1 Supplementary Information

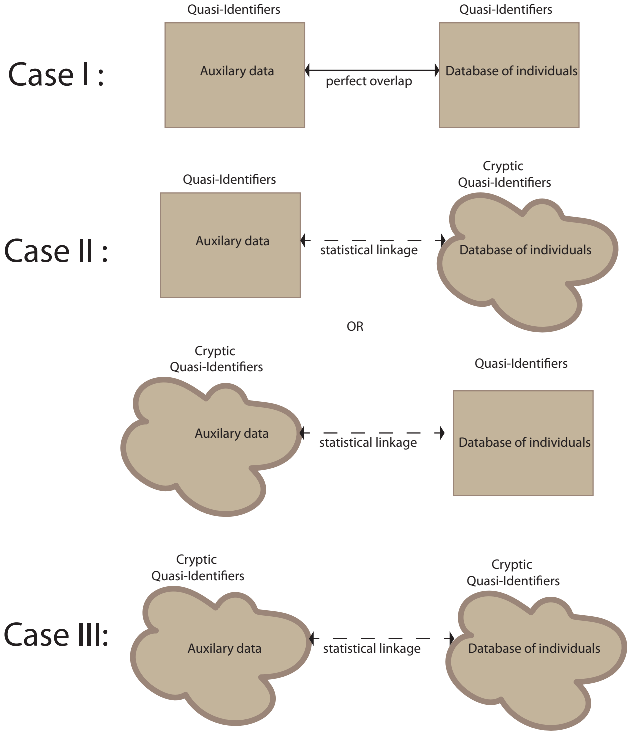
Contents

List of Figures

List of Tables

1.1 Linkage attacks

Linkage attacks can be categorized in three cases. Case I: A perfect auxiliary data is linked to a perfect database to reveal the identity or the preferences of the owner of the auxiliary data. Case II: A perfect auxiliary data is linked to a noisy database to reveal the identity or the preferences of the owner of the auxiliary data. Case III: A noisy auxiliary data is linked to a noisy database to reveal the identity or the preferences of the owner of the auxiliary data. See SI Figure 1.



Supplementary Figure 1: **Different cases of linkage attacks.**

1.2 Similarity and Sparsity

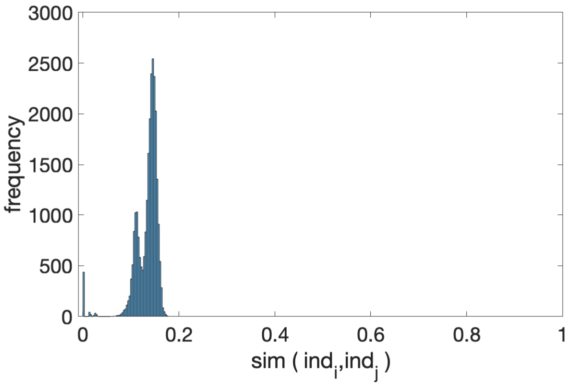
Following Narayanan and Shmatikov [1], one can define similarity between two individuals (ind_i and ind_j) in a phenotype dataset as the total number of genotypes they share normalized by the total number of genotypes inferred for both from the functional genomics data. This could be formalized as:

$$\mathrm{Sim}(\mathrm{ind_i}, \mathrm{ind_j}) = \frac{\sum_{\mathrm{k}} \mathrm{Sim}(\mathrm{ind_i^k}, \mathrm{ind_j^k})}{|\mathrm{supp}(\mathrm{ind_i}) \cup \mathrm{supp}(\mathrm{ind_j})|}$$

$\mathrm{Sim}(\mathrm{ind_i^k}, \mathrm{ind_j^k})$ is 1 if k^{th} genotype on the functional genomics data is the same for both individuals and 0 otherwise. Then a database D is $(\varepsilon, \delta)-$sparse with respect to the similarity measure Sim if

$$Pr[\mathrm{Sim}(\mathrm{ind_i}, \mathrm{ind_j}) > \varepsilon \, \forall \, \mathrm{ind_i} \neq \mathrm{ind_j}] \leq \delta$$

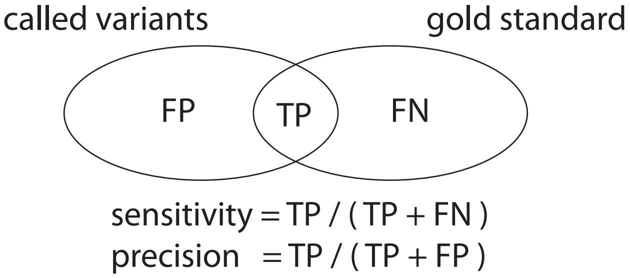
As in the case for Netflix preference dataset, we found that any two individual does not share similarity more than 20% (SI Figure 2).



Supplementary Figure 2: **Sparsity of the phenotype dataset with respect to similarity.**

1.3 Sensitivity and Precision

We defined the sensitivity and the precision of a set of genotypes called from a functional genomics experiments or from the DNA extracted from a coffee cup as the measured of how much of the correct genotypes captured given the individual's full genotype profile and how much of incorrect genotypes captured (noise), respectively (SI Figure 2).



Supplementary Figure 3: **Sensitivity and precision.**

1.4 Experimental Protocols

1.4.1 DNA extraction protocol from coffee-cup lids

We used the QIAamp DNA Investigator Kit from QIAGEN. This kit is design to purify DNA from forensic and human identity samples. We first swabbed the surface of the coffee-cups using a cotton swab dipped into 1 μliter purified water. We followed the QIAamp DNA Investigator Kit protocol suggested for isolation of DNA from surface swab samples without modification. The final amount of DNA isolated from coffee-cups were around 0.9 to 1 ng.

1.4.2 Whole genome amplification

Due to the very low starting amount of purified DNA, we used a single-cell whole genome amplification kit (REPLI-g Single Cell Kit), which allows uniform PCR amplification from single cells or limited sample materials to use in next-generation sequencing applications. We then used Monarch PCR and DNA Cleanup Kit to purify the DNA from PCR reactions.

1.4.3 Illumina sequencing

Amplified DNA samples from coffee-cups as well as the purified PCR-free DNA from blood (as gold standard) were sent to Yale Center for Genome Analysis for Illumina Whole Genome Sequencing. Coffee cup samples were sequenced in 10x coverage and blood samples were sequenced in 30x coverage.

1.4.4 Illumina genotyping arrays

We used Infinium OmniExpressExome-8 BeadChip for the amplified DNA samples from coffee cups. Infinium OmniExpressExome-8 arrays surveys tag SNPs located on exons from all three HapMap phases, which includes 273,000 exonic markers. Each SNP is represented on these chips by on average 30 beads. Yale Center for Genome Analysis performed the BeadChip protocol and calculated the call rates using Illumina BeadStudio.

