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

GG et al

January 20, 2018

Abstract

179 words

[GG2MG: I tried to introduce set point]

Functional genomics experiments provide important insight on genomic activities such as gene expression levels or transcription factor enrichment essential for personalized medicine, thus publicly sharing of such data is extremely valuable for biomedical research. Although functional genome is not necessarily tied to an individual's genotype, extracting information from these experiments require data types, which potentially leak sensitive information. While progressive summarization such as gene expression quantifications provide leakage free data types, current policies regarding functional genomics data sharing are ad-hoc and a systematic study that quantifies the amount of leakage and sets a point, after which breaching privacy is minimal is lacking. Here, we study the quantification of sensitive information in the raw reads of functional genomics data by deriving information theory based measures. We show that functional genomics reads 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 allows accurate quantification of genomic activities 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].

[GG2MG: This paragraph defines the set point and the first motivation of the paper: systematically quantify information to define set point]

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). Inferring biological information from functional genomics experiments is a several steps procedure, in which progresive summarization of the data from raw sequencing reads to the gene quantifications, TF binding peaks or chromatin interaction matrices is performed. Here we introduce the notion of ‘set point”, which allows us to determine the summarization step where amount of sensitive information leakage is minimal. Figure 1) illustrates the trade-off between the amount of data and the information leakage. In detail, functional genomics data analysis 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]. Further summarization of the mapped reads (such as signal profiles or gene expression quantification) still allows researchers to make accurate biological conclusions, while providing further data reduction of a ∼20 fold. Although overall aggregation and averaging reduces biological information, private information leakage also decreases (Figure 1). A hurdle in determining the ‘set point” is the lack of systematic quantification of private information leakage from the functional genomics data. In particular, BAM files are of great interest due to the large amount of biological data they provide as they constitute the most important input of majority of genome annotation pipelines. On the other hand, these files contain sequence information of the individual that may leak sensitive data. Depending on the depth of the functional genomics experiment, raw reads can be used to identify the private SNPs, small indels, and structural variants. However, current policies related to public sharing of the BAM files are somewhat ad-hoc. For example, the genome of HeLa cell line and the raw reads from Hi-C experiments require special access, while reads from ChIP-Seq and RNA-Seq experiments are publicly available [19]. That is, reads from the experiments that do not require substantial depth are sometimes considered to be safe to share without privacy concerns owing to partial and biased sequencing, although it is not clear if these reads are safe to share. Although private information leakage from summary level functional genomics data are quantified previously [**?**, **?**] the lack of a systematic quantification of private data leakage from BAM files makes it difficult for biomedical data sharing policy makers to protect individual's sensitive information in a consistent fashion.

[GG2MG: This paragraph mentions about the challenges related to special access data, utility of the data and the second motivation of the paper: create another layer in the data stack by convertim BAMs to pBAM, I also changed the first figure]

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 bureaucracy 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 bureaucracy. 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]. In an attempt to consider both sides 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 by adding another layer to the data pyramid, which in turn helps with the solving the complexity associated with private biomedical data sharing (Figure 1). 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. Specifically, adversary is interested in investigating whether noisy and partial reads from functional genomics experiments can be used as quasi-identifiers and how accurate the genotyping need to be in order to link individuals to databases.

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 ($\overline{d}$) , 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. We first speculated that the removal of the variants that are unique to NA12878 might be enough to fail at linking. A total of 11,472 variants along with their genotypes are only observed in NA12878, which we refer as 'unique variants" (Fig. 6a). After the removal of unique variants from the NA12878 variant set, we calculated the $gap_{NA12878}$ and surprisingly found that linking accuracy is affected minimally compared to using the all of NA12878 variants (Fig. 6b). We then created another set ('double variants", Fig. 6a), that includes the variants that are observed in NA12878's genome as well as one more individual in the 1000 genomes genotype panel (total of 16,305 genotypes). We again found that individual is extremely vulnerable to linking attacks ($gap_{NA12878} > 2$,Fig. 6b). We then relaxed our cut-off further to remove the variants that are observed in NA12878's genome as well as at most 1.5% of the population ('rare variants", total of 124,093 genotypes, Fig. 6a). This also did not affect the overall linking ($gap_{NA12878} > 2$,Fig. 6b).

These rare genotypes are observed in 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 we can still link NA12878 to the correct genotypes succesfully 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 reads introduce noise to the signal profiles 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. Although they aim at answering questions related to genomic activities such as gene expression, TF binding or 3D organization of genome, public sharing of sequencing data from these experiment can lead to recovery of genotype information and in turn raise privacy concerns. However, the systematic quantification of private information content of the functional genomics BAM files and open access to such data without comprimising individuals' identity have not been well studied. Current policies regarding to public sharing of functional genomics BAM files are ad-hoc. The experiments that require high depth of sequenicng such as Hi-C and sometimes RNA-Seq are considered to be private, while relatively low depth BAM files such as those from ChIP-Seq are often shared publicly. In this study, we derived information thery based measures to systematically quantify the sensitive information leakage in the BAM files of functional genomics experiments in low and high depth experiments.

