1 Suplementary Information

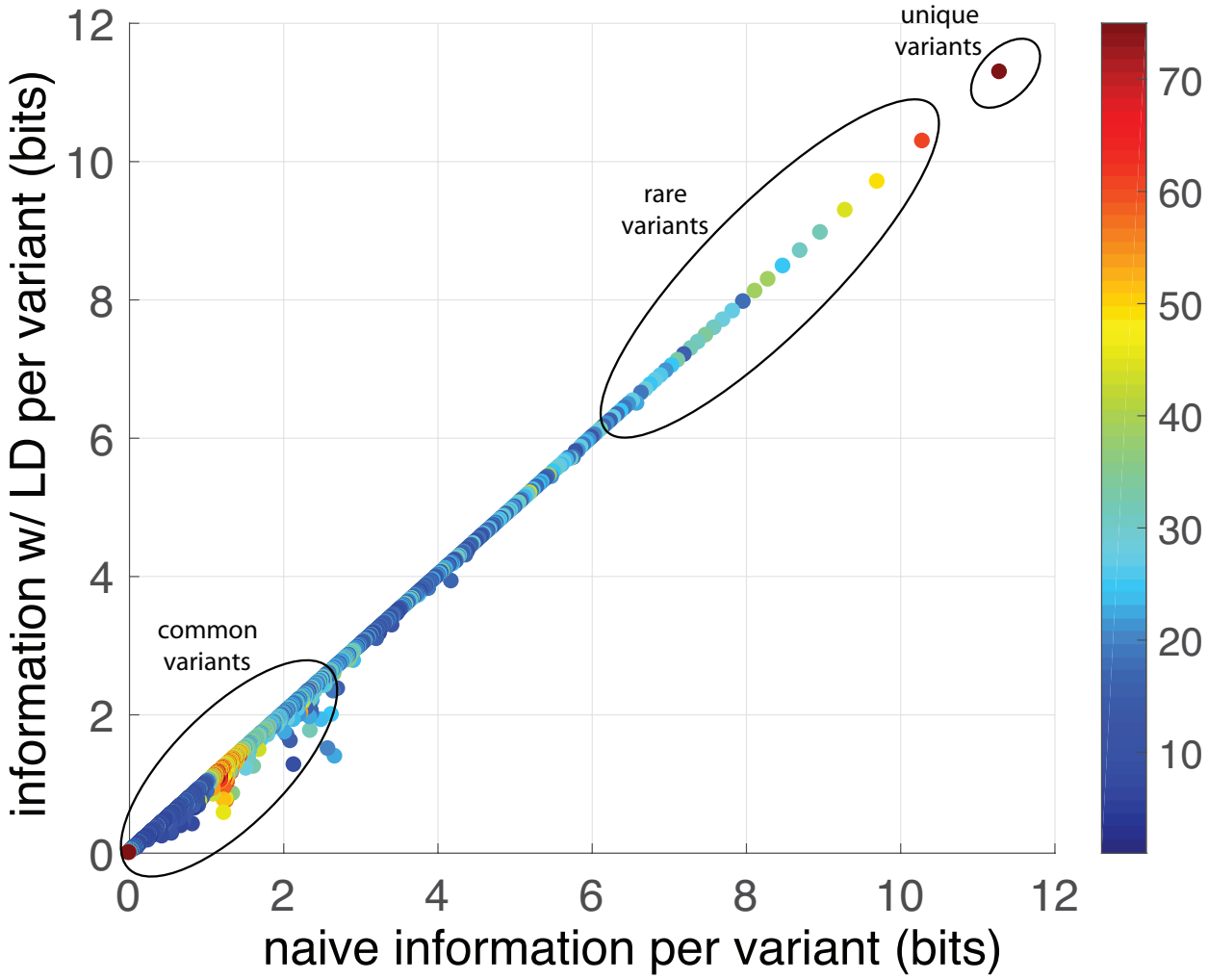
Contents

List of Figures

List of Tables

1.1 Calculation of information with and without LD consideration

We calculated the information per variant ($h(s_i) = -log_2(p(s_i))$ and information per variant with LD consideration ($h^{LD}(s_i) = -(1 - mLD(s_i, s_j))log_2(p(s_i))$) separately and plotted the concordance with respect to the number of times a variant with same $h(s_i)$ and $h^{LD}(s_i)$ occur in the genome of NA12878. We found that information with or without LD consideration does not differ for rare and unique variants, which contribute to the overall information calculation the most. LD consideration changes the information of a variant, when the variant is common in the 1000 genomes panel (Figure 1).



