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

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

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

Abstract

The functional genomics data is emerging as a valuable resource for personalized medicine. Although one might think that the functional genomics data is safe to share, the extent to which they leak sensitive information is not well studied. Here, we show that the signal profiles, which are often publicly shared, for several functional genomics data types can cause concerns for privacy. Genome-wide signal profiles represent measurements of genome-wide activity at each genomic position. For example, in an RNA sequencing based assay, the signal profile is generated by counting the number of reads at each genomic position and represents the transcriptional activity at each position. We show that there is significant leakage from the signal profiles of a number of sequencing based functional assays including RNA-seq, ChIP-Seq, We demonstrate that an adversary can predict indels and structural variants, and use those to accurately identify an individual among a large pool of individuals in a linking attack. We also propose a metric to measure the accuracy of genotyping the deletion variants using signal profiles. To show the practicality of linking attacks through signal profiles, we present several outlier based genomic deletion genotyping methods that lead to accurate linking attacks. We finally present a novel and effective anonymization procedure for protection of signal profiles against genotype prediction based linking attacks. Given that several consortia, for example GTex and TCGA, publicly share signal profiles for personal functional genomics data; our results point to a critical source of sensitive information leakage, which can be potentially protected by our anonymization technique.

1. Introduction

Individual privacy is emerging as an important aspect of biomedical data science. A deluge of genetic data is being generated with the Cancer Moonshot Project[1], Precision Medicine Initiative[2, 3], and UK100K[4, 5] from hundreds of thousands, if not millions, of individuals. Moreover, there is much effort to make genetic data more prevalent in the standard of care[6]. This will increase personal genomic data storage in healthcare providers. Leakage of the genetic information creates many privacy concerns, e.g. genetic predisposition to diseases may bias insurance companies.

The initial studies on genomic privacy has focused on protecting the identities of participants in the early genetic profiling and genotype-phenotype association studies[7, 8]. These focused on whether an

Deleted: pi@gersteinlab.org)
Formatted: Font: 11 pt, Font color: Red
Formatted: Normal
Deleted: for the first time
Deleted: read depth
Deleted: A
Deleted: , and Hi-C
Deleted: small
Deleted: large deletions
Deleted: cross-reference
Deleted: easily
Deleted: protection of single nucleotide polymorphism (SNP) datasets and analysis of privacy of
Deleted:
Deleted: It is worth noting that cryptographic approaches are also utilized for protecting genetic information[9, 10]. The significant increase in available datasets has made genomic linking attacks much more relevant [11–13].

individual's genetic information can be used to reliably predict whether they participated to 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, the differential privacy[9] has been proposed as a theoretically optimal formalism that can fulfill the privacy requirements such that the probability that any individual's participation can be identified made arbitrarily small. In addition, the cryptographic approaches have proven useful for privacy-aware analysis of genomic datasets albeit with high requirements of computational resources[10, 11].

The decrease in cost of DNA sequencing technologies has substantially increased[12] the number and size of available genomic data and has made genomic data much more practically available to hospitals, research institutes, and to individuals. This increase will render the genomic linking attacks much more relevant[13–15]. In a nutshell, the linking attacks are based on cross-referencing and matching of two or more datasets that are released independently. Some of the datasets contain personal identifying information, e.g. names or addresses, while others contain sensitive information, e.g. health information. The immediate consequence of the cross-referencing is that the sensitive information in one dataset gets linked to the identifying information in another dataset, which in turn breaches privacy of individuals whose sensitive phenotypes are revealed. The risks behind linking attacks are becoming high in the recent years because the personal data is generated at exceedingly high pace and these information are independently released and maintained. A rather challenging aspect of linking attacks is that risk assessment is complicated because one dataset that is currently deemed safe to release may become a target for linking attacks when another dataset is released in the future, i.e., a dataset that seems safe to release now may become vulnerable to a linking attack next year.

A well-known example of linking attacks is the Netflix Prize Competition[13]. In this competition, a training dataset was released by the movie rental company Netflix, which was to be used for training new automated movie rating algorithms. The dataset was anonymized by removing names. Although the dataset seemed safe to share at the time, two researchers have shown that this training dataset can be linked to a seemingly independent database of the Internet Movie Database (IMDb). The linking revealed movie preferences and identities of many Netflix users. We believe that similar scenarios will be a major route to breaches in individual genomic privacy and these must be studied well to enable privacy-aware data sharing approaches.

There are two major aspects of genomic privacy that are not well addressed in the previous studies. Firstly, although it is well known that the major portion of individual genomic polymorphism is structural variants, deletion, insertion, translocation, and transversion of large chunks of DNA sequence, these did not receive much attention in the debate of genomic privacy[16]. The structural variants can have much larger effects on the molecular phenotypes (like gene expression) than SNPs simply because they effect a much larger portion of the genome. Secondly, functional genomics data is not in center of the most studies. Especially the newer functional genomics datasets based on sequencing assays, like RNA-Seq[17] and ChIP-Seq[18] are very rich sources of information that can lead to leakage of individual characterizing information. In general, the raw sequenced reads from these experiments are not shared because of privacy concerns. The reads are used to create the genome-wide signal profiles by piling them up along the genome. The signal profiles represent the activity at each genomic position and are therefore fundamental in the analysis of any type of genome-wide functional assay. It is generally assumed that the signal profiles are mostly void of sensitive information and they are publicly shared, for example by the ENCODE Project[20], Roadmap Epigenome Mapping Consortium[21], and GTex[22,

Deleted: or more of the datasets are

Deleted: an individual and this causes a

Deleted: breach.

