Information theory based measures for quantification of private information leakage of functional genomics data and privacy-preserving file formats

GG et al

December 28, 2017

Abstract

[[GG2MG: entirely changed]]

193 words

Functional genomics experiments provide important insight on phenotypes essential for personalized medicine. While publicly sharing of such data is extremely valuable for biomedical research, it invokes privacy concerns. Although phenotypes are not necessarilty tied to an individual's genotype, extracting phenotypical information from these experiments involves using raw sequencing reads that contain a large number of genetic variants. Therefore, reads from functional genomics experiments are considered to be an important set point, which public sharing is only possible in the form of aggregated representations such as signal profiles. There is, however, great desire to share as much data as possible to enhance research reproducibility and enable scientific discovery. Here, we study the quantification of sensitive information in functional genomics data by deriving information theory based measures. We show that functional genomics reads indeed leak a large amount of private information even at small sequencing depths and can be used to construct an individual's complete variant set when combined with imputation. We propose a privacy-enhancing file format enabling public sharing of reads, which constitutes the largest portion of functional genomics data production. Our file format allows accurate quantification of phenotypical data with minimum utility loss.

1 Introduction

With the decreasing cost of DNA sequencing technologies, the number and the size of the available genomic data have exponentially increased and become available to a wider group of audiences such as hospitals, research institutions and individuals [1]. In turn, privacy of individuals has become an important aspect of biomedical data science [2, 3] as availability of genetic information gives rise to privacy concerns such that genetic predisposition to diseases may bias insurance companies [4] or create unlawful discrimination by employers.

Early genomic privacy studies focused on identification of individuals in a mixture by using phenotype-genotype association [5, 6]. These revealed that private information of an individual such as participation to a drug-abuse study [5, 6] can be revealed. With the increase of large-scale genomic projects such as Personal Genome Project (PGP) [7] or recreational/direct-to-consumer genomic databases, researchers showed that multiple datasets can be linked together to infer sensitive information such as pariticipant's surnames [8] or addresses [9]. Such cross-referencing relies on quasi-identifiers, which are pieces of information that are not unique identifiers by themselves, but are well correlated with unique identifiers or can be unique identifiers when combined with other quasi-identifiers [10].

Functional genomics experiments provide a wealth of information on genomic activities related to developmental stages or diseases that are essential for personilized medicine. These are large-scale, high-throughput assays to quantify transcription (RNA-Seq) [11], epigenetic regulation (ChIP-Seq) [12] or 3D organization of genome (Hi-C) [13] in a genome-wide fashion under different conditions (e.g. samples from patients and healthy individuals). The biggest component in inferring biological information from functional genomics experiments is the data analysis step. It starts with the generation of DNA/RNA sequencing reads that are stored in special file formats called fastq [14]. These files are large in size ranging from 5 GB up to 60 GB depending on the

purpose of the experiment. They are then mapped to human reference genome and these mapped reads are stored in compressed binary file types called BAM [15]. While mapping sequences from fastq files to human reference genome is computationally expensive, BAM files provide 1 to 2 fold reduction in data size compared to fastq. The general belief is that BAM files create a set point in the data production process in terms of privacy concern. Especially for high depth experiments such as Hi-C or RNA-Seq, these raw reads can be used to identify the private SNPs, small indels, and structural variants and therefore their accession requires oftentimes special permission. However, it is believed that other aggregated data computed using raw reads are free from genotypes and can be shared. For example, the aggregated data formats for RNA-Seq experiments involve signal profiles and gene/transcript quantifications. Such progressive summarization of the data still allows researchers to make accurate biological conclusions, while providing further data reduction of a $\sim$20 fold. Although overall aggregation and averaging reduces biological information, private information leakage also greatly decreases (Figure 1). On the other hand, recent studies on cross-referencing eQTLs with gene expression levels [17] as well as inferring structural variations from signal profiles [18] showed that there is no static set point in terms of private information leakage and the private information leakage in all levels of the data production process need to be quantified and further processed before making them publicly available. In addition, sequencing data of functional genomics experiments that does not require substantial depth are sometimes considered to be safe to share without privacy concerns as the nature of these data is biased and partial. A good example is that the genome of HeLa cell line requires special access, while we can access the mapped reads from its ChIP-Seq data [19].

On the flip side of the coin is the utility of the mapped reads (BAM files) and challenges related to dealing with private data. Accession to private data require use agreement that has an expiration date and a tremendous amount of burocracy connected to it. Moreover, any secondary data product becomes private and cannot be distributed. Problems associated with the distribution of secondary

data products from private biomedical data is exacerbated due to large file sizes. For example, genome annotations that are derived from private functional genomics data require establishment of their own databases. However, since such annotations are derived from private data, establishment and distribution of these databases require extra levels of privacy related burocracy. Another example to the challenges associated with private data is that big consortia such as ENCODE [20], TCGA [21] or GTEx [22] fund multiple research institutions and enable a collaborative working environment through dedicated phone calls and meetings. In turn, all the participants need to go through required access procedures with their institutions. Otherwise communication based on private data is prohibeted due to data use agreements. Moreoever, when multiple institutions have required access to the same data, they still cannot exchange files with each other. These challanges create a bottleneck and hinder the progress of important biomedical findings. Open data helps the advancement of biomedical data science not only with the easy access to the data, but also helping with the speedy assesment of tools and methods and in turn reproducibility. Funding agencies and research organizations are increasingly supporting new means of data sharing and new requirements for making data publicly available while preserving the participant's privacy [23]. Embracing the both side of the coin, we ask the questions of how much information is enough information to identify individuals and how we can protect the sensitive information with minimum loss of utility in a publicly data sharing mode. This allows to push the set point further down the data analysis, which in turn helps with the solving the complexity associated with private biomedical data. To this end, we derive novel information theory based measures and apply these measures to quantify the amount of leaked information in 24 functional genomic assays from ENCODE [20] at varying coverages. Based on our findings, we develop new file formats that allow the public sharing of read alignments of functional genomics experiments while protecting the sensitive information as well as minimizing the amount of private data that requires special access and storage.

In this study, we use NA12878 as a case example and her 1000 genomes [24] genotypes as gold standard genotypes. We sample reads from the sequencing data of functional genomics experiments at increasing coverages and detect SNVs and indels using Genome Analysis Toolkit (GATK) best practices recommendations [25, 26]. We propose a new metric for qantifying the amount of information that can be obtained from sequencing data with respect to the gold standard. We next present a simple and practical instantiation of a linking attack with the assumption of adversaries accesing increasing amount of the seqencing data. We show that individuals are vulnerable to identifications even at small coverages of sequencing data. We further show that with summation of reads from functional genomics experiments and imputation through linkeage disequilibrium, the leaked number of variants can reach the total number of variants in an indivudal's genome. We then provide a theoretical framework where the amount of leaked information can be estimated from depth and breadth of the coverage as well as the bias of the experiments. Finally, we focus on ways to publicly share alignment data without comprimizing individual's sensitive information. We propose privacy enhancing file formats that hide variant information, are compressed and have minimum amount of utility loss.

2 Results

2.1 Information Theory to quantify private information in an individual's genome

An individual's genome can be represented as a set of variants. Each variant is composed of the chromosome it belongs to, location on that chromosome, the alternative allele and its corresponding genotype. Let $S = \{s_1, s_2, .., s_i, ..s_N\}$ be the set of variants, then each variant can be represented as $s_i = \{v_i, g_i\}$, where v_i consists of the location and alternative allele information and g_i denotes the genotype of the variant as 1 for heterozygous variant and 2 for homozygous variant. We can

then calculate the naive self-information of S in bits as

$$h(S) = -\sum_{i=1}^{i=N} log_2(p(s_i)). \tag{1}$$

In eq.1 N is the total number of variants in an individual's genome, $p(s_i) = n_i/n_T$ is the genotype frequency, in which n_i is the number of individuals with variant $s_i = \{v_i, g_i\}$ and n_T is the total number of individuals in the panel. Note that we denote $h(S)$ as "naive" information, because it is an estimate of the real information in a situation where the population that the individual belongs to is not known and the number of inidivuals are finite. Eq.1 holds only if variants are independent of each other, which is not the case due to the correlation between variants in linkage disequilibrium (LD). In theory, the population that the individual belongs to can easily be predicted by using a few variants. However, from an adversary's perspective, this will add one more layer of calculation, i.e computational and time cost to identification attack. Eq.1 also an estimate to the information when we consider all the individuals in the world (i.e $\lim_{n_t \to \infty} h(S)$).

To be able to understand whether naive information is a good estimate, we first calculate the information with the consideration of LD scores taken from the European population of HapMap project [27]. LD scores are pairwise correlations between variants, which we consider as the prior information on the existence of a variant given other variants in the same LD block exist in a genome. Then the information with LD consideration is calculated as

$$h^{LD}(S) = -\sum_{i=1}^{i=N} (1 - mLD(s_i, s_j))h(s_i) \tag{2}$$

$LD(s_i, s_j)$ is the maximum LD correlation of variant s_i such that $mLD(s_i, s_j) = \max_{i \neq j, j \in (1,..,N)} LD(s_i, s_j)$, where $mLD(s_i, s_j) \neq mLD(s_j, s_i)$.

Figure 2a shows a negligible difference between the naive information and information with LD consideration for NA12878 genome. To understand the lack of difference better, we calculate the self-information of each variant in an LD block with and without LD consideration. We show that highly informative variants do not exhibit any difference due to the low LD correlations (Figure 2b). We further show that the number of variants that have difference between information with and without LD consideration is small compared to highly informative variants having low LD correlations on average.

We then estimate the information when the population size is infinite [28]. We sample fractions in the order of 10%, 20%,..., 100% individuals from the 1000 genomes phase I panel (total of 2504 individuals) and calculate the information using the sampled distribution of genotypes. We repeat this calculation for 100 times and calculate the mean information for each sampled fraction. The relationship between the inverse of the sample fraction and the information fits best to a power function with two terms ($y = ax^b + c$, $R = 0.99$). The y-intercept (c) of the curve is the extrapolation of information when the population size goes to infinity ($1/\infty = 0$, Figure 2c). We again found a negligible difference between the naive information and the information when the population size is infinite (Figure 2a). The information is also calculated by starting from a single individual and adding individuls one by one to the population (SI Figure 1). These individuals are simulated using the genotype frequencies in the 1000 genomes panel and the LD information from HapMap project (see SI methods). Both the information calculation and the KL-divergence between different size populations show that as the size of the population increases, the difference in the information decreases and eventually becomes negligible (SI Figure 1a-b)

In summary, calculations above show that the naive information can be an accurate approximate to the private information content of an individal's genome when the individual's population is not known and the population size is bound by the number of individuals in 1000 genomes panel due to

the relationship of information at $n \to \infty \geq$ naive information $\geq$ information with LD (Figure 2a). That is, an adversary with no prior knowledge on the population of the sample and limited number of individuals in a known genotype panel can accurately approximate the private information in the sample.

