We believe this figure helps illustrate better the idea that phenotypic extremity can lead to privacy breaches in linking attacks.
Excerpt From Revised Manuscript	***Supplementary Material Section 6: 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. 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

Formatted: Not Highlight

Deleted: .

Deleted: gene expression

	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 genotype 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.

Formatted: Normal

-- 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	We thank the reviewer for careful
Excerpt From Revised Manuscript	

Deleted: [[Just

Deleted: introduction]]

-- 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 this 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 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 simplified approach can reach very high linking accuracy. As explained by the reviewer, 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, in k-anonymization,

Formatted: Not Highlight

Formatted: Not Highlight

Deleted: model-free

Formatted: Not Highlight

Deleted: reviewers

Deleted: aim

	one aims to protect published datasets utilizing by imposing statistical indistinguishability to the rare and extreme features using different methods (e.g. censoring, swapping data, adding noise). 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 genotype estimation procedure, the extremity based attack can be implemented easily, which makes it fairly accessible and realistic in most situations. We added a motivation for the extremity attack in the Supplemantary Material Section 1.
Excerpt From Revised Manuscript	***Supplementary Material Section 1: Motivation on Extremity Attack: Outlier Attacks in Privacy*** 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. 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-anonymization[1], 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 predicting the genotypes and distinguishing 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 accessible and realistic in most situations.

Deleted: at

Deleted: of

Deleted: by

Deleted:) so as to protect it.

Deleted: model-free

Formatted: Normal

-- Ref3: The experimental validation needs to be improved.---

Deleted: . [[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. 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	***Section 2.4:*** The previous results show that extremity based linking attacks are highly effective when the eQTLs are identified and linking attack is performed using the same expression and genotype datasets. In order to assess the accuracy when the eQTLs are computed and tested on different datasets, we divided the dataset into a training and a testing dataset. The training dataset, of 210 individuals, is used to discover the eQTLs, using Matrix eQTL[47] method (See Methods Section for details). The testing dataset, of 211 individuals, is utilized for assessing the accuracy of linking. Figure 6a shows the linking accuracy for individuals in testing dataset. The accuracy is very high, around 95%, which suggests that extremity based linking attacks are potentially effective when the datasets where eQTLs are identified do not match the data being tested. This is an important aspect of genotype prediction based linking attacks, as they exploit the generalizability of the correlations between phenotypes and genotypes.

Deleted: tools

Formatted: Normal

-- Ref3: there are a number of biases that can reduce accuracy. --

Deleted: –[[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 divided the samples into 5 populations. For each population, we identified the population specific eQTLs (using Matrix eQTL) then tested the matching accuracy on the expression values of other populations. We observed that for European populations, the linking accuracies are general very high. When the eQTLs are trained on the African population, the accuracies drop to significantly. This result can be attributed to the fact that the

Deleted: have selected 3

Deleted: : GBR, CEU, and YRI.

Deleted: 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.

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.

We also studied how the accuracy gets affected when eQTLs are identified in different tissues than the tested samples. 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.

Deleted: 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.¶
¶

Deleted: from

Excerpt From Revised Manuscript

Results Section 2.4: Individual Characterization using Extremity based Genotype Prediction

We also studied how the linking accuracy changes when the training and testing datasets are measured in different populations. For this, we used the 1000 Genomes Project sample information and divided the GEUVADIS samples into 5 populations. Then we used each population's samples to discover the population specific eQTLs, then used the other populations to test the linking accuracy. Table S1a shows the accuracies in each case. It can be seen that when the eQTLs are disovered in European populations (CEU, GBR, TSI, FIN), the linking accuracies are very high (higher than 95%). When the eQTLs are discovered in YRI (African) population, the linking accuracies are smaller in European populations. Similarly, when eQTLs are discovered on European populations, the linking accuracy in YRI sample is relatively smaller. These results illustrate that extremity attack can still be effective when eQTLs are identified in populations that are genetically close to the population(s) of testing sample and decrease when the discovery and testing populations are diversified. We next studied scenario where the eQTLs are identified in tissues that are different from the tissues on which the expression data is generated. For this, we used the eQTLs that are identified by GTex Project [38]. We downloaded the eQTLs for 6 tissues and performed the linking attack. The results are shown in Table S1b. The accuracy is highest for Whole Blood eQTLs, which is 88%. This is expected since the expression

Formatted: Normal

	levels in GEUVADIS project are measured in blood cell lines. The accuracy is smallest for Muscle Skeletal eQTLs, which is 76%.