Deleted: especially

Deleted: these

Deleted: information

Deleted: speed

Deleted: For example, the maintainers

Deleted: not be aware of each other or some of the datasets may be

Deleted: much earlier/later than

Deleted: other datasets.

Deleted: A very famous example is the Netflix Prize Competition[11]. In this competition, a training dataset was released by the move rental company Netflix, which was to be used for training new automated movie rating algorithms. The dataset was anonymized by removing names. Two researcher have shown that this training dataset can be linked to a seemingly independent database of IMDb web site and revealed movie preferences and identities of many Netflix users. We believe this will be a significant route to breaches in individual genomic privacy. Most of the previous studies focus on leakage of single nucleotide polymorphisms (SNPs) genotypes as a source of sensitive information. There are two major aspects that are not well addressed in the previous studies. Firstly, although it is well known that the major portion of individual genomic polymorphism is structural variants, deletion, insertion, translocation, and transversion of large chunks of DNA sequence, these did not receive much attention in the debate of genomic privacy[14]. The structural variants can have much larger effects on the molecular phenotypes (like gene expression) than SNPs simply because they effect a much larger portion of the genome. This could render the personal SVs more detectable compared to SNPs. Secondly, moreover in a sense more obvious and noticeable???, functional genomics data is not in center of the most studies. Especially the newer functional genomics datasets based on sequencing assays, like RNA-Seq[15] and ChIP-Seq[16] are very rich sources of information that can lead to leakage of individual characterizing information. In general, the raw sequenced reads from these experiments are not shared because of privacy concerns. File formats like MRF[17] and tagAlign can enable removing raw sequence information from reads while keeping the information about read mapping intact. These reads can be used to create the genome-wide signal profiles by piling them up along the genome. Indeed, the genome-wide signal profiles are publicly shared by many projects like ENCODE[18], Roadmap Epigenome Mapping Consortium[19], and GTex[20, 21]. It is urgently necessary to evaluate the ...

23]. Although one might think that these signal profiles are safe to share, there has been no study that systematically analyzed the information leakage in these datasets.

In this paper, we analyze the sensitive individual characterizing information leakage from the signal profiles of several sequencing based functional genomics datasets. By signal profile, we refer to the signal generated by counting the number of reads that overlap with each nucleotide on the genome. Although the signal tracks do not contain any explicit sequence information, an adversary can utilize signal processing techniques to detect the large and small structural variants. The most notable of these variants are the small and large deletions. For example, many methods have been developed to identify genomic deletions and duplications from the DNA-sequencing read depth signal [24, 25]. On the other hand, detection of structural variants from functional genomics datasets is not well-studied. The main reason for this is the dynamic and non-uniform nature of the signal profiles of functional genomics experiments, unlike DNA-sequencing signal profiles that uniformly cover the genome. For example, RNA-seq[17] and ChIP-seq[18] signal profiles concentrate mainly on the exonic regions and promoters of the genome, respectively. Moreover, 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 can bring enough power to an adversary for genotyping structural variants and performing a successful linking attack.

The paper is organized as following: We first present the general scenario of linking attacks that utilize signal profiles. We next propose a new metric for quantifying the extent to which genotypes of small and large deletion variants can be estimated using functional genomics signal profiles. In combination with information content of the deletion variants, we use this new metric for evaluating the extent of characterizing information leakage from functional genomics datasets. We next present several practical instantiations of linking attacks that utilizes deletion variant genotype prediction using outlier signal levels. Finally, we focus on protection of the signal profiles against linking attacks. We present a novel signal processing methodology for anonymizing a signal profile. We show that it is effective in decreasing the predictability of deletion variant genotypes from signal profiles. The source code for linking attacks and anonymization can be downloaded from privaseq2.gersteinlab.org.

Deleted: how correctly

Deleted: To protect

Deleted: , we

Deleted: the

Deleted:

