1 Significance

Privacy is one of the most important topics of debate in data science that stands at the corner of many different fields, including ethics, sociology, law, political science, and forensic science. Recently, genomics has emerged as one of the major foci of studies on privacy. This can mainly be attributed to the advancement of technologies for high throughput biomedical data acquisition that bring about a surge of datasets[1, 2]. Among these, high throughput molecular phenotype datasets, like functional genomic and metabolomic measurements, substantially grow the list of the *quasi-identifiers* (such as birth date, ZIP code, gender[3]) for participating individuals, which can be used by an adversary for re-identification of the identities. With the recent announcement of Precision Medicine Initiative[4], a large body of datasets are to be generated and shared among researchers[5]. The National Institutes of Health also released the plans to encourage public access to biomedical datasets from scientific studies [5–7]. Considering the fact that one does not need many identifiers to uniquely pinpoint an individual[3, 8, 9], these datasets have the potential to exacerbate the risk of privacy breach.

Many consortia, like GTex[10], ENCODE[11], 1000 Genomes[12], and TCGA[13], are generating large amount of personalized biomedical datasets. Coupled with the generated data, sophisticated analysis methods are being developed to discover correlations between genotypes and phenotypes, some of which can contain sensitive information like disease status. Although these correlations are useful for discovering how genotypes and phenotypes interact, they could also be utilized by an adversary in a linking attack for matching the entries in genotype and phenotype datasets. For example, when a phenotype dataset is available, the adversary can utilize the genotype-phenotype correlations to statistically predict the genotypes, compare the predicted genotypes with the entries in another dataset that contains genotypes. For the entries that are correctly matching, he/she can reveal sensitive phenotypes of the individuals and characterize them. Even when the strength of each genotype-phenotype correlation is not high, the availability of a large number of genotype-phenotype correlations increases the scale of linking. In fact, an adversary can perform correct linking with relatively small number of genotypes[14, 15].

Different aspects of privacy have been intensely studied. Recently, genomic privacy is receiving much attention as a result of the deluge of personalized genomics datasets that are being generated[16, 17]. With the increase in the number of large scale genotyping and phenotyping studies, the protection of privacy of participating individuals emerged as an important issue. Homer et al[18] proposed a statistical testing procedure that enables testing whether a genotyped individual is in a pool of samples, for which only the allele frequencies are known. Im et al[19] showed that, given the genotypes of a large set of markers for an individual, an attacker can reliably predict whether the individual participated to a QTL study or not. These attacks, which we refer to as “detection of a genome in a mixture”, are one type of attacks on privacy (Fig S6). There is yet another important attack where the attacker links two or more datasets to pinpoint individuals in datasets and reveal sensitive information. One well-known and illustrative example of these “linking attacks”, although not in a genomic context, is the linking attack that matched the entries in Netflix Prize Database and the Internet Movie Database (IMDB)[20]. For research purposes, Netflix released an anonymized dataset of movie ratings of thousands of viewers, which they thought was secure as the viewers’ names were removed. However, Narayanan et al[20] used IMDB database, a seemingly unrelated and very large database of movie viewers, linked the two databases, and revealed identities and personal information (movie history and choices) of many viewers in the Netflix database. The fact that Netflix and IMDB host millions of individuals in their databases renders the question of detection of an individual in these database irrelevant since any random individual is very likely to be in one or both of these databases but the focus of attacks turns to matching individuals in the databases. Consequently, as the databases grow, the attacks for detection of an individual in a database become unimportant and the linking attacks become more admissible in order to characterize individuals’ sensitive information. In the genomic privacy context, as the size and number of the genotype and phenotype datasets increase, possibility of potentially linkable datasets will increase, which may make scenarios similar to Netflix attacks a reality in genomic privacy.

2 Innovation

There is currently a significant scarcity of tools that enable analysis and protection of genomic and phenotypic datasets. We will focus on characterizability of the individuals' sensitive information in the context of linking attacks, where the adversary exploits the genotype-phenotype correlations to link different datasets and potentially reveal sensitive information. In general, the high dimensional phenotype datasets generated in genomic studies harbor a number of phenotypes that contain sensitive information, like disease status, and other phenotypes, while not sensitive, may have subtle correlations with genomic variant genotypes. We will perform large scale analysis of the potential genotypic information leakage that different QTL datasets can cause. We will build tools that enable reporting of objective measures for genotypic information leakage from phenotype datasets. These tools will enable generating uniform and systematic analysis of privacy risks imposed by releasing new phenotype, genotype, and QTL datasets. We will also evaluate how accurately the predicted genotypes can characterize an individual. We will study the different routes for linking the phenotype and genotype datasets.

For generating a set of preliminary results that will be presented, we will use the expression quantitative trait loci (eQTL) and expression dataset s generated by the GEUVADIS project[21] and the genotype dataset from the 1000 Genomes Project. We will generalize the formalisms, however, to be applicable to any type of QTL, genotype, and phenotype datasets

Specifically, Many quantitative phenotypes can be linked to genotypes using public quantitative trait loci (QTL) datasets. Some of the high dimensional genomic quantitative traits and corresponding QTLs are gene expression levels (eQTLs), protein levels (pQTLs[22, 23]), DNase hypersensitivity site signals (dsQTLs[24]), ribosome occupancy (rQTLs[25]), DNA methylation levels (meQTLs[26]), histone modification levels (haQTLs[27–29]), RNA splicing (sQTLs[30]), and also higher order traits like network modularity (modQTLs[31]). Other QTLs associated with single dimensional non-genomic phenotypes include body mass index[32], basal glucose levels[33], and serum cholesterol levels[22, 34].

3 Approach

We will address the need for new computational approaches for analyzing sensitive information leakage within 3 aims. In the first aim, we will develop statistical formalisms for quantification of the leakage of information that enables pinpointing of individuals in genotype and phenotype datasets with use of QTLs. In the second aim, we will focus on specific linking attacks and work on instantiations of the linking attacks using outliers in the phenotype datasets. In the third aim, we will focus on proposing file formats and methodologies that enable privacy preserving sharing and publishing of the phenotype datasets. Figure 1 shows how we will utilize the methodologies proposed in each aim can be combined for an integrated risk assessment for releasing QTL, phenotype, and genotype datasets.

