bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Private information leakage from functional genomics data: Quantification with calibration experiments and reduction via data sanitization protocols

Gamze Gürsoy[1,2], Prashant Emani[1,2], Otto A. Jolanki[3], Charlotte M. Brannon[1,2], Arif Harmanci[4], J. Seth Strattan[3], Andrew D. Miranker[2,5] and Mark Gerstein*[1,2,6]

[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]Stanford University School of Medicine, Department of Genetics, Stanford, CA, 94305, USA

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

[5]Department of Chemical and Environmental Engineering, Yale University, New Haven, CT 06520, USA

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

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

December 28, 2019

*pi@gersteinlab.org; Corresponding Author

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Functional genomics experiments provide data on aspects of gene function in a variety of conditions and how they relate to organismal phenotype (e.g. "genes upregulated in AIDS"). These experiments do not necessarily concern findings on identifiable individuals, leading to a neglect of their privacy issues; however, for each experiment, it is possible to create "cryptic quasi-identifiers" statistically linking them back to individuals and thereby leaking sensitive phenotypic information (e.g. "HIV status"). Here, we develop metrics for quantifying this leakage and instantiate them in practical linking attacks. As genotyping noise is a crucial quantity for the feasibility of attacks, we perform them both with highly accurate reference genomics datasets as well as by generating RNA and DNA data from more realistic environmental samples. Finally, in order to reduce leakage, we develop a data-sanitization protocol for making principled privacy-utility trade-offs, permitting the sharing of functional genomics data while minimizing risk of leakage.

Keywords: genome privacy, functional genomics, linkage attacks, surreptitious DNA sequencing, data sanitization

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

The decreasing cost of DNA sequencing technologies and the clinical importance of genomic characterization of individuals have resulted in an exponential increase of available genetics data [1], in turn, genomic privacy has recently emerged as an important issue [2]. By its very definition, raw genomic data identifies the owner. As an example, in the famous Wisconsin case the DNA (without the name of the person) was enough to charge an individual with a crime [3]. Genomic data is largely shared with close family members. Therefore, even in instances where patients have provided broad permission to use and access their genomic information, care must be taken to preserve patient privacy as the data implicates not only the immediate owner of the sequence but many third-party relatives as well. Although privacy is enforced fundamentally via laws and social codes, the increasing use of genetic information in new avenues outpaces the legal efforts. For example, easy-to-use, accessible, portable and cheap sequencing technologies such as the Oxford Nanopore, access to large databases by citizen scientists such as the AllofUs program, and growing movements like "do-it- yourself science" [4] create concerns over privacy. For example, one of the missions of AllofUs program is to make data broadly accessible to researchers of all kinds, including citizen scientists.

Recently, there has been a rapid increase in large-scale "omics" datasets such as those for gene expression and chromatin state (i.e., functional genomics data), which are becoming clinically actionable (e.g., The Cancer Genome Atlas). In contrast to DNA sequencing data, due to the targeted nature of functional genomics assays [exons, transcription factor (TF) binding sites, etc.], sequences obtained from functional genomics data often cover only a small portion of the genome and are subject to various biases and base changes (e.g., RNA editing, methylation). Therefore, genotypes obtained from functional genomics data are noisy and incomplete and, hence, may not readily be used for building personal genomes (for example, single-cell RNA-Seq reads contain variants on the exons expressed only in a single cell). Thus, there is less concern over privacy and

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a great desire to share this data, leading to many large-scale projects publicly sharing functional genomics data. Unique privacy issues related to functional genomics stem from the fact that functional genomics data are produced with the intention of learning about the biology and are often coupled with phenotypic information about the subject that is potentially sensitive (e.g., samples taken from tumors or from patients that are HIV positive). 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 private information in the data can be revealing about the individual from which they came. Therefore, as opposed to DNA sequencing data where there are concerns over genotypic information leakage, the key privacy issue with functional genomics data is the leakage of phenotypic information.

A common breach of privacy is known as a "linkage attack", where adversaries combine auxiliary information about an individual from dataset A to link the individual to anonymized data in dataset B that might contain sensitive information. There are a few ways to perform linkage attacks (Figure S1): Case 1: The most common way is to obtain perfect quasi-identifiers about an individual (e.g., data of birth, zip code, gender) from dataset A and overlap these with the information in dataset B that might be sensitive. In this scenario, these quasi-identifiers alone do not reveal information about an individual but are correct pieces of information (noise free) about the individual. A famous example of this type of linkage attack in genomics is the revealing of addresses of Personal Genome Project (PGP) participants by cross referencing birthdate and zip code information present in PGP and Census datasets [5]. Case 2: Linkage attacks become more difficult if the collected auxiliary information about the individual in dataset A may or may not be correct (noisy). Narayanan and Shmatikov, in a non-genomics setting, showed that adversaries with a small amount of, and partially false, background information about an individual can de-anonymize an anonymized dataset without relying on a fixed set of quasi-identifiers by developing formal models for privacy breaches [6]. In this work, the authors used the public Netflix dataset for

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

empirical cross-referencing. Netflix released an anonymized dataset of movie ratings from thousands of viewers. The authors then used the Internet Movie Database (dataset A), in which the identities of many users are public but only some of their movie choices are available, and linked it to the Netflix dataset (dataset B). This revealed the identities and personal movie-preference information of many users in the Netflix dataset. Similar to the case of the IMDB-Netflix linkage attack, due to the presence of noise, the pieces of variant information inferred from functional genomics data may not be quasi-identifiers themselves, but can be statistically linked to other variants in a different dataset in order to reveal sensitive phenotypic information about these individuals. Case 3: Linkage becomes even more difficult if adversaries infer noisy auxiliary information about an individual and link them to a dataset with noisy attributes. An example we investigated here is whether a linkage attack can be performed on a noisy phenotype dataset (functional genomics data; i.e., dataset B), when we use noisy genotypes about a known individual as auxiliary information (dataset A).

In this study, we adopted formal definitions of privacy breaches in Narayanan and Shmatikovs work for functional genomics data [6] and developed statistical measures for the leakage of sensitive information in linkage attacks. We first tested these measures with publicly available data, in which the linked dataset is noisy (i.e., genotypes from gEUVADIS RNA-Seq data [7] with phenotypes) and the background data is perfect (i.e., 1,000 Genomes genotypes [8]). We showed that our measures could conceptually link known genomes to phenotypes linked to functional genomics data using publicly available datasets. This is a scenario where we link "perfect data" to "noisy data". However, with current protections in place, it is unlikely that adversaries would have access to perfect genotype panels. In an effort to validate our formalism in a real-world setting and to understand the robustness of our measures in the presence of real noise, we collected used coffee cups and blood tissues from consented individuals and performed genotyping on extracted DNA from the coffee cups and RNA-Seq on the blood. The DNA from the coffee cups resulted in

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genotypes with different noise profiles. However, we were able to successfully link the majority of the coffee cups to a database of individuals with only RNA-Seq data available. We tried three different technologies (Illumina low coverage sequencing, Illumina SNP chip, Oxford Nanopore sequencing) to obtain noisy genotypes from coffee cups in order to examine the effect of different available technologies.

Despite privacy issues, there is utility of raw functional genomics data and benefits to open data sharing. Access to private data requires complex, often overly bureaucratic, user agreements. Open data helps advance biomedical data science not only by easing access to the data, but also by helping with speedy assessment of tools and methods, and in turn, reproducibility. Funding agencies and research organizations are increasingly supporting new means of data sharing and new requirements for making data publicly available while preserving participants' privacy [9]. It is important to find a way to get around privacy protection issues as large-scale mining of functional genomics data will allow researchers to genetically and environmentally characterize the disease state and susceptibility in detail. For example, one can annotate significant non-coding genome-wide association study (GWAS) hits by looking at the status of the functional genome. Therefore, the more broadly shared these data are, the more likely they are to promote biomedical advances. Hence, converting these datasets to formats that can be shared publicly without compromising patients privacy is important and relevant. To be able to remove sensitive information from data while keeping a high utility of the data, one needs to quantify the amount of sensitive information in the data. To this end, we calculated the amount of information leakage in functional genomics data as a function of sequencing coverage. Based on our findings, we developed data sanitization techniques that allow public sharing of read alignments of functional genomics experiments, while protecting the sensitive information and minimizing the amount of private data that requires special access and storage. Our file format manipulation system is based on optimizing the balance between privacy and utility.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2 Results

2.1 Linkage attack scenario

The attack involves cross-referencing a known individual with in a genotype dataset against the individuals in a functional genomics dataset. Let us assume we have a study that aims to understand the changes in gene expression in the blood for HIV-positive vs. HIV-negative samples. This study creates a database with the RNA-Seq data from a cohort of individuals with HIV-positive and -negative phenotypes. The genotype dataset contains the genotypes of a panel of variants and belongs to a group of individuals whose identities are known. The adversary obtains access to the genotype dataset by lawful or unlawful means (e.g., the adversary might have stolen it or might be allowed to access it but violated the terms of accession). The genotype dataset can also be a set of genotypes extracted from coffee cups or used glasses as there are currently no laws or policies preventing citizens from collecting and sequencing the DNA left on discarded items.

The attack starts with genotyping the raw RNA-Seq data. After genotyping, the adversary builds a data matrix with the observed genotypes, which is denoted by F. This data matrix contains partial and noisy genotypes as genotyping from RNA-Seq may contain errors and cannot cover the entire genome as well as HIV status. The adversary then cross-references the individuals in F to an individual or individuals in the genotype dataset G by comparing the genotypes. The adversary finds the individuals that best match each other based on statistical significance. The results are used to link an individual with background information (i.e., genotypes) to those in the RNA-Seq dataset and the sensitive information (e.g., HIV status of individuals in the genotype dataset are revealed to the adversary).

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.2 Quantification of information leakage and linking reliability

As in the case of preference databases (e.g., Netflix) [6], the functional genomics dataset cohorts are necessarily sparse, meaning each individual in the dataset contains values for only a small fraction of variants (e.g., 20,000 called variants from 4,000,000 possible variants, see Supplemantary Information for definitions of sparsity and Figure S2 for more details). Narayanan and Shmatikov [6] formally and empirically showed that robust de-anonymization of large sparse datasets can be done when large numbers of false positives and false negatives are present in the cross-referenced data. We adopted their matching system to functional genomics data as follows: Let us assume that M variants are called from functional genomics data for an individual j is defined as $S(j)^F = \{s(j)_1^F, .., s(j)_i^F, ..., s(j)_M^F\}$ and $s(j)_i^F = \{v(j)_i, g(j)_i\}$, where $v(j)_i$ consists of the location and alternative allele information and $g(j)_i$ is the genotype of the variant i as 0 for homozygous reference allele, 1 for a heterozygous alternative allele and 2 for a homozygous alternative allele. For a known individual k in the genotype dataset G, we can define N total variants as $S(k)^G = \{s(k)_1^G, s(k)_2^G, .., s(k)_i^G, ..s(k)_N^G\}$. The intersection between the $S(j)^F$ and the $S(k)^G$ can be defined as $I(j,k) = S(k)^G \bigcap S(j)^F$, which contains T number of variants ($I(j,k) = \{s(j,k)_1, .., s(j,k)_i, ..., s(j,k)_T\}$). We then score this intersection for a known individual k by using a bit metric $L(j,k)$, between two sets as $L(j,k) = -\sum_{t=1}^{t=T} log_2(p(s(j,k)_t))$, where $p(s(j,k)_t)$ is a rarity measure weighing each variant by its frequency in the cohort, which is the ratio of the number of individuals with variant $s(j,k)_t$ and the total number of individuals in the cohort, F. Using this estimated information surprisal per genotype with respect to the attacked database allows us to properly weigh the overlapping variants and rewards the score for a given individual j if there is a matching rare genotype. For example, a set that contains many common genotypes will not be very useful for pinpointing individuals, whereas rare variant genotypes will give more information. This is also similar to giving higher weight to statistically rare attributes in the IMDB-Netflix attack [6] to improve the robustness against incorrect observations.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To find the matching individual in the functional genomics data, we then ranked all the $L(j,k)$ scores for a given query individual k in decreasing order. We denoted the individual with the largest score is the queried individual, which reveals the sensitive phenotypic information coupled with the functional genomic data to the attacker. To assess the statistical robustness of this prediction, we defined a measure called gap_k, which is the ratio between the $L(j,k)$ score of the first ranked individual ($max = max(L(j,k))$ for all j) and that of second ranked individual ($max_2 = max_2(L(j,k))$ for all j and $gap_k = max/max_2$). The idea is to determine how separated the predicted individual is from the rest of the individuals in the cohort. We empirically calculated the random chance of observing this separation equal to or greater than gap_k by randomly subsetting N variants (the same number of variants as the background information) from the genotype panel, performing the linking attack, and calculating the associated gap value. If this p-value is statistically significant, then the attacker can rely on the prediction (Figure 1A and 2A).

