Sensitive information leakage from functional genomics data: Theoretical quantifications & practical file formats for privacy preservation

Gamze Gürsoy[1,2], Arif Harmanci[3], Molly E. Green[1,2], Fabio C.P. Navarro[1,2] and Mark Gerstein*[1,2,4]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

[3]School of Biomedical Informatics, Center for Precision Health, University of Texas Health Sciences Center, Houston, TX, 77030, USA

[4]Department of Computer Science, Yale University, New Haven, CT 06520, USA

May 11, 2018

*pi@gersteinlab.org; Corresponding Author

Abstract

Functional genomics experiments on human subjects present a privacy conundrum. On one hand, many of the conclusions we infer from these experiments are not tied to the identity of individuals but represent universal statements about biology and disease. On the other hand, by virtue of the experimental procedure, the sequencing reads are tagged with small bits of patients' variant information, which presents privacy challenges in terms of data sharing. There is great desire to share data as broadly as possible. Therefore, measuring the amount of variant information leaked in a variety of experiments, particularly in relation to the amount of sequencing, is a key first step in reducing information leakage and determining an appropriate "set point" for sharing with minimal leakage. To this end, we derived information-theoretic measures for the private information leaked in experiments and developed various file formats to reduce this during sharing. We show that high-depth experiments such as Hi-C provide accurate genotyping that can lead to large privacy leaks. Counterintuitively, low-depth experiments such as ChIP and single-cell RNA sequencing, although not useful for genotyping, can create strong quasi-identifiers for re-identification through linking attacks. We show that partial and incomplete genotypes from many of these experiments can further be combined to construct an individual's complete variant set and identify phenotypes. We provide a proof-of-concept analytic framework, in which the amount of leaked information can be estimated from the depth and breadth of the coverage as well as sequencing biases of a given functional genomics experiment. Finally, as a practical instantiation of our framework, we propose file formats that maximize the potential sharing of data while protecting individuals sensitive information. Depending on the desired sharing set point, our proposed format can achieve differential trade-offs in the privacy-utility balance. At the highest level of privacy, we mask all the variants leaked from reads, but still can create useable signal profiles that give complete recovery of the original gene expression levels.

1 Introduction

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

Early genomic privacy studies focused on the identification of individuals in a mixture by using phenotype-genotype association [5, 6]. These studies showed that private information of an individual, such as participation in a drug-abuse study, can be revealed [5, 6]. With the increase of large-scale genomics projects such as the Personal Genome Project [7] or recreational/direct-to-consumer genomic databases, researchers showed that multiple datasets can be linked together to infer sensitive information such as pariticipant's surnames [8] or addresses [10]. 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 [9].

Functional genomics experiments provide a wealth of information on genomic activities related to developmental stages or diseases that are essential for personalized medicine. These studies use large-scale high-throughput assays to quantify transcription (RNA-Seq) [11], epigenetic regulation (ChIP-Seq) [12] or the three-dimensonal (3D) organization of genome (Hi-C) [13] in a genome-wide fashion under different conditions (e.g., samples from patients and healthy individuals). Inferring biological information from functional genomics experiments is a multi-step procedure, in which progressive summarization of the data from raw sequencing reads to the gene quantifications, transcription factor (TF) binding peaks or chromatin interaction matrices is per-

formed. Although activities of the functional genome are not necessarily tied to an individual's genotype, reads from these experiments are derived from the biosamples that belong to individuals; hence, they are tagged with individuals' variants. Public sharing of such raw data raises privacy concerns. In order to share high-utility data while preserving 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 determining the set point is the lack of systematic quantification of private information leakage from functional genomics data. Figure 1 summarizes the processing steps of RNA-Seq experiments as an example of how summarization decreases the risk of privacy while greatly decreasing the amount of sharing and the utility of the functional genomics data. In detail, functional genomics data analysis starts with the generation of DNA/RNA sequencing reads that are stored in a special file formats called FASTQ [14]. These files are large in size ranging from 5 GB up to 60 GB depending on the purpose of the experiment. They are then mapped to human reference genome and stored as compressed binary file types called binary alignment map (BAM) and/or compressive alignment map (CRAM) that are derived from the sequence alignment map (SAM) files in text format [15]. File formats such as CRAM have been developed to remedy the ever increasing amount of data; compared to BAM files, CRAM provides up to a ten-fold decrease when information loss is tolerated [16]. Further summarization of the mapped reads (such as signal profiles or gene expression quantification) still allows researchers to make accurate biological conclusions, while providing$\sim$20-fold further data reduction. Although overall aggregation and averaging reduces biological information, private information leakage also decreases (Figure 1).

In particular, read alignment files (SAM / BAM / CRAM) are of great interest due to the large amount of biological data they provide, as they constitute the most important input of the majority of genome annotation pipelines. However, these files contain sequence information of the individuals that may leak sensitive data. Depending on the depth of the functional genomics experiment,

raw reads can be used to identify private single nucleotide polymorphisms (SNPs), small isnertions and deletions (indels), and structural variants. However, current policies related to the public sharing of the BAM files are somewhat ad-hoc. For example, for the genome of the HeLa cell line, the raw reads from Hi-C experiments require special access [17]. By contrast, reads from ChIP-Seq and RNA-Seq experiments are publicly available [18]. 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. However, it is not clear that these reads are leakage free. Although private information leakage from summary-level functional genomics data have been quantified previously [19, 20, 21] the lack of a systematic quantification of private data leakage from BAM files makes it difficult for biomedical data sharing policymakers to protect individuals' sensitive information in a consistent fashion. The CRAM format provides the option for the users to convert BAM files into lossy compression, in which quality scores of the alignments are manipulated. This, in turn, can be used to decrease private information leakage [16]. However, privacy leaks still occur due to the containment of mismatched information of the reads with respect to reference genome [16]. The mapped read format (MRF) was introduced as a conceptual format to remedy privacy concerns; in this case, keeping the sequence of reads is optional [22]. 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 leakage 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 files 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. Access to private data requires use agreements that have expiration dates and a tremendous amount of bureaucracy connected to them. 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 the establishment of their own databases. However, because such annotations are derived from private data, establishment and distribution of these databases require extra levels of privacy-related bureaucracy. Another example of the challenges associated with private data is that big consortia such as the Encyclopedia of DNA elements (ENCODE) [23], the Cancer Genome Atlas (TCGA) [24] or the Genotype-Tissue Expression project (GTEx) [25] are funded to enable a collaborative 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 prohibited according to data use agreements. Moreover, when multiple institutions have required access to the same data, they still cannot exchange files with each other. These challenges create a bottleneck and hinder the progress of important biomedical findings. Open data helps the advancement of biomedical data science not only by easing access to the data, but also by helping with speedy assessment of tools and methods, and in turn, reproducibility. Funding agencies and research organizations are increasingly supporting new means of data sharing and new requirements for making data publicly available while preserving participants' privacy [26]. In an attempt to consider both sides of the coin, we aimed to determine how much information is enough information to identify individuals and how we can protect the sensitive information with minimal loss of utility in a public data sharing mode. To this end, we derived novel information theory-based measures and applied these measures to quantify the amount of leaked information in various functional genomic assays from ENCODE [23] and other sources [17] at varying coverage. Based on our findings, we developed new file formats that allow the public sharing of read alignments of functional genomics experiments, while protecting the sensitive information and minimizing the amount of private data that requires special access and storage. Our file format manupilation system achieves different levels of privacy versus utility balance with an adjustable parameter.

In this study, we used an indivudial (NA12878) as a case example and their 1000 genomes genotypes as the gold standard [27]. We sampled reads from the sequencing data of functional genomics experiments at increasing coverage, and detected SNVs and indels using Genome Analysis Toolkit (GATK) best practices recommendations [28, 29]. We propose a new metric for quantifying the amount of information that can be obtained from sequencing data with respect to the gold standard. We next present a simple and practical instantiation of a linking attack with the assumption of adversaries accessing an increasing amount of the sequencing data. We show that individuals are vulnerable to identifications even at small coverage 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 comprimising an individual's sensitive information. We propose privacy-enhancing file formats that hide variant information, are compressed, and have a minimal amount of utility loss.

2 Results

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

An individual's genome can be represented as a set of variants. Each variant is composed of the chromosome to which it belongs, location on that chromosome, the alternative allele, and the 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 a heterozygous variant and 2 for a homozygous

variant. Note that we calculate the information with respect to reference genome, therefore $g_i = 0$, where the alternative allele and the reference allele are the same is not considered. We can then calculate the naive self-information of S in bits as

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

In eq. 1, N is the total number of variants in an individual's genome, $p(s_i) = n_i/n_T$ is the genotype frequency, in which n_i is the number of individuals with variant $s_i = \{v_i, g_i\}$ and n_T is the total number of individuals in the panel (see Figure 2a). Note that we denote $h(S)$ as "naive" information because it is an estimate of the real information in a situation, in which the population to which the individual belongs is unknown and the number of inidivuals are finite. Eq.1 holds true 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 to which 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 costs) to the identification attack. Eq.1 is also an estimate of the information when we consider all the individuals in the world (i.e., $\lim_{n_t \to \infty} h(S)$).

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 the HapMap project [30]. LD scores are pairwise correlations between variants ($LD(s_i, s_j)$), which we consider as the prior information on the existence of a variant given other variants in the same LD block exist in a genome. Then, the information with LD consideration is calculated as

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

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

Figure 2c shows a negligible difference between the naive information and information with LD consideration for the 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 found that highly informative variants do not exhibit any difference due to the low LD correlations (Figure 2b). We further show that the number of variants with differences between information with or without LD consideration is small compared to the number of variants low LD correlations on average (SI Figure 1). This also shows that information ($h(S)$) is driven by the rare variants.

We then estimated the information when the population size is infinite [31]. We sampled fractions on the order of 10%, 20%,..., 100% individuals from the 1000 Genomes panel (total of 2504 individuals) and calculated the information using the sampled distribution of genotypes. We repeated this calculation 100 times and calculated the mean information for each sampled fraction. The relationship between the inverse of the sample fraction and the information fit best to a power function with two terms ($y = mx^b + n$, $R = 0.99$). The y-intercept of the curve is the extrapolation of information when the population size approaches infinity ($1/\infty = 0$, Figure 2c). We again found a negligible difference between the naive information and the information when the population size is infinite (Figure 2c). We also calculated the information by starting from a single individual and adding individuals one by one to the population (SI Figure 2). These individuals were simulated using the genotype frequencies in the 1000 genomes panel and the LD information from the HapMap project (see SI methods). Both the information calculation and the KL-divergence between different-sized populations showe that as the size of the population increases, the difference in the information decreases and eventually becomes negligible (SI Figure 2)

The calculations above show that naive information can be an accurate approximate to the private information content of an individual's genome when the individual's population is unknown 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 2c). That is, an adversary with no prior knowledge of the population of the sample and limited number of individuals in a known genotype panel can accurately approximate the private information in the genome of the sample.