Instantiation of linking attacks by genotyping of partial or complete functional genomics data showed that even at low coverages of low depth experiments such as ChIP-Seq, linking individuals to the databases can be done without error. When we compare the linking accuracy to the false discovery rate, we found that it is easier to link individuals to the databases than genotyping them accuractely using functional genomics experiments. The implication is that noisy quasi-identifers, i.e bad quality SNP calling, can be used to link the data to the high quality genotypes. For example, according to our calculations, reads from singel-cell RNA-Seq data carry the most amount of noise. This is likely due to the bias towards expressed genes in such small amount of cells, mapping issues of splice sites, false positives from RNA editing sites and amplification bias. However, the noisy genotypes called from small amount of cells, even when the number of reads are only a million, are quasy-identifiers that result in very high linking accuracy. This is worrisome in terms of biomedical data sharing as the number of individuals in genotype databases is increasing exponentially with the decrasing cost of sequencing. Furthermore, rich information about an individual's identity and his/her sensitive phenotypes can also be inferred by combining the reads from low depth functional genomics experiments and through genotype imputation.

In this study, we introduce the concept of ''set point" in determining the data production steps, where sensitive information leakage transitions from unsafe to safe to share (Fig. 1). Setting a ''set point" is possible by systematic genotyping and quantification of information. Although it is obvious that any DNA read contains variants, it is not trivial to understand the amount and the quality of sequencing to do accurate genotyping. Moreoever, we showed that genotyping accuracy of a functional genomics sample and the ability to link individuals to the databases using the same sample are not necessarily linearly correlated. It is easier to link individuals to the databases and infer their complete variant sets than genotyping a sample with accuracy and minimal false discovery. For example, complete set of variants from HeLa's genome may not be obtained by genotyping HeLa BAM files from functional genomic experiments. However, using only a small number of reads from the same BAM files accurate linking attacks are plausable. Nevertheless, policies governing public sharing of HeLa genome vs. HeLa functional genomics reads is ad-hoc and contradictory. Therefore, it is essential to quantify the information in samples and set the ''set point" accurately. On the other hand, 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 an evolving field, hence extensive public sharing of functional genomics data accelerate collaborative research and reproducibility by removing the complexities associated with data accession procedures. Having a clear definition of ‘’set point” helps researcher to develop means to push the ‘’set point” in order to enable the public sharing of functional genomics data without comprimising the privacy of the individuals.

In order to overcome bottlenecks related to data sharing and answer privacy-concerns, researchers proposed solutions such as differential privacy [?, ?, ?]. It has shown that retriving information from private statistical databases without revealing some amount of information is impossible [?]. Furthermore, entire database can be inferred by using a small number of queries. Differential privacy ensures a high level of privacy such that adversary retrieves similar result with and without the addition of the individual's data to the database [?]. An algorithm ($\mathscr{A}$) to a dataset is considered to be ε-differentially private if the results satisfies the condition $Pr[\mathscr{A}(D_1) \in S] \leq e^{\varepsilon} \times Pr[\mathscr{A}(D_2) \in S]$, such that datasets D_1 and D_2 that differ on a single element retrieve similar results. Although, differentially private solutions by definition are applied to databases that contain multiple individual's information and D_1 and D_2 differ on the data of one person, we applied a similar logic to development of means to share raw reads from functional genomics data without comprising individual's sensitive information. That is, we assume a BAM file of a functional genomics experiment is the dataset D_1, where each record is a string that tells the number of matched, mismatched, inserted or deleted basepairs in a read (also known as a ‘’cigar”). The new file format pBAM is the dataset D_2, where each record differs from D_1 by the existence of a variant. For example, a record in D_1 tells that its correponding read has 2 insertion and 70 matches, while the same record in D_2 tells same read has 72 matches. If our algortihm $\mathscr{A}$ calculates the depth signal profiles and the ratio between the results of D_1 and D_1 is $\leq e^{\varepsilon}$, then we can claim that pBAM files are ε-differentially private. The advantage of such formal definition is that the value of ε makes the trade-off between the utility and the privacy of the data adjustable. For example, we showed that we can remove all the variants from BAM files and reach more than 99% recovery in depth signal profiles and gene quantifications, which yields an ε of 0.01 for our pBAM files (see SI for the calculation of this value). Revealing more variants from BAM files will yield smaller ε, while increasing privacy risk.

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]] Presented framework can be used for 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.

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.

Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03). ACM, New York, NY, USA, 202-210.