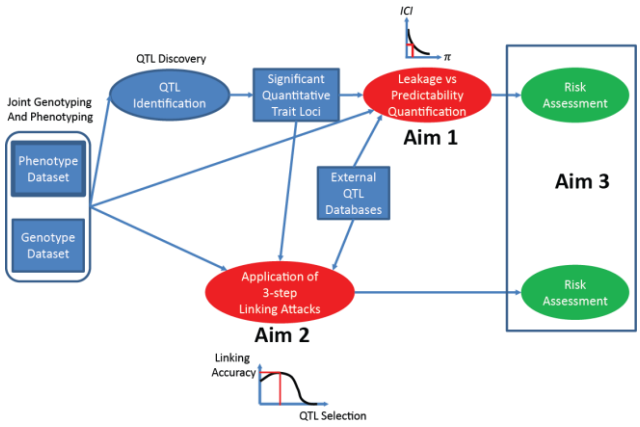

Figure 1: Generalized Risk Assessment Scenario for Genotype-Phenotype Datasets: The joint genotype and phenotype

3.1 AIM1: Development of a Statistical Formalism for Leakage from QTL Sets

In this aim, we will develop a statistical framework for analysis and quantification of the information leakage that can be used for pinpointing and linking individuals in the phenotype and in the genotype datasets using QTL datasets.

3.1.1 Overview of the Individual Characterization Scenario by Linking Attacks

Figure 2 illustrates the general privacy breaching scenario that is considered. There are three datasets in the context of the breach. First dataset contains the phenotype information for a set of individuals. The phenotypes can include sensitive information such as disease status in addition to several molecular phenotypes such as gene expression levels. The second dataset contains the genotypes and the identities for another set of individuals. The third dataset contains correlations between one or more of the phenotypes in the phenotype dataset and the genotypes. In this dataset, each entry contains a phenotype, a variant, and the degree to which these values are correlated. We will focus on the gene expression datasets as the representative phenotype dataset. The abundance of gene expression-genotype correlation (eQTL) datasets makes these datasets most suitable for linking attacks.

Figure 3 illustrates the eQTL, expression, and genotype datasets. The eQTL dataset is composed of a list of gene-variant pairs such that the gene expression levels and variant genotypes are significantly correlated. We will 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 denotes the number of individuals in gene expression dataset and individuals in genotype dataset. The k^{th} row of e, $\boldsymbol{e_k}$, contains the gene expression values for k^{th} eQTL entry and $e_{k,j}$ represents the expression of the k^{th} gene for j^{th} individual. Similarly, k^{th} row of v, $\boldsymbol{v_k}$, contains the genotypes for k^{th} eQTL variant and $v_{k,j}$ represents the genotype ($v_{k,j}$ ϵ

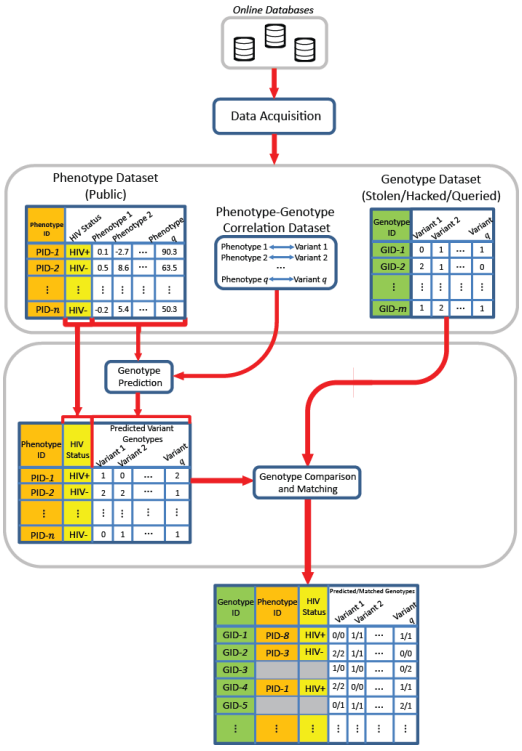

Figure 2: Schematic Representation of a Linking Attack: The attacker links the phenotype and genotype datasets using the genotype predictions. In the predictions, the attacker utilizes the QTL datasets. The resulting attack generates the linked genotype (green), phenotype(orange), and sensitive phenotype (yellow) dataset.

{0,1,2}) of k variant for j^{th} individual. The coding of the genotypes from homozygous or heterozygous genotype categories to the numeric values are done according to the correlation dataset. We assume that the variant genotypes and gene expression levels for the k^{th} eQTL entry are distributed randomly over the samples in accordance with random variables (RVs) which we denote with V_k and E_k, respectively. We denote the correlation between the RVs with $\rho(E_k, V_k)$. In most of the 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. The absolute value of $\rho(E_k, V_k)$ indicates the strength of association between the eQTL genotype and the eQTL expression level. The sign of $\rho(E_k, V_k)$ represents the direction of association, i.e., which homozygous genotype corresponds to higher expression levels. This forms the basis for correct predictability of the eQTL genotypes using eQTL expression levels: The homozygous genotypes associate with the extremes of the gene expression levels and the heterozygous genotypes associate with moderate levels of expression. The eQTL studies utilize linear models to identify the gene and variant pairs whose expressions and genotypes that are significantly correlated. Given this knowledge, the adversary aims at reversing this operation so as to predict genotypes for each individual, using the respective gene expression levels and the genotype-phenotype correlation. For general applicability of the analysis, we will assume that he/she utilizes a prediction model that estimates correctly the *a posteriori* distribution of the eQTL genotypes given the eQTL expression levels, i.e., $p(V_k|E_k)$. This will enable us to perform quantifications independent of the prediction methodology utilized by the attacker.

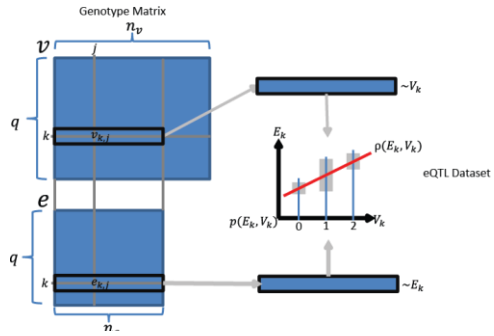

Figure 3: The representation of the genotype and expression datasets.

