Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions

Arif Harmanci[1,2,*], Mark Gerstein[1,2,3,*]

1 Program in Computational Biology and Bioinformatics, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
2 Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
3 Department of Computer Science, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
*Corresponding authors: Arif Harmanci (arif.harmanci@yale.edu), Mark Gerstein (pi@gersteinlab.org)

Abstract

Functional genomics data such as RNA-sequencing are performed to reveal the how gene expression changes in different conditions. Although the main purpose of these data is understanding the dynamic changes in gene expression level, the data also contain a large number of genetic variants in the raw reads. While it is known that raw reads cannot be shared because of privacy concerns, it is generally assumed safe to share other data aggregated representations of functional genomics data, such as read depth signal profiles and quantification of gene expression levels. Here, we focus on the privacy aspects of genome-wide signal profiles of functional genomics experiments, which represent measurement of activity at each genomic position. We show that the signal profiles, which are often publicly shared, can be used to genotype small and large deletions, which can then be used for breaching privacy. We first present measures for predictability of genotypes from signal profiles and information leakage from signal profiles. We then present practical methods for detecting and genotyping large and small genomic deletions, and demonstrate that the genotyped deletions can accurately identify an individual from a large sample. We also present an effective anonymization procedure for protection of signal profiles against genotype prediction based attacks. Given that several consortia, for example GTex and ENCODE, publicly share signal profiles, these results point to one of the critical sources of sensitive information leakage.

1. Introduction

Individual privacy is emerging as an important aspect of biomedical data science[1]. A deluge of genetic data is being generated with the Cancer Moonshot Project[2], Precision Medicine Initiative[3, 4], and the 100K Genome Project [5, 6] from hundreds of thousands, if not millions, of individuals. Moreover, there is much effort to make genetic data more prevalent in the clinical setting[7]. The leakage of genetic information creates many privacy concerns, e.g. genetic predisposition to diseases may bias insurance companies[8].

The initial studies on genomic privacy have focused primarily on protecting the identities of participants in the early genetic profiling and genotype-phenotype association studies[9, 10]. These focused on whether an individual's genetic information can be used to reliably predict whether they participated in a certain cohort of individuals in a genetic study. We refer to these scenarios as detection of a genome in a

Deleted: is emerging as a valuable resource for personalized medicine. Although one might think that these data are safe to share, the extent to which they leak sensitive information is not well studied.

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Deleted: cause concerns

Formatted: Highlight

Formatted: Highlight

Deleted: of

Formatted: Highlight

Formatted: Highlight

Deleted: the

Formatted: Highlight

Deleted: of several sequencing based functional assays including RNA-seq and ChIP-Seq

Formatted: Highlight

Formatted: Highlight

Deleted: TCGA

Formatted: Highlight

Deleted: for personal functional genomics data

Formatted: Highlight

Deleted: a

Formatted: Highlight

Deleted: source

Formatted: Highlight

Deleted: , which can be potentially protected by our anonymization technique

Formatted: Highlight

Deleted: [1]. A deluge of genetic data is being generated with the Cancer Moonshot Project[2], Precision Medicine Initiative[3, 4], and UK100K[5, 6] from hundreds of thousands, if not millions, of individuals. Moreover, there is much effort to make genetic data more prevalent in the clinical setting[7]. The leakage of genetic information creates many privacy concerns, e.g. genetic predisposition to diseases may bias insurance companies[8]

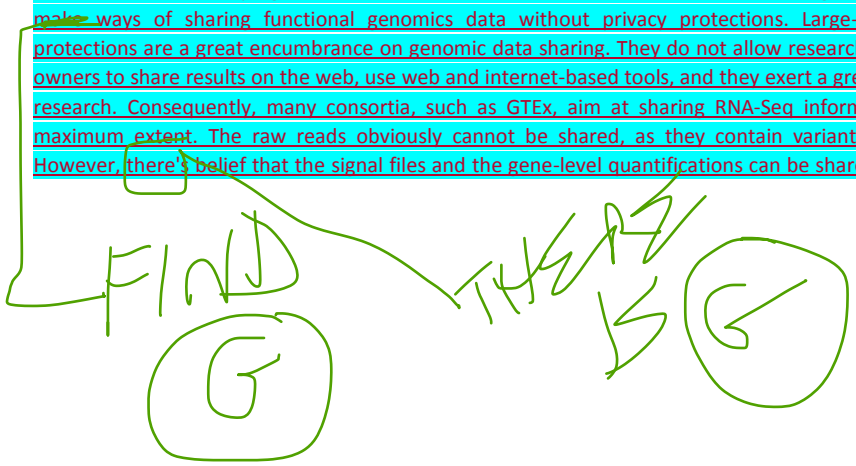
Deleted: The initial studies on genomic privacy have focused primarily on protecting the identities of participants in the early genetic profiling and genotype-phenotype association studies[9, 10]. These focused on whether an individual's genetic information can be used to reliably ...

mixture. In this arena, the differential privacy[11] has been proposed as a theoretically optimal formalism that can fulfill the privacy requirements such that the probability that any individual's participation can be identified made arbitrarily small. In addition, the cryptographic approaches have proven useful for privacy-aware analysis of genomic datasets albeit with high requirements of computational resources[12, 13].

The decrease in cost of DNA sequencing technologies has substantially increased the number and size of available genomic data and has made genomic data much more practically available to hospitals, research institutes, and to individuals[14]. This increase will render new types of attacks much more practical where an adversary can use statistical methods to link multiple datasets to reveal sensitive information. These attacks are termed the linking attacks[15–17]. In a nutshell, the linking attacks are based on cross-referencing and matching of two or more datasets that are released independently. Some of the datasets contain personal identifying information, e.g. names or addresses, while others contain sensitive information, e.g. health information. The immediate consequence of the cross-referencing is that the sensitive information in one dataset gets linked to the identifying information in another, which in turn breaches privacy of individuals whose sensitive information are revealed. The risks behind linking attacks have risen recently because the personal data is generated at exceedingly high pace and these data are independently released and maintained. For this reason, a rather challenging aspect of the linking attacks is that risk assessment is complicated because one dataset that is currently deemed safe to release may become a target for linking attacks when another dataset is released in the future, i.e., a dataset that seems safe to release now may become vulnerable to a linking attack next year.

A well-known example of linking attacks is the Netflix Prize Competition[15] (Supplementary Fig 1a,b). In this competition, a training dataset was released by the movie rental company Netflix, which was to be used for training new automated movie rating algorithms. The dataset was anonymized by removing names. Although the dataset seemed safe to share at the time, two researchers showed that this training dataset can be linked to a seemingly independent database of the Internet Movie Database (IMDb). The linking revealed movie preferences and identities of many Netflix users. We believe that similar scenarios will be a major route to breaches in individual genomic privacy and these must be studied well to enable privacy-aware data sharing approaches.

In this study, we analyze the leakage of sensitive information from the functional genomics data and how they can be used by an adversary in linking attacks. There are a number of motivating key points related to functional genomics data and privacy. First of all, functional genomics data, such as RNA sequencing data, is unique, in that if the data comes from human subjects, the raw reads have genetic variant information, which may be used to identify individuals. However, the main purpose of RNA-Seq data is not related to the variants; main purpose is more related to understanding dysregulation of genes under different conditions, such as cancer. Consequently, there's a great desire to share and study RNA-Seq datasets, to enable helping to find cures for various diseases. Because of this, there is great incentive to make ways of sharing functional genomics data without privacy protections. Large-scale privacy protections are a great encumbrance on genomic data sharing. They do not allow researchers and data owners to share results on the web, use web and internet-based tools, and they exert a great burden on research. Consequently, many consortia, such as GTEx, aim at sharing RNA-Seq information to the maximum extent. The raw reads obviously cannot be shared, as they contain variant information. However, there's belief that the signal files and the gene-level quantifications can be shared. The signal



Deleted: Several studies have addressed aspects of linking attacks in the genomic privacy context[16–19]. Still, there are two major aspects of genomic privacy that are not well addressed in the previous studies in the context of linking attacks. Firstly, the structural variants (SVs), which comprise deletion, insertion, translocation, and transversion of large chunks of DNA sequence, do not receive much attention in the debate of genomic privacy[20]. Rather most of the focus is on the single nucleotide polymorphisms (SNPs). This is well justified because the estimated regulatory effect of SNPs on gene expression is much larger than structural variants[21]. On the other hand, it is known that the major portion of the genomic variation, in terms of the number of nucleotides that are affected, is caused by SVs[22, 23]. Since an SV affects a much larger portion of the genome (in number of nucleotides) than a single nucleotide variant does, its effect on an affected phenotype is expected to be very obvious, if not more

files simply reflect the overall depth of coverage of the RNA-Seq reads at any given point. Ostensibly, they're do not contain variant information. Many of the genomics consortia have decided to openly share RNA-Seq signal information. We show that there is a high degree of private information leakage in the function genomics signal profile data. The gene-level quantifications essentially are averages over the signal profile over exons. Although the overall averaging reduces information, private information leakage. However, there is also private information leakage through the association with variants called eQTLs. It is important to note that this is tackled in the current study, but is looked at elsewhere[16, 18].