2.3 Empirical and Experimental Demonstrations

Case 2 involves linking perfect background information to a noisy dataset. For this case, we used the genotypes inferred from RNA-Seq data taken from 436 individuals of the gEUVADIS project [7] as the functional genomics dataset, and the high-coverage whole-genome sequencing (WGS) of the same individuals from the 1,000 Genomes project [8] as the stolen genotype panel (Figure 1A). Among the 436 individuals, 421 were present in the genotype panel; we linked all of these to the phenotype panel with p-value of $< 10^{-2}$ (Figure 1B). We also showed that 14 remaining individuals that are not in the genotype panel have overlap with the phenotype panel with gap values equal to or smaller than the random gap values, which indicates that our algorithm is capable of detecting individuals that are not in the database. We first assessed the amount of noise in the attacked database and found that among all the called variants from the RNA-Seq samples, on average there was less than 10% sensitivity (ratio of the number of correctly called variants to all variants of an individual); the precision (ratio of correctly called variants to all called variants)

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of the called variants was around 30%. In addition, RNA-Seq missed 66% of the variants on average (Figure 1A and Figure S3). We then investigated the robustness of our algorithm to the false positives in the attacked database by adding an increasing amount of false-positive variants to the variants called from RNA-Seq by keeping the number of true positive variants constant. We found that 418 out of 421 individuals were successfully and significantly linked to the cohort even after adding 100,000 false-positive variants to the dataset (Figure 1C). The linking accuracy decreased to 20% only after adding one million false-positive variants (nearly a quarter of the total number of variants in an individual genome, Figure 1D). Interestingly, we observed that the RNA-Seq data from individuals with African ancestry, despite having the same coverage as the rest of the RNA-Seq data, was more vulnerable to linkage attacks (Figure 1D). This is likely due to a higher number of heterozygous or homozygous alternative alleles in the African genomes compared to the reference genome.

The above attack represents a scenario where the phenotype dataset (RNA-Seq) is noisy but the genotype dataset is noise free. However, in a real-life scenario, one might imagine that obtaining a noise-free, high-coverage genotype dataset could be unlikely with the current protections in place. Instead, we can consider more realistic scenarios: given a phenotype dataset and access to imperfect DNA samples from participants, an attacker can uncover the links between the two, and subsequently, link an individual to a stigmatizing phenotype.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

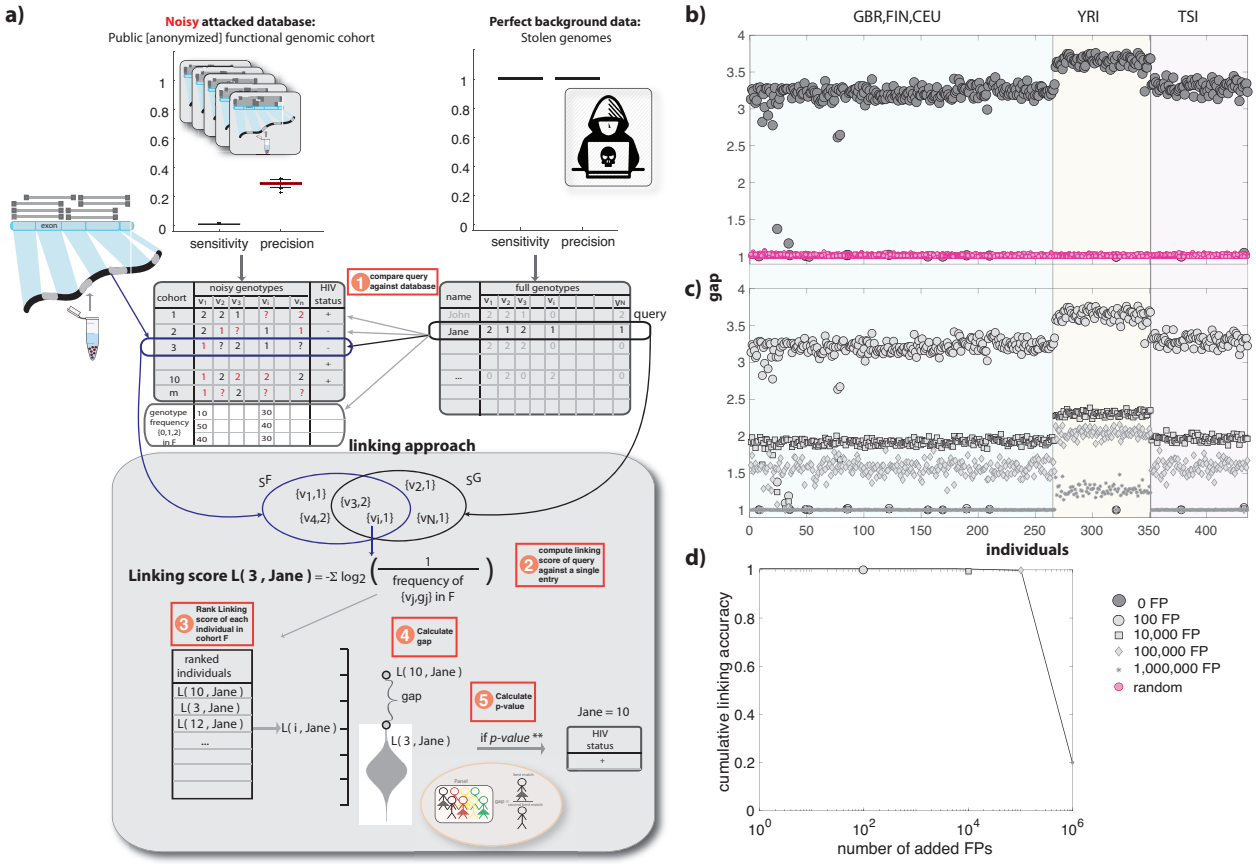

Figure 1: **Linkage Attack Case II: Linking perfect genotypes to noisy functional genomics data.** (**a**) (a) The noisy genotypes can be inferred from the functional genomics experiments. The panel of genotypes contains the variants and associated genotypes that are connected to phenotypes such as disease status for a population of individuals. Perfect (stolen) genotypes for a known individual can be searched in the phenotype panel. To do so, we score each individual in the phenotype panel based on the frequency of the overlapping genotypes between the queried individual and the individual in the phenotype panel. We then rank all the scores and select the top ranked individual as query if the ratio between the top and second best score (*gap*) is greater than random scores. (**b**) *gap* values for the 1,000 Genomes individuals in the gEUVADIS RNA-Seq cohort. A total of 421 out of 436 gEUVADIS individuals were in the 1,000 Genomes Phase III cohort. Red circles are the random gap values obtained by linking a random set of genotypes to the phenotype panel. (**c**) *gap* values for the 1,000 Genomes individuals in the gEUVADIS RNA-Seq cohort after adding false positive genotypes to the phenotype cohort F. As the total number of false positive genotypes increases, the quality of linking decreases. (**d**) Cumulative linking accuracy after adding the false positive genotypes. Linking accuracy is calculated as the ratio between the total number of correctly identified individuals with p-values < 0.01 and total number of queried individuals.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Case 3 involves linking noisy auxiliary information to a noisy dataset. For this, we first obtained blood from two consented individuals and performed RNA-Seq experiments. We combined raw RNA-Seq data from these individuals with RNA-Seq data from gEUVADIS individuals to create a phenotype dataset. We then collected six used coffee cups from the same individuals. Our aim was to extract the DNA left on the surface of the lids of the coffee cups, genotype the extracted DNA, and link the owner of the coffee cup to the phenotype dataset (Figure 2A). In particular, we investigated whether such a privacy breach is possible with little cost and when performed by someone who follows publicly available protocols. To this end, we extracted DNA from the surface of coffee cup lids followed by whole genome amplification using commercially available kits (Supplementary Information). We then used extracted and amplified DNA for Illumina low-coverage sequencing, Illumina genotyping array, and Oxford Nanopore based sequencing to call single-nucleotide variants (SNVs) and insertions and deletions (indels). This mimics the scenario in which an attacker chooses a genotyping technique (based on familiarity or cost) for the DNA extracted from coffee cups and uses the resulting genotypes to link the individuals to the phenotype database. Figure 2A shows the amount of noise both from coffee cups and RNA-Seq data by using the high-coverage PCR-free WGS data from blood as gold standards.

We performed Illumina low-coverage sequencing at 10x coverage, which resulted in an average of 78.18% and 80.36% of the DNA to map to the human reference genome for individual 1 and 2, respectively (see Table S1). After using GATK best practices, we called an average of 216,596 and 186,721 SNVs for individual 1 and 2, respectively. Among them, 55% and 49%of SNVs in the sample belonged to the individual 1 and 2, respectively (see Table S2). We then used all the called SNVs to link these individuals to the phenotype dataset with 438 individuals. We successfully linked all 12 coffee cup samples to the correct individuals in the dataset with an average *gap* of 1.82 and 2.70 for individual 1 and 2, respectively, at p-values of $< 10^{-2}$. We first investigated how much money needs to be spent for an adversary to sequence the DNA from coffee cups in order to

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

perform a successful linking attack. For that, we subsampled the low-coverage sequence down to 5x, 2.5x, 1x, 0.5x, 0.25x, and 0.125x coverage (cost from $\sim$ \$750 down to \$19) and found that our statistical measures can link the majority of these low-coverage WGS data from coffee cups to the functional genomics data (Figure 2C).

Since the adversary aims to link the coffee cups to the RNA-Seq data, an ideal and cheap alternative genotyping method could be exome-based genotyping arrays, as RNA-Seq captures reads overwhelmingly from exons of the genes. To this end, we performed Illumina exome-based genotyping arrays on the same coffee cup samples. We found that genotyping arrays have a relatively low call rate for these samples likely due to fragmented, degraded, and damaged DNA. On average, the call rate was 81% for both individuals (Table S2). Although there were many correctly genotyped SNPs in the coffee cup samples, their overlap with RNA-Seq genotypes were very low, resulting in only 2 out of 12 samples linking to the phenotype dataset (gap=1.98 and 1.73, p-value$< 10^{-2}$). This is likely due to the high number of homozygous reference alleles called from the coffee cups (genotype = 0), which are difficult to infer from RNA-Seq due to uneven depth distribution and noise.

