# ### Notes

<https://grants.nih.gov/grants/guide/pa-files/par-18-844.html>

Nov 16 deadline

Aim to meet next Fri early (11/19)

So goal is to clean the doc by Fri at 7 pm

750 words per page

Paste in the spec aims

≈

|  |  |
| --- | --- |
| sig | 563 |
| inv | 422 |
| pg1 | 477 |
| 2 | 604 |
| 3 | 509 |
|  | 564 |
|  | 574 |
|  | 627 |
|  | 705 |
| tot w/o micro | 5045 |
| div 750 | 6.726666667 |

# ## Spec aims in a sep. doc

==

[**http://bit.ly/pr-r01**](http://bit.ly/pr-r01)

**https://www.ebi.ac.uk/metagenomics/**

**1- Significance (GG: 563 words, 1 page)**

Privacy is one of the most important topics of debate in data science and stands at the corner of many different fields, including ethics, sociology, law, political science, and forensic science. Recently, genomics has emerged as one of the major foci of studies on privacy. This can mainly be attributed to the advancement of technologies for high-throughput biomedical data acquisition that bring about a surge of datasets [1,2,3]. As a result of steep declines in sequencing costs, a growing number of companies now offer to collect, analyze and return genomic information direct to the public. In addition to the commercial entities, there are a number of research organizations that also take advantage of these technological developments to collect and process thousands of genomic datasets for research. However, in addition to concerns that patients provided with their genomic data may either under-estimate or over-estimate their current situation, there are additional concerns that may be underappreciated by the clinical community: in particular, participant privacy.

In contrast to standard medical data, genomic information, often single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and other large scale complex rearrangements like structural variations (SVs) data, is inherently personally identifiable information. By its very definition, raw genomic data identifies the owner. In contrast to other forms of medical data, genomic data is largely shared with close family members. Therefore, even in instances where patients have provided broad permissions to use and access their genomic information, care must be taken to persevere patient privacy as the data implicates not only the immediate owner of the sequence but many third- party relatives as well. In addition, human genome sequencing can result in reads that are sequenced from the exogenous species potentially belong to the microbiome flora of the individual. The amount of these reads can be even more pronounced when tissues are used for sequencing rather than germline. Although not studied as extensively as genomic variants, microbiome codes were also shown to be used for re-identification purposes [4].

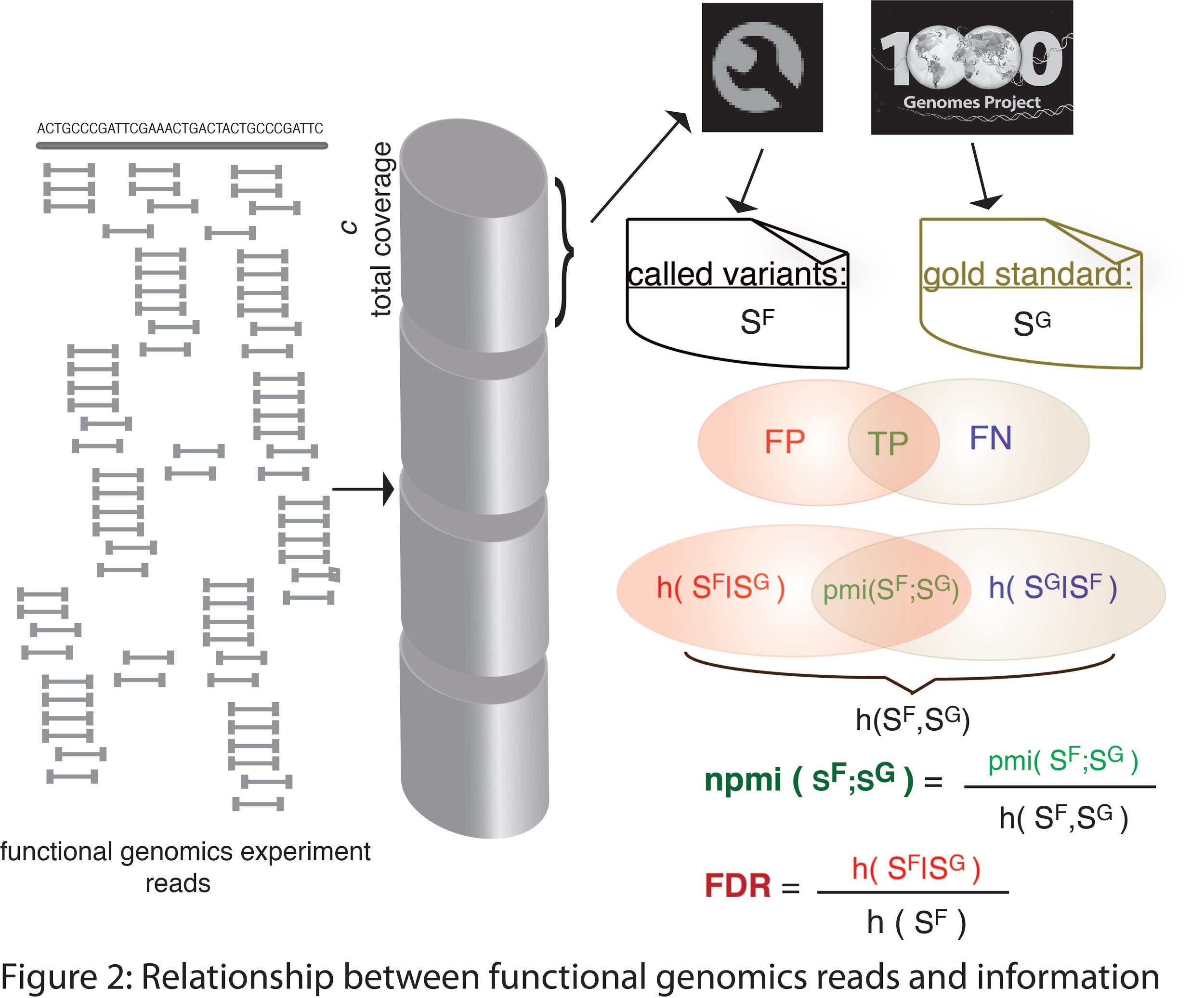
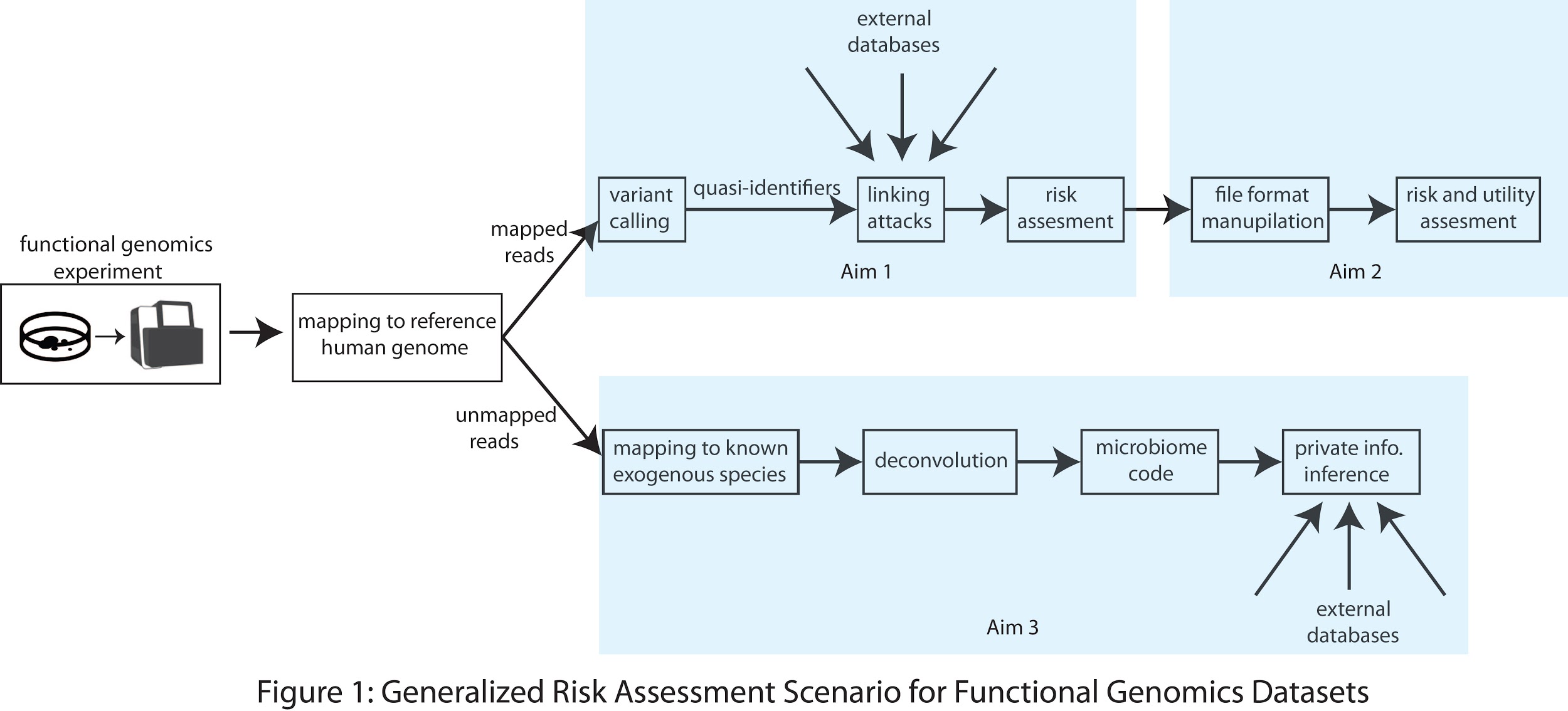
The advancement of sequencing technologies and laboratory techniques has enabled development of many assays to probe, for example, the epigenetic and transcriptomic states of the cell, e.g. gene expression levels and DNA binding protein levels and some of these are medically actionable [5,6]. Many consortia, like GTEx [7], ENCODE [8], 1000 Genomes [9], and TCGA [10], are generating large amount of personalized functional genomics datasets. However, unlike genetic variants uncovered through the more straightforward sequencing route, functional genomics data includes not only legacy data relating to prederminable conditions, traits, sex and race, and the like, but functional genomic data has the potential for being even more intrusive as it can reflect even privately held life choices, for example diet and where one resides through the microbiome inference as these experiments are mostly done on tissues.

