Quantification of private information leakage from phenotype-genotype data: linking attacks

Arif Harmanci[1,2] & Mark Gerstein[1–3]

Studies on genomic privacy have traditionally focused on identifying individuals using DNA variants. In contrast, molecular phenotype data, such as gene expression levels, are generally assumed to be free of such identifying information. Although there is no explicit genotypic information in phenotype data, adversaries can statistically link phenotypes to genotypes using publicly available genotype-phenotype correlations such as expression quantitative trait loci (eQTLs). This linking can be accurate when high-dimensional data (i.e., many expression levels) are used, and the resulting links can then reveal sensitive information (for example, the fact that an individual has cancer). Here we develop frameworks for quantifying the leakage of characterizing information from phenotype data sets. These frameworks can be used to estimate the leakage from large data sets before release. We also present a general three-step procedure for practically instantiating linking attacks and a specific attack using outlier gene expression levels that is simple yet accurate. Finally, we describe the effectiveness of this outlier attack under different scenarios.

Genomic privacy has recently emerged as an important issue, particularly in light of the surge in biomedical data acquisition[1–3]. Molecular phenotype data sets, like functional genomics measurements, substantially grow the list of quasi-identifiers[4], which may lead to re-identification and characterization of individuals[4–6]. In general, statistical analysis methods are used to discover genotype-phenotype correlations[7,8], which can be used by an adversary to link the entries in genotype and phenotype data sets, thereby revealing sensitive information. The availability of a large number of correlations increases the possibility of such linking[9,10].

Protecting the privacy of participating individuals has emerged as an important issue in genotype-phenotype association studies. Several studies have addressed the problem of detecting whether an individual with a known genotype has participated in a study[11], raising privacy concerns[12–15]. We refer to these systematic breaches as 'detection of a genome in a mixture' attacks (**Supplementary Fig. 1**). However, as the number and size of phenotype and genotype data sets increase, the detection of individuals in those sets will become irrelevant, as the genotype or phenotype information for each individual will already be stored in a data set (i.e., participation will already be known). This opens up a new route to breaches of privacy: an adversary can now aim at cross-referencing multiple, seemingly independent genotype and phenotype data sets and pinpointing an individual to characterize his or her sensitive phenotypes. It is most certain that as personal genomics gains more prominence, attackers will aim at linking different data sets in order to reveal sensitive information. We refer to this type of attack here as a 'linking attack'[4,5]. One well-known example is the attack that matched the entries in the Netflix Prize Database and the Internet Movie Database. For research purposes, Netflix released an anonymized data set of movie ratings from thousands of viewers. This data set was assumed to be secure, as the viewer's names were not included. However, Narayanan *et al.*[16] used the Internet Movie Database, in which the identities of many users are public but only some of their movie choices are available, and linked it to the Netflix data set. This revealed the identities and personal movie-preference information of many users in the Netflix data set. The importance of this attack is underpinned by the fact that both Netflix and the Internet Movie Database have millions of individual users, and any individual who is in one data set is very likely to be in the other data set. As the size and number of genotype and phenotype data sets increase, the number of potentially linkable data sets will increase (**Supplementary Note**).

RESULTS

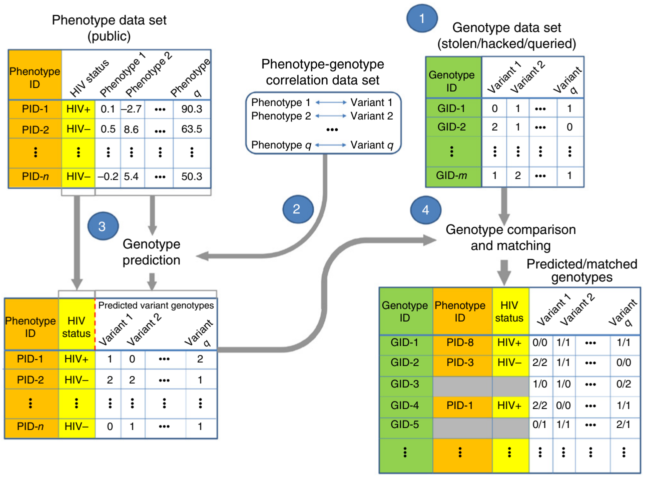
Linking attack scenario

In linking attacks, the attacker aims at characterizing sensitive information about a set of individuals in a stolen genotype data set (**Fig. 1**). Let us assume that the attacker is a woman. For each individual, she aims at querying the publicly available anonymized phenotype data sets to characterize, for example, HIV status. For this she uses a public quantitative trait locus (QTL) data set that contains genotype-phenotype correlations. She statistically predicts genotypes using the phenotypes and QTLs. Then she compares the predicted genotypes to the genotype data set and links the entries that have good genotype concordance. The sensitive information for the linked individuals is thus revealed to the attacker.

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA. [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA. [3]Department of Computer Science, Yale University, New Haven, Connecticut, USA. Correspondence should be addressed to M.G. (mark@gersteinlab.org).

RECEIVED 4 JUNE 2015; ACCEPTED 13 NOVEMBER 2015; PUBLISHED ONLINE 1 FEBRUARY 2016; DOI:10.1038/NMETH.3746

Figure 1 | Illustration of a linking attack. The publicly available anonymized phenotype data set (top left) contains *q* phenotype measurements and HIV status for a list of *n* individuals. The genotype data set (top right) contains the variant genotypes for *m* individuals whose identities are known. The genotype-phenotype correlation data set contains *q* phenotypes, *q* variants and their correlations. The attacker predicts the variant genotypes for *n* individuals in the phenotype data set using the phenotype measurements. The attacker then links the phenotype data set to the genotype data set by matching the predicted genotypes to the genotype data set. The linking potentially reveals the HIV status for the subjects in the genotype data set. The columns for ID and HIV status are shaded to illustrate how the linking combines the entries in the two original data sets. Data in the unshaded columns are not used for linking.