Formatted: Font color: Text 1

-- Ref3: It would also be interesting to understand how these attacks scale with data set size.---

Deleted: . [[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 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	***Results Section 2.4: Individual Characterization using Extremity based Genotype Prediction*** An important practical question is how well the linking accuracy changes with increasing genotype data size. In order to evaluate this, we simulated the genotypes of the eQTLs (discovered in the training set) for 100,000 individuals. The 100,000 simulated individuals are then merged with the testing dataset of 211 individuals to build the large testing dataset. We then performed the extremity attack using the expression levels of the testing dataset and linked them to the merged testing dataset of size 100,211 individuals. The linking accuracy is plotted in Fig 7a with changing eQTL selection criteria. The linking accuracy is very high (Around 96%). This result suggests that the extremity attack can be extended to a large testing sample set. Figure 7b shows the sensitivity versus PPV (with changing first gap distance) for the eQTLs for which the overall linking accuracy is 70% (Yellow dashed lines on Fig. 7b). It can be seen that the attacker can link around 55% of the individuals with PPV higher than 95%. Only the remaining 15% are predicted with accuracy lower than 95%. … We also studied whether having close relatives in the genotype dataset affects the accuracy. To test this, we used the expression and genotype data from 30 CEU trios (mother-father-child) from available from HAPMAP project[48,49]. We first identified the eQTLs from the 90 individuals and performed linking over the same individuals. We then computed the average rank of the (non-self) close relatives in each linking. For example, when the tested individual is a father or mother, we computed the rank of the individual child and if the tested individual is a child, we computed the rank of his/her mother and father. We also selected, for each tested individual, a random individual and computed his/her rank in the linking. The distribution of the ranks are shown in Fig 8. It can be seen that the ranks of the related individuals are significantly shifted to smaller values compared to random individuals. This result shows that the close relatives can get linked to each other. This result indicates that the individuals that are close relatives may potentially be confused with each other. While the correct person may not get characterized, the attacker can still reveal sensitive information about the individual's family, which might extend the reach of privacy breach and cause privacy concerns for the family.

Formatted: Normal

Formatted: Font color: Text 1

-- 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 ?---

Deleted: ? [[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. 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	***Results Section 2.4: Individual Characterization using Extremity based Genotype Prediction*** We evaluated whether the attacker can estimate the reliability of the linkings. This may potentially increase the effectiveness of the linking and increase the risk associated with linking attacks because the attacker can estimate reliability of the linkings and choose the ones that are more likely to be correct. This increases the risk associated with the linking attacks because although he/she may not have a high overall accuracy of linkings, the high ranking linkings may be much higher in accuracy. We observed that the measure we termed, *first distance gap*, denoted by d_{2-1} (See Methods), serves as a good reliability estimate for each linking. For a given linking, d_{2-1} is the difference between the genotype distances of the 1st closest and 2nd closest individuals to the predicted genotypes. When the linking is incorrect, we observed that d_{2-1} is very likely to be smaller than the distance difference when the linking is correct. To evaluate this measure further, we computed the positive predictive value (PPV) versus sensitivity of the linkings of individuals in the testing set with changing d_{2-1} threshold. For this, we first computed d_{2-1} for each linking, then filtered the linkings that did not satisfy the threshold. Then we computed PPV and sensitivity of the linkings (See Methods), which is plotted in Fig 6b. It can be seen that the PPV of linkings can get very high at the same time with high sensitivity. For example, the attacker can link around 79% of the individuals at a PPV higher than 95%. The random sorting of the linkings, on the other hand, have significantly lower PPV (cyan in the plots) at the same sensitivity levels. These results suggest that the attacker can increase the potential risk (accuracy of linkings) of the attack by focusing on a slightly smaller set of linkings with high reliability. ***Methods Section 4.6: First Distance Gap Statistic For Linking Reliability Estimation*** Following the previous section, the attacker computes, for each individual, the distance to all the genotypes in genotype dataset, then identifies the individual with smallest distance. Let $d_{j,(1)}$ and $d_{j,(2)}$ denote the minimum and second minimum genotype distances (among $d^H(\tilde{\boldsymbol{v}}_{\cdot,j}, \boldsymbol{v}_{\cdot,\boldsymbol{a}})$ for all $\boldsymbol{a}$) for jth individual. We propose using the difference between these distances as a measure of reliability of linking. For this, the attacker computes following difference: $$d_{1,2} = d_{j,(2)} - d_{j,(1)} \quad (23)$$ 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. 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.

Formatted: Normal

-- 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 made updates the results section 2.2 to clarify how Section 2.2 relates to the other sections. In summary, the quantification methodology that is presented in Section 2.2 evaluates, for a given list of eQTLs, how much information leakage is expected at different levels of predictability. This way, the data releasing mechanisms can quantify the risks associated with releasing the QTL datasets. The following sections 2.3 and 2.4 focus on how the genotype-phenotype linkages can be made. 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. In addition, we added Figures S5, S6, and S7 that serve to clarify several general aspects of how the scenario that we are presenting, the technical details of linking attack, and also a specific example of how extremity of phenotypes are utilized in a linking attack. We also included a number of experimental validations in the Results Section. We believe these updates clarify how different Sections fit with each other in the manuscript and concretize the experimental validations.
Excerpt From Revised Manuscript	***Supplementary Material Section 1: Motivation on Extremity Attack: Outlier Attacks in Privacy*** 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. 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

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Deleted: conclusion

Deleted: In addition, we

Deleted: updatese

	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 predicting the genotypes and distinguishing 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 accessible and realistic in most situations. ***Section 2.2: Quantification of Tradeoff between Correct Predictability of Genotypes and Leakage of Individual Characterizing Information*** The presented quantification procedure can be utilized for evaluating the risk of information leakage while releasing QTL datasets. For example, the QTLs to be released can be assessed in terms of the characterizing information leakage versus the predictability so as to estimate the size and risk of a linking attack (Fig S8).