One can even combine recent technologies to further estimate the genetic variants from other genomic measurements with a high degree of accuracy [11,12]. For example, the activity levels of some of the genes are very highly correlated with some of the variants, which are referred to as expression quantitative trait loci (eQTL). These eQTL can then be used for re-identification purposes. Although a lot of the studies concentrate on protecting the germline DNA sequence variants, the attention on the privacy protection of functional genomics data has been quite limited [13,14,15].

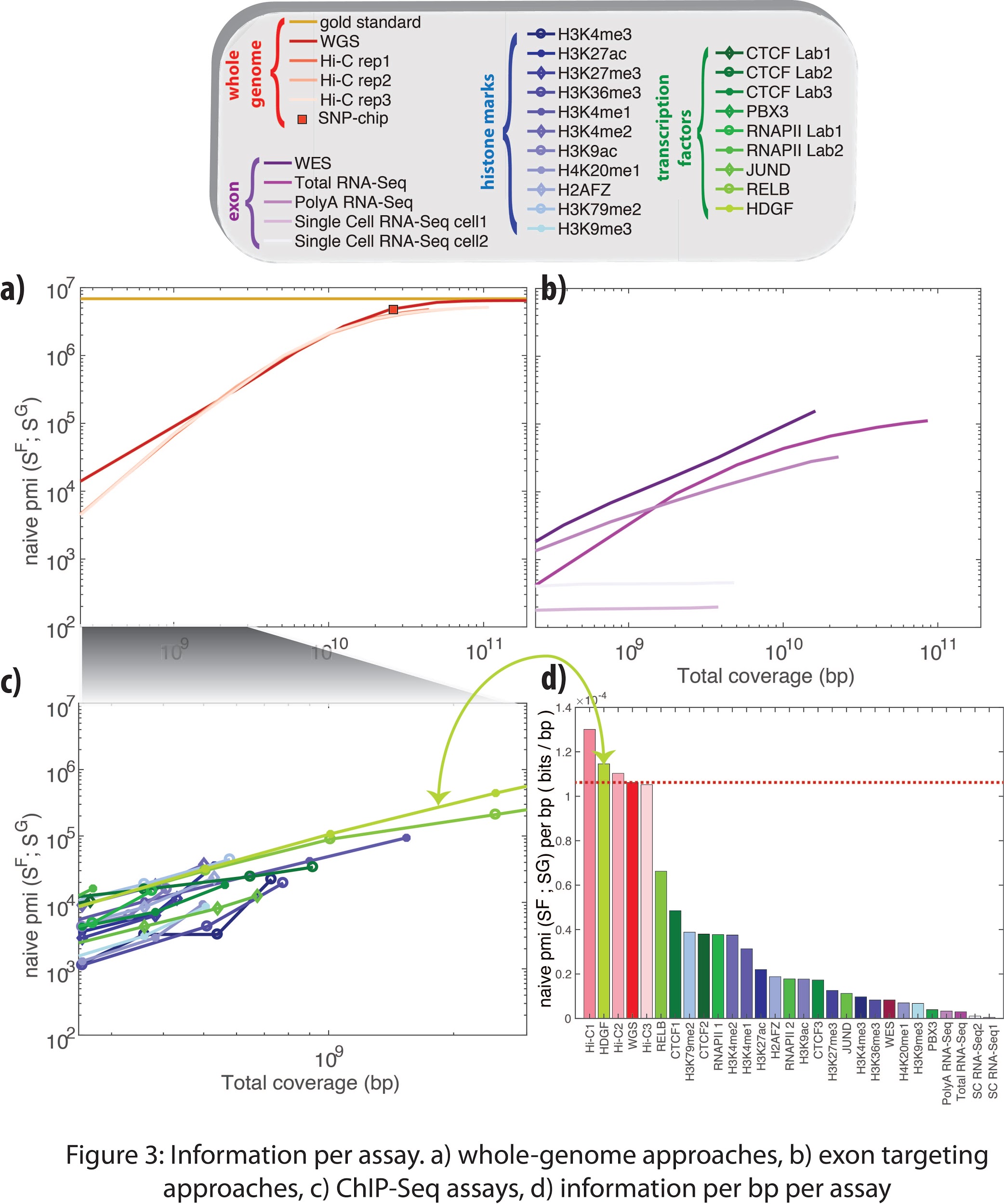
**2- Innovation (GG: 422 words, halfish page, could get rid of the commented part)**

Early privacy studies focused on genomic data that genetic variants of an individual are directly obtained and showed that private information of an individual, such as participation in a drug-abuse study, can be revealed [15,16]. With the advent of large-scale “OMICS” studies, there is a rapid increase in functional genomics datasets such as gene expression, chromatin states and conformation data. Moreover, as more relationships between these epigenetic events and human disease are discovered, the subjects of functional genomics studies are becoming increasingly real individuals. Although, the studies that focus on individuals such as GWAS has been extensively studied by privacy researchers, there is currently a significant scarcity of tools that enable analysis and protection of functional genomics datasets. In addition, there are hidden information that one can infer from any human genome sequencing based technique. For example, reads that cannot be mapped to human genome can be mapped to exogeneous sequences, which, in turn, leak information about individuals’ microbiome flora. The amount of microbiome information leakage is even exacerbated when somatic tissues such as colon or stomach are used. Somatic tissues have been at the center of the functional genomics studies recently. However, to this date, there is no such study that investigates the microbiome information in functional genomics data. Moreover, since the inception of first microbiome re-identification study [4], there is no study investigating the privacy of microbiome leakage. In this proposal, we will focus on characterization of the individuals’ sensitive information in the context of linking attacks, in which the adversary exploits the genotypes inferred from functional genomics data to link inferred noisy and partial genotypes to datasets that potentially reveal sensitive information. We will also use the unmapped reads in functional genomics data to align them on exogenous species to create a microbial code for individuals. Such code will be then used to link individuals to databases that contains relationship between microbial communities and hosts’ life styles such as diet, smoking and environments host has been to. We will build tools that enable reporting of objective measures for information leakage from the functional genomics data. These tools will enable generating uniform and systematic analysis of privacy risks imposed by releasing new functional genomics data. For generating a set of initial results that will be presented, we will use the ENCODE functional genomics dataset from Hi-C, ChIA-PET, ChIP-Seq and RNA-Seq datasets from different individuals and the genotype dataset from the 1000 Genomes Project. We will generalize the formalisms, however, to be applicable to any type of functional genomics datasets.