Phenotype data set (public)

Phenotype ID	HIV status	Phenotype 1	Phenotype 2		Phenotype q
PID-1	HIV+	0.1	–2.7	•••	90.3
PID-2	HIV–	0.5	8.6	•••	63.5
⋮	⋮	⋮	⋮	⋮	⋮
PID-*n*	HIV–	–0.2	5.4	•••	50.3

Phenotype-genotype correlation data set

Phenotype 1 ⟷	Variant 1
Phenotype 2 ⟷	Variant 2
•••	
Phenotype *q* ⟷	Variant *q*

Genotype data set (stolen/hacked/queried)

Genotype ID	Variant 1	Variant 2		Variant q
GID-1	0	1	•••	1
GID-2	2	1	•••	0
⋮	⋮	⋮	⋮	⋮
GID-*m*	1	2	•••	1

Predicted variant genotypes

Phenotype ID	HIV status	Variant 1	Variant 2		Variant q
PID-1	HIV+	1	0	•••	2
PID-2	HIV–	2	2	•••	1
⋮	⋮	⋮	⋮	⋮	⋮
PID-*n*	HIV–	0	1	•••	1

Predicted/matched genotypes

Genotype ID	Phenotype ID	HIV status	Variant 1	Variant 2		Variant q
GID-1	PID-8	HIV+	0/0	1/1	•••	1/1
GID-2	PID-3	HIV–	2/2	1/1	•••	0/0
GID-3			1/0	1/0	•••	0/2
GID-4	PID-1	HIV+	2/2	0/0	•••	1/1
GID-5			0/1	1/1	•••	2/1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

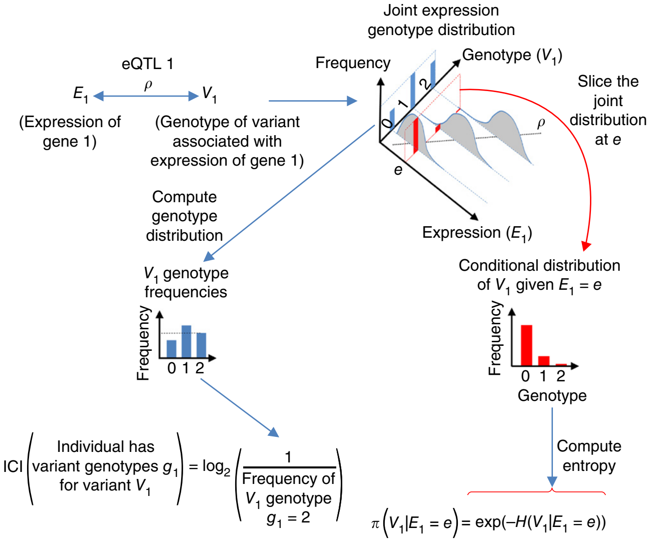
Among the QTL data sets, the abundance of eQTL data sets makes them most suitable for linking attacks. In an eQTL data set, each entry contains a gene, a variant and a correlation coefficient (denoted by ρ) between expression levels and genotypes (**Fig. 2** and **Supplementary Fig. 2**). For reporting results and for performing mock linking attacks, we used the eQTLs and gene expression levels from the GEUVADIS Project[17] and the genotypes from the 1000 Genomes Project[18] as representative data sets.

Genotype predictability and information leakage

We assume that the attacker will behave in a way that maximizes her chances of correctly characterizing the greatest number of individuals. Thus, using the phenotype measurements, she will try to predict the genotypes for the largest set of variants that she believes she can predict correctly. The most obvious way is by first sorting the genotype-phenotype pairs in order of decreasing strength of correlation and then predicting the genotypes for each variant (**Supplementary Fig. 3**). The attacker encounters a tradeoff: as she goes down the list, more individuals can be characterized (as more genotypes can characterize more individuals), but it also becomes more likely that she will make an error in the prediction, because the number of genotype-phenotype correlations will decrease. This tradeoff can also be viewed as the tradeoff between precision (i.e., the fraction of the linkings that are correct) and recall (i.e., the fraction of individuals that are correctly linked). We propose two measures, cumulative individual characterizing information (ICI) and genotype predictability (π), to study this tradeoff.

ICI can be interpreted as the total amount of information in a set of variant genotypes that can be used to pinpoint an individual in a linking attack. This quantity depends on the joint frequency of the variant genotypes. For example, a set that contains many common genotypes will not be very useful for pinpointing individuals, but rare variant genotypes will give more information. Thus the information content of a set of genotypes is inversely proportional to the joint frequency of the genotypes. We use this property to quantify ICI in terms of genotype frequency (Online Methods, **Fig. 2** and **Supplementary Fig. 4**). To estimate the joint frequency of variant genotypes, we assume that the variant genotypes are distributed independently (Online Methods and **Supplementary Note**).

For a set of variants, π measures how predictable genotypes are given the gene expression levels. Because genotypes and expression levels are correlated, knowledge of the expression enables one to predict the genotype more accurately than could be done without such information. To quantify the predictability, we use an information theoretic measure for randomness left in genotypes, given gene expression levels (Online Methods and **Fig. 2**). This has several advantages over using reported correlation coefficients to quantify predictability. Although the correlation coefficient is a measure of predictability, it is computed differently in different studies, and there is no easy way to combine and interpret correlation coefficients for joint predictability of multiple eQTL genotypes. However, joint predictability can be easily quantified using π, as it fits naturally to the information theoretic formulations (Online Methods). Furthermore, the predictability estimated via π can accommodate the nonlinear relationships



eQTL 1
E_1
ρ
V_1
(Expression of gene 1)
(Genotype of variant associated with expression of gene 1)
Joint expression genotype distribution
Frequency
Genotype (V_1)
0
1
2
ρ
e
Expression (E_1)
Slice the joint distribution at e
Compute genotype distribution
V_1 genotype frequencies
Frequency
0 1 2
Conditional distribution of V_1 given $E_1 = e$
Frequency
0 1 2
Genotype
Compute entropy