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

We next aimed to understand the relationship between the leaked information and the coverage to make a fair comparison between different functional genomics experiments. We sampled c amount of total nucleotides from the 24 different functional genomic experiments and from whole genome sequencing (WGS) and whole exome sequencing (WES) data of sample NA12878 (see SI Table 1). We used GATK to call SNVs and indels with the parameters and filtering suggested in the GATK best practices [28, 29]. We used the genotypes in the 1000 Genomes panel for NA1278 as the gold standard. We used "naive" pointwise mutual information (pmi) as a measure to quantify the association between the gold standard and the called variants. If $S^G = \{s_1^*, .., s_i^*, ..., s_M^*\}$ is the set of variants from the gold standard and $S^F(c) = \{s_1, .., s_i, ..., s_M\}$ is the set of variants called from the c total sequencing coverage of a functional genomics experiment, then the set $A = S^G \bigcap S^F(c)$ contains the variants that are called and are in the gold standard set. If $A = \{a_1, .., a_i, .., a_T\}$, then

$$pmi(S^G; S^F(c)) = -\sum_{i=1}^{i=T} log_2(p(a_i)) \tag{3}$$

We then added more coverage to the sampled coverage and repeated the calculation. We repeated this procudere untill we depleted all the reads of a functional genomics experiment. The

overall process is depicted in Figure 2a. Figure 2a also shows how different measures such as self-information, pmi or joint information relate to each other [32]. More detail can be found in SI Methods.

2.3 Private information leakage in 24 functional genomics experiments with different coverage

We calculated naive pmi values for 24 functional genomics experiments with different coverage. The experiments involved whole genome approaches such as Hi-C, transcriptome-wide assays such as RNA-Seq, and targeted assays such as ChIP-Seq of histone modifications and transcription factor binding (SI Table 1). In addition, we calculated the pmi for WGS, WES, and SNP-ChIP for comparison (Figure 3).

As expected, the Hi-C data contained almost as much information as the WGS data and more information than the SNP ChIP array data. The WGS data contained more information than the Hi-C data at the beginning of the sampling process. As we sampled nucleotides between 1.1 and 10 billion bps, the information content of the Hi-C data surpassed the WGS data (Figure 3a). We speculate that this is due to a higher quality of genotyping 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 could not infer as much information from the ChIP-Seq reads (Figure 3b). Surprisingly, many of the ChIP-Seq assays such as the ones targeting CTCF and RNAPII contained a large amount of information at low coverage. Furthermore, comparison between WES and different RNA-Seq experiments showed that none of the RNA-Seq experiments contained as much information as the WES data; this is due to the fact that RNA-Seq captures reads only from expressed genes in a given cell type (Figure 3c). An unexpected observation was that more information could be inferred from polyA RNA-Seq data at low coverage compared to WES and total RNA-Seq data. To make a fair comparison between each of these assays, we calculated the mean pmi per base

pair depicted in Figure 3d. To do so, we normalized the pmi values by the amount of coverage (c). We then averaged each by the number of times (n) we performed sampling on that experiment ($\frac{\sum pmi(S^F(c);S^G)/c}{n}$). The Hi-C and ChIP-Seq experiments targeting the transcription factor HDGF provided more genotyping information per base pair compared to the WGS data. The RNA-Seq experiments provided the least genotyping information per base pair (Figure 3d).

2.4 Genotyping accuracy

In light of our finding that genotyping can be performed using low-depth, biased functional genomics experiments, we next assessed the accuracy of genotyping by calculating the false discovery rate at different coverage. This approach also measures how much noise each assay captures. We defined the false discovery rate as the ratio between the information obtained from the incorrectly called variants ($h(S^F(c) \mid S^G)$) and the information obtained from all the called variants ($h(S^F(c))$) at a given sequencing coverage c, namely

$$FDR(S^F(c)) = h(S^F(c) \mid S^G)/h(S^F(c)) \tag{4}$$

Figure 4a shows that the false discovery rate for Hi-C data was lower compared to WGS data at lower coverage. We attribute this finding to the deeper sequencing of the genomics regions in close spatial proximity. Hence, sampling more reads from regions at low coverage is more likely compared to uniform sampling of reads from WGS. ChIP-Seq data had a comparable false discovery rate to WGS and Hi-C data given the shallow sequencing depth. ChIP-Seq targeting CTCF had the lowest false discovery rate (Figure 4b). We further found that the polyA RNA-Seq experiment had the lowest false discovery rate compared to WES and total RNA-Seq. This could be attributed to the deeper sequencing of regions containing highly expressed genes and deeper sampling from these regions. In general, assays targeting the transcriptome such as WES and RNA-Seq produced noisier genotypes compared to WGS and Hi-C experiments; single-cell

RNA-Seq was the noisiest among all the assays, as expected (Figure 4c).

2.5 Linking attack scenario

Linking attacks aim to re-identify an individual by cross-referencing datasets (Figure 5a). For example, in a hypothetical scenario an attacker aims to query an individuals 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 the adversary obtains access to this dataset by either lawful or unlawful means. Now lets assume that the attacker has access to a biosample. This could be partial or complete mapped reads from functional genomics experiments or a saliva sample taken from a used glass. The idea is to genotype the biosample and find the matching genotypes 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. The question becomes how well an adversary should sequence an individuals genome to be able to perform successful linking. Specifically, the 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 needs 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 *pmi*. This reveals the sensitive information for the linked individual to the attacker. We then consider a scenario in which the attacker has access to 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 coverage, we defined a metric for linking accuracy called gap_{query}. Let assume S_j^{DB} is the set of variants that belongs to the jth individual in the genotype panel and $S_{query}^F(c)$ is the set of variants that was called from the functional genomics experiments of the query individual at c total sequencing coverage. We first calculate the pointwise mutual information between every individual in the panel and the query as $pmi(S^F(c); S_j^{DB})$. We then ranked all the *pmi* values in a decreasing order such that;

$$pmi(S^F(c); S_i^{DB})^{(1)} > pmi(S^F(c); S_j^{DB})^{(2)} > ... > pmi(S^F(c); S_m^{DB})^{(N)}$$

In our linking attack scenario, we calculate a metric called gap_{query}, which is the ratio between the *pmi* of first ranked individual and that of second ranked individual. The idea is that if the first ranked individual is separated from the rest of the population (i.e., gap_{query}¿1), then the first rank individual is predicted as query (Figure 5).

As our query individual (NA12878) was in the panel, we could measure the accuracy of this prediction by further extending the definiton of gap_{query}. We calculate the gap_{query} for three possibilities: (1) First ranked individual is NA12878, (2) first ranked individual is not NA12878, but NA12878 is in the first five ranked individuals, and (3) none of the top five mathing individuals are NA12878. In the possibility (1), the attacker makes a correct prediction. The strength of this prediction is the gap_{query}, which is measured as the fold change difference between the pmi of best matching individual (coorect prediction) and the second best matching individual. In the possibility (2), the strength of this prediction (gap_{query}) is measured as the fold change difference between the pmi of the real individual, that is ranked somewhere between 2nd to 5th and the pmi of the best matching individual, that is the misprediction. In the possibility (3), the attacker makes a false prediction that the query cannot be retrieved from the panel, there gap_{query} becomes 0. We can formulate this as;

$$
\begin{aligned}
gap_{query} &= \frac{pmi(S^F(k); S_i^{DB})^{(t)}}{pmi(S^F(k); S_j^{DB})^{(2)}}, \text{ if } S_i^{DB} = \text{query and } t = 1 \\
gap_{query} &= \frac{pmi(S^F(k); S_i^{DB})^{(t)}}{pmi(S^F(k); S_j^{DB})^{(1)}}, \text{ if } S_i^{DB} = \text{query and } t \in 2,3,4,5 \\
gap_{query} &= 0, \text{ otherwise}
\end{aligned}
\tag{5}
$$

We then defined that if gap_{query} is 0, then the individual cannot be identified as there are other individuals in the panel that have the matching genotypes. If $0 < gap_{query} \leq 1$, then the individual might be vulnerable with auxiliary data such as gender or ethnicity, because he/she is in the top five matching individuals. If $1 < gap_{query} \leq 2$, then the individual is vulnerable as we can identify him/her with a one- to two-fold difference between him/her and the second best match. Lastly, if $gap_{query} > 2$, then the individual is extremely vulnerable with more than a two-fold difference between him/her and the second best match. A detailed flowchart of the linking attack is shown in Figure 5a.

We found that NA12878 was extremely vulnerable even at the lowest sampled coverage for Hi-C and RNA-Seq data (Figure 5b). Interestingly, between $\sim$1.1 and 10 billion base pairs, the Hi-C data exhibited higher linking accuracy than the WGS data, consistent with the previous observation of *pmi* shown in Figure 3a. The total coverage of ChIP-Seq data compared to Hi-C and RNA-Seq data was quite low (SI Table I). However, the linking accuracy of ChIP-Seq was as good as Hi-C and WGS (Figure 5b), showing extreme vulnerability of individuals with respect to a release of a small amount of data. More strikingly, the attacker can link NA12878 by using the reads of single-cell RNA-Seq data, which cover a small portion of the genome in a single cell (Figure 5d). We then added the variants of NA12878s parents to the 1000 Genomes genotype panel and repeated the linking attack. We found that although NA12878 was still extremely vulnerable to re-identification with the presence of her parents in the database, the second-best matching individuals were her parents (SI Figure 3). This shows that, using the metric *gap*, an adversary can

also identify individuals related to the target individual.

2.6 An individuals genome can be accurately approximated from publicly available data by imputation

To determine whether an attacker can correctly assemble an individuals variants by only using the reads from ChIP-Seq and RNA-Seq experiments, we imputed variants by using IMPUTE2 [33, 34, 35] 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 did not contribute to the information due to high correlation with the called variants (SI Methods and SI Figure 3), total number of captured variants increases significantly (Figure 6a). By using shallow sequencing data of ChIP-Seq and RNA-Seq, we were able to call and impute almost as many variants as the gold standard.