3.1.2 Quantification of Tradeoff between Correct Predictability of Genotypes and Leakage of Individual Characterizing Information

We will study the tradeoff the correct predictability of genotypes and the number of individuals that can be characterized with the information leakage (Figure 4). In the context of the linking attack, the attacker aims to correctly characterize n_e individuals in the expression dataset among n_v individuals in the genotype dataset. In order to correctly characterize an individual, he/she should select a set of eQTLs that he/she believes he/she can predict correctly. Next, given the individual's expression levels, the attacker should predict the genotypes for the selected eQTLs correctly such that the predicted set of genotypes are not shared by more than 1 individual, i.e., the predicted genotypes can be matched to the correct individual. In other words, the joint frequency of the set of predicted genotypes for the selected eQTLs should be $\frac{1}{n_v}$. We can rephrase this condition as following in information theoretic terms: Given the genotypes of an individual, if the attacker can correctly predict a subset of genotypes that contain at least $\log_2(n_v)$ bits of information, the individual is vulnerable to characterization of his/her phenotypes. Following this statement, we can quantify the leakage from a set of correctly predicted eQTL variant genotypes as the logarithm of their joint frequency. Assuming that the genotypes of different eQTLs are independent from each other, we can decompose the quantity of individual characterizing information that is leaked for a set of n correctly predicted eQTL genotypes:

$$ICI(\{V_1 = g_1, V_2 = g_2, \dots, V_n = g_n\}) = \overbrace{\sum_{k=1}^{n} \underbrace{-\log\big(p(V_k = g_k)\big)}_{\substack{\text{Convert the genotype} \\ \text{frequency to number of bits} \\ \text{that can be used to characterize} \\ \text{individual}}}}^{\substack{\text{Sum individual characterizing} \\ \text{information from all variants}}} \quad (1)$$

where V_k is the random variable that corresponds to the genotypes for the k[th] eQTL, g_k is a specific genotype, and $p(V_k = g_k)$ denotes the genotype frequency of g_k within the population, and *ICI* denotes the total individual characterizing information. Evaluating the above formula, *ICI* increases as the frequency of the variant's genotype g_k decreases. In other words, the more rare genotypes contribute higher to *ICI* compared to the more common ones. Thus, individual linking information can be interpreted as a quantification of how rare the predicted genotypes are. The attacker aims to predict as many eQTLs as possible such that *ICI* for the predicted genotypes is at least $\log(n_v)$. *ICI* can also be interpreted as the number of rare SNP genotypes that an individual harbors.

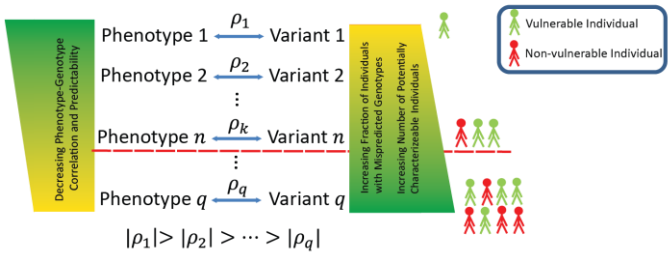

Figure 4:The tradeoff between correct predictability of the genotypes and number of individuals that can be characterized.

In order to maximize the amount of *ICI*, the attacker will aim at correctly predicting as many eQTL genotypes as possible. The (correct) predictability of the eQTL genotypes from expression levels, however, varies over the eQTL dataset as some of the eQTL genotypes are more highly correlated (i.e., more correctly predictable) with the expression levels compared to others, given in $|\rho(E_k, V_k)|$. Thus, the attacker will try to select the eQTLs whose genotypes are the most correctly predictable to maximize *ICI* leakage. Although $\rho(E_k, V_k)$ is a measure of predictability, it is computed differently in different studies. In addition, there is no easy way to combine these correlation values when we would like to estimate the joint predictability of multiple eQTL genotypes. In order to uniformly quantify the joint (correct) predictability of the eQTL genotypes using the expression levels, we use the exponential of entropy of the conditional genotype distribution given gene expression levels. Given the expression levels for j^{th} individual, we compute the predictability of the k^{th} eQTL genotypes as

$$\pi(V_k|E_k = e_{k,j}) = \underbrace{\exp(-1 \times \overbrace{H(V_k|E_k = e_{k,j})}^{\substack{\text{Randomness left in } V_k \\ \text{given } E_k = e_{k,j}}})}_{\substack{\text{Convert the entropy to} \\ \text{average probability}}} \quad (2)$$

where π denotes the predictability of V_k given the gene expression level $e_{k,j}$. π can be interpreted as the average probability (when sampling individuals from the population) that the attacker can correctly predict the eQTL genotype at the given expression level. In the above equation for π, the conditional entropy of the genotypes is a measure for the randomness that is left in genotype distribution when the expression level is known. In the case of high predictability, the conditional entropy is close to 0, and there is little randomness left in the genotype distribution. Taking the exponential of negative of the entropy converts the entropy to average probability of correct prediction of the genotype. In the most predictable case (conditional entropy close to 0), π is close to 1, indicating very high predictability.

As a preliminary study to show how these measures can be used jointly, we considered each eQTL and evaluated the genotype predictability versus the characterizing information leakage. We use the gene expression data from the GEUVADIS project as a representative dataset for this computation. We computed, for each eQTL, average π and average *ICI* over all the individuals (Figure 5). Most of the data points are

spread along the diagonal, which indicate that there is a natural tradeoff between correct predictability and *ICI* leakage. The eQTL variants with high frequency major allele frequencies have high predictability and low *ICI* and vice versa for eQTL variants with lower major allele frequency (Fig 5, left). This is expected because the genotypes of the high frequency variants can be predicted, on average, easily (most individuals will harbor one dominant genotype) and consequently does not deliver much characterizing information. The genotypes for the eQTLs with smaller major frequency alleles, however, are harder to predict as they are mostly uniformly distributed among population. On the other hand, these eQTLs contain high *ICI* on average. The eQTLs with high correlation (Fig 5, right) deviate from the diagonal with high ICI and high predictability. In principle, the adversary will aim at identifying and using these highly informative eQTLs.