In fact, other functional genomics datasets, like ChIP-Seq[19], are rich sources of information and are used extensively in probing gene regulation under different conditions. In the general discussion of privacy, one considers the DNA variants and simply protecting them from an adversary. Once the adversary gets the variants, privacy is breached. Although there is some variant information in the functional genomics datasets, this is often not the main purpose of them. The dataset is collected to make a general statement about how gene regulation and expression relates to some biological phenotype, for example identification of the set of genes that are upregulated in cancer. Thus, unlike the variant data, functional genomics datasets have a more complicated, "Yin-Yang", aspect with relation to privacy.

Furthermore, sometimes the functional genomics datasets are shared with phenotypic information that is potentially of private value, e.g., a particular condition or disease that a person has. This leads to an interesting situation where the data is ostensibly collected and used for non-personal purposes to find general aspects about a condition. But the existence of small amounts of residual private information in the data can potentially be revealing about the individual which they came from. Hence this leads to the complex aspect that we will be probing here.

Several studies have addressed aspects of linking attacks in the genomic privacy[16–18, 20]. One aspect of genomic privacy that is not well addressed in the previous studies is how the structural variants, which comprise deletion, insertion, translocation of large chunks of DNA sequence, much attention in the debate of genomic privacy[21]. In this work, we are studying whether an adversary can use small and large genomic deletions for performing linking attacks. We study whether the adversary can use signal profiles of functional genomics signals to detect and genotype genomic deletions and use them to pinpoint individuals in a large genotype dataset. Most of the previous studies on genomic privacy focus on the single nucleotide polymorphisms (SNPs). This is well justified because the estimated regulatory effect of SNPs on gene expression is much larger than the structural variants[22]. On the other hand, it is known that the major portion of the genomic variation, in terms of the number of nucleotides that are affected, is caused by SVs[23, 24] as shown by 1000 Genomes Project. Since an SV affects a much larger portion of the genome (in number of nucleotides) than a single nucleotide variant does, its effect on a phenotype is expected to be very obvious, if not more than a SNP. For example, homozygous deletion of a gene will cause the total disappearance of its expression.

In this study, we analyze the sensitive information leakage from the signal profiles of several sequencing based functional genomics datasets. By signal profile, we refer to the genome-wide signal computed by counting the number of reads that overlap with each nucleotide on the genome. The signal profiles are just one type of aggregated data that is generated from raw reads. Another type of aggregated data is gene expression quantifications, which are averages of RNA-seq signal profiles over genes. The leakage of information from the gene expression quantifications has been previously studied[16, 18]. Rather, we are

Moved down [1]: . For example, homozygous deletion of a gene will cause the total disappearance of its expression.

Deleted: Putting these together, it is necessary to include SVs in the analysis of sensitive information leakage. ¶ Secondly, functional genomics data is not in center of most studies. Nevertheless newer functional genomics datasets, like RNA-Seq[24] and ChIP-Seq[25] are very rich information sources and can lead to leakage of individual characterizing information. In more general genomic privacy context, one considers the DNA variants and simply protecting them from an adversary. Once the adversary gets the variants, privacy is breached and there is no more discussion. Functional genomics datasets however have a more complicated, "Yin-Yang", aspect with relation to privacy. Although there is some variant information in them, this is often not the main purpose of the data set. Rather the point of collecting the dataset is to make a general statement about the relation of the data and some biological phenotype, for example which genes go up in cancer and which genes go up with AIDS.¶

Deleted:

Deleted: being

Deleted: from

Moved (insertion) [1]

Deleted: In this study, we analyze the sensitive individual characterizing information leakage from the signal profiles of several sequencing based functional genomics datasets. By signal profile, we refer to the signal generated by counting the number of reads that overlap with each nucleotide on the genome. The raw reads from a sequencing experiment contain the nucleotides themselves. Therefore raw reads contain a significant amount of sensitive genetic information. The raw reads can be used to identify the private SNPs, small indels, and structural variants. It is well established that the raw reads must not be released publicly (Supplementary Figure 6). Therefore, in this study, we assume that the raw reads are not publicly shared. We assume that the adversary does not have access to the raw reads and only has access to the signal profiles. Regarding the signal profiles, it is generally assumed that the signal profiles are mostly void of sensitive information. In fact, several large consortia, for example ENCODE Project[26], Roadmap Epigenome Mapping Consortium[27], and GTex[28, 29], publicly share signal profiles (Supplementary Figure 5). Although the signal profiles do not contain any explicit sequence information, signal processing techniques can be utilized to detect the large and small structural variants. There are two quantities that determine how well structural variants can be detected from sequencing data. First is breadth of coverage, which measures how well the genome is covered by signal profiles. Second is depth coverage that measures how deep the sequencing is performed. DNA-sequencing read depth signal[30, 31] is very suitable for detection of structural variants because it attempts to uniformly cover the genor ...

only considering whether the signal profiles have any genotypic information leakage from them. We show that signal profiles do leak a large amount of genotype information for small and large genomic deletions.

As discussed earlier, the raw reads from an RNA sequencing experiment contain the nucleotides themselves. It is well established that the raw reads must not be released publicly (Supplementary Figure 6) because given the raw reads, and adversary can identify a large number of private SNPs and indels. We therefore assume that the raw reads are not publicly shared and that the adversary does not have access to the raw reads. Rather, we assume that the data owners created the signal profiles and made these publicly available. The adversary gains access to these signal profiles. Regarding the signal profiles, it is generally assumed that the signal profiles are mostly void of sensitive information. Several large consortia, for example ENCODE Project[25], Roadmap Epigenome Mapping Consortium[26], and GTex[27, 28] publicly share signal profiles (Supplementary Figure 5). Although the signal profiles do not contain any explicit variant information at the nucleotide level (such as SNPs), signal processing techniques can be utilized to detect the large and small genomic deletions. There are two quantities that determine how well genomic deletions can be detected from sequencing data. First is breadth of coverage, which measures how well the genome is covered by signal profiles. Second is depth coverage that measures how deep the sequencing is performed. DNA-sequencing read depth signal[29, 30] is very suitable for detection of deletions because it attempts to uniformly cover the genome (high breadth coverage) in a deep manner (high depth coverage). On the other hand, detection of genomic deletions from functional genomics datasets is not as straightforward. The main reason for this is the dynamic and non-uniform nature of the signal profiles of functional genomics experiments. For example, RNA-seq[31] signal profiles are concentrated mainly on the exonic regions and promoters of the genome, respectively. In other words, RNA-seq signal profiles generally have high depth coverage but lower breadth coverage. This makes the RNA-seq signal profiles very suitable for detecting small deletions that are in exonic regions. We show that a large number of small deletions can be detected using using RNA-seq signal profiles. On the other hand, ChIP-Seq[19] signal profiles for diffuse histone modifications generally have high breadth coverage but low depth coverage. In addition, these experiments are generally done in combination. This is important because although each experiment assays a different type of genome-wide activity, we show that pooling the signal profiles increases both the depth and breadth coverages and can bring enough power to an adversary for genotyping large deletions and perform successful linking attacks.

The paper is organized as following: We first present the general scenario of linking attacks that utilize signal profiles. We next propose a new metric for quantifying the extent to which genotypes of small and large deletion variants can be estimated using functional genomics signal profiles. In combination with information content of the deletion variants, we use this new metric for evaluating the extent of characterizing information leakage from functional genomics datasets. We next present several practical instantiations of linking attacks that utilizes different practical methods for deletion variant genotyping. Finally, we focus on protection of the signal profiles against linking attacks. We present a novel signal processing methodology for anonymizing a signal profile. We show that it is effective in decreasing the predictability of deletion variant genotypes from signal profiles. The source code for linking attacks and anonymization can be downloaded from http://archive.gersteinlab.org/proj/PrivaSig.

2. Results

2.1. Linking Attack Scenario

Figure 1 summarizes the linking attack scenario. The attack involves cross-referencing the individuals in a signal profile dataset (denoted by S) against the individuals in a genotype dataset, denoted by G. The signal profile dataset is publicly available and it contains, for each individual, a genome-wide signal profile, and an anonymized identifier. The signal profile for an individual represents the measurements of functional activity at each genomic position for this individual. In addition, the signal profile dataset contains sensitive information about each individual, e.g. HIV status. We assume that this dataset is generated for research purposes and is publicly released. The genotype dataset, G, contains, for each individual, the genotypes for a panel of structural variants, denoted by p_G. The genotype dataset also contains the identities of the individuals. Thus, G is normally assumed to be protected. We assume that the adversary obtains access to G. This accession could be established by lawful or unlawful means. For example she (we assume the adversary is a female) might have stolen it or she might be legally allowed to access it but she violates the terms of data accession. The main objective of the adversary is to link G and S by first predicting the structural variant genotypes from signal profiles in S, then matching the predicted genotypes to the genotypes in G. For any matching individuals in G and S, the name and the sensitive information, i.e. HIV status, are revealed to the adversary.