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

2.7 Toy model for estimating the amount of leaked data without variant calling

Genotyping from DNA sequences is the process of comparing the DNA sequence of an individual to that of the reference human genome. To be able to successfully genotype, 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 are sequenced repeatedly, the depth per base pair follows the Poisson distribution with a mean that can be estimated from the read length, number of

reads, and the length of the genome [36]. As functional genomics experiments aim to find highly expressed genes, TF binding enrichment, or 3D interactions of the genome, it is expected that the sequencing depth per base pair does not follow Poisson statistics. Thus, genotyping using reads from functional genomics experiments is biased towards variants that are in the functional regions of the cell types/lines of interest.

To this end, we we hypothesized that genotyping from sequencing-based functional genomics data depends on the average depth per base pair ($\overline{d}$) , and the total fraction of the genome that is represented at least by one read (i.e., the breadth, $b = \sum_{i=1}^{N} \delta(d_i)$, such that $\delta(d_i) = 1$ if $d_i > 0$, $b = 0$ otherwise and 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, Figure 6c). The bias parameter β is composed of two terms: (1) the negative bias $\beta-$ and (2) the positive bias $\beta+$. The negative bias estimates if there is an increase in the number of low depth basepairs relative to the mean with respect to the expected Poisson distribution; the positive bias estimates the increase in the number of high-depth basepairs (see SI for more details).

To quantify the genotyping accuracy from the functional genomics data, we used "naive" normalized *pmi* (*npmi*, see SI for details). This approach takes into account the information from the correctly identified genotypes ($pmi(S^F;S^G)$), the information missed that is in the gold standard ($h(S^G \mid S^F)$) and the information from the incorrectly identified genotypes (i.e FDR, $h(S^F \mid S^G)$) and normalizes it with the joint information of called variants and gold standard variants ($h(S^F,S^G)$ as;

$$npmi(S^F;S^G) = \frac{pmi(S^F;S^G)}{h(S^F,S^G)} = \frac{pmi(S^F;S^G)}{h(S^G \mid S^F) + pmi(S^F;S^G) + h(S^F \mid S^G)} \tag{6}$$

To be able to get a fit for the relationship of $npmi(S^F;S^G) = f(\overline{d_F}, b_F, \beta_F)$, we used Gaussian Process Regression (GPR) [37] to fit 40 training data points and achieved a root mean square

error (RMSE) of 2.60 with the values ranging between [0,100] (Figure 6d). We used five separate data points as a test set and achieved an RMSE of 2.47 was acheieved (Figure 6d). We performed regression learning using a ten-fold cross-validation to protect against overfitting. This toy model represents a conceptual theoretical framework limited to the small sample space available. It shows that the amount of leaked data from functional genomics experiments can be estimated without the need of performing time-consuming genotyping calculation.

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

We next analyzed 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 prevent linking. A total of 11,472 variants along with their genotypes were observed only in NA12878, which we refer as "singletons" (Figure 6a). Please remember that we used the terminology variant not only for the location and minor allele of the SNV but also the genotypes (homozygous or heterozygous). Therefore number of singletons in this context are more than the number of *de novo* variants. After the removal of singletons from the NA12878 variant set, we calculated the $gap_{NA12878}$. Surprisingly, the linking accuracy was affected minimally compared to using the all of the NA12878 variants (Figure 6b). We then created another set (doubletons, Figure 6a), that included the variants observed in NA12878s genome as well as one more individual in the 1000 genomes genotype panel (total of 16,305 genotypes). We again found that the individual was extremely vulnerable to linking attacks ($gap_{NA12878} > 2$,Figure 6b). We then relaxed our cut-off further to remove the genotypes that are observed in NA12878's genome as well as at most 1.5% of the population ("rare genotypes", total of 124,093 genotypes, Figure 6a). This also did not affect the overall linking ($gap_{NA12878} > 2$,Figure 6b).

These rare genotypes were observed in 64 or less individuals including NA12878. A practical solution to the re-identification problem using functional genomics data would be masking or removing such rare genotypes from the reads. However, as iteratively shown here, although rare variants are extremely informative and sufficient enough to achieve re-identification through linking attacks, their removal is not sufficient to prevent re-identification. That is, not only the rare genotypes but also the unique combination of common genotypes are identifiers of the 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 could still link NA12878 to the correct genotypes successfully with an extreme vulnerability ($gap_{NA12878} > 2$, SI Figure 3).

We then analyzed the contribution of small indels to the naive information and whether accurate linking was possible when we removed all the single nucleotide mutations from the data and kept the indels. Figure 6d shows the information contribution of the indels. Although naive *pmi* from indels were much smaller compared to single nucleotide mutations, a high linking accuracy could be achieved by using only indels even at small coverage (Figure 6d). This linking attack is done using one of the noisy data set we have (total RNA-Seq) to make linking more difficult.

2.9 Privacy-preserving file formats for read alignments from functional genomics experiments and relation to *k*-anonymity

Sharing raw read alignments (to the reference genome) from functional genomics experiments is extremely important in developing analysis methods and discovering novel mechanisms about the human genome. Ideally, one would share the maximal amount of information with minimal utility loss while largely maintaining an individual's privacy. As a privacy metric, we aimed to prevent leakage of any variant as well as any quasi-identifier that can lead to identification of the position of variants in the genome. We introduced a user-identified privacy-utility balance that can be adjusted according to the patients consents and institutions policies. By using the concept

of k-anonymity [10], we applied a privacy-preserving transformation to the alignment files such that calling variants from transformed files is largely prevented while quantifications related to the functional genome is possible with minimal error (Figure 8a).

A release of data possesses the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least $k-1$ individuals whose information also appear in the release. Although this concept was developed for the release of datasets with individuals, we can think of a raw alignment file (BAM) as a dataset, where information for each read is contained. Let's assume a BAM file is a dataset D, where each entry is a read. The desire is to release dataset D in a form (say D^*) such that it does not leak variants from the reads, but in the meantime any calculation f based on D and D^* retrieves almost the same result. Two general methods can achieve k-anonymity for some value of k: suppression and generalization. If every column in D is an attribute (such as read length, cigar, sequence, or quality value), then replacing an attribute with an asterisk (*) is suppression and changing an attribute with a more general value is generalization. For example, in our file format transformation, we can replace sequence and sequence quality attributes with an asterisk (suppression), and transform the cigar of the read from partially mapped to fully mapped (generalization) to achieve 3-anonymity with respect to attribute sequence, sequence quality, and cigar (see SI Methods for details). Now let's say the privacy-preserving transformation is done through a function $P_{Q,r}$ such that $P_{Q,r}(D) = D^*$. Q is an operation such as "removal of small indels", "removal of mismatches", "removal of large indels", or "removal of all variants". r is the amount of reads to be manipulated given the operation Q. A calculation f can be a signal depth profile calculation, TF binding peak detection, or gene expression quantification (Figure 8a). Then, we can reconstruct an equation for each unit i as

$$\frac{f(D)}{f(D^*)} = e^{\varepsilon_i}, \tag{7}$$

where a unit i can be a single base pair, an exon, or a gene depending on the function f. In turn, ε_i can be calculated as the log-fold change between the results derived from two datasets. This is also a quantity commonly used to compute differential gene expression [38] or ChIP-Seq binding enrichment over controls [39], and can be used analogously in this context, where log-fold change is the differential signal depth or expression when the manipulated data is used as an input.

Note that $\mid \varepsilon_i \mid$ is a measure of error of the new dataset $D*$. We then calculated the distribution of $\mid \varepsilon_i \mid$ values over every unit and found the mean $\mid \varepsilon \mid$ per unit as the overall error. The level of privacy is controlled by the function $P_{Q,r}$, where Q determines the rmoval of type of entries and r determines the number of entries of the given operation Q that are manipulated. For any particular operation, the obvious threshold could be the size of the indels, minor allele frequency (MAF), or the depth of a particular unit. These thresholds can be converted into the fraction of reads affected. For example, if Q is the removal of indels and r is the reads that contain indels with $MAF < 0.01$, then only reads that have indels with $MAF < 0.01$ will be manipulated in the transformed $D*$.

We constructed the privatized file format pBAM from data D^* as follows. The reads from the BAM files were categorized as perfectly mapped reads and reads with mismatches, insertions, deletions, or soft- and hard-clipping. $P_{Q,r}$ replaces the sequence and the quality string of all of reads with an asterisk and manipulates the cigars, alignment scores (AS tag) and the strings for mismatching positions (MD tags) of the reads that are defined in Q and r. pBAM files can be thought of as scrubbed privacy-preserving binary alignment files and the operation Q and amount r as the level of scrubbing. pBAM files can also be created from BAM files that are obtained by mapping sequences to the transcriptome coordinates, which is essential for gene quantification. Our transformation function $P_{Q,r}$ is general and can be applied to any alignment file type such as SAM, CRAM, and MRF to create a new privatized file format. These files will be concordant to use with tools such as samtools, cramtools, and mrftools.

We calculated the signal depths of each base pair in the genome using an NA12878 polyA RNA-Seq (see SI Table I) BAM file using STAR [40]. We then converted the BAM file into pBAMs with different Qs and calculated the signal depth of each base pair. Figure 8b shows the number of base pairs with $\varepsilon_i > 0$ with respect to the number of base pairs with no change between BAM and pBAM. We did the same calculation by averaging signal over exons as well (Figure 8b). Furthermore, we created pBAM files for the BAM files mapped to the reference transcriptome and compared the gene quantification with the gene expression levels calculated from original BAM files by using RSEM for gene quantification and STAR for transcriptome alignment [40, 41]. We found no difference between the gene expression levels calculated using original BAM files and pBAM files (see Figure 8b and SI Methods for how we treated transcriptome alignments). Overall, when we removed all the variant leakage from the BAM files, we found a 0.18% difference at the base pair resolution, 0.27% difference at the exon resolution, and 0% difference at the gene level. When we removed leakage associated with the mismatches, we did not see any difference (Figure 8b). When we removed leakage associated with indels, we found a 0.0016% difference at the base-pair resolution, 0.0011% at the exon resolution, and 0% at the gene level. When we removed leakage associated with split reads, we found a 0.17% difference at the base-pair resolution, 0.26% at the exon resolution, and 0% at the gene level. Figure 8c shows the change in ε with respect to increasing r for different operations Q. When the mismatches are manipulated, the resulting signal profiles are not affected. Hence, the manipulated dataset will retrieve the same results as the original dataset regardless of the number of reads (r). However, manipulating indels and split reads will result in changes in the utility of the new file formats. It is useful to note that, for example, ENCODE consortium adopts processing pipelines, in which split reads are discarded. Therefore, utility of pBAMs are particularly high when split reads are discarded.