$$\mathrm{ICI}\left(\begin{array}{c}\text{Individual has}\\ \text{variant genotypes } g_1\\ \text{for variant } V_1\end{array}\right) = \log_2\left(\frac{1}{\begin{array}{c}\text{Frequency of}\\ V_1 \text{ genotype}\\ g_1 = 2\end{array}}\right)$$

$$\pi\left(V_1|E_1 = e\right) = \exp(-H(V_1|E_1 = e))$$

Figure 2 | Computation of the ICI and correct predictability of genotypes. Given an eQTL where a genotype of variant V_1 is correlated with expression of gene 1 (E_1), joint distribution of genotype and expression illustrates the correlation (ρ) indicated by the line fit. Computations of marginal and conditional genotype distributions from the joint distribution are illustrated. ICI for the variant genotype g_1 is computed as the logarithm of the reciprocal of the genotype frequency. For n variant genotypes, each genotype contributes to ICI additively with the logarithm of the reciprocal of the genotype frequency: $-\log(p(V_1 = g_1)) - \log(p(V_2 = g_2)) - \cdots -\log(p(V_n = g_n))$. The predictability of the genotype given expression level e is computed in terms of the entropy of the conditional genotype distribution given expression level e. One builds the conditional distribution by slicing the joint distribution at expression level e (indicated by red in the figure).

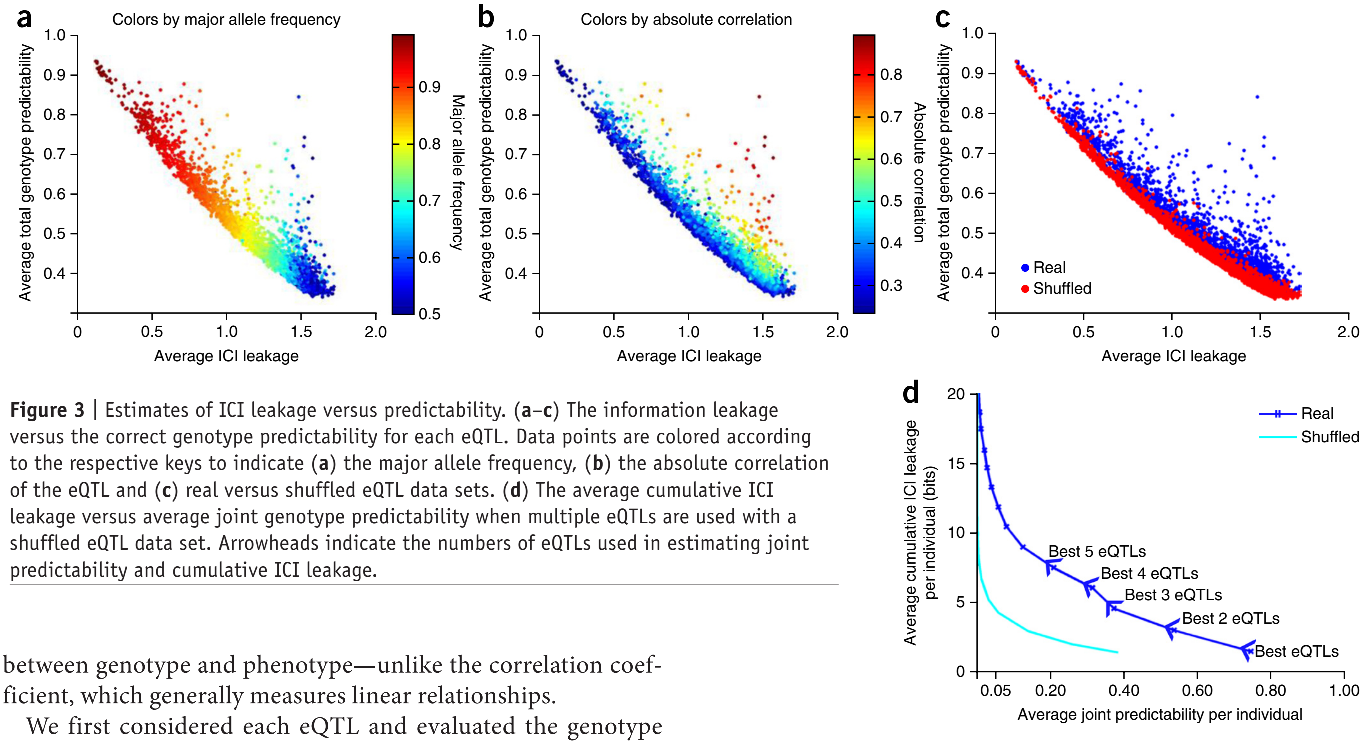


Figure 3 | Estimates of ICI leakage versus predictability. (**a**–**c**) The information leakage versus the correct genotype predictability for each eQTL. Data points are colored according to the respective keys to indicate (**a**) the major allele frequency, (**b**) the absolute correlation of the eQTL and (**c**) real versus shuffled eQTL data sets. (**d**) The average cumulative ICI leakage versus average joint genotype predictability when multiple eQTLs are used with a shuffled eQTL data set. Arrowheads indicate the numbers of eQTLs used in estimating joint predictability and cumulative ICI leakage.

between genotype and phenotype—unlike the correlation coefficient, which generally measures linear relationships.