Dwork, Cynthia (2008-04-25). "Differential Privacy: A Survey of Results". In Agrawal, Manindra; Du, Dingzhu; Duan, Zhenhua; Li, Angsheng. Theory and Applications of Models of Computation. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 119

List of Figures

1 stack 28
2 comparison and procedure 29
3 pmi for all 30
4 FDR 31
5 linking and gap 32
6 imputation, phenotype inference and theoretical framework 33
7 differential privacy 34
8 pBAM 35

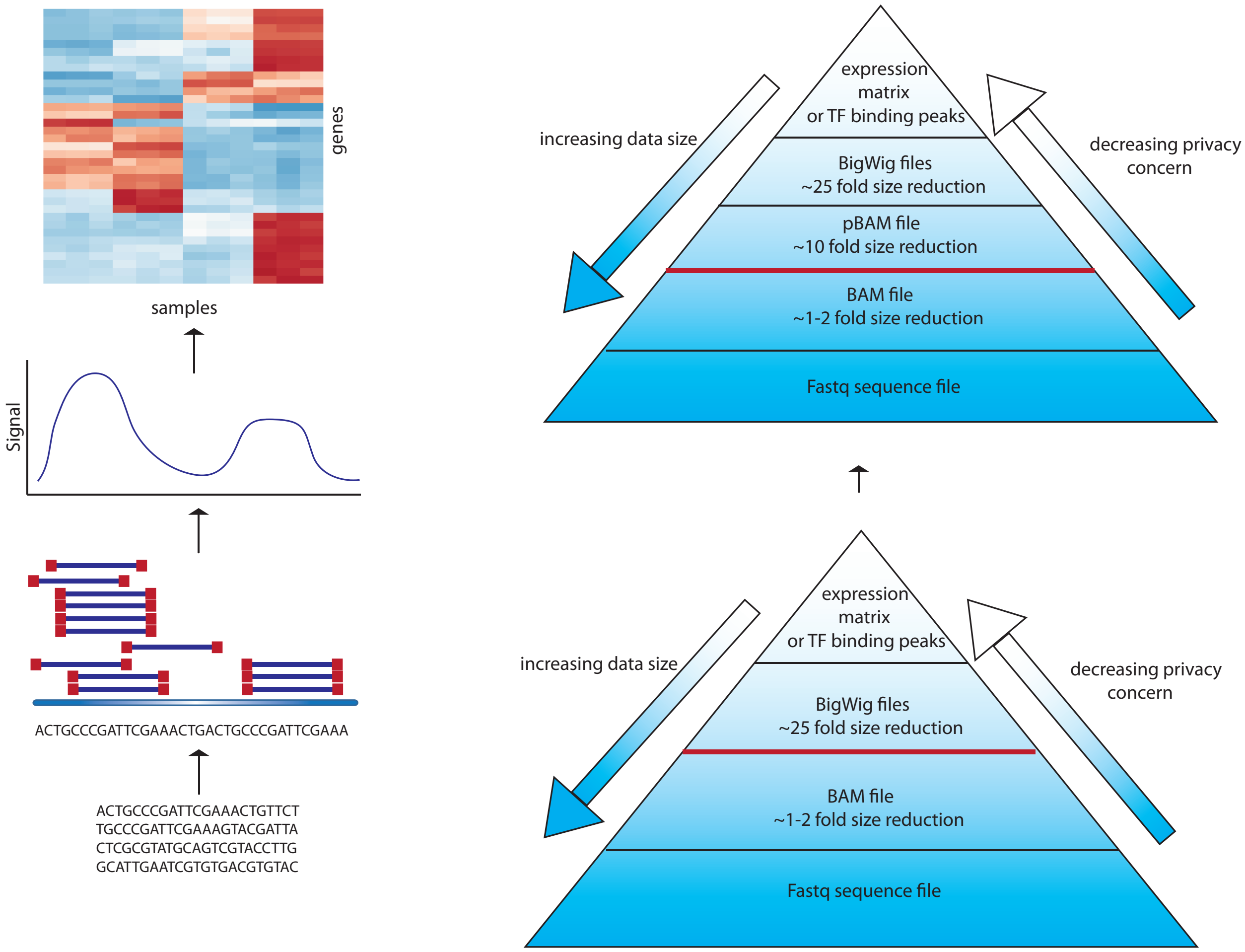

genes
samples
Signal
ACTGCCCGATTCGAAACTGACTGCCCGATTCGAAA
ACTGCCCGATTCGAAACTGTTCT
TGCCCGATTCGAAAGTACGATTA
CTCGCGTATGCAGTCGTACCTTG
GCATTGAATCGTGTGACGTGTAC
increasing data size
decreasing privacy concern
expression matrix or TF binding peaks
BigWig files ~25 fold size reduction
pBAM file ~10 fold size reduction
BAM file ~1-2 fold size reduction
Fastq sequence file
increasing data size
decreasing privacy concern
expression matrix or TF binding peaks
BigWig files ~25 fold size reduction
BAM file ~1-2 fold size reduction
Fastq sequence file

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.