Supplementary Figure 1: **Information per variant with and without LD consideration** LD consideration affects only common variants, in turn does not affect overall information of the genome (h(S)).

1.2 Simulation of individuals

In this study, we simulated individuals that belong to European (CEU) and African (YRI) populations. These individuals are simulated based on the genotype frequency of derived from the population in 1000 genomes panel [1] and the LD relationship of the population in HapMap project [2]. Once we determine the population that the simulated individual belongs to, we draw a random variant with its genotype based on the probability of having that variant with the genotype in the

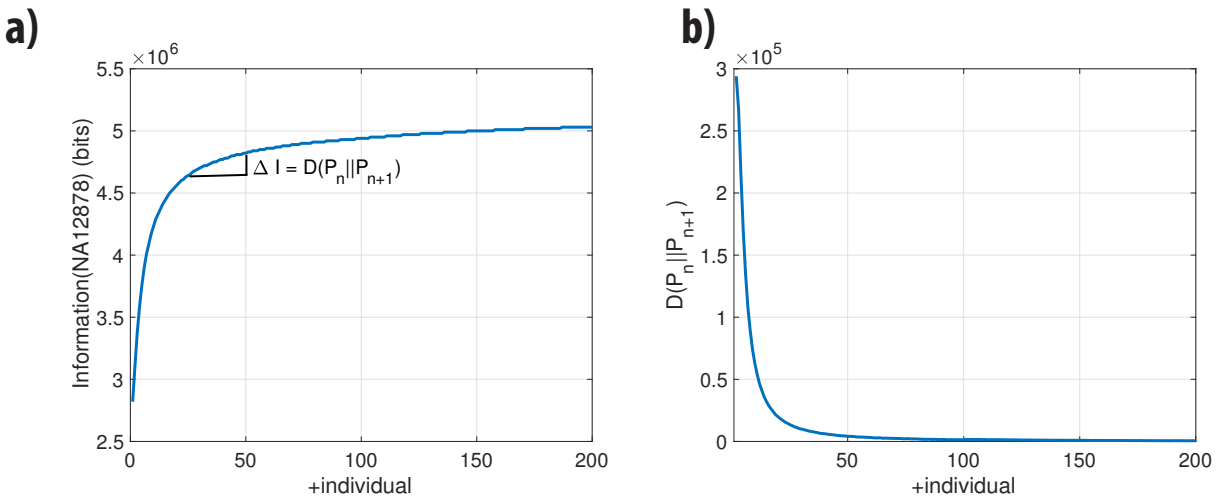
population. We then determine all the other variants that are in LD with the drawn genotype. We use the LD correlation as the joint probability of two variants being observed simultaneously and based on that we decide if the correlated variant will be simulated or not. The joint probability is adjusted based on the genotype of the two variants. If both variants are simulated as homozygous, then the joint probability is assumed to be equal to the LD correlation. If at least one of the variants is heterozygous, then the joint probability is assumed to be equal to the half of the LD correlation. We continue this process till we exhaust all the variants observed in the population. We first simulated 100 individuals that belong to CEU and then 100 more individuals that belong to YRI populations (Figure 2).

1.3 *KL*-Divergence

KullbackLeibler (KL) divergence is a measure to quantify the difference between two probability distribution. For discrete probability distribuions P and Q, the *KL*-divergence from Q to P is calculated as [3]

$$D_{\mathrm{KL}}(P||Q) = -\sum_i P(i) \log \frac{Q(i)}{P(i)} \tag{1}$$

In the context of simulating individuals, we interpreted *KL*-divergence as the information gain achieved by the addition of a new individual to the population, i.e. $D_{\mathrm{KL}}(P^{n+1}||P^n)$, where n is the size of the population before the addition of new individual (Figure 2).



Supplementary Figure 2: **Change in information as the population size increases** (**a**) Information vs. the increasing number of individuals in the population. (**b**) *KL*-divergence between the population with n individuals vs. $n+1$ individuals.

1.4 Overview of the different information measures

In this study, we used self-information and related measures to quantify the private information leakage in functional genomics data. Self-information is also known as surprisal is a measure of surprise of occurance of an event given its probability. In our study, we have the probability of observing a variant in a genome derived from 1000 genomes panel. For example, if a variant is observed in small number individuals in the 1000 genomes panel, then surprisal of calling that variant from the sequencing of a sample will be high. The relation between self-information and entropy is as follows. If we sequence a genome from the same sample say 1000 times. If we do variant calling for each sequencing run, then we can have a probability of calling a variant. The expected value of self-information is then defined as entropy.

