

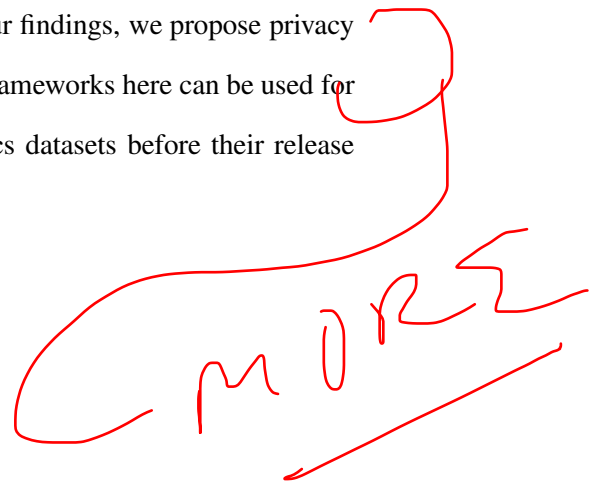
Information theory based measures for quantification of  
private information leakage and privacy-preserving file  
formats for functional genomic experiments

GG et al

December 9, 2017

## Abstract

Functional genomics experiments (FGEs) provide a wealth of insight on genomic activities (i.e. gene expression or transcription factor binding) related to developmental stages or diseases that are essential for personalized medicine. Although these activities are not necessarily tied to an individual's genotype, quantification of such activities is only possible using sequencing data that involves individual's genotypes. Mapped reads from FGEs are considered to be an important set point, in which there is high private information leakage and therefore access to them require special permission. This is not ideal due to the challenges of storing and moving big amount of private data as well as the difficulties in developing privacy-preserving calculations and result sharing. Moreover, in some special cases (low depth ChIP-Seq experiments) reads might sometimes be considered to be safe to share due to inadequate sequencing depth and incompleteness. Here we focus on the quantity of sensitive information in FGEs. 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 at varying coverages. We show that individuals are vulnerable to identification even when small amounts of sequencing data are available to adversaries. We also show that with summation of low depth FGEs and imputation through linkage disequilibrium, the leaked number of variants can reach the total number of variants in an individual'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 experiments. Based on our findings, we propose privacy enhancing file formats with minimal loss of utility. Presented frameworks here can be used for quantification of private information from functional genomics datasets before their release and conversion of sensitive data to private formats.



# 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/commercial genomic databases, researchers showed that multiple datasets can be linked together to infer sensitive information such as participant'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 personalized 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). In turn, these experiments provide stack of data that can be obtained directly from the machines and can further be computed. The first layer of the stack is sequencing data that involves sequencing reads with individual's genotypes and is stored in special file formats called fastq [?]. The second layer

stack

is the mapping of these sequences to the human genome and this data is stored in files formats SAM/BAM/CRAM [?, ?, ?]. Mapped read files also provide around 2 fold reduction compared to fastq files and since there are tools [?] to convert between BAM/SAM/CRAM format to fastq, sharing and storing of fastq files is irrelevant in this context. However, mapped read files are still large and their storage, transfer or computation take computational time and space. Mapped reads contain a significant amount of sensitive genetic information. They can be used to identify the private SNPs, small indels, and structural variants and therefore their accession requires oftentimes special permission. Nonetheless, sequencing data of FGEs 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 CHIP-Seq data [17]. The second layer in the data stack is the signal profiles, in which base pair resolution of sequencing depth is reported. The signal profiles are assumed to be free of sensitive information. However, a recent study has shown that they leak small indels and large deletions and can be used as personally identifying information [?]. The last layer is the quantification based on the signal tracks such as TF binding peaks or gene expression levels. Although gene quantification itself does not reveal sensitive information, when combined with genotype-phenotype correlation successful linking attacks can be made [14, 15]. A detailed description of the data stack and how FGEs work are depicted in Fig. 1.

N O T A C T I V E

oftentimes

?