The attack has two steps. The first step is genotyping of the deletion variants, which is illustrated in Fig 1a. The adversary has access to a genome-wide signal profile dataset (S) for a sample of individuals. This dataset stores, for each individual, a genome-wide signal profile, for example RNA-seq, or ChIP-Seq data. In the first scenario, we assume that the adversary has access to a reference panel of genomic structural variant loci, which are denoted by p_S. For each individual, she utilizes the signal profile and genotypes the deletions in p_S. After the genotyping, the adversary builds a data matrix with the predicted genotypes, which is denoted by $\tilde{G}$. We refer to this scenario, where the adversary has access to a reference panel of structural variants, as linking based on "genotyping only". The second scenario, also illustrated in Fig 1a, is very similar except that the adversary does not have access to the panel of structural variants but discovers the panel of structural variants from the signal profiles. She then uses the signal profiles to genotype the SVs in this de-novo discovered SV panel. We refer to this scenario as linking based on "joint discovery and genotyping". After the genotyping, the genotyped SV matrix ($\tilde{G}$) includes, for each individual, the predicted SV genotypes, and also the sensitive information about HIV status. $\tilde{G}$ can also be thought of as a noisy genotype matrix, since the genotype predictions may contain errors.

The second step of the linking attack is cross-referencing of the individuals in the genotyped SVs ($\tilde{G}$) and the individuals in the genotype dataset, G, illustrated in Fig 1b. The SV genotype dataset G is assumed to contain identifying information about individual's identities. Thus, we assume that this dataset was previously protected and is either leaked or stolen, e.g. variants from a glass. The adversary first compares her genotyped SV panel (p_S) to the SV panel of the genotype dataset, which denoted by p_G. After the matching of the SVs in the two panels, she compares the genotypes of the matching SVs in two panels. She uses this comparison to cross-reference the individuals in two datasets and finds the individuals that best match to each other with respect to genotype match distance, i.e., links the individuals in two datasets. The results are used to link the individuals in genotype dataset to those in the signal profile dataset and the sensitive information, e.g., HIV status of individuals in the genotype dataset are revealed to the adversary (the matched columns in the final linked matrix).

2.2. Information Content and Correct Predictability of Structural Variant Genotypes

In order to assess the correct predictability of SV genotypes, we propose using a measure named genome-wide predictability of SV genotypes, denoted by π_{GW}, from signal profiles. The predictability measures how accurately an SV genotype can be estimated given the signal profile (Methods Section). The predictability of the genotype of a structural variant is the conditional probability of the variant genotype given the signal profile. By this definition, the predictability only depends on the genomic signal levels of an individual and how well they can be used to predict genotypes. For example, Fig 1c illustrates a large deletion that can be easily predictable using the histone modification signal profiles. In principle, the genome-wide predictability is computed for each individual independently from other individuals. Therefore the genome-wide predictability of a variant from signal profile is independent of the population frequency of the variant.

Other than the predictability, an important measure in the linking attacks is the information content each SV genotype supplies. We utilize a previously proposed metric termed individual characterizing information (ICI) to quantify the information content of each SV[16]. For a given variant genotype, ICI measures how much information it supplies for pinpointing an individual in a population. This measure gives higher weight to the genotypes that have low population frequency and vice versa. As we discussed above, the genome-wide predictability is independent of the population frequency of the variants. Therefore the adversary can utilize genome-wide prediction approaches and predict rare variant genotypes to gain high ICI and characterize individuals very accurately.

Deleted: [16].

2.3. Linking Attacks using RNA-Seq Signal Profiles

We first focus on the predictability of short deletions using RNA-seq signal profiles. Fig 1d illustrates a hypothetical example of how the small deletions in RNA-seq signal profiles can be detected as small and sudden dips in the signal. In order to show an example and represent the relevance of small deletions in the RNA-seq signal profiles, we included a screenshot of signal profiles around a small deletion for 6 individuals in the GTex Project (Supp. Figure 5). The 2 base pair deletion, rs34043625, can be easily detected for three of the individuals that are shown. An important aspect of the effect of small deletions on the signal profile is the extent that they affect the total expression of a gene. It is clear from Supplementary Figure 5 that the total signal in the small dips in the RNA-seq signal is much smaller than the perturbations caused by the other genetic factors like eQTLs and sQTLs. In general, an eQTL is associated with a global change in the total signal on the RNA-seq signal profile of a gene. However, a small deletion affects a localized position on the RNA-seq signal profile of the gene with relatively smaller effect on the total expression of the gene, assuming that the small deletion is not an eQTL. It is also worth noting that these signal profiles are publicly available from the UCSC Genome Browser.

Deleted: Figure 5). The 2 base pair deletion, rs34043625, can be easily detected for three individuals that are shown. These

As we mentioned earlier, the RNA-seq signal profiles generally have high depth coverage but low breadth coverage. The short deletions are the type of variants that can be detected most easily using signal profiles that have high depth and low breadth coverage. By small deletions, we refer to the deletions that are smaller than 10 base pairs. Regarding detection of small deletions, the basic observation is that each deletion is manifested as an abrupt dip in the signal profile (Fig 1d). The discovery and genotyping of a deletion rely on detecting these dips in the signal profiles. The genome-wide predictability (π_{GW}) of the small deletions quantifies how well the adversary can identify the dips corresponding to deletions from the signal profile (Methods Section). We first estimated the genome-wide predictability for the panel of short deletions in 1000 Genomes Project using the RNA-seq expression signal profiles from the GEUVADIS

project. Figures 2a,b show π_{GW} vs ICI for short deletions. There is a substantial number of deletions that have much higher predictability compared to a randomized dataset where the signal profile is randomized with respect to the location of deletions. There are also many variants with very high ICI (on the order of 5-6 bits) with high predictability (greater than 80% predictability). This result shows clearly that signal profile based attack scenario is much more powerful that the other approaches like population-wide prediction of variant genotypes (Supplementary Fig 2)

In order to present practicality of small deletion predictability and information content, we propose an instantiation of a linking attack where we utilize outlier signal levels in the signal profiles for discovery and genotyping of the small deletions. As mentioned before, the genotyping of deletions are based on detecting the abrupt dips in the signal profile. In order to detect these dips in the signal profile, the adversary utilizes a quantity we term s*elf-to-neighbor signal ratio*, denoted by $\rho_{[i,j]}$, that measures the extent of the dip in the signal as the fraction of signal on the interval and the signal in the neighborhood,

$$\rho_{[i,j]} = \frac{\text{Average signal within } [i,j]}{\text{Average signal within neighborhood of } [i,j]}.$$

The genomic regions with low $\rho_{[i,j]}$ values point to intervals tend to have dips in them. For each individual, the prediction method sorts the short deletions with respect to *self-to-neighbor signal ratio* and assigns homozygous genotype to a number of deletions with the smallest *self-to-neighbor signal ratio* (Methods Section). The adversary then compares these genotyped deletions to the genotype dataset and identifies the individual whose deletion genotypes that are closest to the predicted genotypes. Using this genotyping strategy, we simulated an attack to link GEUVADIS signal profile dataset to the 1000 Genomes genotype dataset. We used the panel of deletions from the 1000 Genomes Project. In the *genotyping only* scenario, the linking is perfectly accurate when the adversary utilizes more than 40 deletions (Fig 2c). In the scenario where the adversary performs *joint discovery and genotyping*, the linking accuracy is maximized (around 60%) when the attacker utilizes the top 50 deletion candidates in linking (Fig 2d). Next, we studied how accurate the linking is if adversary uses deletions of different lengths. Figure 2e shows the accuracy and number of indels with different lengths. The accuracy of linking decreases substantially for the indels that are longer than 5 base pairs. The decrease in accuracy is affected by both the decrease in the number of indels (i.e., low ICI), shown in Fig 2e, and also decreasing predictability of indels whose lengths are above 5 base pairs. We then asked, for each individual, what the minimum number of indels that are sufficient to accurately link the individual is. Figure 2f and 2g shows the distribution of minimum number of indels for accurately linking each individual in the GEUVADIS dataset, for genotyping only (Fig 2f) and adversary jointly discovery and genotyping (Fig 2g) scenarios. As small as 30 indels are sufficient to correctly link a large fraction of the individuals.

