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

Arif Harmanci[1,2,4,*], Mark Gerstein[1,2,3,*]

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 Department of Computer Science, Yale University, New Haven, CT 06520, USA
4 School of Biomedical Informatics, Center for Precision Health, University of Texas Health Sciences Center, Houston, TX, 77030, USA
*Corresponding authors: Arif Harmanci (arif.harmanci@yale.edu) and Mark Gerstein (pi@gersteinlab.org)

Abstract

Functional genomics experiments such as RNA sequencing are performed to reveal gene expression changes under different conditions and diseases. Although the main purpose of these approaches is to understand ~~the~~ dynamic changes in gene expression levels, the data also contain a large number of genetic variants in the raw reads. Therefore, the raw reads cannot be shared publicly because of privacy concerns. There is, however, great desire to publicly share as much of the data as possible since they are extremely valuable for biomedical and disease research. To enable safe data sharing, researchers often use aggregated representations computed from raw reads, such as read depth signal profiles and gene expression quantifications. These representations do not explicitly reveal variant information and are generally assumed to be safe to share. Here, we study 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 to breach 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 small and large 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 the protection of signal profiles against the presented genotype prediction based attacks. Given that several consortia, such as the GTEx and ENCODE, publicly share signal profiles, these results point to a critical source of sensitive information leakage.

INFORM-THEORETIC

PGT.

IN THE FRAMEWORK OF LINKING

1. Introduction

Individual privacy is emerging as an important aspect of biomedical data science[1]. A deluge of genetic data will be generated from hundreds of thousands, if not millions, of individuals through the Cancer Moonshot Project[2], the Precision Medicine Initiative[3,4], and the 100K Genome Project [5,6]. Moreover, there is much effort to make genetic data more prevalent in the clinical setting[7]. The leakage of genetic

Deleted: 260 Whitney Avenue,

Deleted: 260 Whitney Avenue,

Deleted: 260 Whitney Avenue,

Deleted:),

Deleted: -

Deleted: the how

Formatted ...

Deleted: data

Formatted ...

Formatted ...

Formatted ...

Deleted: in

Formatted ...

Deleted: .

Deleted: data

Formatted ...

Deleted: understanding

Formatted ...

Formatted ...

Deleted: level

Formatted ...

Deleted: While it is known that

Formatted ...

Deleted: , it

Formatted ...

Deleted: generally assumed safe

Formatted ...

Deleted: other data

Formatted ...

Deleted: of functional genomics data

Deleted: quantification of

Formatted ...

Formatted ...

Deleted: levels.

Formatted ...

Deleted: focus on

Formatted ...

Deleted:

Formatted ...

Deleted: for breaching

Formatted ...

Deleted: and small

Formatted ...

Deleted: for example GTex

Formatted ...

Formatted ...

Formatted ...

...

information creates many privacy concerns; for example, genetic predisposition to diseases may bias insurance companies[8].

Initial studies on genomic privacy focused primarily on protecting the identities of participants in the early genetic profiling and genotype-phenotype association studies[9,10]. These studies 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 mixture. In this arena, differential privacy[11] has been proposed as a theoretically optimal formalism that can fulfill privacy requirements such that the probability that any individual's participation can be identified can be made arbitrarily small. In addition, cryptographic approaches have proven useful for privacy-aware analyses 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. In addition, it has made genomic data more available to hospitals, research institutes, and individuals[14]. This increase will render new types of attacks more practical where an adversary can use statistical methods to link multiple datasets to reveal sensitive information. These attacks are termed linking attacks[15–17]. Briefly, 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 cross-referencing is that sensitive information from one dataset becomes linked to the identifying information in another; in turn, this breaches the privacy of individuals whose sensitive information is revealed. The risks behind linking attacks have risen recently because personal data is generated at an exceedingly high pace and because these data are independently released and maintained. For this reason, 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; that is, a dataset that seems safe to release now may become vulnerable to a linking attack next year.

A well-known example of a linking attack is the Netflix Prize Competition[15] (Supplementary Fig. 1a,b). In this competition, the movie rental company Netflix released a dataset that 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 could 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. Therefore, these breaches must be examined to enable privacy-aware data sharing approaches.

In this study, we analyzed the leakage of sensitive information from functional genomics data and how an adversary could use it in linking attacks. Functional genomics data, such as those from RNA sequencing (RNA-Seq), is unique in that if the data comes from human subjects the raw reads have genetic variant information. This information could be used to identify individuals (Supplementary Fig. 2b). However, the main purpose of RNA-Seq data is not related to the variants, but rather understanding how the activity of genes changes under different conditions such as cancer. Thus, unlike the variant data, functional genomics datasets have a more complicated "Yin-Yang" aspect with relation to privacy. In addition, functional genomics datasets are sometimes shared with phenotypic information that is

& IN THE FOLLOWING SENSE.

Deleted: attacks

Deleted: [15] (Supplementary Fig 1a,b). In this competition, a training dataset was released by the movie rental company Netflix, which

Deleted: can

Deleted: and

Deleted: studied well

Deleted: analyze

Deleted: the

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Deleted: they can be used by

Formatted: Highlight

Formatted: Highlight

Deleted: There are a number of motivating key points related to functional genomics data and privacy. First of all, functional

Formatted: Highlight

Formatted: Highlight

Deleted: data,

Formatted: Highlight

Deleted: ,

Formatted: Highlight

Deleted: ,

Formatted: Highlight

Deleted: , which may

Formatted: Highlight

Deleted: .

Formatted: Highlight

Deleted: ; main purpose is more related to

Formatted: Highlight

Deleted: dysregulation

Deleted: ,

Formatted: Highlight

Deleted: Consequently, there's a great

Formatted: Highlight

Formatted: Highlight

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 determine general aspects about a condition. However, the existence of small amounts of residual private information in the data potentially can be revealing about the individual from which they came.

Another important factor is the desire to share and study RNA-Seq datasets to help find cures for various diseases. Because of this, there is great incentive to find ways to share functional genomics data without privacy protections. Large-scale privacy protections are an encumbrance on genomic data sharing. These protections do not allow researchers and data owners to share results on the web or use web- or internet-based tools, ~~exerting a great burden on research~~. Consequently, many consortia, such as the Genotype-Tissue Expression (GTEx) Project, aim to share RNA-Seq information to the maximum extent. Although the raw reads cannot be shared, there is a general belief that other aggregated data computed using raw reads, such as signal profiles and gene-level quantifications, can be shared. Signal profiles simply reflect the overall depth of coverage of the RNA-Seq reads at any given position on the genome. These profiles are computed by counting the number of reads that overlap with each position on the genome (Supplementary Fig. 2a). The profiles ostensibly do not contain variant information. This is why many genomics consortia have decided to openly share RNA-Seq signal profiles in bigwig and wig files. In this study, we focused on leakage from signal profiles. Another commonly shared aggregated data are gene-level quantifications, which are essentially averages of the signal profile over exons. Although overall aggregation and averaging reduces information, private information leakage also decreases. However, private information leakage still occurs from gene expression quantifications through the association of expression levels with variants called expression quantitative trait loci (eQTLs). Although we do not tackle this in the current study, it has been explored elsewhere[16,18].

