

Privacy-preserving file formats and information theory



based measures for sensitive information leakage of raw



functional genomics data

## GG et al

February 23, 2018

Abstract



288 words



Functional genomics experiments provide information about genomic activities such as gene



expression under different conditions, which are not tied to an individual’s genotype. However



sharing of such data raises privacy concerns due to quantifications being tagged with individ-



ual’s variants. Although, progressive summarization such as signal profiles is believed to re-



duce private information leakage, they also reduced the utility and amount of the open-access



data. On the other hand, while granting special permission to access the raw data can remedy



privacy concern, open access to such large amounts of data accelerates biomedical research



and increases reproducibility. Current genome privacy studies lack a systematic quantification



of the amount of leakage in raw functional genomics data and privacy-preserving solutions in



an open-access data-sharing mode. Here we introduce information theory-based measures to



systematically quantify the amount of sensitive information in raw reads of different functional



genomics assays at different sequencing depths. We showed that high depth experiments such



as Hi-C provide accurate genotyping that lead to large privacy leaks. Counter intuitively, noisy



and partial genotypes from low depth experiments such as ChIP-Seq and single-cell RNA-Seq



can be used as strong quasi-identifiers for re-identification purposes through linking attacks. In



order to solve the dilemma between open access vs. privacy leak, we propose a file formatting



system that enables sharing of large amount of data while protecting individuals sensitive in-



formation and preserving the utility of the data. Such file format manipulation can be used in



different levels to achive different levels of privacy, hence provides a differential-privacy like



framework.

# Introduction

With the decreasing cost of DNA sequencing technologies, the number and the size of the avail- able genomic data have exponentially increased and become available to a wider group of audi- ences such as hospitals, research institutions and individuals [[1](#_bookmark2)]. In turn, privacy of individuals has become an important aspect of biomedical data science [[2](#_bookmark3), [3](#_bookmark4)] as availability of genetic informa- tion gives rise to privacy concerns such that genetic predisposition to diseases may bias insurance companies [[4](#_bookmark5)] or create unlawful discrimination by employers.

Early genomic privacy studies focused on identification of individuals in a mixture by using phenotype-genotype association [[5](#_bookmark6), [6](#_bookmark7)]. These revealed that private information of an individual such as participation to a drug-abuse study [[5](#_bookmark6), [6](#_bookmark7)] can be revealed. With the increase of large-scale genomic projects such as Personal Genome Project (PGP) [[7](#_bookmark8)] or recreational/direct-to-consumer genomic databases, researchers showed that multiple datasets can be linked together to infer sensi- tive information such as pariticipant’s surnames [[8](#_bookmark9)] or addresses [[9](#_bookmark10)]. 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](#_bookmark11)].

Functional genomics experiments provide a wealth of information on genomic activities re- lated to developmental stages or diseases that are essential for personilized medicine. These are large-scale, high-throughput assays to quantify transcription (RNA-Seq) [[11](#_bookmark12)], epigenetic regula- tion (ChIP-Seq) [[12](#_bookmark13)] or 3D organization of genome (Hi-C) [[13](#_bookmark14)] in a genome-wide fashion under different conditions (e.g. samples from patients and healthy individuals). Inferring biological in- formation 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.



Although activities of functional genome are not



necessarily tied to an individual’s genotype, quantification of such activities is derived from the



biosamples that are belong to individuals, hence they are tagged with individual’s variants. Public



sharing of such raw data raises privacy concerns. To be able to share high utility data while pre-



serving individuals’ sensitive information, it is essential to determine a ‘’set point”, after which



trade-off between the utility of the data and the privacy risk is balanced. A hurdle in determin-



ing the ‘set point” is the lack of systematic quantification of private information leakage from

[1](#_bookmark33))



the functional genomics data. Figure



summarizes the processing steps of functional genomics



experiments with how summarization decreases the risk of privacy while greatly decreasing the



amount of sharing and the utility of the data. 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](#_bookmark15)].



of the experiment. They are then mapped to human reference genome and these mapped reads are

[[15](#_bookmark16)].



These files are large in size ranging from 5 GB up to 60 GB depending on the purpose



stored in compressed binary file types called BAM



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](#_bookmark33)).

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, for the genome of HeLa cell line, the raw reads from Hi-C experiments require special access, while reads from ChIP- Seq and RNA-Seq experiments are publicly available [[19](#_bookmark18)]. 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 leakage free. 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.



File formats such as CRAM is developed to remedy

[?].



the ever increasing amount of data



Although CRAM greatly reduces the size of the data in

BAM files, there is still privacy leaks due to the containement of the sequence of the reads [?]. MRF

[?].



was introduced as an alternative to CRAM, where keeping the sequence of reads is optional



This does not only reduces the size of the data, but also makes it hard to genotype the individuals



from the information in these files. However, private information leak is not entirely removed from



MRF files, as one can still infer deletions from the information in these files. Moreover, current



quantification pipelines used for gene expression analysis as well as the peak calling softwares



were not designed to take MRF file format as inputs.