Formatted: Normal

-- 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	***Page 4:*** … For example, GTex Project hosts a sizable set of eQTL dataset from multiple studies where the users can view in detail how the genotypes and expression levels are associated [10,38]… ***Page 5:*** … Thus, each time he/she predicts a new genotype, he/she will encounter a tradeoff between the number of genotypes that can be predicted correctly versus the cumulative correctness of all the predicted genotypes…

Formatted: Normal

-- 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 –--

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 reviewer recommends this necessary comparison between the methods. It is first necessary to note that both methods perform linking attacks, so the genotype predictions are performed as middle steps. In fact the source code that we received from the Schadt et al does not give as output the genotype predictions. We therefore compared the linking accuracies of the two methods. For comparison, we 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. 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.

Deleted: [[Schadt Comparison]]

Formatted: Not Highlight

Deleted: mathods is necessary. For this, we first requested

Deleted: of model based method

Deleted: .

Deleted: utilized the

Deleted: For training Schadt et al's method, we used the training set, and evaluated the accuracy of linking on the

Deleted: set. We utilized different number of eQTLs to compare

Deleted: accuracy of methods with different markers.

Deleted: results

Deleted: SXX

Deleted: even at relatively smaller number of markers.

Deleted: model-free

Deleted: (

Deleted:)

Deleted: model based

	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. We would like to emphasize that these comparison results should be interpreted with caution. Our aim in this comparison is to show that the extremity attack has comparable accuracy to the training based attack. When the data is to be published or served, these attacks must be considered altogether (rather than choosing the best performing one) since they represent different paths to a privacy breach. We added the Supplementary Section 3 to report the Comparison Results.
Excerpt From Revised Manuscript	***Supplementary Section 3: Comparison of Extremity based Linking Attack Accuracy with Linking Attack in Schadt et al*** It is worth comparing the accuracies of extremity attack and the attack proposed in Schadt et al[6]. This model based 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. 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

Deleted: [[In order to compare the amount of input to each algorithm, we ran Schadt et al algorithm with different input sizes]]

	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.

Formatted: Normal

-- 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	**Methods Section 4.7: *eQTL Identification on Training Sets with Matrix eQTL*** For identification of eQTLs, we used Matrix eQTL[47] 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. 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 was larger than 5%. We finally removed the redundancy by ensuring that each gene and each SNP is used only once in the eQTL final list. ***Methods Section 5: Datasets*** The normalized gene expression levels for 462 individuals and the eQTL dataset are obtained from gEUVADIS mRNA sequencing project[54]. The eQTL dataset contains all the significant (At most 5% false discovery rate cutoff) 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 genotype, gender, and population information datasets for 1092 individuals are obtained from 1000 Genomes Project [12]. For 421 individuals, both the genotype data and gene expression levels are available.

Formatted: Normal

-- 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	The reviewer is raising a good point. The problem is with the accuracy computation: At the high absolute correlation threshold, there are very small number of SNPs. This makes the genotype accuracy (the fraction) unstable. Although we expect very high accuracy, 1 wrong prediction out of a small number in the fraction pulls accuracy down significantly. The decrease at 0.7 threshold is reflecting this behavior. We added a sentence explaining this in the Results Section.
Excerpt From Revised Manuscript	**Results Section 2.4:** The slight decrease of genotype accuracy at correlation thresholds higher than 0.7 is caused by the fact that the accuracy (fraction of correct genotype predictions within all genotypes) is not robust at very small number of SNPs. Although we expect very high accuracy, even one wrong prediction among small number of total genotypes decreases the accuracy significantly.

Deleted: [[This

Deleted: actually

Deleted: question, the

Deleted: Very

Deleted: make

Deleted: very

Deleted: , although

Deleted: makes it go

Deleted: . I will look into

Deleted: little more and make sure my explanation is correct. Should be just clarification and update.]]

Formatted: Normal

-- 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	The reviewer is bringing up an interesting question about whether the tool supplies the predicted genotypes with extremity based linking. The genotypes predictions are used by our tool to perform compute the genotype distances and perform linking. They are, thus, only intermediate data that is used by the tool, so we do not supply the extremity based genotype predictions. The output from the tool is one tab delimited file that contains vulnerability information for each sample. In each row, there are XX columns that correspond to following: … whether each individual is vulnerable and also the first distance gap statistic corresponding to the individual's linking. We updated the README file to clarify these.
Excerpt From Revised Manuscript	

Deleted: [[Will update

Deleted: .]]

-- Ref4: can your tool address this by being able to use imputed genotypes?–--

Deleted: ? [[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 genotyped SNPs in terms of characterizing information content they provide, so our tool 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 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

Deleted: other

Deleted: is

	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	***Supplemantary Section 4: 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 our tool 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 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 accuracy.

Formatted: Normal

Formatted: Heading 1,Heading 1 Char

REFERENCES

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