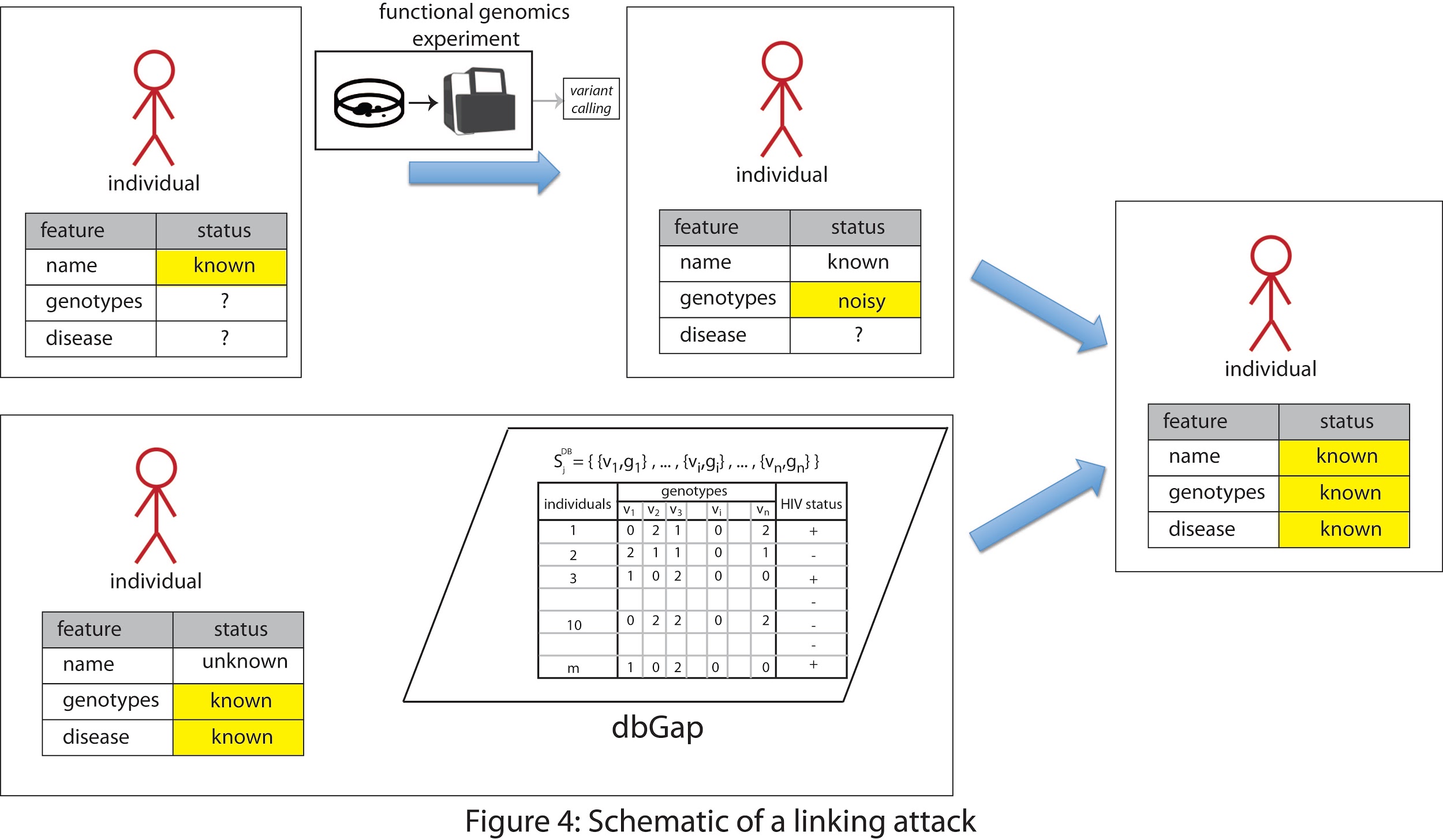
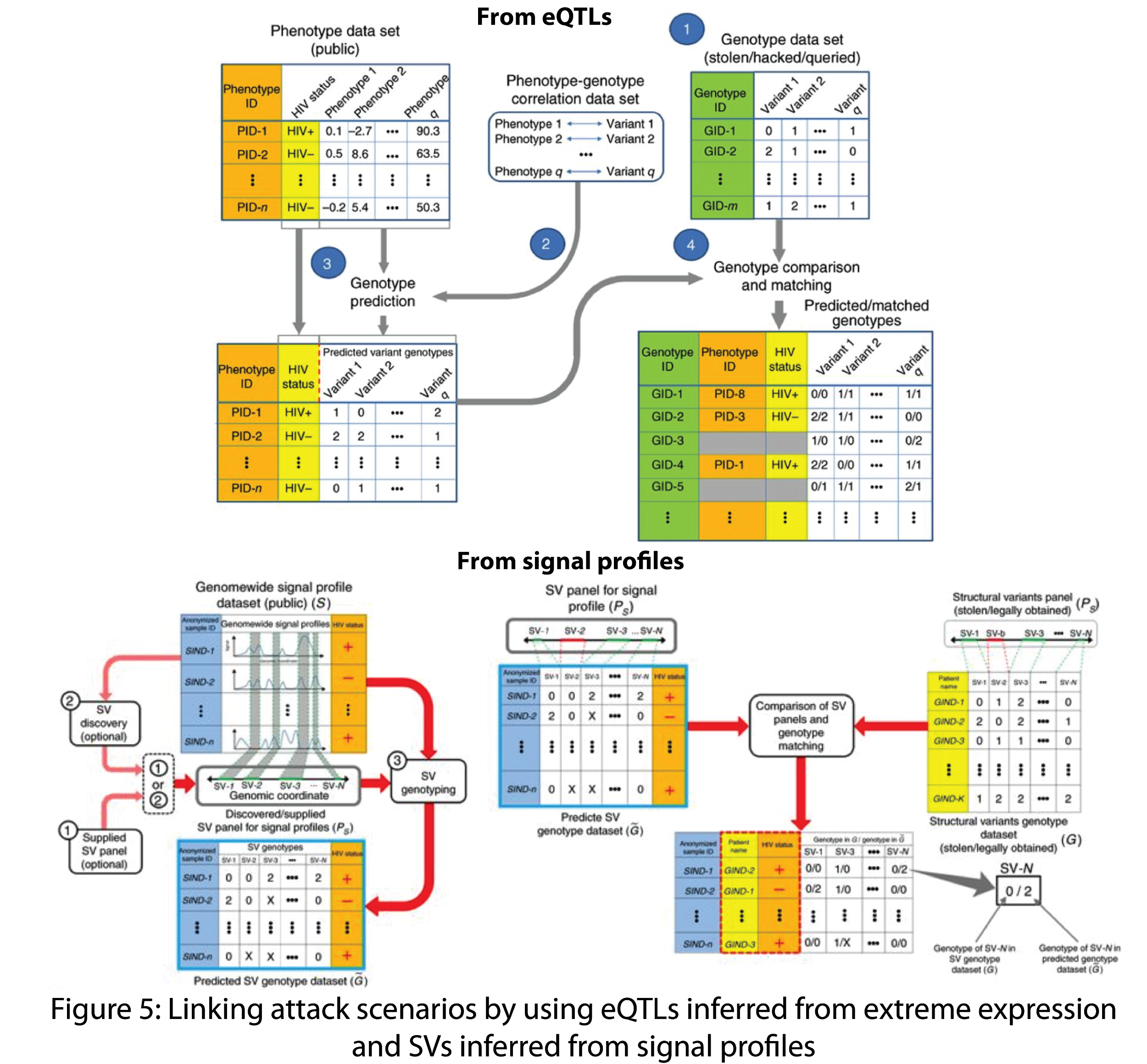
**3- Research strategy (GG: page1=477, page 2=604, page3=509, page4=564, page5=574, page6=627, page7=705)**