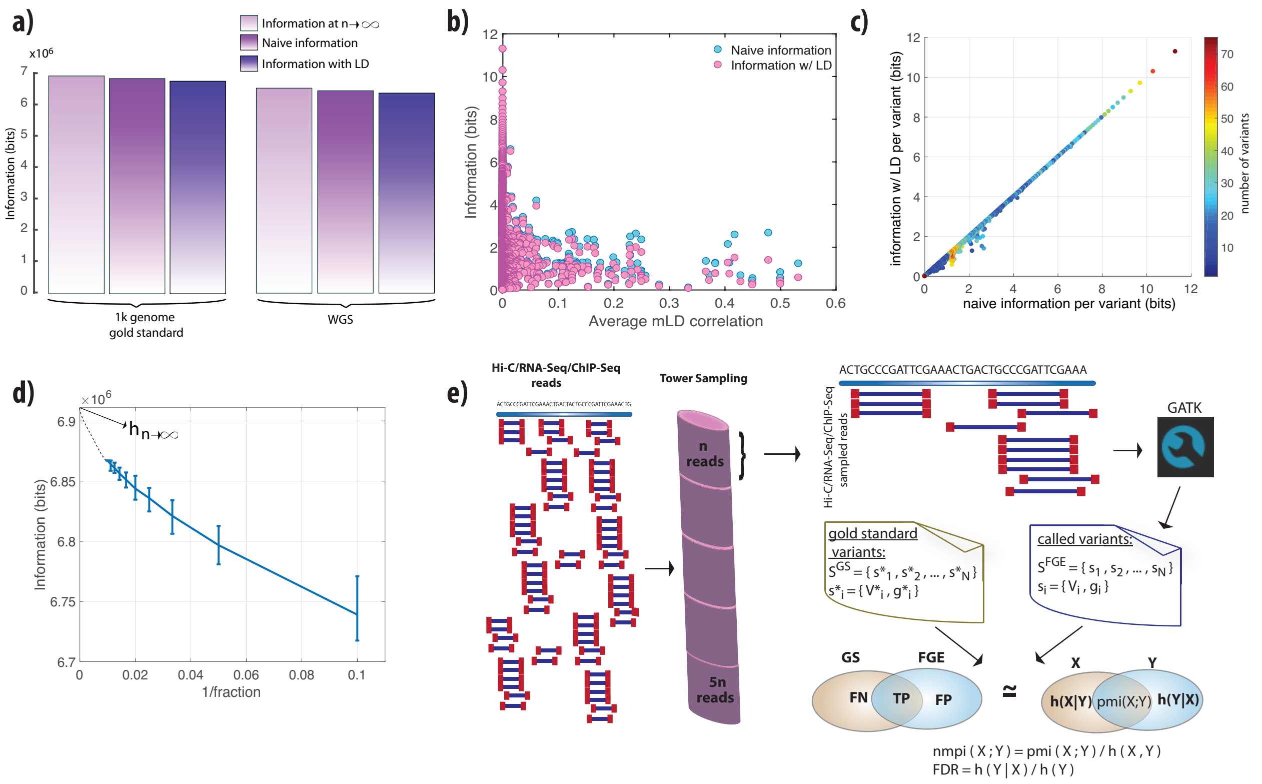


Figure 2: **Comparison of naive information measure with information with LD consideration and sample size correction.** (**a**) Difference between the naive information, information with LD consideration and extrapolated information when population size is infinite. (**b**) The maximum LD score for 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× 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.