Next measure we used is the pointwise mutual information (*pmi*), which is a measure of association. We used this measure to quantify the association between the called variants from a seqencing data of a sample and the gold standard variants of the same sample in the 1000 genomes

panel. The self-information of the variants that are called from the sequencing data and observed in the gold standard data is the *pmi* between the seuqncing data and the gold standard. This measure helps us to quantify how much of the information that is in the gold standard is captured by the sequencing experiment. *pmi* is also used to link a sequencing data from an unknown sample to a database of individuals. The relationship between *pmi* and mutual information is similar to the relationship between self-information and entropy. *pmi* is calculated for single event, whereas mutual information refers to all possible events. Mathematically, *pmi* quantifies the coincidence of event x and y occuring together given their joint distribution. If the joint probability of occurence of x and y is $p(x,y)$ assuming independence, then

$$pmi(x;y) = log\frac{p(x,y)}{p(x)p(y)} = log(p(x,y)) - log(p(x)) = -h(x,y) + h(x) + h(y).$$

Since $p(x,y) = p(x|y)p(y)$, then the equation can be rewritten as

$$pmi(x;y) = log\frac{p(x|y)p(y)}{p(x)p(y)} = \log\frac{p(x|y)}{p(x)} = log(p(x|y)) - log(p(x)) = -h(x|y) + h(x).$$

Another way of expressing the equation is using the relationship of $p(x,y) = p(y|x)p(x)$,

$$pmi(x;y) = log\frac{p(y|x)p(x)}{p(x)p(y)} = \log\frac{p(y|x)}{p(y)} = log(p(y|x)) - log(p(y)) = -h(y|x) + h(y).$$

We quantified the amount of information from the false positive variants that are called from the sequencing data (say x) but not in the gold standard (say y) as the surprisal from the occurence of event x given the event y has occured, which is $h(x|y) = -log(p(x|y))$.

Normalized *pmi* (*npmi*) is also calculated to incorporate the false positive variants ($h(x|y)$) and the gold standard variants that are not called from sequencing data ($h(y|x)$) to the measure.

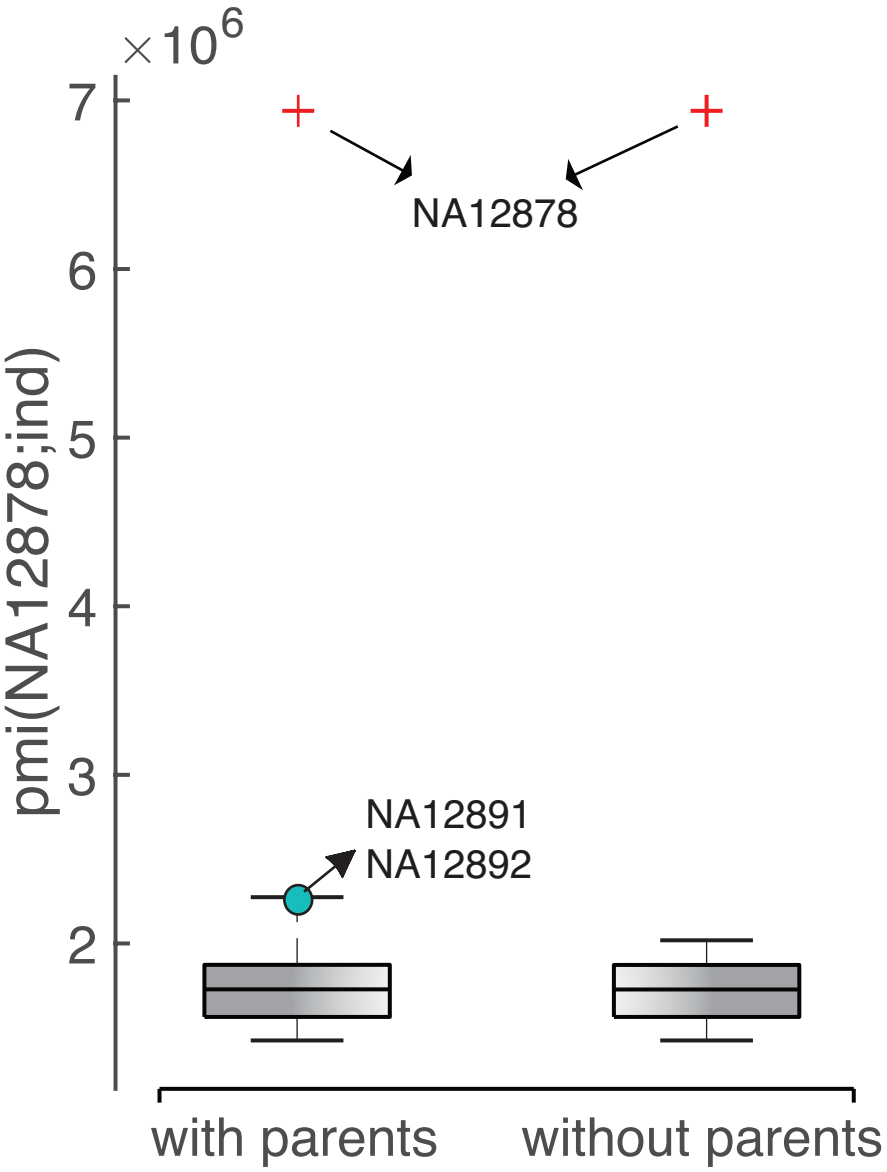
$$npmi(x;y) = \frac{pmi(x;y)}{h(x,y)} = \frac{pmi(x;y)}{h(x)+h(y)-pmi(x;y)}$$

