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

GG et al

November 11, 2017

TOO STRONG

2 SIDED

Abstract

The success of ENCODE project opened the doors to a deeper understanding of functional genome through genome-wide experimental assays. Although identifying individuals using DNA variants from whole genome or exome sequencing data is a major concern of privacy and security, no study on genomic privacy focused on the quantity of sensitive information in functional genomic experiments such as ChIP-Seq, RNA-Seq and Hi-C. Here, we derive novel information theory based measures for quantification of private information in DNA and RNA sequences. We apply these measures to quantify the amount of leaked information in 24 functional genomic assays from ENCODE data portal at varying coverages and provide a comparison between these asssays and WGS, WES and SNP-ChIP data. Based on this quantification, we instantiate linking attacks, in which adversaries have access increasing coverage of the sequencing data from these experiments. We show that individuals are extremely vulnerable to identification even at low coverages. We further show that with summation of functional genomics experiments and imputation through linkeage disequilibrium, the leaked number of variants can reach the total number of variants in an indivudals genome. We then provide a theoretical framework where the amount of leaked information can be predicted from depth and breadth of the coverage as well as the bias of the genome-wide experiment. Presented frameworks here can be used for quantification of private information from large functional genomics datasets before their release. Futhermore, we propose privacy enhancing file formats for the functional genomics experiments based on our findings.

1 Introduction

1. motivation:

   Privacy is important

2. previous work historically:

   (a) Genome Privacy traditionally focuses on DNA variants

      i. Detecting whether an individual with known genotypes in a complex DNA mixture

         Homer et. al, 2008: Distance between genotype and dataset

         Im et. al, 2012: Regression coefficients of GWAS summary statistics can reveal persons participation

      ii. Identification attacks by cross-referencing independent datasets

         Sweeney at al, 2013: Cross-reference PGP profile with public voter list data

         Gymrek et al, 2013: Cross-reference Y-STRs with recreational genetic genealogy database

   (b) Functional genomics era increases the number of quasi-identifiers

      i. define quasi-idenrifier: Quasi-identifiers are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier. Quasi-identifiers can thus, when combined, become personally identifying information. This process is called re-identification. As an example, Latanya Sweeney has shown that even though neither gender, birth dates nor postal codes uniquely identify an individual, the combination of all three is sufficient to identify 87% of individuals in the United States.

      ii. Big consortia like ENCODE, TCGA, GTEx provide a wealth of functional genomics data, which particularly belong to individuals

A. Schadt et. al, 2012: SNP genotypes can be predicted from RNA-Seq expression data using known eQTLs

B. Harmanci and Gerstein, 2016: eQTLs and extreme expression levels can be used to do linking attacks

C. Harmanci and Gerstein, 2017: ChIP-Seq/RNA-Seq/Hi-C signal tracks can also leak genotype information

3. specific motivation for this work:

   (a) Functional genomics era attacks focus on phenotype-genotype relationship

   Functional genomic experiments are large-scale, high-throughput assays to quantify transcription (RNA-Seq), epigenetic regulation (ChIP-Seq) or 3D organization of genome (Hi-C) in a genome-wide fashion under different conditions (e.g. samples from patients and healthy individuals). In turn, these experiments produce sequencing data (often in fastq and bam formats) that involves individual's genotypes in various coverages. The sequencing data of the experiments that require high genome coverage such as Hi-C or RNA-Seq often require special permission for access. However, ...

   i. All the functional genomics data comes with a great deal of sequencing data, most of them (chip-seq) are publicly available

   ii. Is that information enough to identify individuals?

   iii. Mention HeLas genome is locked but all the functional genomics (except Hi-C) reads are available as a part of the motivation

4. introduce MRF format:

   (a) Gerstein lab paper on rseq tools

(b) current MRF needs more work - indels can be inferred

(c) fixing the leakage in the reads is not enough, signal still has leakage - refer to Arif's paper

2 Results

2.1 Information Theory measures 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 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 nmber 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. We used NA12878 and 1000 genomes phase I genotype panel for the calculation of self-information. Note that $h(S)$ is an estimate of the real information in a situation where the ancestral information of the individual 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). Eq.1 also cannot be equal to the information when we consider all the individuals in the world ($n_t \rightarrow \infty$)). Therefore from hereof we will refer Eq.1 as the naive information.