In the previous analysis, the sample set used for discovery of deletion panel and RNA-seq sample set are matching, i.e. 1000 Genomes individuals. This may introduce a bias in linking because the SV genotype dataset may contain rare deletions which may also be in the panel of deletions. This would make it trivial to link some of the individuals. To get around this bias, we studied linking attack where signal profile dataset is generated by the GTex Project Consortium [27, 28] and the panel of small deletions is the deletion set generated by the 1000 Genomes Project. This way, the SVs in the panel are identified among the 1000 Genomes individuals while the RNA-seq signal profiles are generated on a non-matching set of individuals in the GTex Project. In other words, the deletion panel is discovered in a sample set that is totally independent of the sample set that the adversary is linking. In this scenario, the adversary is linking

Deleted: [28, 29] and the panel of small deletions is the deletion set generated by the 1000 Genomes Project.

the signal profile dataset to the genotype dataset that is obtained from the GTex Project. With this setup, we first computed π_{GW} versus ICI for the deletions and observed that there is substantial enrichment of deletions that have high predictability with high ICI compared to randomized datasets (Fig 3a, b). As before, there are many deletions that are highly predictable (>80%) and are very high in ICI (>5bits). In addition there is a substantial increase of predictability in real data compared to the randomized dataset.

We next instantiated the linking attack using the extremity based approach. In the instantiation, we first evaluated the attack based on *genotyping only* scenario. In this scenario, the linking accuracy is close to 100% using a relatively small number of variants, i.e., 30 variants (Fig 3c). An interesting observation is that when the attacker increases the number of variants used in the attack, the linking accuracy decreases. This is caused by the fact that the additional variants after the 30 variants are incorrectly genotyped and decrease the accuracy of linking. In simple terms, the additional variants act as noise and decrease linking accuracy.

Following this, one question that arises is if the adversary can assign reliability scores to the linked individuals. We tested whether the *first distance gap* (Methods Section) measure is suitable for evaluating the reliability of linkings. This is important because when the overall linking accuracy is low, e.g. smaller than 50%, unless the attacker has a systematic way of selecting correct linkings, the risk of linking attack is low. As a test case, we focused on the linking where the adversary uses 200 deletions where the overall linking accuracy is 35% (Fig 3c). Figure 3d shows the sensitivity and positive predictive value (PPV) with changing *first distance gap* metric. The adversary can link 10% of the individuals perfectly and 20% of the individuals are linked with around 90% accuracy, i.e., makes 1 mistake in 10 linkings. Figure 3d also shows the average sensitivity and average PPV over 100 random selections of the linkings. As expected, the PPV is always around 35% and average sensitivity is also always smaller than *first distance gap* based selection of linkings. These results show that even some parameter selections (e.g., number of variants) may show low accuracy, the adversary can increase accuracy by selecting the linkings using first distance gap measure.

2.4. Linking Attacks using ChIP-Seq Signal Profiles

We next focused on predictability versus ICI of large deletions, which are longer than 1000 base pairs. Since the deletions are large, the signal profiles that are suitable for genotyping these deletions must have high breadth coverage. We utilize the ChIP-Seq signal profiles for histone modifications, which generally manifest high breadth and low depth coverage. Several recent studies have generated individual level epigenomic signal profiles through ChIP-Seq experiments [32–34]. These studies aimed at revealing how the genetic variation interacts with the epigenomic signals, mainly the histone modifications. These datasets are very convenient for our study because the majority of the individuals have matching structural variant genotype information in the 1000 Genomes Project. It is worth noting that although we are focusing on the predictability of large deletion genotypes from ChIP-Seq profiles, this does not mean that the small deletions are not detectable in the ChIP-Seq dataset. In fact, the small deletion genotyping based linking attack we presented in the previous section can be applied to ChIP-Seq signal profiles as it is.

Deleted: [32–34].

We use these personalized epigenomic signal profiles for quantifying how much characterizing information leakage they provide. For any individual where there are multiple histone mark ChIP-Seq signals, we pool the signal profiles and compute several features for each large deletion. These are then used for quantifying information leakage (Methods Section). First, we computed π_{GW} versus ICI using the

panel of large deletions in 1000 Genomes Project. Figure 4a,b show π_{GW} versus ICI for the large deletions from the 1000 Genomes. We use the personal epigenome profiling ChIP-Seq datasets presented in studies by Kasowski et al and Kilpinen et al (Methods Section). Similar to the small deletion analysis, it can be seen that for both datasets there are many large deletions with high predictability and high ICI.

We next focused on instantiating linking attacks using ChIP-Seq profiles. We again utilize a variant of the outlier based genotyping in the linking attack. The genotyping of the panel of large deletions is done as follows. The average ChIP-Seq signal on each deletion is computed and the variants are sorted with respect to their average signal in increasing order. The deletions with smallest ChIP-Seq signal are assigned homozygous deletion genotype. For the deletions with assigned genotypes, we identified the individual in the genotype dataset (from the 1000 Genomes project) whose genotypes match closest to the assigned genotypes. We repeated this linking attack with different number of windows and computed the accuracy of linking (Methods Section). Figure 4c shows the accuracy of linking attack based on *genotyping only* scenario, where the adversary is assumed to have access to the large deletion panel from 1000 Genomes. The linking accuracy reaches 100% with a fairly small number of deletions for both datasets. For the *joint discovery and genotyping* scenario where the adversary first discovers deletions and then genotypes them, the accuracy is also very high with small number of identified deletions (Fig 4d).

An interesting question about histone modifications is which combinations of histones leak the highest amount of characterizing information. To answer this question, we studied the individual NA12878, for which there is an extensive set of histone modification ChIP-Seq data from the ENCODE Project[25]. We have evaluated whether different combinations of histone modifications render NA12878 vulnerable against a linking attack among 1000 Genomes individuals, which is illustrated in Fig 4e. In general, we have observed that NA12878 is vulnerable with the dataset combinations that cover the largest space in the genome. This can be simply explained by the fact that when histone marks cover a larger genomic region, higher number of deletions can be detected and genotyped. For example, H3K36me3 and H3K27me3, an activating and a repressive mark respectively, are mainly complementary to each other and they render NA12878 vulnerable. In addition, H3K9me3, a repressive mark that expands very broad genomic regions, renders NA12878 vulnerable in several combinations with other marks. On the other hand, H3K27ac, an activating histone mark that spans punctate regions do not render NA12878 vulnerable.

Deleted: [26].

2.5. Linking Attacks using Hi-C Matrices

We also asked whether a relatively new data type, Hi-C interaction matrices can be used for identification of genomic deletions. Hi-C is a high throughput method for identifying the long range genomic interactions and three dimensional chromatin structure[35]. It is based on proximity ligation of the genomic regions that are close-by in space followed by high throughput sequencing of the ligated sequences. After sequencing data is processed, it is converted to a matrix where the entry (i, j) represents the strength of interaction between i^{th} and j^{th} genomic positions. To study leakage from Hi-C matrices, we again focused on NA12878 individual for whom Hi-C interaction matrices are generated at different resolutions[36]. We first converted the matrix into a genomic signal profile. For this, we summed the interaction matrix along columns and obtained a signal profile along the genome (Fig 5a, Methods Section). This way, we are simplifying the multidimensional nature of the Hi-C contact matrix and treat it as a sequencing assay that spans the entire genome. It is important to emphasize that the standard analysis of Hi-C matrices do not involve such a signal profile generation. We are using this step to convert the Hi-C matrix into a signal

Deleted: [35].

Deleted: [36].

activity profile along the genome. Using the signal profile, we simulated an extremity based linking attack using the outliers in the Hi-C signal profile: For all the large deletions in the 1000 Genomes whose population frequency is higher than 1%, we computed the average Hi-C signal. We next sorted the deletions in increasing order with respect to average signal and assigned top 1000 windows with homozygous deletion genotype. We next compared the predicted genotypes with all the genotypes in the 1000 Genomes project. We observed that NA12878 is vulnerable to this attack when the Hi-C contact matrix resolution (bin length) is 10 kilobases or smaller (Fig 5b).

It is important to clarify that we are focusing on using the final output of Hi-C data, i.e., the Hi-C contact matrix, for generating a genome-wide signal profile and performing a linking attack. We are not studying the possibility of discovering complex structural variants using the paired-end reads of Hi-C experiment, which is a different problem by itself[37]. It also requires access to mapped reads, which we assume the attacker does not have. As we explained above, our attack scenario treats the Hi-C data as any type of sequencing data and uses the linear genomic signal profile to identify deletions for the purpose of linking datasets. We are highlighting the fact that Hi-C interaction matrices themselves leak substantial amount of characterizing information.

Deleted: [37].