If *npmi* is equal to -1, then the events x and y never occur together; if *npmi* is equal to 0 then event x and y are independent; if *npmi* is equal to 1 then event x and y completely co-occur.

Figure 2a in the main text shows the relationship between these measures in a Venn diagram, which is adopted from ref. [4].

1.5 Linking NA12878 to the 1000 genomes panel in the presence of their parents

We first added the genotypes of NA12878's parents (NA12891 and NA12892) to the 1000 genomes panel and then calculated the $pmi(S^G_{NA12878}; S^{DB}_j)$ for all the individuals in the panel. Box plot (Figure 3) shows the distribution of the *pmi* values.



Supplementary Figure 3: **The distribution of pmi values when the parents of NA12878 are added to the 1000 genomes genotype panel.**

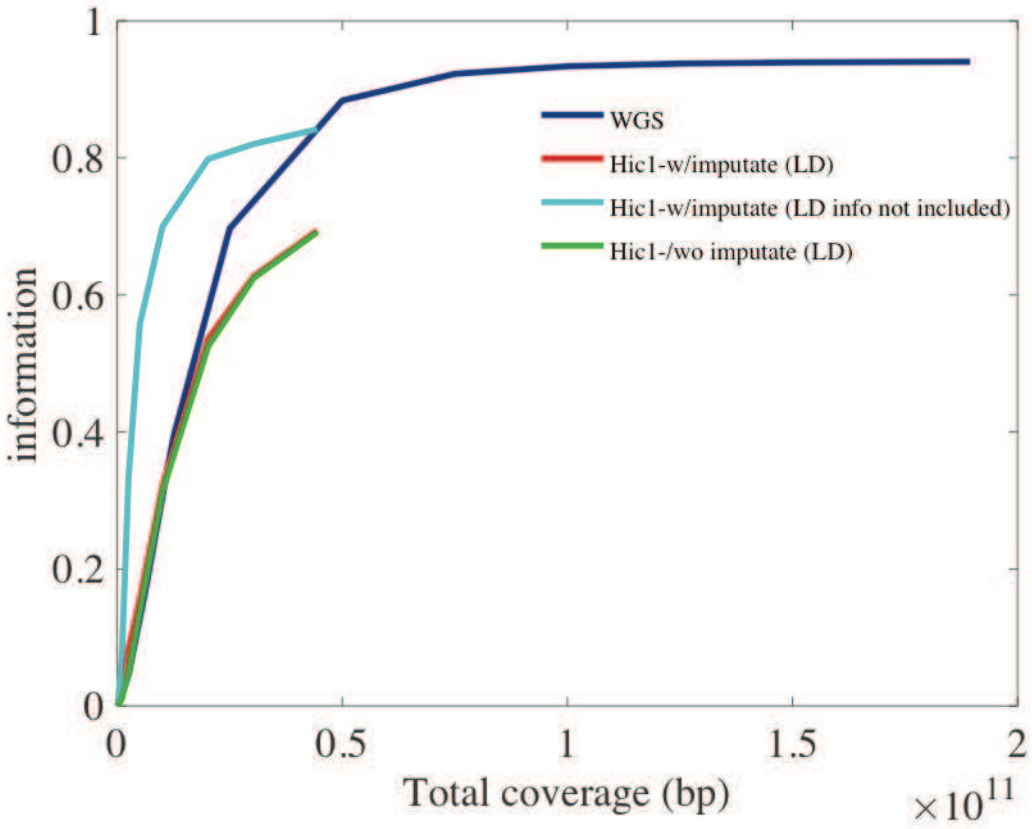
1.6 Calculation of information after imputation

When a variant is imputed, naturally the amount of information gained from imputed variant is low as we have prior information on the probability of observing imputed variant due to the LD correlation. IMPUTE2 [5, 6, 7] prints a probability for each genotype for a given variant and an

info column that is a measure for the confidence of observing the variant. Confidence value (con_i for variant s_i) is reported as a number between 0 and 1. We first removed all the imputed variants that have confidence below 0.3. We then selected the genotypes that have the highest probability for each imputed variant. The confidence is used as a prior information on the probability of observing the imputed variant. We then calculated the information gain from the imputed variant s_i as,

$$h^{im}(i) = (1 - con_i)h(s_i) \tag{2}$$

Figure 4 shows the difference in the information gain with and without considering imputation a priori information.