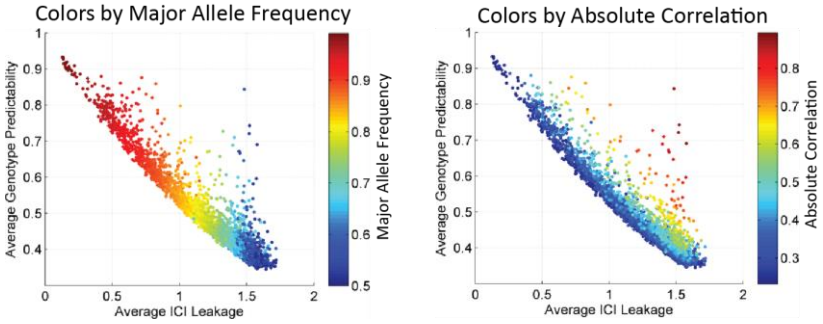

Figure 5: The scatter plot of ICI leakage (x-axis) versus the predictability (y-axis) of SNP genotypes. Each dot represents a SNP. SNPs are colored as per major allele frequencies (left) and per absolute eQTL correlation (right).

3.2 AIM2: Instantiating the Linking Attacks

In this aim, we will study how an attacker can instantiate linking attacks using different techniques for linking the genotype and phenotype datasets.

3.2.1 A General Framework for Analysis of Individual Characterization using Linking Attacks

We first present a tentative 3 step framework for individual characterization in the context of linking attacks. Figure 6 summarizes the steps in the individual characterization for each individual. The input is the phenotype measurements for j^{th} individual. The aim of the attacker is to correctly link the disease state of the individual to the correct identity in the genotype dataset. In the first step, the attacker selects the QTLs, which will be used in linking j^{th} individual. The selection of QTLs can be based on different criteria. As described in the previous section, the most accessible criterion is selection based on the absolute gradient or the absolute strength of association between the phenotypes and genotypes. In the case of eQTLs, this is the reported correlation coefficient, $|\rho(E_k, V_k)|$. We will study different eQTL selection criteria and how they affect the linking accuracy. The second step is genotype prediction for the selected QTLs using a prediction model. We will use different models of genotype prediction with varying complexities and study how the linking accuracy changes when different models are utilized (Figure 9). The third and final step of a linking attack is comparison of the predicted genotypes to the genotypes of the n_v individuals in genotype dataset to identify the

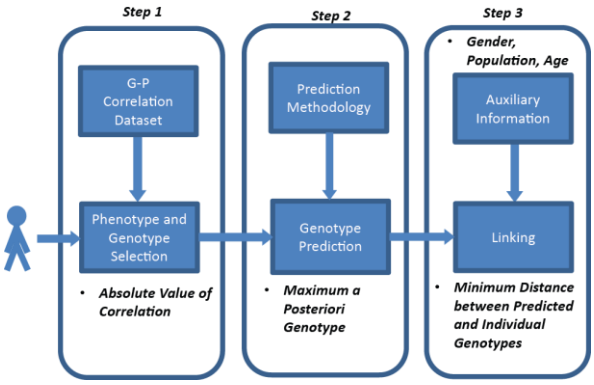

Figure 6: 3 steps of individual characterization with linking attacks.

individual that matches best to the predicted genotypes. In this step, the attacker links the predicted genotypes to the individual in the genotype dataset. We will study different linking methodologies that the adversaries can utilize.

We will study the attacker can utilize a priori knowledge about the relation between gene expression levels and genotypes and build the joint genotype-expression distributions using models with varying complexities and parameters. Even though the genotype prediction with these models may not be very accurate, the attacker can utilize a large number of eQTLs to maximize the accuracy of linking. We will first focus on highly simplified models to evaluate the risk levels associated with simple models for genotype prediction. We will assume the attacker exploits the knowledge that the eQTL genotypes and expression levels are correlated such that the allelic effects on expression are additive and extremes of the gene expression levels (highest and smallest expression levels) are observed with extremes of the genotypes (homozygous genotypes). Therefore, given the gradient of association, the attacker can estimate coarsely the joint distribution of the genotypes and expression levels. This idea is illustrated in Fig 7. 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 will use a statistic we termed $extremity$, which we will briefly introduce here. For the gene expression levels for k^{th} eQTL, $\boldsymbol{e_k}$, $extremity$ of the j^{th} individual's expression level, $e_{k,j}$, is defined as

Deleted: .

$$extremity(e_{k,j}) = \frac{\text{rank of } e_{k,j} \text{in } \{e_{k,1}, e_{k,2}, \dots, e_{k,n_e}\}}{n_e} - 0.5. \quad (3)$$

Extremity can be interpreted as a normalized rank, which is bounded between -0.5 and 0.5. Following from the above discussion, the adversary builds the posterior distribution for k^{th} eQTL genotypes as

$$P(V_k = 0 \mid E_k = e_{k,j}) = \begin{cases} 0 \text{ if } extremity(e_{k,j}) \times \rho(E_k, V_k) > 0 \\ 1 \text{ otherwise} \end{cases} \quad (4)$$

$$P(V_k = 2 \mid E_k = e_{k,j}) = \begin{cases} 1 \text{ if } extremity(e_{k,j}) \times \rho(E_k, V_k) > 0 \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

$$P(V_k = 1 \mid E_k = e_{k,j}) = 0. \quad (6)$$

From the *a posteriori* probabilities, when the sign of the extremity and the reported correlation are the same, the attacker assigns the genotype value 2, and otherwise, genotype value 0. Finally, the genotype value 1 is never assigned in this prediction method, i.e., the *a posteriori* probability is zero. This is expected since we are focusing on the extremes and heterozygous genotype is observed at medium levels of expression. The posterior distribution of genotypes in equations (4-6) can be derived from a simplified model of the genotype-expression distribution that utilizes just one parameter.

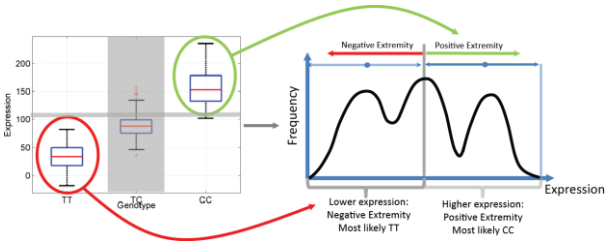

Figure 7: Extremity based genotype for an eQTL. The joint genotype versus expression distribution is shown on left. Given the distribution and the extremity, the genotypes are assigned.