2.2 Information Theory to quantify private information leakage in functional genomics data

In an effort to understand the relationship between the leaked information and the coverage as well as for a fair comparison, k amount of reads were sampled from the 24 different functional genomic experiments and from WGS and WES data of NA12878 (see SI Table 1). Genome Analysis Tool Kit (GATK) is used to call SNVs and indels with the parameters and filtering suggested in GATK best practices [25, 26]. The genotypes in 1000 genomes panel for NA1278 is used as the gold standard. We use "naive" pointwise mutual information (pmi) as a measure to quantify the association between the gold standard and the called variants. If $S^{GS} = \{s_1^*, .., s_i^*, ..., s_M^*\}$ is the set of variants from the gold standard and $S^{FGE}(k) = \{s_1, .., s_i, ..., s_M\}$ is the set of variants called from the k reads of a functional genomics experiment, then the set $A = S^{GS} \bigcap S^{FGE}(k)$ contains the variants that are called and are in the gold standard set. If $A = \{a_1, .., a_i, .., a_T\}$, then

$$pmi(S^{GS}; S^{FGE}(k)) = -\sum_{i=1}^{i=T} log_2(p(a_i)) \tag{3}$$

We then add k more reads to the sampled reads and repeat the calculation. This procudere is repeated till we deplete all the reads of a functional genomics experiment. Overall process is depicted in Figure 2e.

2.3 Private information leakage in 24 functional genomics experiment at different coverages

The pmi values for 24 functional genomics experiments are calculated at different coverages. These experiments involve whole genome approaches such as Hi-C, transcriptome-wide assays such as RNA-Seq and targeted assays such as ChIP-Seq of histone modifications and transcription factor binding. In addition, the pmi is also calculated for WGS, WES, and SNP-ChIP for comparison (Figure 3).

As expected Hi-C data contains almost as much information as WGS and more information than SNP ChIP arrays. WGS data contains more information than Hi-C in the beginning of the sampling process. As we sample nucleotides that are between around 1.1 and 10 billion bps, the information content of Hi-C surpasses the WGS data (Figure 3a). We speculate that this is due to better genotyping quality of the genomics regions that are in spatial proximity, as Hi-C has a bias of sequencing more reads from those regions. As expected, we cannot infer as much information from ChIP-Seq reads (Figure 3b). However, surprisingly many of the ChIP-Seq assays such as the ones targeting CTCF and RNAPII contain a great amount of information at low coverages. Furthermore, comparison between WES and different RNA-Seq experiments show that none of the RNA-Seq experiments contain as much information as WES, which is due to the fact that RNA-Seq captures reads only from expressed genes in a given cell (Figure 3c). The unexpected observation is that more information can be inferred from polyA RNA-Seq data at low coverages compared to WES and total RNA-Seq. To be able to make a fair comparison between all these assays, we calculate the pointwise mutual information per bp at the lowest coverages depicted in Figure 3a–c ($pmi(S^{FGE}(k_{min}); S^{GS})/k_{max}$). We found that ChIP-Seq reads targeting CTCF contains even more information per basepair than WGS data at the lowest coverage we sample (Figure 3d).

2.4 Genotyping accuracy

In light of the above findings, in which genotyping can be done using low depth, biased functional genomics experiments, we asses the accuracy of genotyping by calculating the false discovery rate at different coverages. This also measures how much noise that each assay captures. The false discovery rate is defined as the ratio between the information obtained from the incorrectly called variants ($h(S^{FGE} \mid S^{GS})$) and the information obtained from all the called variants ($h(S^{FGE})$), namely

$$FDR(S^{FGE}(k)) = h(S^{FGE}(k) \mid S^{GS})/h(S^{FGE}(k)) \tag{4}$$

Figure 4a shows that the false discovery rate for Hi-C data is lower compared to WGS data at lower coverages. We attribute it to the deeper sequencing of the genomics regions in close spatial proximity. Hence, sampling more reads from those regions at low coverages is more likely compared to uniform sampling of reads from WGS. ChIP-Seq data has comparable false discovery rate to WGS and Hi-C data, ChIP-Seq targeting CTCF having the lowest FDR (Figure 4b). We further find that assays targeting transcriptome such as WES and RNA-Seq produce the noisest genotypes among all the assays, only around 10% of the called variants being the correctly called variants (Figure 4c).

2.5 Linking attack scenario

Linking attacks aim at re-identification of an individual by cross-referencing datasets (Figure 5a). For example, in an hyphotetical scenario, the attacker aims at querying an individual's HIV status from his/her phenotype data. This phenotype data is released with the individuals' genotype information with an anonymized identifier for each individual. We assume that adversary obtains access to this dataset either lawful or unlawful means. Now let's assume that attacker has access to a biosample. This could be partial or complete mapped reads from functional genomics experiments or a saliva sample taken from a used glass. The idea is to do genotyping to

the biosample and find the matching genotype in the HIV status database. However, individuals share many common variants with each other. The number of shared variants between individuals is large within a racial population and even larger within a family. Then the question becomes how well an adversary has to sequence an individual's genome to be able to do succesful linking.

For this, the attacker calls variants directly from the reads of anonymized functional genomic experiments. Then he/she compares the called noisy and incomplete genotypes to the genotype data panel and finds the entry with the highest pointwise mutual information. This reveals the sensitive information for the linked indivudal to the attacker. We then consider a scenario that the attacker has access partial or increasing amount of reads to find out when the data crosses the set point and becomes private.

Based on the pmi values of each experiment at different coverages, we define a metric for linking accuracy called gap_i. To calculate this metric, we first rank all the $pmi(S^{FGE}(k);S^i)$ where $S^{FGE}(k)$ is the set of called genotypes from the functional genomics experiment at total coverage k and S^i is the set of genotypes of individual i in the panel of genotypes. gap_i for each individual i at total coverage k is calculated as;

$$gap_i = \begin{cases} \frac{pmi(S^{FGE}(k);S^i)}{pmi(S^{FGE}(k);S^j)}, & \text{if } rank(pmi(S^{FGE}(k);S^i) \leq 5 \text{ and } rank(pmi(S^{FGE}(k);S^j) = 2 \\ 0, & \text{otherwise} \end{cases}$$

We then define that if gap_i is 0 for the individual i, whose functional genomics data is used, then the individual cannot be identified as there are other individuals in the panel that have the matching genotypes. If $0 < gap_i \leq 1$, then the individual i might be vulnerable with auxilary data such as gender or ethnicity, because he/she is in the top 5 macthing individuals. If $1 < gap_i \leq 2$, then the individual i is vulnerable as we can identify him/her with 1 to 2 fold difference between him/her

and the second best match. Lastly, if $gap_i > 2$, then the individual is extremely vulnerable with more than 2 fold difference between him/her and the second best match (Figure 5a).

We find that NA12878 is extremely vulnerable even at the lowest sampled coverages for Hi-C and RNA-Seq data (Figure 5b). More interestingly between around 1.1 and 10 billion basepairs, the Hi-C data exhibits higher linking accuracy than WGS data, consistent with the previous observation of pmi shown in Figure 3a. The total of coverage of ChIP-Seq data compared to Hi-C and RNA-Seq is quite low (SI Table I). However, the linking accuracy of ChIP-Seq is as good as Hi-C and WGS (Figure 5b), which shows extreme vulnerability of individuals with respect to release of such small amount of data. More strikingly, attacker can link NA12878 by using the reads of single-cell RNA-Seq data, which cover a small portion of the genome in a single cell (Figure 5d). We then added the variants of NA12878's parents to the 1000 genomes genotype panel and repeated the linking attack. We found that although NA12878 is still extremely vulnrebale to re-identification in the presence of her parents in the database, the second best matching individuals are her parents (SI Figure 2). This shows that using the metric *gap*, an adversary can also identify individuals related to the target individual.

2.6 Individual's genome can be accurately approximated from publicly available data by imputation

To answer the question whether an attacker can correctly assemble an individual's variants by only using the reads from ChIP-Seq and RNA-Seq experiments, we impute variants by using IMPUTE2 [29, 30, 31] and the variants called from ChIP-Seq and RNA-Seq experiments. We then collected all the called and imputed variants in a set. Although imputed variants do not contribute to the information due to high correlation with the called variants (SI Methods and SI Figure 3), total number of captured variants increases significantly (Figure 6a). By using shallow squencing data of ChIP-Seq and RNA-Seq, we were able to call and impute variants almost as many as the

gold standard variants.

We then ask the question if we can infer potentially sensitive phenotypes from these variants. Figure 6b shows a small set of example variants associated with physical traits such as eye color, hair color or freckles. Many of these variants are in the called set of Hi-C, ChIP-Seq and RNA-Seq data. Number of variants associted with traits further increases with imputation as expected.

2.7 Toy model for estimation of amount of leaked data without variant calling

Genotyping from DNA sequences is the process of comparing the DNA sequence of an individial to that of reference human genome. To be able to do succesful genotyping, one needs substantial depth of sequencing reads for each base pair. According to the Lander-Waterman statistics for DNA sequencing, when random chunks of DNA is sequenced repeteadly, the depth per basepair follows Poisson distribution with a mean that can be estimated from the read length, number of reads and the length of the genome [32]. Since functional genomics experiments aim at finding highly expressed genes, TF binding enrichment or 3D interactions of the genome, it is expected that the sequencing depth per basepair does not follow the Poisson statistics. Thus, the genotyping using reads from functional genomics experiments is biased towards the variants that are in the functional regions of the cell types/lines of interest.

To this end, we hyphotesized that the genotyping from the sequencing based functional genomics data depends on the average depth per base pair ($\bar{d}$) , the total fraction of the genome that is represented at least by one read, also called the breadth ($b = \sum_{i=1}^{N} [d_i \geq 1]$, N is the total number of nucleotides in the genome) and a parameter β that estimates the sequencing bias, i.e. how much the distribution of depth per basepair deviates from the Poisson distribution (Fig. 6c). The bias parameter β is composed of two terms: (1) the negative bias $\beta-$ and (2) the positive bias $\beta+$.

Negative bias estimates if there is an increase in the number of low depth basepairs relative to mean with respect to espected Poisson distribution and the positive bias estimates the increase in the number of high depth basepairs (see SI for more details).

To quantify the genotyping from the functional genomics data, we used "naive" normalized pointwise mutual information (npmi). It takes into account the information from the correctly identified genotypes ($pmi(S^{FGE};S^{GS})$), the information missed that is in the gold standard ($h(S^{GS} \mid S^{FGE})$) and the information from the incorrectly identified genotypes, i.e FDR ($h(S^{FGE} \mid S^{GS})$) as;

$$npmi(S^{FGE};S^{GS}) = \frac{pmi(S^{FGE};S^{GS})}{h(S^{FGE},S^{GS})} = \frac{pmi(S^{FGE};S^{GS})}{h(S^{GS} \mid S^{FGE}) + pmi(S^{FGE};S^{GS}) + h(S^{FGE} \mid S^{GS})} \quad (5)$$

With the assumption of $npmi(S^{FGE};S^{GS}) = f(\overline{d_{FGE}}, b_{FGE}, \beta_{FGE})$, we used Gaussian Process Regression (GPR) [**?**] to fit 40 training data points and achieved a root mean square error (RMSE) of 0.06 with the values ranging between [0,35] (Fig. 6d). 5 separate data points were used as test set and an RMSE of 0.07 was acheieved (Fig. 6d),see SI for more details). The regression learning is performed using 10 fold cross-validation to protect against overfitting. This toy model represents a proof of concept suggesting a theoretical framework for the estimation of amount of leaked data from functional genomics experiments without the need of performing time-consuming genotyping calculations.