To be able to understand whether naive information is a good estimate, we first calculated the information with the consideration of LD scores taken from the European population of HapMap project. 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 exists 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 $mLD(s_i, s_j) = \max\limits_{i \neq j, j \in (1,..,N)} LD(s_i, s_j)$, where $mLD(s_i, s_j) \neq mLD(s_j, s_i)$.

2.

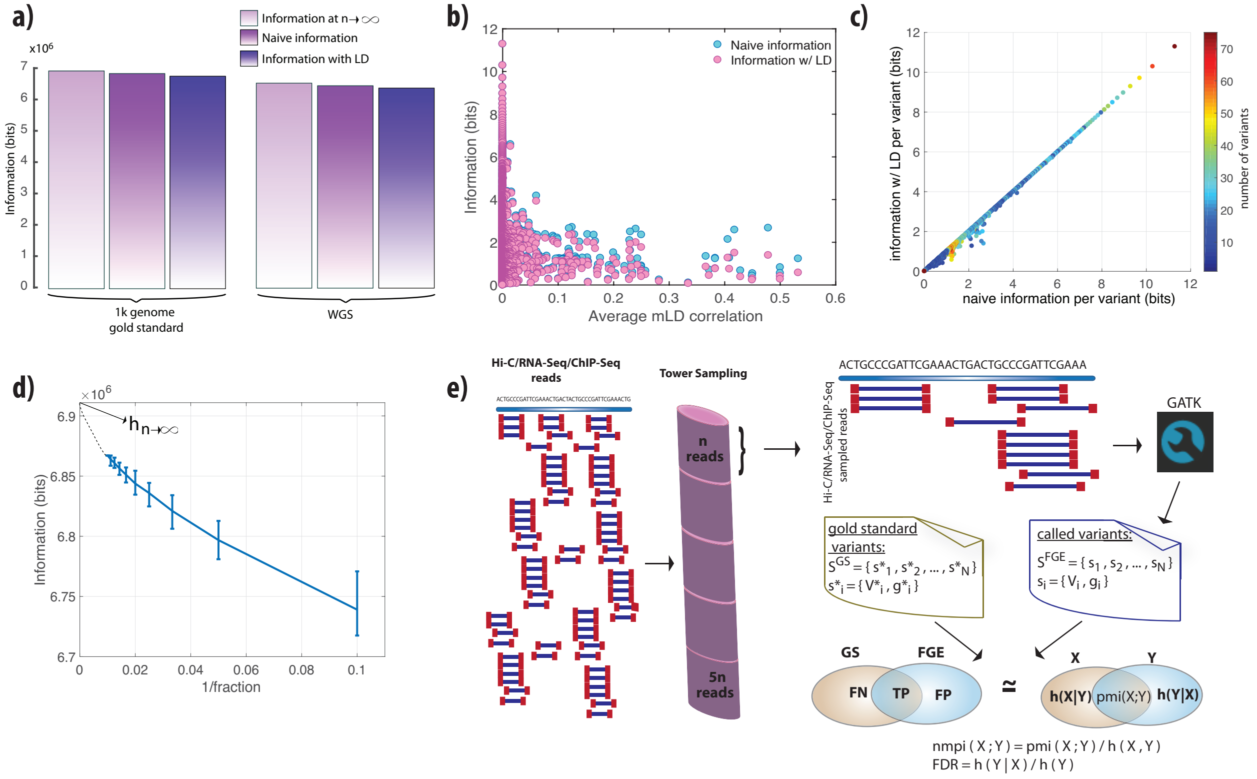


Figure 1: **Comparison of naive information measure with information with LD consideration and sample size correction.** (**a**) Difference between the naive information, information with LD consideration and extrapolated information when population size is infinite. (**b**) The maximum LD score for 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 1a shows a negligible difference between the naive information and information with LD consideration for NA12878 genome. To understand the lack of difference better, we calculated the self-information of each variant in an LD block with and without LD consideration. We showed that highly informative variants do not exhibit any difference due to the low LD correlations (Figure 1b). We further showed that the number of variants that have difference between information with and without LD consideration is small compared to highly informative variants that having low LD correlations on average.