Supplementary Figure 4: **The contribution of imputed variants to the naive information** When we do not consider the a priori information we obtained from imputation, we infliate the information gain from the imputed variants (cyan curve). When we remove the a priori information from the information gain, it shows that there is negligible information gain from the observation of imputed variants due to high correlation (red curve). The information before the imputation is depicted with green curve.

1.7 Gaussian Process Regression (GPR) to estimate information from sequencing properties

In order to increase the number of data points, we used sampled reads as new data points. These downsampled data points are from Hi-C as Hi-C experiments have large number of reads and sampling does not alter the sequence depth distribution. Moreover, if the downsampled data contains reads in the range of 1 million bp to 10 million bp, it can mimick ChIP-Seq data. This allowed us

to have total of 45 data points. We found that normalized pmi for these data points ranges between 0.005 and 0.97. To avoid the problems due to the precision, we magnified *npmi*s by a factor of α as following;

$$npmi(S^{FGE};S^{GS}) = \frac{1}{\alpha} f(\overline{d}_{FGE}, b_{FGE}, \beta_{FGE})$$

We set α to 100. We then randomly selected 5 data points and removed them from the dataset to use it as independent test case. For the remaining 40 data points, we tried many regression learners including linear regression, regression trees, Support Vector Machines, and Gaussian Process Regression. Although they all exhibit good prediction with low root mean square errors (RMSE), Gaussian Process Regression with an exponential kernel reported the lowest RMSE. The GPR is a non-parametric Bayesian approach, which is powerful to capture noisy relationships between inputs and output by optimizing large number of parameters hence allowing the level of complexity to be decided by the data through Bayesian inference [8]. We used MATLAB's Statistics and Machine Learning toolbox to perform fitting.

1.8 Privacy-enhancing file formats for functional genomics experiments

1.8.1 k-anonymity for BAM files

We went through all the attributes of the BAM files and grouped them into three category: (1) attributes to supress with an asteriks, (2) attributes to generalize with a common value and (3) attributes to keep as they are. The first category includes the attributes sequence and quality string for the sequence. The reason for doing suppression instead of generalization to these attributes are to save disk space. The second category includes attributes that are necessary to have for several processing pipelines such as gene expression quantification and TF binding. They are the cigar attribute and optional fields in the BAM files that are tagged with "AS" (alignment score) and "MD" (string for mismatching positions). Cigar gives out information about how many matching and nonmatching nucleotied there are in the read with respect to reference genome. In turn, one can

call variants by lookimg 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. We converted all the alignment scores to the read lengths and all the MD tags to a perfectly matching string (for example "MD:Z:30M" for the example above). the rest of the attributes of the BAM files are designated as the third category and kept as they are.

1.8.2 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 and the string for mismatching positions (see section *k*-anonymity for details). Note that any optional field that leak sesitive information about the sample can be manupilated. We focus on AS and MD tags throughout this paper, since they are the most obvious leakages, but a module to manupilate any other tag can easily be added to p-tools.

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 results since "$start + 10$"th is sequenced, however as a different letter and processing of functional genomics data deals with the depth rather than the letter of the nucleotide.

- 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 6th to 25th are hard-clipped from the fragment, then cigar becomes "5M20H5M". 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 "$start + 6$" and "$start + 25$" gain extra read, i.e depth.

- 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 an 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 does not add any noise to the results, because if there is high depth around these 2 deleted nucleotides, there is functional enrichment in that fragment regardless of the deletion. 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 solution:

  - Cigar '3S15M3I2D7M" becomes "30M", which introduces noise to only to 6 nucleotides that are on the coordinates between "$start + 24$" and "$start + 30$".

  - Cigar "5H10M2X5M3D7M" becomes "30M", which introduces noise to only to 5th nucleotides that are on the coordinates between "$start + 27$" and "$start + 32$".

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 “10M1000N20N”. In this case pBAM will have a cigar of “10M1000N20M” as well.