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 requires use agreements that have expiration dates 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 es- tablishment 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 bureau- cracy. Another example to the challenges associated with private data is that big consortia such as ENCODE [[20](#_bookmark20)], TCGA [[21](#_bookmark21)] or GTEx [[22](#_bookmark22)] fund multiple research institutions and enable a collabo- rative working environment through dedicated phone calls and meetings. In turn, participants have

to go through required access procedures with their institutions. Otherwise communication based on private data is prohibeted according to data use agreements. Moreoever, when multiple institu- tions 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 agen- cies 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](#_bookmark23)]. 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. 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](#_bookmark20)] 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 pri- vate data that requires special access and storage.



different levels of privacy vs. utility balance with an adjustable parameter and provides a differen-



tial-privacy like framework.



Our file format manupilation system achieves

In this study, we use NA12878 as a case example and her 1000 genomes [[24](#_bookmark24)] genotypes as gold standard genotypes. We sample reads from the sequencing data of functional genomics experi- ments at increasing coverages and detect SNVs and indels using Genome Analysis Toolkit (GATK) best practices recommendations [[25](#_bookmark25), [26](#_bookmark26)]. 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.

# Results

## 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 correspond- ing genotype. Let *S* = {*s*1, *s*2, .., *si*, ..*sN*} be the set of variants, then each variant can be represented

as *si* = {*vi*, *gi*}, where *vi* consists of the location and alternative allele information and *gi* 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

*i*=*N*

*h*(*S*) = − ∑ *log*2(*p*(*si*)). (1)

*i*=1

In eq. [1](#_bookmark0) *N* is the total number of variants in an individual’s genome, *p*(*si*) = *ni*/*nT* is the genotype frequency, in which *ni* is the number of individuals with variant *si* = {*vi*, *gi*} and *nT* 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. E[q.1](#_bookmark0) 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](#_bookmark0) also an estimate to the information when

we consider all the individuals in the world (i.e lim*nt* →∞ *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](#_bookmark27)]. 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

*i*=*N*

*hLD*(*S*) = − ∑ (1 − *mLD*(*si*, *s j*))*h*(*si*) (2)

*i*=1

*LD*(*si*, *s j*) is the maximum LD correlation of variant *si* such that *mLD*(*si*, *s j*) = max *LD*(*si*, *s j*),

where *mLD*(*si*, *s j*) /= *mLD*(*s j*, *si*).

*i*/= *j*, *j*∈(1,..,*N*)

Figure [2](#_bookmark34)a 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 [2](#_bookmark34)b). 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](#_bookmark28)]. 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 frac- tion. The relationship between the inverse of the sample fraction and the information fits best to

a power function with two terms (*y* = *axb* + *c*, *R* = 0.99). The *y*-intercept (*c*) of the curve is the

extrapolation of information when the population size goes to infinity (1/∞ = 0, Figure [2](#_bookmark34)c). We

again found a negligible difference between the naive information and the information when the population size is infinite (Figure [2](#_bookmark34)a). 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* → ∞ ≥ naive information ≥ information with LD (Figure [2](#_bookmark34)a).

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.

## Information Theory to quantify private information leakage in func- tional 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 ge- nomic 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](#_bookmark25), [26](#_bookmark26)]. 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 *SG* = {*s*∗, .., *s*∗, ..., *s*∗ } is the set

1 *i M*

of variants from the gold standard and *SF* (*k*) = {*s*1, .., *si*, ..., *sM*} is the set of variants called from the *k* reads of a functional genomics experiment, then the set *A* = *S* n *S* (*k*) contains the variants

*G F*

that are called and are in the gold standard set. If *A* = {*a*1, .., *ai*, .., *aT* }, then

*i*=*T*

*pmi*(*SG*; *SF* (*k*)) = − ∑ *log*2(*p*(*ai*)) (3)

*i*=1

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 [2](#_bookmark34)e.

## 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 transcrip- tion factor binding. In addition, the pmi is also calculated for WGS, WES, and SNP-ChIP for

comparison (Figure [3](#_bookmark35)).

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 [3](#_bookmark35)a). 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 [3](#_bookmark35)b). 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 [3](#_bookmark35)c). 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

[3](#_bookmark35)d



the mean pointwise mutual information per bp depicted in Figure



( ∑ *pmi*(*SF* (*k*);*SG*)/*k* ). We found

(number of samples)



that Hi-C experiments and ChIP-Seq experiments targeting the transcription factor HDGF provide

(Figure [3](#_bookmark35)d).



more genotyping information per basepair compared to WGS data

## Genotyping accuracy

In light of the above findings, in which genotyping can be done using low depth, biased func- tional genomics experiments, we asses the accuracy of genotyping by calculating the false dis- covery 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 incor- rectly called variants (*h*(*SF* | *SG*)) and the information obtained from all the called variants (*h*(*SF* )),

namely

*F DR*(*SF* (*k*)) = *h*(*SF* (*k*) | *SG*)/*h*(*SF* (*k*)) (4)

Figure [4](#_bookmark36)a 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 [4](#_bookmark36)b). 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 [4](#_bookmark36)c).