Another popular genotyping method due to its low cost and portability is the Oxford Nanopore. Although it is well known that SNV genotype quality is quite low with single-pass sequencing with standard protocols, there is no study to date that investigates the role of easy access to privacy. DNA extracted from coffee cups is likely partially degraded, highly fragmented, and damaged. We used the simplest and most standard sequencing kit recommended by the Oxford Nanopore with multiplexing (due to the low amounts of input DNA) without the additional steps of DNA repair or suggested quality control (Supplementary Information). This was to minimize the cost and to mimic the act of a curious scientist surreptitiously gathering DNA. On average, we called 65 and 45 SNVs for individual 1 and 2, respectively. Among those, four and five SNVs were correct

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for individual 1 and 2, respectively. Among these correct SNVs, only one SNV was present in the RNA-Seq genotype data (Table S2). Moreover, only a few variants were present in the 1,000 Genomes panel, suggesting that these were false calls. Therefore, linking the coffee cups to RNA-Seq data using a standard Nanopore kit without DNA quality control or damage repair was not successful.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

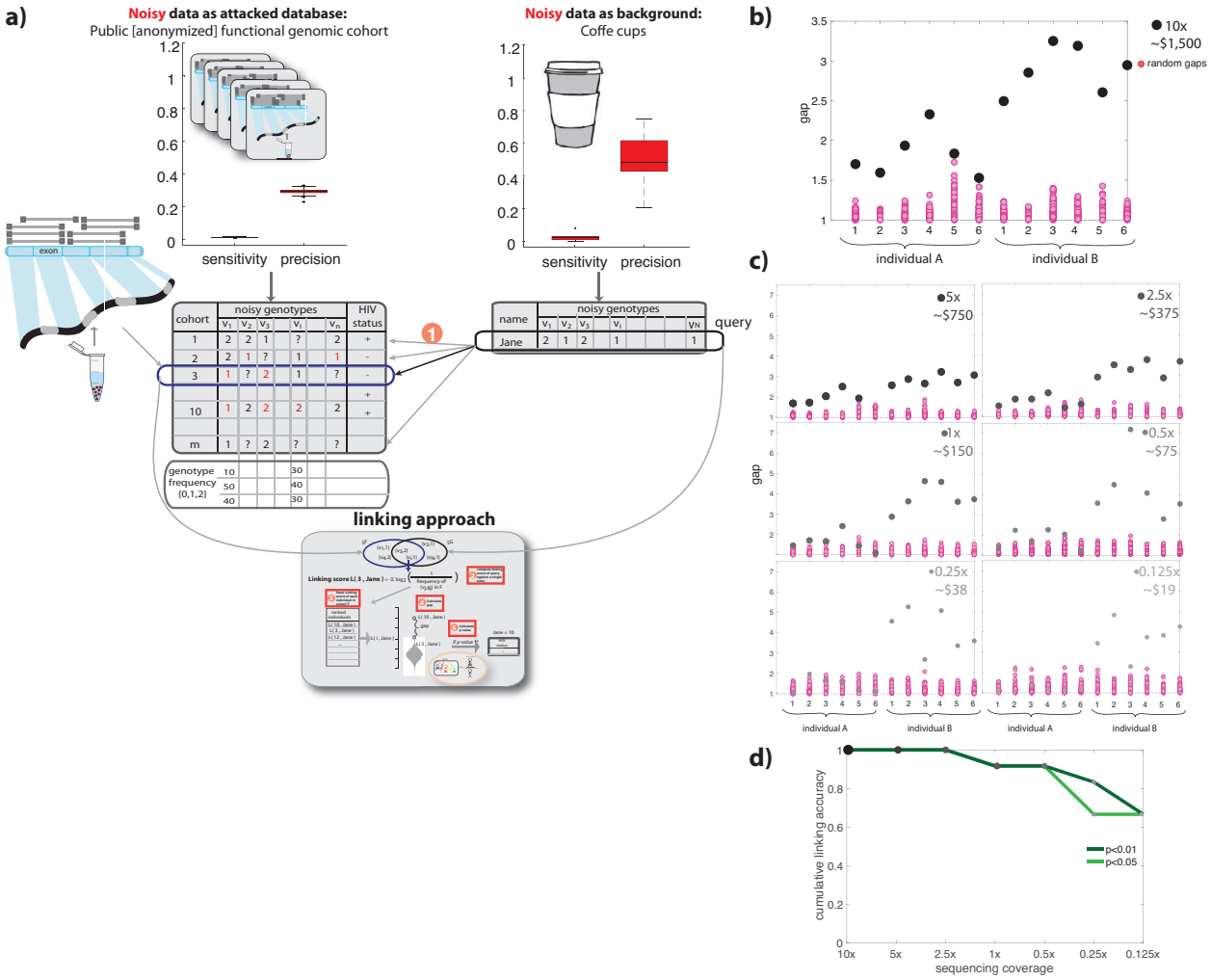

Figure 2: **Linkage Attack Case III: Linking noisy genotypes to noisy functional genomics data.** (**a**) The noisy genotypes can be inferred from the functional genomics experiments. The noisy genotypes from a known individual can be obtained by surreptitiously sequencing the DNA swabbed from a used coffee cup. We can then link this coffee cup to the phenotypes associated with the functional genomics cohort using the linking technique described in Figure 1A. (**b**) *gap* values for two individuals and four coffee cups each at 10x sequencing coverage. All of the coffee cups were successfully linked to the owner of the cups in the phenotype cohort with p-values< 0.01. (**c**) *gap* values for two individuals and four coffee cups each at different sequencing coverage and comparison with random *gap* values. (**d**) Cumulative linking accuracy after subsampling the sequencing coverage. Linking accuracy is calculated as the ratio between the total number of correctly identified coffee cups with p-values < 0.01 or p-values < 0.05 and total number of queried coffee cups.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In summary, we showed that potential privacy breaches to infer sensitive phenotypes can be done via linkage attacks through designing statistical measures. Our surprisal-based measures that depend on the frequency of the genotypes in a cohort are robust to the addition of false positive and low rates of true positives. We also showed that if adversaries use environmental samples such as coffee cups for genotyping purposes, Illumina-based next-generation sequencing yields high privacy risks when connected to functional genomics data. By contrast, genotyping arrays and the Oxford Nanopore require collection of large amounts and high-quality isolated DNA.

2.4 Information leakage in other functional genomics data and the effect of sequencing coverage

We can extend the above scenario to other types of functional genomics data [e.g., ChIP-Seq, Hi-C, ChIA-PET (a hybrid technique of ChIP-Seq and Hi-C), and single-cell based assays] and answer other questions such as how well an adversary must sequence an individuals genome to be able to perform a successful linkage. Specifically, if an adversary obtains permission to perform functional genomics experiments on a biosample, then can the adversary use the reads from these experiments to link the individual to a genotype panel? Answering questions such as this is essential to designing better counter-measures for data dissemination. To this end, we used 36 real functional genomics datasets taken from 11 real individuals (11 from the 1,000 Genomes project and 36 from the ENCODE functional genomics dataset, see also Table S3) to estimate the information leakage as a function of assay coverage. We performed a linking attack by calculating the $L(j,k)$ score between the genotypes in the database and the genotypes inferred from functional genomics experiments and evaluating them by empirical p-values based on the gap_k metric at every functional genomics sequencing coverage c.

We calculated $L(j,k)$ as a function of coverage using various functional genomics experiments. The experiments involved whole-genome approaches such as Hi-C, ChIA-PET, and Repli- Seq,

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

transcriptome-wide assays such as RNA-Seq, and targeted assays such as ChIP-Seq of histone modifications and TF binding (Table S3). In addition, we calculated $L(j,k)$ for WGS, WES, and SNP-ChIP data of individual NA12878 for comparison.

We pooled c total nucleotides from the BAM files generated by aligning the raw reads to the human reference genome and used GATK to call SNVs and indels with the parameters and filtering levels suggested in the GATK best practices (Figure 3A) [10, 11]. We aimed to estimate the information leakage in relation to sequencing coverage (c). We defined aggregate sequencing coverage c as the total number of nucleotides sequenced in a functional genomics experiment. To understand the relationship between the leaked information and the coverage, we first randomly pooled c base pairs of sequence and calculated $L(j,k)$ at that coverage (Figure 3B). We then added more coverage to the pooled reads and repeated the calculation. We repeated this procedure until we depleted all the reads of a functional genomics experiment. The overall process is depicted in Figure 3A.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

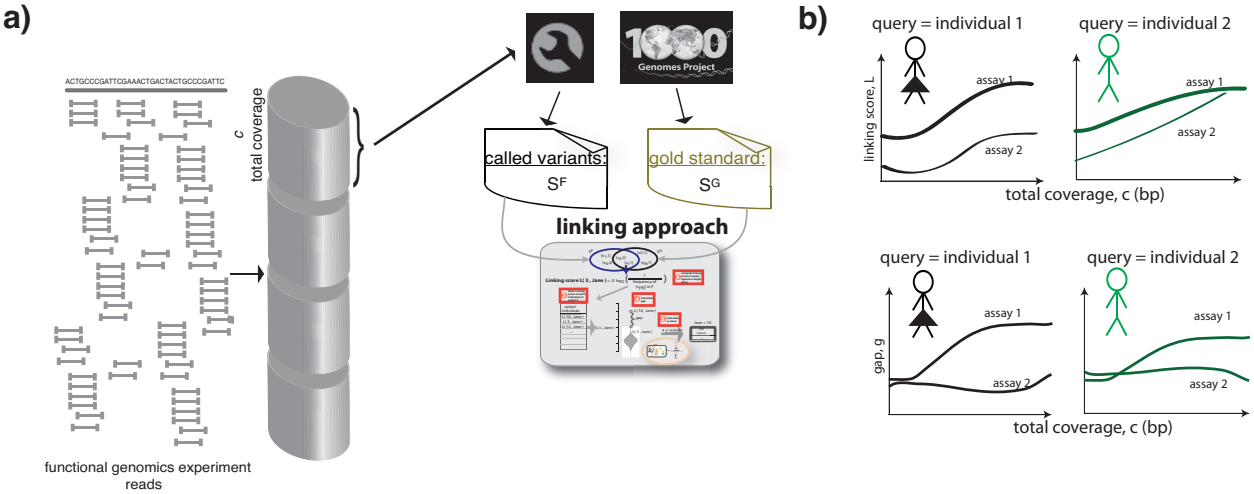

Figure 3: **The effect of sequencing coverage on linking scores and** *gap* **values** (**a**) The process of pooling reads from functional genomics experiments to calculate linking scores between 1,000 Genomes gold standard variants and called variants from functional genomics reads at different coverage. *c* amount of reads were pooled and genotyped using GATK. These genotypes were then compared against the gold standard of 1,000 Genomes genotypes. The self-information of the gold standard genotypes that are called from the functional genomics reads (true positives) is the estimated linking score. The genotypes that are called from functional genomics reads but are not in the gold standard variants are false positives. The genotypes that are in the gold standard but are not called from the functional genomic reads are false negatives. (**b**) The illustration of linking scores and *gap* as a function of sequencing coverage.