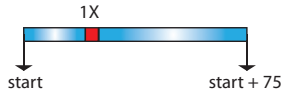
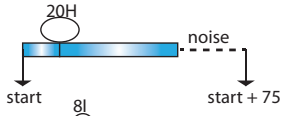
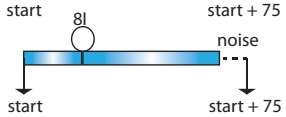
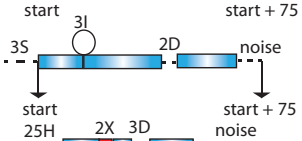
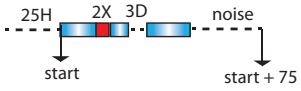
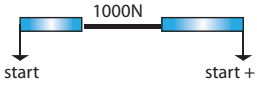
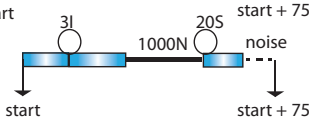
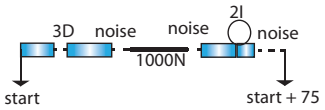
- If the reads are split in the mapped regions due to mismatch, insertion, deletion or clipping, then pBAM deals with them such that splice sites are as accurate as possible. Here are few examples;
  - Cigar “3S15M1000N10M2D” becomes “18M1000N12M”, which shifts the intronic region to right by 3 nucleotides.
  - Cigar “10M3I3M1000N2S2M” becomes “16M1000N4M”, which shifts the intronic region to left by 2 nucleotides.
  - Cigar “10M3D3M1000N3M2I9M” becomes “16M1000N14M”, which does not add any noise to the plice site.

Details of these examples are depicted in Figure 5.

1.8.2.1 Transcriptome alignments Since RSEM requires sequences to be present in BAM files, we can no longer put random strings to the sequence column in pBAM file. Therefore, we manipulated the sequences in BAM files and reported the reference transcriptome sequences in the pBAM.

1.8.3 .diff files

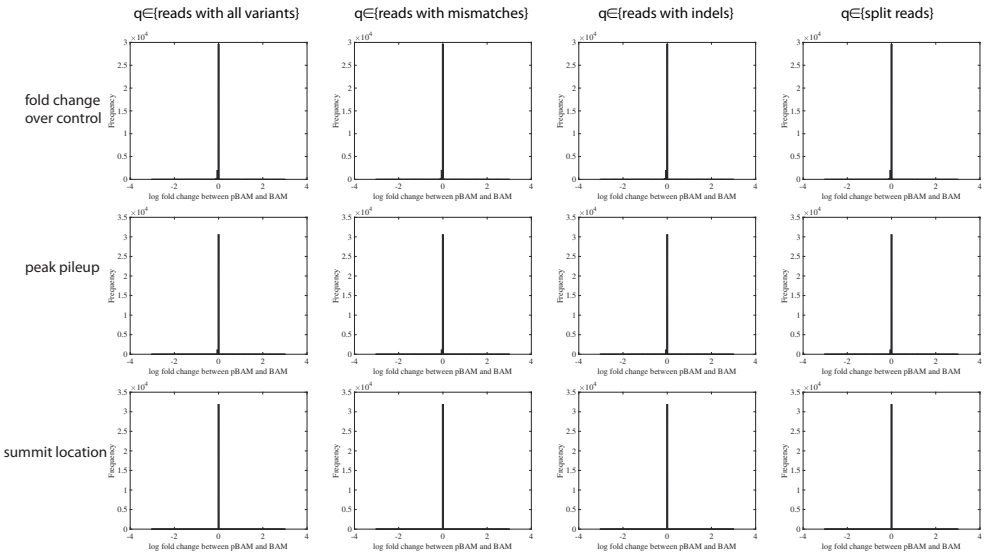
.diff files contain the difference between the original BAM files and the pBAM files in a compact form. If the information is already available in the reference human genome such as sequence of the fragment, then the .diff file does not report it. This is done to keep the .diff files as small as possible. 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.

BAM	pBAM	
76M	76M	
9M1X66M	76M	
5S71M	76M	
5M20H51M	76M	
22M8I46M	76M	
12M2D62M	76M	
3S15M3I40M2D13M	76M	
25H30M2X5M3D11M	76M	
30M1000N46N	30M1000N46N	
3S30M1000N15M2D26M	30M1000N46N	
10M3I23M1000N20S20M	33M1000N43N	
10M3D23M1000N30M2I8M	33M1000N43N	

Supplementary Figure 5: **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.

1.8.4 Utility of the pBAM files

Figure 6 shows the difference of various quantification metrics from ChIP-Seq data when BAM vs. pBAM files are used.