Several studies have examined aspects of linking attacks in genomic privacy[16–19]. One aspect of genomic privacy that has not yet been addressed extensively is how structural variants (SVs), such as deletions, insertions, and translocations of large chunks of DNA sequences, might affect genomic privacy[20]. In this study, we explored whether an adversary could use signal profiles of functional genomics signals to detect and genotype genomic deletions and use them to pinpoint individuals in a large genotype dataset in a linking attack. Most previous studies on genomic privacy focus on 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]. However, the major portion of genomic variation, in terms of the number of nucleotides that are affected, is caused by SVs[22,23], as shown by The 1,000 Genomes Project. Since an SV affects a much larger portion of the genome than a SNP, we expect a phenotype caused by an SV to be very obvious. For example, homozygous deletion of a gene will cause the total disappearance of its expression.

In this study, we analyzed sensitive information leakage from signal profiles of several sequencing-based functional genomics datasets. Signal profiles are currently at the junction between public and private information, and where genomic information has begun to be shared publicly. Hence, it is particularly important to probe the leakage from the signal profile representation of functional genomics data. It might be the case that this type of information will not be publicly shareable at all in the future. We emphasize that in this paper we are not trying to look at all sources of leakage from functional genomics data, but just the sources right at the decision boundary of sharing and not sharing.

INVERT

THESE STUDIES HAVE EXCLUSIVELY FOCUSSED ON SNVS( )

Deleted: , to enable helping

Formatted ...

Formatted ...

Deleted: make

Formatted ...

Deleted: of sharing

Formatted ...

Deleted: a great

Formatted ...

Deleted: They

Formatted ...

Deleted: ,

Formatted ...

Deleted: and

Deleted: and they exert

Formatted ...

Formatted ...

Deleted: at sharing

Formatted ...

Deleted: The

Formatted ...

Deleted: obviously

Formatted ...

Deleted: as they contain variant information. However, ...

Formatted ...

Deleted: the

Formatted ...

Deleted: files

Formatted ...

Deleted: the

Deleted: The signal files

Formatted ...

Formatted ...

Deleted: point. Ostensibly, they're

Formatted ...

Deleted: Many of the

Formatted ...

Deleted: information. We show that there is a high degre ...

Formatted ...

Deleted: in the function genomics

Formatted ...

Deleted: profile

Formatted ...

Deleted: . The

Formatted ...

Formatted ...

Formatted ...

Formatted ...

THUS, EASY

As we introduced earlier, the raw reads from an RNA-Seq experiment contain the nucleotides themselves. We assume that the data owners created the signal profiles and made them publicly available. Several large consortia, for example the Encyclopedia of DNA Elements (ENCODE) project[24], the Roadmap Epigenome Mapping Consortium[25], and GTEx[26,27], publicly share signal profiles (Supplementary Fig. 3). An adversary can ~~then~~ gain access to the signal profiles. Although signal profiles do not contain any explicit variant information at the nucleotide level (such as SNPs), signal processing techniques can be utilized to detect small and large genomic deletions. Two quantities can determine how well genomic deletions can be detected from sequencing data. The first is breadth of coverage, which measures how well the genome is covered by signal profiles. The second is depth OF coverage, which measures how deep the sequencing is performed. DNA-sequencing read depth signal[28,29] is very suitable for the 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, the 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[30] 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 but low-breadth coverage. This makes RNA-Seq signal profiles very suitable for detecting small deletions in exonic regions. We show that a large number of small deletions can be detected using RNA-Seq signal profiles. Chromatin immunoprecipitation sequencing (ChIP-Seq)[31] signal profiles for diffuse histone modifications (SUCH AS...) generally have high-breadth 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, pooling the signal profiles increases both the depth and breadth of coverage and can bring enough power to an adversary for genotyping large deletions and performing successful linking attacks.

THIS IS SUITABLE FOR BIG SVS...

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 utilize different practical methods for deletion variant genotyping. Finally, we focus on protecting the signal profiles against linking attacks. We present a novel signal processing methodology for anonymizing a signal profile. We show that this method 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://privasig.gersteinlab.org.

2. Results

2.1. Linking Attack Scenario

Figure 1 summarizes the linking attack scenario that we are studying. 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 contains a genome-wide signal profile and an anonymized identifier for each individual. The signal profile for an individual represents the measurements of functional activity at each genomic position. In addition, the signal profile dataset contains sensitive information about each individual (e.g., HIV status). We assume that the dataset is generated for research purposes and is publicly released. The genotype dataset, G, contains, for each

Deleted: utilizes

Deleted: protection of

Deleted: it

Deleted: archive

Deleted: /proj/PrivaSig

Deleted: it

Deleted: , for each individual,

Deleted: ,

Deleted: for this individual.

Deleted: ,

Deleted: .

Deleted: .

Deleted: this

individual, the genotypes of 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 the adversary might have stolen it or might have been legally allowed to access it but violated 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, and then matching the predicted genotypes to the genotypes in G. For any matching individuals in G and S, the name and sensitive information are revealed to the adversary.

The attack has two steps. The first step is genotyping the deletion variants, which is illustrated in Figure 1a. The adversary has access to a genome-wide signal profile dataset (S) for a sample of individuals. This dataset stores a genome-wide signal profile for each individual, for example containing 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, the adversary utilizes the signal profile and genotypes the deletions in p_S. After genotyping, the adversary builds a data matrix with the predicted genotypes, which is denoted by $\tilde{G}$. We refer to this scenario as linking based on "genotyping only". The second scenario, also illustrated in Figure 1a, is very similar except that the adversary does not have access to, but discovers the panel of structural variants from the signal profiles. The adversary 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 genotyping, the genotyped SV matrix ($\tilde{G}$) includes, for each individual, the predicted SV genotypes and the sensitive information (e.g., 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 the individuals in the genotyped SVs ($\tilde{G}$) and the individuals in the genotype dataset, G, illustrated in Figure 1b. The SV genotype dataset G is assumed to contain identifying information about the individuals. Thus, we assume that this dataset was previously protected but was either leaked or stolen (e.g., variants from a glass). The adversary first compares the genotyped SV panel (p_S) to the SV panel of the genotype dataset, which is denoted by p_G. After matching the SVs in the two panels, the adversary compares the genotypes of the matching SVs in the two panels. The adversary uses this comparison to cross-reference the individuals in two datasets and find the individuals that best match each other with respect to genotype match distance (i.e., links 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 SV Genotypes