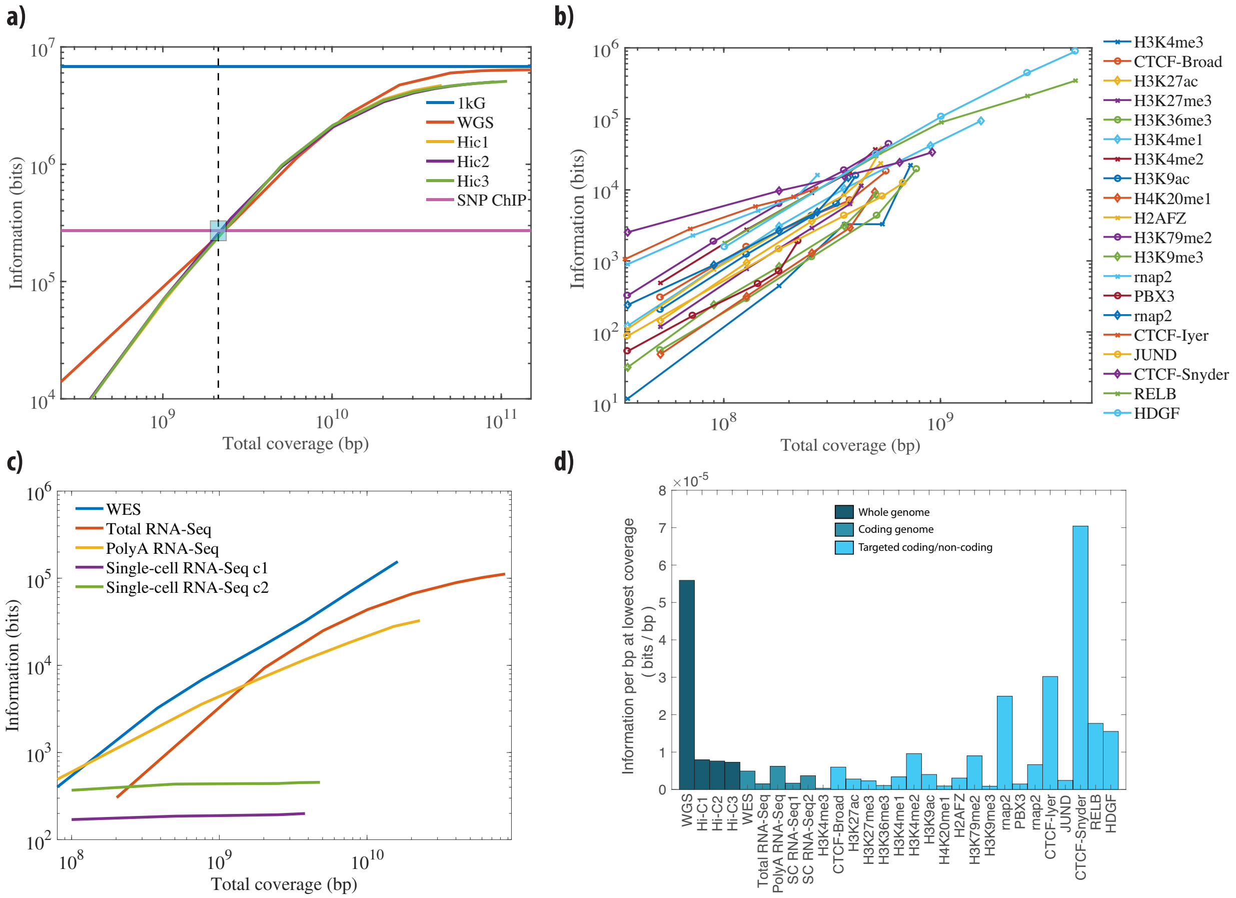


Figure 3: **The pointwise mutual information calculated for 24 different functional genomics assays and WGS, WES and SNP ChIP data using NA12878 1000 genomes variants as gold standard.** (**a**) The pmi values for WGS and three different primary Hi-C experiments plotted at different coverages. The information 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.

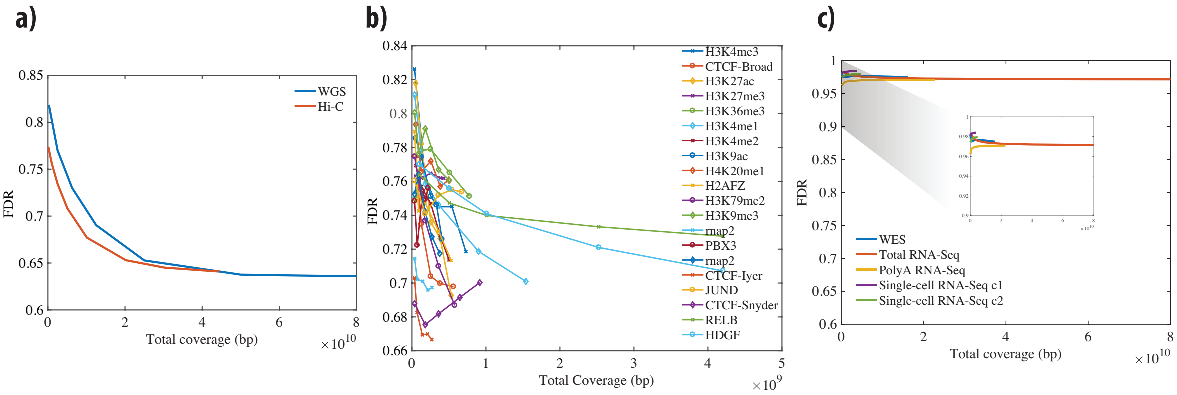


Figure 4: **False discovery rate of functional genomics experiments at different coverages** (**a**) FDR comparison for Hi-C and WGS data at different sampled coverages. (**b**) FDR comparison for different ChIP-Seq experiments at different coverages. (**c**) FDR comparison for WES and different RNA-Seq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