As expected, the Hi-C data contained almost as much estimated information leakage as the WGS data and more information than the SNP-ChIP array data. The WGS data contained more information leakage than the Hi-C data at low coverage. As we added more reads accumulating between 1.1 and 10 billion base pairs, the information content of the Hi-C data surpassed the WGS data (Figure 4A). We speculate that this is due to a higher quality of genotyping of the genomics regions that are in spatial proximity, as Hi-C has a bias of sequencing more reads from those regions. Furthermore, comparing between WES and different RNA-Seq experiments showed that none of the RNA-Seq experiments contained as much information as the WES data; this is due to the fact that RNA-Seq captures reads only from expressed genes in a given cell type (Figure 4A). An unexpected observation was that more information could be inferred from polyA RNA-Seq data

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

at low coverage compared to WES and total RNA-Seq data. As expected, we could not estimate as much information leakage from the ChIP-Seq reads (Figure 4). Surprisingly, many of the ChIP-Seq assays such as the ones targeting CTCF and RNAPII contained a large amount of information at low coverage (Figure 4). We performed the same calculations for the other individuals and assays as well (Figure 4D–F). For example, we found that ChIA-PET leaked a smaller amount of information compared to ChIP-Seq from CTCF at low coverage (Figure 4B). To make a fair comparison between each of these assays, we calculated the $L(j,k)$ score per base pair depicted in Figure 4A for NA12878. To do so, we normalized the $L(j,k)$ values by the amount of coverage (c). We then averaged each by the number of times (n) c was withdrawn ($\frac{\sum L_c(j,k))/c}{n}$). The Hi-C and ChIP-Seq experiments targeting the TF HDGF provided more leakage per base pair compared to the WGS data. The RNA-Seq experiments provided the least leakage per base pair (Figure 4B).

We next assessed the noise levels in these experiments by calculating the false discovery rate at each coverage level (Supplementary Information). We found that Hi-C data had a higher genotyping accuracy at lower coverage compared to WGS, whereas single-cell RNA-Seq showed the highest noise at higher coverage among all functional genomics assays (Figure S4).

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

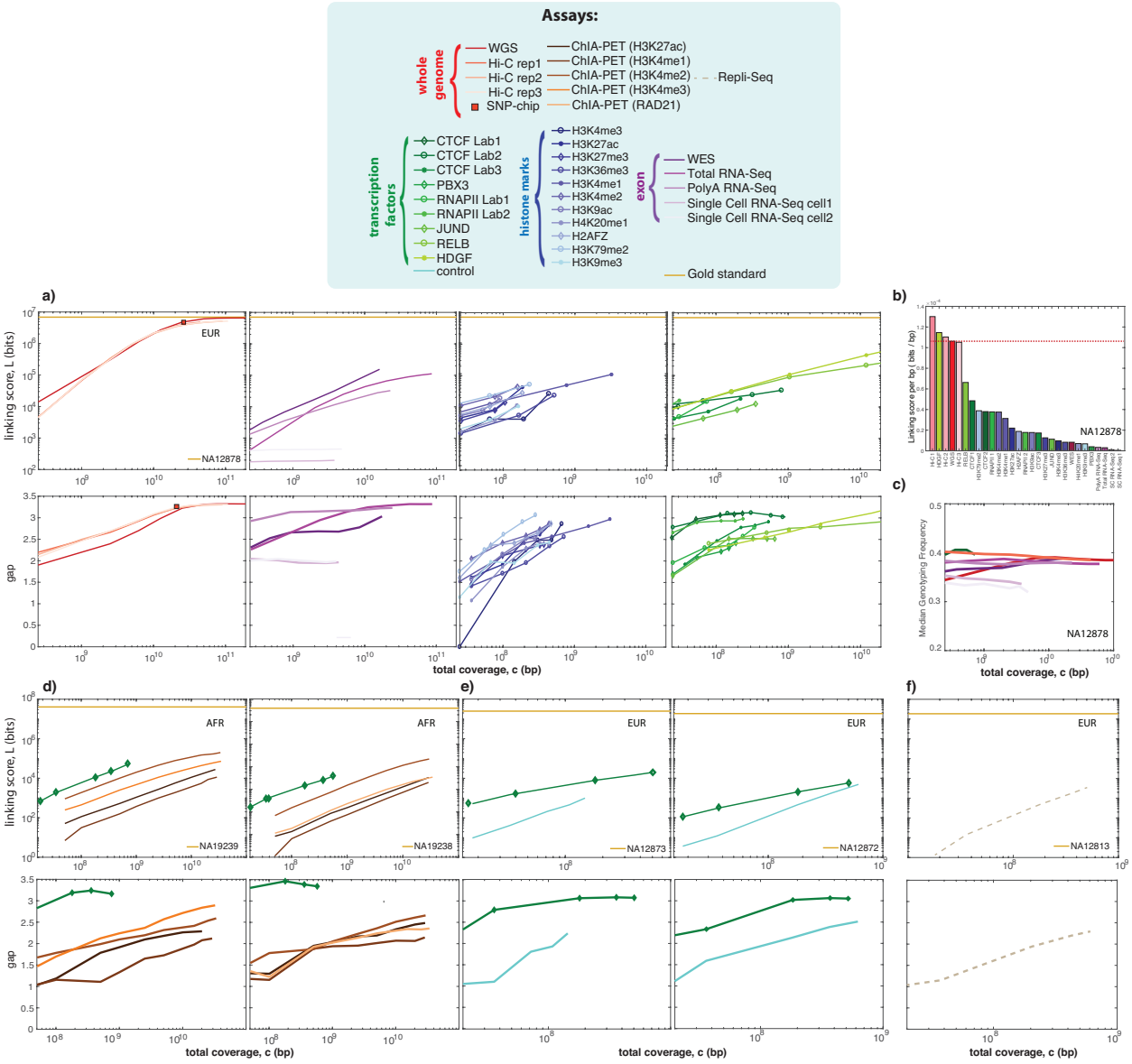

Figure 4: **Linking scores and** *gap* **as a function of coverage.** (**a**) The linking score and *gap* values for different functional genomics assays from individual NA12878 with the 1,000 Genomes genotypes as the gold standard. (**b**) The linking score per base pair for different functional genomics experiments from individual NA12878. (**c**) The median genotyping frequency of observed genotypes for different functional genomics experiments from individual NA12878. (**d**) The linking score and *gap* values values for ChIA-PET and ChIP-Seq experiments from individuals NA19239 and NA19238 with the 1,000 Genomes genotypes as the gold standard. (**e**) The linking score and *gap* values for ChIP-Seq experiments from individuals NA12873 and NA12872 with the 1,000 Genomes genotypes as the gold standard. (**f**) The linking score and *gap* values for repli-Seq experiment from individual NA12813 with the 1,000 Genomes genotypes as the gold standard.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.5 Linking attacks using other functional genomics data as a function of sequencing coverage

Next, we assessed if we can link these individuals by using functional genomics data as the background information and 1,000 Genomes panel as the linked database. We found that we could successfully link individual NA12878 to the database even at very low coverage with Hi-C, total RNA-Seq, polyA RNA-Seq, single-cell RNA-Seq, and ChIP-Seq TF binding data. Overall, ChIP-Seq for histone modification data showed lower gap values at low coverage compared to other assays (Figure 4A). Although the total linking score from single-cell RNA-Seq data was lower compared to other assays, the *gap* values were surprisingly high even at the lowest coverage (Figure 4A). In general, experiments targeting exons (WES, RNA-Seq) demonstrated comparable gap values to whole-genome approaches even though the linking scores were lower. To investigate the reasons behind this, we calculated the median frequency of genotypes called from different functional genomics assays at different coverage (Figure 4C). We found that genotypes from assays targeting exons (especially single-cell RNA-Seq) were slightly more rare in the cohort than genotypes from other assays (Hi-C, WGS, ChIP-Seq TF binding and histone modification), and hence had higher linking quality (Figure 4C, see Supplementary Information for contributions of rare and common genotypes to linking scores).

We then performed linking attacks for other individuals from which we had less functional genomics data. We found that ChIP-Seq experiments targeting the CTCF had high linking quality even at very low coverage. This held true for every individual in our cohort regardless of their ancestry (Figure 4D-4F). We found that non-obvious data types such as ChIP-Seq control experiments and Repli-Seq can be used for linking purposes after coverage of around 10 million bp (if we consider a typical experiment having on average 100 bp read length, then this would correspond to roughly 100,000 reads). We also found that, surprisingly, with some of the ChIP-Seq TF binding experiments it was not possible to link the individuals to the databases despite their relatively high

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

depth (see Figure S5).

We also considered what happens to gap values when we have a trio in the panel. To do so, we added the genotypes of NA12891 and NA12892 (parents of NA12878) to the 1,000 Genomes panel. Although the second best matching individuals were the parents of NA12878, the gap value for linking quality of NA12878 was still statistically significant ($p < 10^{-2}$). We believe this is because we used genotypes as auxiliary information rather than variants. Although individuals share a large amount of variants with their parents, the heterozygosity or homozygosity of the variant naturally varies between a parent and a child (Figure S6).

We then considered a scenario where we had a panel with a completely different genotype frequency distribution than the 1,000 Genomes panel. To create such a panel, we used genotypes from an AFR population (108 individuals) and added two EUR individuals including NA12878. We found that gap values were still statistically significant and the individual was still vulnerable to linking (Figure S7). We then examined how linking would be affected if we removed NA12878 from the panel and left in 108 AFR and 1 EUR. Since the genotype frequency of the AFR population is vastly different than the EUR population (Figures S7 and S8), we misidentified the remaining EUR individual as NA12878 with a statistically significant gap value ($gap \sim 1.6$, Figure S7). Although the existence of such a misbalanced panel is unlikely, this shows that when we have a single individual from the same ancestry as the query individual in the panel, while the rest of the panel is from a different ancestry, mispredictions are possible with noisy genotype-based linking. However, this would be the case only if the ancestry of the query and the ancestry of the rest of the individuals have very different genotype frequency distributions (i.e., AFR vs. the rest of the population).

Here, we tried to make clear that there is a difference between accurate genotyping and sensitive information leakage. Unlike the genotypes we infer from genomic data, the genotypes we obtain

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from functional genomics data are not accurate; hence, they leak enough information to conceptually identify individuals. Moreover, the linking concept from noisy genotypes can also be used for other purposes otherwise relevant to the field such as identifying mislabeled samples during experimental protocols.

2.6 Privacy-preserving file formats for read alignments from functional genomics experiments

Sharing raw read alignments files from functional genomics experiments is extremely important for developing analysis methods and discovering novel mechanisms about the human genome. Ideally, one would share the maximal amount of information with minimal utility loss while largely maintaining an individual's privacy. The aim is to balance the efficiency and effectiveness of the data anonymization process with the usability of the anonymized dataset. Thus, we propose versatile data sanitization such that privacy and usability can be tuned (Figure 5A).

We can think of a raw alignment file (BAM) as a dataset, where information for each read is contained. Let us assume a BAM file is a dataset D, where each entry is a read. The desire is to release dataset D in a form (say D^*) such that it does not leak variants from the reads, but for which any calculation f based on D and D^* retrieves almost the same result. Now let us say the privacy-preserving transformation is done through a function $P_{Q,r}$ such that $P_{Q,r}(D) = D^*$. Q is an operation such as "removal of variants" and r is the parameter, which is the number variants to be removed.

2.6.1 Practical Sanitization Process

We can practically construct a sanitized file pBAM from a BAM/SAM/CRAM file using generalization (a technique commonly used in data sanitization) for the BAM features. Some of the features in BAM files contain large amounts of variant information that can directly tell us the

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

location of a variant (CIGAR and SEQ strings). There are also features that cause subtle leakages such as inferring the presence of a variant in a read (alignment scores, string for mismatching positions, and string for distance to the reference). In general, we replace the BAM tags that leak the presence of a variant by generalizing them. For example, we replace the sequence with corresponding sequence from the reference genome, convert the CIGARs into a format that does not contain variant information, and generalize the other tags. Specifics of pBAM construction can be found in Methods and Supplementary Information.