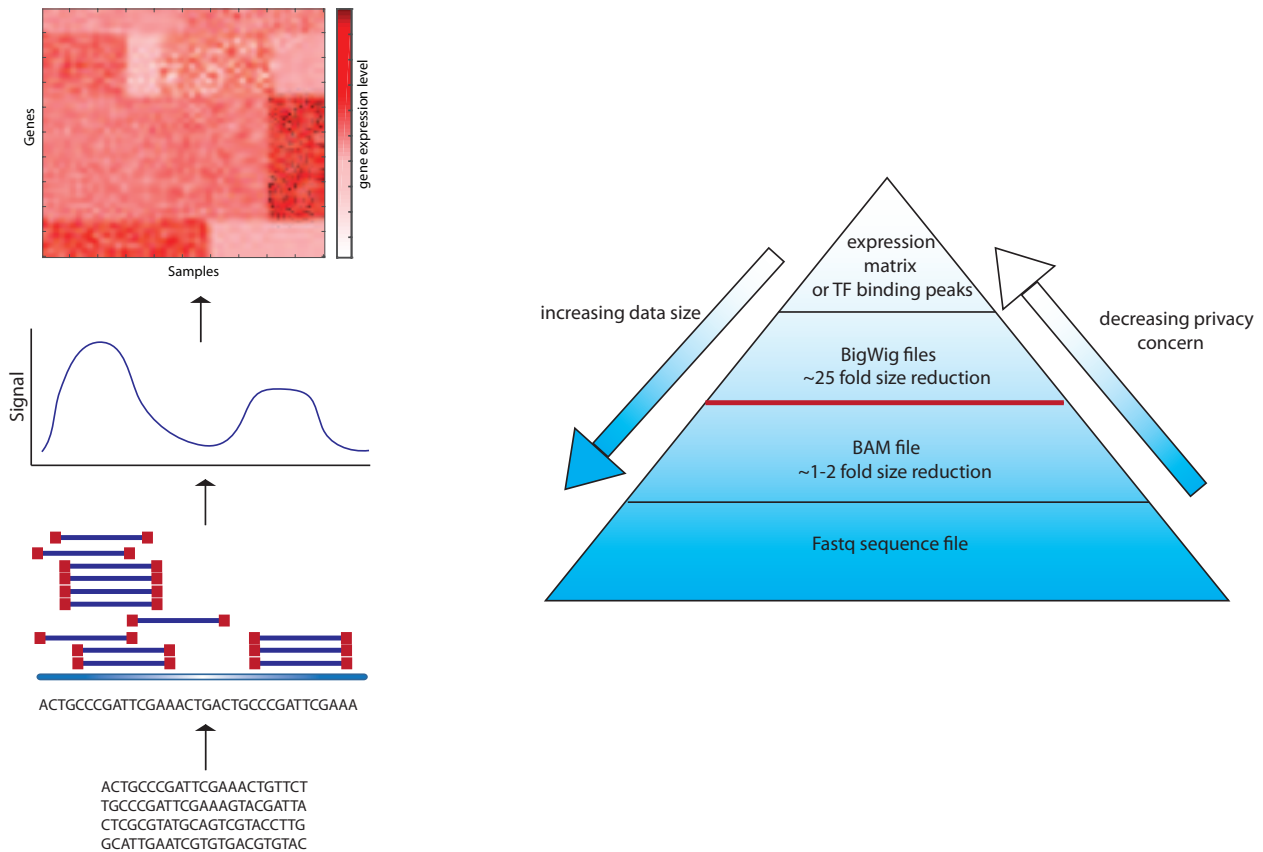


Figure 1: Schematic of data types from functional genomics experiments.

On the flip side of the coin is the utility of the sequencing data, which is the base layer in the data stack and challenges related to moving, storage of private data that are computationally costly and sharing of results based on private data. For example, big consortia such as ENCODE [18], TCGA [?] or GTEx [?] fund multiple research institutions and enable a collaborative working environment through dedicated phone calls and meetings. In turn, all the participants are expected to have special permissions to any external or internal genomics data, otherwise any computation based on the private data cannot be shared or presented. This creates a bottleneck and hinder the progress of important biomedical findings. Moreover, 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 organi-

BETTER

zations are increasingly supporting new means of data sharing and new requirements for making data publicly available while preserving the participant's privacy [?]. 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. 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 [18] 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 [20] 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 [21, 22]. We propose a new metric for quantifying 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 accessing only a portion of the sequencing data. We show that individuals are vulnerable to identifications even at small coverages of sequencing data. We further show that with summation of functional genomics experiments and imputation through linkage disequilibrium, the leaked number of variants can reach the total number of variants in an individual'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 compromising 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  $S = \{s_1, s_2, \dots, s_i, \dots, s_N\}$  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

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

In eq.1  $N$  is the total number of variants in an individual's genome,  $p(s_i) = n_i/n_T$ , where  $n_i$  is the number of individuals with variant  $s_i$  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 individuals 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  $\lim_{n_i \rightarrow \infty} h(S)$ ).