2.6. Anonymization of RNA-Seq Signal Profiles

An important aspect of the genomic privacy is risk management and protection of datasets. For protection, anonymization of the datasets is the most effective way so that the data can be shared publicly in a safe manner. The personal RNA-seq datasets are currently by far the most abundant datasets compared to other functional genomic datasets. For example, the RNA-seq signal profiles are being publicly shared from the GTex project while the genotypes are not in public access. In addition, RNA-seq is becoming commonly used in the clinical settings and new RNA-seq based assays are being developed to probe gene expression, for example single cell RNA-sequencing. Altogether these make protection of RNA-seq data urgent. We therefore focus on protection of the RNA-seq datasets. The most effective way to protect against linking attack scenario is to ensure that the deletion genotypes cannot be inferred from the signal profiles. As we showed in previous sections, the small deletions are major source of leakage of genetic information from RNA-seq signal profiles. We propose systematically removing the dips in the signal profiles as a way to anonymize the signal profiles against the prediction of small deletions. Specifically, we propose smoothing the signal profile using median filtering locally around a given panel of deletions (Methods Section). We have observed that median filtering removes the dips in the signal very effectively while conserving the signal structure fairly well. To evaluate the effectiveness of this method, we applied signal profile anonymization to the RNA-seq signal profiles generated by the GEUVADIS Project consortium and the GTex Project Consortium. After application of the signal profile anonymization, we observed that the large fraction of the leakage is removed for GTex datasets (Fig 2b and 3b). We also observed that the extremity based linking attack proposed in the previous section is ineffective in characterizing individuals such that no individuals are vulnerable for GTex project and only 1% of the individuals are vulnerable for GEUVADIS dataset. It is worth noting that this procedure can be used anonymizing not only RNA-seq signal profiles but also other signal profiles against attacks that are based on small deletion genotyping. The anonymization is, however, not as effective for large deletions. This is not a major concern for RNA-seq signal profiles as we observed that large deletions are not easily genotyped using RNA-seq data. However, as we showed in the previous section, the linking attacks can be successful when they use the large deletions that are genotyped using ChIP-Seq datasets.

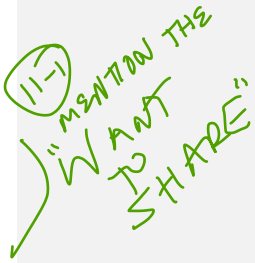
3. Discussion

The sequencing based functional genomics assays provide very large amount of biological information. Within this, much of the variant genotype information is within the raw reads (Supplementary Figure 6). In fact an adversary that gains access to the raw reads can easily call SNPs, indels, and structural variants. This is why raw reads are always protected from public access. The gene expression levels have also been shown to leak enough genotype data that can be used in linking attacks[16, 18]. The privacy concerns around sharing signal profiles are not well studied yet.

We have systematically analyzed a critical source of sensitive information leakage from the signal profile datasets, which were previously thought to be largely secure to share. Specifically, our results show that an adversary can perform fairly accurate linking attacks for characterizing individuals by prediction of structural variants using functional genomics signal profiles. These results reflect how the rich nature of functional genomics data can cause privacy concerns in an unforeseen manner. This is an interesting aspect of the data. Although there may be some variant information in functional genomics signal profiles, these data are not generated mainly for detecting variant information. The main purpose of them is to reveal how they change under different conditions and how they relate to phenotypes, which may be sensitive. The existence of residual variant information, as we showed, may enable an adversary to reveal sensitive information about individual.

Although we are focusing mainly on RNA-seq and ChIP-Seq signal profiles, the linking attack scenario and the measures that we presented are data-driven and are generally applicable to any type of genome-wide signal profile. For example, although it is obvious, the linking attacks can easily be carried out on the DNA-sequencing signal profiles. Also, signal profiles from genome-wide profiling techniques other than sequencing based assays, like ChIP and expression tiling arrays[38, 39] can be vulnerable to the linking attack scenario that we presented. Different genome-wide data representation, e.g., Hi-C interaction matrices, can be utilized for generation of genome-wide signal profiles and these can in turn leak sensitive information. We believe that many more genome-wide omics technologies will be developed in the near future[40]. The genome-wide signal profiles will be a vital source of information in the analysis of these datasets. The framework we presented here can be utilized for assessing the leakage and protection of these datasets. In addition, albeit the focus is on the small and large deletion variants, the analyses of predictability and practical linking attacks can be extended for other structural variants, for example, genomic insertions.

We showed that the linking can be done by predicting a fairly small number of variants (generally less than 100 variants). Our results show that these data leak enough information for individual characterization among a large set of individuals. This can cause practical privacy issues because several large consortia are making signal profiles publicly available. For example GTex signal profiles are publicly available through the UCSC Genome Browser. Given the extent of public sharing of datasets, we believe that the anonymization of the RNA-seq signal profiles using the signal processing technique that we proposed is very useful. The technique we proposed applies a signal smoothing around all the known deletions and removes a significant amount of characterizing information. The anonymization procedure can be easily integrated into existing functional genomics data analysis pipelines. We believe that this anonymization technique can complement other approaches for removing genetic information from shared datasets. For example file formats like MRF[41] and tagAlign[25] can enable removing raw sequence information from reads while keeping the information about read mapping intact. It is worth



Deleted: [16, 18].

Deleted: [38, 39]

Deleted: [40].

Deleted: [41] and tagAlign[26] can enable removing raw sequence information from reads while keeping the information about read mapping intact. It is worth noting that the anonymization method that we presented does not close all the sources of leakage. The anonymization procedure aims to close the leakages caused by the genotyping of genomic deletion using the dips in the signal profile. These leakages are very accessible to and adversary and we believe that they must be urgent closed because they can be detected directly from the signal profiles. Given other types of data, there can still be other sources of genotype information leakage after the anonymization is applied. For example, the gene expression levels can be used to infer genotype information, which was demonstrated in earlier studies. In addition, the effects of variants on the activity levels of pathways are not well known yet.

noting that the anonymization method that we presented does not close all the sources of leakage. The anonymization procedure aims to close the leakages caused by the genotyping of genomic deletion using the dips in the signal profile. These leakages are very accessible to and adversary and we believe that they must be urgent closed because they can be detected directly from the signal profiles. Given other types of data, there can still be other sources of genotype information leakage after the anonymization is applied. For example, the gene expression levels can be used to infer genotype information, which was demonstrated in earlier studies[16, 18]. In addition, the effects of variants on the activity levels of pathways are not well known yet. The complex machine learning frameworks, such as deep learning methods, have great potential to reveal the correlations between variants and activity levels of pathways. The leakage from the pathway level activity can be analyzed by using deep learning based approaches.

Overall, at this point, it is useful to review all the sources of information leakage from functional genomics experiments, such as RNA-Sequencing, and point out the sources that we probed in this paper. First, there is the leakage directly from the reads. This is the most obvious leakage, and this leakage is avoided with by simply not sharing the raw reads. Next source of leakage is from the signal profile. This leakage is addressed in this paper. There is yet another source of leakage though, when one averages over the signal file, and produces quantifications in particular regions such as genes. These quantifications can be subtly connected with variants through the notion of eQTLs. This is not addressed in this paper, and there can be substantial leakage from these quantifications.

Furthermore, one can envision additional sources of leakage beyond that, in these main areas. For instance, one can imagine complex and subtle correlations between the levels of gene expression of many genes within pathways and networks. Although there has been interest in identifying these higher order QTLs, these are not yet extensively studied[28]. Complex machine learning techniques, such as deep learning, can reveal subtle correlations of gene expression at the network level with variants. Also, eQTLs traditionally have been linked to genes; ostensibly, one might imagine by averaging over various intergenic regions, some of the more highly expressed intergenic regions might also show correlations with variants. This is another source of information not studied in this work. Finally, an additional source of information is, while we do look at calling of particular types of structural variants, such as small and large deletions, there may be very large-scale, megabase-scale deletions, which affect many genes. This is particularly the case for somatic events in cancer samples. This case is also not covered by our procedure.

Finally, we would like to emphasize that we focused on a particular type of leakage of private information in functional genomics data, such as RNA-Seq data, such that the leakage stems from the signal profile. There are many other sources of information, however the signal file is currently at the junction between public and private information, and is where genomic information is begun to be shared publicly. Hence, we believe it is particularly important to probe the leakage from the signal profile representation of functional genomics data. It might unfortunately be the case that this type of information is not able to be shared publicly in the future, perhaps only sharing gene-level quantifications, or even worse, nothing at all. We wish to emphasize that, in this paper, we are not trying to look at all sources of leakage from functional genomics data, but just the sources of leakage right at the decision boundary of sharing and not sharing.

4. Methods

We provide the details of the computational methodologies. We first introduce the notations. The genomic deletions are intervals of genomic coordinates. We refer to them simply as intervals, e.g. a deletion between genomic positions i and j by $[i,j]$. The genotype of a genomic deletion at $[i,j]$ is denoted by $G_{[i,j]}$, which is a discrete random variable distributed over the 3 values $\{0,1,2\}$. These values correspond to the three genotypes of the deletion and they represent how many copies of the genomic sequence is deleted. The functional genomics read depth signal is denoted by $\boldsymbol{S}$, which is a vector of values corresponding to each genomic position**.** The signal level at genomic position at i is denoted by S_i. An important quantity that we utilize in formulating methods is the multi-mappability profile of the deletion regions. The multi-mappability is a signal profile that measures, for each position in the genome, how uniquely we can map reads. The multi-mappability signal is denoted by $\boldsymbol{M}$**,** which is a vector of multi-mappability signals for all the genomic positions and the signal at genomic position i is denoted by M_i. The multi-mappability signal profile is generated as follows: The genome is cut into fragments and the fragments are mapped back to the genome using bowtie2[42] allowing the multi-mapping reads. We then generate the read depth signal of the mapped reads. In this signal profile, the uniquely mapping regions receive low signal while the multi-mapping regions receive high signal[43].