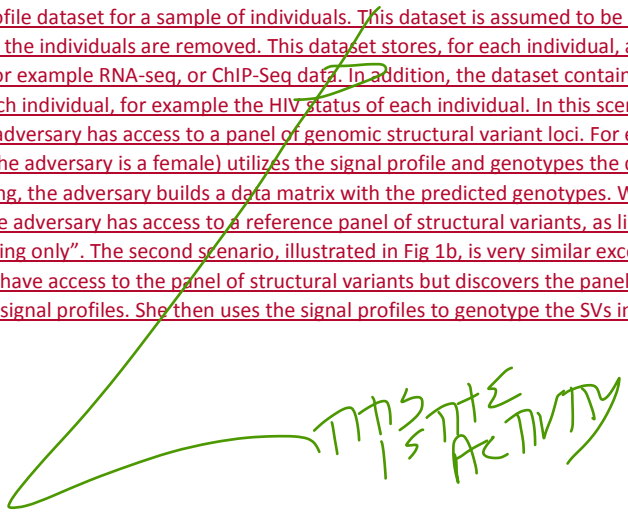
2. Results

2.1. Linking Attack Scenario

Figure 1 summarizes the linking attack scenario. The attack has two steps. The first step is genotyping of the deletion variants, which is illustrated in Figures 1a. The adversary has access to a genome-wide signal profile dataset for a sample of individuals. This dataset is assumed to be shared publicly after the names of the individuals are removed. This dataset stores, for each individual, a genome-wide signal profile, for example RNA-seq, or ChIP-Seq data. In addition, the dataset contains sensitive information about each individual, for example the HIV status of each individual. In this scenario, we also assume that the adversary has access to a panel of genomic structural variant loci. For each individual, she (we assume the adversary is a female) utilizes the signal profile and genotypes the deletions. After the genotyping, the adversary builds a data matrix with the predicted genotypes. We refer to this scenario, where the adversary has access to a reference panel of structural variants, as linking based on "genotyping only". The second scenario, illustrated in Fig 1b, is very similar except that the adversary does not have access to the panel of structural variants but discovers the panel of structural variants from the signal profiles. She then uses the signal profiles to genotype the SVs in this de-novo discovered

Deleted: Genome-wide

Deleted: Figure 1a summarizes the linking attack scenario that we focus. The adversary has access to a leaked structural variation (SV) dataset and another molecular phenotype dataset that contains genome-wide functional genomics signal profiles for example RNA-seq or ChIP-Seq signal profiles. The SV dataset contains identifying sample IDs for each individual and SV genotypes for multiple locations. We assume that the SV dataset comprises different types of variants like deletions, duplications, and translocations. The phenotype dataset also contains very sensitive information, i.e. HIV status, about the individuals. He/She uses the signal profiles to perform SV genotype prediction. He/She then compares the predicted SV genotypes and the leaked genotype dataset.



SV panel. We refer to this scenario as linking based on "joint discovery and genotyping". After the genotyping, the genotyped SV matrix includes, for each individual, the predicted SV genotypes, and also the sensitive information about HIV status.

The second step of the linking attack is linking of the genotyped SV dataset and the SV genotype dataset. The SV genotype dataset is assumed to contain identifying information about individual's identities. We assume that this dataset was either leaked or stolen. The adversary first compares her genotyped SV panel to the SV panel of the genotype dataset. For example, she may overlap the loci of the SVs in two panels. After the matching of the SVs in the two panels, she compares the genotypes of the matching SVs in two datasets. She uses this comparison to cross-reference the individuals in two datasets and find the individuals that best match to each other with respect to genotype match distance, i.e., links the individuals in two datasets. The results are used to link the genotype samples to the phenotype samples and the HIV status of genotype samples are revealed to the adversary (the matched columns in the final linked matrix).

In the analysis below, we are focusing on the small and large deletion variants. So we assume that the adversary focuses only on the deletion variants in the panel of SVs that she uses for performing the attacks.

2.2. Information Content and Correct Predictability of Structural Variant Genotypes

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

Other than the predictability, an important measure in the linking attacks is the information content each SV genotype supplies. We utilize a previously proposed metric termed individual characterizing information (ICI) to quantify the information content of each SV. This measure gives higher weight to the genotypes that have low population frequency and vice versa. For a given variant genotype, ICI measures how much information it supplies for pinpointing an individual in a population. As we discussed above, the genome-wide predictability is independent of the population frequency of the variants. Therefore the adversary can utilize genome-wide prediction approaches and predict rare variant genotypes to gain high ICI and characterize individuals very accurately. This is one of the major differences between genome-wide prediction approach and the sample-wide prediction[14] based approach (Supplementary Fig 1).

2.3. Linking Attacks using RNA-Seq Signal Profiles

We first focus on predictability of short deletions using RNA-seq signal profiles (Fig 1b). By small deletions, we refer to deletions smaller than 10 base pairs. The basic observation is that each deletion is manifested as an abrupt dip in the signal profile. The discovery and genotyping of a deletion relies on detecting these dips in the signal profiles. The genome-wide predictability (π_{GW}) of the small deletions

Deleted: .

Deleted: It has been observed that the prediction of SV genotypes from functional genomics signal profiles has relatively low accuracy.

Deleted: Given

Deleted: , the predictability

Deleted: separately and independently

Deleted: Because of this fact,

Deleted: variants.

Deleted: This is one of the major differences between genome-wide prediction approach proposed in this study and the recently proposed sample-wide prediction [12] based approach (Supplementary Fig 1). To compare these two approaches, we computed the sample-wide predictability of all the genomic deletions from the 1000 Genomes Project using the gene expression quantifications from GEUVADIS project [14, 26] (Supplementary Fig 1). ICI versus π_{SW} plot shows that there are not many SVs that have high predictability and high information content. One reason for this is that a significant number of SVs that impact gene expression levels have low population frequency and their sample-wide predictability are rather low. This implies that the gene expression levels can be shared without high risk of individual characterization using SV genotype prediction. However, as we will show later, a large fraction of these low frequency SVs have high genome-wide predictability and they can be used in individual characterization and identification.

Deleted: Genome-wide

Deleted: with Short Deletion Prediction from

Deleted: Each

Deleted: prediction

Deleted: is done by

quantifies how well the adversary can identify the dips from the signal profile (Methods Section). We first estimated the genome-wide predictability for the panel of short deletions in 1000 Genomes Project using the RNA-seq expression signal profiles from the GEUVADIS project. Figure 2a,b show π_{GW} vs ICI for short deletions. There is a substantial number of deletions that have much higher predictability compared to a randomized dataset where the signal profile is randomized with respect to location of deletions. There are also many more variants with very high ICI (on the order of 5-6 bits) with high predictability (greater than 80% predictability).