The pBAM file format contains necessary information to be used in functional genomics pipelines such as gene expression quantification and TF binding peak calling. The difference between the

results of the ENCODE Chip-Seq TF binding peak calling pipeline (MACS2 [39]) was even more negligible when BAM and pBAM were used as input (SI Figure 6). We then created a .diff file format containing the original information that was manipulated in the pBAM file. With the motivation of keeping the size of private file formats relatively small, we report only differences between BAM and pBAM in the .diff file by avoiding printing any sequence information of the reads that can be found in the reference human genome (see SI Methods). The .diff files are private files that require special permission for access. A user is able to retrieve the original BAM file when they have access to the .diff file by using our collection of scripts called "p-tools" that can convert pBAM + .diff + reference genome into the original BAM file (Figure 8d).

2.10 Comparison of private information leakage from different layers of data stack of RNA-Seq

To demonstrate the extent of leakage in raw alignment files, we reviewed all the known sources of information leakage from different points of the data summarization process (Figure 1a,b). As we pointed out throughout this study, the most obvious leakage is directly from the reads and can be largely avoided by converting the BAM files into a pBAM format. The next source of leakage is from the signal profiles, which we studied extensively [21]. Additionally, leakage can come from the quantification of expression values. Given a population of individuals, these gene expression values can be related to variants through eQTLs, and hence can create leakage [20].

In Figure 1c, we calculated the potential number of variants one can obtain from a typical RNA-Seq experiment: (1) At the read level, we can potentially observe all the SNVs on the exons; however, only a fraction is accessible through RNA-Seq depending on which gene is expressed in which cell line/type. (2) At the pBAM level, the information leakage will vary based on the operation Q. At the highest level of privacy, we theoretically remove all the variants that can be observed from the reads. However, we do not discard the possibility of discovering leakage from pBAMs

using more complicated algorithms. (3) At the signal profile level, we can potentially observe all the deletions on the exons. However, only a fraction is accessible to the experiment depending on the expressed transcripts in the given cell type/line. Therefore, following the genotyping described in ref. [21], we calculated the number of deletions that can be genotyped from the signal profile of the polyA RNA-Seq experiment of the individual NA12878 as the accessible deletions in Figure 1c. (4) At the gene expression quantification level, the potential number of variants that can be observed are all the eQTLs connected to the genes. We calculated the accessible variants through gene expression quantification as the average number of eQTLs per individual based on the calculations in ref. [20].

In summary, the amount of private information leakage from raw alignment files is almost $\sim 1,000$ times more than the amount of leakage from the signal profiles and gene expression levels. This shows the extent to which pBAMs protect the sensitive information while allowing sharing of such a large amount of data.

2.11 Implementation

Conversion of BAM files to pBAM and pBAM+.diff files back to BAM files are implemented as a series of scripts in bash, awk and Python. .diff files are encoded in a compressed format to save disk space. We offer different size .diff files for further compression. similar to the idea of lossy CRAM, we generated lossy .diff files that do not encode for the quality string. For example, a private BAM file that has a size of 12 Gb can be reduced to a 10 Mb lossy .diff file (see Table 1). For convenience, pBAM files are saved as BAM files with manipulated content and with a p.bam extension. That is, any pipeline that uses BAM as an input can take p.bam as an input as well. CPU times (calculated using a single 2.3 GHz AMD Opteron processor) and associated file sizes for alignments from RNA-Seq experiments and ChIP-Seq experiments are documented in Table 1. Our file format manipulation has been adopted by the ENCODE Consortium Data Coordination

Center. Codes for the calculation of information leakage, scripts for file manipulation as well as examples of BAM, pBAM, and .diff files can be found at privaseq3.gersteinlab.org.

Table 1: p-tools performance and associated file sizes

Experiment	Total RNA-Seq	PolyA RNA-Seq	ChIP-Seq (CTCF)	ChIP-Seq (H3K4me1)
BAM size (bytes)	42,833,835,295	18,394,473,842	285,068,500	1,220,248,527
pBAM size (bytes)	7,049,078,124	3,747,104,948	100,690,055	413,710,488
.diff size (bytes)	22,953,435,346	9,470,880,178	76,742,518	393,118,845
lossy .diff size (bytes)	1,576,658,875	667,227,325	5,703,202	10,310,777
BAM to .diff compression	46.4%	48.5%	73.0%	67.8%
BAM to lossy .diff compression	96.3%	96.4%	98.0%	99.15%
BAM to pBAM+.diff CPU time	14:29:44	9:24:38	0:38:17	0:43:09
pBAM+.diff+hg to BAM CPU time	14:20:12	9:21:52	0:29:27	1:25:43

3 Discussion

Functional genomics experiments using large-scale, high-throughput, sequencing-based assays provide a large amount of biological data. Although these experiments aim to answer questions related to genomic activities such as gene expression, TF binding, or the 3D organization of the genome, public sharing of sequencing data from these experiments can lead to recovery of genotype information and, in turn, raise privacy concerns. The systematic quantification of private information content of the functional genomics BAM files and open access to such data without compromising individuals identity have not been well studied. Current policies regarding public sharing of functional genomics BAM files are ad-hoc. The experiments that require a high depth of sequencing such as Hi-C and sometimes RNA-Seq are considered to be private, whereas rela-

tively low-depth BAM files such as those from ChIP-Seq are often shared publicly. In this study, we derived information theory-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 coverage of low-depth experiments such as ChIP-Seq, linking individuals to the databases can be done without error. When we compared the linking accuracy to the false discovery rate, we found that it is easier to link individuals to the databases than genotyping them accurately using functional genomics experiments. The implication is that noisy quasi-identifiers (i.e., low-quality SNP calling) can be used to link the data to the high-quality genotypes. For example, according to our calculations, reads from single-cell RNA-Seq data carry the largest amount of noise. This is likely due to the bias towards expressed genes in such small amounts of cells, mapping issues of splice sites, false positives from RNA editing sites, and amplification bias. However, the noisy genotypes called from a small amount of cells, even when the number of reads is only a million, are quasi-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 decreasing 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.

Another implication of the false discovery rate of genotyping in privacy is the relationship between the accuracy of the genotypes and the amount of information gained from the genotypes. For example, if the query individual is not in the genotype panel, any called genotypes of the query that are not in the panel will be *de novo* variants and will greatly contribute to the information gained. However, these *de novo* variants can be rich in artifacts and sequencing errors. Conversely, any called common genotypes of the query will be highly accurate while poor in information. Con-

sequently, from an adversarys perspective, the most valuable genotypes will be the rare ones in the panel to make accurate inferences about the querys identity and sensitive phenotypes, despite the fact that most information is gained from the *de novo* variants by definition. One way to correct for this is to count any *de novo* variant genotyped as a false discovery, which changes the false discovery rate values in Figure 4 greatly for different functional genomics experiments (see SI Figure 8).

In this manuscript, we also discuss the concept of a set point in determining the data production steps, where sensitive information leakage and utility of the data are balanced (Figure 1). Setting a set point is possible by systematic genotyping and quantification of information. Although it is obvious that any DNA read contains variants, it is not trivial to understand the amount and the quality of sequencing to perform accurate genotyping. Moreover, 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 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, a complete set of variants from HeLas 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 plausible. That is, noisy and incomplete genotyping from partial sequencing experiments can serve as strong quasi-identifiers, which is not straightforward to predict at first. Nevertheless, policies governing the public sharing of HeLa genome versus HeLa functional genomics reads is ad-hoc and contradictory. Therefore, it is essential to quantify the information in samples and determine the set point accurately. Importantly, functional genomics experiments advance our understanding of health and disease by revealing functions of the genome under different conditions. The quantification, analysis, and interpretation of functional genomics data is still an evolving field; hence, extensive public sharing of functional genomics data will accelerate collaborative research and reproducibility by removing

the complexities associated with data accession procedures.

The increasing incentive to share data for the advancement of biomedical research and the corresponding increasing 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 have been proposed [42, 43, 44]. Studies have shown that retrieving summary information from private statistical databases without revealing some amount of an individual's information is impossible [45]. Furthermore, an entire database can be inferred by using a small number of queries. Differential privacy ensures a high level of privacy such that an adversary retrieves a similar result with or without the addition of the individual's data to the database by adding perturbations or noise to the queries [45]. We further studied if the concept of differential privacy can be utilized to create leakage-free raw functional genomics data (see SI Methods). Although such a concept is useful for sharing summary statistics of functional genomics data from multiple individuals, it is conceptually hard to apply to the raw mapped read sharing from functional genomics experiments taken from a single individual. Although further research will be fruitful on how to extract useful information from genomics data that are noisy and perturbed, we envision there will be more applications of privacy concepts like differential privacy in genomics data sharing such as releasing population-based genotype-phenotype data.

To enable public sharing of raw alignments from functional genomics experiments, we designed a privacy-preserving transformation and created privacy-preserving binary alignment files (pBAM). We developed a framework with which 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 functional genomics experiments. To ease the challenges associated with moving and storing large special-access files, we created a lightweight .diff file format that consists of the differences between pBAM and BAM

files in a compact format. This allows us to not repeat the sequence information in the human reference genome files in .diff files and reduces the size of the private files significantly. The presented framework can be used for quantification of sensitive information from the raw reads of functional genomics experiments and conversion of raw files to privacy-preserving file formats. We address the most obvious leakage and provide solutions for quick quantification and safe data sharing. However, it is useful to review all the sources of information leakage from functional genomics experiments. For example, the next source of leakage is from the signal profiles in RNA-Seq, which was addressed elsewhere [21]. There is also leakage from gene expression quantifications, which was shown to be connected with variants through the eQTLs [20]. We also anticipate more leakages to be discovered as new functional genomics experiments are developed. Combined with the increasing attention to genomic privacy, we expect future studies will lead to novel privacy-preserving solutions in an open data-sharing mode.

4 Figure Legends

Figure 1: **Schematic of data types from the progressive summarization of functional genomics experiments.** (**a**) Data summarization flow for RNA-Seq data processing from mapped reads to gene quantifications. Mapped reads to the reference genome were stacked together to create signal profiles. The signal profiles were then averaged over exons to calculate the gene expression quantification. Darker color indicates larger privacy leaks for that particular data type. (**b**) Different layers of data produced from functional genomics experiments. Raw reads from the sequencer that are stored as FASTQ files are mapped to the reference genome. These are the data types that leak the greatest amount of sensitive information, while also possessing the highest utility. Moving upwards through the pyramid levels, there is less privacy concern but largely reduced utility and amount of data. Although modified reads have a large amount of data and utility, their privacy leak can be minimal depending on the amount and the operation of the modification. The purple line denotes the set point, where the privacy and utility trade-off is balanced. (**c**) Quantification of the number of leaked variants and sensitive information leakage in different layers of the data stack of RNA-Seq data processing. The variants in the 1,000 Genomes panel are overlapped with exons to calculate the number of potential leaking variants. The average and maximum information from these variants are quantified. The number of variants that can be genotyped from a typical RNA-Seq experiment is calculated as the number of accessible variants. The number of accessible variants is then multiplied with the average leakage per variant to quantify the total amount of leakage. This procedure is repeated to calculate the leakage from other sources such as signal profiles and gene expression quantifications. Raw reads leak approximately $1,000$ times more information than signal profiles; the amount of leakage from signal profiles and gene expression quantifications are comparable.