In order to assess the correct predictability of SV genotypes, we propose using genome-wide predictability of SV genotypes, denoted by π_{GW}, from signal profiles. 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, predictability only depends on the genomic signal levels of an individual and how well they can be used to predict genotypes. For example, Figure 1c illustrates a large deletion that can be easily predictable using histone modification signal profiles. In principle, genome-wide predictability is computed for each individual independently. Therefore, the

Deleted: for

Deleted: she (we assume …he adversary is a female) …ig

Deleted: the …ensitive information, i.e. HIV status,

Deleted: of …he deletion variants, which is illustrated in

Deleted: , for each individual, …a genome-wide signal

Deleted: she

Deleted: the

Deleted: , where the adversary has access to a reference panel of structural variants,…as linking based on

Formatted: Font: Italic

Deleted: -

Formatted: Font: Italic

Deleted: …discovered SV panel. We refer to this scenari

Deleted: ,…and also …he sensitive information about…e.

Deleted: of

Deleted: Fig

Deleted: individual's identities.…he individuals. Thus, we

Deleted: the …atching of …he SVs in the two panels,

Deleted: Structural Variant

Deleted: a measure named

Deleted: The predictability…redictability measures how genome-wide predictability of a variant from a signal profile is independent from the population frequency of the variant.

Deleted: of

Other than predictability, an important measure in linking attacks is the information content each SV genotype supplies. We utilized 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 genotypes that have low population frequency. As the genome-wide predictability is independent of the population frequency of the variants, the adversary can utilize genome-wide prediction approaches and predict rare variant genotypes to gain high ICI and characterize individuals accurately.

Deleted: the ...redictability, an important measure in the ...

2.3. Linking Attacks using RNA-Seq Signal Profiles

We first focused on the predictability of small deletions using RNA-Seq signal profiles. Figure 1d illustrates a hypothetical example of how small deletions in RNA-Seq signal profiles can be detected as small and sudden dips in the signal. As an example showing the relevance of small deletions in RNA-Seq signal profiles, we include a screenshot of signal profiles around a small deletion for six individuals in the GTEx Project (Supplementary Fig. 3). The two base pair deletion, rs34043625, can be easily detected for three of the individuals shown. An important aspect of the effect of small deletions on the signal profile is the extent to which they affect the total expression of a gene. It is clear from Supplementary Figure 3 that the total signal in the small dips in the RNA-Seq signal is much smaller than the perturbations caused by other genetic factors like eQTLs and splicing QTLs. In general, an eQTL is associated with a global change in the total signal on a RNA-Seq signal profile of a gene. However, a small deletion affects a localized position on the RNA-Seq signal profile with a relatively smaller effect on the total expression of the gene, assuming the small deletion is not an eQTL.

ANONYMIZE

Deleted: focus...ocused on the predictability of short...m ...

As mentioned above, RNA-Seq signal profiles generally have high-depth but low-breadth coverage. Such signal profiles can most easily detect small deletions under 10 base pairs. In this case, each small 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. The genome-wide predictability (π_{GW}) of small deletions quantifies how well an adversary can identify the dips corresponding to deletions in the signal profile (Methods Section). We first estimated the genome-wide predictability for the panel of small deletions in 1,000 Genomes Project using the RNA-Seq expression signal profiles from the GEUVADIS project. Figures 2a and b show π_{GW} versus ICI for small deletions. A substantial number of deletions has much higher predictability compared to a dataset where the signal profile is randomized with respect to the location of deletions. In addition, many variants have very high ICI (on the order of 5-6 bits) with high predictability (greater than 80%). This result shows that the signal profile-based attack scenario is much more powerful that other approaches like population-wide prediction of variant genotypes (Supplementary Fig. 4).

Deleted: we...mentioned earlier, the...bove, RNA-seq...e ...

Deleted: the ...mall deletions quantifies how well the...n ...

Deleted: vs...ersus ICI for short...mall deletions. There is ...

In order to present the 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 the discovery and genotyping of small deletions. As mentioned above, the genotyping of deletions is based on detecting abrupt dips in the signal profile. In order to detect these dips, the adversary utilizes a quantity we term the *self-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,

Deleted: the ...mall deletions. As mentioned before...bov ...

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

The genomic regions with low $\rho_{[i,j]}$ values point to intervals that tend to have dips in them. For each individual, the prediction method sorts the small deletions with respect to the *self-to-neighbor signal ratio* and assigns a 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 are closest to the predicted genotypes. Using this genotyping strategy, we simulated an attack to link the RNA-seq signal profile dataset from the GEUVADIS Project[32] to the 1000 Genomes Project genotype dataset. In the *genotyping only* scenario, the linking is perfectly accurate when the adversary utilizes more than 40 deletions (Fig. 2c). In a 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 would be if the adversary used deletions of different lengths. Figure 2e shows the accuracy and number of insertions and deletions (indels) with different lengths. The accuracy of linking decreases substantially for indels that are longer than five base pairs. The decrease in accuracy is affected by both the decrease in the number of indels (i.e., low ICI), shown in Figure 2e, and the decreasing predictability of indels whose lengths are above 5 base pairs. We then determined the minimum number of indels that is sufficient to accurately link an individual. We evaluated the distribution of the minimum number of indels for accurately linking each individual in the GEUVADIS dataset for genotyping only (Fig. 2f) and for adversary jointly discovery and genotyping (Fig. 2g) scenarios. As small as 30 indels are sufficient to correctly link a large fraction of individuals.

In the previous analysis, the sample set used for discovery of the deletion panel and the RNA-Seq sample set were matching (i.e., 1000 Genomes individuals). This may cause a bias in linking because the SV genotype dataset may contain rare deletions that are also in the panel of deletions. This would make it trivial to link some of the individuals. To get around this bias, we studied linking attacks where the signal profile dataset was generated by the GTEx Project[26,27] and the panel of small deletions was generated by the 1000 Genomes Project, thus utilizing a non-matching set of individuals. 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. With this setup, we first computed π_{GW} versus ICI for the deletions and observed substantial enrichment of deletions with high predictability and high ICI compared to randomized datasets (Fig. 3a, b). As before, many deletions were highly predictable (>80%) and very high in ICI (>5bits). In addition there is a substantial increase of predictability when compared to the randomized dataset.

We next instantiated the linking attack using an extremity-based approach. We first evaluated the attack based on a *genotyping only* scenario. In this scenario, the linking accuracy was close to 100% using a relatively small number of variants (30 variants) (Fig. 3c). An interesting observation was that when the attacker increased the number of variants used in the attack, the linking accuracy decreased. This was caused by the fact that the additional variants beyond the initial 30 variants were incorrectly genotyped, decreasing the accuracy of linking. Thus, the additional variants act as noise and decrease the linking accuracy.