We will address the need for new computational approaches for analyzing sensitive information leakage within 3 aims. In this proposal, we hypothesize that to be able to characterize individuals, one does not need to obtain high-quality, accurate genotypes. We based this hypothesis to an important process in privacy literature called “re-identification”, which is combining quasi-identifiers to create personally identifying information. Quasi-identifiers are pieces of data that are not necessarily unique identifiers by themselves, but are sufficiently they can be combined with other quasi-identifiers to create a unique identifier.In the first aim, we will develop statistical formalisms for quantification of the leakage of information that enables pinpointing of individuals in genotype and phenotype datasets with use of raw functional genomics data. We will focus on specific linking attacks and work on instantiations of the linking attacks using noisy and partial genotypes as quasi-identifiers from the aligned reads of functional genomics data. In the second aim, we will focus on proposing file formats and methodologies that enable privacy preserving sharing and publishing of the functional genomics data starting from the raw aligned reads to signal profiles and gene expression levels. In the third aim, we will focus on the unmapped reads of functional genomics data and aligned them to the known microbial species to instantiate attacks on life styles or known microbiome related diseases of the patients. Figure 1 shows how the methodologies proposed in each aim can be combined for an integrated privacy software package for functional genomics data.

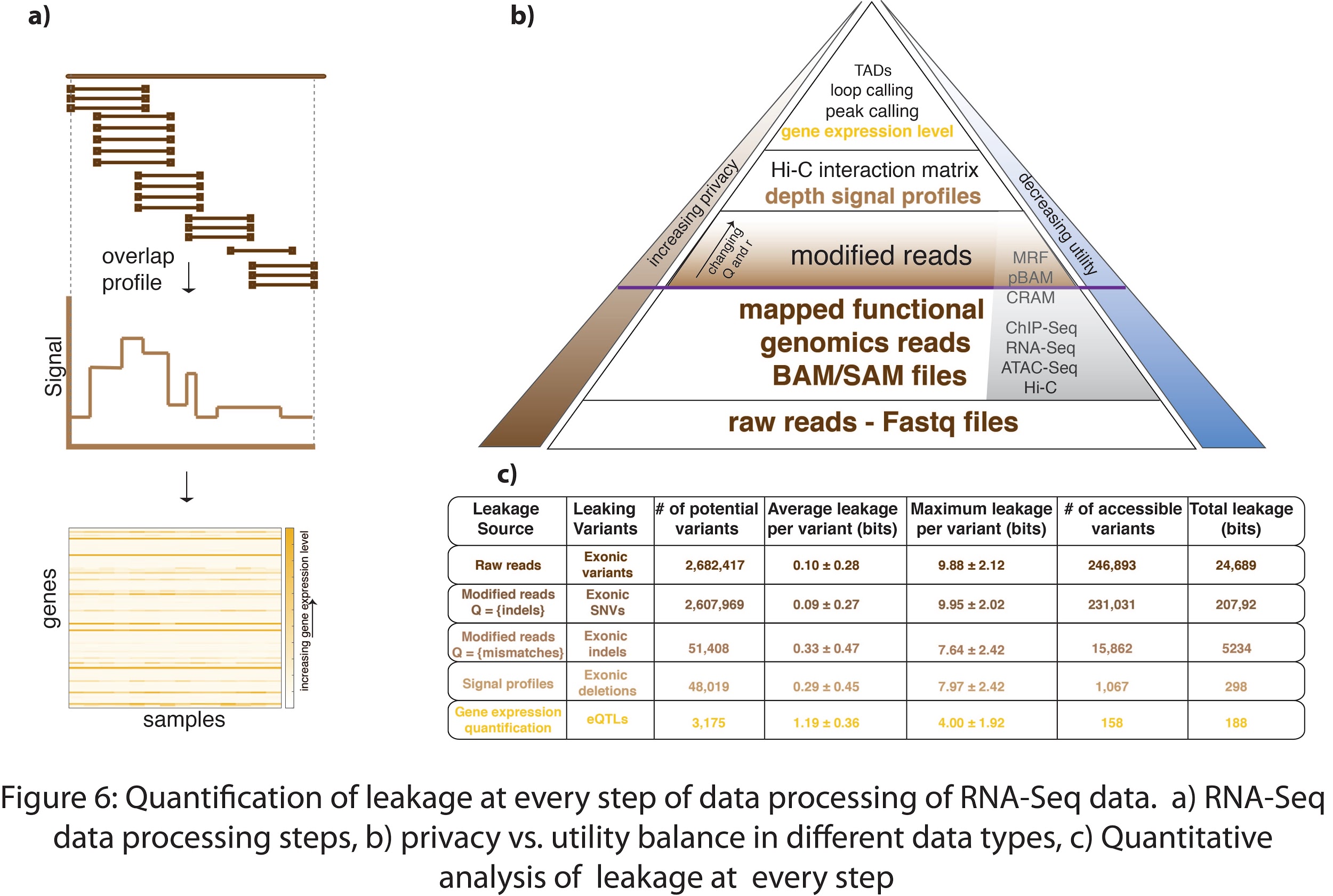
**3.1 Aim 1: Formalism for quantifying genetic leakage in the raw reads and linking attacks.**

**3.1.1 Sub-aim 1: Leakage from raw reads:** We will derive information-theoretic measures to quantify the amount of information in an individual’s functional genome. We will first represent an individual's genome as a set of genotypes. Each genotype is composed of the chromosome to which it belongs, location on that chromosome, and whether the variant is homozygous and heterozygous. Let be the set of genotypes. Then each genotype can be represented as where consists of the location and alternative allele information and is *1* for a heterozygous and *2* for a homozygous genotype. We can then calculate the self-information of *S* in bits as , where is the frequency of the genotype in a population of individuals. Next, we will quantify the information in functional genomics data in relation to sequencing coverage. We plan to use over 20 functional genomics assays from more than 10 different individuals using ENCODE data. The quantification will be take into three features that account for the sequencing differences between functional genomics data and traditional whole genome sequencing data. These three features are the depth of the coverage, breadth of the coverage and a bias term we introduce as the deviation of the depth distribution of the functional genomics sequencing from the Poisson distribution expected from whole-genome sequencing. To capture the amount of leakage with respect to amount of sequencing, we will sample reads from the sequencing data of functional genomics experiments at increasing coverage. We will then create a pipeline based on GATK best practices to call variants from DNA and RNA sequences. The procedure is as follows. We will sample *c* amount of total nucleotides from different functional genomic experiments taken from different individuals and from whole genome sequencing (WGS), whole exome sequencing (WES) and genotyping array data as well if available. We will use GATK to call SNVs and indels with the parameters and filtering suggested in the GATK best practices [17,18]. We will use the genotypes in the 1000 Genomes panel for these individuals as the gold standard. We will use pointwise mutual information (pmi) as a measure to quantify the association between the gold standard and the called variants (Figure X). If s the set of genotypes from the gold standard and is the set of genotypes called from the *c* total sequencing coverage of a functional genomics experiment, then the set contains the genotypes that are called and are in the gold standard set. If , then .