## Linking attack scenario

Linking attacks aim at re-identification of an individual by cross-referencing datasets (Fig- ure [5](#_bookmark37)a). 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 adver- sary 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 ge- nomics 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 *gapi*. To calculate this metric, we first rank all the *pmi*(*SF* (*k*); *Si*) where *SF* (*k*) is the set of called genotypes from the functional genomics experiment at total coverage *k* and *Si* is the set of genotypes of individual *i* in the panel of genotypes. *gapi* for each individual *i* at total

coverage *k* is calculated as;

 *pmi*(*SF* (*k*);*Si*) , if *rank*(*pmi*(*SF*(*k*); *Si*) ≤ 5 and *rank*(*pmi*(*SF*(*k*); *S j*) = 2

*gapi*

=  *pmi*(*SF* (*k*);*S j*)

****0, otherwise

We then define that if *gapi* 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 < *gapi* ≤ 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 < *gapi* ≤ 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 *gapi* > 2, then the individual is extremely vulnerable with more than 2 fold difference between him/her and the second best match (Figure [5](#_bookmark37)a).

We find that NA12878 is extremely vulnerable even at the lowest sampled coverages for Hi-C and RNA-Seq data (Figure [5](#_bookmark37)b). 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 ob-

servation of pmi shown in Figure [3](#_bookmark35)a. 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 [5](#_bookmark37)b), 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 (Fig- ure [5](#_bookmark37)d). 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 individ- uals are her parents (SI Figure 2). This shows that using the metric *gap*, an adversary can also identify individuals related to the target individual.

## Individual’s genome can be accurately approximated from publicly avail- able 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 IM- PUTE2 [[29](#_bookmark29), [30](#_bookmark30), [31](#_bookmark31)] 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 [6](#_bookmark38)a). 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 [6](#_bookmark38)b 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.

## Toy model for estimation of amount of leaked data without variant call- ing

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](#_bookmark32)]. 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 (*d*) , the total fraction of the genome that is

represented at least by one read, also called the breadth (*b* = ∑*N*

*i*=1

[*di* ≥ 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. [6](#_bookmark38)c). The bias parameter β is composed of two terms: (1) the negative bias β − and (2) the positive bias β +.

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).

*d* = *i*=1 *di N*

∑

*N*

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*(*SF* ; *SG*)), the information missed that is in the gold standard (*h*(*SG* | *SF* ))

and the information from the incorrectly identified genotypes, i.e FDR (*h*(*SF* | *SG*)) as;

*F G pmi*(*SF* ; *SG*)

*pmi*(*SF* ; *SG*)



*npmi*(*S* ; *S*

) = *h*(*SF* , *SG*) = *h*(*SG* | *SF* ) + *pmi*(*SF* ; *SG*) + *h*(*SF* | *SG*) (5)

With the assumption of *npmi*(*SF* ; *SG*) = *f* (*dF* , *bF* , β*F* ), 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. [6](#_bookmark38)d). 5 separate data points were used as test set and an RMSE of 0.07 was acheieved (Fig. [6](#_bookmark38)d),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.

## 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. [6](#_bookmark38)a). from the NA12878 variant set,



After the removal of unique variants

we calculated the *gapNA*12878 and surprisingly found that linking accuracy is affected minimally compared to using the all of NA12878 variants (Fig. [6](#_bookmark38)b). We then created another set (‘double variants”, Fig. [6](#_bookmark38)a), 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 (*gapNA*12878 > 2,Fig. [6](#_bookmark38)b). 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. [6](#_bookmark38)a). This also did not affect the overall linking (*gapNA*12878 > 2,Fig. [6](#_bookmark38)b).

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 re- moving 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 at- tacks, 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 (*gapNA*12878 > 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. [6](#_bookmark38)c 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. [6](#_bookmark38)d). This linking attack is done using the most noisy data set we have (total RNA-Seq) to make linking more difficult.

## Privacy-preserving file formats for alignments from functional genomics experiments and their relation to differential privacy



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 the utility measures, we borrowed con-



cepts from differential privacy and introduced a user identified privacy-utility balance that can be



adjusted according to the patient consents and policies.



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 *A* that



retrives results *A*(*D*) from database *D* is considered ε-differentially private if the results satisfies



the condition

*prob*(*A*(*D*) = *C*) *prob*(*A*(*D*±*i* = *C*)

= expε , (6)



where *D*±*i* indicates the addition or subtraction of *ith* individual to the database.