We store removed information in a structured way (see Supplementary Information) in a compressed file format called ".diff". These .diff files are small files to be kept behind controlled access. With the motivation of keeping the size of private file formats relatively small, we report only differences between BAM and pBAM in the .diff file by avoiding printing any sequence information of the reads that can be found in the reference human genome. This allows us to share the majority of the data with minimal utility loss. If the users find the data useful for their research, they can then apply for controlled access to access the ".diff" files. We provide an easy-to-use software suite that can successfully convert the pBAMs to their original BAM format. This also allows us to avoid keeping the entire BAM dataset behind controlled access and enables sharing more data that is otherwise locked.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

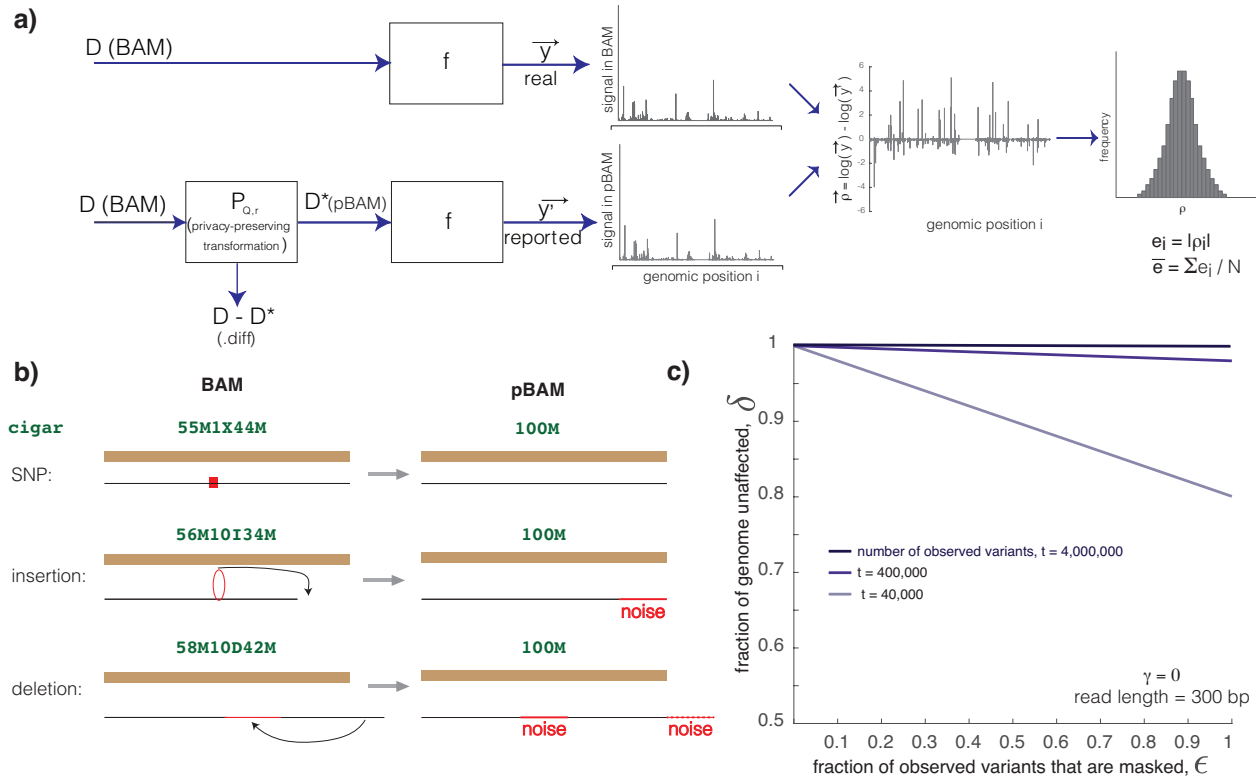

Figure 5: **Privacy-preserving file formats for alignment files from functional genomics experiments.** (**a**) The schematic of the privacy-preserving transformation of an alignment file, the difference between the signal calculated from original BAM and transformed pBAM files, and the concept of ε and δ for the error and privacy. (**b**) A schematic of how different reads and corresponding CIGAR strings are treated in pBAM files. (**c**) The numerical bounds for the privacy-utility relationship. In the very extreme case of obtaining all four million variants from a functional genomics dataset in an assay with 300 bp read length, the maximum utility loss is 20%.

2.6.2 Privacy of pBAM

In practice, the privacy that pBAM provides can be assessed by applying different variant calling software to determine if any of the sanitized variants are in the call set.

In our sanitizer $P_{Q,r}$, r is the number of variants to be sanitized. r includes the variants that are within the linkage disequilibrium (LD) of the variants to be sanitized as well. Note that if the goal is to sanitize all the variants from the BAM file, then the LD variants will be irrelevant because all of the variants will be removed. If one can observe t number of variants from a functional genomics

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BAM file, then the resulting $D^* = P_{Q,r}(\mathrm{D})$ can be viewed as δ-private with respect to operation Q, if $\delta = r/t$, which is the ratio of non-observable variants in pBAM over all observed variants in BAM. We can reach 100% privacy when $r = t$. Here, variants are defined as homozygous and heterozygous alternative alleles. This definition assumes that all r number of variants are guaranteed to be deleted in the resulting pBAM. To investigate if this assumption is theoretically true, we can use the theories behind variant calling algorithms (see Methods).

2.6.3 Utility of pBAM

We defined the the utility of the pBAM as such that any calculation f that is based on the signal from a BAM file should result in similar results when pBAM is used instead. A calculation f can be a signal depth profile calculation, TF binding peak detection, or gene expression quantification (Figure 5A-5B). If the log fold-change between the transformed and original data is $\rho = log(f(D)) - log(f(D^*))$. Then, we can reconstruct an equation for each unit i as

$$e_i = |log(f(D)) - log(f(D^*))| \tag{1}$$

where a unit i can be a single base pair, an exon, or a gene depending on the function f. In turn, e_i can be calculated as the log-fold change between the results derived from two datasets.

Note that e_i is a measure of error of the new dataset D^*. $D^* = P_{Q,r}(\mathrm{D})$ can be viewed as having ε-γ-utility with respect to operation Q if $\varepsilon = (G-m)/G$, where m is the total number of units with $e_i > \gamma$ and G is the total number of the units. We can obtain 100% utility if error is 0 for every unit in the genome. More details on how sanitizing different types of variants affect key utilities such as the coverage can be found in Methods.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.6.4 Privacy-Utility Relationship

The relationship between privacy and utility can be derived through the mathematical relationship between δ and ε. Let us assume our unit for the utility calculation is single bases in the genome, as this will give us the upper bound. Let us also assume that the function for which we want to measure the utility loss f is signal depth calculation (i.e., more utility loss in higher resolution). Sanitization is done over three kinds of variants: SNPs, insertions, and deletions. (1) In the case of SNPs, when we change a letter from the alternative allele to the reference allele, the resulting signal profile at that location does not change (Figure 5B). (2) In the case of an insertion at position x, when we delete the insertion from a read (since an insertion is not represented in the reference), we have to append l_{ins} number of bases to the end of the read (l_{ins} is the length of the insertion). This adds error to all of the bases between position x and $x+L_R$, where L_R is the length of the reads (Figure 5B). That is, for each insertion, the total number of bases with $e_i > \gamma$ will be at most L_R, when γ is equal to 0 for the upper bound (see Methods for details). (3) In the case of a deletion at position x, when we fill the deletion with the reference, we have to delete l_{del} number of bases from the end of the read (l_{del} is the length of the deletion). This adds error to all of the bases between position x and $x+L_R+l_{del}-1$, where L_R is the length of the reads (Figure 5B). Maximum detected indel length varies by the aligner settings. In most extreme cases, l_{del} can be as large as L_R-1. That is, for each deletion, the total number of bases with $e_i > \gamma$ will be at most $2 \cdot L_R - 2$, when γ is equal to 0 and l_{del} is equal to L_R-1 for the upper bound (see Methods for details).

If r is the total number of variants to be sanitized, then $r = r_{snp} + r_{ins} + r_{del}$. The number of bases with m such that $e_i > 0$ are at most

$$m \leq L_R \cdot r_{ins} + (2 \cdot L_R - 2) \cdot r_{del}$$

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Since $\varepsilon = (G-m)/G$, then $m = -\varepsilon \cdot G + G$. We can then say

$$(-\varepsilon \cdot G + G) \leq L_R \cdot r_{ins} + (2 \cdot L_R - 2) \cdot r_{del}$$

If we replace r_{ins} with $r - r_{snp} - r_{del}$ and r with $\delta \cdot t$, then our relationship becomes

$$(-\varepsilon \cdot G + G) \leq L_R \cdot (\delta \cdot t - r_{snp} - r_{del}) + (2 \cdot L_R - 2) \cdot r_{del}$$

See Methods for further details. Figure 5C and Figure S10 show the privacy and utility balance at different δ and ε values, when we use the commonly known values for r, r_{del}, t and L_R. In the most extreme case (for the upper bound calculation), if we assume the read length is 300 bp and we can observe all 4 million SNPs and indels from a functional genomics dataset and 25% of them are deletions, then the utility loss will be at maximum 20% when all the variants are removed. Considering we cannot observe all of the variants of an individual from a functional genomics data, the utility loss will always be lower than 20%, going down to the values of 1% in realistic cases (Figure S10).

2.6.5 Empirical Calculations

We calculated the signal depths of each base pair in the genome with an NA12878 polyA RNA-Seq BAM file using STAR [12]. We then converted the BAM file into pBAMs with different Qs and calculated the signal depth of each base pair. Figure 6A shows the distribution of the log-fold changes in the signal depth with respect to the number of base pairs between BAM and pBAM. We did the same calculation by averaging signal over exons (Figure 6A). Furthermore, we created pBAM files for the BAM files mapped to the reference transcriptome and compared the gene quantification with the gene expression levels calculated from the original BAM files by using RSEM for gene quantification and STAR for transcriptome alignment [12, 13]. We found no difference between the gene expression levels calculated using the original BAM files and the

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pBAM files (see Figure 6A and Supplementary Information for how we treated transcriptome alignments). Overall, when we removed all of the variant leakage from the BAM files, we found a 0.18% difference at the base-pair resolution, 0.27% difference at the exon resolution, and 0% difference at the gene level. When we removed leakage associated with the mismatches, we did not see any difference (Figure 6A). When we removed leakage associated with indels, we found a 0.0016% difference at the base-pair resolution, 0.0011% at the exon resolution, and 0% at the gene level. When we removed leakage associated with split reads, we found a 0.17% difference at the base-pair resolution, 0.26% at the exon resolution, and 0% at the gene level. Figure 6B shows the change in error with respect to an increasing number of modified reads for different operations Q. When the mismatches are manipulated, the resulting signal profiles are not affected. Hence, the manipulated dataset will retrieve the same results as the original dataset regardless of the number of reads. However, manipulating indels and split reads will affect the utility of the new file formats. It is useful to note that the ENCODE consortium, for example, adopts processing pipelines in which split reads are discarded. Therefore, the utility of pBAMs is particularly high when split reads are discarded in the context of gene expression. We also calculated error values for TF binding peak calling using ENCODE processing pipelines (MACS2 [14]) and found negligible error between BAM and pBAM files (Figure S11).