We then estimated the information when the population size is infinite. We sampled fractions of 1000 genomes phase I panel (total of 2504 individuals) and calculated the information using the sampled distribution. We repeated this calculation for 100 times and calculated the mean information for each sampled fraction. We found a power relationship with two terms ($y = ax^b + c$) between the inverse sampled fraction and the information ($R = 0.99$). The y-intercept (c) of the curve is the extrapolation of information when the population size goes to infinity ($1/\infty = 0$, Figure 1c). We again found a negligible difference between the naive information and the information when the population size is infinite (Figure 1a). The information is also calculated by starting from a single individual and adding individuls 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 showed that as the size of the population increases, the differences between in the information decreases (SI Figure 1a-b)

In summary, calculations above showed that the naive information can be an accurate approximate to the private information content of an individal's genome when the ancestral information 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 1a). That is, an adversary with no prior ancestral information of the sample and limited number of individuals in a known genotype panel can accurately approximate the private information in an individual's genome.

2.2 Information Theory measures 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. The genotypes in 1000 genomes panel for NA1278 is used as the gold standard. We used 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 $S_k^{FGE} = \{s_1, .., s_i, ..., s_M\}$ is the set of variants called from the k reads of a functional genomics experiment, then the set $A = S^{GS} \bigcap S_k^{FGE}$ contains the variants that are called and are in the gold standard set. If $A = \{a_1, .., a_i, .., a_T\}$, then

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

We then added k more reads to the sampled reads and repeated the calculation. This procudere were repeated till we depleted all the reads of a functional genomics experiment. Overall process is depicted in Figure 1e.

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 involved whole genome approaches such as Hi-C, transcriptome-wide assays such as RNA-Seq and targeted assays such 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 2).