Our preliminary analysis using the individual NA12878 captured the differences between the high depth experiments such as Hi-C and low depth experiments such as ChIP-Seq histone modification. Figure 3 shows how much sequencing is needed to achieve WGS level of genotyping. A similar formalism can be constructed based on the SVs from the raw functional genomics sequencing. To do so, we will first call deletions from the sampled reads using our previously developed software CNVnator [19]. Current softwares for SV calling are based on the depth differences in the sequencing with the assumption that whole genome sequencing data samples the genome uniformly. However, the sequences from the functional genomics data are targeted to regions of interest, hence the depth differences are due to the functional activity of the genome. Therefore, SV calling using current pipelines directly from functional genomics data will result in a great amount of false positives. We will take two different approaches to show the degree to which functional genomics raw sequences can leak SVs. (1) We will use CNVnator as is and use the raw sequences from functional genomics data as is to determine if SVs enriched with false positive that are called from the data are adequate enough to characterize individuals. This will highlight the importance of quasi-identifiers in making inferences. (2) We will create additional filtering in our pipeline such that it will assign a confidence level to the called deletions by assessing the sequencing depth of the neighboring basepairs. This will allow us to determine the regions of the genome that are low depth due to SVs rather than lack of function. We will use the high confidence SV calls to perform the linking attacks, which is explained in the next section.

**3.1.2 Sub-aim 2: Linking attacks using genotyping from raw functional genomics sequences:** Linking attacks aim to re-identify an individual by cross-referencing datasets. For example, in a hypothetical scenario an attacker aims to query an individual’s HIV status from his/her genotype data. This genotype data is released with anonymized identifier for each individual. We assume that the adversary obtains access to this dataset by either lawful or unlawful means. Now let’s assume that the attacker has access to a biosample. This could be partial or complete mapped reads from functional genomics experiments or a saliva sample taken from a used glass. The idea is to genotype the biosample and find the matching genotypes in the HIV status database. However, individuals share many common genotypes with each other. The number of shared genotypes between individuals is large within a population and even larger within a family. The question becomes how well an adversary should sequence an individual’s genome to be able to perform successful linking. Specifically, if an adversary obtains permission to only perform functional genomics experiment on a biosample, then can the adversary use the reads from these experiments to link the individual to a genotype-phenotype panel? For this, the attacker calls SNVs/SVs directly from the reads of functional genomics experiments. Then he/she compares the called noisy and incomplete genotypes to the genotype data panel and finds the entry with the highest match. This reveals the sensitive information for the linked individual to the attacker. We then consider a scenario in which the attacker has access to partial or increasing amount of reads to find out how much sequencing is enough to successfully perform linking attacks. Figure 5 shows a hypothetical linking scenario. In this aim, we will derive formalism for the genotypes called from raw functional genomics sequences and perform various linking attack to show the degree of leakage. Our preliminary results showed that with the design of various different linking attack scenarios [11,12], we are able to link individuals to databases. To do so, we first inferred the correlated eQTLs with the extreme gene expression values and used them to genotype individuals. We showed that these genotypes yield more than 90% linking accuracy when applied to gEUVADIS dataset. We also showed that an adversary can detect small and large genomic deletions in signal profiles as well with high predictability. The breadth of coverage and depth of coverage, which are inferred from the signal profiles can be used to genotype deletions. Our analysis on ChIP-Seq and RNA-Seq profiles showed that RNA-seq is concentrated on exonic regions and has high depth of coverage but low breadth of coverage therefore can used to genotype small deletions, while ChIP-Seq signal profiles generally have high breadth of coverage but low depth of coverage and can be used for detecting large deletions.

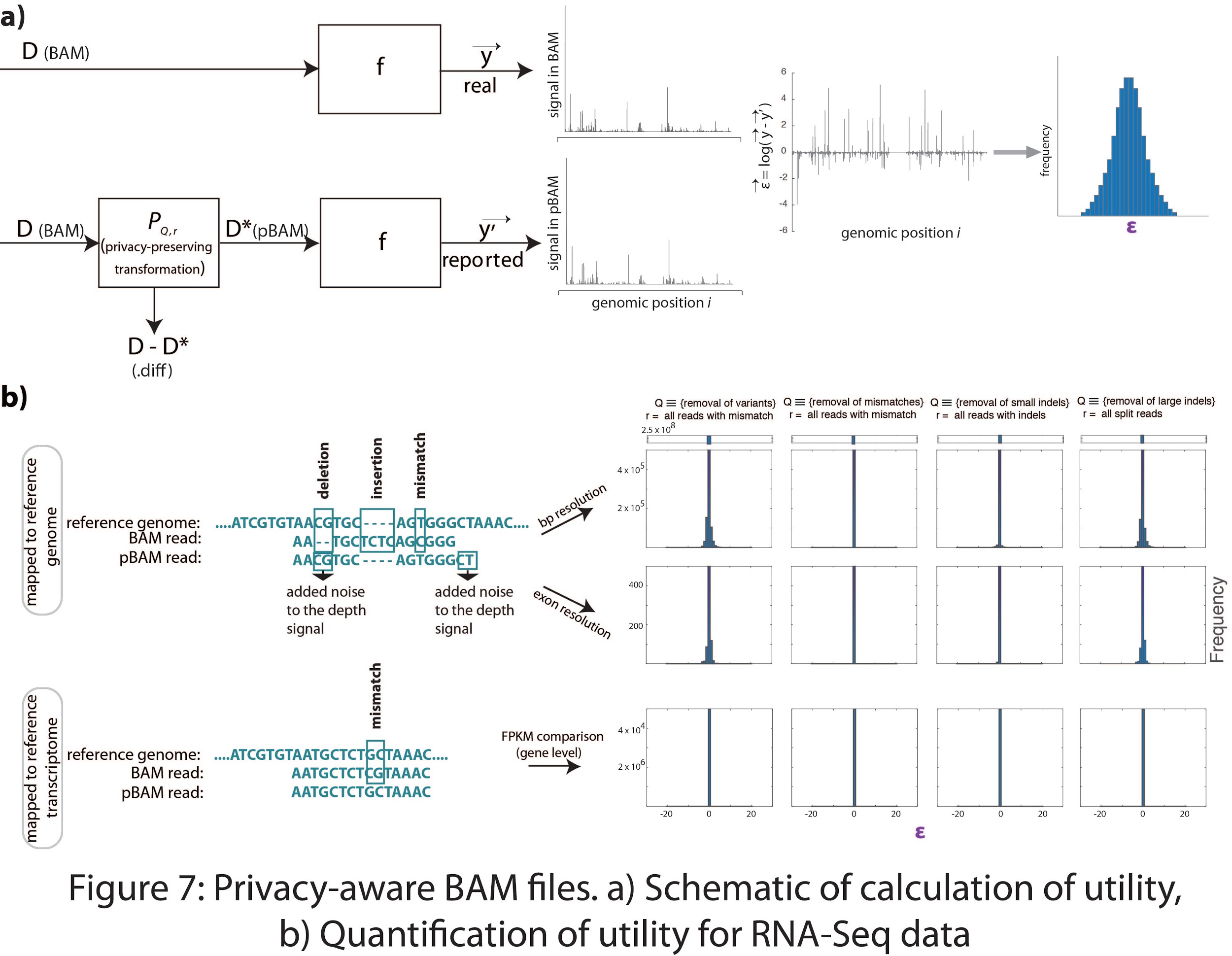
**3.1.3 Sub-aim 3: Linking attacks using genotyping from Hi-C interaction matrices**: Much effort has been devoted to capturing copy number variations and large structural variants of cancer cell lines using Hi-C contact matrices [20, 21, 22]. Our previous work showed that we can collapse the Hi-C interaction matrices into one dimension to reveal genomic deletions in germline. In this sub-aim, we aim to develop methods that can infer (1) germline copy number variations (2) large SVs such as translocations from the interaction matrices. To do (1), we will create signal profiles from every row of the Hi-C interaction matrix, also known as virtual 4C plots [23]. These plots represent the interactions between the selected row and the rest of the genome. Highest peak in these plots correspond to the diagonal in the Hi-C matrices, which represent the proximity between the selected row and its immediate neighbors due to the chain connectivity in the polymer properties. In an ideal experimental setup, these peak are expected to be the same height for every row, i.e., one expects to see proximity ligation happening in every cell of the population. Large variations between these peaks can be attributed to copy number variations in the genome. For example, if a region in the genome (that corresponds to a row) is duplicated, then we can expect to sequence that region twice more than others, hence it will double the value in the diagonal. Same logic will apply to deletions in the genome. To do (2), we will base our calculations on the fact that the expected interactions between chromosomes are less compared to the expected intra-chromosomal interactions. If regions in different chromosomes are interacting with high frequency, this is likely due to translocations between these chromosomes. Both (1) and (2) will not generate accurate SV calls but will be enough to use as quasi-identifiers to do linking attacks.