2.8 Unique combination of common variants contribute significantly to the information leakage and linking accuracy

We next analyze whether a linking attack can be prevented by removing rare variants from the datasets as their contribution to the information is the highest. To understand the cut-off for the frequency of the variants that need to be removed in order to fail at linking, we calculate the naive information of the gold standard genotypes in the 1000 genomes panel with and without

the presence of the NA12878 (Fig. 6a). The genotypes that deviate from the diagonal the most are the ones that contribute the most to the overall naive information. We group the genotypes into three categories based on their contribution to the overall naive information as depicted in Fig. 6a. We first remove genotypes in the category I from the NA12878 variant list. These are the genotypes that are unique to NA12878. We expected that removal of category I genotypes will affect the linking accuracy greatly. However, we surprisingly found that removal of these unique genotypes affects linking minimally and the individual is still extremely vulnerable to linking attacks ($gap_{NA12878} > 2$,Fig. 6b). Therefore removing genotypes that are unique to the individual from functional genomics data would not be enough to protect the individual's sensitive information. We then relaxed our criteria and removed the variants in category II, which includes the variants in category I and more. We again found that individual is extremely vulnerable to linking attacks ($gap_{NA12878} > 2$,Fig. 6b). We finally relaxed the cut-off further and removed any variant that has information contribution 6 bits or higher. This also did not affect the overall linking ($gap_{NA12878} > 2$,Fig. 6b).

The genotypes in these 3 categories are rare in the 1k genome panel. They are seen 64 or less individuals including NA12878. A practical solution to the re-identification problem using functional genomics data would be masking or removing such rare genotypes from the reads. However, as iteratively shown here that although rare variants are extremely informative and sufficient enough to do re-identification through linking attacks, their removal is not sufficient to fail at re-identification. That is, not only the rare genotypes but also the unique combination of common genotypes are identifiers of genetic make-up of an individual. To further support this calculation, we added the genotypes of the parents of NA12878 to the panel and found that linking is still succesful with an extreme vulnerability ($gap_{NA12878} > 2$,SI Fig.2).

We then analyze the contribution of small indels to the naive information and whether accurate linking is possible when we remove all the single nucleotide mutations from the data and keep the indels. Fig. 6c shows the information contribution of the indels. Although naive pointwise mutual information from indels are much smaller compared to single nucleotide mutations, a high linking accuracy can be achived by using only indels even at small coverages (Fig. 6d). This linking attack is done using the most noisy data set we have (total RNA-Seq) to make linking more difficult.

2.9 Privacy-enhancing file formats for functional genomics experiments

After discovering neither common variants nor indels can be publicly shared, we seek for ways to share the mapped reads of functional genomics data. The purpose is to share maximum amount of information with minimum utility lost while maintaining the individual's privacy. As a privacy metric, we aim to prevent leakage of any variants as well as any quasi-identifier that can lead to identification of position of variants in the genome. For utility measure, we used the following equation:

$$U = \frac{\overline{d} - RMSE}{\overline{d}} \tag{6}$$

In Eq. 7, $\overline{d}$ is the mean number of reads that overlap with a basepair or with a functional unit such as exons. *RMSE* is the root mean square error between the real depth and the depth after distorting the file to make it private. For a genome with N number of basepairs or exons, it can be calculated as:

$$RMSE = \frac{\sum_{i=1}^{N} |d_i - d_i^p|}{N} \tag{7}$$

d_i is the real number of reads for ith basepair, wheras d_i^p is the number of reads obtained from the distorted file. U measures the percentage of the depth per bp that are correctly reported on the privatized file format, while *RMSE* is the mean difference between the depth of a nucleotide between privatized and original files. According to U and *RMSE*, the high depth regions of genome, i.e. the functional regions, will be penalized more if the new file format reports the depth different than the

original depth, while the low depth regions such as low expression genes will be penalized less. As the purpose of functional genomcis experiments is to annotate functional genome, the utility metric measures the quality of the annotation when the analysis tools are performed using privatized file formats.

The reads from the SAM/BAM/CRAM files are categorized as perfectly mapped reads that includes also intronic reads and reads with mismatches, insertions, deletions, soft- and hard-clipping. We remove the sequence of the reads and keep mapping quality, start coordinates, fragment lengths and flags related to mappability of the reads while adjusting the cigar and alignment scores such that leakage of variants are masked (Fig. 8). The details of how new file format deals with reads are reported in detail in SI Methods with a detailed figure (SI Fig. 3).

Such treatment of cigars introduce noise to the reads especially with deletions since the start coordinates and total length of the fragments are unchanged (Fig. 8a-b, see SI Methods and SI Figure 4). However, our utility analysis showed $> 99.9\%$ accordance with the original depth of the nucleotides. As can be seen from the scatter plots, noise is mostly introduced to basepairs with low depth (Fig. 8c-d). We call this file format pSAM/BAM/CRAM. The pBAM file format contains the necessary information to be used in functional genomics pipelines such as gene expression quantification and transcription factor binding peak calling. We then create a ".diff" file that contains the information that are distorted in the pBAM files, except the sequence of the reads. Instead of reporting the entire sequence of a fragment, we reported the nucleotides that are different than the reference sequence (see SI). ".diff" files are private files that require special permission for access. The advantage of locking up the ".diff" files instead of the entire BAM files is that they are smaller in size, hence it is easier to store and move these files. A user is able to reach the original BAM file when they have access to the .diff file and a script can convert pBAM + .diff + reference genome into the original BAM file (Fig 8b).

3 Discussion

Functional genomics experiments provide large amount of biological data. These are large-scale, high-throughput assays based on sequencing. In turn, a great amount of genotype information can be recovered from the raw reads. Therefore, raw reads almost always require special permission for access. The size of these protected files can go up to few gigabytes. Thus, private accesion to this data creates one more layer of complexity in terms of processing, analysis and data transfer by posing an excessive administrative burden on academic. Moreover, functional genomics experiments advanced our undertsanding of health and disease by revealing function of the genome under different conditions. The quantification, analysis and the interpretation of functional genomics data are still not entirely developed, hence extensive sharing of functional genomics data accelerate collaborative research and reproducibility.

Although re-identification of individuals using DNA variants is a well-studied area of genomic privacy, the quantification of private information content of the functional genomic data and open access to such data without comprimising individuals' identity have not been well studied. We have derived information thery based measures to systematically quantify the sensitive information leakage of functional genomics data. Our results demonstrate that extremely accurate linking attacks can be performed when an adversary have access to fairly small amount of reads from functional genomics experiments. Although these experiments are not performed for the purpose of detecting variants, our results show that the nature of functional genomics data is similar to whole genome sequencing data and rich information about an individual's identity can easily be inferred when combined with imputation. These results reflect the reasoning behind the necessity of special permission acces to this data. We further showed that the information leakage of functional genomics data can accurately be estimated by using only three variables that are easily accesible from the sequencing data. This framework is likely fruitful for early estimation of private information leakage from any sequencing data before their release without the need of doing costly

genotyping calculations.

Analysis and interpretation of functional genomics data start with the alignment of sequencing products obtained from functional regions of genome to the reference human genome. In this study, we showed that even a small portion of these reads leaks great amount of sensitive information. SNPs, small indels and structural variants can be identified using these reads. Consequently, the files that contain mapped raw reads are not publicly available, while aggregated data computed using raw reads are considered to be free of sensitive information and can be shared publicly. This marks a set point in the data processing chain, where below the set point data contains sensitive information and cannot be shared, wheras above set point the data is freely available (Fig. 1). Although the idea of set point is effective in sharing data without comprimising individual's identity, in practice it does not work due to couple of reasons. Firstly, aggregated data such as signal profiles or gene expression levels are shown to be strong quasi-identifiers that leak sensitive information. Secondly, functional genomics data analysis is still an evolving field. Different analysis techniques that generates aggregated data such as gene expression levels have been implemented. Thus, protecting mapped reads is detrimental to progress of biomedical research as well as implementation of reproducible pipelines.

[[GG2MG: Following paragraph about differential privacy adopted from JP's genoshare paper]]

In order to overcome bottlenecks related to data sharing and answer privacy-concerns, researchers proposed solutions such as differential privacy [**?**, **?**, **?**]. Differential privacy solutions often supress and add noise to the data, so that the dataset retrieves the same result with and without the individual's information in it [**?**]. It was suggested that solutions based on differential privacy are not suitable for sharing raw reads from a single individual, because they aim at protecting sensitive information when the data from multiple individuals are aggregated in databases [**?**]. Another concern about the solutions based on differential privacy is the utility of the privatized data as,

for example, association studies based on relationship between genotypes and phenotypes require complete accuracy [**?**]. However, solution becomes less complex for sharing raw reads from functional genomics experiments as the main purpose of functional genomics data is not related to genotypes, but rather annotation of functional regions of the genome under different conditions such as cancer. Therefore, privacy-preserving solutions for sharing raw reads of an individual's functional genomics data can dismiss the accurate sharing of genotypes and focus on the utility of the data in terms of annotation of functional genome.

To this end, inspired by the essence of differentially private solutions, we propose means to share raw reads from functional genomics data without comprising individual's sensitive information. We created privatized data formats, in which the functional annotation of genome is as accurate as possible with and without the genotype leakage from the reads. We showed that we can accurately compute the signal profiles and gene quantifications with more than 99% recovery. This new file format called pBAMs enable researchers to share the mapped reads, which are largest data product of functional genomics experiments. To easen the challenges associated with moving and storing of large special access files, we created light-weight .diff file format that consists of the differences between pBAM and BAM files in a compact format. This allows us not to repeat the sequence information in the human reference genome files in .diff files and reduces the size of the private files significantly. [[GG2MG: I will add "ENCODE uses these files" after we have the call with them]]

[[GG2MG A little bit recycling from PrivaSig but we can change the sentences more if needed]]Presented framework can be used to quantification of sensitive information from the raw reads of functional genomics experiments and conversion of raw files to privacy-preserving file formats. We address the most obvious leakage and provide solutions for quick quantification and safe data sharing. However, it is useful to review all the sources of information leakage from functional genomics

experiments. For example, the next source of leakage is from the signal profiles in RNA-Seq, which was addressed elsewhere [18]. There is also leakage from gene expression quantifications, which was shown to be connected with variants through the eQTLS [17]. We also anticipate more leakages to be discovered as new functional genomics experiments are developed. Combined with the increasing attention to genomic privacy, we expect future studies will lead to novel privacy-preserving solutions in an open data sharing mode.

-[[GG2MG we can also claim that signal profiles created from pBAMs are safe to share? there is tehcnically no dip in the signals]]

References

[1] Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biology*, 2011;12(8):125.

[2] Joly Y, Dyke SOM, Knoppers BM, Pastinen T. Are Data Sharing and Privacy Protection Mutually Exclusive? *Cell*, 2016;167(5):1150-1154.

[3] Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.*, 2014;15(6):409-421.

[4] Joly Y, Feze IN, Song L, Knoppers BM. Comparative Approaches to Genetic Discrimination: Chasing Shadows? *Trends Genet*, 2017;33(5):299-302.