As a next set of preliminary evaluation of how effective the proposed approach is, we utilized extremity based genotype prediction in the 2nd step of the individual characterization framework (Fig 6) and evaluated the fraction of characterizable individuals in the GEUVADIS dataset. We utilized the correlation based eQTL selection in step 1, then extremity based genotype prediction in step 2. In order to demonstrate the utility of the 3-step analysis framework; we evaluated two different distance measures for linking the predicted genotypes to the

individuals in genotype dataset in the 3rd step of the attack. First is based on comparison of the predicted genotypes to all the genotypes in genotype dataset. Second is based on comparison of the predicted genotypes to only the homozygous genotypes in the genotype dataset. The motivation for using this distance measure is that the extremity based genotype prediction never assigns heterozygous genotypes. Thus the heterozygous genotypes are excluded from distance computation.

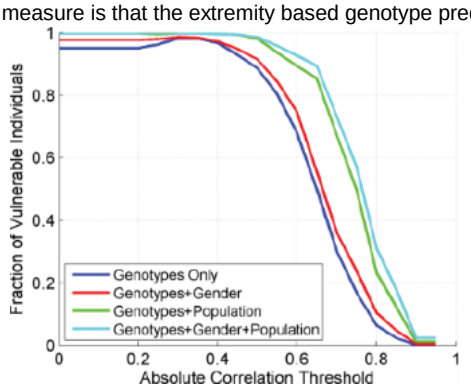

Figure 8: The accuracy of extremity based linking attack with changing eQTL selection correlation threshold.

For each measure, the attacker links the predicted genotypes to the individual whose genotypes minimize the selected distance measure. Figure 8 left and right show the fraction of vulnerable individuals for both distance measures. More than 95% of the individuals are vulnerable for most of the parameter selections for both distance measures. The homozygous genotype matching distance measure has slightly higher linking accuracy. When the gender and/or population information is present as auxiliary information (red and green plots), the fraction of vulnerable individuals increases to 100% for most of the eQTL selections. These results show that linking attack with extremity based genotype prediction, although technically simple, can be extremely effective in characterizing individuals. We will focus on homozygous genotype matching based distance computation in the rest of the paper for simplicity of presentation. We will extend the linking attack analysis on different QTL, phenotype, and genotype datasets.

3.2.2 Modeling Genotype-Expression Distribution for Genotype Prediction

In the second step of a linking attack, the attacker performs genotype predictions. The genotype predictions are used, as an intermediate information, as input to the 3rd step (Fig 3), where linking is performed. The main aim of 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 3rd step. Other than the accuracy of linking, another important consideration, for risk management purposes, is the amount of auxiliary input data (like training data for prediction model) that the genotype prediction takes. The prediction methods that require high amount of auxiliary data would decrease the applicability of the linking attack as the attacker would need to gather extra information before performing the attack. On the other hand, the prediction methods that require little or no auxiliary data makes the linking attack much more realistic and prevalent. We will, therefore study complexities of genotype prediction methods and evaluate how these translate into assessing the accuracy and applicability of the linking attack. We will study different simplifications of genotype prediction, and illustrate different levels of complexity for genotype prediction.

We will study several models of the genotype-phenotype distribution, which can be used in genotype prediction. Figure 9a shows the joint genotype-expression distribution for an eQTL. Figure 9b shows the modeling of the joint distribution using 3 conditional distributions of expression levels at each genotype. First, the means and variances of the distributions are assumed independent. Assuming that mean and variance are sufficient statistics for the conditional distributions (e.g., normally distributed), the joint distributions can be modeled when the 6 parameters (3 means and 3 variances) are trained. We will study different approaches for training the model, e.g. unsupervised methods and evaluate the . This would, however, increase the required auxiliary data and decrease the applicability of the linking attack. Figure 9c shows a simplification of the model by assuming the variances of the conditional expression distributions are same for each genotype. This decreases the number of parameters to be trained to 4 (3 means and 1 variance). Figure 9d shows an equally complex model with 4 parameters where the conditional distributions are uniform at non-overlapping ranges of expression for each genotype. This model requires 4 parameters to be trained corresponding to the expression range limits. Figure S9e shows the final simplification of the genotype prediction, which requires only one parameter to be trained. In this model, the prediction only assigns uniform probability for homozygous

Deleted: ¶
¶
[[We're at ~10pages – we think we could a para here on modellign the joint distribution]]¶

Formatted: Font: (Default) Arial

Formatted: Heading 3

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

Formatted: Font: (Default) Arial, Not Highlight

genotypes when expression levels higher or lower than e_{mid} and assigns 0 conditional probability to the heterozygous genotypes, which brings up an important point: This simplified model is exactly the distribution that is utilized in the extremity based genotype prediction. In the extremity based prediction, we estimate e_{mid} simply as the mid-point of the range of gene expression levels within the expression dataset (Equations 3 and 4-6).

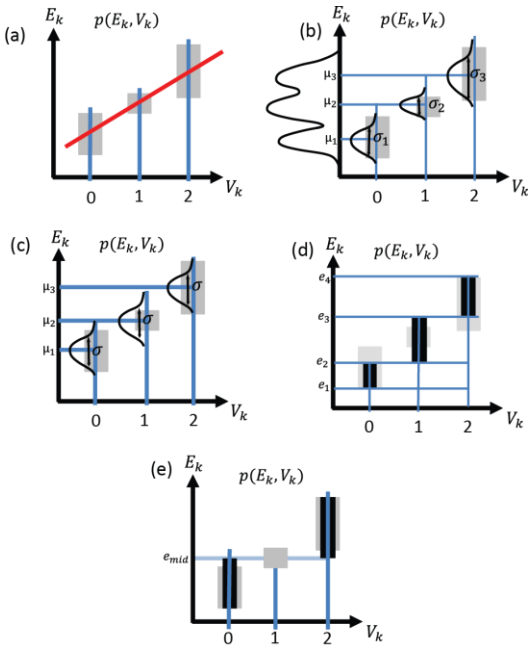

Figure 9: Different Models of Joint Genotype-Expression Distribution with changing Complexities

3.3 AIM3: Building Privacy Reducing File Formats

3.3.1 Rationale

In this section of the grant, we describe the development of a file format that protects privacy, and still maintains a high level of usability of genomic data. We will also describe a practical software implementation to simulate a privacy attack on a dataset and give a consenting subject a sense of how much information leaks in various presentations of a dataset.