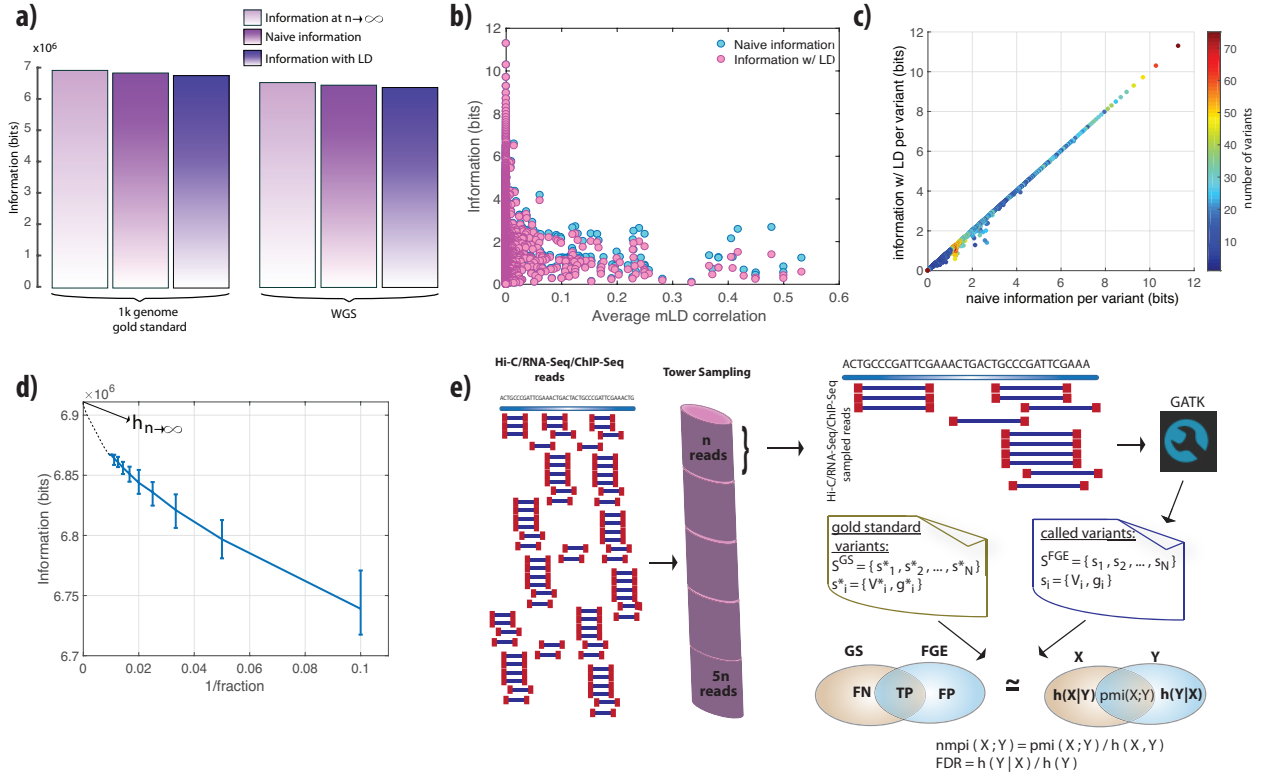
To be able to understand whether naive information is a good estimate, we first calculate the information with the consideration of LD scores taken from the European population of HapMap

project [23]. 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

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

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





**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 calculated 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 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 (Fig-

ure 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 [24]. 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 individuals one by one to the population (SI Figure 1a). 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 individual'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  $n \rightarrow \infty \geq \text{naive information} \geq \text{information with LD}$  (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 a functional genomics experiment

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 NA1278 (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 [21, 22]. 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  $S^{GS} = \{s_1^*, \dots, s_i^*, \dots, s_M^*\}$  is the set of variants from the gold standard and  $S^{FGE}(k) = \{s_1, \dots, s_i, \dots, s_M\}$  is the set of variants called from the  $k$  reads of a functional genomics experiment, then the set  $A = S^{GS} \cap S^{FGE}(k)$  contains the variants that are called and are in the gold standard set. If  $A = \{a_1, \dots, a_i, \dots, a_T\}$ , then