We further calculated the empirical privacy-utility balance in the pBAMs with different error and privacy levels using the total RNA-Seq data of NA12878. We set Q to "removal of SNPs and indels" and systematically ramped up r to a number with an increasing amount of genotyping frequencies and generated pBAMs for each Q. We then calculated the gap_{query} to quantify the risk of re-identification and number of overlapping GWAS SNPs to quantify the risk of characterization. We also calculated the error from the basepair resolution signal profiles for each Q and r as a utility metric. Figure 6C shows how privacy (both risk of re-identification and risk of characterization) decreases while the amount of error introduced to the signal profile increases.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

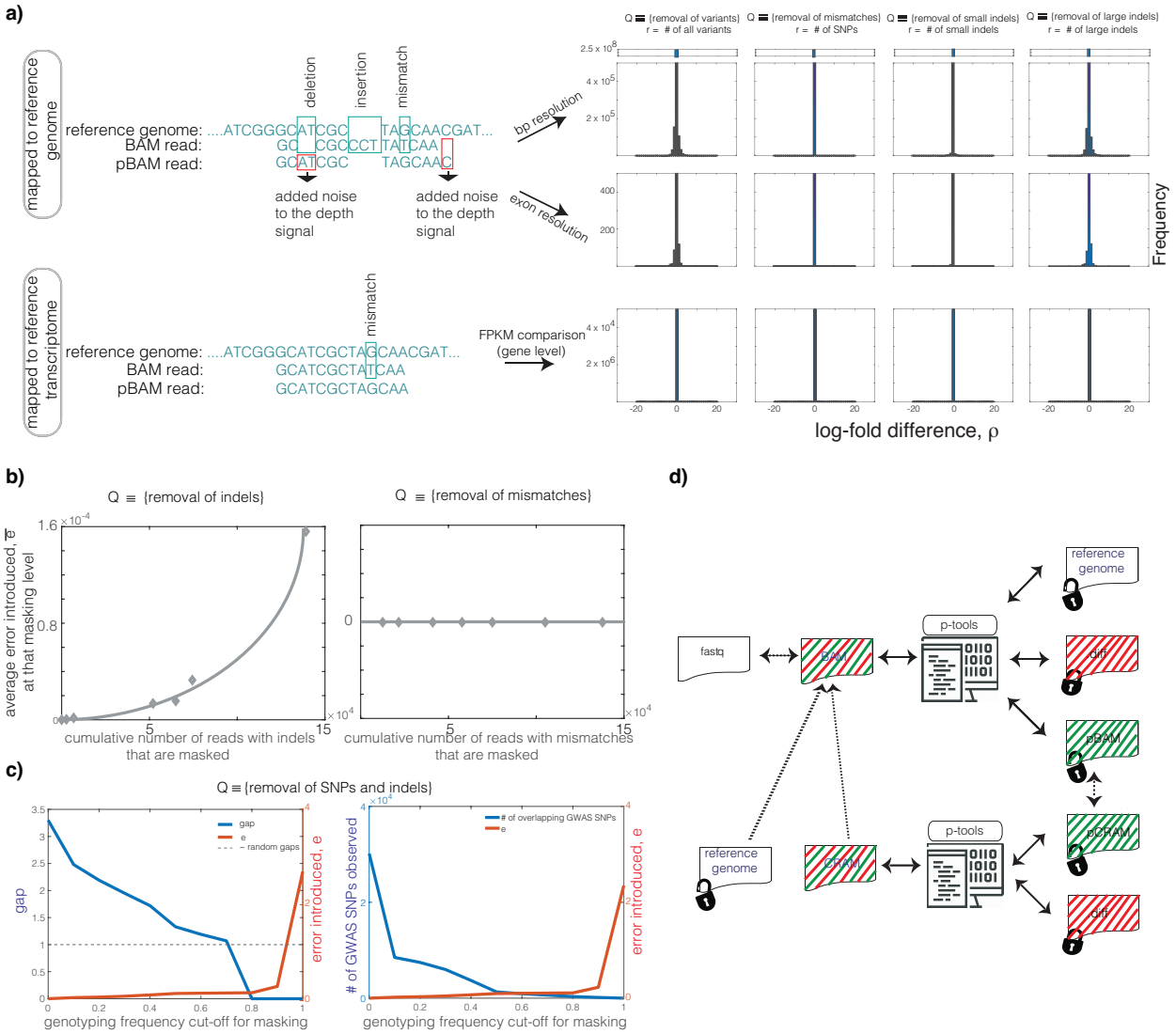

Figure 6: **Privacy-preserving file formats for alignment files from functional genomics experiments.** (**a**) The difference between the BAM and pBAM files is shown at the read level, when the reads were mapped to a reference genome and reference transcriptome. The added noise to the depth signals due to the BAM to pBAM transformation is shown in an example read with a deletion, insertion, and mismatch. The difference of the depth signal when calculated from BAM and pBAM is shown using different operations Q and r values for the polyA RNA-Seq experiment of individual NA12878. The depth signal was compared between different file formats at the single base pair, exon, and gene resolution. (**b**) The change in error with increasing number of manipulated reads for different operations Q. When Q is the removal of mismatches, no noise is added to the depth signal. However, when the Q is the removal of indels, the error increases with an increasing number of manipulated reads. (**c**) The empirical values for privacy-utility balance using NA12878 total RNA-Seq BAM files. We removed the variants with increasing genotyping frequency from the BAM files and calculated the *gap* values and total error in the signal profiles for each pBAM. We also calculated the number of overlapping GWAS variants vs. the total error of the resulting pBAM files. (**d**) The schematic of how p-tools work with different file formats.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.6.6 How to find the privacy-utility balance?

The linkage between a set of variants from known individuals to a database of functional genomics data is highly dependent on the structure of the databases. As we have shown in previous sections, a linkage attack using perfect data with all genotypes from individuals compared to using data with noisy and incomplete data is an easier task. The population composition also affects the linkage. Therefore, if the goal before the dissemination of the data is to protect against present and future linkage attacks while still providing open-access data with high utility, then the proposed solution is to remove all the split reads (to avoid inference of structural variants) and to convert the reads with SNPs and indels to the reference as described above to provide 100% privacy.

However, the consent from the study participants can be more flexible. For example, participants may wish to mask only the certain variants that leak information on susceptibility to stigmatizing phenotypes or diseases that can be used against them by insurers or employers. For that, researchers can convert BAM files to pBAM by only masking the desired variants and the variants within LD of them and theoretically calculate the utility loss by using the above bounds. Moreover, participants may ask for protection against an ideal linkage attack with current reference databases.

We derived a formalism to show the loss of utility under different coverage along the genome (see Methods). For example, masking an indel overlapping with a highly expressed gene will result in greater utility loss than masking an indel overlapping with a non-expressed gene. The genomic coordinates of a highly expressed gene will be represented by more reads, hence the noise added to these reads will be large. The genomic coordinates of a non-expressed gene will not be sequenced in RNA-Seq, so removing any indel overlapping with these locations will not result in any utility loss. Our formalism can help to find an optimal combination of variants to be masked to preserve high utility. For the remaining unmasked variants, our linkage attack software can be used to assess the empirical privacy risks of sharing.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.7 Implementation

We implemented our conversion pipeline of BAM files to pBAM and pBAM+.diff files back to BAM files in bash, awk, and Python. The .diff files are encoded in a compressed format to save disk space. For convenience, pBAM files are saved as BAM files with manipulated content and with a p.bam extension. That is, any pipeline that uses BAM as an input can take p.bam as an input as well. CPU times (calculated using a single 2.3 GHz AMD Opteron processor) and associated file sizes for alignments from RNA-Seq and ChIP-Seq experiments are documented in Table 1. Our file format manipulation has been adopted by the ENCODE Consortium Data Coordination Center and deployed in the ENCODE Uniform Pipeline Framework using workflow description language scripts and docker images, accompanied by appropriate documentation for computational reproducibility on multiple platforms (Google Cloud, Slurm Scheduler, LINUX servers, etc.) under ENCODE Data Processing pipelines. Codes for calculating information leakage, scripts for file manipulations, examples, and file specifications of BAM, pBAM, pCRAM and .diff files can be found at privaseq3.gersteinlab.org and github.com/ENCODE-DCC/ptools.

3 Discussion

Functional genomics experiments using large-scale, high-throughput, sequencing-based assays provide a large amount of biological data. Although these experiments aim to answer questions related to genomic activities such as gene expression, TF binding, or the three-dimensional organization of the genome, public sharing of sequencing data from these experiments can lead to the recovery of genotype information, raising privacy concerns. That is, by mining the "data exhaust [15] resulting from these experiments one can ascertain private facts that were not the intended reasons for the experiments. No study has yet systematically quantified the private information content of functional genomics BAM files or explored approaches to provide open access to such data without compromising an individuals identity.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: p-tools performance and associated file sizes

Experiment	Total RNA-Seq	PolyA RNA-Seq	ChIP-Seq (CTCF)	ChIP-Seq (H3K4me1)
BAM size (bytes)	35,219,346,385	16,301,017,652	954,993,667	2,230,202,265
pBAM size (bytes)	31,986,293,946	14,057,962,755	876,237,603	1,838,044,304
CRAM size (bytes)	21,317,709,703	10,084,587,074	425,193,485	1,120,941,477
pCRAM size (bytes)	20,436,579,905	966,7758,517	378,736,942	883,158,959
.diff size (bytes)	623,943,745	260,978,063	6,991,519	19,438,692
BAM to .diff compression	99.98%	99.98%	99.99%	99.99%
CRAM to .diff compression	96.94%	97.30%	99.98%	99.98%
BAM+hg to pBAM+.diff CPU time	23:21:17	17:05:26	00:37:11	01:08:00
pBAM+.diff+hg to BAM CPU time	12:34:23	04:21:05	00:27:47	00:46:50

We demonstrated how functional genomics data cohorts can be de-anonymized by instantiating linking attacks using publicly available data as well as surreptitiously gathered DNA samples. We showed that, as opposed to databases with full genotypes of individuals, functional genomics cohorts leak phenotypic information that can be sensitive or stigmatizing (i.e., HIV status). Instantiation of linking attacks by genotyping partial or complete functional genomics data showed that even in low coverage experiments, such as ChIP-Seq, an attacker can link individuals to the databases without error. We found that it is easier to link individuals to the databases than genotyping them accurately using functional genomics experiments. The implication is that cryptic quasi-identifiers (i.e., low-quality SNP calling) can be used to link the data to the high-quality genotypes. For example, according to our calculations, reads from single-cell RNA-Seq data carry the largest amount of noise. This is likely due to the bias towards expressed genes in such small amounts of cells, mapping issues of splice sites, false positives from RNA editing sites, and am-

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

plification bias. However, the noisy genotypes called from a small amount of cells, even when the number of reads is only one million, are useful information that result in very high linking accuracy. This is worrisome in terms of biomedical data sharing as the number of individuals in genotype databases is increasing exponentially with the decreasing cost of sequencing.