3.3.2 Previous experience in tool development and file formats for anonymizing sequence information

The Gerstein lab has developed a number of tools and data formats to handle the increasingly large quantities for data generated by RNA-Seq experiments. For example, we have developed the Mapped Read Format (MRF), a compact data summary format for short, long and paired-end read alignments that enables the anonymization of confidential sequence information. We have developed RSEQtools, which is a suite of tools that use the MRF format for the analysis of RNA-Seq experiments. \cite{21134889}. These tools consist of a set of modules that perform common tasks such as calculating gene and exon expression values, generating signal tracks of mapped reads and segmenting that signal into actively transcribed regions. RSEQtools is implemented in C and the source code is available at http://rseqtools.gersteinlab.org/.

3.3.3 Previous experience in RNA-seq and ChIP-seq computational technology development and data analyses

We have extensive experience in ChIP-seq and RNA-seq tool development and analysis \cite{19015660}. For ChIP-seq, notably, we developed two ChIP-seq peak calling tools: PeakSeq \cite{19122651} and MUSIC \cite{25292436}. PeakSeq is a widely used and highly cited tool for the identification of transcription factor (TF) binding sites. It is also one of the standard peak calling programs used by the ENCODE and modENCODE consortia for numerous ChIPSeq datasets \cite{22955616, 21177976}. MUSIC has just been recently introduced for the identification of enriched regions in ChIP-seq data, especially where the signals are broad and strict peaks are difficult to detect.

For RNA-seq, we have developed MRF and RSEQtools, a suite of tools that enables anonymization of sequence information and quantification of annotated RNAs and identification of splice sites and gene models \cite{21134889}. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms

Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Font: (Default) Arial, Not Highlight
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Font: (Default) Arial, Not Highlight
Formatted: Not Highlight
Deleted: ¶

for alternatively spliced transcripts \cite{22238592}. Our Database of Annotated Regions with Tools (DART) package contains tools for identifying unannotated genomic regions enriched for transcription, as well as a framework for storing and querying this information \cite{17567993}. We developed incRNA, a method that uses known ncRNAs of various classes as a gold standard training set to predict and analyze novel ncRNAs \cite{21177971}.

Furthermore, we continue to play a substantial role in large consortia. We have been heavily involved in the ENCODE consortium \cite{17568003}. For example, a recent ENCODE publication involved the processing and integration of all ENCODE and modENCODE RNA-seq and ChIP-seq data, involving 575 experiments and more than 65 billion reads from three organisms \cite{25164755}. The Gerstein lab is also the data integration hub in the exRNA consortium (http://exrna.org/) that is generating hundreds of RNA-seq and small RNA-seq samples. Other notable consortia for which we have been involved in pipeline construction and big data processing and analyses include the BrainSpan project (http://www.brainspan.org/), which collected RNA-seq data for 8-16 brain structures in each of 13 brain developmental stages \cite{24695229}, as well as the PsychENCODE project (http://psychencode.org/).

3.3.4 Previous work on various social and practical aspects of privacy

We also been an active voice in raising privacy concerns with regards to large-scale genomic datasets. Genomic information axiomatically uniquely and unerringly identifies its owner. Moreover, and perhaps more problematic, individuals represented in genomic datasets share much of their genomic information with their close relatives who likely have not consented to having their genomic data included in the dataset. The Gerstein Lab has suggested in a number of publications that a combination of technological, regulatory and policy changes might best serve to protect individuals described in this arguably unannonymizable data. The policy and regulatory changes ought to be designed to reflect changing norms where we, as a society, no longer dogmatically desire anonymity for every aspect of our lives, or at the minimum, have come to peace with the lack of privacy in the modern age. In acknowledging the changing realities, instead of regulating how to seek out data, we suggest that regulations ought to focus on how that data can be used to harm, for example, in limiting employment or insurance opportunities, thereby further reducing the need for anonymity of formerly sensitive data. Corresponding technological changes include considering both how data is stored as well as where that data should be stored. We have suggested, for example, using cloud based storage options to control and monitor access to data sets and limiting the ability and need to download data to inherently more insecure computers. We have also proposed creating "stub-datasets" that have the look and feel of the typical online data sets, but that would be freely available to all researchers. Holding no personal information, these data sets, while sharing many of the same statistical characteristics, with their larger cousins, would not present privacy concerns, and consequently, could be used to develop and profile code before deployment on real datasets.

Deleted: ¶
¶

3.3 Approach

3.3.1 Practically Quantifying and Minimizing information Leakage in an RNA-Seq File, using MRF and Outlier Removal

Deleted: seq
Deleted: file
Deleted: &
Deleted: outlier
Deleted: removal

In the first step, we will develop a software to efficiently measure the amount of information leakage in an RNA-seq dataset. This quantification of information leakage requires the knowledge of the number of accessible variants that can be obtained from a typical RNA-seq dataset. The ‘accessibility’ can be dependent on a number of factors, including: the type and the coverage of the RNA-seq dataset, and the parameters when aligning the RNA-seq reads.

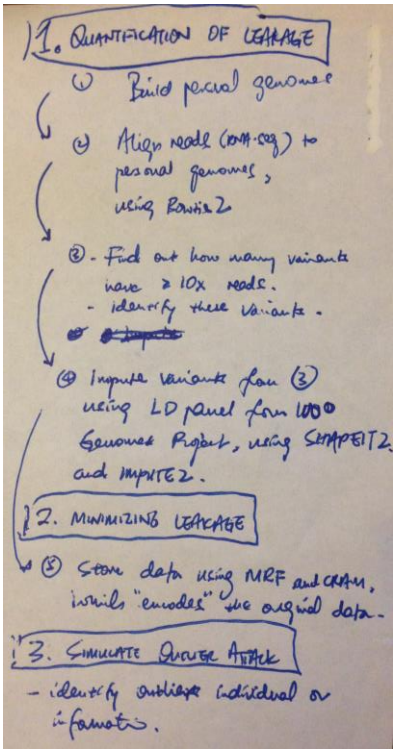

Figure 10. Flowchart of Aim 3 approach

For instance, poly-A RNA-seq data limits the variants to mostly found in the exome, while the total cell RNA-seq data can encompass variants from both the exome and non-coding genes. The coverage of the RNA-seq dataset will determine how reliable a variant call is. GEUVADIS data is mostly poly-A data, and we will use only variants with at least 10X coverage as a quality control. Using the software SHAPEIT2 \cite{24743097} and IMPUTE2 \cite{22384356}, we can expand the accessible variant pool, by imputing variants and inferring their genotypes based on linkage disequilibrium (LD), which is their haplotypic association with the accessible variants. This expanded variant pool (accessible variants and those estimated from LD) will be used to quantify the information leakage of each RNA-seq dataset.

