Response to reviewers for “Analysis of Information Leakage in Phenotype and Genotype Datasets”

Response Letter

-- Ref1: Introduction –--

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

Formatted Table

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

Reviewer Comment	The reviewer suspects that the authors are unaware that very similar work was published in 2012 with a fair amount of discussion and attention showing the core principles of this work on eQTL under what the reviewer considers a more broadly applicable mathematical framework. While the author's focus on using extremes or outliers as information sources has some unique aspects, the innovative work was in the original work by Im, Cox and colleagues in the American Journal of Human Genetics. Indeed it was a complete surprise at that time to those who read and went to meetings where this work was presented. I am sure the authors of this paper are in no doubt aware that Dr. Cox leads one of the largest NIH funded efforts putting forth eQTL data. Thus its reassuring to see that her team prospectively put for the careful analytical consideration of risk for the community to vet at that time in 2012.
Author Response	We thank the reviewer for pointing us to the Im and Cox et al 2012 paper, which is a very important study. We have carefully reviwed the Im, Cox et al paper in detail. Firstly, in the Im, Cox et al paper, the problem that is addressed by the authors is different from our manuscript: In 2012 paper, the authors address “detection of a genome in a mixture” in the setting of GWAS studies: When the attacker gains access to the allelic dosages (from genotyping arrays or DNA sequencing) or 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. We are undertaking the “Linking Attack” problem. In this attack, the attacker aims at characterizing as many individuals as possible. In our setting, as described in Figure 1, 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

Formatted Table

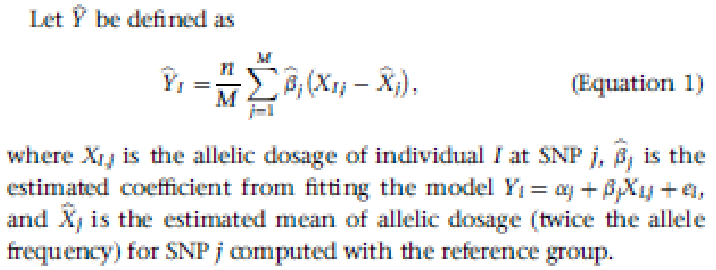
Deleted: studied

mixture”) but rather tries to characterize as many individuals as possible. The accuracy and size estimation is the main focus of our study. In Section 2.2, we are aiming to jointly quantify the correct predictability of genotypes versus the amount of characterizing information leakage. Im-Cox et al do not address the issue of “linking”, which is the 3rd step in the individual characterization.

[[Genotype Prediction, Model inversion aspect, 3 steps approach, how this helps analyze the linking attack problem easier]]

This final point is important for following reason: Let’s consider that our study is redundant because of Im-Cox et al’s study. This would suggest that an attacker could utilize Im-Cox et al attack to perform a linking attack. However, if an attacker tried to perform the linking attack as per Im-Cox et al study, the input and outputs of the method does not support a linking attack: The attacker could certainly utilize the Im-Cox 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-Cox 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-Cox et al’s study.

Secondly, In Im-Cox et al perform classification of class membership (Participated/Not participated) using a statistical test that uses a statistic defined as following (taken from the 2012 paper):

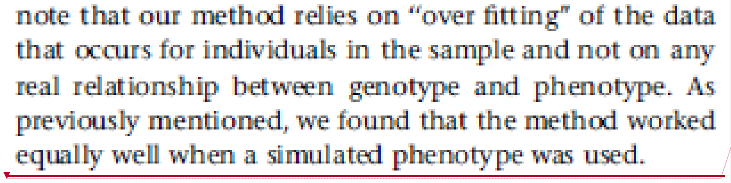


Let $\hat{Y}$ be defined as

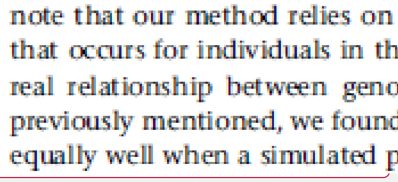
$$\hat{Y}_I = \frac{n}{M}\sum_{j=1}^{M}\hat{\beta}_j (X_{Ij} - \hat{X}_j), \qquad \text{(Equation 1)}$$

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

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

	[[Following may not be very clear, remove maybe?]] [[Another technical difference between the two methods is that the statistical test in Im-Cox 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. Here, the information leakage happens through this "biological channel", unlike the Im-Cox study, where the leakage happens through a "statistical channel".]]* [[Following is a rather technical point, and may sound strong, not sure if we should put this here: Im-Cox et al attack works well when M>>n>>1. Authors use M/n=300 in their experiments. For eQTLs, however, M/n=300 means, for GEUVADIS dataset where n=462, this value turns out to be 138,600 regression coefficients for each gene. From the available files, the largest M for any gene goes upto 20,000 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. Moreover, the attacker also needs to ensure M>>n*>>1; which indicates that these have to be met with respect to the reference population. Considering n*=1092 as in 1000 Genomes, the required number of regression coefficients are even much higher. These issues render the attack almost non-applicable on GEUVADIS dataset.]] We believe this confusion is caused on our part as we may not have to clarified the attack setting. [[We have added citation to Im-Cox et al paper and made updates to the introduction and methods section to ensure that our manuscript is clearer]]
Excerpt From Revised Manuscript	