One question that arises following these results is whether the adversary can assign reliability scores to the linked individuals. We tested whether the *first distance gap* (Methods Section) measure is suitable

Deleted: 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.¶ ...

Deleted: the

Deleted:

Deleted: In the instantiation, we

Deleted: is

Deleted: , i.e.,

Deleted: is

Deleted: increases

Deleted: decreases

Deleted: is

Deleted: after

Deleted: are

Deleted: and decrease

Deleted: In simple terms

Deleted: Following this, one

Deleted: is if

for evaluating the reliability of linking. 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 a linking attack is low. As a test case, we focused on a linking where the adversary used 200 deletions with an overall linking accuracy of 35% (Fig. 3c). Figure 3d shows the sensitivity and positive predictive value (PPV) with a changing *first distance gap* metric. The adversary could link 10% of the individuals perfectly, and 20% with around 90% accuracy. Figure 3d also shows the average sensitivity and average PPV over 100 random selections of the linkings. As expected, the PPV was always around 35% and average sensitivity was always smaller than for a *first distance gap*-based selection of linkings. These results show that even though some parameter selections (e.g., number of variants) may show low accuracy, the adversary can increase accuracy by selecting the linkings using the first distance gap measure.

Deleted: linkings

Deleted: ,

Deleted: .

Deleted: %,

Deleted: the

Deleted: uses

Deleted: where the

Deleted: is

Deleted: can

Deleted: of the individuals are linked

Deleted: , i.e., makes 1 mistake in 10 linkings

Deleted: is

Deleted: is also

Deleted:

2.4. Linking Attacks using ChIP-Seq Signal Profiles

WHERE THERE is enough cov.

We next focused on predictability versus ICI of deletions over 1,000 base pairs. Since the deletions are large, the signal profiles that are suitable for genotyping them must have high-breadth coverage. We utilized 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[33–35]. These studies aimed at revealing how genetic variation interacts with epigenomic signals, mainly 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. Although we are focusing on the predictability of large deletion genotypes from ChIP-Seq profiles, this does not mean that 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.

Deleted: 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. ¶ ...

We used these personalized epigenomic signal profiles to quantify how much characterizing information leakage they provide. For any individual in which there were multiple histone-marked ChIP-Seq signals, we pooled the signal profiles and computed several features for each large deletion. These were then used for quantifying information leakage (Methods Section). First, we computed π_{GW} versus ICI using the panel of large deletions in 1,000 Genomes Project. Figure 4a and 4b show π_{GW} versus ICI for the large deletions in the 1000 Genomes Project. We used the personal epigenome profiling ChIP-Seq datasets presented in studies by Kasowski et al.[35] and Kilpinen et al.[34] (Methods Section). Similar to our small deletion analysis, for both datasets there were many large deletions with high predictability and high ICI.

We next focused on instantiating linking attacks using ChIP-Seq profiles. We again utilized a variant of the outlier-based genotyping in the linking attack. The genotyping of the panel of large deletions was performed as follows. The average ChIP-Seq signal on each deletion was computed and the variants were sorted with respect to their average signal in increasing order. The deletions with smallest ChIP-Seq signal were assigned a homozygous deletion genotype. For the deletions with assigned genotypes, we identified the individual in the genotype dataset (from the 1000 Genomes project) whose genotype matched closest to the assigned genotypes. We repeated this linking attack with a different number of windows and computed the accuracy of linking (Methods Section). Figure 4c shows the accuracy of a linking attack based on the *genotyping only* scenario, where the adversary was assumed to have access to the large deletion panel from 1000 Genomes. The linking accuracy reached 100% with a fairly small

Deleted: utilize

Deleted:

Deleted: is done

Deleted: is

Deleted: are

Deleted: are

Deleted: genotypes match

Deleted: is

Deleted: reaches

number of deletions for both datasets. For the *joint discovery and genotyping* scenario where the adversary first discovered deletions and then genotyped them, the accuracy was also very high with a 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[24]. We evaluated whether different combinations of histone modifications render NA12878 vulnerable against a linking attack among 1000 Genomes individuals, as illustrated in Figure 4e. In general, we 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 a histone marks cover a larger genomic region, a larger 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 render NA12878 vulnerable. In addition, H3K9me3, a repressive mark that expands very broad genomic regions, renders NA12878 vulnerable in several combinations with other marks. By contrast, H3K27ac, an activating histone mark that spans punctate regions, does not render NA12878 vulnerable.

2.5. Linking Attacks using Hi-C Matrices

We next tested whether a relatively new data type, Hi-C interaction matrices, can be used for the identification of genomic deletions. Hi-C is a high-throughput method for identifying long-range genomic interactions and three-dimensional chromatin structure[36]. Hi-C is based on proximity ligation of 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 the NA12878 individual for whom Hi-C interaction matrices were generated at different resolutions[37]. 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). In this way, we are simplifying the multidimensional nature of the Hi-C contact matrix and treating it as a sequencing assay that spans the entire genome. It is important to emphasize that the standard analysis of Hi-C matrices does not involve such a signal profile generation. We are using this step to convert the Hi-C matrix into a signal 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 the top 1000 windows with a homozygous deletion genotype. We next compared the predicted genotypes with all the genotypes in the 1,000 Genomes project. We observed that NA12878 was vulnerable to attack when the Hi-C contact matrix resolution (bin length) was 10 kilobases or smaller (Fig. 5b).

It is important to clarify that we are focusing on using the final output of Hi-C data, that is 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 a Hi-C experiment, which is a separate issue[38]. This would require 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

Deleted: discovers

Deleted: genotypes

Deleted: is

Deleted: Project[25]. We have

Deleted: which is

Deleted: Fig

Deleted: have

Deleted: higher

Deleted: they

Deleted: On the other hand

Deleted: do

Deleted: also asked

Deleted:

Deleted: the

Deleted:

Deleted:

Deleted: [35]. It

Deleted: the

Deleted:

Deleted: are

Deleted: [36].

Deleted: This

Deleted: treat

Deleted: do

Deleted:

Deleted: :

Deleted: 1000

Deleted: is

Deleted: this

Deleted: is

Deleted: i.e.,

Deleted: Hi-C experiment, which is a different problem by itself[37]. It also requires of sequencing data and uses the linear genomic signal profile to identify deletions for the purpose of linking datasets. We highlight the fact that Hi-C interaction matrices themselves may leak substantial amounts of characterizing information.