1.4.5 Nanopore Sequencing

Due to the low quality DNA obtained from coffee-cups, we did not perform size selection of fragments from the PCR-based libraries we obtained using Oxford Nanopore (ONT) rapid sequencing kit. Total of 12 libraries from 6 coffee cups per individual were barcoded using the ONT rapid barcoding kit. Libraries were sequenced across an individual R9.4 flow cell on a single MinION instrument. A total of 844,599 reads were successfully base-called and demultiplexed using Guppy. The recommended MinION run-time was 48 h, therefore run was terminated after 48 h. SNP calling is performed using Nanopolish software.

1.4.6 RNA extraction protocol and RNA-Seq

Blood samples from individuals were sent to Yale Center for Genome Analysis for RNA purification and Illumina high coverage total RNA-Seq sequencing following the suggested protocols by Illumina. Total RNA-Seq data yielded more genotypes than the gEUVADIS data. To do a fair comparison in linkage attacks, we downsampled the total number of captured variants to the average number of variants observed in gEUVADIS dataset.

1.5 Mapping and Genotyping Statistics for Coffee-cup Samples

Supplementary Table 1: Mapping Statistics for WGS data.

Sample Name	# of Total Reads	Percentage of Mapped Reads
Ind1-Cup1	127,850,309	82.36
Ind1-Cup2	137,723,180	82.14
Ind1-Cup3	160,138,265	61.50
Ind1-Cup4	215,379,662	75.98
Ind1-Cup5	157,769,539	83.56
Ind1-Cup6	121,329,654	83.54
Ind2-Cup1	207,414,742	83.58
Ind2-Cup2	151,139,298	62.65
Ind2-Cup3	122,164,133	83.76
Ind2-Cup4	220,846,681	77.93
Ind2-Cup5	140,335,759	85.26
Ind2-Cup6	172,391,527	88.97

1.6 Data

The functional genomics data except Hi-C were downloaded from the ENCODE Data Portal. Hi-C data was taken from ref. [2]. WGS and WES data were taken from the 1000 Genomes Data Portal. The accession codes and references of the data are summarized in SI Table. 3.

1.7 Genotyping accuracy

We calculated the rate of false information for all the functional genomics assays for the individual NA12878. Rate of false information (RFI) is calculated as the ratio between the linking score of the false positive SNPs and all SNPs inferred from functional genomics data at a given coverage. SI Figure 4a shows that the noise for Hi-C data was lower compared to WGS data at lower coverage. We attribute this finding to the deeper sequencing of the genomics regions in close spatial proximity. Hence, sampling more reads from regions at low coverage is more likely compared

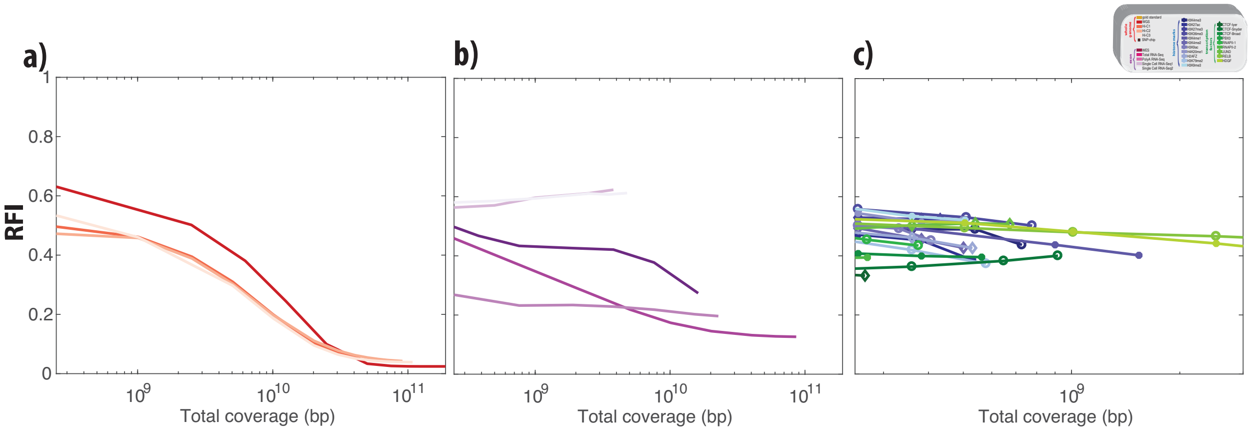
Supplementary Table 2: Genotyping statistics for different technologies.

	WGS			Genotyping Array				ONT		
samples	Called variants	Correct variants	Overlap with RNA-Seq	Called variants	Correct variants	Overlap with RNA-Seq	Call rate	Called variants	Correct variants	Overlap with RNA-Seq
ind1-cup1	219,887	127,389	2,843	781,807	544,932	4	84%	14	0	0
ind1-cup2	526,798	366,539	5,764	805,034	572,318	11	86%	287	9	0
ind1-cup3	252,452	104,710	2,314	761,566	532,966	1	81%	79	14	0
ind1-cup4	217,840	84,418	2,323	672,857	452,859	2	72%	12	4	4
ind1-cup5	25,822	6,984	272	808,505	600,175	5	86%	0	0	0
ind1-cup6	56,577	30,580	601	743,598	574,474	3	79%	0	0	0
ind2-cup1	267,844	127,606	9,872	751,349	534,890	12	80%	122	15	3
ind2-cup2	243,783	136,480	9,831	788,263	562,419	15	84%	0	0	0
ind2-cup3	35,133	6,848	949	823,790	608,231	5	88%	0	0	0
ind2-cup4	85,613	56,716	5,014	821,752	500,897	11	88%	0	0	0
ind2-cup5	228,913	93,103	7,450	696,447	515,111	8	74%	67	12	2
ind2-cup6	229,167	102,538	7,734	702,263	474,838	6	75%	83	6	2

to uniform sampling of reads from WGS. ChIP-Seq data had a comparable noise levels to WGS and Hi-C data given the shallow sequencing depth. ChIP-Seq targeting CTCF had the lowest noise (SI Figure 4b). We further found that the polyA RNA-Seq experiment had the lowest noise compared to WES and total RNA-Seq. This could be attributed to the deeper sequencing of regions containing highly expressed genes and deeper sampling from these regions. In general, assays targeting the transcriptome such as WES and RNA-Seq produced noisier genotypes compared to WGS and Hi-C experiments; single-cell RNA-Seq was the noisiest among all the assays, as expected (Figure 4c).