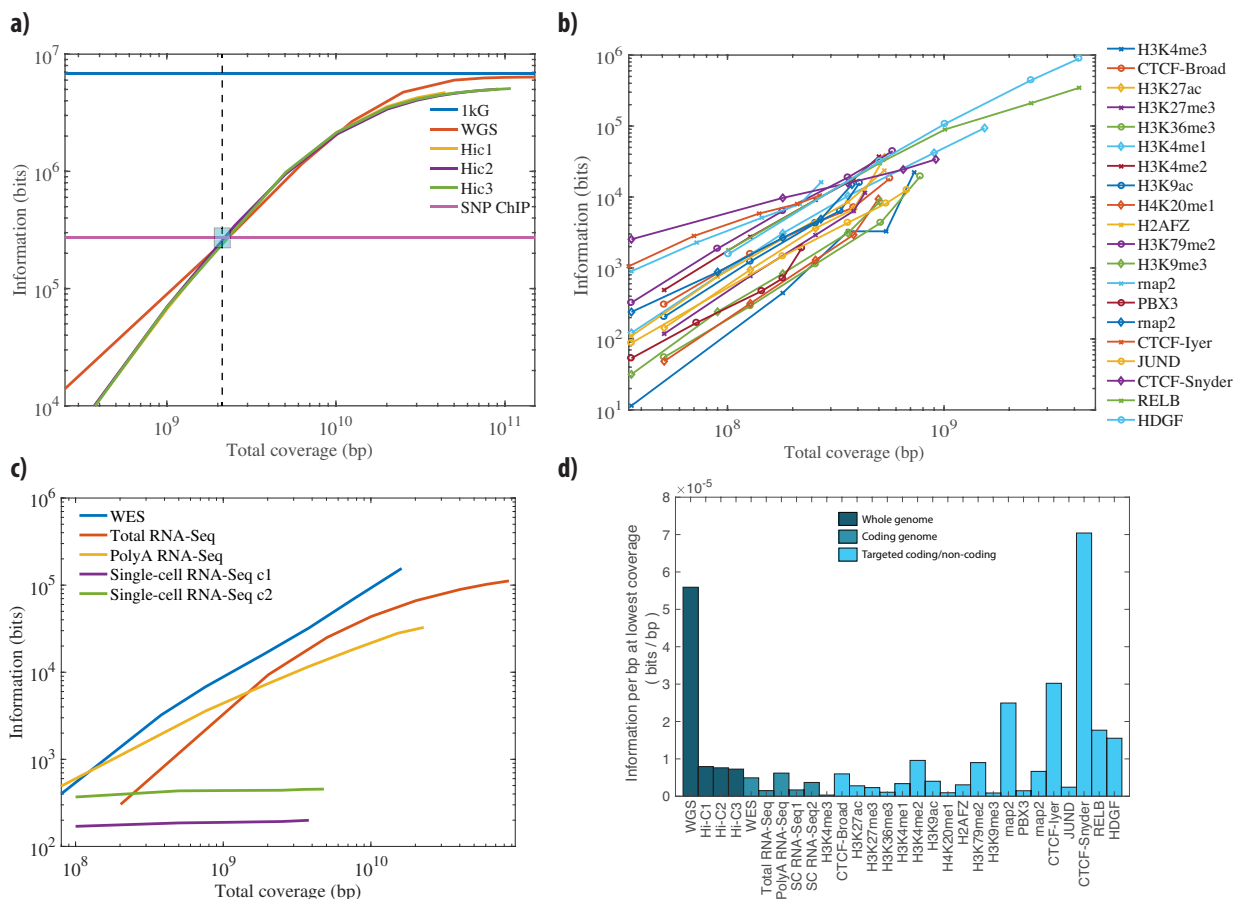
$$pmi(S^{GS}; S^{FGE}(k)) = - \sum_{i=1}^{i=T} \log_2(p(a_i)) \quad (3)$$

We then add  $k$  more reads to the sampled reads and repeat the calculation. This procedure 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).