Figure 2: **Comparison of naive information measure with information with LD consideration and sample size correction.** (**a**) The process of sampling reads from functional genomics experiments for the calculation of pmi between 1,000 Genomes gold standard variants and called variants from functional genomics reads at different coverage. c amount of reads were sampled and genotyped using GATK. These genotypes were then compared against the gold standard 1,000 Genomes genotypes. The Venn diagrams show the relationship between different information measures used to quantify the leakage. The self-information of the gold standard genotypes that are called from the functional genomics reads (true positives) is the pmi. The genotypes that are called from functional genomics reads but are not in the gold standard variants are false positives. The genotypes that are in the gold standard but are not called from the functional genomic reads are false negatives. (**b**) The variants are grouped based on their self-information. For each group, the maximum LD (mLD) score was averaged. These averaged mLD scores are plotted against the self-information calculated using a naive approach (blue dots) and the self-information with LD consideration (pink dots). The average mLD score for high self-information variants is small, whereas low self-information variants have a larger mLD correlation. As the overall information quantification is dominated by high self-information variants, consideration of LD in the information quantification makes a negligible difference. (**c**) The individuals from the 1,000 Genomes panel sampled in increasing fractions from 10% to 100%. Genotyping frequencies were calculated for each sampled sub-population and information calculated based on those frequencies using a naive approach. The naive information vs. $1/\text{sampled}-\text{fraction}$ (e.g., 1/10 for 10%) is plotted. The y-intercept is extrapolated from the fitted curve and denotes the information when the population size is infinite (i.e., 1/inf=0). Error bars were calculated using 100 times bootstrapping. (**d**) Difference between the naive information, information with LD consideration, and extrapolated information when population size is infinite. There is negligible difference between the overall information quantification when comparing the three approaches.

Figure 3: **The pmi calculated for 24 different functional genomics assays and WGS, WES, and SNP ChIP data using NA12878 1,000 Genomes variants as the gold standard.** (**a**) The *pmi* values for WGS and three different primary Hi-C experiments plotted at different coverage. The information contents of the gold standard and SNP ChIP are included for comparison. The genotyping from Hi-C experiments compared to the gold standard is almost the same as the genotyping from WGS. (**b**) The *pmi* values for WES, total RNA-Seq, polyA RNA-Seq, and single-cell RNA-Seq of two different cells plotted at different coverage. The *pmi* for the RNA-Seq experiments never reach the level of the *pmi* for WES. WES and total RNA-Seq cannot capture variants as well as Hi-C experiments. (**c**) The *pmi* values for 20 different ChIP-Seq experiments targeting histone modifications and TF binding plotted at different coverage. The *pmi* for some of the ChIP-Seq experiments is larger than the *pmi* for WES and RNA-Seq experiments even at smaller coverage. (**d**) The *pmi* per base pair for different functional genomics experiments. The overall *pmi* values are normalized by the amount of coverage to quantify the amount of information gained from genotyping one base pair in any functional genomics experiment. The dashed red line is the *pmi* per base pair of the WGS for an easier comparison.

Figure 4: **False discovery rate of functional genomics experiments at different coverage** (**a**) FDR comparison for Hi-C and WGS data at different coverage. As the amount of coverage increases, the FDR decreases. Overall, variants from Hi-C consistently have lower FDR than the variants from WGS until the coverage reaches its maximum. (**b**) FDR comparison for WES and different RNA-Seq experiments at different coverage. In general, there is a decreasing FDR trend with increasing coverage, as seen in panel (a), except for single-cell RNA-Seq. The noise increased for single-cell RNA-Seq experiments as more reads were included. (**c**) FDR comparison for different ChIP-Seq experiments at different coverage. There was a general trend of decreasing FDR with increasing coverage.

Figure 5: **Illustration of a linking attack and the accuracy of linking.** (**a**) The publicly available anonymized reads from functional genomics experiments contain a set of variants and the HIV status for the sample that the functional genomics experiment was performed. The panel of genotypes contains the variants and associated genotypes for m individuals. The attacker links the called genotypes from the functional genomics experiment to the panel of genotypes by using the best matched $pmi(S^F; S_j^{DB})$. The linking potentially reveals the HIV status for the linked individual. (**b**) The measure gap is used to quantify the accuracy of linking. Comparison of gap for NA12878 at different coverage for Hi-C and WGS. NA12878 is more vulnerable to re-identification through linking at low coverage using the Hi-C reads than the WGS reads. (**c**) Comparison of gap for NA12878 at different coverage for different RNA-Seq experiments and WES. Although, single-cell RNA-Seq variants have low pmi and large FDR, NA12878 is still vulnerable to re-identification through linking at any coverage. (**d**) Comparison of gap for NA12878 at different coverage for ChIP-Seq experiments. Some of the TF binding ChIP-Seq variants provide re-identification as accurately at Hi-C reads, and more accurately than WES reads at lower coverage.

Figure 6: **An individual's full variant list can be assembled and sensitive phenotypes can be inferred from publicly available data using imputation, and a theoretical framework for prediction of amount of leaked data** (**a**) Comparison of the number of SNVs called from WGS data and from the combination of all of the ChIP-Seq and RNA-Seq data with and without imputation. The total set of variants of NA12878 can be assembled by combining all the available ChIP-Seq and RNA-Seq reads together. (**b**) The variants associated with physical traits and whether they are present in the called variants from different functional genomics experiments before and after imputation. Hi-C and total RNA-Seq reads largely reveal sensitive phenotypes even before imputation. (**c**) The schematic of the features that are used for the regression learner. The first feature is the average depth of the sequencing experiment. The number of reads representing each base pair is calculated as the depth of that base pair. The second feature is the breadth of the sequencing experiment, which is the number of base pairs that are sequenced at least once (d=1). The third feature is the bias of the sequencing experiment. This is the deviation from the WGS data. We first calculated the depth distribution of the reads from functional genomics experiment. We then assumed a Poisson distribution with the same mean depth and calculated the difference between the assumed Poisson and the calculated empirical distribution. (**d**) After using the features explained in panel (c), we fit a regression model to predict the normalized pmi. This panel shows the prediction accuracy over 40 randomly drawn data points used as training. Blue data points are from ChIP-Seq, pink data points are from Hi-C, and the red data point is from WGS data. The root mean square error for this prediction is 2.60 for the true values ranging from 0 and 100. The R^2 for the correlation between predicted and true values is 0.9. (**e**) The prediction accuracy over five randomly drawn data points used to make blind predictions by using the model fitted in panel (d). Blue data points are from ChIP-Seq and pink data points are from Hi-C data. The root mean square error for this prediction is 2.47 for the true values ranging from 0 and 100. The R^2 for the correlation between predicted and true values is 0.9.

Figure 7: **Removal of rare variants and linking** (**a**) Self-information of NA12878 gold standard variants vs. genotyping frequency of these variants. The pink color code corresponds to the number of variants with a given self-information and genotyping frequency observed in the NA12878 gold standard set. We divided these variants into three categories: (1) Singleton genotypes, which were only observed in NA12878 variants, (2) doubleton genotypes, which were observed in at most one additional individual in the 1,000 Genomes panel, (3) rare genotypes, which were observed in 1.5% of the 1,000 Genomes individuals. (**b**) The singleton, doubleton, and rare genotypes were iteratively removed from the NA12878 variant set and a linking attack for each iteration was performed. *gap* was still large and NA12878 was still vulnerable to re-identification even after removal of all rare genotypes from the variant set. (**c**) The *pmi* of all the variants called from total RNA-Seq reads vs. the *pmi* of the indels called from total RNA-Seq reads. Indels reveal much less information than SNVs. (**d**) Two linking attacks were performed by using all the variants called from total RNA-Seq and by using only indels called from total RNA-Seq. A *gap* comparison between these two attacks shows that the individual is vulnerable to re-identification even when only indels are used for linking.

Figure 8: **Privacy-preserving file formats for alignment files from functional genomics experiments.** (**a**) The schematic of the privacy-preserving transformation of an alignment file, the difference between the signal calculated from original BAM and transformed pBAM files, and the concept of ε for the error. (**b**) The difference between the BAM and pBAM files is shown at the read level, when the reads were mapped to a reference genome and reference transcriptome. The added noise to the depth signals due to the BAM to pBAM transformation is shown in an example read with a deletion, insertion, and mismatch. The difference of the depth signal when calculated from BAM and pBAM is shown using different operations Q and r values for the polyA RNA-Seq experiment of individual NA12878. The depth signal was compared between different file formats at the single base pair, exon, and gene resolution. (**c**) The change in ε with increasing number of manipulated reads (r) for different operations Q. When Q is the removal of mismatches, no noise is added to the depth signal. However, when the Q is the removal of indels, the ε increases with the increasing number of manipulated reads. (**d**) The schematic of how p-tools work with different file formats. We focused on the BAM to pBAM transformation in this study, however p-tools can be adopted for transformation between CRAM and pCRAM (shaded schematic). Arrows with solid lines demonstrate conversions that are currently within the scope of p-tools; arrows with dashed lines indicate the conversion will require additional tools.

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

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

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

[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. Simple demographics often identify people uniquely. *Carnegie Mellon University, unpublished*, 2000.

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

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

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

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

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

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

[16] Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, 2011;21(5):734-740.

[17] Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 2014;159(7):1665-1680.

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

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

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

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

[22] Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M. RSE-Qtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, 2011;27(2):281-283.

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

[24] 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.

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

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

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

[28] 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.

[29] 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.

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

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

[32] Cover TM and Thomas JA Elements of information theory. *John Wiley & Sons*, 2012.

[33] 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.

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

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

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

[37] Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. *MIT Press*, 2006.

[38] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak M, Gaffney D, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biol.*, 2016;17:13-14.

[39] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 2008;9(9):R137.

[40] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 2013;29(1):15-21.