Deleted:



note that our method relies on
that occurs for individuals in th
real relationship between geno
previously mentioned, we found
equally well when a simulated p

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

Reviewer Comment	Again, a major aspect of this 2012 work was indeed privacy risk via eQTL, and indeed at that time it was a major shock to myself and other colleagues how powerful eQTL data really can be. In comparison of the two papers, the 2012 seems focused on a broader problem building from eQTL in line with Nature Methods as premier journal to publish methodological firsts. The review views the incremental advancements over the 2012 paper do not support the far-reaching conclusions that the work by Harmanci and Gerstein for changing legistlative decision making process in a way that the Im et al paper did not. I remain more impressed to see how Cox and colleagues in 2012 provider a broader framework and a bit stunned that p-values and odds ratios from enough SNPs limit absolute privacy. This generalizable framework intuitively makes sense - when asking one question about a person's membership in a cohort can we use thousands and thousands of correlated measurements to infer correctly the answer. The privacy risk management issue covered elsewhere then is towards what is the probability of this impacting a specific person's privacy.
Author Response	We thank the reviewer for articulating on our suggestions for changing the legislative decision making processes. We believe that our study supports and advances the results of Im-Cox et al and many other authors' studies. Our study concentrates on characterizability of individuals in a world where the biomedical phenotyping datasets will significantly increase in number. We believe that linking attacks represent a source of potential privacy breach that may be exacerbated with these datasets. Because of reasons we explained above, we believe that our study is sufficiently different from studies of Homer et al, Im-Cox et al, and many other studies on "detection of a genome in a mixture". [[We have reworded the legislative clauses to ensure that this study advances on all the previous studies]]
Excerpt From Revised Manuscript	

Formatted Table

-- 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	[[The reviewer finds our scenario anecdotal and unrealistic. To be honest, he is being a little bullyish with all the arguments on “risk management”, which I think he is using in the wrong context here. What he comes down to at the end is that he does not think that our scenario is reasonable. What I don’t get is this scenario is utilized by Schadt et al in 2012, too. Counter argument: 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. In addition, if this is just an anecdotal/non-practical example, how can one explain why Netflix was sued (https://en.wikipedia.org/wiki/Netflix_Prize#Privacy_concerns) over the privacy concerns that stem from the linking attack performed by researchers who linked the IMDB records and Netflix Prize competition database to reveal identities of Netflix users? Similarly, Sweeney’s public stunt which characterized the governor of Massacusetts, in addition to many other individuals, by linking

Formatted Table

Deleted: https://en.wikipedia.org/wiki/Netflix_Prize#Privacy_concerns

	the voter registration list to the Group Information Comission using several common columns in these databases. I agree that our study is not the whole story about privacy but it surely is an important aspect of it. Why does the reviewer think that our scenario is not reasonable? Are we wording our legislative propositions too strongly?]]
Excerpt From Revised Manuscript	

-- 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 need to come up with a way to evaluate the PPV; given the predictions that we made, what fraction of the predictions are correct.]] One argument: that we can make is that extremity based linking is fairly accurate; thus PPV can be estimated]]

Formatted Table

	[[We can set a threshold on the predicted genotype-matched genotype distance and reject some of the linkings to control our false positive rate. This way we would have a way to control PPV.]]
Excerpt From Revised Manuscript	

-- 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 can also remove this, I guess he is being extremely sarcastic, as generally he was in his review. I am pretty sure this is Y. Erlich. It resembles his style of writing from twitter/blog posts.]]
Excerpt From Revised Manuscript	

Formatted Table

-- 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

Formatted Table

	(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 here. This is here to be complete. Probably going to probably 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	[[Rephrasing, etc]] We understand that the reviewer find extremity based genotype prediction unclear. The extremity based prediction exploits the outliers; i.e, the outliers in the expression levels of eGenes are associated with the outliers in the genotypes of eSNPs, i.e, the homozygous genotypes. The heterozygous genotypes of eSNPs, however, do not co-incide with the extremes of the expression levels, i.e., they co-incide with the mean level expression levels. Thus, we do not assign the heterozygous genotype in the genotype prediction. We clarified the explanation of genotype prediction by extremity attack in the Results Section.
Excerpt From Revised Manuscript	

Formatted Table

Formatted: Highlight

Deleted: It should be noted that the extremity attack

-- 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	
Excerpt From Revised Manuscript	