[5] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, 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(8):e1000167.

[6] 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(4):591-598.

[7] Church GM. "The Personal Genome Project". *Molecular Systems Biology*, 2005;1(1):E1E3.

[8] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*, 2013;339(6117):321-324.

[9] Sweeney L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002;10(5):557-570.

[10] Sweeney L. Simple demographics often identify people uniquely. *Carnegie Mellon University, unpublished*, 2000.

[11] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 2009;10(1):57-63.

[12] Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat. Rev. Genet.*, 2009;6:S22S32.

[13] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 2009;326(5950):289-293.

[14] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 2009;38(6):1767-1771.

[15] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009;25(16):2078-2079.

[16] Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. *Science*, 2012;44(5):603-608.

[17] Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature Methods*, 2016;13(3):251-256.

[18] Harmanci A, Gerstein M. Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions. *Nature Communications*, 2017

[19] Beskow LM. Lessons from HeLa Cells: The Ethics and Policy of Biospecimens. *Annu Rev Genomics Hum Genet.*, 2016;17:395-417

[20] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012;489(7414):57-74.

[21] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 2013;45(10):1113-1120.

[22] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 2013;45(6):580-585.

[23] National Institute of Health data sharing policy. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-110.html

[24] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 2010;467(7319):1061-1073.

[25] DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 2011;43(5):491-498.

[26] Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 2013;43:11.10.1-33.

[27] International HapMap Consortium. The International HapMap Project. *Nature*, 2003;426(6968):789-796.

[28] Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.*, 1998;80:197.

[29] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 2009;80:5(6):e1000529.

[30] Howie BN, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3: Genes, Genomics, Genetics*, 2011;1(6):457-470.

[31] Howie BN, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 2012;44(8):955-959.

[32] Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 1988;2(3):231-239.

[33] Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. *MIT Press*, 2006;ISBN 0-262-18253-X.

S. Fienberg, A. Slavkovic, and C. Uhler. Privacy preserving GWAS data sharing. In ICDM, pages 628635, 2011.

A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In KDD, pages 10791087, 2013.

F. Yu, S. E. Fienberg, A. B. Slavkovi, and C. Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. Journal of Biomedical Informatics, 2014.

List of Figures

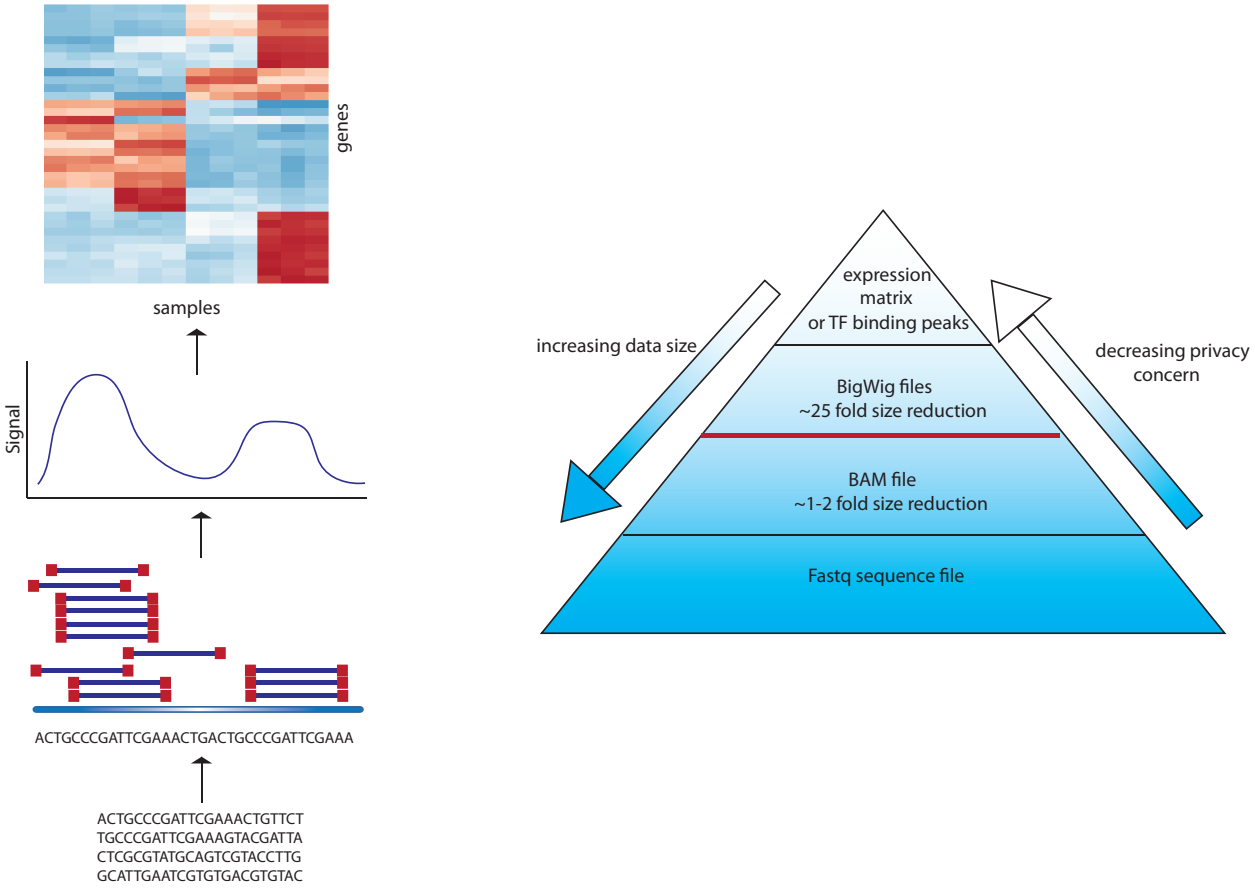


Figure 1: **Schematic of data types from functional genomics experiments.** (**a**) The flow for RNA-Seq data processing from mapped reads to the gene quantifications. (**b**) Different layers of produced data from RNA-Seq pipeline. Red line denotes the set point, where privacy concern vanishes afterwards.

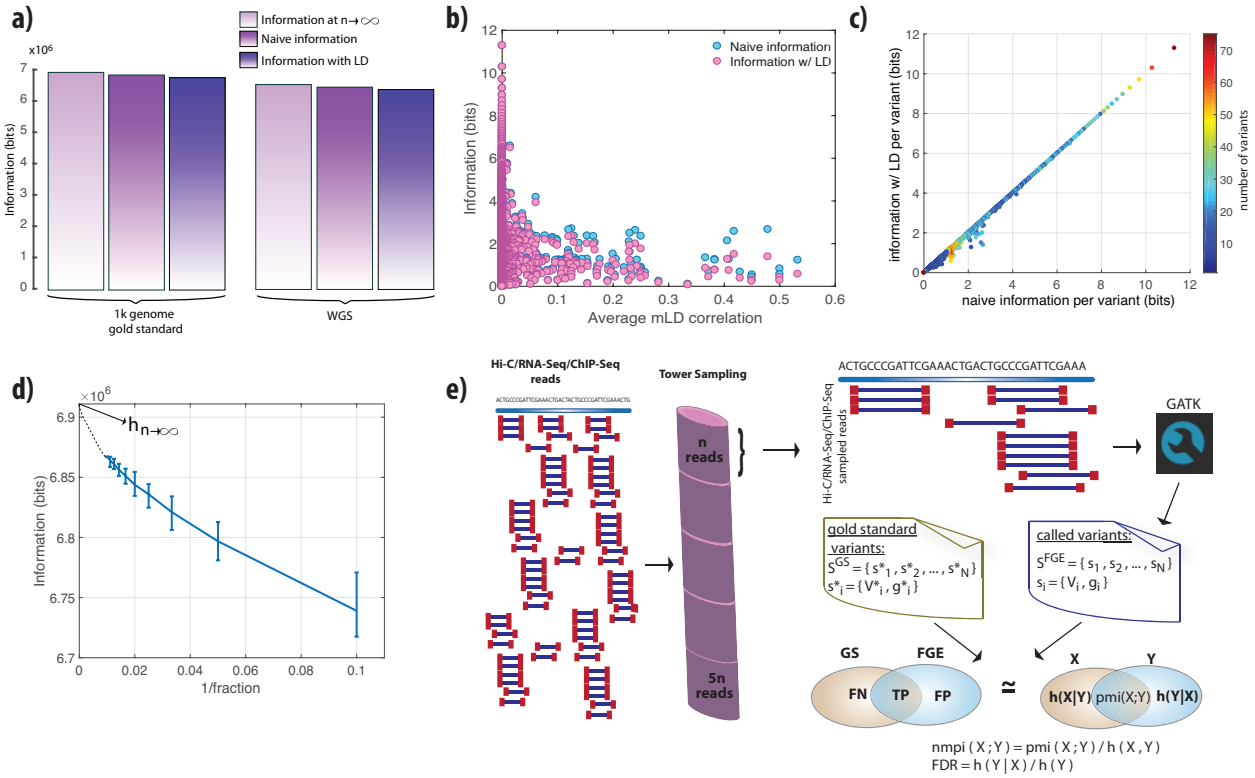


Figure 2: **Comparison of naive information measure with information with LD consideration and sample size correction.** (**a**) Difference between the naive information, information with LD consideration and extrapolated information when population size is infinite. (**b**) The maximum LD score for each variant are averaged over per information and plotted against information. Highly informative variants do not exhibit difference when information is calclated sing naive approach vs. with LD consideration. (**c**) Naive information vs. information with LD consideration per each variant in an LD block. Only low information variants show slight difference between two approaches. (**d**) Naive information vs. inverse fraction of the data sampled from the 1000 genomes population. y-intercept is extrapolated from the fitted curve and denotes the information when the population size is infinite. Error bars are calculated using 100$\times$ bootstrapping. (**e**) The process of sampling reads from functional genomics experiments for the calculation of pointwisw mutual information between 1000 genomes gold standard variants for NA12878 in different coverages.