Supplementary Table 3: The functional genomics experiments used in this study with their total coverage

Individual	ENCODE ID/Source	Experiment	# of Reads	Read Length
NA12878	1kG	WGS	757,704,193	255
NA12878	1kG	WES	212,461,381	76
NA12878	Rao et al. 2014	Hi-C exp 1 PE1	219,616,072	101
NA12878	Rao et al. 2014	Hi-C exp 1 PE2	220,087,882	101
NA12878	Rao et al. 2014	Hi-C exp 2 PE1	448,843,710	101
NA12878	Rao et al. 2014	Hi-C exp 2 PE2	451,088,484	101
NA12878	Rao et al. 2014	Hi-C exp 3 PE1	536,684,803	101
NA12878	Rao et al. 2014	Hi-C exp 3 PE2	536,101,709	101
NA12878	ENCSR000CVT	Total RNA-Seq	227,501,266	202
NA12878	ENCSR000COQ	PolyA RNA-Seq	267,602,146	76
NA12878	ENCSR000AJA	Single-cell RNA-Seq1	38,377,124	100
NA12878	ENCSR000AJH	Single-cell RNA-Seq2	47,896,396	100
NA12878	ENCSR000AKF	H3K4me1	42,763,056	36
NA12878	ENCSR145XQO	HDGF	41,626,373	101
NA12878	ENCSR387QUV	RELB	25,652,682	101
NA12878	ENCSR000DZN	CTCF-Snyder	25,463,397	36
NA12878	ENCSR000AKA	H3K4me3	20,221,959	36
NA12878	ENCSR000DYS	JUND	18,701,295	36
NA12878	ENCSR000AOW	H3K79me2	16,073,184	36
NA12878	ENCSR000AKE	H3K36me3	15,239,685	51
NA12878	ENCSR000AOV	H2AFZ	14,724,790	36
NA12878	ENCSR000AOX	H3K9me3	14,049,420	36
NA12878	ENCSR000AKB	CTCF-Broad	11,026,086	51
NA12878	ENCSR000BIF	rnap2	10,428,778	36
NA12878	ENCSR000AKC	H3K27ac	10,410,928	51
NA12878	ENCSR000AKG	H3K4me2	9,815,194	51
NA12878	ENCSR000AKI	H4K20me1	9,757,368	51
NA12878	ENCSR000AKD	H3K27me3	8,454,639	51
NA12878	ENCSR000AKH	H3K9ac	7,981,456	51
NA12878	ENCSR000DKV	CTCF-Iyer	7,614,943	35
NA12878	ENCSR000BGD	rnap2	7,516,461	36
NA12878	ENCSR000BGR	PBX3	6,119,046	36
NA19239	ENCSR018VOS	ChIA-PET (H3K4me1)	335,232,702	PE 101
NA19239	ENCSR332ZHA	ChIA-PET (H3K4me2)	289,328,492	PE 101
NA19239	ENCSR952NXC	ChIA-PET (H3K4me3)	322,739,907	PE 101
NA19239	ENCSR761FUE	ChIA-PET (H3K27ac)	271,351,477	PE 101
NA19239	ENCSR000DLE	CTCF-Iyer	9,999,915	36
NA19238	ENCSR823TEV	ChIA-PET (H3K4me1)	286,387,111	PE 101
NA19238	ENCSR380UPB	ChIA-PET (H3K4me2)	294,881,881	PE 101
NA19238	ENCSR029IXY	ChIA-PET (H3K27ac)	289,564,091	PE 101
NA19238	ENCSR527RXH	ChIA-PET (RAD21)	339,707,301	PE 101
NA19238	ENCSR000DLD	CTCF-Iyer	16,368,229	36
NA12812	ENCSR281KLF	Repli-Seq	8,082,874	36
NA12813	ENCSR834FTN	Repli-Seq	9,999,915	36
NA10847	ENCSR000DYO	POLR2A	6,476,857	28
NA10847	ENCSR000DYM	RELA	19,376,644	28
NA18505	ENCSR000EAU	POLR2A	28,951,453	28
NA18505	ENCSR000EAW	RELA	22,274,656	28
NA18526	ENCSR000EAY	POLR2A	5,058,348	28
NA18526	ENCSR000EBA	RELA	6,353,939	28
NA18951	ENCSR000EBC	POLR2A	8,729,371	28
NA18951	ENCSR000EBD	RELA	5,514,493	28
NA19099	ENCSR000EBG	POLR2A	7,759,177	28
NA19099	ENCSR000EBI	RELA	11,961,302	28