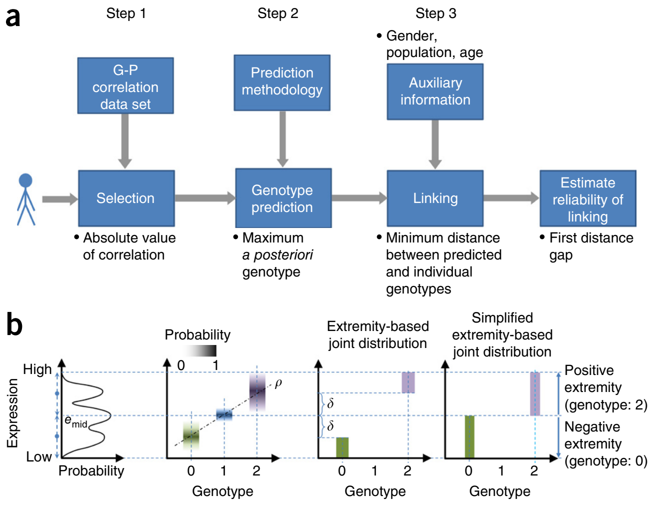
We first considered each eQTL and evaluated the genotype predictability versus the leakage of characterizing information. We computed, for each eQTL in the GEUVADIS data set, the average predictability and average ICI over all individuals (**Fig. 3a**). Most of the data points were spread along the antidiagonal: eQTL variants with high major allele frequencies had high predictability and low ICI, and those with lower frequencies had low predictability and high ICI (**Fig. 3b**). This was expected because on average the genotypes of high-frequency variants can be predicted easily (most individuals will harbor one dominant genotype) and consequently do not deliver much characterizing information (the opposite is true for eQTLs with lower-frequency alleles). To evaluate how much gene expression levels contribute to the predictability of genotypes, we used a shuffled eQTL data set. The predictability versus the ICI leakage for the eQTLs in the shuffled eQTL data set (Online Methods) was predominantly on the antidiagonal (**Fig. 3c**). This was also expected, as the predictabilities for shuffled eQTL genotypes depend mainly on how frequently the genotypes occur in the population (very high-frequency genotypes are associated with low ICI). In contrast, the real eQTLs (**Fig. 3b**) deviated from the antidiagonal compared to the shuffled eQTLs, which shows that expression supplies much information for predicting eQTL genotypes (**Fig. 3c**). The eQTLs with high correlation had substantially more ICI and greater predictability. These results illustrate the fact that π measures the total effect of genotype frequencies and expression levels on the predictability of genotypes.

When multiple genotypes are used, the information leakage is greatly increased. To study this, we computed the ICI and predictability for increasing numbers of eQTLs (**Supplementary Note** and **Fig. 3d**). As expected, the predictability decreased with increasing ICI leakage. Inspection of mean predictability versus mean cumulative ICI enabled us to estimate the number of vulnerable individuals at different predictability levels. For example, at 20% predictability there were approximately 8 bits of cumulative ICI leakage. At this level of leakage, the attacker could pinpoint an individual with 20% accuracy in a sample of $2^8 = 256$ individuals. Thus, in any sample of 256 individuals, we would expect the attacker to correctly link 51 (20% of 256) individuals. Although the attacker would not know which individuals were correctly linked, she could estimate the reliability of the linkings, as discussed later, and focus on the most reliable ones. At 5% predictability, the leakage would be 11 bits, and the attacker could pinpoint an individual in a sample of $2^{11} = 2,048$ individuals. This corresponds to approximately 100 individuals getting correctly linked (5% of 2,048). Auxiliary information can be easily added into ICI. For example, gender information, which can be predicted with high accuracy from many molecular phenotype data sets, brings 1 bit of additional auxiliary information to the ICI (**Supplementary Note**).

Framework for linking attacks

We present a three-step framework for practical instantiation of linking attacks (**Fig. 4a**). This framework can be used to perform mock linking attacks on data sets to assess their privacy risks. We used this framework to simulate attacks and assess the accuracy of the different scenarios. As input, the framework uses the phenotype measurements for an individual being queried for a match to individuals in the genotype data set (**Fig. 1**). In the first step, the attacker selects the QTLs that will be used in linking. The selection of QTLs can be based on different criteria. As discussed earlier, genotype predictability (π) is the most suitable QTL-selection criterion. Although the attacker cannot practically compute predictability using only the QTL list, any function of predictability will still be useful to the attacker for selecting QTLs. For example, the most accessible criterion is selected on the basis of the absolute strength of association ($|\rho|$) between the phenotypes and genotypes. In the second step, the attacker predicts genotypes for the selected QTLs by modeling the genotype-phenotype distribution (**Fig. 4b**). The third and final

Figure 4 | Genotype-expression associations and linking attacks. (**a**) Illustration of the three-step linking process: selection of a genotype-phenotype (G-P) data set for use in linking (step one), prediction of genotypes (step two) and linking of predicted genotypes to the genotype data set (step three). The attacker can also estimate the reliabilities of the linkings using the first distance gap metric. (**b**) Schematic of expression-genotype relationships and simplifications. The trimodal gene expression distribution and the joint genotype-expression distribution are shown. The conditional distribution of expression given each genotype is illustrated with different-colored rectangles corresponding to the different genotypes. The genotypes and expression levels are correlated (ρ) as indicated by the line fit. In the extremity-based joint distribution, when the genotype value is 0, a uniform probability is assigned for expression values where the extremity is less than δ (green rectangle). For a genotype value of 1, no probability is assigned. When the genotype value is 2, the probability is uniformly distributed over expression values for which the extremity is greater than δ (purple rectangle). The simplified extremity-based model uses the same distribution by setting δ to 0. In this case, when the genotype is 0, the joint probability is distributed uniformly over expression levels with negative extremity (green rectangle). When the genotype value is 2, uniform probability is assigned to expression levels with positive extremity (purple rectangle).


a
Step 1
Step 2
Step 3
• Gender, population, age
G-P correlation data set
Prediction methodology
Auxiliary information
Selection
Genotype prediction
Linking
Estimate reliability of linking
• Absolute value of correlation
• Maximum *a posteriori* genotype
• Minimum distance between predicted and individual genotypes
• First distance gap
b
Probability
0
1
High
Low
Expression
e_{mid}
Probability
ρ
0 1 2
Genotype
Extremity-based joint distribution
δ
δ
0 1 2
Genotype
Simplified extremity-based joint distribution
0 1 2
Genotype
Positive extremity (genotype: 2)
Negative extremity (genotype: 0)

step of a linking attack is comparison of the predicted genotypes to the genotypes of the individuals in genotype data set to identify the individual that best matches the predicted genotypes. In this step, the attacker links the predicted genotypes to an individual in the genotype data set (Online Methods).