Deleted: are highlighting…ighlight the fact that Hi-C

2.6. Anonymization of RNA-Seq Signal Profiles

An important aspect of genomic privacy is risk management and the protection of datasets. Anonymization of the datasets is the most effective way to ensure safe protection when sharing data publicly. Personal RNA-Seq datasets are currently by far the most abundant functional genomic datasets. For example, RNA-Seq signal profiles are being publicly shared from the GTEx project, although 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-Seq. Altogether, these factors make the protection of RNA-Seq data urgent. The most effective way to protect against a linking attack scenario is to ensure that deletion genotypes cannot be inferred from signal profiles. As we showed in the previous sections, small deletions are a major source of leakage of genetic information from RNA-Seq signal profiles. We propose systematically removing the dips in signal profiles as a way to anonymize the profiles against the prediction of small deletions. Specifically, we propose smoothing the signal profiles using median filtering locally around a given panel of deletions (Methods Section). We observed that median filtering removes the dips in the signal 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 and GTEx Project consortia. After applying signal profile anonymization, the large fraction of the leakage was removed for the GTEx datasets (Fig. 2b and 3b). We also observed that the extremity-based linking attack proposed in the previous section was ineffective in characterizing individuals such that no individuals were vulnerable for the GTEx project and only 1% of the individuals were vulnerable for the GEUVADIS dataset. Importantly, this procedure can be used for anonymizing not only RNA-Seq signal profiles but also other signal profiles against attacks based on small deletion genotyping. However, the anonymization is not as effective for large deletions. This is not a major concern for RNA-Seq signal profiles, as we observed that large deletions were not easily genotyped using RNA-Seq data. However, as we showed in the previous section, linking attacks can be successful when they use large deletions that are genotyped using ChIP-Seq datasets.

Deleted: the …enomic privacy is risk management and th

3. Discussion

Sequencing-based functional genomics assays provide a large amount of biological information for understanding the dynamic nature of gene activity and epigenetic regulation. This information is extremely valuable for understanding genetic mechanisms behind disease initiation and progression. Thus, data producers and owners want to share these data as openly as possible. At the same time, genomic data can contain variant genotype information within the raw reads that may cause concerns for privacy. These two competing factors, the incentive to share and privacy concerns, make it necessary to carefully evaluate the sharing mechanisms of functional genomics data. To decrease genetic variant leakage in sequencing data, aggregate data formats have been widely used. Two examples are signal profiles and gene expression quantifications. Unlike raw reads, these data do not immediately reveal variant information and are generally accepted to be safe for public data sharing. However, gene expression levels have been shown to leak enough genotype data to be used in accurate linking attacks[16,18]. In this study, we evaluated the possible privacy concerns around sharing signal profiles.

Deleted: 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 systematically analyzed a critical source of sensitive information leakage from signal profile datasets, which were previously thought to be largely secure. Our results show that an adversary can perform very accurate linking attacks for characterizing individuals by the genotyping 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 an individual.

Although we focused 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, linking attacks can be easily carried out on DNA sequencing signal profiles. In addition, signal profiles from genome-wide profiling techniques other than sequencing-based assays, like ChIP and expression tiling arrays[39,40], can be vulnerable to the linking attack scenario that we presented. Different genome-wide data representations (e.g., Hi-C interaction matrices) can be utilized for the generation of genome-wide signal profiles; these can in turn leak sensitive information. We believe that many more genome-wide omics technologies will be developed in the near future[41]. 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 our focus is on small and large deletion variants, the analyses of predictability and practical linking attacks can be extended for other structural variants such as genomic insertions.

We showed that linking can be done by predicting a fairly small number of variants (generally less than 100). 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 RNA-Seq signal profiles are publicly available through the University of California, Santa Cruz (UCSC) Genome Browser. Given the extent of public sharing of datasets, we believe that the anonymization of RNA-Seq signal profiles using the signal processing technique that we proposed is very useful. Our method applies 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[42] and tagAlign[24] can enable removing raw sequence information from reads while keeping the information about read mapping intact. We note that the anonymization method that we presented does not close all sources of leakage. The procedure aims to close the leakages caused by the genotyping of genomic deletions using the dips in the signal profile. These leakages are accessible to an adversary and can be detected directly from the signal profiles. Thus, we believe that they must be urgently closed. For other types of data, additional sources of genotype information leakage could be present after the anonymization is applied. For example, 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. Complex machine learning frameworks, such as deep learning and neural networks, have great potential to reveal the correlations between variants and activity levels of

Deleted: have

Deleted: the

Deleted: to share. Specifically, our

Deleted: fairly

Deleted: prediction

Deleted: 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 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 ...

pathways. Although there has been interest in identifying these higher-order QTLs, these are not yet extensively studied[27].

At this point, it is useful to review all the sources of information leakage from functional genomics experiments, such as RNA-Seq, and point out the sources that we probed in this paper. First, there is leakage directly from the reads. This is the most obvious leakage, and can be avoided by simply not sharing the raw reads. The next source of leakage is from the signal profile. We address this leakage is in this paper. There is yet another source of leakage, 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 eQTLs and can create substantial leakage. Furthermore, one can envision additional sources of leakage beyond these main areas. For instance, although the eQTLs traditionally have been linked to genes, highly expressed intergenic regions[43] may also be linked to eQTLs. In addition, while we consider a particular class of structural variants (i.e., small and large deletions), there may be very large, megabase-scale deletions that affect many genes. This is particularly the case for somatic events in cancer samples. These cases are not addressed in our study.

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, for example, 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 three values $\{0,1,2\}$. These values correspond to the three genotypes of the deletion and represent how many copies of the genomic sequence are 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 i is denoted by S_i. An important quantity that we utilize in formulating methods is the multi-mappability profile of the deletion regions. 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[44] 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[45].

Deleted: e.g.

Deleted: 3

Deleted: they

Deleted: is

Deleted: at

Deleted: The multi

Deleted: [42]

Deleted: [43].

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 (Supplementary Fig. 5). Given a deletion $[i, j]$, an important feature for genotyping the deletion is the average functional genomic signal within the deletion:

Deleted: Supp

Deleted: 3

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

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, that is, higher $\eta_{[i,j]}$, the accuracy of deletion genotyping is higher. Next, we computed the average signal in the neighborhood as

Deleted: i.e.,

Deleted: compute

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

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

Deleted: define

Deleted: 5

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

This corresponds to the conditional probability (over all deletions within the genome) that we observed for 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; that is, we estimate the probability for all the deletions in the genome. For this, we computed five features for every deletion in the genome, and then estimated the conditional probability using this set as the sample of deletions.

Deleted: the

Deleted: observe the

Deleted: , i.e.,

Deleted: compute 5

Deleted: estimate

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 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 and then rounding this value to the closest smaller integer value.

Deleted: ,

Deleted: .