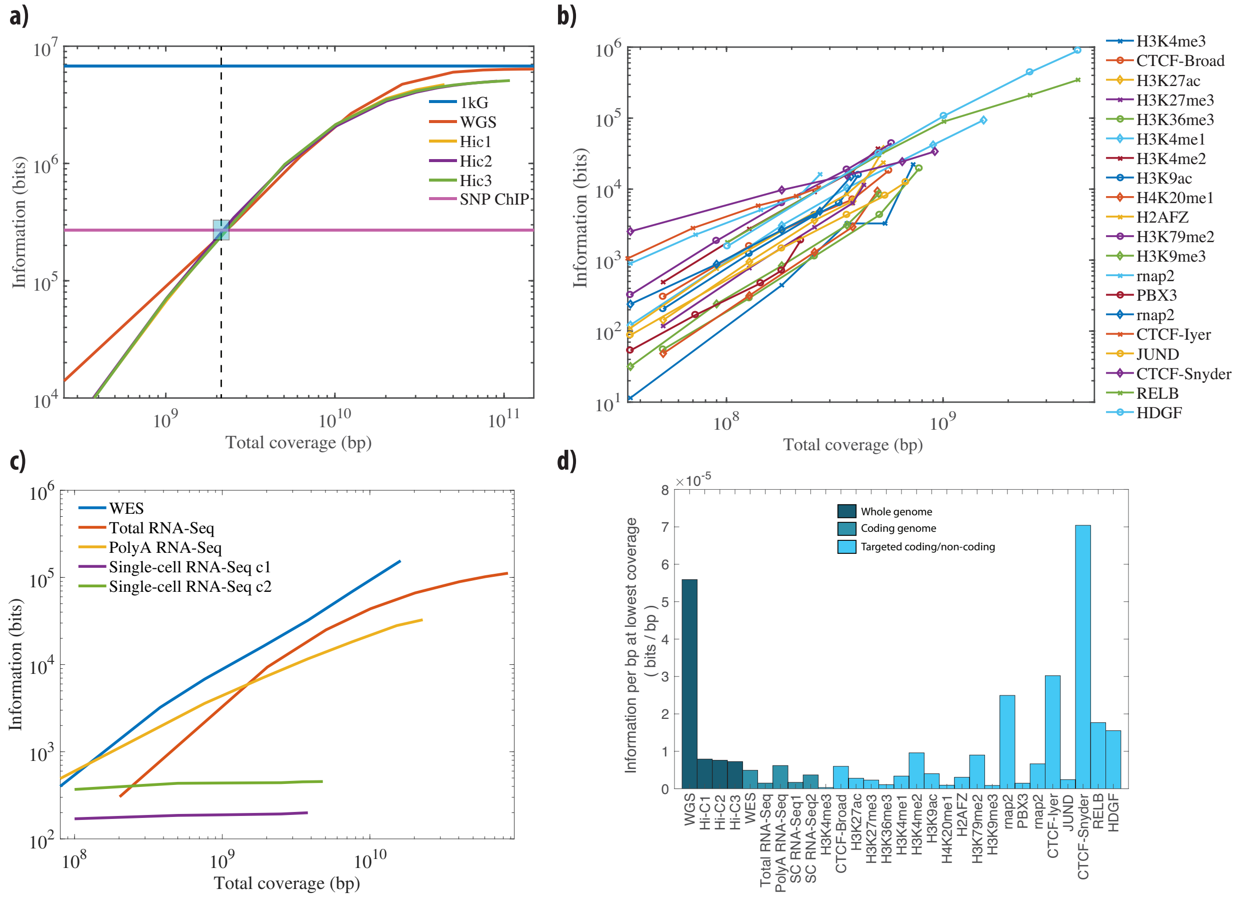


Figure 2: **The pointwise mutual information calculated for 24 different functional genomics assays and WGS, WES and SNP ChIP data using NA12878 1000 genomes variants as gold standard.** (**a**) The pmi values for WGS and three different primary Hi-C experiments plotted at different coverages. The information 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.

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 sampling is between around 1.1 and 10 billion bps, the information content of Hi-C surpass the WGS data (Figure 2a). 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 2b). 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. Comparison between WES and different RNA-Seq experiments showed that none of the RNA-Seq experiments contains 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 2c). The more unexpected observation was 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 calculated the pointwise mutual information per bp at the lowest coverages depicted in Figure 2a–c (pmi(FGE;GS)/total coverage). We found that CTCF data from Snyder lab contains even more information per basepair than WGS data at the lowest coverage we sampled (Figure 2d).

NO ?

PRAC APP ?.

2.4 Linking attack scenario

Linking attacks aim at determination of sensitive information about an indiviudal in a stolen genotype data set (Figure 3a). For example, in an hyphotetical scenario, the attacker aims at querying an individual's HIV status from his/her phenotype data available through functional genomics experiments. In majority of linking attacks, the attacker finds the relationship between the phenotype and genotype data and use this relationhip 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 RNA-Seq and 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 indivudal to the attacker. We also considered 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.

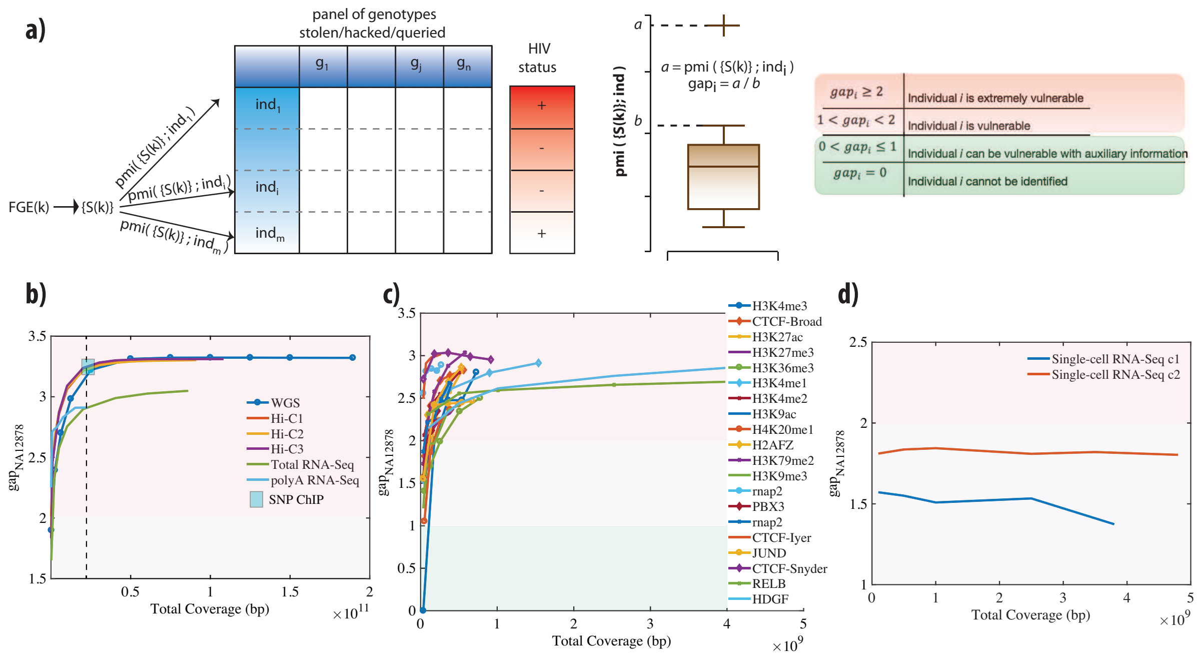


Figure 3: **Illustration of a linking attack and the accuracy of linking.** (**a**) The publicly available ananoymized reads from functional genomics experiments contains a set of variants and HIV status for the sample that the functional genomics experiment was performed at increasing coverages. The 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.