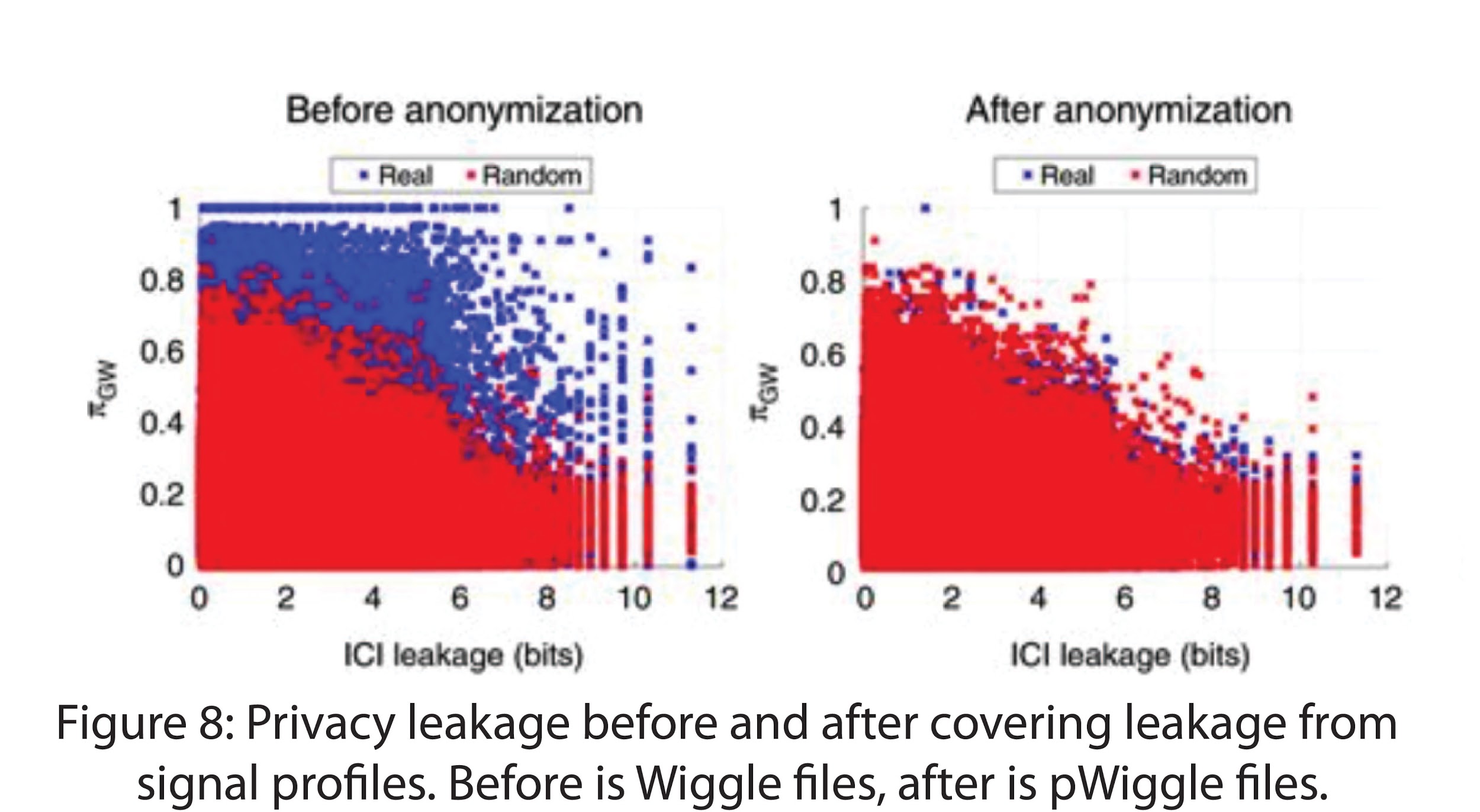
**3.1.4 Sub-aim 4: Linking attacks using genotyping from gene expression levels using sQTLs and isoQTLs:** Based on the success of re-identification through expression-eQTL relationship [11], we will apply the same formalism to sQTL-transcript abundance and isoQTL-isoform expression relationship. By using publicly available sQTL, isoQTL, transcript abundance and isoform expression data, we will create a relationship between the transcript expression and sQTLs as well as isoform expression and isoQTLs. This will allow finding the distribution of transcript expression level given an sQTL and the isoform expression level given an isoQTL. Given these expression levels from a population of individuals, we will guess which individuals have which sQTLs or isoQTLS using the conditional probability distribution ofand . We will then use inferred sQTLs and isoQTLs as quasi-identifiers to link these individuals to genotype-phenotype databases. 

As a preliminary results to aim 1, in Figure x, we calculated the potential number of variants one can obtain from a typical RNA-Seq experiment: (1) At the read level, we can potentially observe all the SNVs on the exons; however, only a fraction is accessible through RNA-Seq depending on which gene is expressed in which cell line/type. (2) At the signal profile level, we can potentially observe all the deletions on the exons. However, only a fraction is accessible to the experiment depending on the expressed transcripts in the given cell type/line. Therefore, following the genotyping described in ref. [12], we calculated the number of deletions that can be genotyped from the signal profile of the polyA RNA-Seq experiment of the individual NA12878 as the accessible deletions in Figure~\ref{stack}c. (3) At the gene expression quantification level, the potential number of variants that can be observed are all the eQTLs connected to the genes. We calculated the accessible variants through gene expression quantification as the average number of eQTLs per individual based on the calculations in ref. [11].

**3.2 Aim 2: Privacy-preserving file formats for raw and processed functional genomics data.**

In this aim, we plan to create privacy-aware file formats for every step of data processing steps of functional genomics data. We will have two measures to evaluate our file format systems: privacy and utility. For privacy, we will use the measures we defined in aim 1. For utility, we will compute the functional genome from original files and privacy-aware files and find the dissimilarity. This will be different for each data type. Details are as following.