Deleted: more

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

The practical instantiation of the linking attacks that we studied are based on genotyping the panel of small deletions, p_S, using 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 performed 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 the *self-to-neighbor signal ratio* is very useful for genotyping small deletions. For all small deletions, we computed *self-to-neighbor signal ratio*, $\rho_{[i,j]}$, neighbor signal balance, $\eta_{[i,j]}$, and average neighbor signal. We then selected the deletions that satisfied the following criteria:

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

We sorted the set of small deletions that passed these criteria with respect to increasing $\rho_{[i,j]}$. The deletions that are at the top of the sorted list correspond to deletions that are highly mappable (low multi-mappability signal), have strong neighbor signal support (high average neighbor signal), and 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 (Supplementary Fig. 6), 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 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 selected candidate large deletions using average multi-mappability signal:

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

We sorted the deletions that satisfied the 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 a substantial effect on accuracy of linking attacks as long as they were 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 fragmented the genome into windows and used these windows as candidate deletions. We utilized the above procedure for selection of the candidate deletions, which were assigned homozygous deletion genotypes. For small

Deleted: study
Deleted: the
Deleted: GTex
Deleted: perform
Deleted: the
Deleted: are computed.
Deleted: select
Deleted: satisfy
Formatted Table
Deleted: For
Deleted: pass
Deleted: , we sorted the deletions
Deleted: which
Deleted: the
Deleted: which
Deleted: finally they
Deleted: Supp
Deleted: 4
Deleted: the
Deleted: select
Formatted Table
Deleted: satisfy
Deleted: are
Deleted: fragment
Deleted: use
Deleted: Above
Deleted: is utilized
Deleted: are

deletions, we used five base pair windows within the exonic regions. For large deletions, we used 1,000 base pair windows over the whole genome.

Deleted: use 5

Deleted: use 1000

Deleted: all

4.3. Instantiations of Genome-wide Linking Attack

Following genotyping of the deletions in p_S, we used 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 compared these deletions to the panel of deletions in the genotype dataset, p_G. The comparison was performed by overlapping the deletions in p_S and p_G. Any two deletions that overlapped at least one base pair were assumed to be common in the two panels. For the "common" set, $\{[i_1,j_1],[i_2,j_2],\ldots,[i_n,j_n]\}$, we computed the genotype distance by matching the genotypes,

Deleted: the

Deleted: use

Deleted: compare

Deleted: is

Deleted: in

Deleted: 1

Deleted: are

Deleted: set of deletions that are

Deleted: in two panels

Deleted: compute

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

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 $d\left(G_{[i',j']},G_{[i\prime,j\prime]}\right)$ 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 computed the genotype distance of the 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 the genotype dataset. The individual in the genotype dataset that has the smallest genotype distance was linked to k^{th} individual:

Deleted: compute

Deleted: is

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

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

Deleted: matches

Deleted: mark

Deleted: compute

Deleted: [16]

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

where $d_k^{(1)}$ and $d_k^{(2)}$ is the minimum and second minimum genotype distance among all the genotype distances computed between the 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 PPV of linkings when the linkings are selected using the first distance gap measure, we used following formula:

Deleted: positive predictive value (

Deleted:)

Deleted: use

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

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

where $d_{1,2}^{min}$ represents the minimum first distance gap measure that was used to select individuals. In these formulae, sensitivity represents the fraction of all individuals that the adversary correctly links. PPV represents the fraction of individuals that are correctly linked among the individuals whose linking satisfies the 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 median filtering[46]-based signal processing to locally smooth the signal profile around the deletion. This signal processing technique has been used to remove shot noise in two-dimensional imaging data and one-dimensional audio signals[45,47]. For each genomic a in the deletion $[i, j]$, we replaced 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.

5. Datasets

The mapped reads for the RNA-Seq data from the GEUVADIS project were obtained from the GEUVADIS project website (http://geuvadis.org/). The RNA-Seq mapped reads from the GTEx project were obtained from the dbGAP portal. We used only the RNA-Seq datasets from whole blood tissue to create signal profiles. The structural variant panel and genotypes were obtained from the 1,000 Genomes Project. Very low frequency SVs may introduce bias since they can uniquely identify an individual. In order to get around this bias, we removed the SVs of which the minimum genotype frequency was larger than 0.01. In addition, we extended the genotype dataset by re-sampling the 1,000 Genomes deletion dataset and created genotype data for 10,000 simulated individuals.

We utilized randomized datasets to compare predictability with real data. In order to create randomized data, we shuffled the signal profiles circularly. In this way, the association between the SV genotypes and signal profiles were 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 a genome-wide signal profile and also sensitive information (e.g., HIV status) for each individual. The names are anonymized into IDs as shown in the blue shaded column. The

Deleted: are

Deleted: the

Deleted: [44]

Deleted: 2

Deleted: 1

Deleted: [43, 45]. For each genomic a in the deletion $[i, j]$, we replace

Deleted: seq

Deleted: gEUVADIS

Deleted: are

Deleted: gEUVADIS

Deleted: web site

Deleted: seq

Deleted: GTex

Deleted: are

Deleted: seq

Deleted: are

Deleted: 1000

Deleted: The very

Deleted: and

Deleted: whose

Deleted: is

Deleted: Also

Deleted: 1000

Deleted: have

Deleted: for comparison of

Deleted: This

Deleted: are

Deleted: , for each individual,

Deleted: ,

Deleted: .

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. The adversary then genotypes the SVs (3) in the panel and builds the dataset for genotyped SVs ($\tilde{G}$). b) The adversary acquires an SV panel (p_G) and genotype dataset (G), which contains the 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 their SV panel (p_S) to the SV panel (p_G). For the matching SVs, the adversary compares the genotypes. The individuals in G who have good matches with respect to genotype distance are linked to signal profile individuals, as indicated by the matching of colored columns. This linking reveals the HIV status of the individuals in the genotype dataset. c) This example shows a large deletion in the NA12878 individual and how it affects signal profiles. A 70kb long region is deleted in the NA12878 individual and the decrease in signal profiles show the loss of signal along the deletion. d) This schematic shows large and small deletions and how they are manifested in signal profiles. The large deletions show a large decrease in the signal profiles, while the 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 (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) This plot demonstrates the accuracy of linking with genotyping of a known panel. The number of variants used in the attack is shown on the x-axis, while accuracy is shown on the y-axis. The variants are sorted with respect to decreasing predictability. d) This shows the linking accuracy when the adversary performs joint discovery and genotyping of deletions to achieve linking. e) The blue plot shows the accuracy of linking when indels of a specific length are used in the attack. The green plot shows the distribution of indel lengths. f) For the genotyping only scenario, the plot shows the distribution of the minimum number of variants required to identify each individual. The x-axis shows the number of indels and the 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 the minimum number of variants required to identify each individual.