Although, differentially private solutions by definition are applied to the databases of individ-



uals, a similar logic can be applied to raw sequence alignment data from functional genomics



experiments. Our hyphotesis is that since gene quantification or TF binding peak calling do not



necessarily depen on the individual’s genotype, masking or removal *q* amount of variants from the



alignment files minimally (on the level of ε) affect the functional genome quantification. That is,



we assume a BAM file of a functional genomics experiment is the dataset *D*, where each record is



a string with the information of number of matched, mismatched, inserted or deleted basepairs in



a read (also known as a cigar). If we create a new file format (called pBAM) that is refered as *D*±*q*



where we change *q* number of records such that cigars no longer carry any labels corresponding



mismatches, insertions or deletions. If we calculate depth signal for each genomic location *i* in *D*



and *D*−*q*, then we obtain

*prob*( *f* (*D*) = *C*) *prob*( *f* (*D*−*q* = *C*)

= *e*ε*i* , (7)



which can be rewritten as

ε*i* = ln *prob*( *f* (*D*) == *C*) − ln *prob*( *f* (*D*−*q* == *C*) (8)

[8](#_bookmark1)



In equation



*f* (*D*) is the function that computes the depth of a genomic location *i* from the



BAM/pBAM file *D*/*D*−*q*. Since depth of a genomic location *i* is not informative without the depth



of the rest of the genome, we used z-scores of the depth for *prob*( *f* (*D*) == *C*), as





ε = ln( *di* − *d* ) − ln( *di*−*q* − *d*−*q* ) (9)

*i* σ (*D*)

σ (*D*−*q*)



We then calculated the ε*q* for the new file format as ε = max ε*q*. As the ε is a privacy mea-

*i*∈(1,..,*N*) *i*



sure, we used *U q* = 1 as the utility measure of the new file format. As the purpose of functional

ε*q*



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, which could be



signal depth profile in basepair and exon resolutions or gene quantifications.



The privatized file format pBAM is constructed as following. The reads from the BAM 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 map-



ping quality, start coordinates, fragment lengths and flags related to mappability of the reads while

(Fig. [8](#_bookmark40)). The



adjusting the cigar and alignment scores such that leakage of variants are masked

details of how new file format deals with reads are reported in detail in SI Methods with a detailed



figure (SI Fig. 3). We also created pBAM files for the BAM files that are obtained by aligning



thereads to transcriptome which is necessary for gene quantification (in SI Methods).



Such treatment of reads introduce noise to the signal profiles especially with deletions since the

(Fig. [8](#_bookmark40)a-b, see SI Methods and



start coordinates and total length of the fragments are unchanged

SI Figure 4).



However, our utility analysis showed even when *q* is the all SNVs in the BAM files,



*x*% accordance with the original depth of the nucleotides, *y*% accordance with the original depth



averaged over exons and 100% recovery for gene quantifications. [[GG: will be updated as:As can

(Fig. [8](#_bookmark40)c-



be seen from the scatter plots, noise is mostly introduced to basepairs with low depth

d).]]



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 our collec-



tion of scripts called ptools can convert pBAM + .diff + reference genome into the original BAM

file (Fig [8](#_bookmark40)b).

# 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 in- formation 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,

* 1. 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 also discuss the concept of ‘’set point” in determining the data production

(Fig. [1](#_bookmark33)). 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.



incomplete genotyping from partial sequencing experiments can serve as strong quasi-identifiers,



steps, where sensitive information leakage and utility of the data are balanced



That is, noisy and

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.



which is not straightforward to predict at first.



The increasing incentive to share data for the advancement of biomedical research and the cor-



related increrasing privacy concerns have led researchers to look for more complex solutions to



overcome the bottleneck between data-sharing and privacy preserving means. Solutions such as

[?, ?, ?].



differential privacy has been proposed



It has shown that retriving summary information



from private statistical databases without revealing some amount of individuals’ 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



Although such concept might



be useful for sharing summary statistics of functional genomics data from multiple individuals, it



is conceptually hard to envision the applications to the raw mapped read sharing from functional



genomics experiments taken from a single individual. By borrowinf language from the concept of



differential privacy, we created privacy-preserving binary allignment files (pBAM) to enable public



sharing of raw data. We developed a framework, where researchers can tune the level of privacy



and utility balance they want to achieve based on the policies and consents of the donors.

pBAMs enable researchers to share the mapped reads, which are largest data product of func- tional 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 signif- icantly.



for sharing raw reads of experiments from individuals, who do not consent public sharing of their



Our file formatting system is currently adopted by ENCODE Data Coordination Center

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 pro- files in RNA-Seq, which was addressed elsewhere [[18](#_bookmark19)]. There is also leakage from gene expression quantifications, which was shown to be connected with variants through the eQTLS [[17](#_bookmark17)]. 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



genotype data.

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 Uncer-* *tainty, Fuzziness and Knowledge-based Systems*, 2002;10(5):557-570.
    10. Sweeney L. Simple demographics often identify people uniquely. *Carnegie Mellon Univer-* *sity, 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 map- ping 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 Re-* *search*, 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.

* + 1. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. *Science*, 2012;44(5):603-608.
    2. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype- genotype data: linking attacks. *Nature Methods*, 2016;13(3):251-256.
    3. Harmanci A, Gerstein M. Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions. *Nature Communications*, 2017
    4. Beskow LM. Lessons from HeLa Cells: The Ethics and Policy of Biospecimens. *Annu Rev* *Genomics Hum Genet.*, 2016;17:395-417
    5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012;489(7414):57-74.
    6. 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.
    7. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 2013;45(6):580-585.
    8. National Institute of Health data sharing policy. https://grants.nih.gov/grants/guide/notice- files/NOT-OD-17-110.html
    9. The 1000 Genomes Project Consortium. A map of human genome variation from population- scale sequencing. *Nature*, 2010;467(7319):1061-1073.
    10. 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.
    11. 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.