In order to present practicality of small deletion predictability and information content, we propose an instantiation of a linking attack where we utilize outlier signal levels in the signal profiles for discovery and genotyping of the small deletions. As we explained, the genotyping of deletions are based on detecting the abrupt dips in the signal profile. In order to detect these dips in the signal profile, the adversary utilizes a quantity we term *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,

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

The genomic regions with low $\rho_{[i,j]}$ values point to intervals tend to have dips in them. For each individual, the prediction method sorts the short deletions with respect to *self-to-neighbor signal ratio* and assigns homozygous genotype to a number of deletions with 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 order to minimize the bias on the deletion panel, we used the deletions with minor allele frequency greater than 1% in this analysis. Also, we extended the genotype dataset by re-sampling 1000 Genomes deletion dataset and created genotype data for 10,000 simulated individuals. 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 (Figure 2d).

In the previous analysis, the sample set used for discovery of deletion panel and RNA-seq sample set are matching, i.e. 1000 Genomes individuals. This may introduce a bias in linking because the SV genotype dataset may contain rare indels which may also be in the panel of deletions. This would make it trivial to link some of the individuals. To get around this bias, we studied linking attack where signal profile dataset is generated by the GTex Project Consortium [22, 23] and the panel of small deletions is the deletion set generated by the 1000 Genomes Project. This way, the SVs in the panel are identified in 1000 Genomes individuals while the linking is performed for the individuals in GTex Project datasets. In other words, the deletion panel is discovered in a sample set that is totally independent of the sample set that the adversary is linking. In this scenario, the adversary is linking the signal profile dataset to the genotype dataset that is obtained from the GTex Project. With this setup, we first computed π_{GW} versus ICI for the deletions and observed that there is substantial enrichment of deletions that have high predictability with high ICI compared to randomized datasets (Fig 3a). We also instantiated the linking attack using the previously presented extremity based approach. In the instantiation, we first evaluated the attack based on *genotyping only* scenario. In this scenario, the linking accuracy is close to 100%

Deleted: ¶

Deleted: computed

Deleted: , for genotyping of known deletions (Methods Section). For both cases, there

Deleted: . In comparison to sample-wide predictability of genotypes, there are a lot of deletions that provide deletions with very high ICI (higher than 5 bits) with high genome-wide predictability (Supplementary Figure S1

Deleted: prediction

Deleted: deletion genotype prediction (Methods Section).

Deleted: *deletion*

Deleted: *deletion*

Deleted: the assigned homozygous deletion loci

Deleted: Thus, the attacker utilizes the outliers in *deletion-*

Formatted: Font: Not Italic

Deleted: *-neighbor*

Formatted: Font: Not Italic

Deleted: *ratio*

Deleted: predict genotypes and identify individuals.

Deleted: variant call set

Deleted: known

Deleted: . Figure 2c shows the accuracy of linking versus number of deletions used in linking attack. The

Deleted: perfect

Deleted: . The attacker can also perform linking by first predicting *existence of deletion*

Deleted:) and using the identified deletions to perform linking (Methods Section). When he/she

Deleted: this criteria, around 60 deletions are required for perfect linking.

Deleted: We also studied the scenario where the adversary does not have access to the deletion loci but aimed at finding deletions and estimating their genotypes at the same time. This is a harder linking problem because the adversary must also correctly find deletion variants. We call this, linking attack based on joint deletion discovery and genotype prediction. Figure 2d shows that the linking accuracy is maximized (around 60%) when the attacker utilizes the top 50 deletion candidates in linking. If the attacker uses the existence of variant criteria in linking, the linking accuracy decreases. ¶
In the previous analysis, the SV discovery set and RNA-seq sample set are matching. Since this may introduce a bias, we studied linking attack where signal profiles are generated by the GTex Project Consortium [20, 21] and th...

using a relatively small number of variants, i.e., 20 variants (Fig 3c). An interesting observation is that when the attacker increases the number of variants used in the attack, the linking accuracy decreases. This is caused by the fact that the additional variants after the 20 variants are incorrectly genotyped and decrease the accuracy of linking. In simple terms, the additional variants act as noise and decrease linking accuracy.

Following this, one question that arises is whether the adversary can assign reliability score to the linked individuals. We used whether *first distance gap* (Methods Section) is suitable for evaluating the reliability of linkings. This is important because when the overall linking accuracy is low, e.g. smaller than 50%, unless the attacker has a systematic way of selecting correct linkings, there is not high risk. As a test case, we focused on the linking where the adversary uses 200 deletions where the overall linking accuracy is 35%. Figure 3d shows the sensitivity and specificity with changing *first distance gap* metric. The adversary can link 10% of the individuals with perfect specificity and 20% of the individuals are linked with around 90% specificity. Figure 3d also shows the average sensitivity and specificity over 100 random selections of the linkings. As expected, the specificity is always around 35% and average sensitivity is also always smaller than *first distance gap* based selection of linkings.