Figure 3: The accuracy of linking attack on GTEx dataset. a-b) The ICI leakage versus predictability for all the indels before (a) and after (b) signal profile anonymization. c) The linking attack accuracy with a changing number of variants used in the attack. The x-axis shows the number of variants used in the attack and the y-axis shows the accuracy of linking. d) When the adversary uses 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-b) Scatter plots show ICI leakage versus predictability for Kasowski (a) and Kilpinen (b) datasets. c) The accuracy of linking attack on the two datasets for a genotyping only scenario. The x-axis shows the changing number of variants used in the attack and the 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. The middle column indicates whether NA12878 is identifiable among 1,000 genomes samples, represented by green

Deleted: She

Deleted: dataset

Deleted: (G)

Deleted: her

Deleted: that

Deleted: the

Deleted: Example of

Deleted: The

Deleted: representation of

Deleted: , denoted by a

Deleted: .

Deleted:)

Deleted: c) The

Deleted: in

Deleted: The

Deleted: of linking

Deleted: perform

Deleted: Green

Deleted: indels

Deleted: that is

Deleted: X

Deleted: that is

Deleted: GTex

Deleted: X

Deleted: the

Deleted:) The scatter plot of

Deleted: X

Deleted: Middle

Deleted: 1000 Genomes

check for yes and red cross for no. The third column is a schematic representation of 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. A green check indicates that NA12878 is vulnerable while a red cross indicates it is not vulnerable.

Figure S1: Illustration of the 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 an average of 200 movies and each movie was rated on average by 5,000 users. b) The training dataset was linked to 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 S3 (was 2): A 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 1,000 Genomes SV and the population-wide predictability represents how correctly predictable the SV genotypes are given the gene expression levels. Gene expression levels were obtained from the GEUVADIS dataset. The ellipses point to the small number SVs that have high predictability and high ICI leakage.

Figure S4 (was 3): Feature sets that are used to genotype and discover deletions. A candidate deletion is located between the 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 $[j+1,\ 2j-i+1]$, respectively) to compute the features. $\rho_{[i,j]}$ represents the deepness of the dip 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 S5 (was 4): Accuracy of genotype predictions that are used in instantiating the linking attacks. The x-axis shows the number of variants used and the y-axis shows the genotype accuracy. The GEUVADIS signal profiles are used with a known panel of 1,000 Genomes small indels.

Figure S6 (was 5): A screenshot of the region surrounding the two 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 bottom figure shows a zoomed view around 100 base pairs of the deletion. The dip in the RNA-Seq signal that is caused by the deletion is visible by eye in three individuals, XV7Q, 14BMU, and 139D8. Other individuals shown in the figure do not have this deletion. It is also worth noting that these signal profiles are publicly available for download from the UCSC Genome Browser.

Figure S6: 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.

Deleted: the

Deleted: Green

Deleted: on

Deleted: the Internet Movie Database (IMDb)'s

Deleted: ***S2:*** The

Deleted: 1000

Deleted: The

Deleted: are

Deleted: ***S3:***

Deleted: set

Deleted:

Deleted: how deep

Deleted: is

Deleted: ***S4:***

Deleted: 1000

Deleted: ***S5:***

Deleted: 2

Deleted: GTex

Deleted: seq

Deleted: botton

Deleted: version

Deleted: pair

Deleted: seq

Deleted: can be easily seen

Deleted: 3

Deleted: ***Figure S6: [[TO BE ADDED]]***¶

REFERENCES

1. Joly, Y., Dyke, S. O. M., Knoppers, B. M. & Pastinen, T. Are Data Sharing and Privacy Protection Mutually Exclusive? *Cell* **167,** 1150–1154 (2016).

2. Singer, D. S., Jacks, T. & Jaffee, E. A U.S. &quot;Cancer Moonshot&quot; to accelerate cancer research. *Science* **353,** 1105–6 (2016).

3. Collins, F. S. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372,** 793–795 (2015).

4. Handelsman, J. The Precision Medicine Initiative. *White House, Off. Press Secr.* 1–5 (2015). doi:10.1177/1557988315574512

5. Caulfield, M. *et al.* The 100,000 Genomes Project Protocol. *Genomics Engl.* (2015).

6. Chisholm, J., Caulfield, M., Parker, M., Davies, J. & Palin, M. Briefing- Genomics England and the 100K Genome Project. *Genomics England* (2013). Available at: http://www.genomicsengland.co.uk/briefing/.

7. Feero, W. G., Guttmacher, A. E., Feero, W. G., Guttmacher, A. E. & Collins, F. S. Genomic Medicine — An Updated Primer. *N. Engl. J. Med.* **362,** 2001–2011 (2010).

8. Joly, Y., Feze, I. N., Song, L. & Knoppers, B. M. Comparative Approaches to Genetic Discrimination: Chasing Shadows? *Trends Genet.* **33,** 299–302 (2017).

9. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4,** (2008).

10. Im, H. K., Gamazon, E. R., Nicolae, D. L. & Cox, N. J. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet.* **90,** 591–598 (2012).

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

12. Vaikuntanathan, V. Computing Blindfolded: New Developments in Fully Homomorphic Encryption. *2011 IEEE 52nd Annu. Symp. Found. Comput. Sci.* 5–16 (2011). doi:10.1109/FOCS.2011.98

13. Fienberg, S. E., Slavković, A. & Uhler, C. Privacy preserving GWAS data sharing. in *Proceedings - IEEE International Conference on Data Mining, ICDM* 628–635 (2011). doi:10.1109/ICDMW.2011.140

14. Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biology* **12,** 125 (2011).

15. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. in *Proceedings - IEEE Symposium on Security and Privacy* 111–125 (2008). doi:10.1109/SP.2008.33

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

17. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339,** 321–4 (2013).

18. Schadt, E. E., Woo, S. & Hao, K. Bayesian method to predict individual SNP genotypes from gene

Formatted

Deleted: …Joly, Y,…, Dyke SOM,… S. O. M., Knoppers

Formatted

Deleted: 2016:…**67,** 1150–1154.

Deleted: …Singer DS,… D. S., Jacks, T,… & Jaffee, E:

Formatted

Deleted: 2016, …**53**:… 1105–6.

Deleted: …Collins FS:

Formatted

Deleted: 2015,… **372**:… 793–795.

Deleted: …Handelsman, J:

Formatted

Deleted: … 1–5 (2015:1–5.

Deleted: …Caulfield M, Davies J, Dennys M, Elbahy L,

Deleted: … (2015(February

Formatted

Formatted

Deleted: [… *Genomics England* (2013). Available at:

Deleted: …Feero WG,… W. G., Guttmacher AE,… A. E.,

Formatted

Deleted: 2010,… **362**:… 2001–2011.