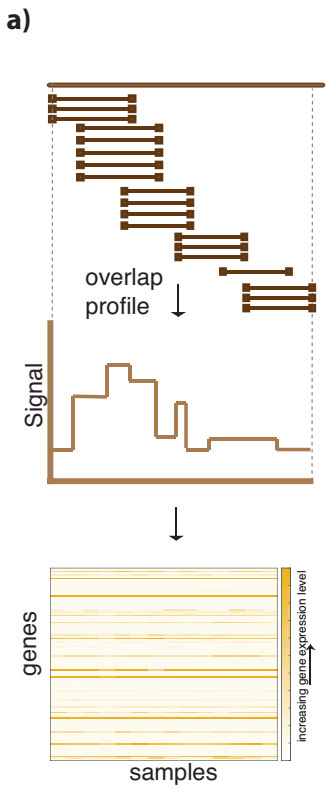
[41] Li B and Dewey C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 2011;12:323.

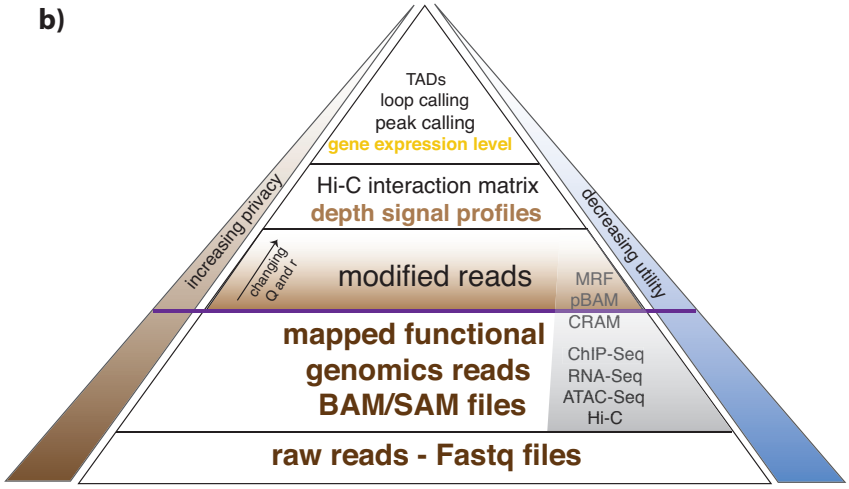
[42] Fienberg S, Slavkovic A, Uhler C Privacy preserving GWAS data sharing. *In ICDM*, 2011:628635.

[43] Johnson A and Shmatikov V. Privacy-preserving data exploration in genome-wide association studies. *In KDD*, 2013:1079-1087.

[44] Yu F, Fienberg SE, Slavkovic AB, Uhler C. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 2014;50:133-141.

[45] Dwork C. Differential Privacy: A Survey of Results. *Springer Berlin Heidelberg*, 2008;Theory and Applications of Models of Computation. Lecture Notes in Computer Science. pp. 1-19





c)

Leakage Source	Leaking Variants	# of potential variants	Average leakage per variant (bits)	Maximum leakage per variant (bits)	# of accessible variants	Total leakage (bits)
Raw reads	Exonic variants	2,682,417	0.10 ± 0.28	9.88 ± 2.12	246,893	24,689
Modified reads Q = {indels}	Exonic SNVs	2,607,969	0.09 ± 0.27	9.95 ± 2.02	231,031	207,92
Modified reads Q = {mismatches}	Exonic indels	51,408	0.33 ± 0.47	7.64 ± 2.42	15,862	5234
Signal profiles	Exonic deletions	48,019	0.29 ± 0.45	7.97 ± 2.42	1,067	298
Gene expression quantification	eQTLs	3,175	1.19 ± 0.36	4.00 ± 1.92	158	188

Figure 1

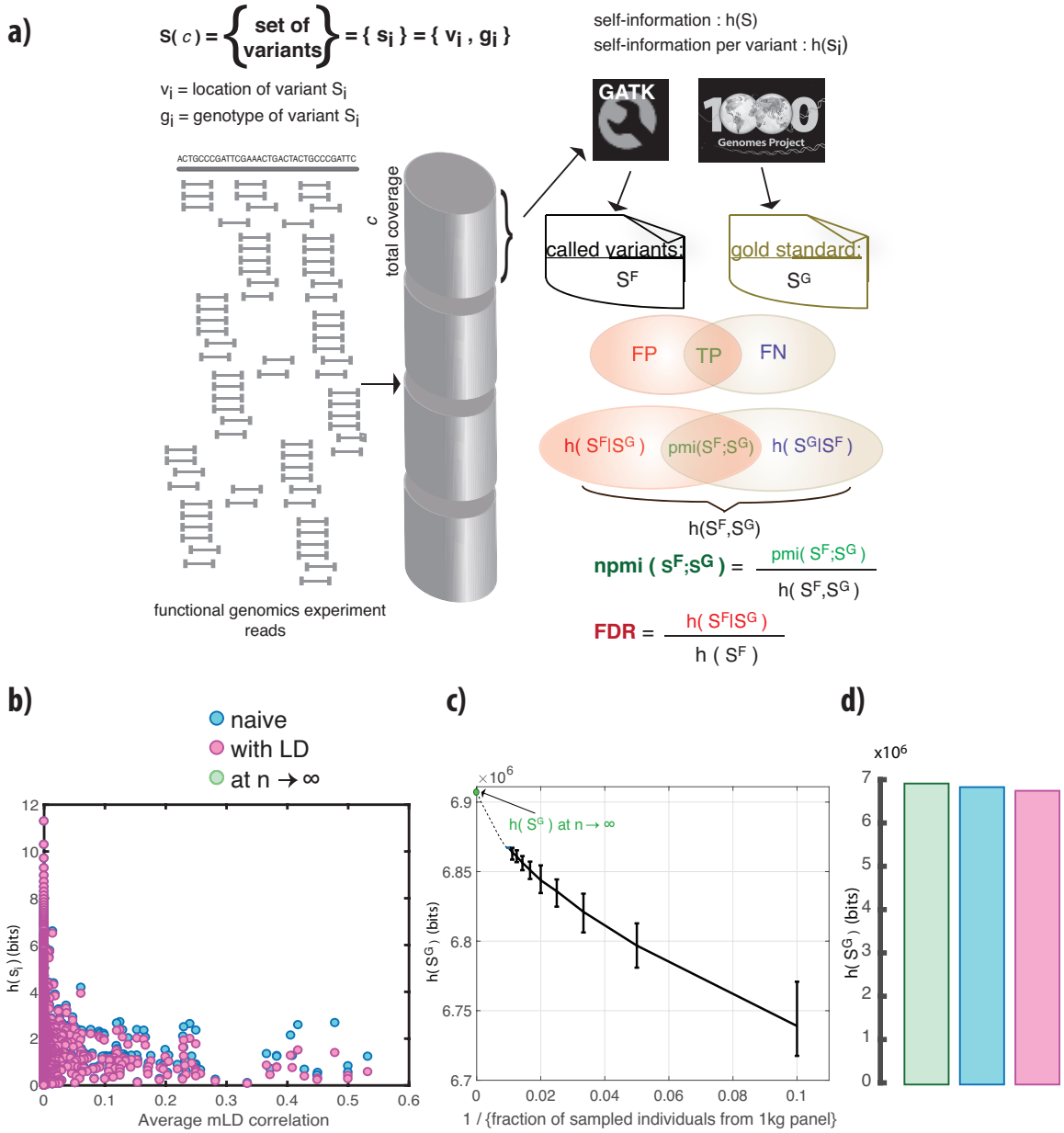

a)
S(c) = { set of variants } = { s_i } = { v_i , g_i }
v_i = location of variant S_i
g_i = genotype of variant S_i
self-information : h(S)
self-information per variant : h(s_i)
GATK
1000 Genomes Project
ACTGCCCGATTCGAAACTGACTACTGCCCGATTC
c
total coverage
called variants: S^F
gold standard: S^G
FP
TP
FN
h(S^F|S^G)
pmi(S^F;S^G)
h(S^G|S^F)
h(S^F,S^G)
npmi (S^F;S^G) = pmi(S^F;S^G) / h(S^F,S^G)
FDR = h(S^F|S^G) / h (S^F)
functional genomics experiment reads
b)
naive
with LD
at n → ∞
h(s_i) (bits)
Average mLD correlation
c)
×10^6
h(S^G) at n → ∞
h(S^G) (bits)
1 / {fraction of sampled individuals from 1kg panel}
d)
x10^6
h(S^G) (bits)