Deleted: [42]

Field Code Changed

4.1. Genome-wide Predictability of Deletion Genotypes and Individual Characterizing Information

The genome-wide predictability, π_{GW}, of a deletion genotype refers to how well a deletion can be genotyped given the functional genomics signal ($\boldsymbol{S}$) of interest. We assume that the adversary employs a prediction methodology based on statistical modeling of the deletion genotypes with respect to read depth signal profile such that the adversary utilizes features from the functional genomics signal profile. We define here the features that are most useful for genotyping deletions (Supp Fig 3). Given a deletion $[i,j]$, an important feature for genotyping the deletion is the average functional genomic signal within the deletion:

$$\bar{s}_{[i,j]} = \frac{\sum_{i'=i}^{j} S_{i'}}{j-i+1}.$$

Another feature is the average multi-mappability signal within the deletion:

$$\bar{m}_{[i,j]} = \frac{\sum_{i'=i}^{j} M_{i'}}{j-i+1}.$$

In order to measure the extent of the dip within the signal, we observed that a measure we termed *self-to-neighbor signal ratio* and *neighbor signal balance ratio* are very useful for genotyping. Given a deletion $[i,j]$, *self-to-neighbor signal ratio*, denoted by $\rho_{[i,j]}$, is computed as

$$\rho_{[i,j]} = \frac{2 \times \bar{s}_{[i,j]}}{\bar{s}_{[2i-j+1,i-1]} + \bar{s}_{[j+1,2j-i+1]}}.$$

This is simply twice the ratio of total signal on the deletion and the total signal in the neighborhood of the deletion. The *neighbor signal balance ratio*, is computed as

$$\eta_{[i,j]} = \min\left(\frac{\bar{s}_{[j+1,2j-i+1]}}{\bar{s}_{[2i-j+1,i-1]}}, \frac{\bar{s}_{[2i-j+1,i-1]}}{\bar{s}_{[j+1,2j-i+1]}}\right).$$

Finally, we observed that the average signal on the neighborhood of the deletion coordinates are useful in genotyping deletions. This is because when the neighbor signals are more balanced around a dip, i.e., higher $\eta_{[i,j]}$, the accuracy of deletion genotyping is higher. Next, we compute the average signal in the neighborhood as

$$\tau_{[i,j]} = 0.5 \times \left(\bar{s}_{[2i-j+1,i-1]} + \bar{s}_{[j+1,2j-i+1]}\right).$$

We define π_{GW} as the conditional probability of a deletion genotype g given the 5 features computed from functional genomics signal profile:

$$\pi_{GW}\left(G_{[i,j]} = g, \boldsymbol{S}_{[i,j]}\right) = P_{GW}\left(G_{[i,j]} = g \,\middle|\, \begin{matrix} \log_2\left(\bar{s}_{[i,j]}\right), \\ \log_2\left(\bar{m}_{[i,j]}\right), \\ \log_2(\rho_{[i,j]}), \\ \log_2(\eta_{[i,j]}), \\ \log_2(\tau_{[i,j]}) \end{matrix}\right).$$

This corresponds to the conditional probability (over all the deletions within the genome) that we observe the genotype g for a deletion at $[i,j]$ given the average functional genomics signal and average multi-mappability signal over the interval $[i,j]$. The probability is defined over the genome, i.e., we estimate the probability for all the deletions in the genome. For this, we compute 5 features for every deletion in the genome, then estimate the conditional probability using this set as the sample of deletions.

The basic idea behind the formulation of predictability is the observation that the regions with low functional genomics signal, low multi-mappability (i.e., uniquely mappable), low *self-to-neighbor signal ratio*, and high average neighbor signal are more likely to be deleted, i.e., their probability is large. Therefore, π_{GW} is higher for deletions that are more easier to identify than the deletions with lower π_{GW}. In order to estimate the conditional probabilities, we binned the feature values by computing the logarithm then rounding this value to the closest smaller integer value.

4.2. Discovery and Genotyping of Small and Large Deletions from Signal Profiles

The practical instantiation of the linking attacks that we study are based on genotyping the panel of small deletions, p_S, using the functional genomics data. In addition, when the deletions panel p_S is not available, the adversary also discovers the deletions using the signal profile. For GEUVADIS and GTex datasets, we perform small deletion genotyping using RNA-Seq signal profiles. The basic idea behind genotyping of deletions is the fact that there is a sudden dip in signal profile whenever there is a deletion (Fig 1d). In order to detect these dips, we observed that *self-to-neighbor signal ratio* is very useful for genotyping small deletions. For all the small deletions, *self-to-neighbor signal ratio*, $\rho_{[i,j]}$, neighbor signal balance, $\eta_{[i,j]}$, and average neighbor signal are computed. We then select the deletions that satisfy following criteria:

$$\begin{aligned} \bar{m}_{[i,j]} &< \bar{m}_{max} && \text{(High Mappability)} \\ \tau_{[i,j]} &> \tau_{min} && \text{(High Neighbor Signal)} \\ \eta_{[i,j]} &> \eta_{min} && \text{(High Neighbor Signal Balance)} \end{aligned}$$

For the set of small deletions that pass these criteria, we sorted the deletions with respect to increasing $\rho_{[i,j]}$. The deletions which are at the top of the sorted list correspond to the deletions which are highly mappable (low multi-mappability signal), have strong neighbor signal support (high average neighbor signal), and finally they have a strong signal dip on them (Low $\rho_{[i,j]}$, and high $\eta_{[i,j]}$). We selected the top n deletions and assigned them homozygous genotypes, i.e., $G_{[i,j]} = 0$. The basic idea is that the deletions with strongest signal dips are enriched in homozygous deletions. It is worth noting that this genotyping method only assigns homozygous genotypes. Although this might result in low genotyping accuracy (Supp Fig 4), these genotyping predictions have enough information for accurate linking attacks.

We utilize pooled ChIP-Seq read depth signal profiles and Hi-C signal profiles for genotyping large deletions. For genotyping the large deletions, we first computed the average signal ($\bar{s}_{[i,j]} = \frac{\sum_{i'=i}^{j} \boldsymbol{s}_{i'}}{j-i+1}$) and average multi-mappability signal ($\bar{m}_{[i,j]} = \frac{\sum_{i'=i}^{j} \boldsymbol{M}_{i'}}{j-i+1}$) on each large deletion. We select candidate large deletions using average multi-mappability signal:

$$\bar{m}_{[i,j]} < \bar{m}_{max} \qquad \text{(High Mappability)}$$

We sorted the deletions that satisfy above criteria with respect to increasing average signal, $\bar{s}_{[i,j]}$. For the top n deletions, we assigned homozygous genotypes, i.e., $\tilde{G}_{[i,j]} = 0$.

We generally observed that the parameter selection for filtering variants did not have substantial effect on accuracy of linking attacks as long as they are not made too stringent. In the computational experiments, we used $\bar{m}_{max} = 1.5$, $\tau_{min} = 10$, $\eta_{min} = 0.5$ as the parameter set.

For the case when the adversary does not have access to the deletion panel, we fragment the genome into windows and use these windows as candidate deletions. Above procedure is utilized for selection of the candidate deletions, which are assigned homozygous deletion genotypes. For small deletions, we use 5 base pair windows within the exonic regions. For large deletions, we use 1000 base pair windows over all genome.