**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 assays. (d) The pmi values per basepair plotted using the lowest total coverage for all the assays.**

As expected Hi-C data contains almost as much information as WGS and more information than SNP ChIP arrays. In the beginning of the sampling process, WGS data contains more information than Hi-C. 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 assess 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 ( $h(S^{FGE} | S^{GS})$ ) and the information obtained from all the called variants ( $h(S^{FGE})$ ), namely

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

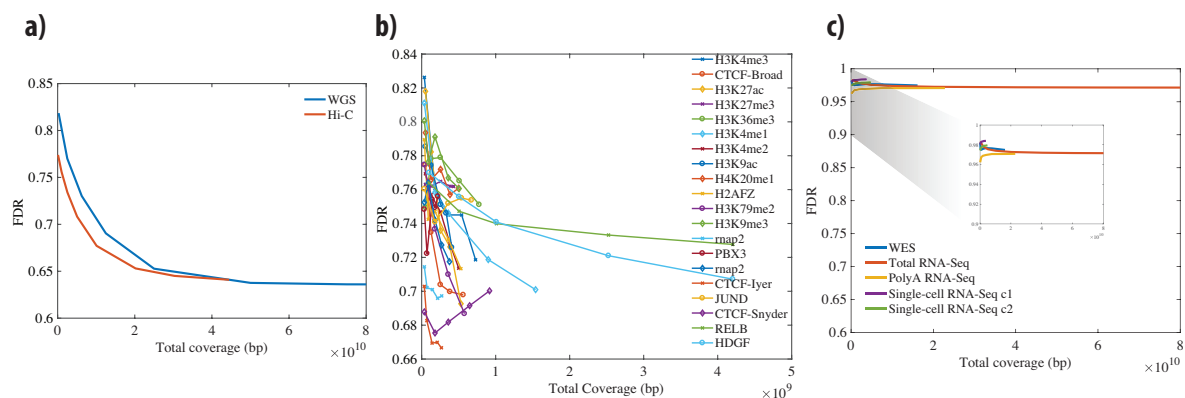


Figure 4: **False discovery rate of functional genomics experiments at different coverages (a)** FDR comparison for Hi-C and WGS data at different sampled coverages. **(b)** FDR comparison for different ChIP-Seq experiments at different coverages. **(c)** FDR comparison for WES and different RNA-Seq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

Figure 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 noisiest 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 hypothetical scenario, the attacker aims at querying an individual's HIV status from his/her phenotype data available through functional genomics experiments. The majority of linking attacks to this date focused on phenotypes, which the attacker finds the relationship between the phenotype and genotype data and use this relationship to link the HIV status

to the genotype data set. However, in this study, we go one step back from the phenotype data and directly inferred genotypes from the read files associated with the phenotype as, for example, majority of fastq and bam files of ChIP-Seq experiments are publicly available. 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 that have the highest pointwise mutual information. This reveals the sensitive information for the linked individual to the attacker. We also consider a scenario that the attacker has access increasing amount of reads in situations such as the attacker can query the sequencing data from a consortium certain amount at a time or has limited computing power.