**3.2.1 Sub-Aim 1:** **Privacy-Aware Binary Alignment Files (pBAM):** We aim to base our file format to the existing file format system SAM/BAM. This will allow users to use the existing functional genomics pipeline using pBAMs as input. We will apply a privacy-preserving transformation to the alignment files such that calling variants from transformed files is largely prevented while quantifications related to the functional genome is possible with minimal error. We can think of a raw alignment file~(BAM) as a dataset, where information for each read is contained. Let's 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 in the meantime any calculation *f* based on *D* and *D\** retrieves almost the same result. We will use two general methods to achieve this: suppression and generalization. If every column in *D* is an attribute (such as read length, cigar, sequence, or quality value), then replacing an attribute with an asterisk (\*) is suppression and changing an attribute with a more general value is generalization. For example, in our file format transformation, we can replace sequence and sequence quality attributes with an asterisk (suppression), and transform the cigar of the read from partially mapped to fully mapped (generalization). Now let's say the privacy-preserving transformation is done through a function such that . *Q* is an operation such as “removal of small indels”, “removal of mismatches”, “removal of large indels”, or “removal of all variants”. *r* is the amount of reads to be manipulated given the operation *Q*. A calculation *f* can be a signal depth profile calculation, TF binding peak detection, or gene expression quantification. Then, we can reconstruct an equation for each unit *i* as , where a unit *i* can be a single base pair, an exon, or a gene depending on the function *f*. In turn, can be calculated as the log-fold change between the results derived from two datasets. This is also a quantity commonly used to compute differential gene expression [24] or ChIP-Seq binding enrichment over controls [25], and can be used analogously in this context, where log-fold change is the differential signal depth or expression when the manipulated data is used as an input. Note that || is a measure of error of the new dataset *D\**. We can then calculate the distribution of || values over every unit and found the mean || per unit as the overall error. The level of privacy is controlled by the function , where *Q* determines the removal of type of entries and *r* determines the number of entries of the given operation *Q* that are manipulated. For any particular operation, the obvious threshold could be the size of the indels, minor allele frequency (MAF), or the depth of a particular unit. These thresholds can be converted into the fraction of reads affected. For example, if *Q* is the removal of indels and *r* is the reads that contain indels with MAF<0.01, then only reads that have indels with MAF<0.01 will be manipulated in the transformed *D\**. We will construct the privatized file format pBAM from data *D\** as follows. The reads from the BAM files will be categorized as perfectly mapped reads and reads with mismatches, insertions, deletions, or soft- and hard-clipping. will replace the sequence and the quality string of all of reads with an asterisk and manipulates the cigars, alignment scores (AS tag) and the strings for mismatching positions (MD tags) of the reads that are defined in *Q* and *r*. We will also create pBAM files from BAM files that are obtained by mapping sequences to the transcriptome coordinates, which is essential for gene quantification. This transformation function will be general and can be applied to any alignment file type such as SAM, CRAM, and MRF to create a new privatized file format. These files will be concordant to use with tools such as samtools, cramtools, and mrftools. Suppression of sequences and quality strings can be done to all reads regardless of *Q* and *r*, since these features can be recovered using reference genome, cigar string and MD:Z tags if needed. This way an adversary cannot tell which reads are manipulated in the data. This also allows us to create much smaller alignment files. Our other approach to cover the unforeseen leakages in cases where we manipulate only a subset of reads is to replace the sequences with the corresponding sequences in the reference genome. This will result in larger pBAM files. Our preliminary analysis showed that pBAMs provide high utility while protecting privacy in basepair, exon and gene expression level when applied to RNA-Seq data (Figure x).

**3.2.2 Sub-Aim 2:** **Privacy-Aware Signal Track Files (pWiggle):** The most effective way to protect against a linking attack scenario is to ensure that deletion genotypes cannot be inferred from signal profiles. As we showed earlier [12], small deletions are a major source of leakage of genetic information from RNA-Seq signal profiles. We propose systematically removing the dips in signal profiles as a way to anonymize the profiles against the prediction of small deletions. To remove these dips systematically, we propose using median filtering based signal processing to locally smooth the signal profile around the deletion. This signal processing technique has been used to remove shot noise in two-dimensional imaging data and one-dimensional audio signals. For each genomic region *a* in the deletion [*i, j*], we will replace the signal level using the median filtered signal level as , where refers to the signal level at the genomic position *a*, *l* = *j* – 1 + 1, refers to the smoothed signal level at position *a*, and median refers to the median of all the signal values in the genomic region [𝑎−𝑙/2, 𝑎+𝑙/2]. The median will be computed by sorting all the signal levels and choosing the value in the middle of the sorted list of signal levels. We will then use the utility metric we defined in section 3.2.1 to measure the utility of pWiggle files. This will be done by calculating the gene expression levels from the pWiggle and Wiggle files and calculate the log-fold difference both in exon and gene level. Our preliminary analysis showed that after anonymization of signal profiles with this method, the information leakage is largely covered (Figure x).

**3.2.3 Sub-Aim 3:** **Privacy-Aware Hi-C contact matrices (pHiC):** We aim to cover the leakages in the Hi-C interaction matrices, while keeping the overall properties of the matrix the same. To do so, we will apply two different transformation to matrices. First one will correct for the copy number variation that can be inferred from the elevated interactions seen in the diagonal and the second one will correct for the translocations. (1) We will first determine the distribution of the diagonal values of the matrix. Any outlier in this distribution will correspond to the copy number variation. For these outliers, we will sample values based on (a) the original value, (b) mean of the distribution, (b) the standard deviation of the distribution. For example, if the original value is in the lower end of the distribution, the sampled value will be from the left hand side of the distribution, while it is not more than 3 standard deviation away from the mean. (2) We will apply another sampling scheme for the translocations. For this, we will find unexpectedly high inter-chromosomal interactions. We will again determine the distribution of the inter-chromosomal interactions. We will truncate the distribution such that only mean and 3 standard deviation away from the mean is included. We will replace the values of translocations partner with the values we sampled from this truncated distribution. The privacy of this new matrix will be determined by applying the SV calling described in section **3.1.3** followed by a linking attack. It is difficult to quantify the true utility of the new file formats, as there are myriad of softwares that call significant interactions, topologically associated domains (TADs) or higher order organization of the chromatin and the results from different softwares does not necessarily overlap with each other. Therefore, we will determine the utility of the pHiC files in a practical way by performing a number of loop calling algorithms (juicer [26], fit Hi-C [27]) and a number of TAD calling algorithms (mrTADfinder [28], TADtreefinder?). We will calculate the differences between loop anchors and TAD boundaries called from original matrix and called from pHiC files. We will also perform a similarity matrix using our reproducibility software Hi-CSpector [29] to determine the utility. For this, we will generate a random Hi-C matrix, a pHi-C matrix from the cell type A and a pHi-C matrix from cell type B. We will calculate the HiCSpecter score for the similarity matrix between Hi-C of cell type A, pHi-C of cell type A, Hi-C of cell type B, pHi-C of cell type B and the random matrix.

**3.2.4 Sub-Aim 4:** **Privacy-Aware gene-expression quantification matrix (pExpQuant):** Our previous work showed that there are leakages when the gene expression level from multiple individuals are shared [12]. We showed that we can use the extreme expression levels of a gene to link them to the publicly available eQTL data. Once we discover the potential eQTL an individual has, we can link these eQTLs (noisy quasi-identifiers) to a phenotype-genotype panel to infer private information about the individual [12]. In this sub-aim, we plan to prevent leakages by manipulating the gene expression matrices from multiple individuals. We will use k-anonymization [30], which will enable one to anonymize the matrix by ensuring none of the extreme gene expression values (hence the predicted genotypes through eQTLs) can be used to match a record in the dataset to fewer than k individuals. We will select entries from the dataset based on the genotype predictability and private information leakage to determine the leaky gene expression values. We will then apply a generalization technique by replacing these values with a value sampled from the distribution of gene expression values for that gene, such that sampled values will not exceed mean -/+ 3 standard deviation. This maximizes the utility of the anonymized data set by allowing one to focus only on the data points that leak the most characterizing information. In addition, because the anonymization process can focus only on the sources of highest leakage, this approach cuts down on computing requirements and increases the efficiency of anonymization. Same anonymization process will be applied to isoform expression and transcript abundance data to cover leakages from sQTLs and isoQTLs.

**Aim 3: Formalism to quantify the amount of leakage in human functional genome sequencing through microbiome.**

**[[GG notes:**

**Microbiome from RNA-Seq (mostly gtex) papers:**

[**https://www.nature.com/articles/s41398-018-0107-9**](https://www.nature.com/articles/s41398-018-0107-9)

[**https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1403-7**](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1403-7)

**Microbiome contaminants in tissues (list of bacteria to remove from the output because they are identified to be lab contamination:**

[**https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187044**](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187044)

**]]**

Here we hypothesize that when functional genome is sequenced, there are by-products, which are the reads from exogenous species that reside in tissues. These exogenous reads can be used to construct a microbiome code for individuals and used for re-identification purposes.