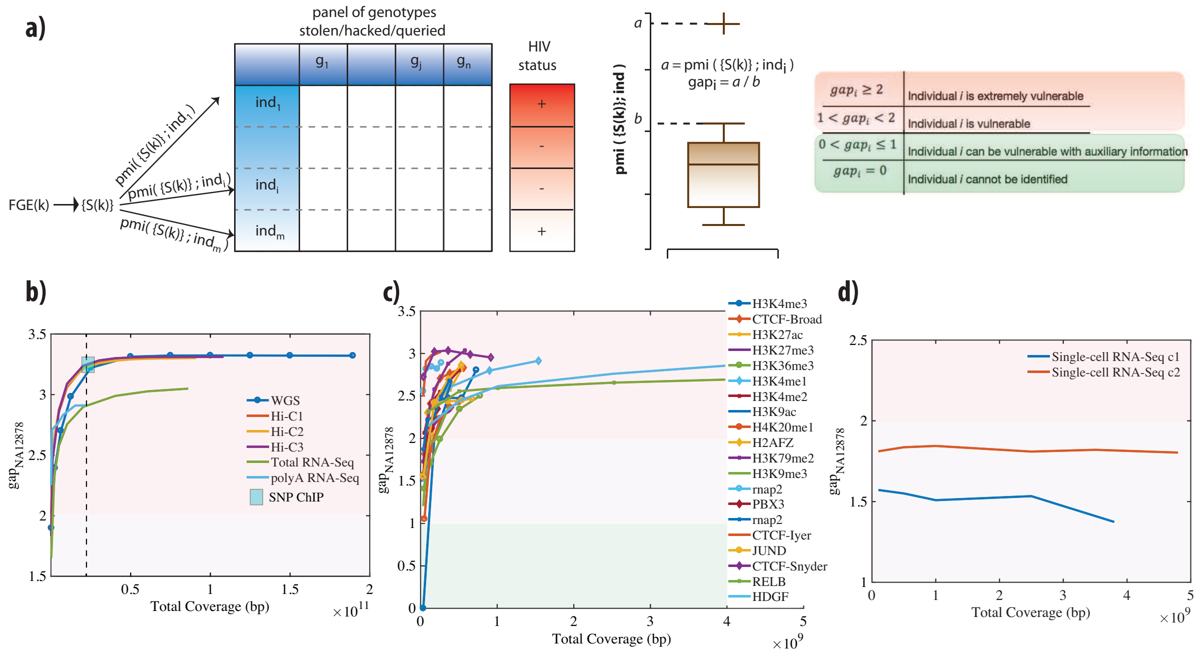
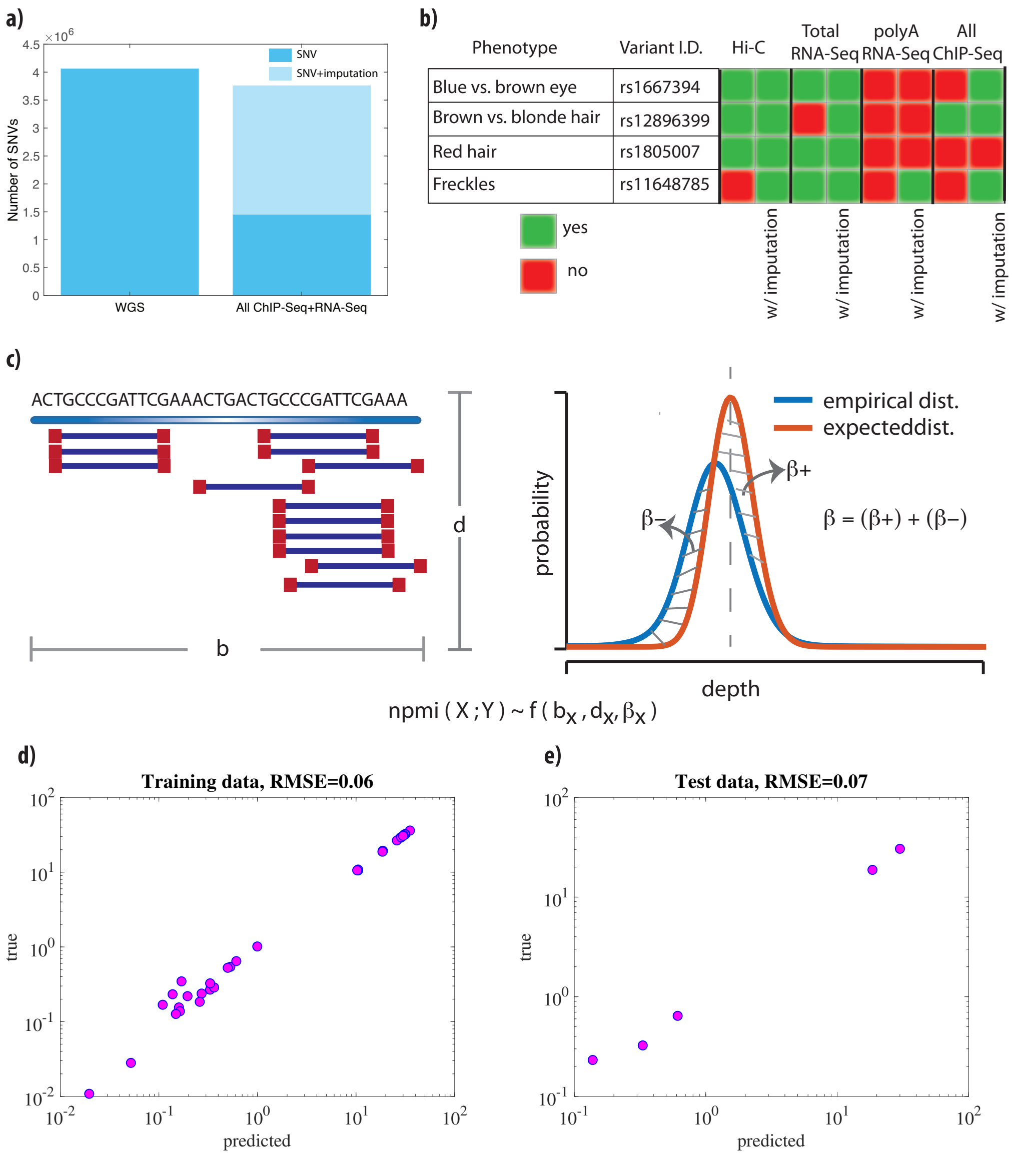