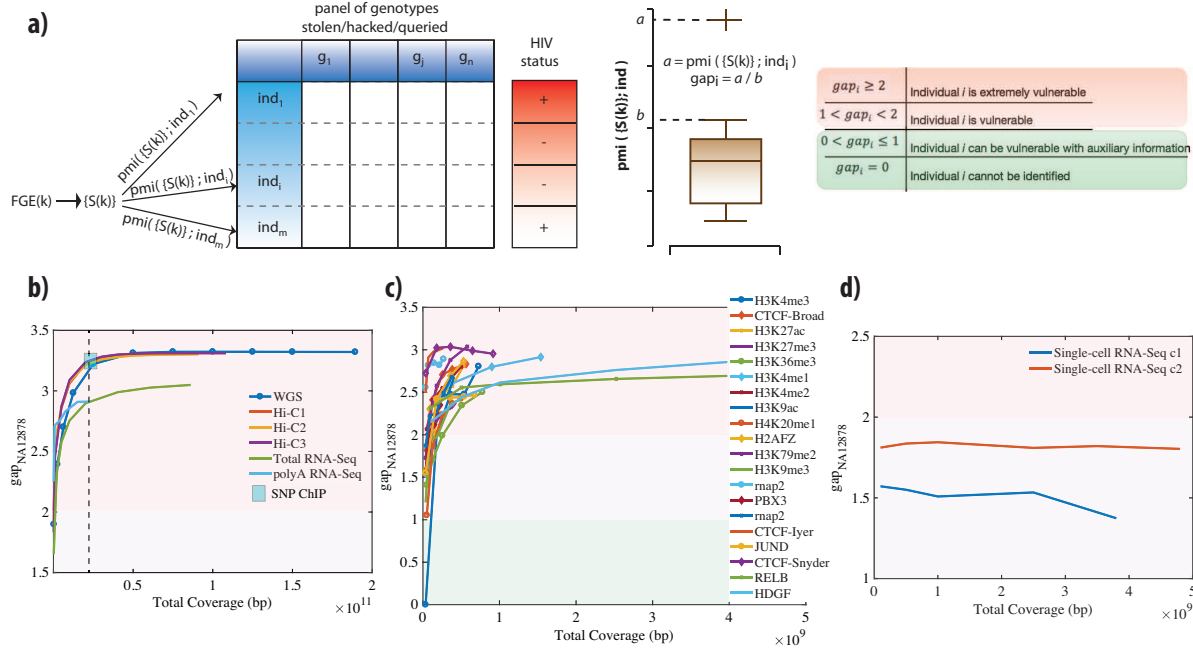


Figure 5: **Illustration of a linking attack and the accuracy of linking.** (a) The publicly available anonymized reads from functional genomics experiments contains a set of variants and HIV status for the sample that the functional genomics experiment was performed at increasing coverages. The panel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

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 auxiliary data such as gender or ethnicity, because he/she is in the top 5 matching 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).

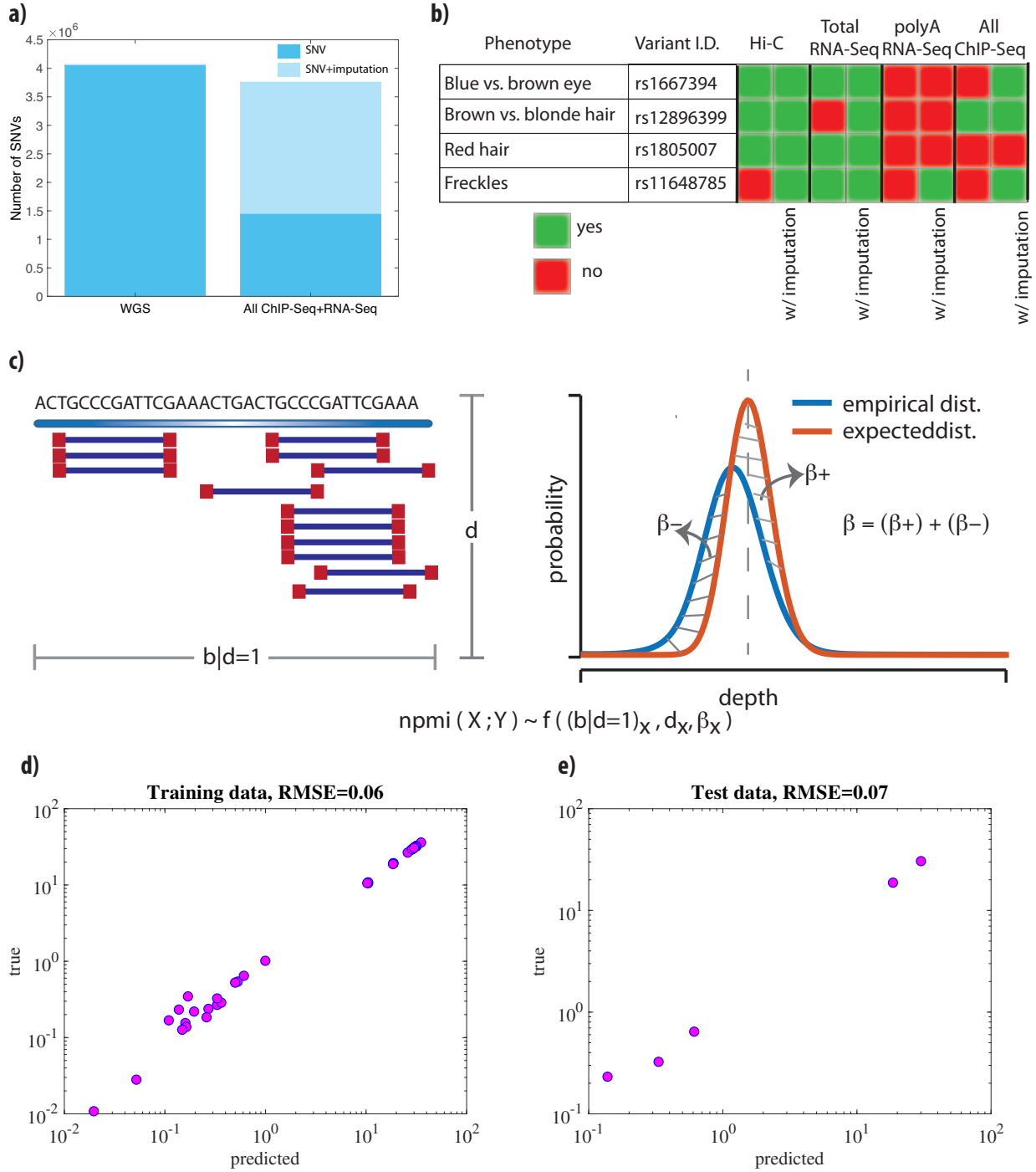
We find that NA12878 is extremely vulnerable even at the lowest sampled coverages for Hi-C and RNA-Seq data (Figure 5b). More interestingly between around 1.1 and 10 billion basepairs, the Hi-C data exhibits higher linking accuracy than WGS data, consistent with the previous observation of  $pmi$  shown 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 has small coverages with high accuracy (Figure 5d).