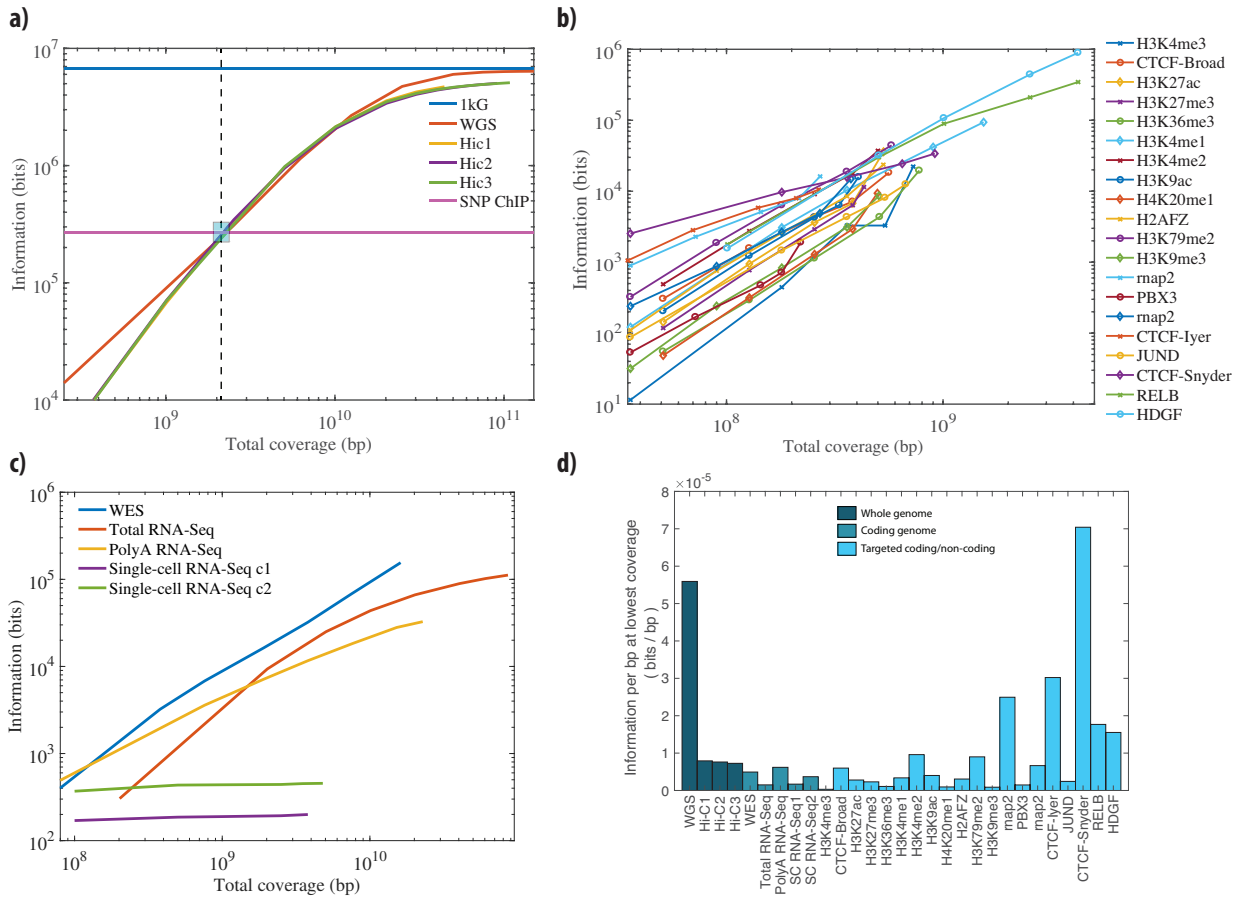


Figure 3: **The pointwise mutual information calculated for 24 different functional genomics assays and WGS, WES and SNP ChIP data using NA12878 1000 genomes variants as gold standard.** (**a**) The pmi values for WGS and three different primary Hi-C experiments plotted at different coverages. The information contents of the gold standard (1kG in blue) and SNP ChIP (in pink) are added for comparison. (**b**) The pmi values for 20 different ChIP-Seq experiments targeting histone modifications and transcription factor binding plotted at different coverages. (**c**) The pmi values for WES, total RNA-Seq, polyA RNA-Seq and single-cell RNA-SEq from two different cells plotted at different coverages. (**d**) The pmi values per basepair plotted using the lowest total coverage for all the assays.

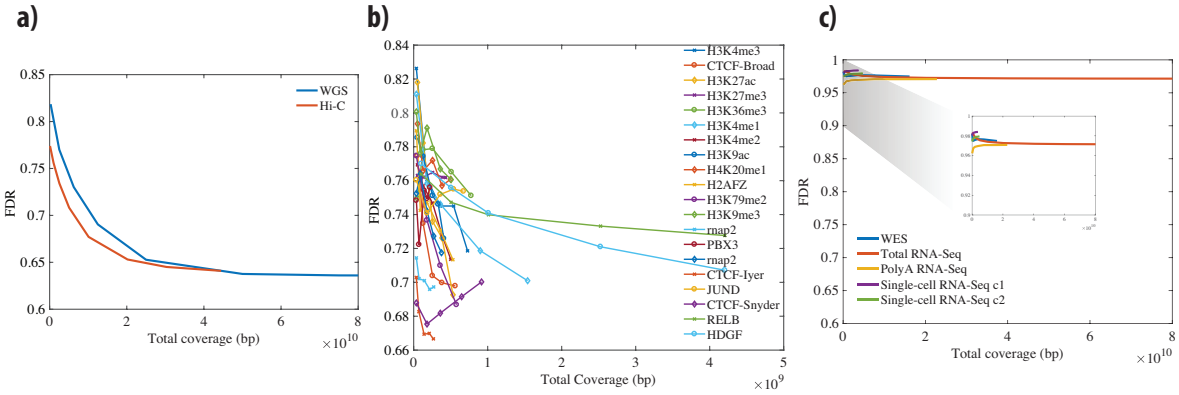


Figure 4: **False discovery rate of functional genomics experiments at different coverages** (**a**) FDR comparison for Hi-C and WGS data at different sampled coverages. (**b**) FDR comparison for different ChIP-Seq experiments at different coverages. (**c**) FDR comparison for WES and different RNA-Seq experiments.

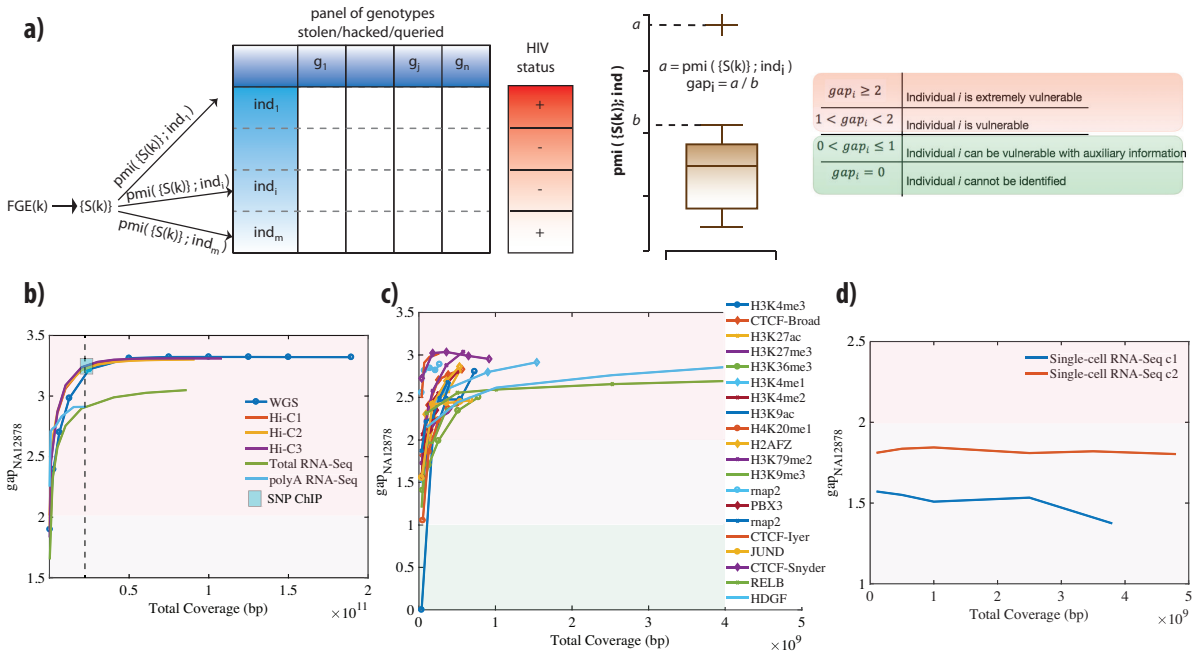


Figure 5: **Illustration of a linking attack and the accuracy of linking.** (**a**) The publicly available ananoymized reads from functional genomics experiments contains a set of variants and HIV status for the sample that the functional genomics experiment was performed at increasing coverages. The panel of genotypes contains the variants and associated genotypes for m individuals. The attacker links the inferred variants and genotypes to the panel of genotypes by using the best matched pointwise mutual information. The linking potentially reveals the HIV status for the linked individual. (**b**) Comparison of *gap* for NA12878 at different coverages for Hi-C and Total/PolyA RNA-Seq reads. WGS and SNP-ChIP are also added for comparison. (**c**) Comparison of *gap* for NA12878 at different coverages for 20 different ChIP-Seq experiments. (**d**) Comparison of *gap* for NA12878 at different coverages for single-cell RNA-Seq experiments.

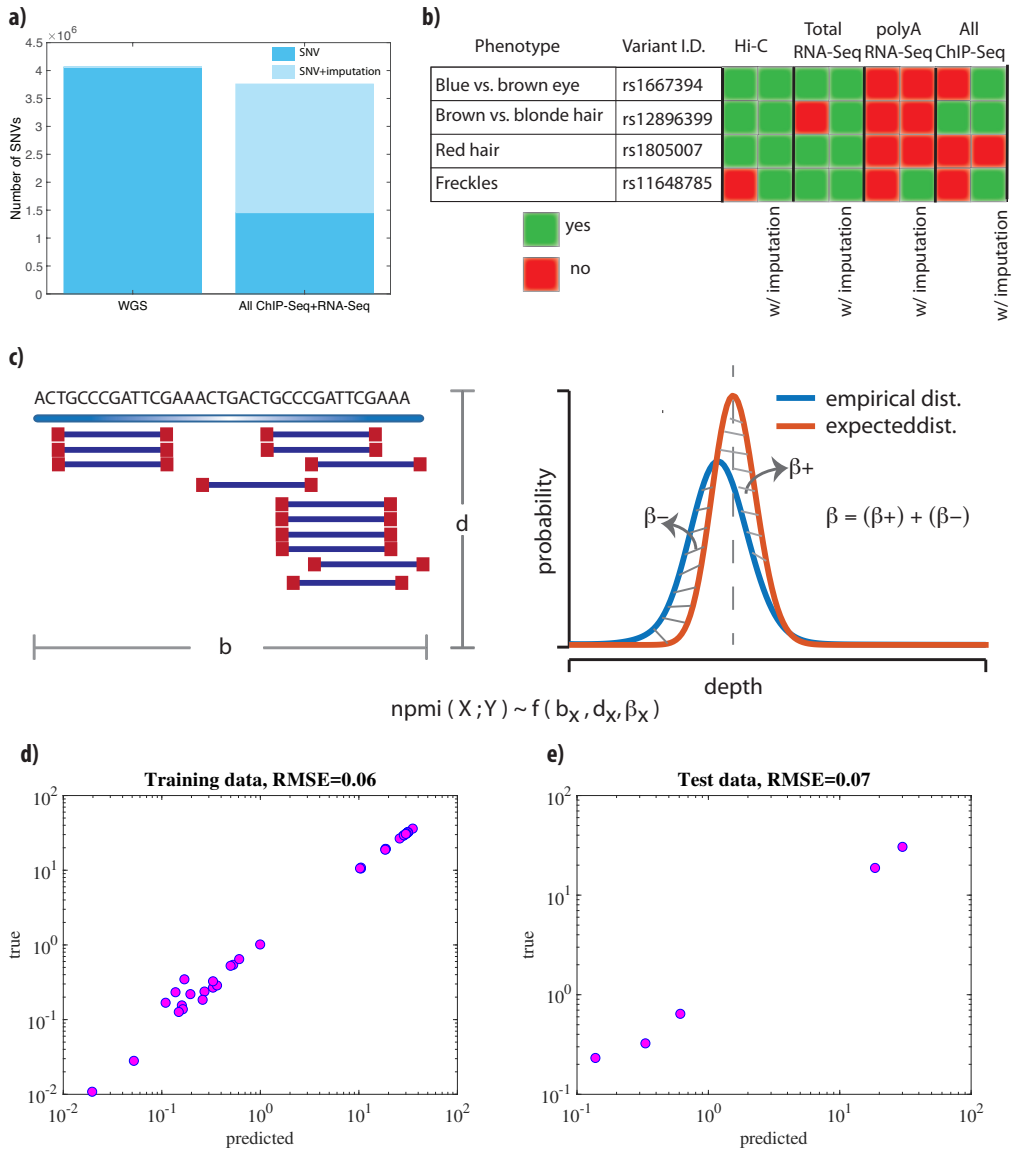


Figure 6: **Individual's genome can be approximated and sensitive phenotypes can be inferred from publicly available data by imputation and a theoretical framework for prediction of amount of leaked data** (**a**) Number SNVs called from WGS data and all of the ChIP-Seq and RNA-Seq data together with and without imputation. (**b**) Variants associated with physical traits and if they present in the called variants from different functional genomics experiments before and after imputation. (**c**) Features of the theoretical framework - write more. (**d**) Accuracy of fitted model on training set- write more (**e**) Accuracy of fitted model on test set - write more