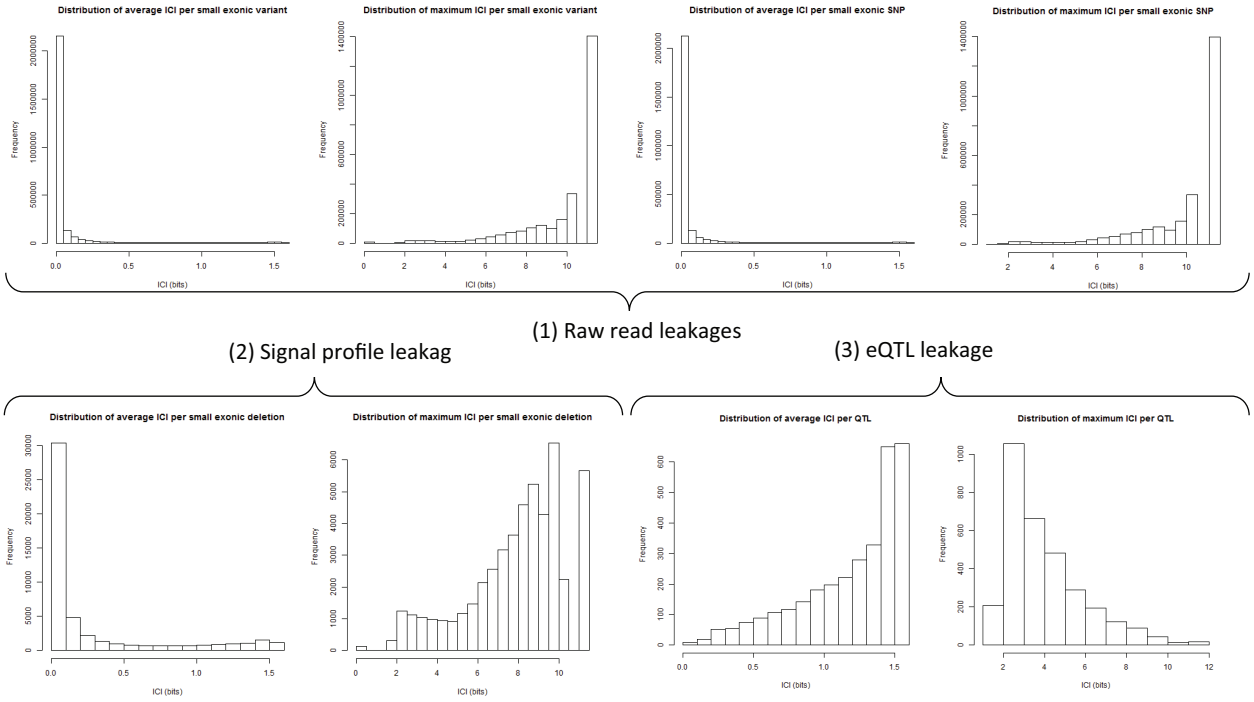


Supplementary Figure 6: **The difference between ChIP-Seq peak calling using BAM and pBAM as input for the fold over change compared to control, the number of reads that pile up on the location of peak and the location of the peak summit.**

1.9 Calculation of average and maximum leakage per variant

In the section "Comparison of private information leakage from different layers of data stack of RNA-Seq" of main text, we first overlapped the 1000 genomes variants with the exon annotations. We classified the variants into the categories of exonic variants, exonic SNVs (excluding indels), exonic indels and exonic small deletions. For each category, we calculated the self-information of the variant for all three possible genotypes (0, 1 and 2) as $h(s_0)$,$h(s_1)$ and $h(s_2)$. The average of self-information for each variant in each catorgory is the average information leakage and the $max(h(s_0), h(s_1), h(s_2)$ is the maximum information leakage for that particular variant. We then calculated the mean and standard deviation for all the variant in each category. Total information leakage is calculated as the product of the total number of accesible variants and the average