Individual characterization by linking attacks

Using the three-step approach, we first evaluated the accuracy of linking using a genotype-prediction model in which the attacker knows the exact joint genotype-expression distribution (**Supplementary Note**). Although not very realistic, this scenario is useful as a baseline reference for comparison of linking accuracy. The attacker builds the posterior distribution of genotypes using expression levels from the joint distribution. Finally, she predicts each genotype by selecting the genotype with the maximum *a posteriori* probability (**Supplementary Note** and **Supplementary Fig. 5**) and links the predicted genotypes to the individual whose genotypes match best. For several eQTL selections with changing correlation thresholds, the linking accuracy was greater than 95% and approached 100% when auxiliary information was available (**Fig. 5a**).

In general, knowledge or correct reconstruction of the exact joint genotype-expression distribution might not be possible because the genotype-phenotype correlation coefficient alone is not sufficient to reconstruct the genotype distribution given the expression levels. The attacker can, however, use *a priori* knowledge about the genotype-expression relation and build the joint distributions using models with varying complexities and parameters (Online Methods, **Supplementary Note** and **Supplementary Fig. 6**). Here we focus on a highly simplified model in which the attacker exploits the fact that the extremes of the gene expression levels (i.e., the highest and lowest expression levels) are observed with the extremes of the genotypes (homozygous genotypes). We use a measure termed extremity to quantify the 'outlierness' of expression levels (Online Methods, **Supplementary Note** and **Supplementary Figs. 7** and **8**). On the basis of the extremity of the expression level and the gradient of association, the attacker builds an estimate of the joint genotype-expression distribution, constructs the posterior distribution of genotypes and finally chooses the genotype with the maximum *a posteriori* probability (Online Methods, **Supplementary Note** and **Fig. 4b**).

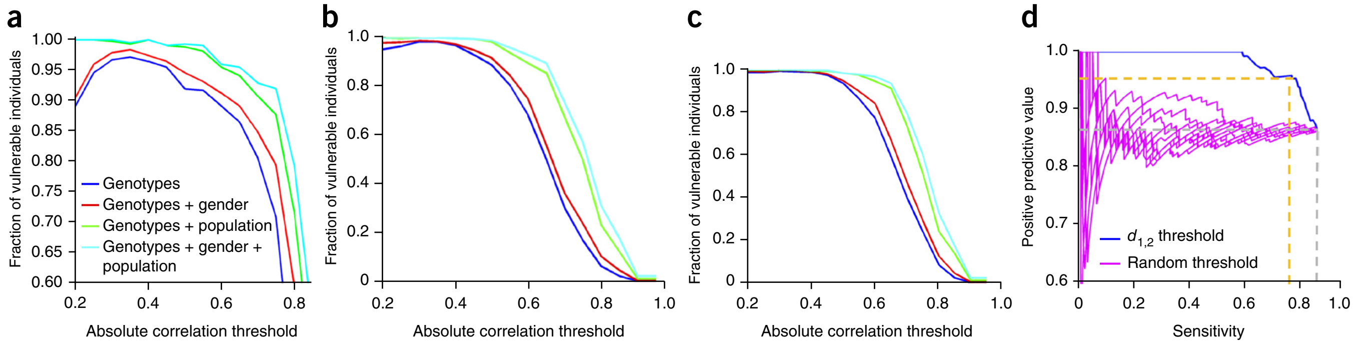


Figure 5 | Accuracy of linking attacks. (**a**) Accuracy of linking with genotype predictions where exact genotype-expression distributions are known (baseline attack). (**b**) Linking accuracy with extremity-based linking with all genotypes. (**c**) Linking accuracy with extremity-based linking with homozygous genotypes. In **a**–**c**, red, green and cyan curves show the linking accuracy with gender, population, and both gender and population used as auxiliary information, respectively. (**d**) Sensitivity versus positive predictive value (PPV) of linkings chosen with different $d_{1,2}$ thresholds for eQTL selection where the overall linking accuracy is 84% (indicated by gray dashed lines). The sensitivity versus PPV for random selections of linkings is shown by the magenta curves. The yellow dashed lines show the sensitivity when the PPV is 95%.

The extremity-based prediction methodology assigns zero probability to the heterozygous genotype. Thus it assigns only homozygous genotypes to variants for which the associated gene's expression level has an absolute extremity higher than a specified threshold. With this approach, the genotype-prediction accuracy increases with increasing absolute correlation threshold (**Supplementary Fig. 7c**). We performed a mock linking attack using this prediction method (in the second step of linking). In the first step of the attack, we used thresholds of absolute correlation ($|\rho|$) and extremity ($|\delta|$) for eQTL selection. The linking accuracy was higher than 95% for most eQTL selections (**Fig. 4b** and **Supplementary Fig. 7d**). We also observed that changing the extremity threshold did not affect the linking accuracy substantially compared to changing the absolute correlation threshold. We thus focused on attack scenarios in which the absolute extremity threshold was set to zero. This also simplified the attack scenario by removing one parameter from the genotype prediction. In the mock linking attack performed with this model, we used the correlation-based eQTL selection in step 1 and extremity-based genotype prediction in step 2. In step 3, we evaluated two distance measures for linking the predicted genotypes to individuals in the genotype data set (Online Methods and **Supplementary Fig. 9**). More than 95% of the individuals (**Fig. 5b,c**) were vulnerable for most of the parameter selections, which is more accurate than in the baseline linking attack (**Fig. 5a**). When the auxiliary information was used, the fraction of vulnerable individuals was 100% for most of the eQTL selections. We also observed that the extremity attack had the potential to link close relatives to each other, which could create potential privacy concerns for family members (**Supplementary Note** and **Supplementary Fig. 10a**). These results show that a linking attack with extremity-based genotype prediction, although technically simple, can be extremely effective in characterizing individuals.

We evaluated whether an attacker can estimate the reliability of linkings. We observed that the measure we termed the first distance gap, denoted by $d_{1,2}$, served as a good reliability estimate for each linking. We computed the positive predictive value versus the sensitivity of the linkings with varying $d_{1,2}$ thresholds. We found that for eQTL selection with an overall linking accuracy of 84%, an attacker could link a large fraction (79%) of the individuals at a positive predictive value higher than 95% (Online Methods, **Fig. 5d** and **Supplementary Fig. 10b**).