## **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 [25, 26, 27] using 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 Figure 2), total number of captured variants increases significantly (Figure 6a). By using shallow sequencing 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 associated with traits further increases with imputation as expected.



**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

## 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 individual to that of reference human genome. To be able to do successful genotyping, one needs substantial depth of sequencing reads for each base pair. According to the Waterman-Lander statistics for DNA sequencing, when random chunks of DNA is sequenced repeatedly, 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 [?]. 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 hypothesized that the genotyping from the sequencing based functional genomics data depends on the depth per base pair ( $d$ ), the total fraction of the genome that is represented at least by one read ( $b \mid d = 1$ ) and a parameter  $\beta$  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  $\beta$  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 expected 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 ( $h(S^{GS} \mid h(S^{FGE}))$ ) and the information from the incorrectly identified genotypes, i.e FDR ( $h(S^{FGE} \mid h(S^{GS}))$ )

as;

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

With assumption of  $npmi(S^{FGE}; S^{GS}) = f(d_{FGE}, b_{FGE} | d = 1, \beta_{FGE})$ , we used generalized linear models 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 achieved ((Fig. 6d), see SI for more details). 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 individual's genotypes can be predicted 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, we calculate the naive information of the gold standard variants in the 1000 genomes with and without the presence of the NA12878 (Fig. 6a). The variants that deviate the most from the diagonal are the ones that contribute the most to the overall naive information. We group the variants into three categories based on their contribution to the overall naive information. We remove the variants in these categories and instantiate a linking attack based on the naive information of the remaining variants. Comparison between the linking accuracy when we consider all of the variants and when we removed them categorically shows that individuals are extremely vulnerable to linking attacks even when the half of the variants that have the highest contribution to the overall naive information are removed (Fig. 6b). Our conclusion from this cal-

ulation is that not only the rare variants but also the unique combination of common variants are identifiers of genetic make-up of individuals.

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 information from indels are much smaller compared to single nucleotide mutations, a high linking accuracy can be achieved by using only indels even at small coverages (Fig. 6d).