* + 1. International HapMap Consortium. The International HapMap Project. *Nature*, 2003;426(6968):789-796.
    2. Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.*, 1998;80:197.
    3. 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.
    4. Howie BN, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3:* *Genes, Genomics, Genetics*, 2011;1(6):457-470.
    5. Howie BN, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate geno- type imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 2012;44(8):955-959.
    6. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathemat- ical analysis. *Genomics*, 1988;2(3):231-239.
    7. 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 associa- tion 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 Pro- ceedings 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](#_bookmark33) 30
      2. [comparison and procedure](#_bookmark34) 31
      3. [pmi for all](#_bookmark35) 32
      4. [FDR](#_bookmark36) 33
      5. [linking and gap](#_bookmark37) 34
      6. [imputation, phenotype inference and theoretical framework](#_bookmark38) 35
      7. [differential privacy](#_bookmark39) 36
      8. [pBAM](#_bookmark40) 37

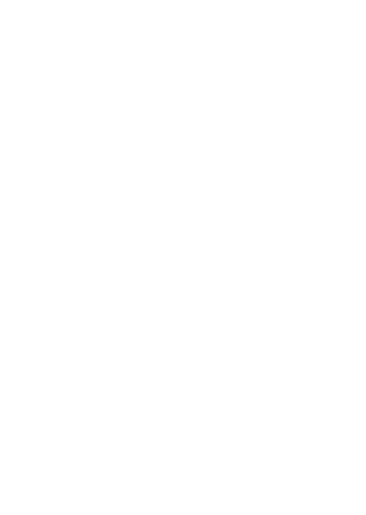
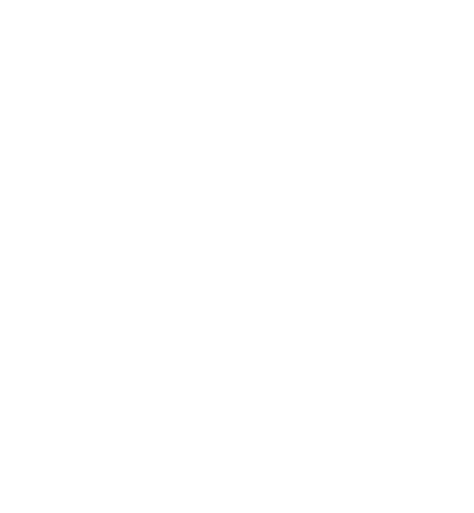
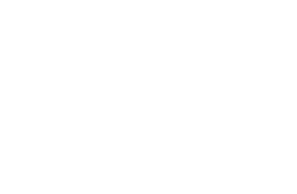
Signal

a) b)



overlap

profile



TADs

loop calling

peak calling

gene expression level

Hi-C interaction matrix

depth signal profiles

modified reads

mapped functional genomics reads BAM/SAM files

MRF

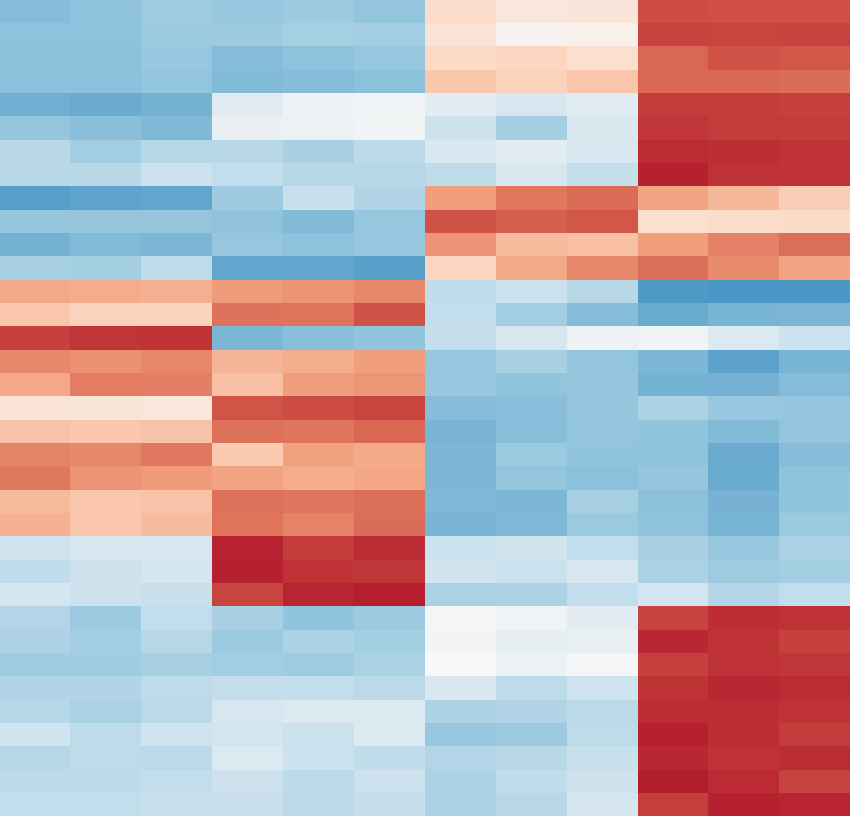
pBAM CRAM

ChIP-Seq

RNA-Seq ATAC-Seq Hi-C

Raw reads - Fastq files



samples

genes

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 and utility

trade-off balanced.