We also studied several biases that can affect linking accuracy. First, when the eQTL discovery sample set was different from the sample set on which the linking attack was performed, the accuracies were still very high (**Supplementary Note** and **Supplementary Fig. 10c**). Moreover, attacks were accurate when there was mismatch between the tissue or population of the eQTL discovery sample set and those of the linking attack sample set (**Supplementary Note** and **Supplementary Table 1**). In addition, we observed that the extremity attack was still effective when the genotype sample size was very large (**Supplementary Note** and **Supplementary Fig. 10d**).

DISCUSSION

To ensure genomic privacy, it is necessary to consider a basic aspect of sharing any type of information: there is some leakage of sensitive information from every released data set[19]. It is therefore essential that mechanisms for sharing and publishing genomic data incorporate statistical quantification methods to objectively quantify risk estimates before the data sets are released. The quantification methodology and the analysis framework presented here can be used for analysis of information leakage when the correlative relations between data sets can be exploited for linking attacks (**Supplementary Note** and **Supplementary Fig. 11**).

In the context of linking attacks, the presence of data about a given individual in two seemingly independent databases (for example, phenotype and genotype databases) can lead to privacy risks if an attacker statistically links the databases using *a priori* information about correlation of the entries in the databases. The methods that we propose can be integrated directly into existing risk assessment and management strategies. One such strategy is the use of *k*-anonymization and its extensions[20–22], which enables one to anonymize data sets by ensuring that no combination of the features (e.g., predicted genotypes) can be used to match a record in the data set to fewer than *k* individuals. This is done by censoring the entries or by adding noise to the data set. The estimates of genotype predictability and ICI leakage can be used to select which entries in the phenotype data set should be anonymized so as to achieve anonymity. This maximizes the utility of the anonymized data set by allowing one to focus only on the data points that leak the most characterizing information. In addition, because the anonymization process can focus only on the sources of highest leakage, this approach cuts down on computing requirements[23] and increases the efficiency of anonymization. Another approach is to serve phenotypic data from a statistical database. In this context, differential privacy has been proposed as optimal for privacy-aware data serving[24]. In a differentially private database, release mechanisms are used to query the database and share statistics of the included data. The individual records in the database are not shared. To ensure the privacy of the database, the release mechanisms keep track of the leakage in the past queries and limit access to the database. For phenotype databases, the ICI leakage can be incorporated into the release mechanisms so that the total leakage can be tracked. It is also worth noting that mechanisms for publishing and serving anonymized data may substantially decrease the biological utility of the data[25]. Thus, it is necessary to integrate measures of the biological utility of anonymized data sets as another quantity in risk assessments.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge support from the US National Institutes of Health and from the A.L. Williams Professorship funds. We thank A. Serin Harmanci for constructive comments and discussions on the study design and running of external tools. We also thank D. Clarke for comments on the manuscript.

AUTHOR CONTRIBUTIONS

A.H. designed the study, gathered data sets, performed experiments and drafted the manuscript. M.G. conceived the study, oversaw the experiments and wrote the manuscript. Both authors approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Sboner, A., Mu, X., Greenbaum, D., Auerbach, R.K. & Gerstein, M.B. The real cost of sequencing: higher than you think! *Genome Biol.* **12**, 125 (2011).
2. Rodriguez, L.L., Brooks, L.D., Greenberg, J.H. & Green, E.D. The complexities of genomic identifiability. *Science* **339**, 275–276 (2013).
3. Erlich, Y. & Narayanan, A. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15**, 409–421 (2014).
4. Sweeney, L., Abu, A. & Winn, J. Identifying participants in the Personal Genome Project by name. *Social Science Research Network* doi:10.2139/ssrn.2257732 (2013).
5. Golle, P. Revisiting the uniqueness of simple demographics in the US population. in *Proc. 5th ACM Workshop on Privacy in Electronic Society* 77–80 (ACM, 2006).
6. Golle, P. Revisiting the uniqueness of simple demographics in the US population. in *Proc. 5th ACM Workshop on Privacy in Electronic Society* 77–80 (ACM, 2006).
7. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
8. Ardlie, K.G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
9. Pakstis, A.J. *et al.* SNPs for a universal individual identification panel. *Hum. Genet.* **127**, 315–324 (2010).
10. Wei, Y.L., Li, C.X., Jia, J., Hu, L. & Liu, Y. Forensic identification using a multiplex assay of 47 SNPs. *J. Forensic Sci.* **57**, 1448–1456 (2012).
11. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).
12. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).
13. Im, H.K., Gamazon, E.R., Nicolae, D.L. & Cox, N.J. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet.* **90**, 591–598 (2012).
14. Lunshof, J.E., Chadwick, R., Vorhaus, D.B. & Church, G.M. From genetic privacy to open consent. *Nat. Rev. Genet.* **9**, 406–411 (2008).
15. Church, G. *et al.* Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet.* **5**, e1000665 (2009).
16. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. in *Proc. IEEE Symposium on Security and Privacy* 111–125 (IEEE, 2008).
17. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
18. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2015).
19. Erlich, Y. *et al.* Redefining genomic privacy: trust and empowerment. *PLoS Biol.* **12**, e1001983 (2014).
20. Sweeney, L. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10**, 557–570 (2002).
21. Ninghui, L., Tiancheng, L. & Venkatasubramanian, S. *t*-closeness: privacy beyond *k*-anonymity and *l*-diversity. in *Proc. IEEE 23rd International Conference on Data Engineering* 106–115 (IEEE, 2007).
22. Machanavajjhala, A., Gehrke, J., Kifer, D. & Venkitasubramaniam, M. λ-diversity: privacy beyond *k*-anonymity. *Proc. 22nd International Conference on Data Engineering* 24 (IEEE, 2006).
23. Meyerson, A. & Williams, R. On the complexity of optimal K-anonymity. in *Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* 223–228 (ACM, 2004).
24. Dwork, C. Differential privacy. in *Proc. 33rd International Colloquium on Automata, Languages and Programming* 1–12 (Springer-Verlag, 2006).
25. Fredrikson, M. *et al.* Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. in *Proc 23rd USENIX Security Symposium* 17–32 (USENIX, 2014).