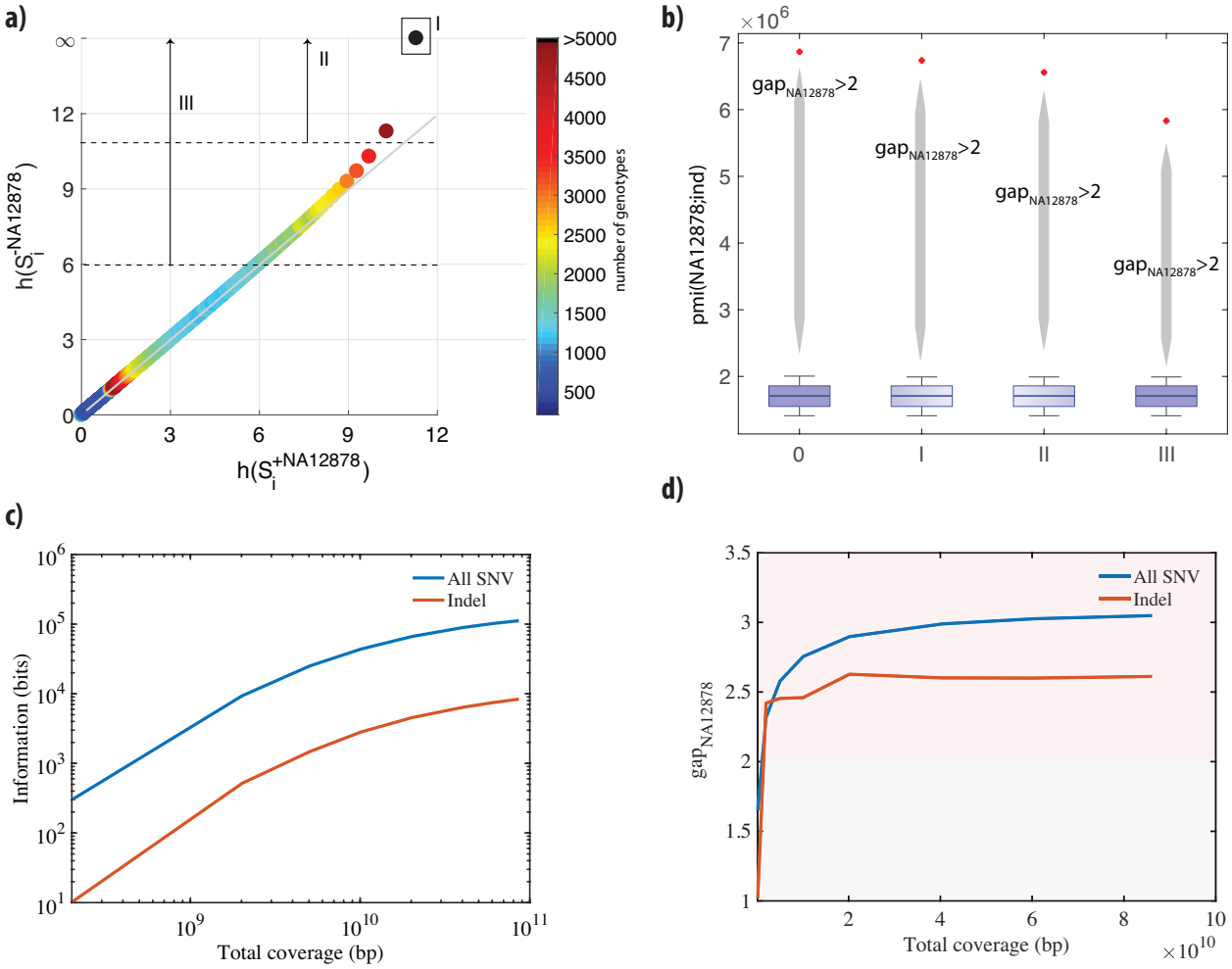


Figure 7: **Removal of rare variants and linking** (**a**) Information of the variant before and after addition of NA12878 to the population. We iteratively removed variants from the set as (I) only the variants that is only NA12878 specific, (II) the variants that have an information of 11 or higher bits after removal of NA12878 from the population, (III) the variants that have an information of 6 or higher bits after removal of NA12878 (**b**) Linking accuracy for every iteration of removal of NA12878 variants from the set. (**c**) Information of all the variants that are called from Total RNA-Seq reads vs. the information of the indels that are called from Total RNA-Seq reads. (**d**) Linking accuracy when we consider all the variants that are called from Total RNA-Seq rads vs. the linking accuracy when we consider only indels called from Total RNA-Seq reads.

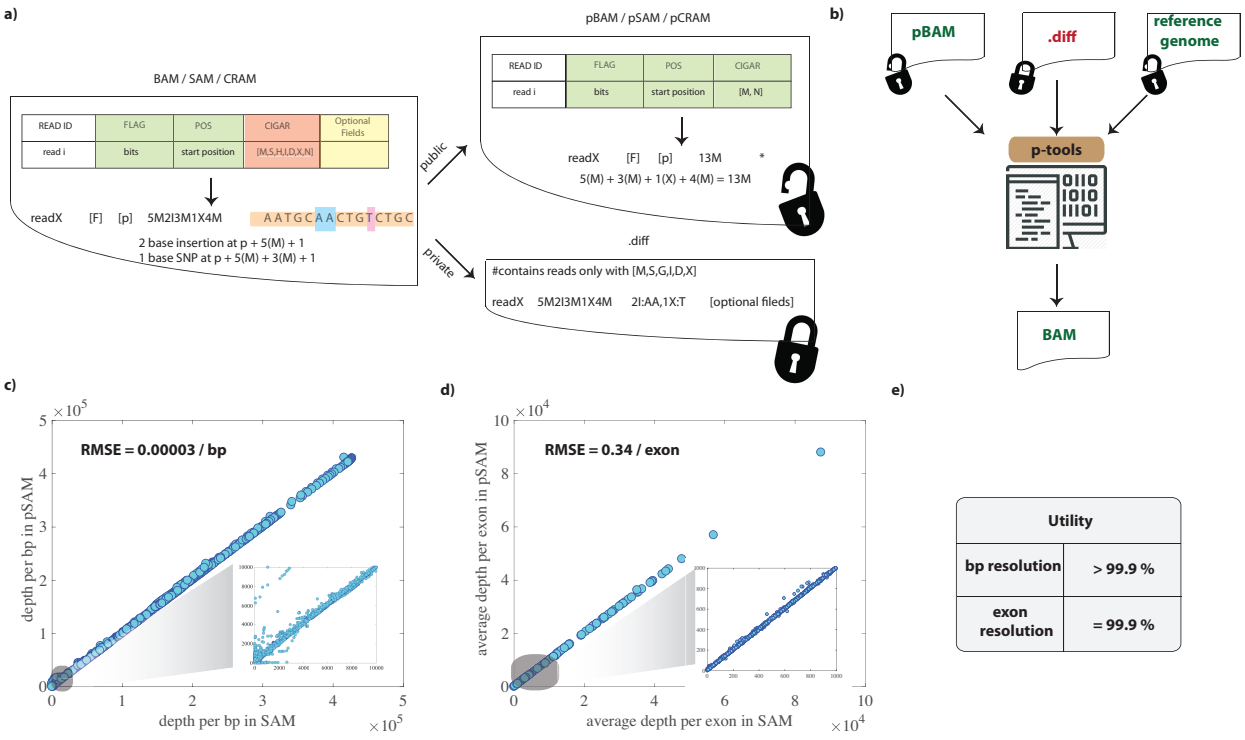


Figure 8: **Privacy-preserving file formats for mapped reads** (**a**) The generation of public pSAM and private .diff files. (**b**) Schematic of how to go between pBAM and BAM formats by utilizing the human reference (**c**) Comparison of nmber of reads for each basepair in the original SAM file and the distorted pSAM file. Noise is mostly introduced to basepairs with low depth. (**d**) Comparison of nmber of reads for each exon in the original SAM file and the distorted pSAM file. Noise is mostly introduced to exons with low expression.

1 Suplementary Information

1.1 Simulation of individuals

In this study, we simulated individuals that belong to European (CEU) and African (YRI) population. These individuals are simulated based on the genotype frequency of derived from the population in 1000 genomes panel [**?**] and the LD relationship of the population in HapMap project [**?**]. Once we determine the population that the simulated individual belongs to, we draw a random variant with its genotype based on the probability of seen that variant with the genotype in the population. We then determine all the other variants that are in LD with the drawn genotype. We use the LD correlation as the joint probability of two variants being observed simultaneously and based on that decide if the correlated variant will be simulated or not. The joint probability is adjusted based on the genotype of the two variants. If both variants are simulated as homozygous, then the joint probability is assumed to be equal to the LD correlation. If at least one of the variants is heterozygous, then the joint probability is assumed to be equal to the half of the LD correlation. We continue this process till we exhaust all the variants observed in the population. We first simulated 100 individuals that belong to CEU and then 100 more individuals that belong to YRI populations.

1.2 *KL*-Divergence

KullbackLeibler (KL) divergence is a measure to quantify th difference between the distribution of two probability distribution. For discrete probability distribuions P and Q, the *KL*-divergence from Q to P is calculated as [**?**]

$$D_{\mathrm{KL}}(P||Q) = -\sum_i P(i) \log \frac{Q(i)}{P(i)} \tag{1}$$

In the contxt of simulating individuals, we interpreted *KL*-divergence as the information gain achieved by the addition of a new individual to the population, i.e. $D_{\mathrm{KL}}(P^{n+1}||P^n)$, where n is

the size of the population before the addition of new individual.