### a)

x106

7



|  |  |  |
| --- | --- | --- |
|  |  |  |

6

5

Information (bits)

4

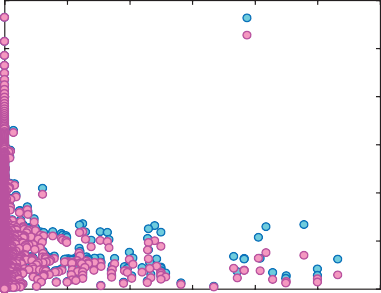
3

2

1

Information at n

Naive information 12



Naive information

Information w/ LD

**b)**

Information with LD

10



|  |  |  |
| --- | --- | --- |
|  |  |  |

8

Information (bits)

6

4

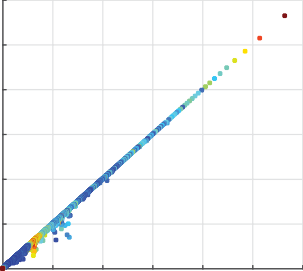
2

0

1k genome gold standard

**c)**

12



variants

rare

variants

common

variants

WGS

unique **e)**



0

0 0.1 0.2 0.3 0.4 0.5 0.6

Average mLD correlation

70

information w/ LD per variant (bits)

{

10 S =

set of

} = { si } = { vi , gi }

60

8 50

6 40

Hi-C / RNA-Seq / ChIP-Seq reads

ACTGCCCGATTCGAAACTGACTACTGCCCGATTC

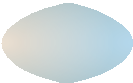
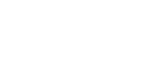
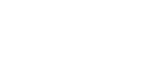
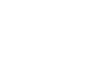
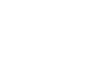
number of variants

variants

GATK

1Kgenome

30



called ariants

SF

gold s andard:

SG

FN TP FP

h(SF|SG) pmi(SF;SG) h(SG|SF)



n reads

}

5n reads

4

20

2

10

0

variants:

tandard:

0 2 4 6 8 10 12

### d)

6.9

106

naive information per variant (bits)

6.85



Information (bits)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | |  |  |  |
|  | hn |  | |  |  |  |
|  |  |  | |  |  |  |
|  |  |  |  |  |  |  |
|  | |
|  |  |  | |  |  |  |

6.8

6.75

6.7

0 0.02 0.04 0.06 0.08 0.1

1/fraction

h(SF,SG)

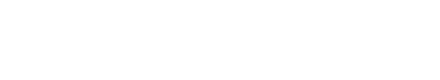
npmi ( SF;SG ) = pmi(SF;SG) / h(SF,SG)

FDR = h(SF|SG) / h ( SF )

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.

**a)** 107



**b)**

109

1010

1011

**d)**

1.4

109

1010

Total coverage (bp)

1011

1.2

1

0.8

0.6

0.4

0.2

0

106

105

naive pmi (SF; SG)

104

103

102

**c)**

107

106

105

naive pmi (SF; SG)

104

103

gold standard

WGS

Hi-C1 Hi-C2

Hi-C3

WES

Total RNA-Seq PolyA RNA-Seq

Single Cell RNA-Seq1 Single Cell RNA-Seq2 H3K4me3

H3K27ac



H3K27me3 H3K36me3 H3K4me1 H3K4me2

10-4



H3K9ac

H4K20me1 H2AFZ H3K79me2 H3K9me3



CTCF-Iyer CTCF-Snyder

CTCF-Broad



PBX3

RNAPII-1

RNAPII-2 JUND RELB HDGF



102

naive pmi (SF ; SG) per bp ( bits / bp )

109

Hi-C1 HDGF Hi-C2 WGS Hi-C3 RELB

CTCF-Snyder H3K79me2

CTCF-Iyer RNAPII - 1 H3K4me2 H3K4me1 H3K27ac H2AFZ RNAPII - 2

H3K9ac CTCF-Broad H3K27me3

JUND H3K4me3 H3K36me3

WES H4K20me1 H3K9me3 PBX3

PolyA RNA-Seq Total RNA-Seq

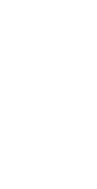
Single-cell RNA-Seq-2 Single-cell RNA-Seq-1

Total coverage (bp)

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 mean of all the ratios between the pmi and the corresponding coverage.