Based on the pmi values of each experiment at different coverages, we defined a metric for linking accuracy called gap_i. To calculate this metric, we first ranked all the $pmi(S_k^{FGE}; individual_i)$ where $S^F GE_k$ is the set of called genotypes from the functional genomics experiment at total coverage k and $individual_i$ 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;

$$gap_i = \begin{cases} \frac{pmi(S_k^{FGE};individual_i)}{pmi(S_k^{FGE};individual_j)}, & \text{if } rank(pmi(S_k^{FGE};individual_i) \leq 5 \text{ and } rank(pmi(S_k^{FGE};individual_j) = 2 \\ 0, & \text{otherwise} \end{cases}$$

We then defined 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 3a).

We found that NA12878 is extremely vulnerable even at the lowest sampled coverages for Hi-C and RNA-Seq data (Figure 3b). 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 2a. 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 3b), which shows extreme vulnerability of individuals when 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 with high accuracy (Figure 3d).

2.5 Genotyping accuracy

BEFORE LINK

After realizing that the genotyping can be done using low depth, biased functional genomics experiments, we asseses the accuracy of genotyping by calculating the false discovery rate at different coverages. This also measures how much noise that each assay capture. The false discovery rate is defined as the ratio between the information obtained from the incorrectly called variants (h(FGE—GS)) and the information obtained from all the called variants (h(S)), namely

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

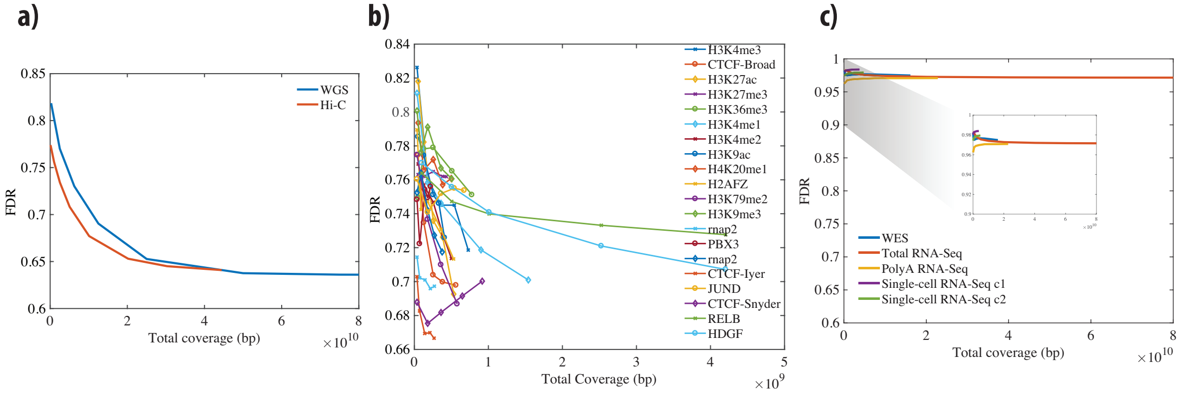


Figure 4: **False discovery rate of functional genomics experiments at different coverages** (**a**) FDR comparison for Hi-C and WGS data at different sampled coverages. (**b**) FDR comparison for different ChIP-Seq experiments at different coverages. (**c**) FDR comparison for WES and different RNA-Seq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