ONLINE METHODS

Genotype, expression and eQTL data sets. The eQTL, expression and genotype data sets contain the information for a linking attack (**Supplementary Fig. 2**). The eQTL data set is composed of a list of gene-variant pairs such that the gene expression levels and variant genotypes are significantly correlated. Here we denote the number of eQTL entries with q. The eQTL (gene) expression levels and eQTL (variant) genotypes are stored in $q \times n_e$ and $q \times n_v$ matrices e and v, respectively, where n_e and n_v denote the number of individuals in the gene expression data set and in the genotype data set, respectively. The kth row of e, $\boldsymbol{e_k}$, contains the gene expression values for the kth eQTL entry, and $e_{k,j}$ represents the expression of the kth gene for the jth individual. Similarly, the kth row of v, $\boldsymbol{v_k}$, contains the genotypes for the kth eQTL variant, and $v_{k,j}$ represents the genotype ($v_{k,j}$ {0,1,2}) of the k variant for the jth individual. The coding of the genotypes from homozygous or heterozygous genotype categories to the numeric values is done according to the correlation data set. We assume that the variant genotypes and gene expression levels for the kth eQTL entry, which we denote with V_k and E_k, respectively, are distributed randomly over the samples in accordance with random variables (RVs). We denote the correlation between the RVs with $\rho(E_k, V_k)$. In most eQTL studies, the value of the correlation is reported in terms of a gradient (or the regression coefficient) in addition to the significance of association (P value) between genotypes and expression levels.

Quantification of characterizing information and predictability. The genotype RV V_k takes one of three different values, {0,1,2}, with the genotype coding done by counting the number of alternate alleles in the genotype. Given that the genotype is $g_{k,j}$, we quantify the ICI in terms of self-information[26] of the event in which the RV takes the value $g_{k,j}$:

$$\begin{aligned}\mathrm{ICI}(V_k = g_{k,j}) &= I(V_k = g_{k,j}) \\ &= -\log_2(p(V_k = g_{k,j}))\end{aligned} \tag{1}$$

where V_k is the RV that represents the kth eQTL genotype and $p(V_k = g_{k,j})$ is the probability (frequency) of V_k taking the value $g_{k,j}$. Given multiple eQTL genotypes, assuming that they are independent, the total ICI is simply a summation of those values.

$$\begin{aligned}&\mathrm{ICI}(\{V_1 = v_{1,j}, V_2 = v_{2,j}, \ldots, V_N = v_{N,j}\}) \\ &= -\sum_{k=1}^{N} \log_2(p(V_k = v_{k,j}))\end{aligned} \tag{2}$$

The genotype probabilities are estimated on the basis of the frequency of genotypes in the genotype data set. We measure the predictability of eQTL genotypes using an entropy-based measure. Finally, the base of the logarithm that is used determines the units in which ICI is reported. When the base two logarithm is used as above, the unit of ICI is bits.

Given the genotype RV V_k and the correlated gene expression RV E_k,

$$\pi(V_k \mid E_k = e) = \exp(-H(V_k \mid E_k = e)) \tag{3}$$

where π denotes the predictability of V_k given the gene expression level e, and H denotes the entropy of V_k given e for E_k. The extension to multiple eQTLs is straightforward. For the kth individual, given expression levels $e_{k,j}$ for all the eQTLs, the total predictability is computed as

$$\begin{aligned}\pi(\{V_k\}, \{E_k = e_{k,j}\}) &= \exp(-H(\{V_k\} \mid \{E_k = e_{k,j}\})) \\ &= \exp\left(-\sum_k H(V_k \mid E_k = e_{k,j})\right)\end{aligned} \tag{4}$$

In addition, this measure is guaranteed to be between 0 and 1, where 0 represents no predictability and 1 represents perfect predictability. The measure can be thought of as a way of mapping the prediction process to a uniform random guess for which the average correct prediction probability is measured by π.

Extremity-based MAP genotype prediction. Using an estimate of the joint distribution, the attacker can compute the *a posteriori* distribution of genotypes given gene expression levels. To quantify the extremeness of expression levels, we use a statistic we term extremity. For the gene expression levels for the kth eQTL, $\boldsymbol{e_k}$, the extremity of the jth individual's expression level, $e_{k,j}$, is defined as

$$\mathrm{ext}(e_{k,j}) = \frac{\text{Rank of } e_{k,j} \text{ in} \{e_{k,1}, e_{k,2}, \ldots, e_{k,n_e}\}}{n_e} - 0.5 \tag{5}$$

Extremity can be interpreted as a normalized rank that is bounded between −0.5 and 0.5. The average median extremity is uniformly distributed among individuals (**Supplementary Fig. 7a**). In addition, around half of the genes (10,000) in each individual have an extremity value greater than 0.3, and around 1,000 genes have an absolute extremity greater than 0.45 (**Supplementary Fig. 7b**). In other words, each individual harbors a substantial number of genes with expressions at the extremes for the population. These can potentially serve as quasi-identifiers. It is worth noting, however, that not all of these extreme genes are associated with eQTLs.