Formatted Table

Formatted: Highlight

-- Ref3: Introduction –--

Reviewer Comment	Genomic privacy is an increasingly important direction of research. One of the aspects of work on genomic privacy has focused on ways to breach privacy by linking different kinds of data. This paper presents an attack that can be used to link a phenotype (in their specific case, gene expression) to a genotype and possibly to other identifying information. The study presents simulations to show the feasibility of this attack. The authors consider the following setup: an attacker has access to an individual genotype (this could be part of a larger dataset), a dataset of individual-level gene expression (but no genotypes) and a list of variants that are known to affect expression of specific genes. The attack consists of predicting the genotypes at the list of expression SNPs corresponding to the the gene expression data and then testing if the target individual genotype matches any of the predicted genotypes. They consider two variants. In the first (2.3), the attacker needs a prediction model to predict genotypes from expression. This, in turn, implies that the attacker would need access to data where individuals have genotypes as well as gene expression. In the second (2.4), termed Extremity-based genotype prediction, the attacker only has access to the correlation between genotype and gene expression. The authors show that for both variants, a large fraction of individuals (>=95%) are vulnerable as assessed by simulation experiments on the GEUVADIS dataset.
Author Response	[[Just the introduction]]
Excerpt From Revised Manuscript	

Formatted Table

-- 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	[[I am not sure how we can explain better that extremity based attack is realistic. In addition, Schadt et al attack assumes that there is access to the population panels, etc so that the attacker knows the a-priori distribution of genotypes]]
Excerpt From Revised Manuscript	

Formatted Table

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

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

Formatted Table

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

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

Formatted Table

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

Reviewer Comment	c. It would also be interesting to understand how these attacks scale with data set size. For example, how feasible is this attack within a dataset of 100,000 genotypes that are now being generated. Another interesting question is whether the method can discriminate close relatives that are likely to be present in large datasets.
Author Response	We agree that these are important points for illustrating the general applicability of the extremity attack. To evaluate how the matching genotype sample size affects the accuracy, we simulated 100,000 individuals using the 1000 Genomes genotype frequencies for the eQTL SNPs. The eQTLs are identified from the training set of 210 individuals. The 100k simulated individual genotypes are then merged with the 211 testing sample set to generate the 100,211 individual sample set. We then used the expression levels (from GEUVADIS dataset) for the test sample and performed the extremity based attack on this larger dataset to check the characterizability of individuals in testing set. We observed that the matching accuracy is very high, around 99%. This result indicates that extremity attack can potentially be effective in very large sample sizes. [[Close relatives: Unfortunately, we do not have the relationship information in 1kG dataset, we can only comment on this by saying that since we use predicted genotype distance as the metric of choice for linking, it would not be able to discriminate the close relatives well. But on the other hand, predicting the family of a person correctly would still be useful for the attacker In many circumstances.]]
Excerpt From Revised Manuscript	

Formatted Table

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

Reviewer Comment	d. The authors declare an individual to be vulnerable if pred_j = j. This is only a first step in documenting its utility. For a realistic attack, the attacker would need some threshold on the distance function to decide if a test individual is linked to a given predicted genotype. How should this threshold be chosen ? Does it

Formatted Table

	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” (1st-to-2nd distance difference) 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 1st-to-2nd distance difference and select the ones that have high genotype distance to focus the attack on highly reliable linkings. [[Also, assign a notation for this distance measure]] [[We need a supplementary figure to illustrate this: Each linking is basically computing distance to all the individuals in the genotype dataset.]]
Excerpt From Revised Manuscript	

Deleted: 90

Deleted: XX

Deleted: 90

Deleted: 86

Deleted: predicted

Deleted: 98

Deleted:]]

-- 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	[[Rephrase, move, clarify]]
Excerpt From Revised Manuscript	

Formatted: Highlight

Formatted Table

Formatted: Highlight

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

Formatted Table

-- 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	

Formatted Table

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

Reviewer Comment	The authors rightfully cite a previous publication (Schadt et al, Nature Genetics 2012) that relates to their study, as they also developed a method to predict genotypes from eQTL gene expression. Nevertheless, the authors do not compare the performance of their method with this previous one. This should be done, as to assess the importance of this new method with the current state-of-the-art tools addressing the same issue.
Author Response	[[The problem here is that Schadt et al does not provide source code. I can try and do my best to change the first part of the paper to match Schadt et al's model based prediction, using part of the data for "model" building, and other parts for testing. This is also useful since Ref3 also asked something similar. On the other hand, this may not be a fair comparison since it may not capture all the details of Schadt et al. We can thus just spin it by saying that we the model based method (as an alternative to Schadt et al's method) and the extremity based prediction and model based prediction are very similar in performance.]]
Excerpt From Revised Manuscript	

Formatted Table

-- 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, 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 data section in detail to explain how eQTL selection was performed.

Formatted Table

Excerpt From Revised Manuscript	

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

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

Formatted Table

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

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

Formatted Table

-- Ref4: can your tool address this by being able to 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. Currently our tool does not consider the dependencies between eQTL genotypes. This is why we are considering the list of eQTLs that are non-redundant in terms of both genes and SNPs, i.e., a gene/SNP can be seen once in the eQTL dataset. This enables decreasing dependencies and maximize the amount of characterizing information. One could, however, evaluate the dependencies between genotypes

Formatted Table

	and build a more complicated model of genotype prediction (step 2) and also include this information in linking (step 3). We have added a paragraph of these points in the Discussion Section.
Excerpt From Revised Manuscript	