4.3. Instantiations of Genome-wide Linking Attack

Following the genotyping of the deletions in p_S, we use the genotyped deletions to link the individual to the individuals in the SV genotype dataset. Given the genotyped deletions for the k^{th} individual in the signal profile dataset, we first compare these deletions to the panel of deletions in the genotype dataset, p_G. The comparison is performed by overlapping the deletions in p_S and in p_G. Any two deletions that overlapped at least 1 base pair are assumed to be common in the two panels. For the set of deletions that are common in two panels, $\{[i_1, j_1], [i_2, j_2], \dots, [i_n, j_n]\}$, we compute the genotype distance by matching the genotypes,

$$d_{k-l} = \sum_{\substack{a=[i',j'] \in \\ \{[i_1,j_1], \\ \dots \\ [i_n,j_n]\}}} d(\tilde{G}^{(k)}_{[i',j']}, G^{(l)}_{[i',j']})$$

where d_{k-l} represents the genotype distance of k^{th} individual in the signal profile dataset to the l^{th} individual in the genotype dataset and $d\left(G_{[i',j']}, G_{[i',j']}\right)$ is the distance function:

$$d\left(\tilde{G}^{(k)}_{[i',j']}, G^{(l)}_{[i',j']}\right) = \begin{cases} 1 \; if \; \tilde{G}^{(k)}_{[i',j']} \neq G^{(l)}_{[i',j']} \\ 0 \; if \; \tilde{G}^{(k)}_{[i',j']} = G^{(l)}_{[i',j']} \end{cases}.$$

We next compute the genotype distance of k^{th} individual to all the individuals in the genotype dataset; d_{k-l} for all l in $[1, K]$ where K represents the number of individuals in genotype dataset. The individual in the genotype dataset that has the smallest genotype distance is linked to k^{th} individual:

$$\text{linked individual's index} = \underset{l' \in [1,K]}{\text{argmin}}(d_{k-l'})$$

Finally, if the linked individual in the genotype dataset matches the individual in signal profile dataset, we mark the individual in the signal profile as a vulnerable individual. We also compute the *first distance gap*, $d_{1,2}$, for each linked individual[16] to evaluate the reliability of linking. For a linked individual, first distance gap is computed as

$$d_{1,2} = d_k^{(1)} - d_k^{(2)}$$

where $d_k^{(1)}$ and $d_k^{(2)}$ is the minimum and second minimum genotype distance among all the genotype distances computed between k^{th} individual and all the genotype dataset individuals.

4.4. Computation of Sensitivity and Positive Predictive Value

In order to compute the sensitivity and positive predictive value (PPV) of linkings when the linkings are selected using first distance gap measure, we use following formula:

$$\text{Sensitivity} = \frac{\text{Number of correctly linked individuals with } d_{1,2} > d_{1,2}^{min}}{\text{Number of All Individuals}}$$

$$\text{PPV} = \frac{\text{Number of correctly linked individuals with } d_{1,2} > d_{1,2}^{min}}{\text{Number of Individuals with } d_{1,2} > d_{1,2}^{min}}$$

where $d_{1,2}^{min}$ represents the minimum first distance gap measure that are used to select individuals. In these formulae, sensitivity represents the fraction of all individuals that adversary correctly links. PPV represents the fraction of individuals that are correctly linked among the individuals whose linking satisfies minimum first distance gap threshold.

4.5. Anonymization of Signal Profile Datasets

The anonymization of the signal profile datasets refers to the process of protecting the signal profile data against correct predictability of the genotypes for deletion variants. As we discussed earlier, the large and small dips in the functional genomics signal profiles are the main predictors of deletion variant genotypes. To remove these dips systematically, we propose using the median filtering[44] based signal processing to locally smooth the signal profile around the deletion. This signal processing technique has been used

Deleted: [16] to evaluate the reliability of linking.

Deleted: [44]

to remove shot noise in 2 dimensional imaging data and 1 dimensional audio signals[43, 45]. For each genomic a in the deletion $[i, j]$, we replace the signal level using the median filtered signal level:

$$\tilde{x}_a = \text{median}\left(\{x_b\}, b \in \left[a - \frac{l}{2}, a + \frac{l}{2}\right]\right)$$

where x_a refers to the signal level at the genomic position a, $l = j - 1 + 1$, $\tilde{x}_a$ refers to the smoothed signal level at position a, and median refers to the median of all the signal values in the genomic region $\left[a - \frac{l}{2}, a + \frac{l}{2}\right]$. The median is computed by sorting all the signal levels and choosing the value in the middle of the sorted list of signal levels.

Deleted: [43, 45].

5. Datasets

The mapped reads for the RNA-seq data from gEUVADIS project are obtained from gEUVADIS project web site (http://geuvadis.org/). The RNA-seq mapped reads from the GTex project are obtained from dbGAP portal. We used only the RNA-seq datasets from whole blood tissue to create signal profiles. The structural variant panel and genotypes are obtained from the 1000 Genomes Project. The very low frequency SVs may introduce bias since they can uniquely identify and individual. In order to get around this bias, we removed the SVs whose minimum genotype frequency is larger than 0.01. Also, we extended the genotype dataset by re-sampling 1000 Genomes deletion dataset and created genotype data for 10,000 simulated individuals.

We have utilized randomized datasets for comparison of predictability with real data. In order to create randomized data, we shuffled the signal profiles circularly. This way, the association between the SV genotypes and signal profiles are randomized.

Figure Legends

Figure 1: Illustration of the attack scenario**.** a) The adversary starts the attack with a signal profile dataset(S). This dataset contains, for each individual, a genome-wide signal profile and also sensitive information, e.g., HIV status. The names are anonymized into IDs as shown in blue shaded column. The adversary uses an SV panel (p_S) in the attack. This panel can be obtained from outside (1) or the adversary can use the genome-wide signal profiles to discover the panel (2), as denoted by the shaded red arrows. She then genotypes the SVs (3) in the panel and builds the genotyped SVs dataset ($\tilde{G}$). b) The adversary acquires an SV panel (p_G) and genotype dataset (G) which contains genotypes of the SVs in the panel for a large number of individuals. In order to link the genotyped SV dataset ($\tilde{G}$) to the SV genotype dataset, the adversary compares her SV panel (p_S) to the SV panel (p_G). For the matching SVs, the adversary compares the genotypes. The individuals in G that have good matches with respect to genotype distance are linked to signal profile individuals, as indicated by the matching of the colored columns. This linking reveals the HIV status of the individuals in genotype dataset. c) Example of a large deletion in NA12878 individual and how it affects signal profiles. 70kb long region is deleted in NA12878 individual and the decrease in signal profiles show the loss of signal along the deletion. d) The schematic representation of large and small deletions and how they are manifested in signal profiles. The large deletions show a large decrease in the signal profiles while small deletions have much smaller footprints.

Figure 2: The accuracy of linking attack on GEUVADIS dataset. a) The scatter plot shows the ICI versus predictability for each deletion, denoted by a dot. The real data (blue dots) show a much higher

predictability compared to randomized data (red dots) b) After anonymization of signal profiles, the predictability of real data is decreased substantially. c) The accuracy of linking with genotyping of a known panel. The number of variants used in the attack is shown in x-axis while accuracy is shown on y-axis. The variants are sorted with respect to decreasing predictability. d) The accuracy of linking when adversary performs joint discovery and genotyping of deletions to perform linking. e) The blue plot shows the accuracy of linking when indels of specific length are used in the attack. Green plot shows the distribution of indels lengths. f) For the genotyping only scenario, the plot shows the distribution of minimum number of variants that is required to identify each individual. X-axis shows the number of indels and y-axis shows the frequency of individuals that can be identified. g) For the scenario where adversary discovers the SV panel first and performs genotyping on the discovered panel, the plot shows the distribution of minimum number of variants that is required to identify each individual.

Figure 3: The accuracy of linking attack on GTex dataset. The ICI leakage versus predictability for all the indels before (a) and after (b) signal profile anonymization. c) The linking attack accuracy with changing number of variants used in the attack. X-axis shows the number of variants used in the attack and y-axis shows the accuracy of linking. d) When the adversary uses the 200 variants in (c) and selects linking based on thresholding $d_{1,2}$ (shown on x-axis), the plot shows on the y-axis the sensitivity (black) and positive predictive value (red) of linkings for real (solid) and random (dashed) datasets while $d_{1,2}$ is changed.

Figure 4: a) The scatter plot of ICI leakage versus predictability for Kasowski (a) and Kilpinen (b) datasets. c) The accuracy of linking attack on the two datasets for genotyping only scenario. X-axis shows the changing number of variants used in the attack and y-axis shows the linking accuracy. d) The accuracy of linking on the two datasets when the adversary performs the attack by joint discovery and genotyping of deletions. e) The accuracy of linking of NA12878 when adversary utilizes different combinations of histone modifications. The first column shows different combinations. Middle column indicates whether NA12878 is identifiable among 1000 Genomes samples, represented by green check for yes and red cross for no. The third column is a schematic representation of the signal profiles for each combination.

Figure 5: Representation of the linking attack that utilizes Hi-C interaction matrix data. a) Schematic representation of how genome-wide signal profile is computed from the interaction matrix. Each column i of the matrix is summed along the rows and the total value is recorded at the i^{th} entry of the signal profile. b) Table shows whether NA12878 is vulnerable when different resolutions of the interaction matrix is used in linking. Green check indicates that NA12878 is vulnerable while red cross indicates not vulnerable.

Figure S1: Illustration of Netflix Prize competition and linking to IMDb. a) Netflix released an anonymized training dataset that contained the movie identifiers, ratings, dates of ratings, and anonymized user identifiers. This dataset contained more than 100 million ratings for 500,000 users where each user had rated on average 200 movies and each movie was rated on average by 5,000 users. b) The training dataset was linked to the Internet Movie Database (IMDb)'s database. The linking is based on matching the movie rating, the date of rating and other features in the databases. For the individuals whose names can be found in the IMDb database, the movie ratings are made public.