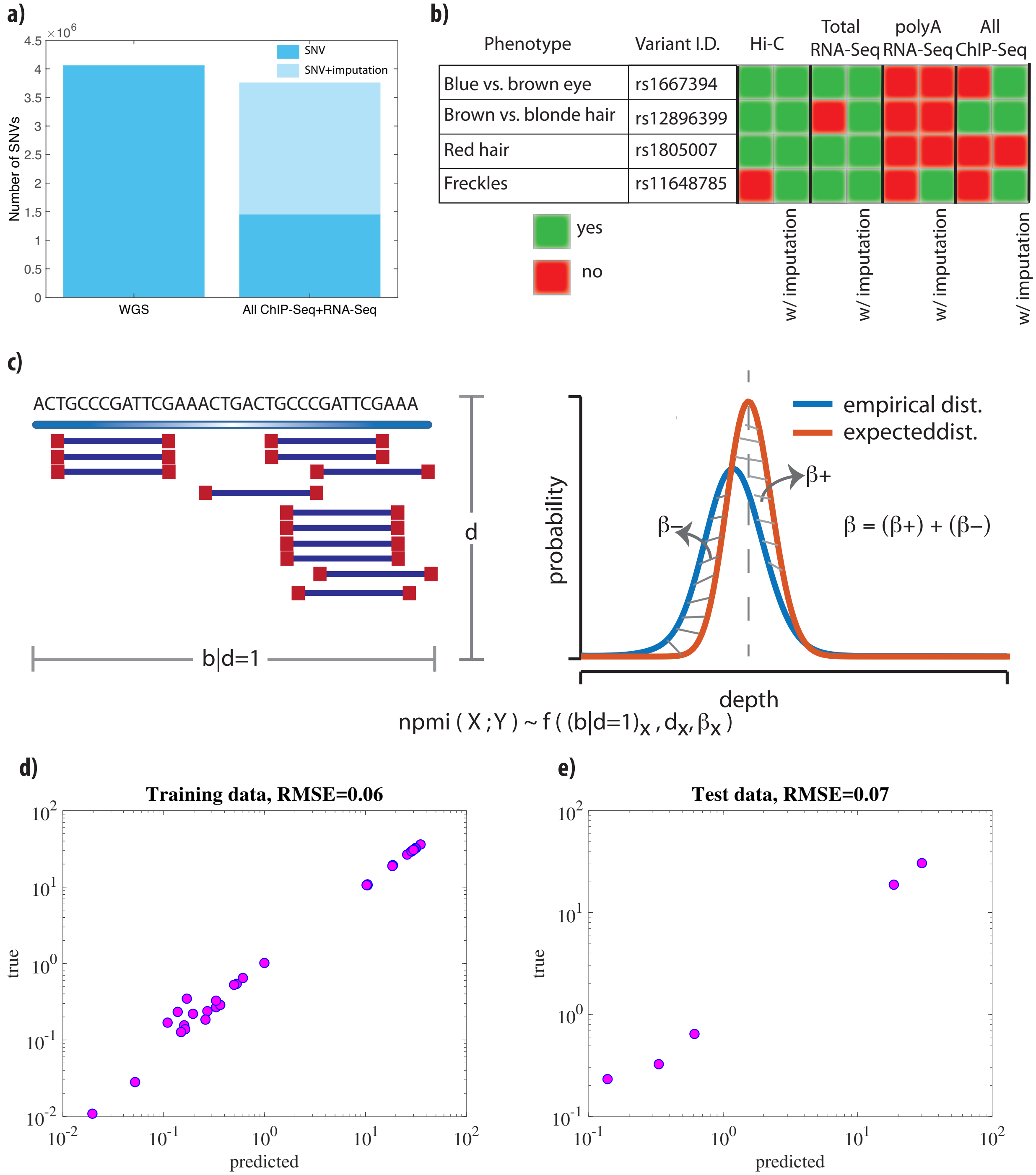
Figure 4a shows that the false discovery rate for Hi-C data is lower compared to WGS data at lower coverages. We attribute it to the deeper sequencing of the genomics regions in close spatial proximity. Hence, sampling more reads from those regions at low coverages is more likely compared to uniform sampling of reads from WGS. ChIP-Seq data has comparable false discovery rate to WGS and Hi-C data, ChIP-Seq targeting CTCF from Snyder Lab having the lowest FDR (Figure 4b). We further found 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).

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

To answer the question whether an attacker correctly assembled an individual's variants by only using the publicly available reads from ChIP-Seq and RNA-Seq experiments, we imputed variants by using IMPUTE2 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 increase significantly (Figure 5a). By using shallow squencing data of ChIP-Seq and RNA-Seq, we were able to call and impute variants almost as many as the variants called from WGS data.

We then asked the question if we can infer potentially sensitive phenotypes from these variants. Figure 5b 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.

develop



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

🟩 yes

🟥 no

Figure 5: **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 prediction of amount of leaked data without variant calling

Explain Figure 5 c, d and e.