1



1

0.99

0.98

0.97

0.96

0.95

0.9

0.8

FDR (h (SF | SG))

0.7

0.6

0.5

109 1010 1011

Total coverage (bp)

109 1010 1011

Total coverage (bp)

109

Total coverage (bp)

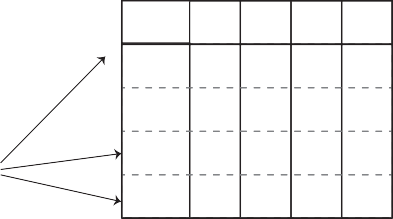
Figure 4: False discovery rate of functional genomics experiments at different coverages (a) FDR comparison for Hi-C and WGS data at different sampled coverages. (b) FDR comparison for different ChIP-Seq experiments at different coverages. (c) FDR comparison for WES and different RNA-Seq experiments.

#### a)

FGE(k)  {S(k)}

#### b)

indi

panel of genotypes stolen/hacked/queried

g1

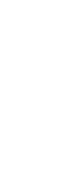
gj

gn

ind1

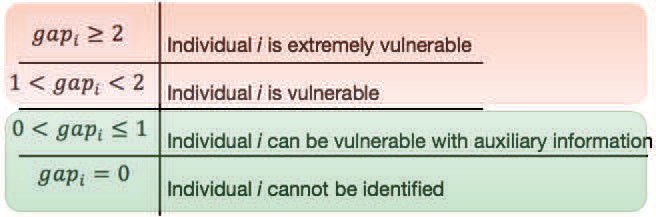
indm

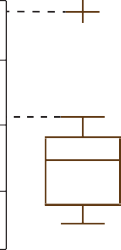
HIV

status

|  |
| --- |
| + |
| - |
| - |
| + |

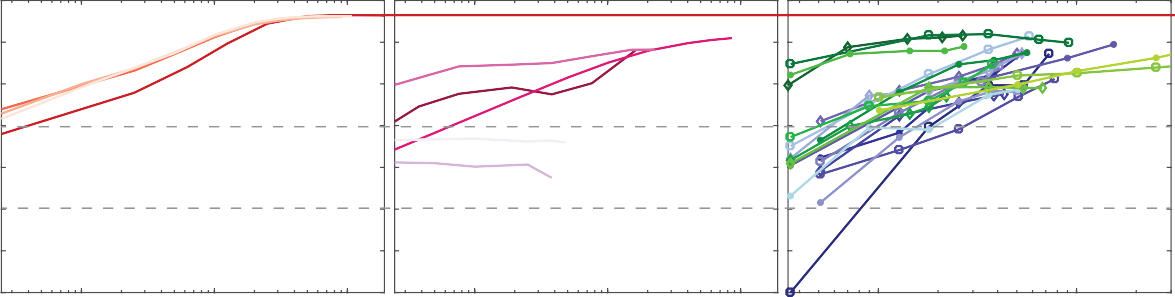
*a*

*a =* pmi ( {S(k)} ; indi ) gapi = *a* / *b*



**pmi ( {S(k)}; ind )**

*b*

3.5

3

2.5

NA12878

2

1.5

gap

1

0.5

0

109 1010 1011

Total coverage (bp)

109 1010 1011

Total coverage (bp)

108 109

Total coverage (bp)

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 sta- tus 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 at- tacker 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 indi-

vidual. (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.

**a) b)**

Total RNA-Seq

polyA RNA-Seq

All ChIP-Seq

106

SNV

SNV+imputation

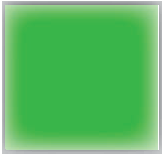
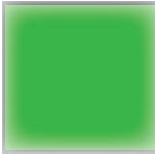
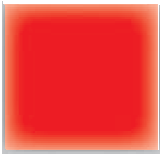
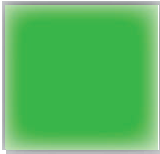
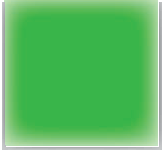
4.5

4

Phenotype Variant I.D.

Hi-C

3.5



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Blue vs. brown eye | rs1667394 |  |  |  |  |  |  |  |  |
| Brown vs. blonde hair | rs12896399 |  |  |  |  |  |  |  |  |
| Red hair | rs1805007 |  |  |  |  |  |  |  |  |
| Freckles | rs11648785 |  |  |  |  |  |  |  |  |

3

Number of SNVs

2.5



2

1.5

1

0.5

0

WGS All ChIP-Seq+RNA-Seq



yes no



w/ imputation

w/ imputation

w/ imputation

w/ imputation

**c)**

ACTGCCCGATTCGAAACTGACTGCCCGATTCGAAACGAA

di = 3 



di >0

b =  (di = 3) b =  (di > 0)

probability

N d

b = ∑N (di )

i=1

 (di ) = 1 if di > 0

= 0 otherwise

∑i=1 di

d =

N

nmpi ( SF ; SG ) ~ f ( b, d,  )

empirical dist.

expected dist.