Figure S2: The scatter plot of sample-wide predictability versus ICI leakage of the SV genotypes when gene expressions are used to genotype SVs. Each dot represents a 1000 Genomes SV and the population-wide predictability represents how correctly predictable the SV genotypes are given the gene expression levels.

The expression levels are obtained from GEUVADIS dataset. The ellipses point to the small number SVs that have high predictability and high ICI leakage.

Figure S3: Feature set that are used to genotype and discover deletions. A candidate deletion is between i and j indices. The attacker uses the signal profiles within the deletion region and the left and right neighboring regions ($[2i - j - 1, i - 1]$ and $\lfloor j + 1,\ 2j - i + 1 \rfloor$, respectively) to compute the features. $\rho_{[i,j]}$ represents how deep the dip is in the signal profile along the deletion. $\eta_{[i,j]}$ represents how balanced the signal levels in the neighboring regions are. $\tau_{[i,j]}$ represents how high the signal levels are in the neighboring regions.

Figure S4: Accuracy of genotype predictions that are used in instantiating the linking attacks. The x-axis shows the number of variants used and y-axis shows the genotype accuracy. The GEUVADIS signal profiles are used with known panel of 1000 Genomes small indels.

Figure S5: A screenshot of the region surrounding the 2 base pair indel rs24043625 in GTex RNA-seq signal profile hub on UCSC Genome Browser. The figure shows the profiles for 6 individuals. The top figure shows 6 kb region around the deletion. The botton figure shows a zoomed version around 100 base pair of the deletion. The dip in the RNA-seq signal that is caused by the deletion can be easily seen by eye in 3 individuals XV7Q, 14BMU, 139D8. Other individuals shown in the figure do not have this deletion.

Figure S6: [[TO BE ADDED]]

REFERENCES

1. Joly Y, Dyke SOM, Knoppers BM, Pastinen T: **Are Data Sharing and Privacy Protection Mutually Exclusive?** *Cell* 2016:1150–1154.

2. Singer DS, Jacks T, Jaffee E: **A U.S. "Cancer Moonshot" to accelerate cancer research.** *Science* 2016, **353**:1105–6.

3. Collins FS: **A New Initiative on Precision Medicine**. *N Engl J Med* 2015, **372**:793–795.

4. Handelsman J: **The Precision Medicine Initiative**. *White House, Off Press Secr* 2015:1–5.

5. Caulfield M, Davies J, Dennys M, Elbahy L, Fowler T, Hill S, Hubbard T, Jostins L, Maltby N, Mahon-Pearson J, McVean G, Nevin-Ridley K, Parker M, Parry V, Rendon A, Riley L, Turnbull C, Woods K: **The 100,000 Genomes Project Protocol**. *Genomics Engl* 2015(February).

6. **Briefing- Genomics England and the 100K Genome Project** [http://www.genomicsengland.co.uk/briefing/]

7. Feero WG, Guttmacher AE, Feero WG, Guttmacher AE, Collins FS: **Genomic Medicine — An Updated Primer**. *N Engl J Med* 2010, **362**:2001–2011.

8. Joly Y, Feze IN, Song L, Knoppers BM: **Comparative Approaches to Genetic Discrimination: Chasing Shadows?** *Trends Genet* 2017, **33**:299–302.

9. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson J V., 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**.

10. 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**:591–598.

11. Dwork C: **Differential privacy**. *Int Colloq Autom Lang Program* 2006, **4052**:1–12.

12. Vaikuntanathan V: **Computing Blindfolded: New Developments in Fully Homomorphic Encryption**. *2011 IEEE 52nd Annu Symp Found Comput Sci* 2011:5–16.

13. Fienberg SE, Slavković A, Uhler C: **Privacy preserving GWAS data sharing**. In *Proceedings - IEEE International Conference on Data Mining, ICDM*; 2011:628–635.

14. Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB: **The real cost of sequencing: higher than you think!** *Genome Biology* 2011:125.

15. Narayanan A, Shmatikov V: **Robust de-anonymization of large sparse datasets**. In *Proceedings - IEEE Symposium on Security and Privacy*; 2008:111–125.

16. Harmanci A, Gerstein M: **Quantification of private information leakage from phenotype-genotype data: linking attacks.** *Nat Methods* 2016, **13**:251–256.

17. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: **Identifying personal genomes by surname inference.** *Science* 2013, **339**:321–4.

18. Schadt EE, Woo S, Hao K: **Bayesian method to predict individual SNP genotypes from gene expression data**. *Nature Genetics* 2012:603–608.

19. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies.** *Nat Methods* 2009, **6**:S22–S32.

20. Backes M, Berrang P, Bieg M, Eils R, Herrmann C, Humbert M, Lehmann I: **Identifying Personal DNA Methylation Profiles by Genotype Inference**. In *Proceedings - IEEE Symposium on Security and Privacy*; 2017:957–976.

21. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Jasmine Mu X, Alkan C, Antaki D, et al.: **An integrated map of structural variation in 2,504 human genomes**. *Nature* 2015, **526**:75–81.

22. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315**:848–853.

23. The 1000 Genomes Project Consortium: **An integrated map of genetic variation**. *Nature* 2012, **135**:0–9.

24. The 1000 Genomes Project Consortium: **A global reference for human genetic variation**. *Nature* 2015:68–74.

25. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.

26. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G: **Epigenomics: Roadmap for regulation**. *Nature* 2015, **518**:314–316.

Deleted: Backes M, Berrang P, Bieg M, Eils R, Herrmann C, Humbert M, Lehmann I: **Identifying Personal DNA Methylation Profiles by Genotype Inference**. In *Proceedings - IEEE Symposium on Security and Privacy*; 2017:957–976

Deleted: 20.

Deleted: 21

Deleted: 22

Deleted: 23

Deleted: 24.

Moved down [2]: Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57–63.¶

Deleted: 25. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies.** *Nat Methods* 2009, **6**:S22–S32.¶
26.

Deleted: 27

27. Consortium TG: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45**:580–5.

28. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, et al.: **The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans**. *Science (80-)* 2015, **348**:648–660.

29. Abyzov A, Urban AE, Snyder M, Gerstein M: **CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing**. *Genome Res* 2011, **21**:974–984.

30. Handsaker RE, Korn JM, Nemesh J, McCarroll SA: **Discovery and genotyping of genome structural polymorphism by sequencing on a population scale**. *Nat Genet* 2011, **43**:269–276.

31. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57–63.

32. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK: **Identification of genetic variants that affect histone modifications in human cells.** *Sci (New York, NY)* 2013, **342**:747–749.

33. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, Yurovsky A, Lappalainen T, Romano-Palumbo L, Planchon A, Bielser D, Bryois J, Padioleau I, Udin G, Thurnheer S, Hacker D, Core LJ, Lis JT, Hernandez N, Reymond A, Deplancke B, Dermitzakis ET: **Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription.** *Science* 2013, **342**:744–7.

34. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek D V, Li J, Xie D, Olarerin-George A, Steinmetz LM, Hogenesch JB, Kellis M, Batzoglou S, Snyder M: **Extensive variation in chromatin states across humans.** *Science (New York, NY)* 2013:750–752.

35. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES: **Hi-C: a method to study the three-dimensional architecture of genomes.** *J Vis Exp* 2010, **6**:1869.

36. Rao SSP, 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**:1665–1680.

37. Korbel JO, Lee C: **Genome assembly and haplotyping with Hi-C.** *Nat Biotech* 2013, **31**:1099–1101.

38. Euskirchen GM, Rozowsky JS, Wei CL, Wah HL, Zhang ZD, Hartman S, Emanuelsson O, Stolc V, Weissman S, Gerstein MB, Ruan Y, Snyder M: **Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies**. *Genome Res* 2007, **17**:898–909.

39. Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M: **Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping**. *Trends in Genetics* 2005:466–475.

40. Berger B, Peng J, Singh M: **Computational solutions for omics data**. *Nat Rev Genet* 2013, **14**:333–346.

Deleted: 28

Deleted: 29

Deleted: 30

Deleted: 31

Moved (insertion) [2]

41. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M: **RSEQtools: A modular framework to analyze RNA-Seq data using compact, anonymized data summaries**. *Bioinformatics* 2011, **27**:281–283.

42. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012:357–359.

43. Harmanci A, Rozowsky J, Gerstein M: **MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework.** *Genome Biol* 2014, **15**:474.

44. Chan RH, Ho C-W, Nikolova M: **Salt-and-Pepper noise removal by median-type noise detectors and detail-preserving regularization.** *IEEE Trans Image Process* 2005, **14**:1479–1485.

45. Wang ZWZ, Zhang D: **Progressive switching median filter for the removal of impulsenoise from highly corrupted images**. *IEEE Trans Circuits Syst II Analog Digit Signal Process* 1999, **46**.