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

- Inspired from differential privacy, we did Figure 6a - Iteratively removed rare variants and calculated information and linking accuaracy (Figure 6b) - We can still link the individual

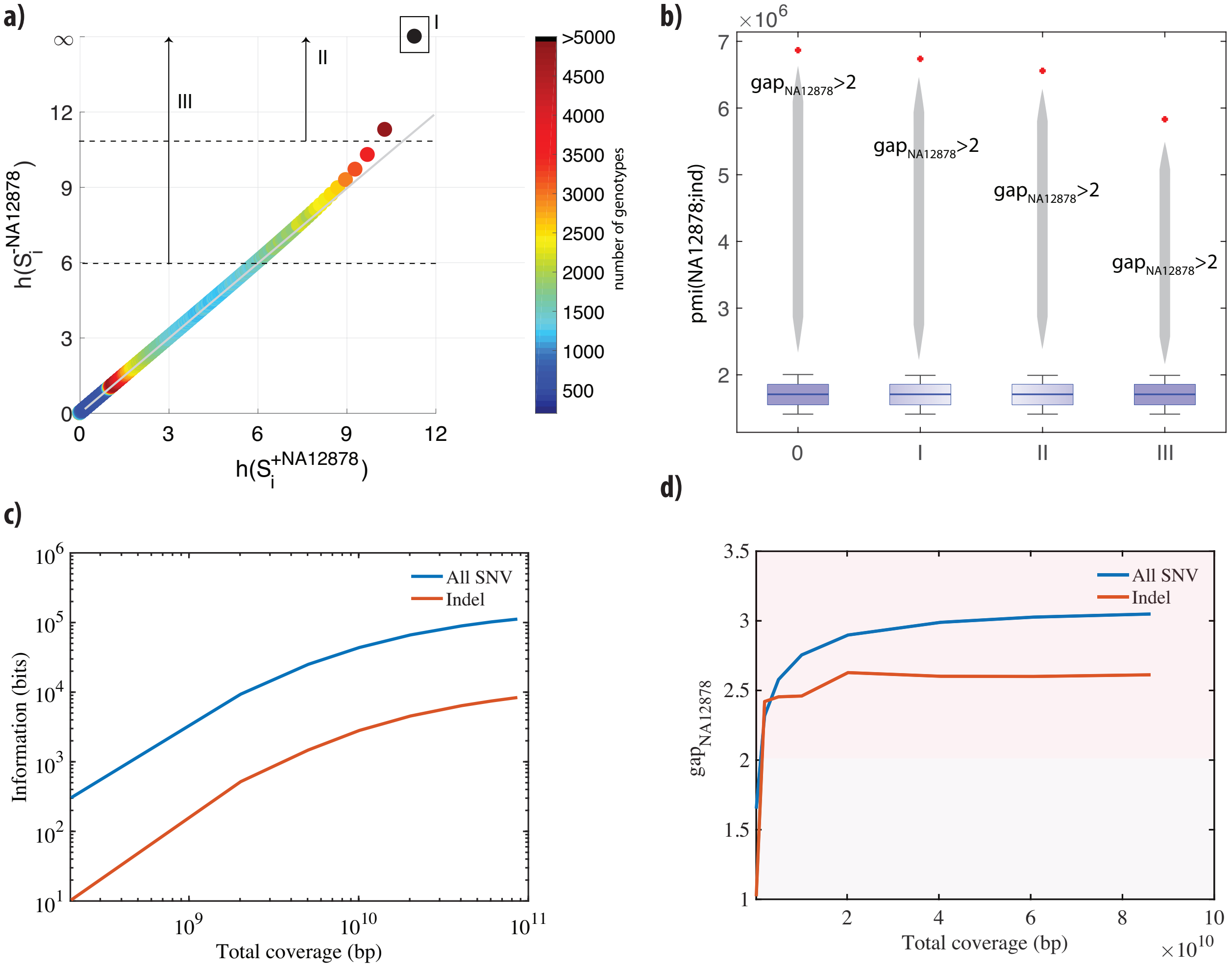
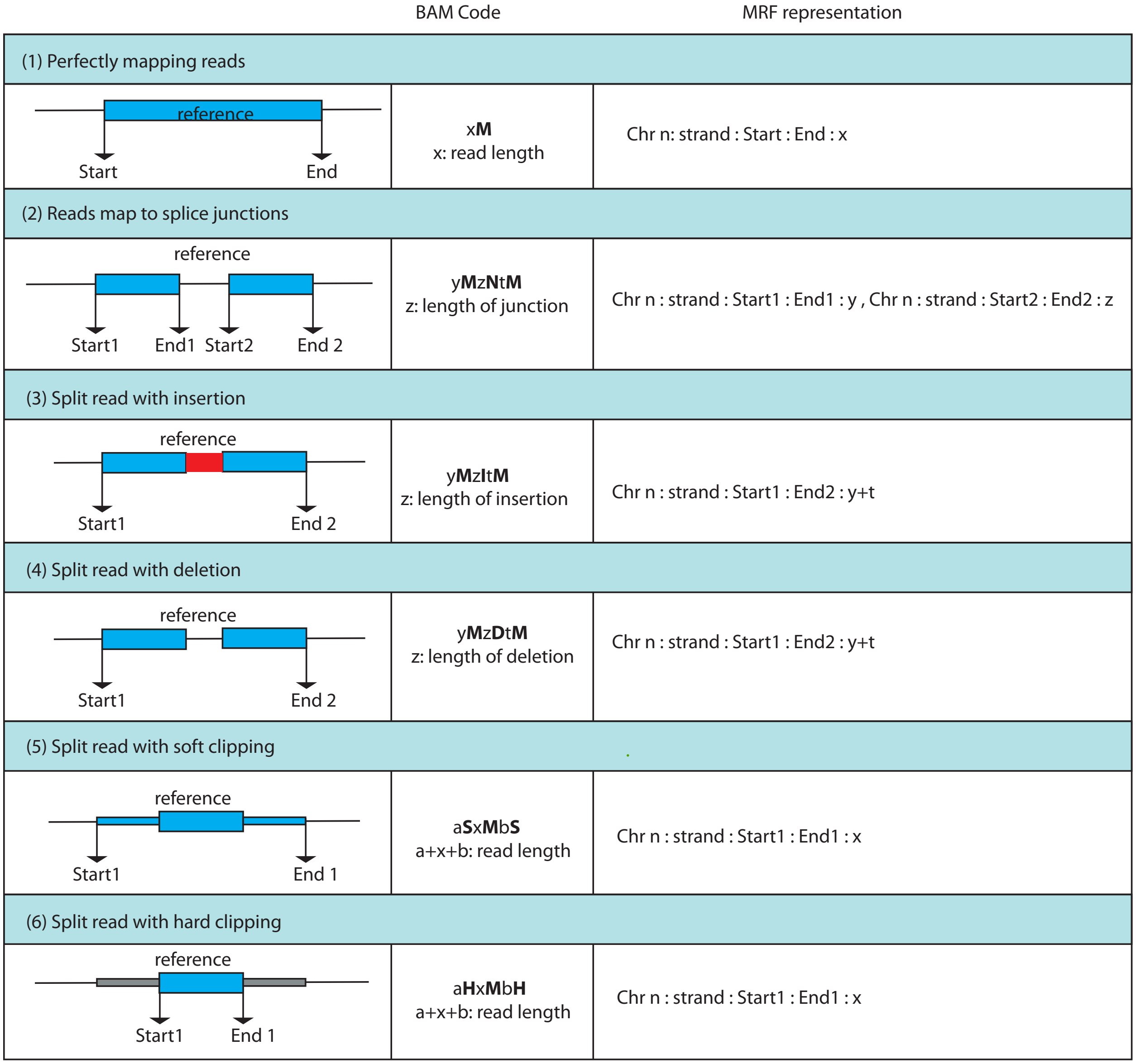