In this manuscript, we also discussed the concept of principled trade-offs in determining the data production steps, where sensitive information leakage and utility of the data are balanced. Determining a balance is possible by systematically genotyping and quantifying information. Although it is obvious that if given at a substantial amount of depth, as raw sequences contain variants, it is not trivial to determine the amount and the quality of sequencing necessary to perform accurate genotyping. Moreover, we showed that genotyping accuracy of a functional genomics sample and the ability to link individuals to the databases using the same sample are not necessarily correlated. It is easier to link individuals to the databases and infer their complete variant sets than genotyping a sample with accuracy and minimal false discovery. For example, a complete set of variants of a genome may not be obtained by genotyping BAM files from functional genomics experiments. However, using only a small number of reads from the same BAM files, accurate linking attacks are plausible. That is, noisy genotyping from partial sequencing experiments can serve as strong auxiliary information, which is not straightforward to predict at first. Therefore, it is essential to estimate the information in samples accurately. However, functional genomics experiments advance our understanding of health and disease by revealing functions of the genome under different conditions. The quantification, analysis, and interpretation of functional genomics data is still an evolving field; hence, extensive public sharing of functional genomics data will accelerate collaborative research and reproducibility by removing the complexities associated with data accession procedures.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this study, we showed that when it comes to sharing raw sequences from functional genomics experiments, there is no "one rule fits for all approach for determining the appropriate balance by only looking at the amount of sequencing. Therefore, we suggest a framework through which a data producer can estimate the amount of leakage and the ability of re-identification of functional genomics data before their release.

The increasing incentive to share data to advance biomedical research and the corresponding increasing privacy concerns have led researchers to look for more complex solutions to overcome the bottleneck between data sharing and privacy preserving means. Researchers have proposed solutions such as differential privacy [16, 17, 18]. Studies have shown that retrieving summary information from private statistical databases without revealing some amount of an individual's information is impossible [19]. Furthermore, an entire database can be inferred by using a small number of queries. Differential privacy ensures a high level of privacy such that an adversary retrieves a similar result with or without the addition of the individual's data to the database by adding perturbations or noise to the queries [19]. We further studied if we could utilize the concept of differential privacy to create leakage-free raw functional genomics data (see Supplementary Information). Although such a concept is useful for sharing summary statistics of functional genomics data from multiple individuals, it is conceptually hard to apply to raw mapped read sharing from functional genomics experiments taken from a single individual. Although further research will be fruitful on how to extract useful information from genomics data that are noisy and perturbed, we envision privacy concepts like differential privacy will be useful in genomics data sharing applications such as releasing population-based genotype-phenotype data.

To enable public sharing of raw alignments from functional genomics experiments, we designed a privacy-preserving transformation and created privacy-preserving binary alignment files (pBAM). We developed a framework with which researchers can tune the level of privacy and util-

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ity balance they want to achieve based on the policies and consents of the donors. pBAMs enable researchers to share the mapped reads, which are the largest data product of functional genomics experiments. To ease the challenges associated with moving and storing large special-access files, we created a lightweight .diff file format that consists of the differences between pBAM and BAM files in a compact format. This allows us not to repeat the sequence information in the human reference genome files in .diff files and reduces the size of the private files significantly. The presented framework can also be combined with other privacy-preserving solutions such as k-anonymity before releasing functional genomics data from a cohort of individuals. For example, a user can create a data frame with the individual identifying features (SNPs, indels, SVs, etc.) that are in the functional genomics data. Based on the desired k-anonymity, the user then can decide to suppress or generalize any combination of these features. This will then be converted into a specific Q and r, which in turn instantiate a pBAM with the desired coverage of leakage that will satisfy the k-anonymity property of the table that can be released. We developed a mathematical formalism showing the relationship between utility and privacy in key genomics applications. There are also other types of utility of functional genomics data, which are closely tied to the genetic information in the raw files. For example, allelic imbalance in gene expression can only be detected if SNPs are present in the data. Hence, utility and privacy in that sense have a different mathematical relationship compared to the utility of gene expression quantification and privacy, as utility in the sense of gene expression is not directly tied to the SNPs in the data.

We addressed the most obvious leakage and provide solutions for quick quantification and safe data sharing. However, it will be useful to review all the sources of information leakage from functional genomics experiments. For example, the next source of leakage is from the signal profiles in RNA-Seq, which researchers have addressed [21]. Leakage can also occur from gene expression quantification, which studies have shown is connected with variants through eQTLs [20]. We also anticipate more leakages to be discovered as new functional genomics experiments are developed.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Combined with the increasing attention on genomic privacy, we expect future studies will lead to novel privacy-preserving solutions in an open data-sharing mode.

4 Methods

4.1 Method Details

4.1.1 Sample Selection

We present short variant calls on 478 samples. Of these, 16 are newly sequenced for this study (2 DNA samples from 2 individuals, 2 RNA samples from the same individuals, 6 coffee cup DNA samples from two individuals), and the remaining 462 RNA samples were obtained from the gEUVADIS study [7]. Genomic materials for newly sequenced samples were obtained through collection of blood and used coffee cups. DNA samples from the blood were sequenced using high-coverage Illumina sequencing (30x) and used as the gold standard. RNA samples from the blood were sequenced using the Illumina total RNA-Seq protocol. Extracted DNA from coffee cups were sequenced using low-coverage Illumina sequencing (10x), Oxford Nanopore Technologies (ONT), and were genotyped with the Illumina Infinium OmniExpressExome-8 v1.6.

4.1.2 Genotyping blood DNA and coffee-cup samples

Illumina low-coverage sequencing: Raw fastq files were processed by mapping them to the hg19 reference genome (b37 assembly) using bwa [25]. The resulting BAM files were processed using Picard tools to remove PCR duplicates. Deduplicated files were then genotyped using GATK best practices [10, 11].

Genotyping Arrays: Twelve coffee cup samples were genotyped using Illumnina Infinium OmniExpressExome 8 v1.6 and UV coated chips were scanned using IScan. Scanned output files were analyzed and

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

call rates were calculated using Illumina BeadStudio.

ONT: Twelve coffee-cup samples were prepared and sequenced using the rapid barcoding kit and minION following the manufacturer suggestions. After converting fast5 files into the fastq format using Guppy software for base-calling and de-multiplexing, each sample was aligned to the hg19 reference genome (b37 assembly) using bwa "mem -x ont2d options [25]. Aligned reads were used for variant calling using nanopolish software.

4.1.3 Genotyping functional genomics data

Raw RNA-Seq fastq files (from this study, gEUVADIS and ENCODE) were processed by mapping them to the hg19 reference genome (b37 assembly) and gencode v19 transcriptome assembly using STAR [12]. Other raw functional genomics (e.g., Hi-C, ChIP-Seq) fastq files were mapped to the hg19 reference genome (b37 assembly) using bwa [25]. The resulting BAM files were processed using Picard tools to remove PCR duplicates. Deduplicated files were then genotyped using GATK best practices [10, 11] for RNA and DNA for RNA-Seq and other functional genomics data, respectively.

4.1.4 Empirical p-values for gap

We estimated the empirical p-values for a particular gap value by linking a set of random genotypes that do not belong to a particular individual to the database. To do so, we calculated empirical p-values as follows.

a. We selected N random genotypes from a database of nearly 50 million genotypes. N is the total number of genotypes that was linked to the database for gap calculation.

b. We calculated the $L(j,k)$ between these random N genotypes and every individual in the database.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

c. We calculated the gap as the ratio between the first ranked and the second ranked $L(j,k)$.

d. We repeated the above steps 1,000 times and obtained a distribution of random gap values.

e. The total number of random gap values that are equal or greater than the real gap divided by 1,000 is the probability of observing the real gap value by chance.

4.1.5 pBAM details

Below is a practical guideline on how to convert BAM files into pBAM files.

1. Let us assume the variants that need to be sanitized from the BAM file are $V_s = \{s_1, .., s_i, ..s_n\}$ and the variants that are in LD with the variants in V_s are $V_s^{LD} = \{s_1^{LD}, .., s_i^{LD}, .., s_k^{LD}\}$. The total number of variants that need to be sanitized are then $r = n + k$.

2. We first find all the reads that contain the variants in the $V_s \cup V_s^{LD}$ such that $R = \{R_1, .., R_T\}$.

3. 3. We apply one of the sanitization techniques (i.e., generalization) to the BAM fields so that an adversary can infer the existence of these variants:

   (a) CIGAR: Convert the cigar of each R_i to a perfectly mapped read cigar (e.g. 100M, where 100 is the read length and M denotes that each base on the read perfectly mapped to reference genome).

   (b) SEQ: Replace the sequence of each R_i with the sequence in the reference genome.

   (c) QUAL: Convert all the base qualities of R_i to perfectly called base phred scores.

   (d) There are also optional tags in the BAM files such as AS (alignment scores), MD (string for mismatching positions) and NM (edit distance to reference) that should be sanitized. They can be generalized to the values for perfectly mapped reads if they are present in the BAM files.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(e) We found that there might be extremely subtle leakages through MAPQ scores (mapping quality, see Figure S12). In particular, if the goal is to prevent the large leakages such as structural variants, then a data sanitization procedure such as suppression or generalization might be suitable for this field as well.

In addition, we treat intronic reads differently to be able to capture the splicing accurately. Details can be found in Supplementary Information.

4.1.6 Privacy bounds

Due to the errors and biases in sequencing and alignment processes, variant calling algorithms often take a probabilistic approach based on Bayess theorem. Let us assume a position on the genome is represented by a pile of reads R. We can observe the probability of reads having a letter given a genotype from the data ($P(R|G)$). This then help us to compute the genotype likelihood given the reads R as

$$P(G|R) = \frac{P(R|G)P(G)}{P(R)} = \frac{P(R|G)P(G)}{\sum_i^n P(R|G_i)P(G_i)}$$

Here R is the observed data, i.e the aligned reads at that particular position; G is the genotype whose probability is being calculated; G_i is the ith possible genotype out of n possibilities; and $P(G)$ is the prior probability for the genotype whose probability is being calculated.

Let us assume that we have three possible genotypes for a position on the genome: **AA**, **BB** and **AB**. In a error-free sequencing and alignment with the unsanitized BAM file, the observed distribution of the letters in a location of a genome would follow a binomial distribution with $P(R = A|G = AA)$=1 for **AA** genotype; $P(R = B|G = BB)$=1 for **BB** genotype and $P(R = A|G = AB)$=$P(R = B|G = AB)$=0.5 for **AB** genotype. Since the sequencing and alignment are not precise, these probabilities can server as upper bounds in our calculation. Now we suppose this position in the genome is in our variants to sanitize list. During the sanitization process, we then convert letter

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

B in any read R_i of R to a letter **A**, which means we never observe any read $R_i = B$. This results in the observation of $P(R = A|G) = 1$ for any genotype G and $P(R = B|G) = 0$ for any genotype G. Then the genotype likelihoods will be calculated as

$$P(G=AA|R) = \frac{P(R|G=AA)P(G=AA)}{P(R|G=AA)P(G=AA)+P(R|G=AB)P(G=AB)+P(R|G=BB)P(G=BB)} = 1$$

$$P(G=AB|R) = \frac{P(R|G=AB)P(G=AB)}{P(R|G=AA)P(G=AA)+P(R|G=AB)P(G=AB)+P(R|G=BB)P(G=BB)} = 0$$

$$P(G=BB|R) = \frac{P(R|G=BB)P(G=BB)}{P(R|G=AA)P(G=AA)+P(R|G=AB)P(G=AB)+P(R|G=BB)P(G=BB)} = 0$$

According to above formalism, no matter what genotype and variant we have in the BAM file, it will always be observed as the homozygous reference allele in the resulting pBAM file.