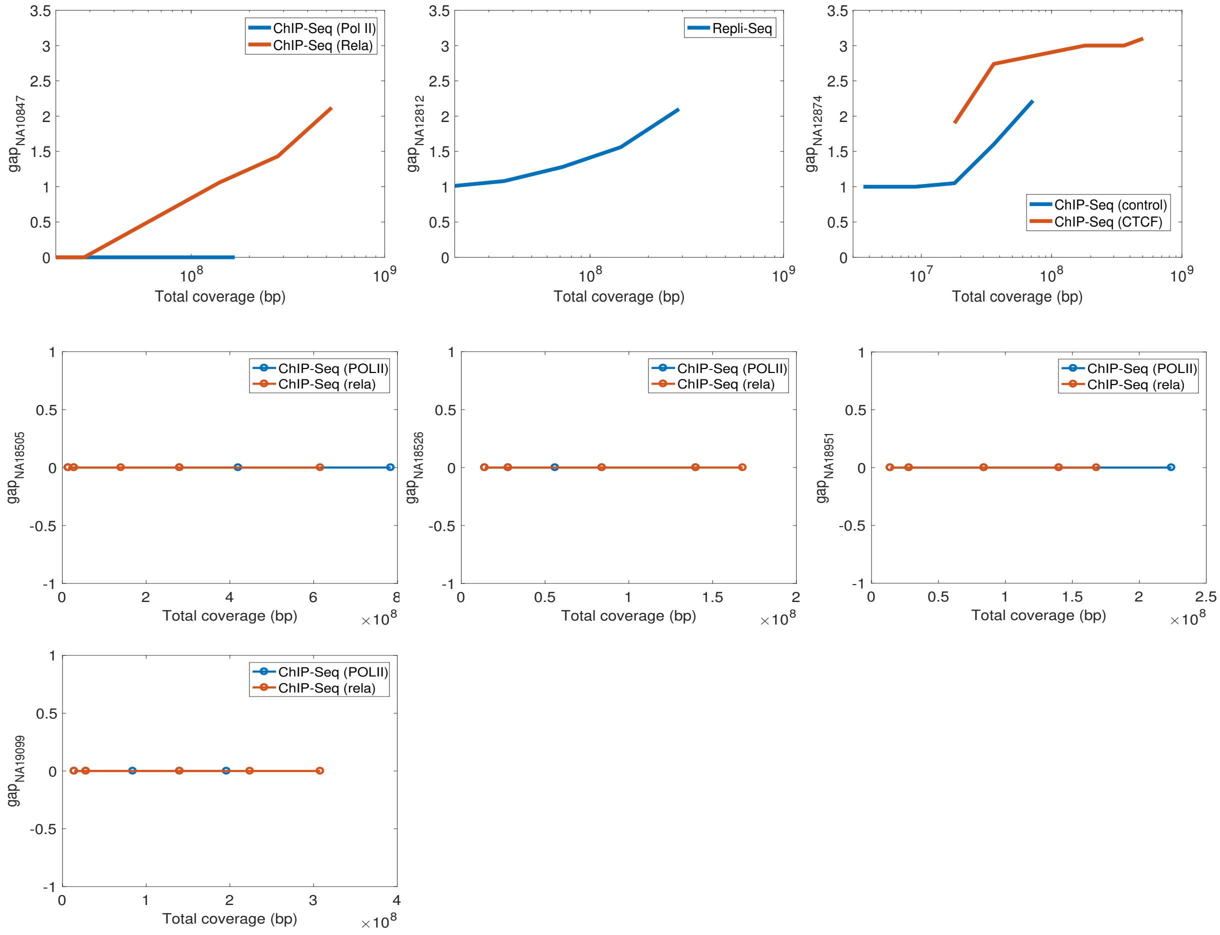
Supplementary Figure 4: **Rate of False Information of functional genomics experiments from NA12878 at different coverage** (**a**) RFI comparison for Hi-C and WGS data at different coverage. As the amount of coverage increases, the RFI decreases. Overall, variants from Hi-C consistently have lower RFI than the variants from WGS until the coverage reaches its maximum. (**b**) RFI comparison for WES and different RNA-Seq experiments at different coverage. In general, there is a decreasing RFI trend with increasing coverage, as seen in panel (a), except for single-cell RNA-Seq. The noise increased for single-cell RNA-Seq experiments as more reads were included. (**c**) RFI comparison for different ChIP-Seq experiments at different coverage. There was a general trend of decreasing RFI with increasing coverage.

1.8 Contribution of very rare and unique genotypes to $L(j,k)$ score

We calculated the number of unique, very rare and common genotypes for every individual in the 1000 genomes panel. We observed around 15,000 unique genotypes per individual. This contributes around $11 \times 15,000 = 165,000$ bits of information. We observed around 670,000 very rare genotypes, which have contribution of $7 \times 670,000 = 4,690,000$ bits of information on average. In total, the contribution of unique and very rare genotypes is $4,855,000$ bits of information. We then calculated the information in the genomes of all the individuals in the 1000 genomes phase III panel. Mean information per individual is around $2x10^7$ bits. The contribution of unique and very rare variants then becomes around 24% of the total information in an individual's genome, despite that the number of unique and very rare variant is only 3% of the total number of variants in an individual's genome. Note that this calculation is based on our scoring system adopted from Narayanan and Shmatikov [1], which assumes independence between variants.

1.9 Gap values for the individuals

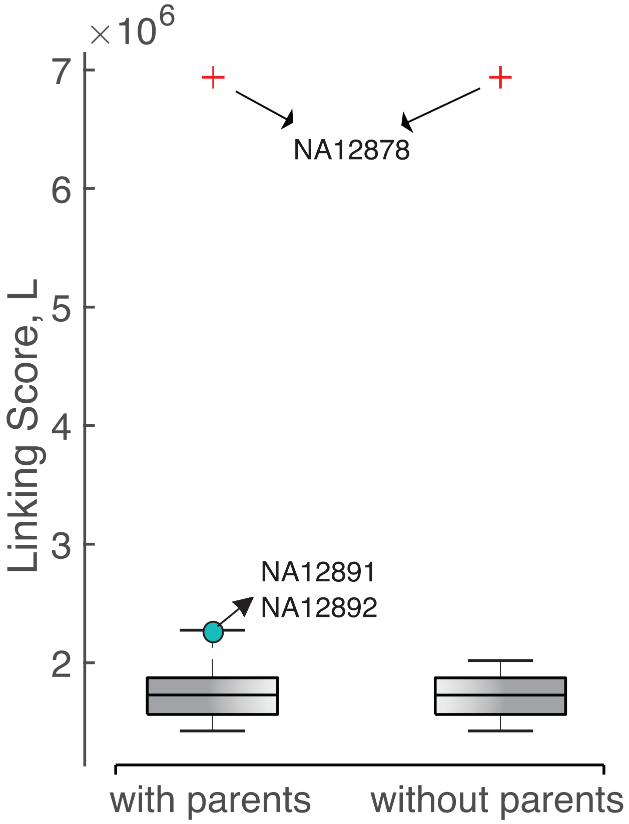
Gap values for the remaining 7 individuals with different functional genomics assays are shown in SI Figure 5.



Supplementary Figure 5: *gap* **values for the remaining 7 individuals with different functional genomics assays**