+

-  = (+) + (-)

**d) e)**

Training data, RMSE=0.06

**Test data, RMSE=0.07**

102 102

WGS

Hi-C (decreasing depth) ChIP-Seq (decreasing depth)

R2 ~ 0.9



Hi-C

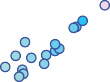
ChIP-Seq

101

101

true npmi

true npmi

100

100

10-1

10-2

10-2 10-1 100 101 102

predicted npmi

10-1

10-1 100 101 102

predicted npmi

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

##### a) b)



106

gapNA12878>2 gap >2

NA12878

gapNA12878>2

gapNA12878>2

naive self-information per variant (bits)

unique

12

10

8

6

4

2

0

##### c)

106

All SN V

Indel

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| d | ouble |  |  |  |
|  |  |  |  |  |
|  |  |  | r var | are iants |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

0 0.5% 1% 1.5% 2%

genotyping frequency

5000

4500



4000

3500

3000

2500

2000

1500

1000

7

6

number of variants

pmi( NA12878 ; indi )

5

4

3

2

##### d)

3.5

all NA12878 genotypes

* unique genotypes
* double genotypes

- rare genotypes

105 3

All SNV

Indel

All SNV

Indel

naive pmi ( rnaseq ; 1kG ) (bits)

104 2.5

gapNA12878

103 2

102 1.5

101

109 1010 1011

Total coverage (bp)

1

2 4 6 8 10

Total coverage (bp) 1010

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.

**a) b)**



**.diff**

**reference**

**genome**

**p-tools**



**pBAM**

reference genome: **....ATCGTGTAACGTGC - - - -AGTGGGCTAAAC....**

mapped read:

**AA - - TGCTCTCAGCGGG**

**deletion insertion mismatch**



**BAM read: AATGCTCTCAGCGGG**

**pBAM read: AACGTGCAGTGGGCT**

added noise to the depth signal

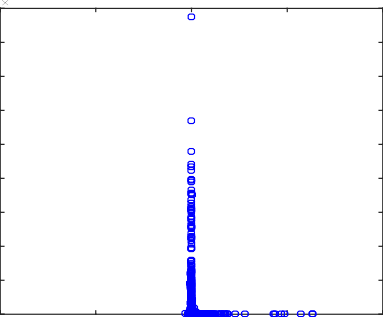
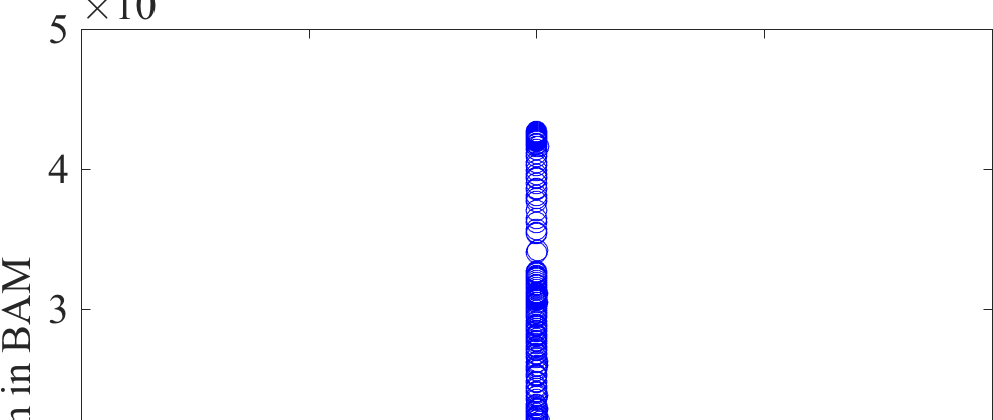
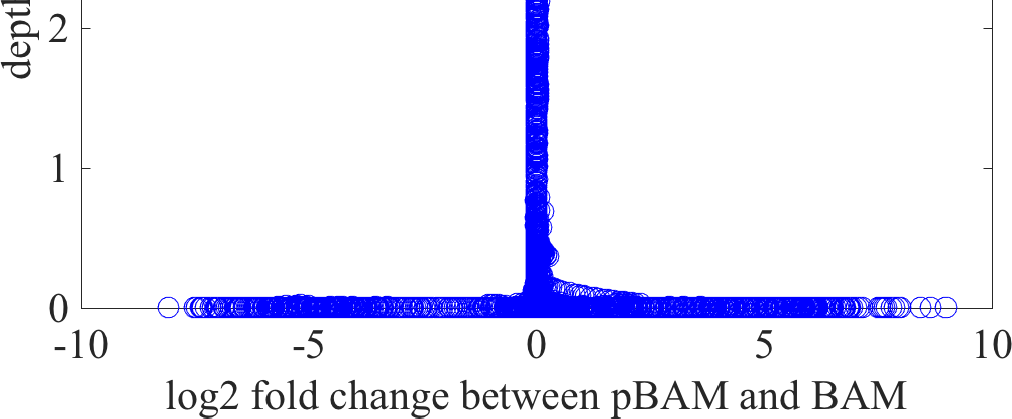
added noise to the depth signal

**c)**



**BAM**

4



10

9

8

7 **d)**

6

**Utility**

5

depth in BAM

|  |  |
| --- | --- |
| **bp resolution** | **> 99.9 %** |
| **exon resolution** | **= 99.9 %** |

4

3

2

1

0

-10 -5 0 5 10

log2 fold change between pBAM and BAM

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.