4.1.7 Utility bounds

The goal is to consider the impact of sanitizing BAM files through modifications applied to identified variants. The sanitization procedure utilized here is inherently asymmetric, as bases are added or removed at only one end of the read. We make a distinction between a personal genome and the reference genome: the personal genome includes all SNPs and indels; the reference genome is the standard external metric. While the personal genome is not necessarily actually constructed, it serves as a useful conceptual tool to understand the impact of the transformation involved in the sanitization procedure. We discuss three types of variants and the chosen method of sanitization applied in the pBAM format:

SNP: A single nucleotide variant/polymorphism is changed by mutating the variant to the reference allele in every read in which the mutation is observed.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Insertion: An insertion is sanitized by removing any fraction of the new inserted segment and adding the equivalent number of reference nucleotides to one end of the corresponding read.

Deletion: A deletion is sanitized by filling in the reference nucleotides into the part of the deleted segment occurring on any read, and then removing the equivalent number of nucleotides from that read.

Definitions

The genome is indexed by discrete positions i. The coverage prior to and after sanitization are functions of the genomic position, and are labeled as $c^{pre}(i)$ and $c^{post}(i)$, respectively. The read length is fixed at L_R. The size of the insertion is labeled l_i^{ins}, while the size of the deletion is labeled as l_i^{del}, where in both cases the position i marks the start position of the insertion or deletion. In addition, define $N(i) =$ Number of reads that start at position i in the mapping to the personal genome.

SNPs: Every read containing a SNP will be modified, with the alternate allele replaced by the corresponding reference allele. Under the assumption that the presence of this SNP does not alter the mapping efficiency (say, if the other variants within a particular read sum to m^{mis-1}, then this SNP will lead to that read being dropped) and thus read dropout, we see that $c^{pre}(i) = c^{post}(i)$. So no impact will be observed, unless one looks through the mapping QC and finds all the reads overlapping a given locus have slightly lower quality. This might be possible unless the QC metadata is being modified explicitly in the sanitization procedure (see Supplementary Information)..

Short Indels: For indels, we consider the mapping changes due to the sanitization procedure in the following. For convenience sake, we examine the mapping changes with respect to a single strand. The impact will proceed in an equivalent manner for the reverse strand as well. It is

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

important to keep in mind that the total "genomic footprint" of the sanitization procedure (i.e., the region of the genome that is impacted by the procedure) is therefore double the number quoted below for one strand.

(1) Insertions: The variant is indexed by position i, where the insertion occupies the base pairs from $i+1$ to $i+l_i^{ins}$. We consider the following cases:

- $l_i^{ins} < L$: No individual read will dropout in this case due to the presence of the insertion. Consider the case that the added nucleotides are on the end of higher genomic position for the sake of clarity. The process of sanitization leads to an additional build-up of reads downstream of the insertion point. This happens due to the replacement process discussed above. Certain reads that would have been mapped to the insertion in the personal genome of the individual are now, instead, added downstream of the insertion in the mapping to the reference genome. This allows us to quantify the read build-up in terms of the start positions of the reads. Thus for all reads that overlap with the insertion the following transformation occurs:

  - $\forall x : i - L_R + 2 \leq x \leq i - L_R + l_i^{ins}$, all reads starting at position x in the personal genome are newly mapped to the reference in the interval $[i+1, x+L_R+1]$.
  - $\forall x : i - L_R + l_i^{ins} + 2 \leq x \leq i$, all reads starting at position x in the personal genome are newly mapped to the reference in the interval $[x+L_R-l_i^{ins}, x+L_R-1]$.
  - $\forall x : i+1 \leq x \leq i+l_i^{ins}$, all reads starting at position x in the personal genome are newly mapped to the reference in the interval $[x+L_R-l_i^{ins}+1, i+L_R]$.

  The genomic footprint of the sanitization procedure is the interval $[i+1, i+L_R]$. The resultant read build-up of the sanitized BAM relative to the original BAM is thus given by the integral/discrete sum over all the accumulated contributions described above (again, x is the position along the personal genome, with the insertion in place; α in the equation below is the position along the reference genome in the downstream interval $[i+1, i+L_R-1]$ that is

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

impacted by read build-up due to post-sanitization remapping):

$$\text{Change in the raw read count} = \Delta c(i+1 \leq \alpha \leq i+L_R) = \sum_{x=\alpha-L_R+1}^{\alpha-L_R+l_i^{ins}} N(x)$$

For example, imagine an ideal case where all the reads are uniformly distributed in the given region. This means that $c(x) = \text{constant} = c$ and $N(x) = \text{constant} = \frac{c}{L_R}$ across the original mapping. Note that in the ideal case of uniformly distributed reads, the number of reads that begin at a locus is the total number of reads that overlap with a locus, divided by the length of each read. Thus we get

$$\text{Change in the raw read count} = \Delta c(i+1 \leq \alpha \leq i+L_R) = \sum_{x=\alpha-L_R+1}^{\alpha-L_R+l_i^{ins}} N(x) = c \cdot \frac{l_i^{ins}}{L_R}$$

Thus, the fold-change in the coverage is given by $FC = \frac{c+\frac{c \cdot l_i^{ins}}{L_R}}{c} = 1 + \frac{l_i^{ins}}{L_R}$ in the interval $i+1 \leq \alpha \leq i+L$ If we define, $FC(\alpha) = \exp(e_\alpha)$, we end up with $e_\alpha = log(1+\frac{l_i^{ins}}{L_R})$ in the interval $i+1 \leq \alpha \leq i+L$. The general formula is

$$e_\alpha = log(\frac{c(\alpha) + \sum_{x=\alpha-L_R+1}^{\alpha-L_R+l_i^{ins}} N(x)}{c(\alpha)})$$

- $l_i^{ins} \geq L_R$: This case will be slightly different from the above case, as some of the reads will completely overlap with the insertion region and never contribute to the remapping build-up downstream. The calculation would then ignore the reads that overlap significantly with the insertion. We do not discuss this situation here, as these reads will have soft or hard clipping in their cigars (split reads) and will be treated differently as discussed above.

(2) Deletions: The variant is indexed by position i, where the deletion removes the reference base pairs from $i+1$ to $i+l_i^{del}$. We exclusively consider the case of $l_i^{del} < L_R$ in this case, with the un-

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

derstanding that longer deletions would require slight modifications of the following calculations. The mapping to the reference genome after sanitization results in the loss of coverage from regions downstream of the deletion, and an equal gain in regions of the deletion. The mapping changes are as follows:

- $\forall x : i - L_R + 2 \leq x \leq i - L_R + 1 + l_i^{del}$, all reads starting at position x in the personal genome are removed from $[i + l_i^{del} + 1, x + l_i^{del} + L_R - 1]$ and newly mapped to the reference in the interval $[i+1, x+L_R-1]$.

- $\forall x : i - L_R + 2 + l_i^{del} \leq x \leq i$, all reads starting at position x in the personal genome are removed from $[x + L_R, x + l_i^{del} + L_R - 1]$ and newly mapped to the reference in the interval $[i+1, i+l_i^{del}]$.

The genomic footprint of the sanitization procedure is the interval $[i+1, i+l_i^{del}+L_R-1]$. The change in coverage can be calculated in a manner similar to the case of insertions, with the additional notion that reads that are remapped to the deleted segment are drawn from downstream portions of the genome:

If the gain in the raw read count in the deleted segment of the reference genome is *Gain*, then

$$Gain = \Delta c(i+1 \leq \alpha \leq i + l_i^{del}) = \sum_{x=\alpha-L_R+1}^{i} N(x)$$

If the loss in the raw read count downstream of the deleted segment is *Loss*, then

$$Loss = \Delta c(i + l_i^{del} + 1 \leq \alpha \leq i + l_i^{del} + L_R - 1) = -\sum_{x=\alpha-l_i^{del}-L_R+1}^{min(\alpha-L_R,i)} N(x)$$

It is not possible to compute the fold-change in the coverage in the deleted segment, as the coverage is 0 pre-sanitization. However, it can be calculated by adding a pseudo count to the coverage pre-

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sanitization. In the downstream segment, the fold-change in the coverage is given by

$$\mathrm{FC} = \frac{\mathrm{c}(\alpha) - \sum_{\mathrm{x}=\alpha-\mathrm{l}_{\mathrm{i}}^{\mathrm{del}}-\mathrm{L}_{\mathrm{R}}+1}^{\min(\alpha-\mathrm{L}_{\mathrm{R}},\mathrm{i}} \mathrm{N}(\mathrm{x})}{\mathrm{c}(\alpha)}$$

for $i + l_i^{del} + 1 \leq \alpha \leq i + l_i^{del} + L_R - 1$, and with $FC(\alpha) = e^{e_\alpha}$, we have

$$e_\alpha = log(\frac{c(\alpha) - \sum_{x=\alpha-l_i^{del}-L_R+1}^{min(\alpha-L_R,i)} N(x)}{c(\alpha)}) = log(1 - \frac{\sum_{x=\alpha-l_i^{del}-L_R+1}^{min(\alpha-L_R,i)} N(x)}{c(\alpha)})$$

Again, considering the example of constant coverage discussed for the insertions, we have

$$\mathrm{FC} = 1 - \frac{\sum_{\mathrm{x}=\alpha-\mathrm{l}_{\mathrm{i}}^{\mathrm{del}}-\mathrm{L}_{\mathrm{R}}+1}^{\min(\alpha-\mathrm{L}_{\mathrm{R}},\mathrm{i})} \mathrm{N}(\mathrm{x})}{\mathrm{c}(\alpha)} = 1 - \frac{\sum_{\mathrm{x}=\alpha-\mathrm{l}_{\mathrm{i}}^{\mathrm{del}}-\mathrm{L}_{\mathrm{R}}+1}^{\min(\alpha-\mathrm{L}_{\mathrm{R}},\mathrm{i})} \frac{\mathrm{c}}{\mathrm{L}_{\mathrm{R}}}}{\mathrm{c}} = 1 - \frac{\min(\mathrm{l}_{\mathrm{i}}^{\mathrm{del}}, \mathrm{i} + \mathrm{l}_{\mathrm{i}}^{\mathrm{del}} + \mathrm{L}_{\mathrm{R}} - \alpha)}{\mathrm{L}_{\mathrm{R}}}$$

The upper bound on FC is given by $\mathrm{FC} \leq 1 - \frac{\mathrm{l}_{\mathrm{i}}^{\mathrm{del}}}{\mathrm{L}_{\mathrm{R}}}$. Note that the formulae and the genomic footprint are different from those in the insertion case.

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biology*, 2011;12(8):125.

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

[3] $https://www.washingtonpost.com/news/morning-mix/wp/2018/10/17/the-culprits-name-remains-unknown-but-he-licked-a-stamp-and-now-his-dna-stands-indicted/?utm_term=.25eba675732b$

[4] $https://www.economist.com/christmas-specials/2017/12/19/do-it-yourself-science-is-taking-off$

[5] Sweeney L. Simple demographics often identify people uniquely. *Carnegie Mellon University, unpublished*, 2000.

[6] Narayanan A. and Shmatikov V. Robust de-anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy*, 2008;111-115.

[7] Lappalainen T et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* , 2013;501:506-511

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

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

[10] DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K,

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Gabriel S, Altshuler D, Daly M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 2011;43(5):491-498.

[11] Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 2013;43:11.10.1-33.

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

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

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

[15] Kosciejew, M. The individual and big data. *Feliciter*, 2013;59(6):47.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/345074; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

[21] Harmanci A, Gerstein M. Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions. *Nature Communications*, 2018

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

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

[24] Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, Rowe LD, Dreszer TR, Roe G, Podduturi NR, Tanaka F, Hong EL, Cherry JM. ENCODE data at the ENCODE portal. *Nucleic Acids Res.*, 2016;44(D1):D726–D732.

[25] Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 2009;25:1754–1760.