1.10 Linking individuals to a panel in the presence of trios

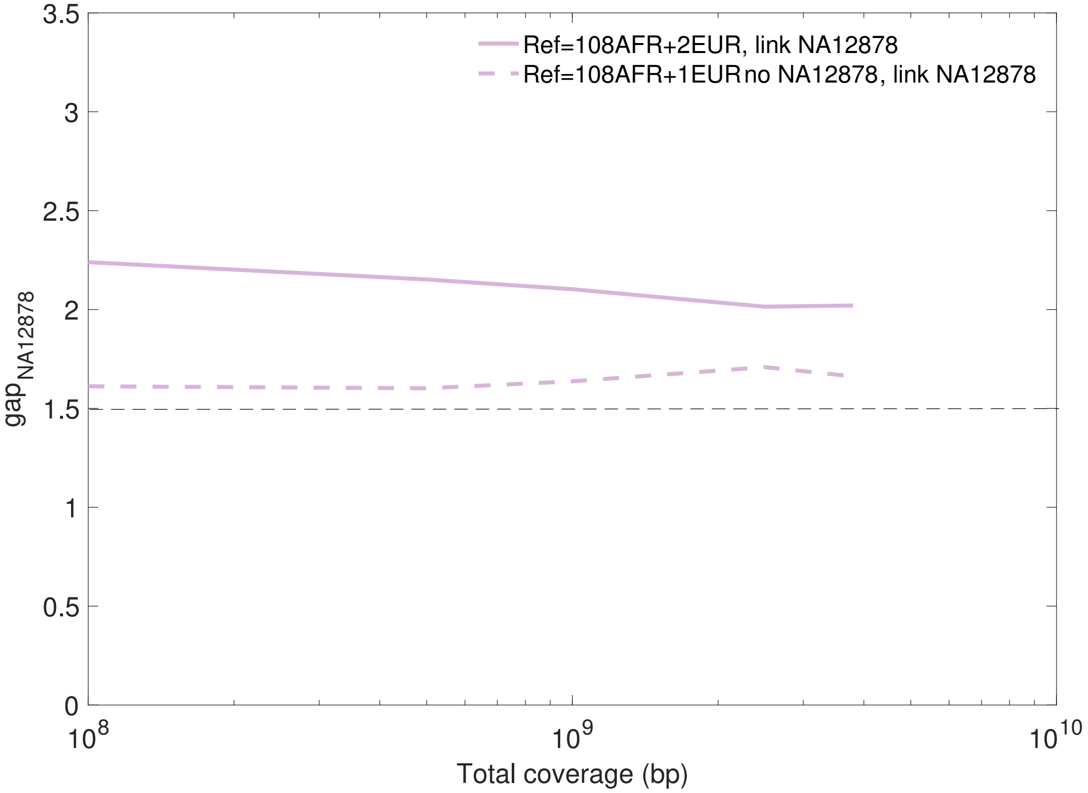
We first added the genotypes of NA12878's parents (NA12891 and NA12892) to the 1000 genomes panel and then calculated the $L(S^G_{NA12878}, S^{DB}_k)$ for all the individuals in the panel. Box plot (SI Figure 6) shows the distribution of the $L(S^G_{NA12878}, S^{DB}_k)$ values.



Supplementary Figure 6: **The distribution of $L(S^{G}_{NA12878}, S^{DB}_{k})$ values when the parents of NA12878 are added to the 1000 genomes genotype panel.**

1.11 Linking NA12878 to a different panel with or without NA12878 using functional genomics data

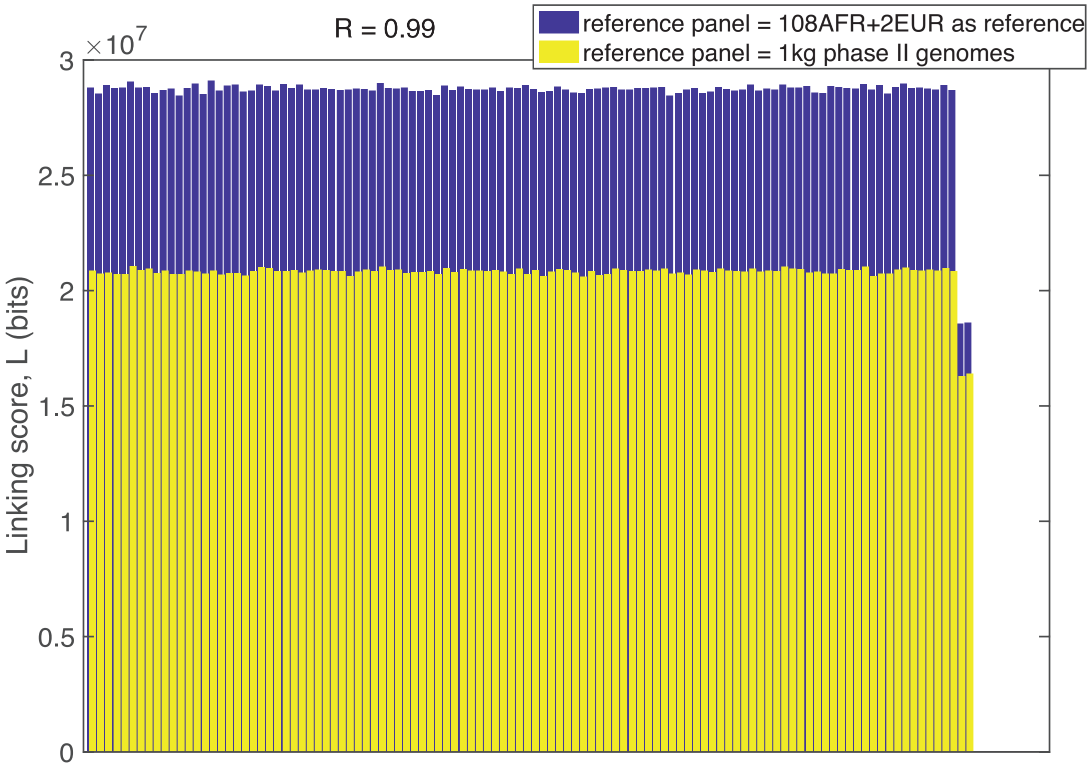
We showed that if we have a panel of individuals with a vastly different genotyping frequencies (108 AFR and 2 EUR individuals), we are still able to link NA12878 to the panel in all of coverage using the noisiest functional genomics assay (single-cell RNA-Seq). If we remove NA12878 from this panel, the we identify the other EUR individual as NA12878 due to the large difference between the EUR and AFR populations (SI Figure 7).



Supplementary Figure 7: **Linking NA12878 to a panel with 108 AFR and another EUR individuals, with and without NA12878 in the dataset.**

1.12 Comparison of linking scores using different reference panels

The linking scores of the individuals are calculated using the 1000 genomes genotyping frequency distribution as reference and a new panel consisting of 108 AFR and 2 EUR individuals as reference (Figure 8).



Supplementary Figure 8: **Linking score estimates for two different reference panels**

1.13 Privacy-enhancing file formats for functional genomics experiments

1.13.1 Anonymizing the BAM files

We went through all the attributes of the BAM files and grouped them into two category: (1) attributes to generalize with a common value and (2) attributes to keep as they are. The first category includes attributes that are leaking variants. They are the sequence of the read, cigar attribute and optional fields in the BAM files that are tagged with "AS" (alignment score) , "MD" (string for mismatching positions) and "NM" (string for distance to reference). MAPQ values can also be revealing at times and suggested to be sanitized in certain cases. Cigar gives out information about how many matching and nonmatching nucleotide there are in the read with respect to reference genome. As a result, one can call variants by looking at the non-matching nucleotides. We converted all the cigars to perfectly matching strings. For example, if the read length is 35 and the cigar is 14M1X15M, then the cigar is converted to 35M. AS reveals information about the number of matching positions in a read. An adversary can predict if a read contains variant by looking at the alignment score and subtracting it from the read length. MD reveals information about the mismatching positions and deletions in the reads and their corresponding nucleotides. For example, if there is a nucleotide in the read that is "A" in the 15th position of 30 bp long read, and if the reference allele for this position is G, then the MD tag will look like "MD:Z:14MA15M", which directly reveals the variant position in the read. NM reveals how many bases are different than the reference, which in turn gives away how many SNPs there are in the read. We converted all the alignment scores to the read lengths and all the MD, AS and NM tags to a perfectly matching string (for example "MD:Z:30M" for the example above). For the sequence attribute, we find the position of the read in the reference genome and replace the sequence attribute with the sequence in the reference. The rest of the attributes of the BAM files are designated as the second category and kept as they are.