Figure 2

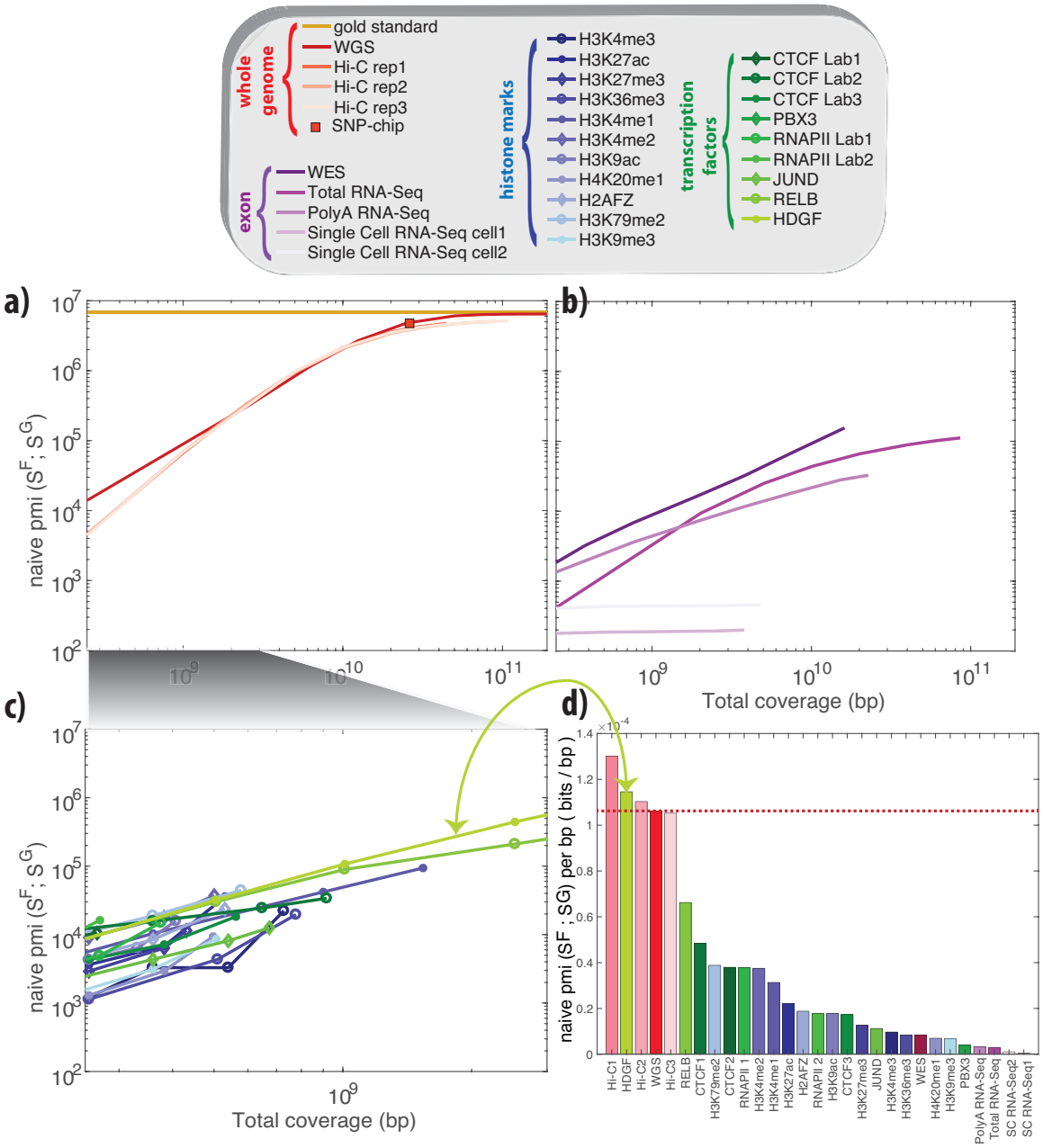

whole genome
gold standard
WGS
Hi-C rep1
Hi-C rep2
Hi-C rep3
SNP-chip
exon
WES
Total RNA-Seq
PolyA RNA-Seq
Single Cell RNA-Seq cell1
Single Cell RNA-Seq cell2
histone marks
H3K4me3
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me2
H3K9ac
H4K20me1
H2AFZ
H3K79me2
H3K9me3
transcription factors
CTCF Lab1
CTCF Lab2
CTCF Lab3
PBX3
RNAPII Lab1
RNAPII Lab2
JUND
RELB
HDGF
a)
b)
c)
d)
naive pmi (S^F ; S^G)
Total coverage (bp)
naive pmi (SF ; SG) per bp (bits / bp)
Hi-C1
HDGF
Hi-C2
WGS
Hi-C3
RELB
CTCF1
H3K79me2
CTCF2
RNAPII 1
H3K4me2
H3K4me1
H3K27ac
H2AFZ
RNAPII 2
H3K9ac
CTCF3
H3K27me3
JUND
H3K4me3
H3K36me3
WES
H4K20me1
H3K9me3
PBX3
PolyA RNA-Seq
Total RNA-Seq
SC RNA-Seq2
SC RNA-Seq1

Figure 3

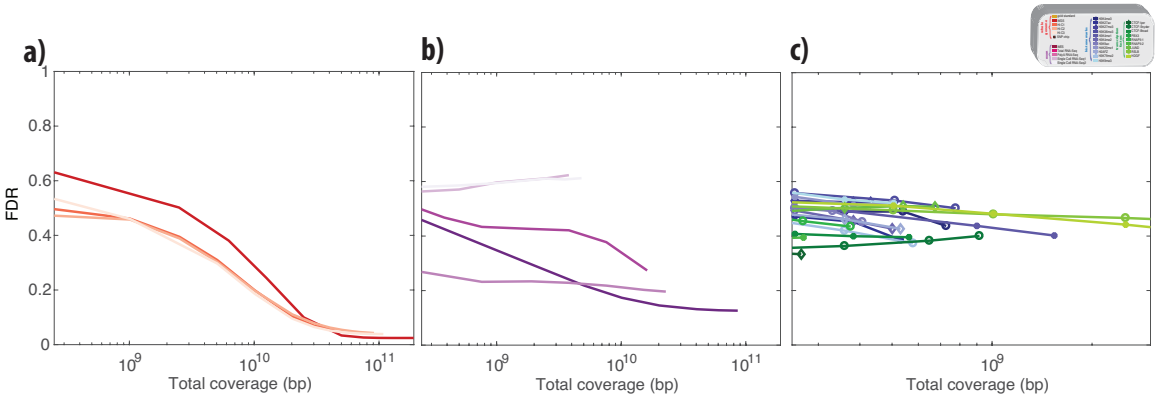


Figure 4

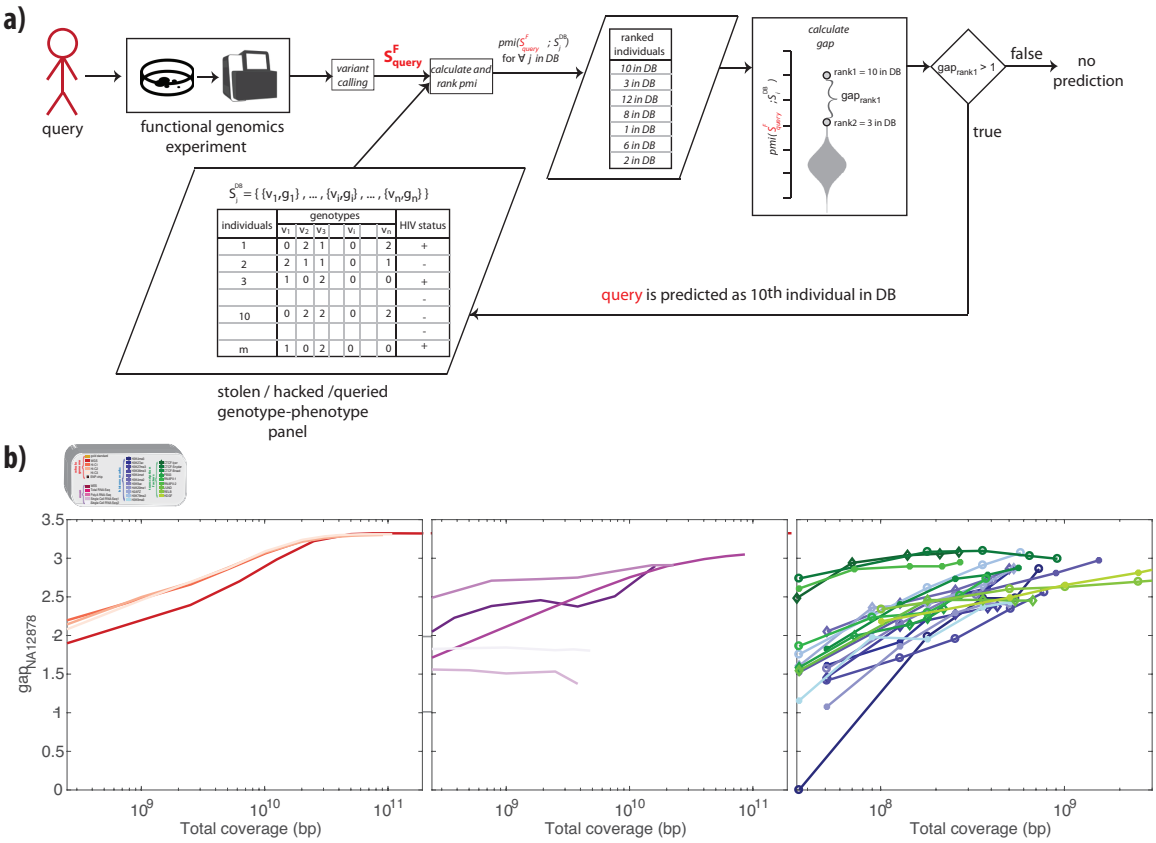

a)
query
functional genomics experiment
variant calling
calculate and rank pmi
for ∀ j in DB
ranked individuals
10 in DB
3 in DB
12 in DB
8 in DB
1 in DB
6 in DB
2 in DB
calculate gap
rank1 = 10 in DB
rank2 = 3 in DB
false
no prediction
true
query is predicted as 10th individual in DB
individuals
genotypes
HIV status
stolen / hacked /queried genotype-phenotype panel
b)
Total coverage (bp)
Total coverage (bp)
Total coverage (bp)