Following from the above discussion, the adversary builds the posterior distribution for the kth eQTL genotype as

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

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

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

From the *a posteriori* probabilities, when the sign of the extremity is the same as that of the reported correlation, the attacker assigns the genotype a value of 2; otherwise the genotype value is 0. Finally, a value of 1 is never assigned to the genotype in this prediction method (i.e., the *a posteriori* probability is zero). In yet another interpretation approach, one can interpret the genotype prediction as a rank correlation between the genotypes and expression levels and choose the homozygous genotypes that maximize the absolute values of the rank correlation. Thus, this process can be generalized as a rank correlation–based prediction. The posterior

doi:10.1038/nmeth.3746

distribution of genotypes in equations (6)–(8) can be derived from a simplified model of the genotype-expression distribution that uses just one parameter. We used the posterior genotype probabilities in extremity-based predictions and assessed the accuracy of genotype prediction. As expected, the accuracy increased with increasing correlation thresholds (**Supplementary Fig. 7c**). There is a slight decrease in genotype accuracy at correlation thresholds higher than 0.7 because the accuracy (i.e., the fraction of correct genotype predictions among all genotypes) is not robust at very small numbers of single-nucleotide polymorphisms. Although very high accuracy is expected, even one wrong prediction among a small number of total genotypes can decrease the accuracy significantly.

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

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

The first distance gap can be computed without knowledge of the true genotype and is immediately accessible by the attacker, with no need for auxiliary information (**Supplementary Fig. 9**). The basic motivation for using this statistic comes from the observation that the first distance gap for correctly linked individuals is much higher than that for incorrectly linked individuals.

eQTL identification with Matrix eQTL. To identify eQTLs, we used the Matrix eQTL[27] method. We first generated the testing and training sample lists by randomly picking 210 and 211 individuals, respectively. We then separated the genotype and expression matrices into training and testing sets. We ran Matrix eQTL to identify the eQTLs using the training data set. To decrease the run time, we ran Matrix eQTL in *cis*-eQTL identification mode. After the eQTLs had been generated, we filtered out the eQTLs that had a false discovery rate (as reported by Matrix eQTL) greater than 5%. We finally removed the redundancy by ensuring that each gene and each SNP was used only once in the final eQTL list. To accomplish this, we selected the eQTL that was correlated with the highest association with each gene. The association statistic reported by Matrix eQTL was used as the measure of the strength of association between expression levels and genotypes. A similar procedure is applied when eQTLs for 30 trios are identified.

Modeling the genotype-phenotype distribution. In the second step of the linking attack, the genotype predictions are performed. As intermediary information, the genotype predictions are used as input for the third step (**Fig. 4a**), in which linking is performed. The main aim of the attacker is to maximize the linking accuracy (not the genotype-prediction accuracy), which depends jointly on the genotype-prediction accuracy and the accuracy of the genotype matching in the third step. In addition to the accuracy of the linking, another important consideration for risk-management purposes is the amount of auxiliary input data (such as training data for a prediction model) that the genotype prediction uses. Prediction methods that require a high amount of auxiliary data would be less useful in a linking attack because the attacker would need to gather extra information before carrying out the attack. In contrast, prediction methods that require little or no auxiliary data make mock linking attacks much more realistic. It is therefore useful, in the context of risk-management strategies, to study the complexities of genotype-prediction methods and evaluate how these are related to the accuracy and applicability of linking attacks. We studied different simplifications of genotype prediction, and here we illustrate different levels of complexity for genotype prediction.

The attacker estimates the posterior distribution of genotypes and uses the maximum *a posteriori* estimate of the genotype for the general prediction method. For this, she must first model the joint genotype-phenotype distribution and then build the posterior distribution of genotypes (**Supplementary Fig. 6a**). The first level of the model can be built via decomposition of the conditional distribution of expression (given genotypes) with independent variances and means (**Supplementary Fig. 6b**). Assuming that the mean and variance are sufficient statistics for conditional distributions (e.g., normal distributions), the joint distributions can be modeled when the six parameters (three means and three variances) are trained. The training can be done by means of unsupervised methods such as expectation maximization, or it can be performed using training data. This would, however, increase the required amount of auxiliary data and decrease the applicability of the linking attack. To introduce a simplification of the model, we assume that the variances of the conditional expression distributions are the same for each genotype (**Supplementary Fig. 6c**). This decreases the number of parameters to be trained to four (three means and one variance). We can build an equally complex model with four parameters by assuming that the conditional distributions are uniform at nonoverlapping ranges of expression (**Supplementary Fig. 6d**). This model requires training of four parameters (e_1, e_2, e_3 and e_4). This model can be further simplified into a model that requires only one parameter (**Supplementary Fig. 6e**). In such a model, uniform probability is assigned when homozygous genotypes are observed and the expression level is higher or lower than e_{mid}. In addition, zero probability is assigned when heterozygous genotypes are observed. Depending on the direction of the genotype-expression gradient, expression levels higher than e_{mid} associate with one of the homozygous genotypes, and expression levels lower than e_{mid} associate with the other homozygous genotype. This simplified model is exactly the distribution that is used in the extremity-based genotype prediction. In the extremity-based prediction, we estimate e_{mid} simply as the midpoint of the range of gene expression levels in the expression data set (**Supplementary Note**).

Data sets. The normalized gene expression levels for 462 individuals and the eQTL data set were obtained from the GEUVADIS mRNA Sequencing Project[17]. The eQTL data set contained all the significant (i.e., those identified with a false discovery rate of no more than 5%) gene-variant pairs with high genotype-expression correlation. To ensure that there were no dependencies between the variant genotypes and expression levels, we used the eQTL entries for which gene and variants were unique. In other words, each variant and gene was found exactly once in the final eQTL data set. We generated the shuffled (randomized) eQTL data sets used in comparisons by shuffling the gene names in the gene-variant pairs in eQTL data set. This randomized the gene and variant matchings. The genotype, gender and population data sets for 1,092 individuals were obtained from the 1000 Genomes Project[18]. For 421 individuals, both genotype data and gene expression levels were available. For the tissue analysis, we downloaded the publicly available significant eQTLs for six tissues, computed by the GTEx Project, from the GTEx Portal. The HAPMAP CEU trio expression and genotype data sets were obtained from the HAPMAP Project website.

Code availability. The analysis code used to generate the results can be obtained from http://privaseq.gersteinlab.org.

26. Cover, T.M. & Thomas, J.A. *Elements of Information Theory* 2nd edn. (John Wiley & Sons, 2005).

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

doi:10.1038/nmeth.3746