For alignment, we will construct the diploid personal genomes of possibly all the individuals in GEUVADIS and then align RNA-seq reads to them using Bowtie2 \cite{22388286}. We have previously shown that reads from functional genomics assays such as ChIP-seq and RNA-seq map better to the personal genomes than the reference genome \cite{21811232}. Better alignments will aid us by providing more reads for variant calling.

The quantification will also enable the detection of actual variants (or the quantity of variants) that are the most disruptive and the most identifiable. These two pieces of information can be managed in two ways: (1) they can be systematically removed, or (2) they can serve to inform the pertinent individual(s) regarding the extent to which his identity will be compromised; these can be, for instance, conveyed in a consent form. In order to maintain the overall usability and public accessibility of the dataset, the original data and information will be stored in a Mapped Read Format (MRF), which we previously published as part of our RSEQtools \cite{21134889}. The MRF is a compact data file format for storing both short and long reads in functional genomics assays. It decouples sequence and alignment information, and stores only the latter, thereby anonymizing confidential sequence information. CRAM, a highly optimized and widely used data compression tool, is very similar to MRF, but specifically for BAM files \cite{21245279}. We will adapt both MRF and CRAM to current context, by decoupling actual genotype or variant information from their genomic coordinates and storing only the coordinate information. Ultimately, using the file format, we will generate an anonymized dataset that could possibly be more easily shared with less risk of privacy issues.

In addition to GEUVADIS, we intend to apply our software to all major functional genomic datasets, such as GTex[10], ENCODE[11], and TCGA[13], and variant calling datasets such as the 1000 Genomes Project [12] and the Hapmap 3 project \cite{20811451}.

The second step creates a simulation of a privacy attack on gene expression levels that are the most extreme, which can potentially be used to identify the variants associated with such extreme gene expression levels, i.e. the greatest outliers in the dataset. For this, we will implement practically the 'extremity' attack described in Aim 2 and find the gene expression level that has the greatest predictability. Then, reads associated with these outlier gene expression levels can either be removed or the read counts can be modified to the mean gene expression level in a variety of fashion.

[[Pitfalls]]3.3.3 Investigating other sources of 'extremities'

Here we've focused on variants & outlier expression levels but there might be other sources of identifying information in RNA-seq data. Here we will attempt to survey these and get a sense of hteir magnitude….

Formatted: Centered

Deleted: ¶

Deleted: ¶

Moved down [1]: Our study focuses on the individual privacy breaches in the context of linking attacks, where an individual's existence in two seemingly independent databases (e.g., phenotype and the genotype) can cause a privacy concern when an attacker links statistically the databases using the a priori information about correlation of different entries in the databases. The fact that the available molecular phenotypes are (i.e., gene expression levels) generally very high in dimension makes this attack much more probable. The obvious risk management strategy against these attacks is restricting access to the phenotype datasets. The statistical techniques like k-anonymization and differential privacy can also be utilized. These, however, have associated drawbacks about loss of biological utility, and high computational complexity. Moreover, some studies also demonstrated that there are still risks associated with linkability of the anonymized data[35–38]. We believe new studies should address protection and risk management strategies for serving utility-maximized and privacy-aware high dimensional phenotype datasets. ¶

Finally, we will look for other types of extremity-identifying information in gene expression datasets. These may be cryptic information that can be teased from gene expression datasets and exploited to identify individuals. For example, very rare splice transcript isoforms resulting in aberrant gene expression profiles or rare non-coding gene expression in specific individuals can potentially be exploited. These are not as well-studied but are nonetheless worth exploring, especially to investigate the degree to which privacy can be compromised.

3.3.5 Risk Management Strategies for Phenotype Datasets

After quantification of the privacy leakage levels, we will evaluate several risk management strategies to control the leakage of sensitive information in the context of linking attacks.As explained earlier, the privacy leakage is caused by an individual's existence in two seemingly independent databases (e.g., phenotype and the genotype). An attacker can statistically links the databases using the a priori information about correlation of different entries in the databases. The fact that the available molecular phenotypes are (i.e., gene expression levels) generally very high in dimension makes this attack much more probable. The obvious risk management strategy against these attacks is restricting access to the phenotype datasets. The statistical techniques like k-anonymization and differential privacy can also be utilized. These, however, have associated drawbacks about loss of biological utility, and high computational complexity. Moreover, some studies also demonstrated that there are still risks associated with linkability of the anonymized data[35–38]. We believe new studies should address protection and risk management strategies for serving utility-maximized and privacy-aware high dimensional phenotype datasets.

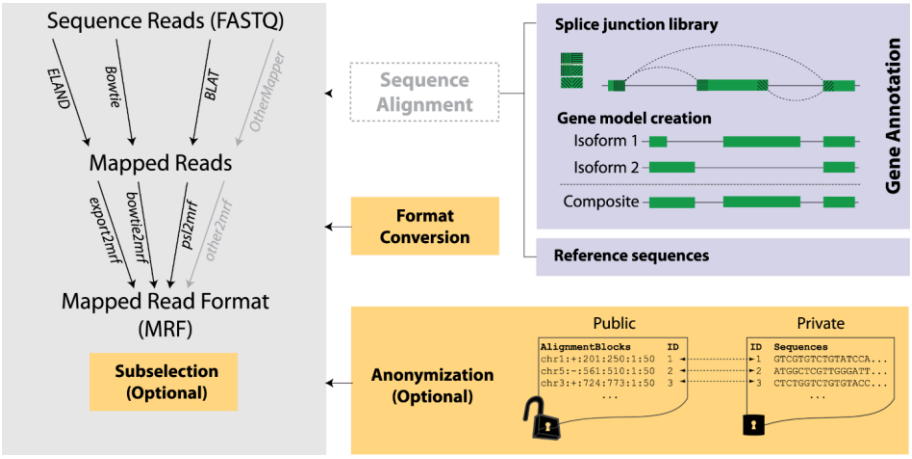

Figure 11: Anonymization Strategy for protection of RNA-seq datasets

4 REFERENCES

1. Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB: **The real cost of sequencing: higher than you think!** *Genome Biology* 2011:125.