2.4. Linking Attacks using ChIP-Seq Signal Profiles

We next focused on predictability versus ICI of large deletions, which are longer than 1000 base pairs. In this analyses, we utilize the ChIP-Seq signal profiles. Several recent studies have generated individual level epigenomic signal profiles through ChIP-Seq experiments [27–29]. 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 majority of the individuals have matching structural variant genotype information in the 1000 Genomes Project. The histone modifications are especially useful for identifying deletion genotypes because some of them cover a large portion of the genome, which is useful for predicting deletion genotypes. In addition, the histone modification ChIP-Seq signals create different profiles such that they can be complementary to or overlapping with each other. It is worth nothing 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 use these personalized epigenomic signal profiles for quantifying how much characterizing information leakage they provide. For any individual where there are multiple histone mark ChIP-Seq signals, we pool the signal profiles and compute several features for each large deletion. These are then used for quantifying information leakage (Methods Section). First we computed π_{GW} versus ICI using the panel of large deletions in 1000 Genomes Project. Figure 4a,b show π_{GW} versus ICI for the large deletions from the 1000 Genomes. We use the personal epigenome profiling ChIP-Seq datasets presented in studies by Kasowski et al and Kilpinen et al (Methods Section). Similar to the small deletion analysis, it can be seen that for both datasets there are many large deletions with high predictability and high ICI.

We next focused on instantiating linking attacks using ChIP-Seq profiles. We again utilize a variant of the outlier based genotyping in the linking attack. The genotyping of the panel of large deletions is done as follows. The average ChIP-Seq signal on each deletion is computed and the variants are sorted with respect to their average signal in increasing order. The deletions with smallest ChIP-Seq signal are

Deleted: with Large Deletion Prediction from

Deleted: long

Deleted: variants interact

Deleted: them then

Deleted: ¶

Deleted: for

Deleted: performed practical

Deleted: attack utilizing the genotyping of known deletions and deletion discovery followed by genotyping.

Deleted: genotype prediction

Deleted: genotype prediction

Deleted: pooled

Deleted: pooled

assigned homozygous deletion genotype. For the deletions with assigned genotypes, we identified the individual in the genotype dataset (from the 1000 Genomes project) whose genotypes match closest to the assigned genotypes. We repeated this linking attack with different number of windows and computed the accuracy of linking (Methods Section). Figure 4c shows the accuracy of linking attack based on *genotyping only* scenario, where the adversary is assumed to have access to the large deletion panel from 1000 Genomes. The linking accuracy reaches 100% with fairly small number of deletions for both datasets. For the *joint discovery and genotyping* scenario where the adversary first discovers deletions then genotypes them, the accuracy is also very high with small number of identified deletions (Fig 4d).

An interesting question about histone modifications is which combinations of histones leak the highest amount of characterizing information. To answer this question, we studied the individual NA12878, for which there is an extensive set of histone modification ChIP-Seq data from the ENCODE Project[20]. We have evaluated whether different combinations of histone modifications render NA12878 vulnerable against a linking attack among 1000 Genomes individuals, which is illustrated in Fig 4e. In general, we have observed that NA12878 is vulnerable when the dataset combinations that cover the largest space in the genome. This can be simply explained by the fact that when histone marks cover more space, higher number of deletions can be predicted. For example, H3K36me3 and H3K27me3, an activating and a repressive mark respectively, are mainly complementary to each other and they render NA12878 vulnerable. In addition, H3K9me3, a repressive mark that expands very broad genomic regions, renders NA12878 vulnerable in several combinations with other marks. On the other hand, H3K27ac, an activating histone mark that spans punctate regions do not render NA12878 vulnerable.

2.5. Linking Attacks using Hi-C Matrices

We also asked whether a relatively new data type, Hi-C signal profiles can be used for identification of genomic deletions. Hi-C is a high throughput method for identifying the long range genomic interactions and three dimensional chromatin structure[26]. It is based on proximity ligation of the genomic regions that are close-by in space followed by high throughput sequencing of the ligated sequences. After sequencing data is processed, it is converted to a matrix where the entry (i, j) represents the strength of interaction between i^{th} and j^{th} genomic positions. To study leakage from Hi-C datasets, we again focused on NA12878 individual for whom Hi-C interaction matrices are generated at different resolutions[30]. In order to convert the matrix into a genomic signal profiles, we summed the interaction matrix along columns and obtained a signal profile along the genome (Fig 5a, Methods Section). This way, we are simplifying the multidimensional nature of the Hi-C contact matrix and treat it as a sequencing assay that spans the entire genome. 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, we computed the average Hi-C signal. We next sorted the deletions in increasing order and assigned top 1000 windows with homozygous deletion genotype. We next compared the predicted genotypes with all the genotypes in the 1000 Genomes project. NA12878 is vulnerable to this attack when the Hi-C contact matrix resolution (bin length) is 10 kilobases or smaller (Fig 5b).

It is important to clarify that we are focusing on using the final output of Hi-C data, i.e., the Hi-C contact matrix, for generating a genome-wide signal profiles and performing a linking attack. We are not studying the possibility of discovering complex structural variants using the paired-end reads of Hi-C experiment, which is a different problem by itself[31]. It also requires access to mapped reads, which we

Deleted: accuracy when known

Deleted: deletions are used in linking

Deleted: , i.e. deletion loci are not given

Deleted: [18].