1.13.2 pBAM

Privacy-enhancing file formats can be generated for SAM, BAM and CRAM files. For simplicity, we will refer the regular files as BAM and the privatized file format as pBAM. The difference between the regular files and the privatized files are on the fields of cigar, sequence, alignment score, the string for mismatching positions and the string for the distance to the reference. Note that any optional field that leak sensitive information about the sample can be manipulated. We focus on AS, MD and NM tags throughout this paper, since they are the most obvious leakages, but a module to manipulate any other tag can easily be added to pTools.

Let's assume read length for the sequencing experiment is 30, which is the total number of nucleotides in a fragment. Below are itemized description of how cigars are converted to privatized cigars along with examples.:

Cigars in non-intronic reads (i.e cigars with no 'N"):

- Cigar for perfectly mapped reads is a number of read length followed by the letter "M", indicating every nucleotide in the read is mapped to the reference human genome. This also means that there is no variant in this read (unless indicated in the MD tag). In this case, regular BAM has "30M" in the cigar and pBAM will have '30M" in the cigar as well.
- Cigar for reads that contain a mismatch is marked with the letter "X". For example, if the 10th nucleotide in the fragment has a mismatch, then the cigar in the regular BAM becomes "9M1X20M". This usually means that there is a SNP on the 10th nucleotide of the fragment. Since we know the start coordinate of the read from the regular BAM, an adversary can easily infer that there might be a SNP on the "$start + 10$"th coordinate of the

genome of the sample. To prevent that we convert "9M1X20M" to "30M" in the pBAM file. This conversion does not add any noise to the calculation of depth since "$start+10$"th is sequenced, however as a different letter.

- Cigar for reads that contain soft-clipping is marked with the letter "S". For example, if the first 5 nucleotides are soft-clipped from the fragment, then cigar becomes "5S25M". The start coordinate reported as the beginning of mapped nucleotides, which is the 6th nucleotide of the fragment. In this case, we report the cigar as "30M" and keep the start coordinate as it is. This is because soft-clipping can be due to a structural variant, insertion or a deletion. The associated noise with this conversion is that the coordinates between "$start+26$" and "$start+30$" gain extra read, i.e depth.

- Above point applies for the reads with hard-clipping that are marked by the letter "H". For example, if the nucleotides from 1st to 21st are hard-clipped from the fragment, then cigar becomes "20H10M". In this case, we report the cigar as "30M" ignoring the hard-clipped nucleotides. The associated noise with this conversion is that the coordinates between "$start$" and "$start+20$" gain extra read, i.e depth.

- Note that some analysis pipelines such as ENCODE RNA-Seq and ChIP-Seq processing pipelines do not include these clipped reads in their analysis, in which case these reads can be filtered out as they add the most noise to the signal after sanitization.

- Cigar for reads that contain an insertion is marked with the letter "I". For example, if the 23th to 30th nucleotide in the fragment is an insertion, then the cigar in the regular BAM becomes "22M8I". Since we know the start coordinate of the read from the regular BAM, an adversary can easily infer that there is an insertion on the "$start+23$"th coordinate of the genome of the sample. To prevent that we convert "22M8I" to "30M" in the pBAM file. The associated noise with this conversion is that the coordinates between "$start+22$" and "$start+22+8$" gain extra read, i.e depth.

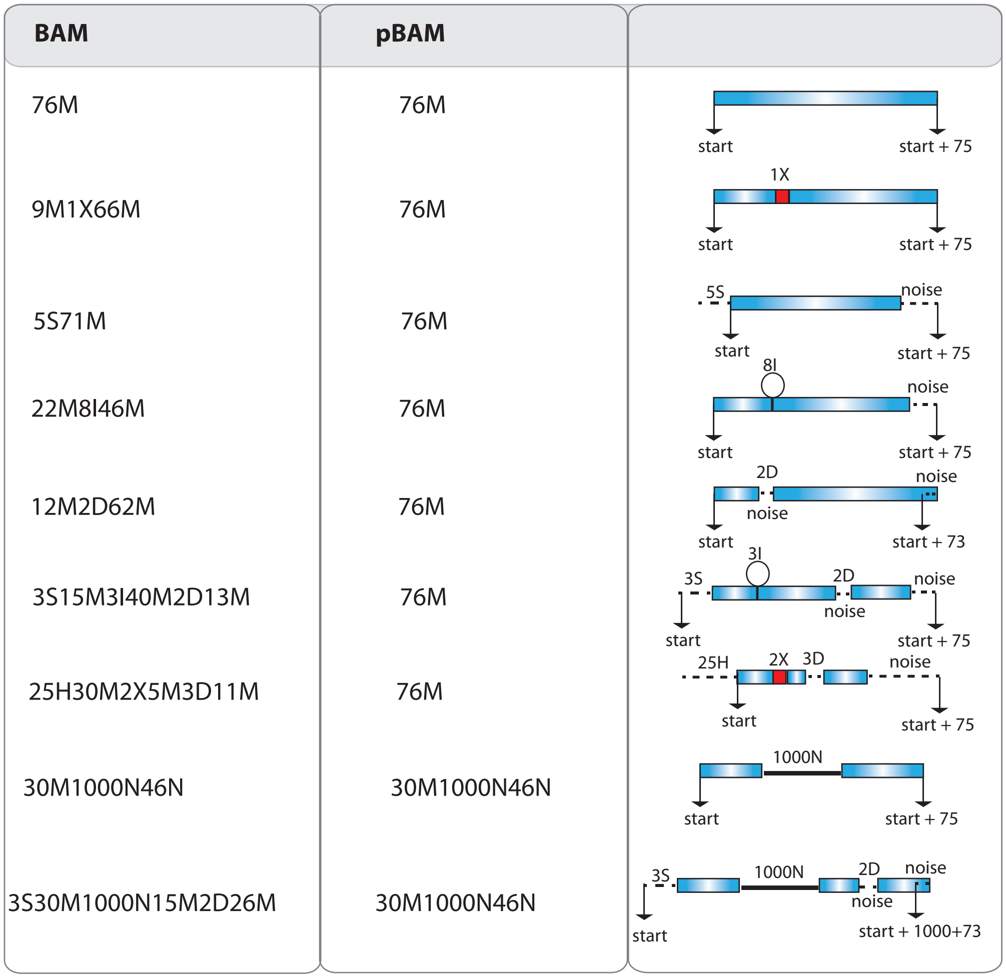
- Cigar for reads that contain a deletion is marked with the letter "D". For example, if the 13th to 14th nucleotide in the fragment is a deletion, then the cigar in the regular BAM becomes "12M2D16M". Since we know the start coordinate of the read from the regular BAM, an adversary can easily infer that there is an deletion on the "$start+12$"th coordinate of the genome of the sample. To prevent that we convert "12M2D16M" to "30M" in the pBAM file. This conversion add any noise to the deleted coordinates as well as the last 2 nucleotide at the end of the read as total read length has to be capped at 30.This also prevents signal profiles to leak the small deletions as the curve that corresponds to the deletion will look smooth based on its neighboring nucleotides.