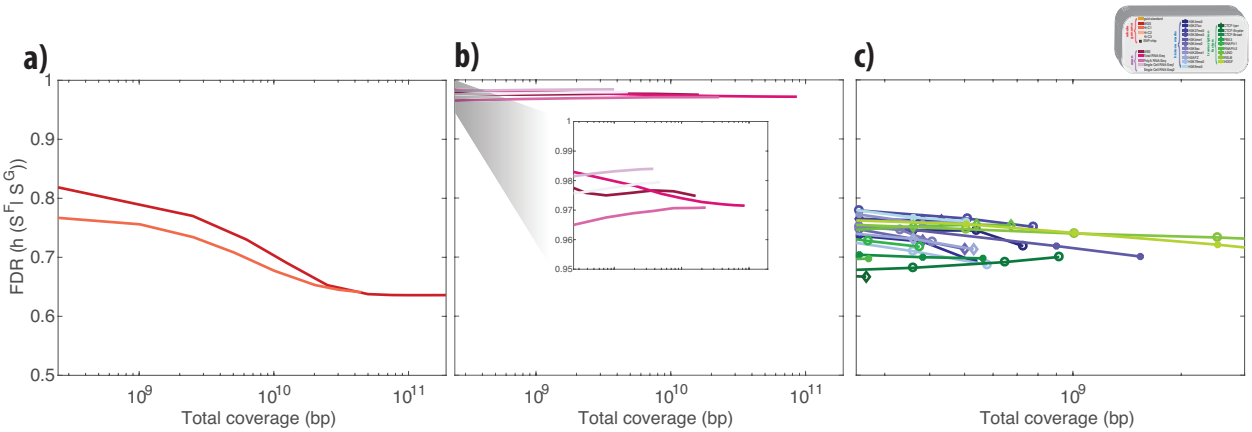
information leakage per variant. The distributions of the lakage can be seen in Figure 7.



Supplementary Figure 7: **The distributions of the information leakage per variant in different levels of the data stack. Individual characterizing information (ICI) is calculated based on ref [9].**

1.10 Contribtuion of *de novo* variants to FDR

Figure 8 shows how FDR values inflated when *de novo* variants are assumed to be false positives. This relates to genotyping a noisy sequencing experiment when the correct genotypes are unknown. In that case, any new genotype prediction can be seen as false positives even when they are actually *de novo* variants.



Supplementary Figure 8: **False discovery rates when *de novo* variants are assumed to be false positives**

1.10.1 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 [10]. A randomized algorithm A that retrives results $A(D)$ from database D is considered ε-differentially private if the results satisfies the condition

$$\frac{prob(A(D) = C)}{prob(A(D_{\pm i}) = C)} = e^{\varepsilon}, \tag{3}$$

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 retrive 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 quantifiying 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. Therefore, we decided to apply k-anonymity to BAM files to create pBAMs, where the user has the freedom to have all the reads at once and use the new file formats with any software that works with BAM format.

References

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

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

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

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

[5] 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.

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

[7] 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.

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

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

[10] 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

Supplementary Table 1: The functional genomics experiments used in this study with their total coverage

ENCODE ID/Source	Experiment	# of Reads	Read Length
1kG	WGS	757,704,193	255
1kG	WES	212,461,381	76
Rao et al. 2014	Hi-C exp 1 PE1	219,616,072	101
Rao et al. 2014	Hi-C exp 1 PE2	220,087,882	101
Rao et al. 2014	Hi-C exp 2 PE1	448,843,710	101
Rao et al. 2014	Hi-C exp 2 PE2	451,088,484	101
Rao et al. 2014	Hi-C exp 3 PE1	536,684,803	101
Rao et al. 2014	Hi-C exp 3 PE2	536,101,709	101
ENCSR000CVT	Total RNA-Seq	227,501,266	202
ENCSR000COQ	PolyA RNA-Seq	267,602,146	76
ENCSR000AJA	Single-cell RNA-Seq1	38,377,124	100
ENCSR000AJH	Single-cell RNA-Seq2	47,896,396	100
ENCSR000AKF	H3K4me1	42,763,056	36
ENCSR145XQO	HDGF	41,626,373	101
ENCSR387QUV	RELB	25,652,682	101
ENCSR000DZN	CTCF-Snyder	25,463,397	36
ENCSR000AKA	H3K4me3	20,221,959	36
ENCSR000DYS	JUND	18,701,295	36
ENCSR000AOW	H3K79me2	16,073,184	36
ENCSR000AKE	H3K36me3	15,239,685	51
ENCSR000AOV	H2AFZ	14,724,790	36
ENCSR000AOX	H3K9me3	14,049,420	36
ENCSR000AKB	CTCF-Broad	11,026,086	51
ENCSR000BIF	rnap2	10,428,778	36
ENCSR000AKC	H3K27ac	10,410,928	51
ENCSR000AKG	H3K4me2	9,815,194	51
ENCSR000AKI	H4K20me1	9,757,368	51
ENCSR000AKD	H3K27me3	8,454,639	51
ENCSR000AKH	H3K9ac	7,981,456	51
ENCSR000DKV	CTCF-Iyer	7,614,943	35
ENCSR000BGD	rnap2	7,516,461	36
ENCSR000BGR	PBX3	6,119,046	36