1.3 Calculation of information after imputation

Using partial variant set when a variant imputed, naturally the amount of information gained from imputed variant is low as we have prior information on the probability of observing imputed variant due to the LD correlation. IMPUTE2 [**?**] prints a probability for each genotype for a given variant and an info column that is a measure for the confidence of observing the variant. Confidence value (c_i for variant s_i) is reported as a number between 0 and 1. We first removed all the imputed variants that have confidence below 0.3. We then selected the genotypes that have the highest probability for each imputed variant. As the confidence value gives a priori information on the probability of observing the imputed variant. We then calculated the information gain from the imputed variant s_i as,

$$h^{im}(i) = (1 - c_i)h(s_i) \tag{2}$$

1.4 Gaussian Process Regression (GPR) to estimate information from sequencing properties

In order to increase the number of data points, we addede sampled reads to the 24 functional genomics data. The added sample reads are from Hi-C as Hi-C experiments have large number of reads and sampling does not alter the depth distribution. Moreover, if the sampled reads are between the range of 1 million bp to 10 million bp, it can mimick ChIP-Seq data. This allowed us to have total of 45 data points. We found that normalized pmi for these data points ranges between 0.005 and 0.35. To avoid the problems due to the precision, we magnified npmis by a factor of α as following;

$$npmi(S^{FGE}; S^{GS}) = \frac{1}{\alpha} f(\overline{d}_{FGE}, b_{FGE}, \beta_{FGE})$$

We set α to 100. We then randomly selected 5 data points and removed them from the dataset to use it as independent testing case. For the reamining 40 data points, we tried many regression learners including linear regression, regression trees, Support Vector Machines, and Gaussian Process Regression. Although they all exhibit good prediction with low root mean square errors (RMSE), Gaussian Process Regression with an exponential kernel reported the lowest RMSE. The GPR is a non-parametric Bayesian approach, which is powerful to capture noisy relationships between inputs and output by optimizing large number of parameters hence allowing the level of complexity to be decided by the data through Bayesian inference [**?**]. We used MATLAB's Statistics and Machine Learning toolbox to perform fitting.

Privacy-enhancing file formats for functional genomics experiments

Privacy-enhancing file formats can be generated for SAM, BAM and CRAM files. For simplicity, we will refer the regular files as BAM and the privatized file format as pBAM. The difference between the regular files and the privatized files are on the fields of cigar, sequence and the alignment score. Let's assume read length for the sequencing experiment is 76, which is the total number of nucleotides in a fragment. Below are itemized description of how cigars are converted to privatized cigars along with examples.:
Cigars in non-intronic reads (i.e cigars with no 'N")

- Cigar for perfectly mapped reads is a number of read length followed by the letter 'M", indicating every nucleotide in the read is mapped to the reference human genome. This also means that there is no variant in this read. In this case, regular BAM has '76M" in the cigar and pBAM will have '76M" in the cigar as well.

- Cigar for reads that contain a mismatch is marked with the letter 'X". For example, if the 10th nucleotide in the fragment has a mismatch, then the cigar in the regular BAM becomes '9M1X66M". This usually means that there is a SNP on the 10th nucleotide of the fragment.

Since we know the start coordinate of the read from the regular BAM, an adversary can easily infer that there might be a SNP on the '$start + 10$"th coordinate of the genome of the sample. To prevent that we convert '9M1X66M" to '76M" in the pBAM file. This conversion does not add any noise to the results since '$start + 10$"th is observed in the functional genome, however as a different letter and processing of functional genomics data deals with the depth rather than the letter of the nucleotide.

- Cigar for reads that contain soft-clipping is marked with the letter 'S". For example, if the first 5 nucleotides are soft-clipped from the fragment, then cigar becomes '5S71M". The start coordinate reported as the beginning of mapped nucleotides, which is the 6th nucleotide of the fragment. In this case we report the cigar as '76M" and keep the start coordinate as it is. This is because soft-clipping can be due to a structural variant, insertion or a deletion. The associated noise with this conversion is that the coordinates between '$start + 72$" and '$start + 76$" gain extra read, i.e depth.

- Above point applies for the reads with hard-clipping that are marked by the letter 'H". For example, if the nucleotides from 6th to 25th are hard-clipped from the fragment, then cigar becomes '5M20H51M". In this case we report the cigar as '76M" ignoring the hard-clipped nucleotides. The associated noise with this conversion is that the coordinates between '$start + 6$" and '$start + 25$" gain extra read, i.e depth.

- Cigar for reads that contain an insertion is marked with the letter 'I". For example, if the 23th to 30th nucleotide in the fragment is an insertion, then the cigar in the regular BAM becomes '22M8I46M". Since we know the start coordinate of the read from the regular BAM, an adversary can easily infer that there is an indel on the '$start + 23$"th coordinate of the genome of the sample. To prevent that we convert '22M8I46M" to '76M" in the pBAM file. The associated noise with this conversion is that the coordinates between '$start + 22 + 46$" and '$start + 22 + 46 + 8$" gain extra read, i.e depth.

- Cigar for reads that contain an insertion is marked with the letter ‘D”. For example, if the 13th to 14th nucleotide in the fragment is a deletion, then the cigar in the regular BAM becomes ‘12M2D62M”. Since we know the start coordinate of the read from the regular BAM, an adversary can easily infer that there is an indel on the ‘$start + 12$”th coordinate of the genome of the sample. To prevent that we convert ‘12M2D62M” to ‘76M” in the pBAM file. This conversion does not add any noise to the results, because if there is high depth around these 2 deleted nucleotides, there is functional enrichment in that fragment regardless of the deletion. This also prevents signal profiles to leak the small deletions as the curve that corresponds to the deletion will look smooth based on its neighboring nucleotides.

- There are also cigars that may have multiple of the above letters. Here are a few examples and the solution:

  - Cigar ‘3S15M3I40M2D13M” becomes ’76M”, which introduces noise to only to 6th nucleotides that are on the coordinates between ‘$start + 70$” and ‘$start + 76$”.

  - Cigar ‘25H30M2X5M3D11M” becomes ’76M”, which introduces noise to only to 6th nucleotides that are on the coordinates between ‘$start + 52$” and ‘$start + 76$”.

Cigars in intronic reads (i.e cigars with ‘N”)

- Cigar for perfectly mapped reads but split due to the introns are split by the letter ‘N”. For example, if there is a 1000 nucleotide long intronic region between mapped regions, it can have a cigar as ‘30M1000N46N”. In this case pBAM will have a cigar of 30M1000N46N” as well.

- If the reads are split in the mapped regions due to mismatch, insertion, deletion or clipping, then pBAM deals with them such that splice sites are accurate. Here are few examples;

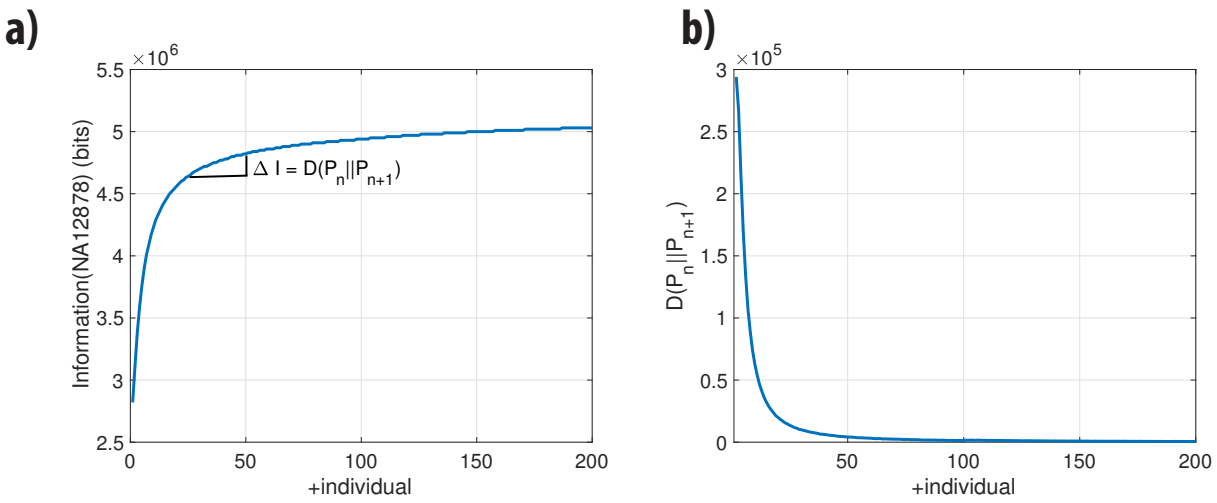
  - Cigar ‘3S30M1000N15M2D26M” becomes ’30M1000N46M”. This does not add any noise to the final signal.

– Cigar ‘10M3I23M1000N20S20M” becomes ’33M1000N43M”, which introduces noise to 20 nucleotides that are on the coordinates between ‘$start + 33 + 1000 + 21$” and ‘$start + 33 + 1000 + 40$”.

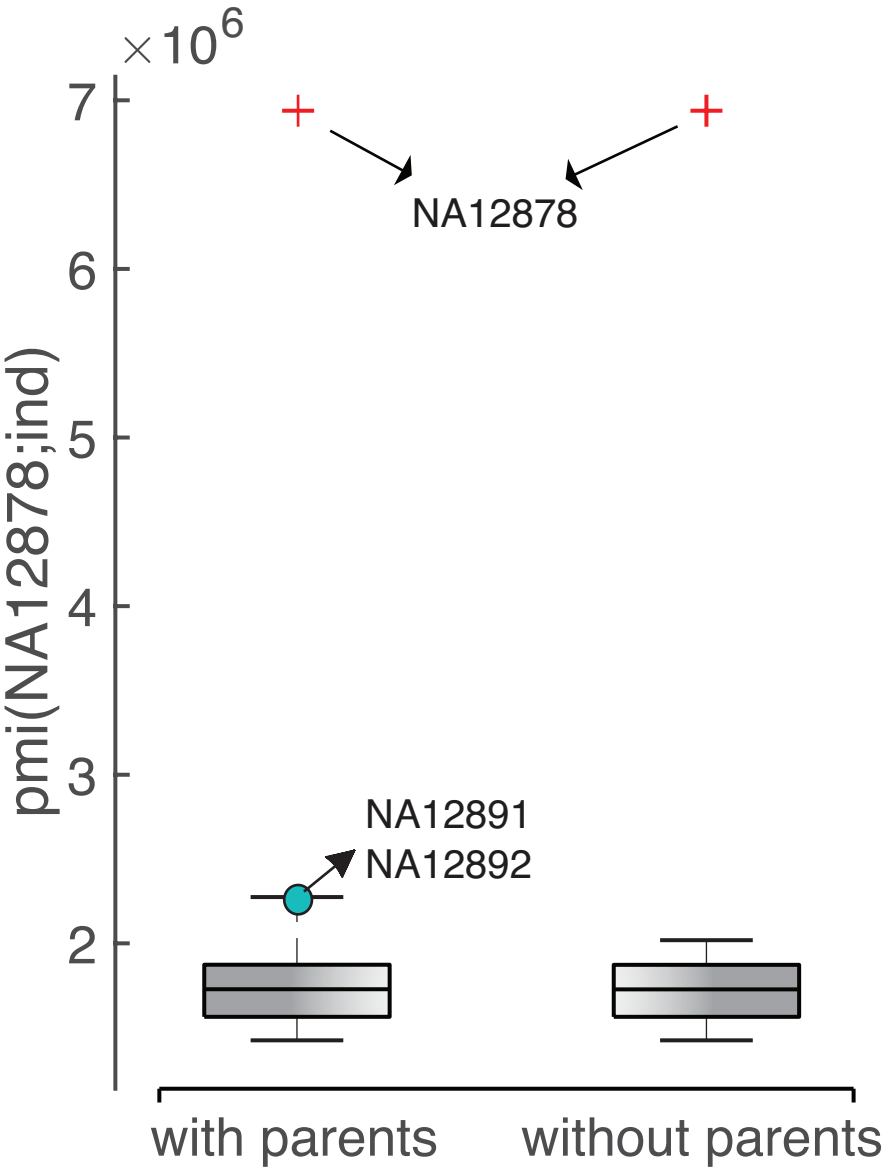
– Cigar ‘10M3D23M1000N30M2I8M” becomes ’33M1000N43M”, which shifts the intronic regions 3 nucleotides to the right.