- There are also cigars that may have multiple of the above letters. Here are a few examples and the solutions:

Cigars in intronic reads (i.e cigars with ‘N”):

- Cigar for perfectly mapped reads but split due to the introns are split by the letter “N”. For example, if there is a 1000 nucleotide long intronic region between mapped regions, it can have a cigar as “10M1000N20M”. In this case pBAM will have a cigar of “10M1000N20M” as well.
- If the reads are split in the mapped regions due to mismatch, insertion, deletion or clipping, then pBAM deals with them such that splice sites are as accurate as possible. Here are few examples;
  - Cigar “3S15M1000N10M2D” becomes “18M1000N12M”, which does not add any noise to the splice site.
  - Cigar “10M3D3M1000N3M2I9M” becomes “16M1000N14M”, which does not add any noise to the splice site.

Details of these examples are depicted in Figure 9.



BAM	pBAM	
76M	76M	
9M1X66M	76M	
5S71M	76M	
22M8I46M	76M	
12M2D62M	76M	
3S15M3I40M2D13M	76M	
25H30M2X5M3D11M	76M	
30M1000N46N	30M1000N46N	
3S30M1000N15M2D26M	30M1000N46N	

Supplementary Figure 9: **Visual representation of mapped fragments before and after converting the cigars for pBAM file format.** The insertions, deletions, soft and hard-clipping as well as intronic reads are depicted. The noise that is added to the pBAM file in order to enhance privacy is also depicted in the fragments.

1.13.2.1 Transcriptome alignments

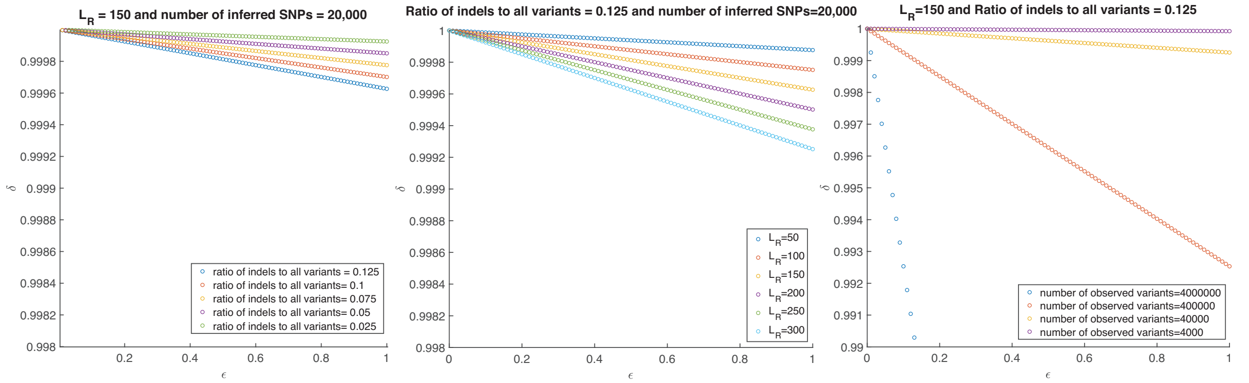
pTools searches the reference transcriptome for the position of the transcripts and reports the reference transcriptome sequences in the pBAM. We used the reference transcriptome files that are generated by RSEM software.

1.13.3 .diff files

.diff files contain the difference between the original BAM files and the pBAM files in a compact form. If the information is already available in the reference human genome such as sequence of the fragment, then the .diff file does not report it. This is done to keep the .diff files as small as possible. These are the files that require special permission to access and contains the private information about the individual. To be able to go back and forth between BAM and pBAM files using the .diff files, the BAM and pBAM files are required to be coordinate sorted.

1.13.4 Utility-privacy balance

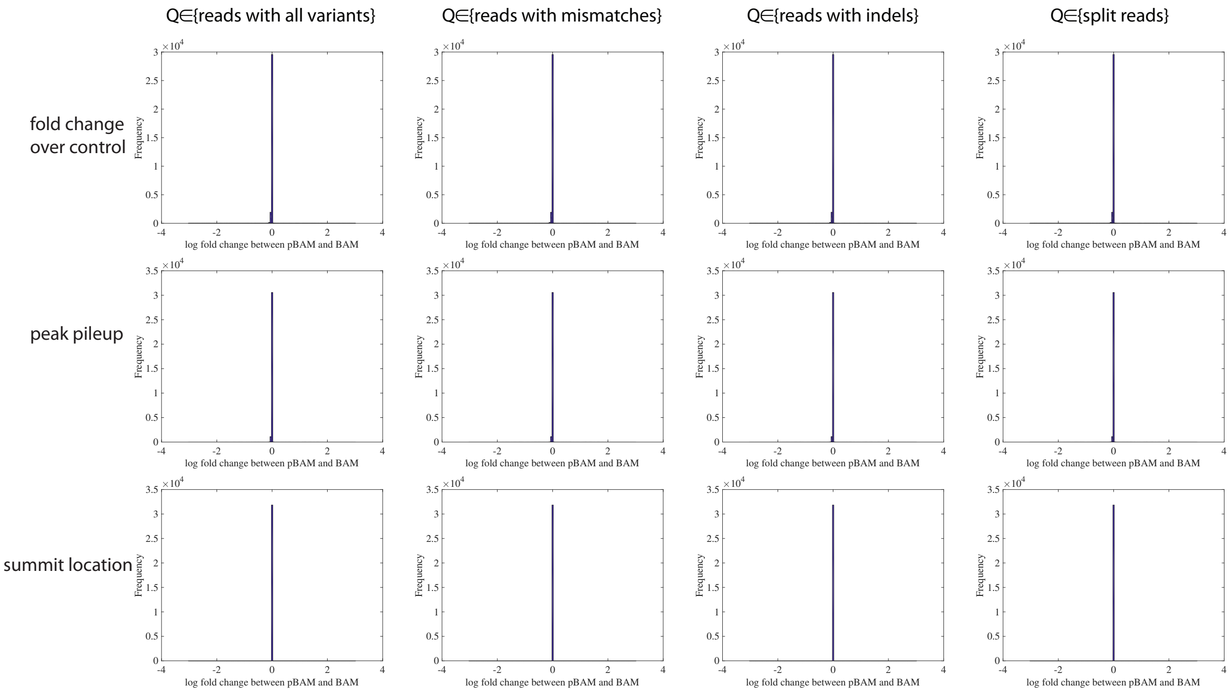
SI Figure 10 shows the utility-privacy balance under different read lengths, observed and sanitized number of variants.