Deleted: Genome-wide

Deleted: with Large Deletion prediction from

Deleted: [24]. It is based on proximity ligation of the genomic sequences that are close-by and

Deleted:

Deleted: Next we

Deleted: this

assume the attacker does not have. As we explained above, our attack scenario treats the Hi-C data as any type of sequencing data and uses the linear genomic signal profile to identify deletions for the purpose of linking datasets. We are highlighting the fact that Hi-C interaction matrices themselves leak substantial amount of characterizing information.

2.6. Anonymization of Signal Profiles

An important aspect of the genomic privacy is risk management and protection of datasets. For protection, anonymization of the datasets is the most effective way to share the data publicly in a safe manner. The most effective way to protect against linking attack scenario is to ensure that the deletion genotypes are not predictable from the signal tracks. We believe RNA-seq signals are currently the most vulnerable against the linking attacks and protection of these datasets against prediction of deletion variants is most immediate. As we showed in previous sections, the small deletions are major source of leakage of genetic information from RNA-seq signal profiles. We propose systematically removing the dips in the signal profiles as a way to anonymize the RNA-seq signal profiles against prediction of small deletions. Specifically, we propose smoothing the signal profile using median filtering locally around a given panel of deletions (Methods Section). We have observed that median filtering removes the dips in the signal very effectively while conserving the signal structure fairly well. To evaluate the effectiveness of this method, we applied signal profile anonymization to the RNA-seq signal profiles generated from the datasets generated by GEUVADIS Project consortium and the GTex Project Consortium. After application of the signal profile anonymization, we observed that the large fraction of the leakage is removed for GTex datasets (Fig 2b and 3b). For GEUVADIS datasets, there is still some leakage but the genome-wide predictability of the variants are decreased substantially (Fig 2a). We also observed that the extremity based linking attack proposed in the previous section is ineffective in characterizing individuals such that no individuals are vulnerable for GTex project and at most 1% of the individuals are vulnerable for GEUVADIS dataset. The anonymized signal profiles for GTex and GEUVADIS individuals can be downloaded from privaseq2.gersteinlab.org/Anonymized_Signal_Profiles. It is worth noting that the anonymization of ChIP-Seq datasets against small deletion genotyping based linking attacks can be performed using above approach. However as we have shown in previous section, there is significant leakage when large deletions are genotyped using ChIP-Seq datasets. We observed that the median based anonymization of signal profiles is not very effective against large deletion genotyping based attacks.

3. Discussion

We have systematically analyzed a critical source of sensitive information leakage from the signal profile datasets, which were previously thought to be largely secure to share. Specifically, our results show that an adversary can perform fairly accurate linking attacks for characterizing individuals by prediction of structural variants using functional genomics signal profiles. Although we are focusing mainly on RNA-seq and ChIP-Seq signal profiles, the linking attack scenario and the measures that we presented 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[32, 33] can be vulnerable to the linking attack scenario that we presented. On another note, the practical linking attack instantiations that we presented are data-driven and can be applied to any signal profile. We believe that many more genome-wide omics technologies will be developed in the near

Deleted: against Linking Attacks

Deleted: that indicate deletions

Deleted: and

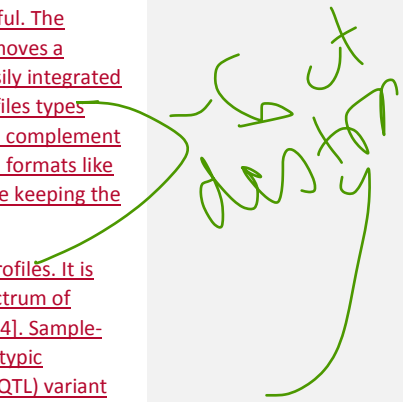
Deleted: Supplementary Figure 6).

Deleted: <#>Discussion¶
We have systematically analyzed a critical source of sensitive information leakage from the signal profile datasets, which were previously thought to be largely secure to share. Specifically, our results show that an adversary can perform fairly accurate linking attacks for characterizing individuals by prediction of structural variants using functional genomics signal profiles. In addition, we also showed that the linking can be done by predicting fairly small number of variants (generally less than 100 variants). Although the functional genomics assays do not reveal the full spectrum of structural variants, our results show that these data leak enough information for individual characterization among a fairly large set of individuals. This can be rather problematic because several large consortia are offering these signal tracks publicly. For example GTex signal profiles are publicly available through the UCSC Genome Browser. In addition, ENCODE RNA-Seq and ChIP-Seq signal profiles for several personal genomes (NA12878 and HeLa-S3) are downloadable through the UCSC Genome Browser and ENCODE Project's portal. Given the extent of public sharing of datasets, we believe that the anonymization of signal profiles using the signal processing technique that we proposed is very useful. The technique we proposed applies a minor signal smooting around all the known deletions and removes a significant amount of information. The anonymization procedure can be easily integrated into existing functional genomics data analysis pipelines. It can handle all the widely used files types including bigwig, wig, and bedGraph.¶
We also proposed a new metric for measuring the predictability of deletions from signal profiles. This measure of predictability is complementary to the sample-wide genotype predictability measure proposed earlier [12]. Sample-wide predictability measure is computed when genotypes are predicted from sample-wide datasets, for example from a sample-wide gene expression profiling datasets. Sample-wide predictability is suitable when adversary utilizes sample-wide phenotypic measurements to predict genotypes in a linking attack (Supplementary Fig 1). This scenario is meaningful when the variant genotype is of high frequency and affects phenotype among samples, e.g. quantitative trait loci. For the rare variants sample-wide predictability will not be effective because variant genotypes do not show much variation among samples. For these variants, genome-wide predictability can be computed for each individual separately (Supplementary Fig 1). The genome-wide and sample-wide predictability of genotypes must be studied together in a risk assessment procedure while functional genomics datasets are being shared. ¶

future[34]. The genome-wide signal profiles will be vital source of information in the analysis of these datasets. The framework we presented here can be utilized for assessing the leakage and protecting these datasets.