Figure 5

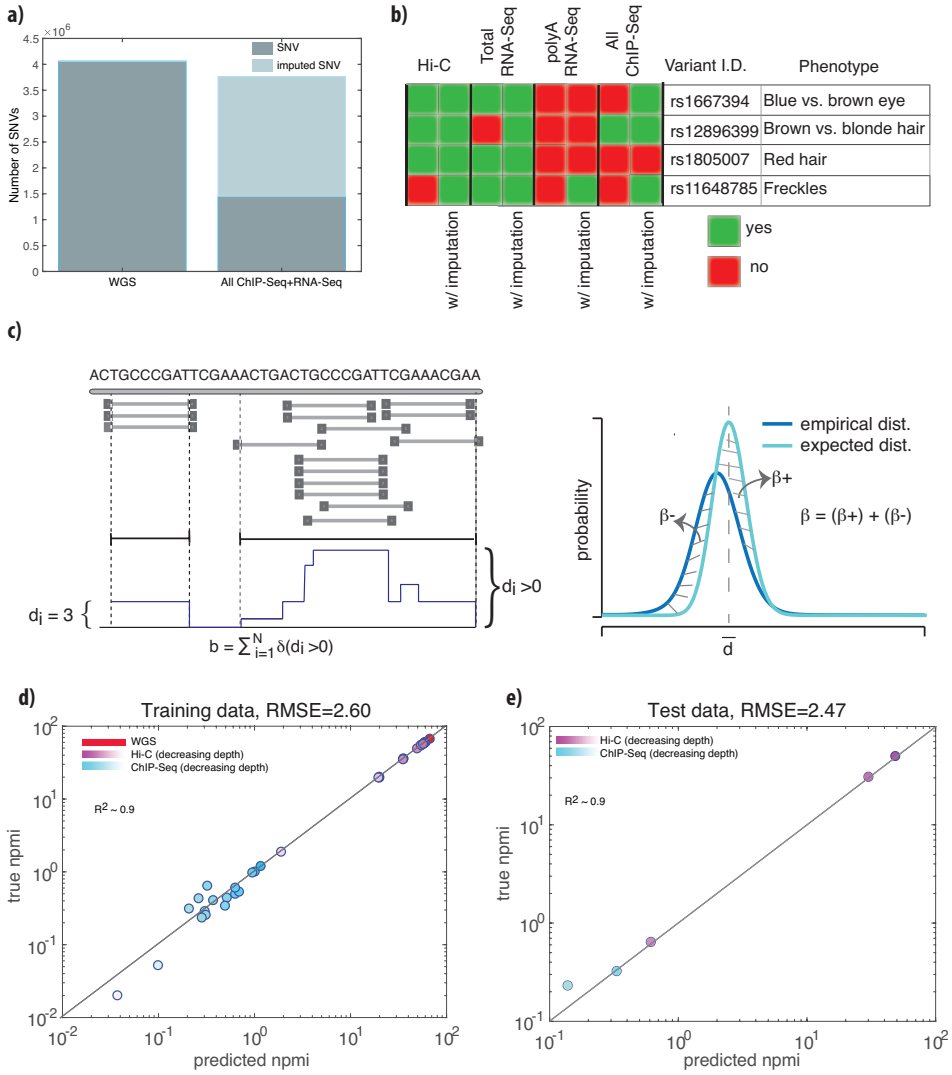

a)
Number of SNVs
SNV
imputed SNV
WGS
All ChIP-Seq+RNA-Seq
b)
Hi-C
Total RNA-Seq
polyA RNA-Seq
All ChIP-Seq
Variant I.D.
Phenotype
rs1667394 Blue vs. brown eye
rs12896399 Brown vs. blonde hair
rs1805007 Red hair
rs11648785 Freckles
w/ imputation
yes
no
c)
ACTGCCCGATTCGAAACTGACTGCCCGATTCGAAACGAA
$d_i = 3$
$d_i > 0$
$b = \sum_{i=1}^{N} \delta(d_i > 0)$
empirical dist.
expected dist.
probability
β+
β-
β = (β+) + (β-)
d
d)
Training data, RMSE=2.60
WGS
Hi-C (decreasing depth)
ChIP-Seq (decreasing depth)
$R^2 \sim 0.9$
true npmi
predicted npmi
e)
Test data, RMSE=2.47
Hi-C (decreasing depth)
ChIP-Seq (decreasing depth)
$R^2 \sim 0.9$
true npmi
predicted npmi

Figure 6

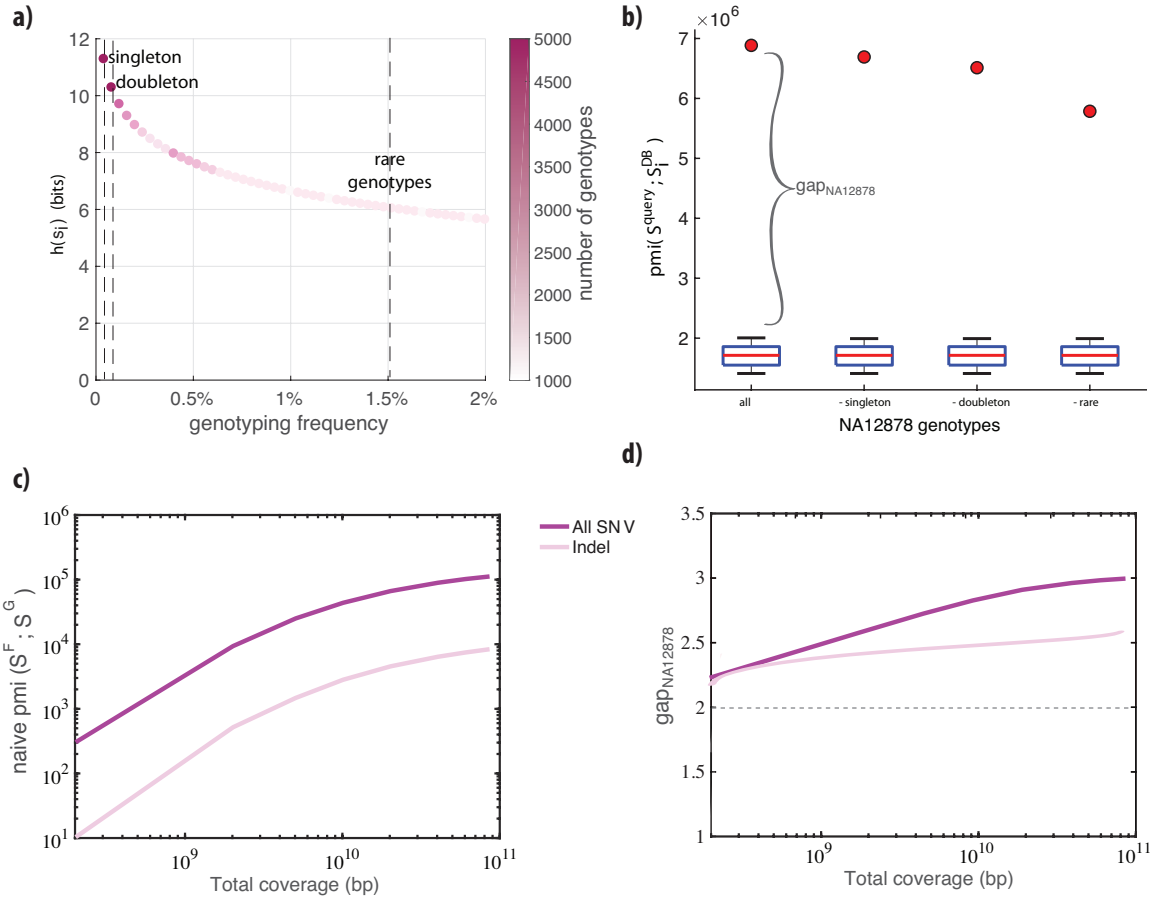

a)
singleton
doubleton
rare
genotypes
h(s_i) (bits)
genotyping frequency
number of genotypes
b)
×10^6
pmi(S^query ; S_i^DB)
gap_NA12878
all
- singleton
- doubleton
- rare
NA12878 genotypes
c)
naive pmi (S^F ; S^G)
Total coverage (bp)
All SNV
Indel
d)
gap_NA12878
Total coverage (bp)

Figure 7

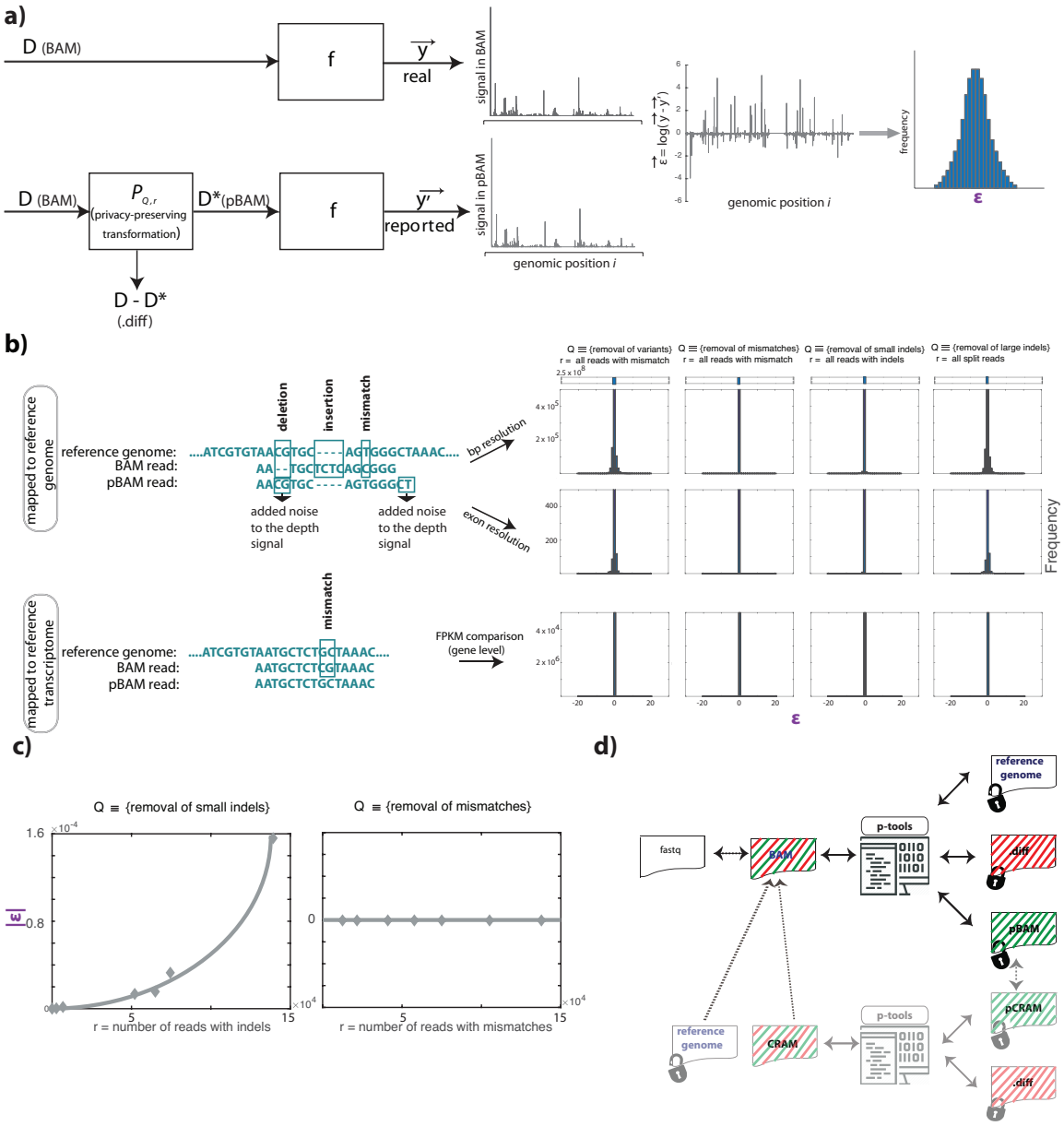

a)
D (BAM)
f
real
signal in BAM
D (BAM)
(privacy-preserving transformation)
D*(pBAM)
reported
signal in pBAM
genomic position i
D - D*
(.diff)
genomic position i
frequency
ε
b)
mapped to reference genome
reference genome:
BAM read:
pBAM read:
deletion
insertion
mismatch
added noise to the depth signal
bp resolution
exon resolution
Q ≡ {removal of variants}
Q ≡ {removal of mismatches}
Q ≡ {removal of small indels}
Q ≡ {removal of large indels}
Frequency
mapped to reference transcriptome
FPKM comparison (gene level)
c)
Q ≡ {removal of small indels}
Q ≡ {removal of mismatches}
r = number of reads with indels
r = number of reads with mismatches
d)
reference genome
fastq
BAM
p-tools
.diff
pBAM
pCRAM
CRAM

Figure 8