Figure 5: **Illustration of a linking attack and the accuracy of linking.** (**a**) The publicly available ananoymized reads from functional genomics experiments contains a set of variants and HIV status for the sample that the functional genomics experiment was performed at increasing coverages. The 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.



Phenotype	Variant I.D.	Hi-C	w/ imputation	Total RNA-Seq	w/ imputation	polyA RNA-Seq	w/ imputation	All ChIP-Seq	w/ imputation
Blue vs. brown eye	rs1667394	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟩
Brown vs. blonde hair	rs12896399	🟩	🟩	🟥	🟩	🟥	🟥	🟩	🟩
Red hair	rs1805007	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥
Freckles	rs11648785	🟥	🟩	🟩	🟩	🟥	🟩	🟥	🟩

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

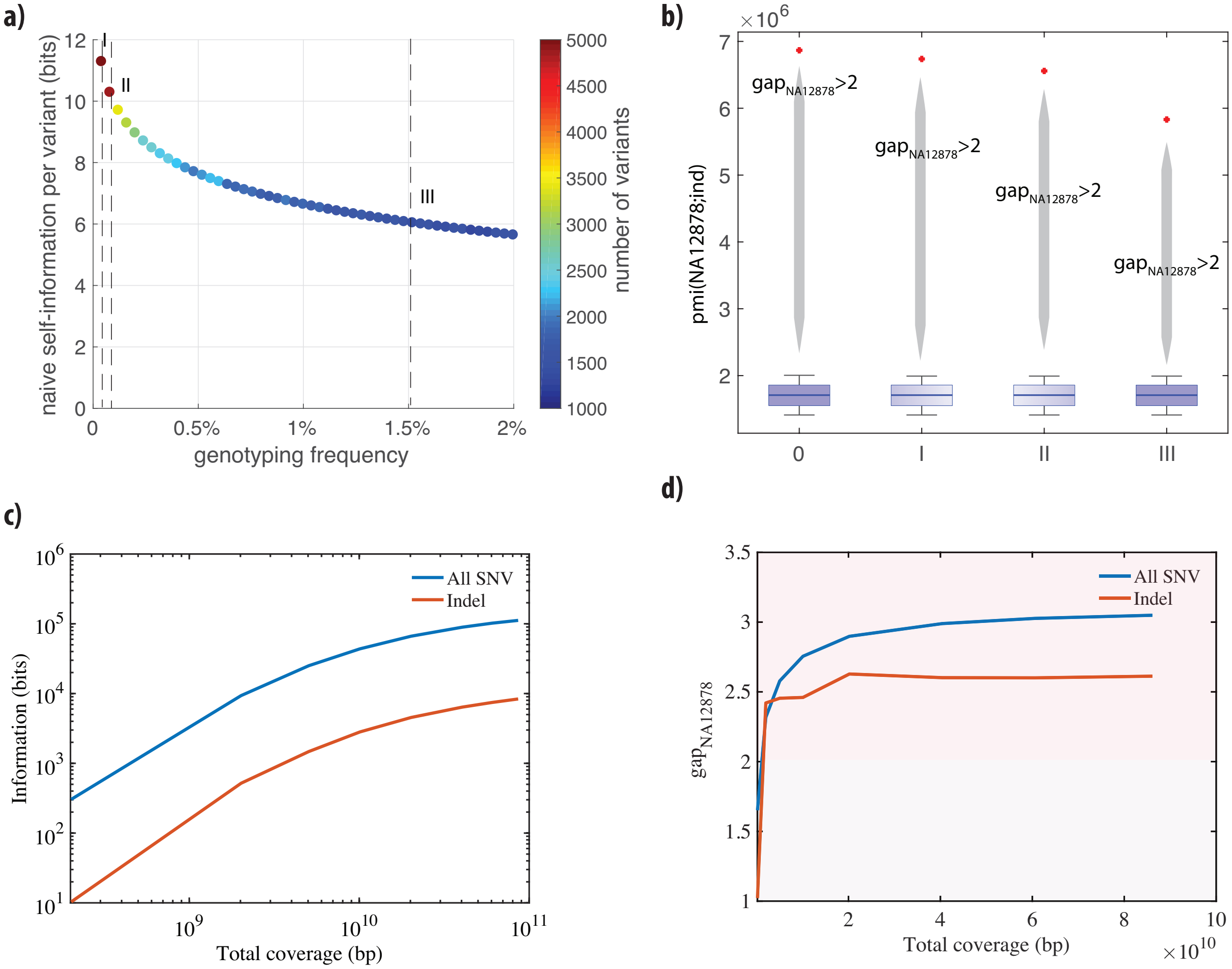


Figure 7: **Removal of rare variants and linking** (**a**) Information of the variant before and after addition of NA12878 to the population. We iteratively removed variants from the set as (I) only the variants that is only NA12878 specific, (II) the variants that have an information of 11 or higher 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.

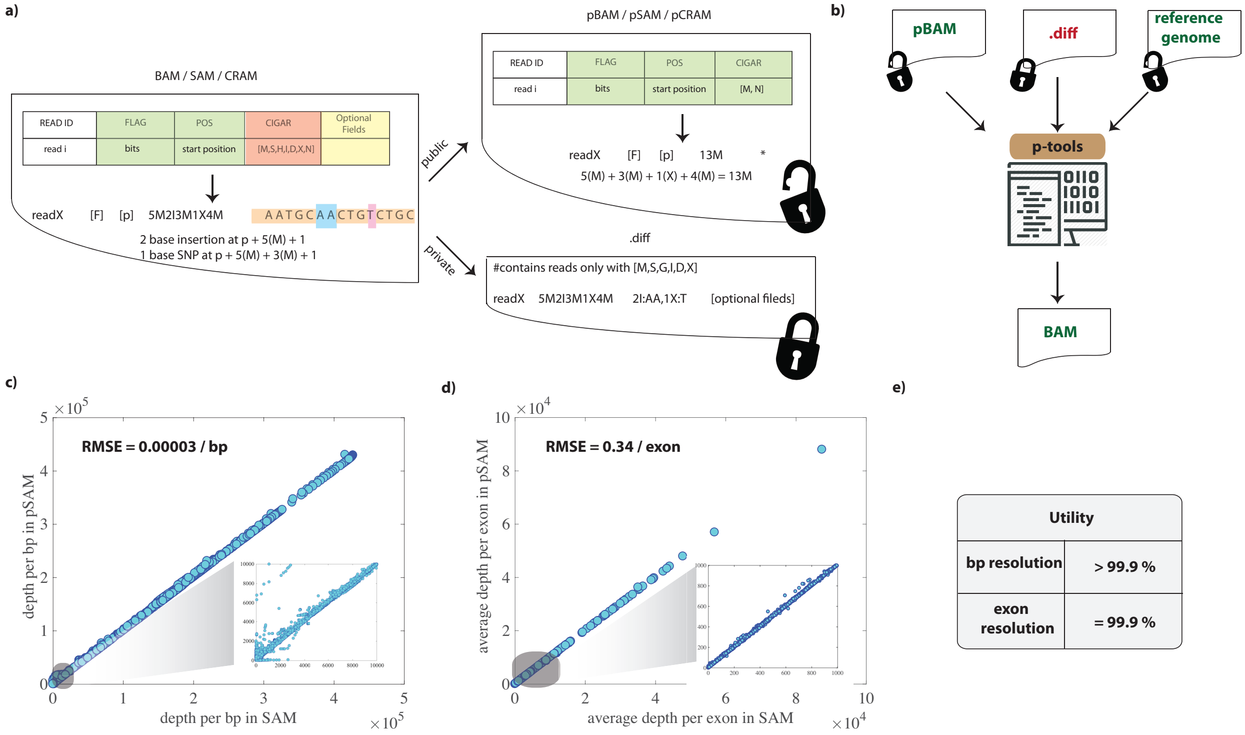


Figure 8: **Privacy-preserving file formats for mapped reads** (**a**) The generation of public pSAM and private .diff files. (**b**) Schematic of how to go between pBAM and BAM formats by utilizing the human reference (**c**) Comparison of nmber of reads for each basepair in the original SAM file and 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.