Details of these examples are depicted in **??**.

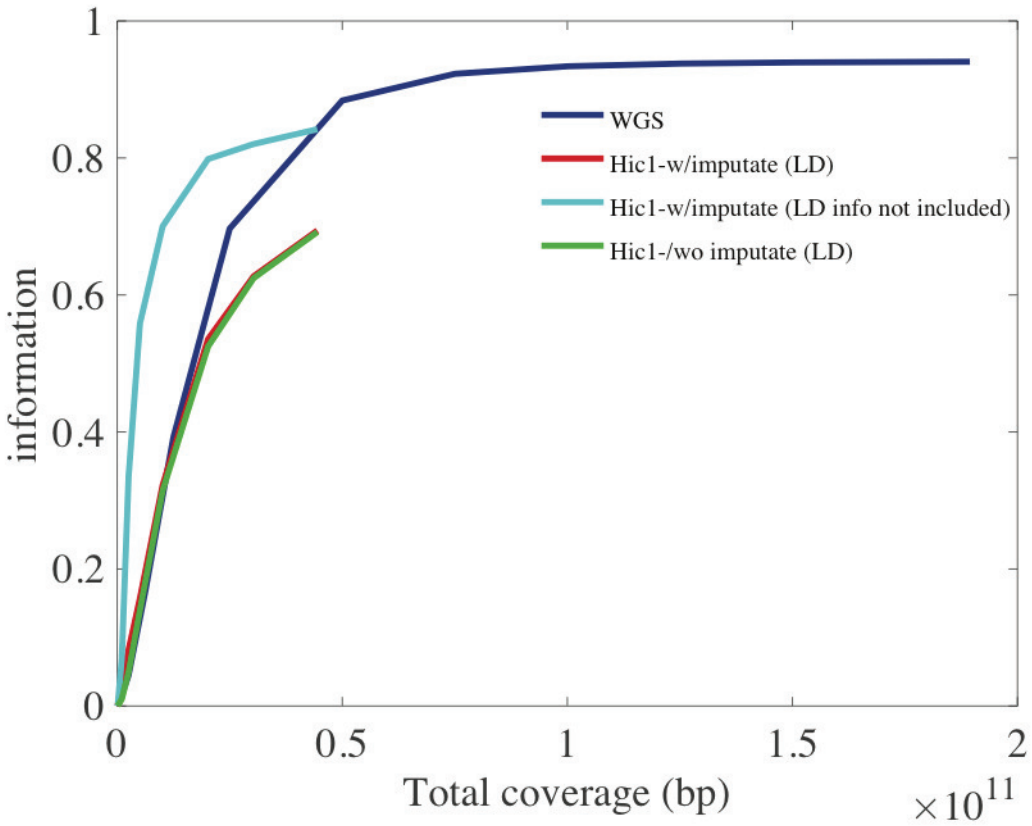
We remove the sequences when we convert BAM files to pBAMs. We also change the alignment scores in the pBAM files. Alignment scores vary depending on the sequencer. But, typically highest score corresponds to the perfectly mapped fragments. Therefore, if we see a alignment score smaller than the maximum in the files, we convert it to the maximum score. For example, if the maximum alignment score for the sequencer is 152 for 76bp length read. If there is any read that is reported as ‘AS:0:140”, the it is reported as ‘AS:0:152” in pBAM.



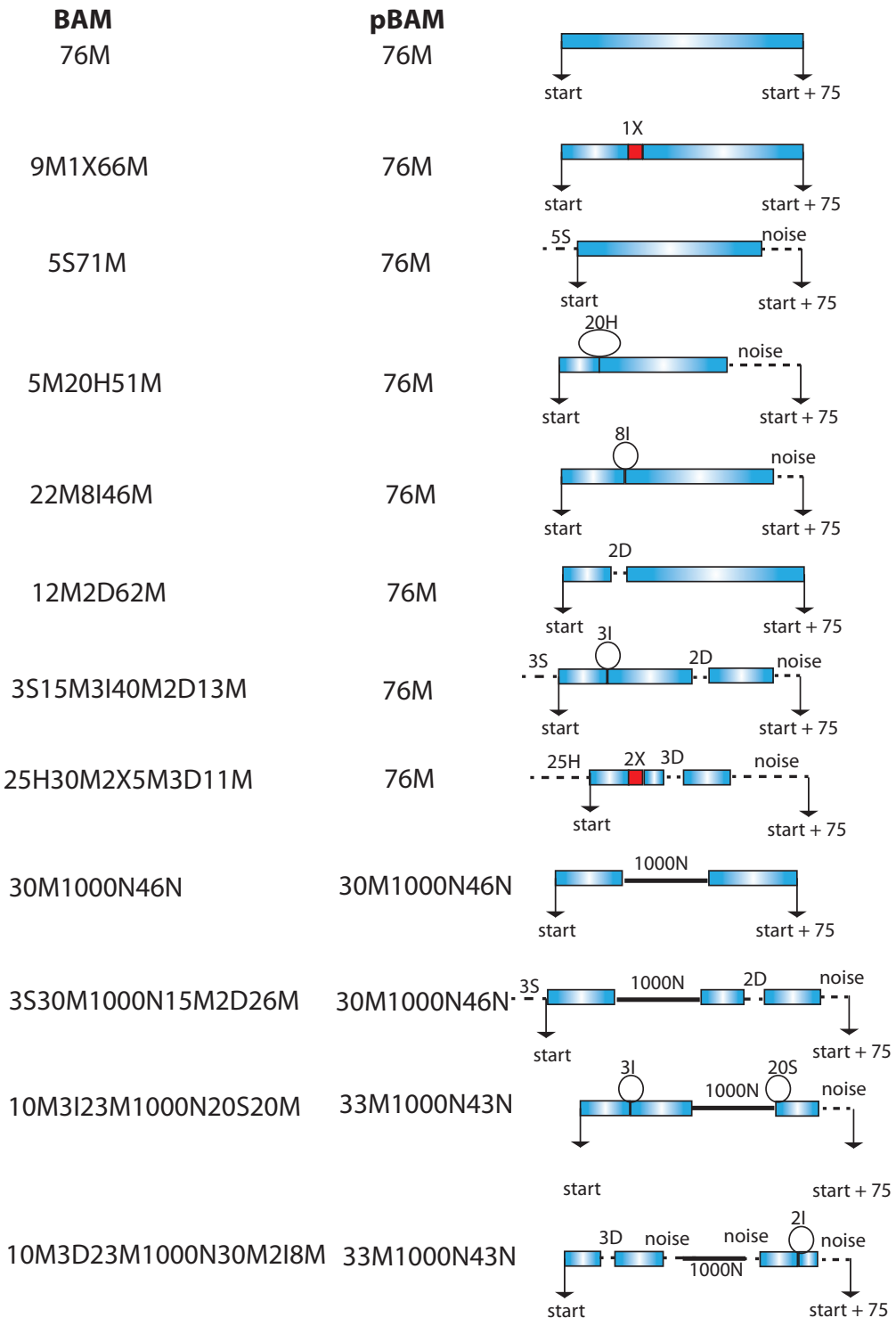
Supplementary Figure 1: **Change in information as the population size increases** (**a**) Information vs. the increasing number of individuals in the population. (**b**) *KL*-divergence between the population with n individuals vs. $n+1$ individuals.



Supplementary Figure 2: **The distribution of pmi values when the parents of NA12878 are added to the 1000 genomes genotype panel.**



Supplementary Figure 3: **The contribution of imputed variants to the naive information** When we do not consider the a priori information we obtained from imputation, we infliate the information gain from the imputed variants (cyan curve). When we remove the a priori information from the information gain, it shows that there is negligible information gain from the observation of imputed variants due to high correlation (red curve). The information before the imputation is depicted with green curve.



Supplementary Figure 4: **Visual representation of mapped fragments before and after converting the cigars for pBAM file format.** The insertions, deletions, soft and hard-clipping as well as intronic reads are depicted. The noise that is added to the pBAM file in order to enhance privacy is also depicted in the fragments.

Supplementary Table 1: The functional genomics experiments used in this study with their total coverages

ENCODE ID/Source	Experiment	# of Reads	Read Length
1kG	WGS	757,704,193	255
1kG	WES	212,461,381	76
Rao et al. 2014	Hi-C exp 1 PE1	219,616,072	101
Rao et al. 2014	Hi-C exp 1 PE2	220,087,882	101
Rao et al. 2014	Hi-C exp 2 PE1	448,843,710	101
Rao et al. 2014	Hi-C exp 2 PE2	451,088,484	101
Rao et al. 2014	Hi-C exp 3 PE1	536,684,803	101
Rao et al. 2014	Hi-C exp 3 PE2	536,101,709	101
ENCSR000CVT	Total RNA-Seq	227,501,266	202
ENCSR000COQ	PolyA RNA-Seq	267,602,146	76
ENCSR000AJA	Single-cell RNA-Seq1	38,377,124	100
ENCSR000AJH	Single-cell RNA-Seq2	47,896,396	100
ENCSR000AKF	H3K4me1	42,763,056	36
ENCSR145XQO	HDGF	41,626,373	101
ENCSR387QUV	RELB	25,652,682	101
ENCSR000DZN	CTCF-Snyder	25,463,397	36
ENCSR000AKA	H3K4me3	20,221,959	36
ENCSR000DYS	JUND	18,701,295	36
ENCSR000AOW	H3K79me2	16,073,184	36
ENCSR000AKE	H3K36me3	15,239,685	51
ENCSR000AOV	H2AFZ	14,724,790	36
ENCSR000AOX	H3K9me3	14,049,420	36
ENCSR000AKB	CTCF-Broad	11,026,086	51
ENCSR000BIF	rnap2	10,428,778	36
ENCSR000AKC	H3K27ac	10,410,928	51
ENCSR000AKG	H3K4me2	9,815,194	51
ENCSR000AKI	H4K20me1	9,757,368	51
ENCSR000AKD	H3K27me3	8,454,639	51
ENCSR000AKH	H3K9ac	7,981,456	51
ENCSR000DKV	CTCF-Iyer	7,614,943	35
ENCSR000BGD	rnap2	7,516,461	36
ENCSR000BGR	PBX3	6,119,046	36