Figure 6: **to discuss** (**a**) Information of the variant before and after addition of NA12878 to the population. We iteratively removed variants from the set as (I) only the variants that is only NA12878 specific, (II) the variants that have an information of 11 or higher bits after removal of NA12878 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.

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

MRF (Figure 7)



	BAM Code	MRF representation
(1) Perfectly mapping reads		
reference Start End	**xM** x: read length	Chr n: strand : Start : End : x
(2) Reads map to splice junctions		
reference Start1 End1 Start2 End 2	**yMzNtM** z: length of junction	Chr n : strand : Start1 : End1 : y , Chr n : strand : Start2 : End2 : z
(3) Split read with insertion		
reference Start1 End 2	**yMzItM** z: length of insertion	Chr n : strand : Start1 : End2 : y+t
(4) Split read with deletion		
reference Start1 End 2	**yMzDtM** z: length of deletion	Chr n : strand : Start1 : End2 : y+t
(5) Split read with soft clipping		
reference Start1 End 1	**aSxMbS** a+x+b: read length	Chr n : strand : Start1 : End1 : x
(6) Split read with hard clipping		
reference Start1 End 1	**aHxMbH** a+x+b: read length	Chr n : strand : Start1 : End1 : x

Figure 7: **MRF: to discuss**

3 Discussion