We showed that the linking can be done by predicting fairly small number of variants (generally less than 100 variants). Although the functional genomics assays do not reveal the full spectrum of structural variants, our results show that these data leak enough information for individual characterization among a fairly large set of individuals. This can be rather problematic because several large consortia are offering these signal tracks publicly. For example GTex signal profiles are publicly available through the UCSC Genome Browser. In addition, ENCODE RNA-Seq and ChIP-Seq signal profiles for several personal genomes (NA12878 and HeLa-S3) are downloadable through the UCSC Genome Browser and ENCODE Project's portal. 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. It can handle all the widely used files types including bigwig, wiggle, and bedGraph. We believe that this anonymization technique can complement other approaches for removing genetic information from shared datasets. For example file formats like MRF[19] and tagAlign[20] can enable removing raw sequence information from reads while keeping the information about read mapping intact.

We also proposed a new metric for measuring the predictability of deletions from signal profiles. It is important to note that this measure of predictability is more inclusive, in terms of the spectrum of variants that it can be applied to, than the sample-wide genotype predictability measure[14]. Sample-wide predictability measures how well variants can be genotyped given a sample of phenotypic measurements from multiple individuals. For example, expression quantitative trait loci (eQTL) variant genotypes of multiple individuals can be predicted from the gene expression levels of the individuals. Sample-wide predictability is suitable when adversary utilizes sample-wide, i.e. measurements from multiple individuals, phenotypic measurements to predict genotypes in a linking attack. By definition, the sample-wide predictability of rare variants will not be high because sample-wide predictability relies on the fact that the genotype-phenotype relation is statistically detectable within a sample of individuals. Since the phenotypic effects of rare variants are not easily detected, their sample-wide predictability is not high. For example, the structural variants that affect gene expression have very low population frequency and thus they have low effect on the expression level when the effect is averaged over all the individuals. Thus, the sample-wide predictability of these are rather low (Supplementary Fig 1). Genome-wide predictability, on the other hand, is for each individual separately. A variant, independent of its population frequency, can have a high genome-wide predictability. Following previous example, the structural variants have high genome-wide predictability because they do have very obvious effect on the genome-wide signal profiles. Thus, genome-wide prediction strategy can predict both high and low frequency variant. The sample-wide and genome-wide prediction approaches underpin different paths to linking attacks. They must be studied together in a risk assessment procedure while functional genomics datasets are being shared.



Formatted: Font: 11 pt

4. Methods

We provide the details of the computational methodologies. We first introduce the notations. The genomic deletions are intervals of genomic coordinates. We refer to them simply as intervals, e.g. a deletion between genomic positions i and j by $[i, j]$. The genotype of a genomic deletion at $[i, j]$ is denoted by $G_{[i,j]}$, which is a discrete random variable distributed over the 3 values $\{0,1,2\}$. These values correspond to the three genotypes of the deletion and they represent how many copies of the genomic sequence is deleted. The functional genomics read depth signal is denoted by $\boldsymbol{S}$, which is a vector of values corresponding to each genomic position**.** The signal level at genomic position at i is denoted by $\boldsymbol{S_i}$. An important quantity that we utilize in formulating methods is the multi-mappability profile of the deletion regions. The multi-mappability is a signal profile that measures, for each position in the genome, how uniquely we can map reads. The multi-mappability signal is denoted by $\boldsymbol{M}$**,** which is a vector of multi-mappability signals for all the genomic positions and the signal at genomic position i is denoted by $\boldsymbol{M_i}$. 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[35] 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[36].

Deleted: [31]

Field Code Changed

Deleted: 32

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. We assume that the adversary performs prediction by extracting features from the functional genomics signal profile. We define here the features that are most useful for genotyping deletions (Supp Fig XX). Given a $[i, j]$, an important feature for genotyping the deletion is the average functional genomic signal within the deletion:

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

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

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

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. We compute the average signal in the neighborhood as