Formatted: Font: (Default) Arial, 11 pt
Formatted: Heading 3
Formatted: Font: (Default) Arial
Moved (insertion) [1]
Deleted: Our study focuses on the individual privacy breaches in the context of linking attacks
Deleted: where
Deleted: can cause a privacy concern when a
Deleted: statistically
Formatted: Caption, Centered
Deleted: ¶
Formatted: Font: (Default) +Body (Calibri)
Deleted: ¶

2. Rodriguez LL, Brooks LD, Greenberg JH, Green ED: **The Complexities of Genomic Identifi ability**. *Science (80-)* 2013, **339**(January):275–276.

3. Sweeney L, Abu A, Winn J: **Identifying Participants in the Personal Genome Project by Name**. *SSRN Electron J* 2013:1–4.

4. **infographic-printable.pdf** [http://www.nih.gov/precisionmedicine/infographic-printable.pdf]

5. Collins FS: **A New Initiative on Precision Medicine**. *N Engl J Med* 2015, **372**:793–795.

6. **Plan for Increasing Access to Scientific Publications - NIH-Public-Access-Plan.pdf** [https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf]

7. **GENOMIC DATA SHARING (GDS) Home** [http://gds.nih.gov/index.html]

8. Sweeney L: *Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4*. 2000.

9. Golle P: **Revisiting the uniqueness of simple demographics in the US population**. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*; 2006:77–80.

10. Consortium TG: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45**:580–5.

11. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.

12. The 1000 Genomes Project Consortium: **An integrated map of genetic variation**. *Nature* 2012, **135**:0–9.

13. Collins FS: **The Cancer Genome Atlas (TCGA)**. *Online* 2007:1–17.

14. Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, Kidd KK: **SNPs for a universal individual identification panel**. *Hum Genet* 2010, **127**:315–324.

15. Wei YL, Li CX, Jia J, Hu L, Liu Y: **Forensic Identification Using a Multiplex Assay of 47 SNPs**. *J Forensic Sci* 2012, **57**:1448–1456.

16. Church G, Heeney C, Hawkins N, De Vries J, Boddington P, Kaye J, Bobrow M, Weir B: **Public access to genome-wide data: Five views on balancing research with privacy and protection**. *PLoS Genetics* 2009.

17. Lunshof JE, Chadwick R, Vorhaus DB, Church GM: **From genetic privacy to open consent.** *Nat Rev Genet* 2008, **9**:406–411.

18. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson J V., Stephan DA, Nelson SF, Craig DW: **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays**. *PLoS Genet* 2008, **4**.

19. Im HK, Gamazon ER, Nicolae DL, Cox NJ: **On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy**. *Am J Hum Genet* 2012, **90**:591–598.

20. Narayanan A, Shmatikov V: **Robust de-anonymization of large sparse datasets**. In *Proceedings - IEEE Symposium on Security and Privacy*; 2008:111–125.

21. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P,

Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, et al.: **Transcriptome and genome sequencing uncovers functional variation in humans.** *Nature* 2013, **501**:506–11.

22. Holdt LM, von Delft A, Nicolaou A, Baumann S, Kostrzewa M, Thiery J, Teupser D: **Quantitative trait loci mapping of the mouse plasma proteome (pQTL)**. *Genetics* 2013, **193**:601–608.

23. Stark AL, Hause RJ, Gorsic LK, Antao NN, Wong SS, Chung SH, Gill DF, Im HK, Myers JL, White KP, Jones RB, Dolan ME: **Protein Quantitative Trait Loci Identify Novel Candidates Modulating Cellular Response to Chemotherapy**. *PLoS Genet* 2014, **10**.

24. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK: **DNase I sensitivity QTLs are a major determinant of human expression variation**. *Nature* 2012:390–394.

25. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y: **Impact of regulatory variation from RNA to protein**. *Science (80-)* 2014, **347**:664–667.

26. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK: **DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.** *Genome Biol* 2011, **12**:R10.

27. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK: **Identification of genetic variants that affect histone modifications in human cells.** *Sci (New York, NY)* 2013, **342**:747–749.

28. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, Yurovsky A, Lappalainen T, Romano-Palumbo L, Planchon A, Bielser D, Bryois J, Padioleau I, Udin G, Thurnheer S, Hacker D, Core LJ, Lis JT, Hernandez N, Reymond A, Deplancke B, Dermitzakis ET: **Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription.** *Science* 2013, **342**:744–7.

29. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek D V, Li J, Xie D, Olarerin-George A, Steinmetz LM, Hogenesch JB, Kellis M, Batzoglou S, Snyder M: **Extensive variation in chromatin states across humans.** *Science (New York, NY)* 2013:750–752.

30. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464**:768–772.

31. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, et al.: **The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans**. *Science (80-)* 2015, **348**:648–660.

32. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Mägi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segrè A V, Estrada K, Liang L, Nemesh J, Park J-H, Gustafsson S, Kilpeläinen TO, et al.: **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.** *Nat Genet* 2010, **42**:937–948.

33. Cheverud JM, Ehrich TH, Hrbek T, Kenney JP, Pletscher LS, Semenkovich CF: **Quantitative trait loci for obesity- and diabetes-related traits and their dietary responses to high-fat feeding in LGXSM recombinant inbred mouse strains**. *Diabetes* 2004, **53**:3328–3336.

34. Beekman M, Heijmans BT, Martin NG, Whitfield JB, Pedersen NL, DeFaire U, Snieder H, Lakenberg N, Suchiman HED, de Knijff P, Frants RR, van Ommen GJB, Kluft C, Vogler GP, Boomsma DI, Slagboom PE: **Evidence for a QTL on chromosome 19 influencing LDL cholesterol levels in the general population.** *Eur J Hum Genet* 2003, **11**:845–850.

35. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M: **L -diversity**. *ACM Trans Knowl Discov Data* 2007, **1**:3–es.

36. Ninghui L, Tiancheng L, Venkatasubramanian S: **t-Closeness: Privacy beyond k-anonymity and ℓ-diversity**. In *Proceedings - International Conference on Data Engineering*; 2007:106–115.

37. Wong RC-WW, Fu AW-CC, Wang K, Pei J: **Minimality attack in privacy preserving data publishing**. In *Proceedings of the 33rd international conference on Very large data bases*; 2007:543–554.

38. Fredrikson M, Lantz E, Jha S, Lin S: **Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing**. In *23rd USENIX Security Symposium*; 2014.