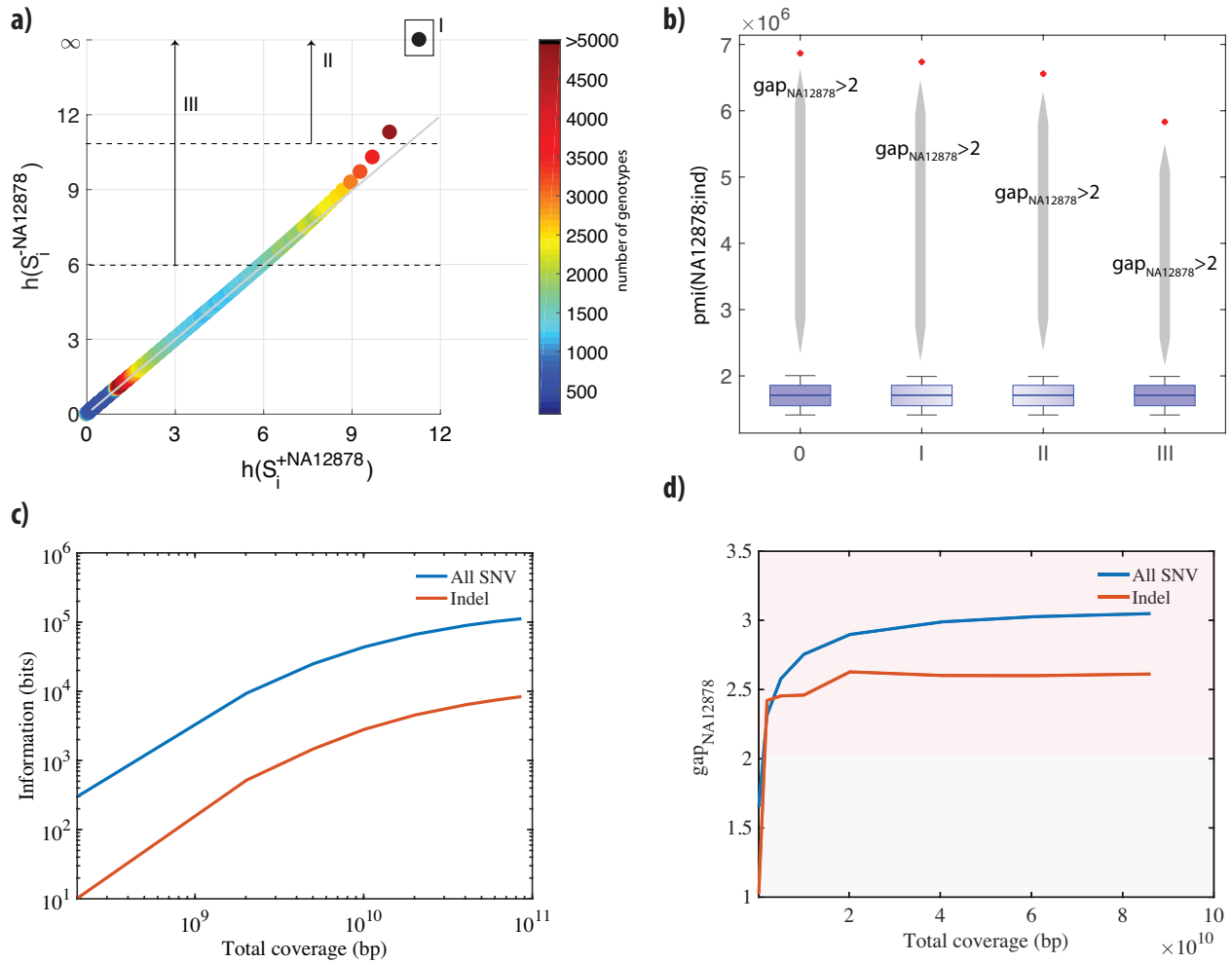


Figure 7: **Removal of rare variants and linking** (a) Information of the variant before and after addition of NA12878 to the population. We iteratively removed variants from the set as (I) only the variants that is only NA12878 specific, (II) the variants that have an information of 11 or higher bits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

## 2.9 Privacy-enhancing file formats for functional genomics experiments

- Indels can be inferred from the current MRF - Refer to Figure 6c and 6d for the possibility of linking with using only indels of the noisiest data set we have - total rna-seq - Describe the new

p-BAM (new figure)

### 3 Discussion

#### 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, Szelling 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] Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. *Science*, 2012;44(5):603-608.
- [15] Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature Methods*, 2016;13(3):251-256.
- [16] Harmanci A, Gerstein M. Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions. *Nature Communications*, 2017
- [17] Beskow LM. Lessons from HeLa Cells: The Ethics and Policy of Biospecimens. *Annu Rev Genomics Hum Genet.*, 2016;17:395-417
- [18] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012;489(7414):57-74.



- [19] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 2010;467(7319):1061-1073.
- [20] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 2010;467(7319):1061-1073.
- [21] 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.
- [22] 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.
- [23] International HapMap Consortium. The International HapMap Project. *Nature*, 2003;426(6968):789-796.
- [24] Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.*, 1998;80:197.
- [25] 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.
- [26] Howie BN, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3: Genes, Genomics, Genetics*, 2011;1(6):457-470.

[27] 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.