Supplementary Figure 10: The balance between the privacy and utility under different conditions

1.13.5 Utility of the pBAM files

SI Figure 11 shows the difference of various quantification metrics from ChIP-Seq data when BAM *vs.* pBAM files are used.



Supplementary Figure 11: **The difference between ChIP-Seq peak calling using BAM and pBAM as input for the fold over change compared to control, the number of reads that pile up on the location of peak and the location of the peak summit.**

1.14 Relation to differential privacy

Differential privacy ensures a high level of privacy such that adversary retrieves similar result with and without the addition of the individual's data to the database [13]. A randomized algorithm A that retrieves results $A(D)$ from database D is considered ε-differentially private if the results satisfies the condition

$$\frac{prob(A(D) = C)}{prob(A(D_{\pm i}) = C)} = e^{\varepsilon}, \tag{1}$$

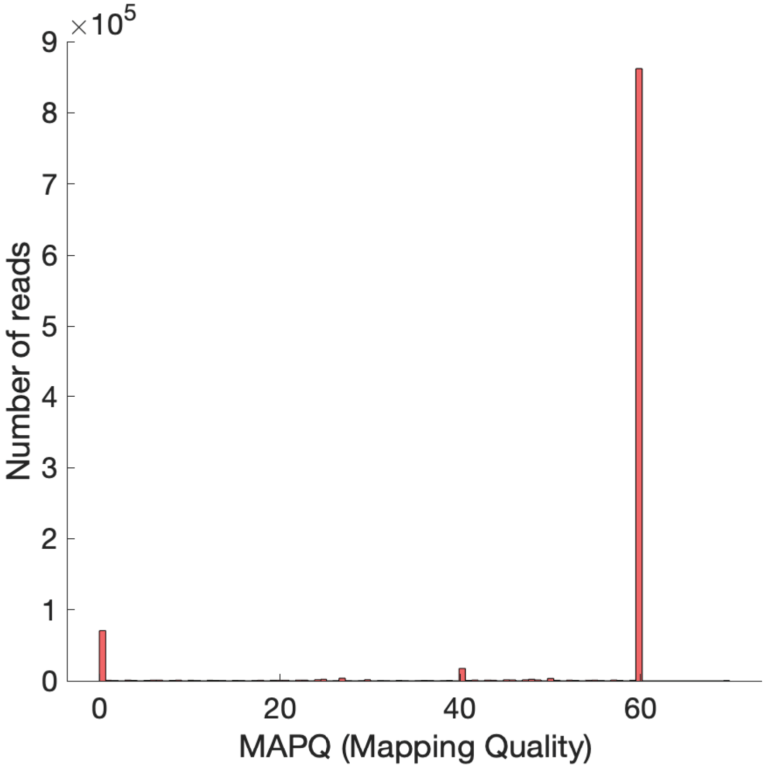
where $D_{\pm i}$ indicates the addition or subtraction of i^{th} individual to the database. This concept applies to databases of individuals, in which database itself is not released and calculations from this database (i.e algorithm A) is randomized such that adversary cannot infer information about individuals in the database.

We fist tried to see if we can apply differential privacy to BAM files, where we consider each read in the BAM file as an entry and the file itself as a database. The idea is that everytime we retrieve a read from BAM file, it will be manipulated such that with or without the retrieved read, when genotyping is performed the results will be the same, hence one cannot infer the variant in that retrieved read.

However since our desire is to be able to use the data for further processing such as testing a newly developed algorithm or quantifying gene expression without the need to go through special access process, retrieving information from BAM files one read at a time, while satisfying the differential privacy is not practical. Moreover, ensuring that the final pile of reads will have high enough utility to make any biological conclusions is challenging as randomizing the data might affect the conclusions.

1.14.1 Leakage from MAPQ

We found that read with MAPQ values that are smaller than mean MAPQ contain insertions, deletions, soft and hard clipping more than expected (Figure 12), hence they might leak the location of large SVs. In Figure 12, we analyzed a subsampled BAM file from a whole genome sequencing data. The BAM files from functional genomics data are noisier than WGS, however the MAPQ values could still potentially be a source of variant leakage.



Observed = # of reads with cigar feature below cut-off / # of reads with cigar feature total
Expected = # of reads below cut-off / # of total reads

MAPQ <= 10

	X	I	D	S	H
Observed / Expected	0	2.59	1.69	2.31	10.66

MAPQ <= 20

	X	I	D	S	H
Observed / Expected	0	2.58	1.79	2.42	10.33

MAPQ <= 30

	X	I	D	S	H
Observed / Expected	0	2.41	1.78	2.40	9.27

Supplementary Figure 12: **Potential variant leakage from MAPQ scores. As can be seen, the reads with potential large SVs have smaller than expected MAPQ scores. An adversary can sort the MAPQ scores in a BAM file and guess the location of these SVs that are mapped with low MAPQs.**

References

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

[2] 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.

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

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

[5] Kullback S. Information Theory and Statistics. *John Wiley & Sons.*, 1959.

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

[7] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 2009;80:5(6):e1000529.

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

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

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

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

[12] Bertran M, Martinez N, Papadaki A, Qiu Q, Rodrigues M, Guillermo Sapiro G. Learning Representations for Utility and Privacy: An Information-Theoretic Based Approach https://ppml-workshop.github.io/ppml/papers/15.pdf *NeurIPS Conference*, Privacy Preserving Machine Learning Workshop, accepted

[13] 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