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

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

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

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

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

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

The practical instantiation of the linking attacks that we study are based on genotyping of small deletions using extremity based statistics of functional genomics data. In addition, when a panel of deletions is not available, the adversary also discovers the deletions using the signal profile. For GEUVADIS and GTex datasets, we perform small deletion genotyping using RNA-Seq signal profiles. The basic idea behind genotyping of deletions is the fact that there is a sudden dip in signal profile whenever there is a deletion (Fig XX). In order to detect these dips, we observed that *self-to-neighbor signal ratio* is very useful for genotyping small deletions. For all the small deletions, *self-to-neighbor signal ratio*, $\rho_{[i,j]}$, neighbor signal balance, $\eta_{[i,j]}$, and average neighbor signal are computed. We then filter out the small deletions whose multi-mappability signal is larger than 1.5 or average neighbor signal (τ) is smaller than 10 or $\eta_{[i,j]}$ is smaller than 0.5. For the remaining set of small deletions, we sorted the deletions with respect to increasing $\rho_{[i,j]}$. The deletions which are at the top of the sorted list correspond to the deletions which are highly mappable (low multi-mappability signal), have strong neighbor signal support (high average neighbor signal), and finally they have a strong signal dip on them (Low $\rho_{[i,j]}$, and high $\eta_{[i,j]}$). We selected the top n deletions and assigned them homozygous genotypes, i.e., $G_{[i,j]} = 0$. The

basic idea is that the deletions with strongest signal dips are enriched in homozygous deletions. It is worth noting that this genotyping method only assigns homozygous genotypes. Although this results in low genotyping accuracy (Supp Fig XX), these genotyping predictions have enough information for accurate linking attacks.

We utilize pooled ChIP-Seq read depth signal profiles and Hi-C signal profiles for genotyping large deletions. For genotyping the large deletions, we first computed the average signal ($\frac{\sum_{i'=i}^{j} \boldsymbol{S}_{i'}}{j-i+1}$) and average multi-mappability signal ($\frac{\sum_{i'=i}^{j} \boldsymbol{M}_{i'}}{j-i+1}$) on each large deletion. Then we filtered out the large deletions for which the average multi-mappability signal is larger than 1.5. We then sorted the remaining deletions with respect to increasing average signal profiles. For the top n deletions, we assigned homozygous genotypes, i.e., $G_{[i,j]} = 0$.

For the case when the adversary does not have access to the deletion panel, we fragment the genome into windows and use these windows as candidate deletions. For small deletions, we use 5 base pair windows within the exonic regions. For large deletions, we use 1000 base pair windows over all genome.

Deleted: loci are not known to the adversary

4.3. Details of the Instantiations of Genome-wide Linking Attack

Following the genotyping of the deletions, we use the genotyped deletions to link the individual to the individuals in the SV genotype dataset. Given the genotyped deletions $\{[i_1, j_1], [i_2, j_2], \dots, [i_n, j_n]\}$ for the k^{th} individual in the signal profile dataset, we compute the genotype distance by comparing the genotyped deletions to the individuals in the genotype dataset:

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

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

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

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

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

Deleted: *linked individual*

Finally, if the linked individual in the genotype dataset matches the individual in signal profile dataset, we mark the individual in the signal profile as a vulnerable individual. We also compute the *first distance gap*, $d_{1,2}$, for each linked individual[14] to evaluate the reliability of linking. For a linked individual, first distance gap is computed as

Deleted: [12] to evaluate the reliability of linking.

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

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

4.4. Anonymization of Signal Profile Datasets

The anonymization of the signal profile datasets refers to the process of protecting the signal profile data against correct predictability of the genotypes for deletion variants. As we discussed earlier, the large and small dips in the functional genomics signal profiles are the main predictors of deletion variant genotypes. To remove these dips systematically, we propose using the median filtering[37] based signal processing to locally smooth the signal profile around the deletion. This signal processing technique has been used to remove shot noise in 2 dimensional imaging data and 1 dimensional audio signals[36, 38]. For each genomic a in the deletion $[i, j]$, we replace the signal level using the median filtered signal level:

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

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

5. Datasets

The mapped reads for the RNA-seq data from gEUVADIS project are obtained from gEUVADIS project web site (http://geuvadis.org/). The RNA-seq mapped reads from the GTex project are obtained from dbGAP portal. The structural variant loci and genotypes are obtained from the 1000 Genomes Project.

[[Randomized Data Generation for \Pi?]]

[[We filtered out variants with allele freq > 0.01]]

[[GTex data is just blood, we did not use other data]]

Figure Legends

Figure 1:

Figure 2:

Figure 3:

Figure 4:

Figure 5:

Figure S1:

Figure S2:

Deleted: [33]

Deleted: [34] and in genomic signal smoothing[32].

Figure S3:

Figure S4:

REFERENCES

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

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

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

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

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

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

7. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson J V., Stephan DA, Nelson SF, Craig DW: **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays**. *PLoS Genet* 2008, **4**.

8. Im HK, Gamazon ER, Nicolae DL, Cox NJ: **On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy**. *Am J Hum Genet* 2012, **90**:591–598.

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

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

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

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

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

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

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

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

Deleted: 10

Deleted: 11

Deleted: 12

Deleted: 13

Deleted: 14

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

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

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

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

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

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

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

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

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

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

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

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

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

30. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping**. *Cell* 2014, **159**:1665–1680.

Deleted: 15

Deleted: 16

Deleted: 17

Deleted: 18

Deleted: 19

Deleted: 20

Deleted: 21

Deleted: 22

Deleted: 23

Deleted: 24

Deleted: 25. Korbel JO, Lee C: **Genome assembly and haplotyping with Hi-C.** *Nat Biotech* 2013, **31**:1099–1101.¶
26. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, et al.: **Transcriptome and genome sequencing uncovers functional variation in humans.** *Nature* 2013, **501**:506–11.¶

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

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

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

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

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

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

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

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

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

Deleted: 32.

Deleted: 33

Deleted: 34