Deleted: …Joly, Y,…, Feze IN,… I. N., Song, L,… &

Formatted

Deleted: 2017,… **33**:… 299–302.

Deleted: 9. Homer N, Szelinger S, Redman M, Duggan D

Formatted

Deleted: … **4,** (2008, **4**.

Deleted: …Im HK,… H. K., Gamazon ER,… E. R., Nicolae

Formatted

Deleted: 2012,… **90**:… 591–598.

Deleted: …Dwork, C:

Formatted

Deleted: 2006,… **4052**:… 1–12.

Deleted: …Vaikuntanathan, V:

Formatted

Deleted: 2011:… 5–16.

Deleted: …Fienberg SE,… S. E., Slavković, A,… & Uhler,

Formatted

Deleted: In…n *Proceedings - IEEE International*

Deleted: …Sboner, A,…, Mu, X,…, Greenbaum, D,…,

Formatted

Deleted: 2011:…**2,** 125.

Deleted: …Narayanan, A,… & Shmatikov, V:

Formatted

Deleted: In…n *Proceedings - IEEE Symposium on Securi*

Formatted

Formatted

expression data. *Nature Genetics* **44,** 603–608 (2012).

19. Backes, M. *et al.* Identifying Personal DNA Methylation Profiles by Genotype Inference. in *Proceedings - IEEE Symposium on Security and Privacy* 957–976 (2017). doi:10.1109/SP.2017.21

20. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

21. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).

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

23. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

24. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

25. Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L. & Almouzni, G. Epigenomics: Roadmap for regulation. *Nature* **518**, 314–316 (2015).

26. Consortium, T. G. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–5 (2013).

27. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).

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

29. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).

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

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

32. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–11 (2013).

33. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Sci. (New York, NY)* **342**, 747–749 (2013).

34. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–7 (2013).

35. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science (New York, NY)* **342,** 750–752 (2013).

36. van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **6**, 1869 (2010).

Deleted: 2012:...**4,** 603–608.

Deleted: Pepke S, Wold B, Mortazavi A: **Computation fo**

Deleted: In...n *Proceedings - IEEE Symposium on Securi*

Formatted

Deleted: 21. ...0. Sudmant PH, Rausch T, Gardner EJ,

Formatted

Deleted: 2015, ...**26:**... 75–81.

Deleted: 22. ...1. Stranger BE, Forrest MS, Dunning M

Formatted

Deleted: 2007, ...**15:**... 848–853.

Deleted: 23. ...2. The 1000 Genomes Project

Formatted

Deleted: , **135**:0–9.

Deleted: 24. ...3. The 1000 Genomes Project

Formatted

Deleted: 2015:...**26,** 68–74.

Deleted: 25. ...4. Bernstein BE, Birney E, Dunham I,

Formatted

Deleted: 2012, ...**89:**... 57–74.

Deleted: 26. ...5. Romanoski CE,... C. E., Glass CK,... C.

Formatted

Deleted: 2015, ...**18:**... 314–316.

Deleted: 27. ...6. Consortium TG:

Formatted

Deleted: 2013,... **45:**... 580–5.

Deleted: 28. Ardlie KG, Deluca DS, Segre A V., Sullivan T

Formatted

Deleted: - *)* 2015,.... *).* **348:**... 648–660.

Deleted: 29. ...8. Abyzov, A,..., Urban AE,... A. E., Snyd

Formatted

Deleted: 2011,... **21:**... 974–984.

Deleted: 30. ...9. Handsaker RE,... R. E., Korn JM,... J. M

Formatted

Deleted: 2011,... **43:**... 269–276.

Deleted: 31. ...0. Wang, Z,..., Gerstein, M,... & Snyder,

Formatted

Deleted: 2009,... **10:**... 57–63.

Deleted:

Formatted

Deleted: , van de Geijn B, Degner JF, Cain CE, Banovich

Formatted

Deleted: 2013, ...**42:**... 747–749.

Deleted: 33. Kilpinen H, Waszak SM, Gschwind AR,

Formatted

Deleted: 2013, ...**42:**... 744–7.

Formatted

Formatted

37. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).

38. Korbel, J. O. & Lee, C. Genome assembly and haplotyping with Hi-C. *Nat Biotech* **31,** 1099–1101 (2013).

39. Euskirchen, G. M. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies. *Genome Res.* **17,** 898–909 (2007).

40. Royce, T. E. *et al.* Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends in Genetics* **21,** 466–475 (2005).

41. Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nat. Rev. Genet.* **14,** 333–346 (2013).

42. Habegger, L. *et al.* RSEQtools: A modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27,** 281–283 (2011).

43. Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512,** 445–8 (2014).

44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

45. 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.* **15,** 474 (2014).

46. Chan, R. H., Ho, C.-W. & Nikolova, M. Salt-and-Pepper noise removal by median-type noise detectors and detail-preserving regularization. *IEEE Trans. Image Process.* **14,** 1479–1485 (2005).

47. Wang, Z. W. Z. & Zhang, D. Progressive switching median filter for the removal of impulsenoise from highly corrupted images. *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.* **46,** (1999).

Deleted: 36. ...7. Rao SSP, Huntley MH, Durand NC, ...

Formatted: Font: Not Bold

Deleted: 2014,...**159**:... 1665–1680. ...

Deleted: 37. ...8. Korbel JO,... J. O. & Lee, C: ...

Formatted: Font: Not Bold

Deleted: 2013, ...**1**:... 1099–1101. ...

Deleted: 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:

Formatted: Font: Not Bold

Deleted: 2007,... **17**:... 898–909. ...

Deleted: 39. Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M:

Formatted: Font: Not Bold

Deleted: 2005:...**1,** 466–475. ...

Deleted: 40. ...1. Berger, B,..., Peng, J,... & Singh, M: ...

Formatted: Font: Not Bold

Deleted: 2013,... **14**:... 333–346. ...

Deleted: 41. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M:

Formatted: Font: Not Bold

Deleted: 2011, ...**7**:... 281–283. ...

Deleted: 42.

Formatted: Indent: Left: 0", Hanging: 0.44"

Deleted: ,... & Salzberg SL: ...

Formatted: Font: Not Bold

Deleted: 2012:...**,** 357–359. ...

Deleted: 43. ...5. Harmanci, A,..., Rozowsky, J,... & ...

Formatted: Font: Not Bold

Deleted: 2014,... **15**:... 474. ...

Deleted: 44. ...6. Chan RH,... R. H., Ho, C-...-W,... & ...

Formatted: Font: Not Bold

Deleted: 2005,... **14**:... 1479–1485. ...

Deleted: 45. ...7. Wang ZWZ,... Z. W. Z. & Zhang, D: ...

Formatted: Font: Not Bold

Deleted: ... **46,** (1999, **46**. ...