**3.3.2 Sub-aim 1:**- Building a RNA-Seq pipeline for the identification of potential exogenous (bacterial) RNAs - 2 pages [JR]

Here we present here a RNA-Seq pipeline for the identification of exogenous (including bacterial) RNA sequences that are present in a human derived RNA-Seq sample. Identification of exogenous sequences present in human cellular RNA-Seq profiling data faces some unique challenges. Computational methods for identification of exogenous sequences are much more vulnerable to experimental laboratory contamination and artifacts, in large part due to their relative low abundance. The variable presence of rRNA; deterministic cleavage of structured smallRNA (tRNAs and piRNAs) and longer RNA molecules; and imperfect annotation of miRNAs, piRNAs, and tRNAs all pose challenges for quantification and functional interpretation. For these reasons, existing computational tools capable of analysing RNA-seq data are not as well suited to the new field of exRNA analysis.

By providing an optimized and standardized bioinformatics platform, the pipeline will reduce technical bias will in a set of cascading steps will confidently remove all possible human derived sequences before alignment against databases of exogenous ribosomal RNA sequences and other RNAs from a comprehensive set of non-human genomes available from NCBI. In order to conservatively and relatively confidently identify exogenous RNA sequences present in a human sample the pipeline will first remove possible contaminant laboratory sequences followed by a liberal mapping to the host human genome in order to minimize possible human derived sequences being misidentified as exogenous sequences.

[RJ1]Need to be careful – people get picky about the distinction between EV and exosomes

[RJ2]Add ref to freedman paper in Nat Comm that we were on

Filtering and Quantification Cascade:

We plan on developing a pipeline that is composed of a cascade of serial computational filters and alignments, whose order is designed to reflect an appropriate level of confidence in the various endogenous and exogenous sequences that may exist in the extracellular fraction (Figure XX). To combat potential contamination in an RNA-seq library, reads aligned to known contaminants are removed before mapping to the host genome and to the exogenous sequences. The pipeline is highly modular (constructed as a makefile file containing shell, Java, and R modules), allowing the user to define the order of which RNA annotations that are used during read-mapping; it includes support for random-barcoded libraries and spike-in sequences for calibration. The general workflow comprises steps for preprocessing, endogenous alignment, and exogenous alignment (Figure XX).

Preprocessing:

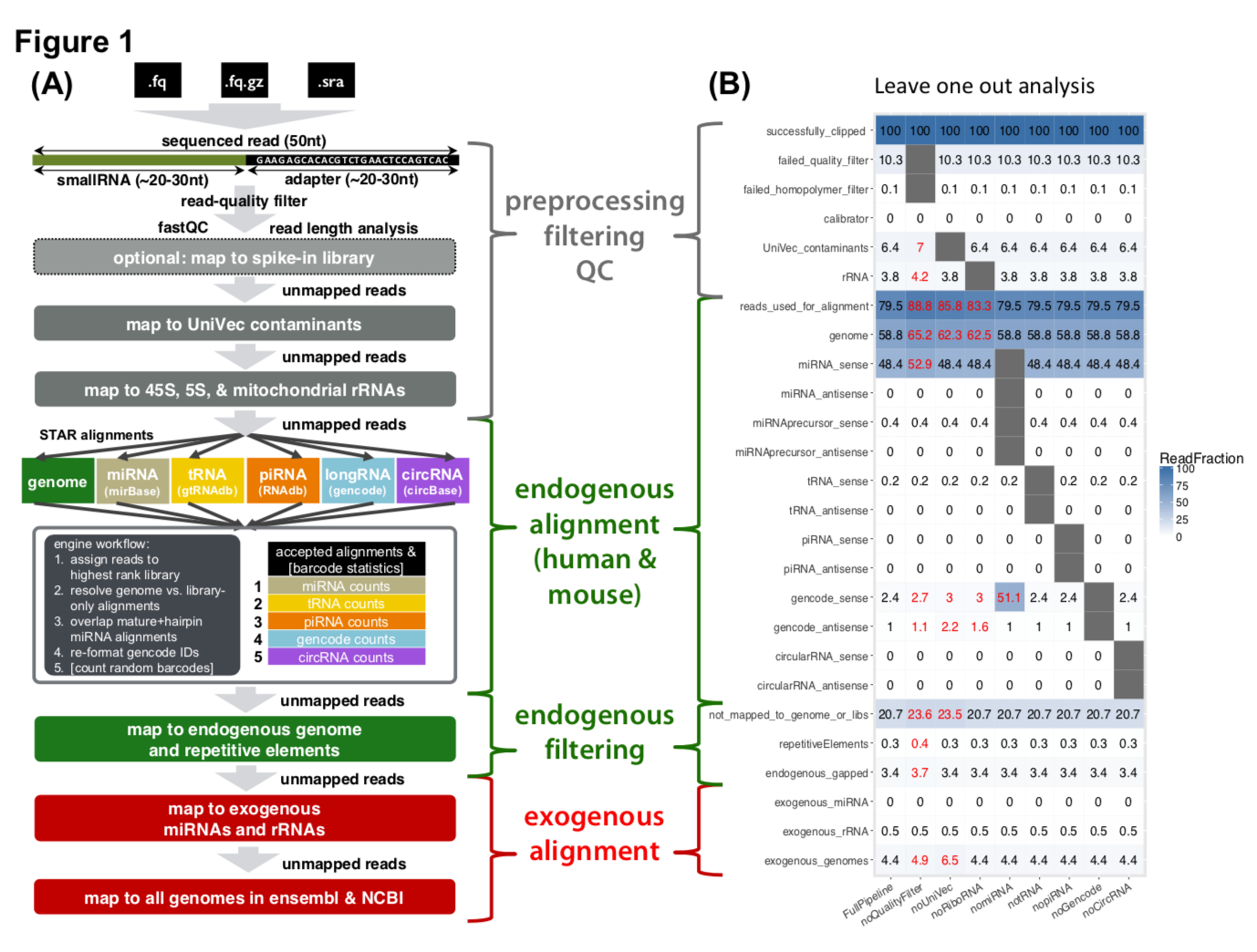
First, the pipeline will automatically identifies and removes 3’ adapter sequences. If specified, it will also remove and store information from random-barcoded insert sequences, which are increasingly being used in RNA sequencing in an attempt to identify and compensate for ligation and/or amplification artifacts that have the potential to affect downstream quantification (Fu et al., 2014). If spike-in sequences were added to the library, the pipeline will align adapter- and barcode-clipped reads against either standard NIST/ERCC sequences or sets of user-specified sequences. RNA-seq reads of consistently low-quality reads are removed, as are reads aligned to sequences in the NCBI’s UniVec/Vecscreen database in order to filter out common laboratory contaminants. Finally, reads aligned to all primary endogenous ribosomal RNAs (5S, 5.8S, 18S, 28S, and 45S), many of which are highly variable in abundance in RNA library preparations.

Endogenous Alignment:

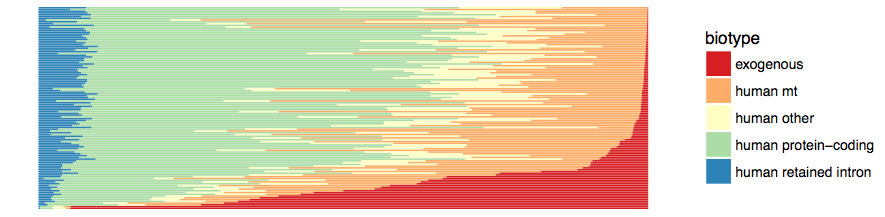
RNA-seq reads are aligned to the host genome and transcriptome of human or mouse, and transcript abundances are calculated (RNAs are quantified using both raw read counts and normalized reads per million (RPM)). Because of the variety of RNA preparations available (totalRNA, smallRNA, miRNA), the pipeline will allow the user to prioritize the order in which annotations (miRBase, tRNAscan, piRNA, GENCODE, circRNA) are used for quantification based on our confidence in the presence of a given annotation in a given sample. For example, reads from a miRNA-seq prep can be assigned to miRBase miRNA annotations before piRNA annotations. Likewise, reads from long- or total-RNA preparations can be assigned to longer GENCODE transcripts before (or instead of) the other RNA libraries. This feature is particularly relevant for lower-confidence RNA annotations; piRNAs, for example, are generally given lower priority than tRNAs to ensure correct read assignments.

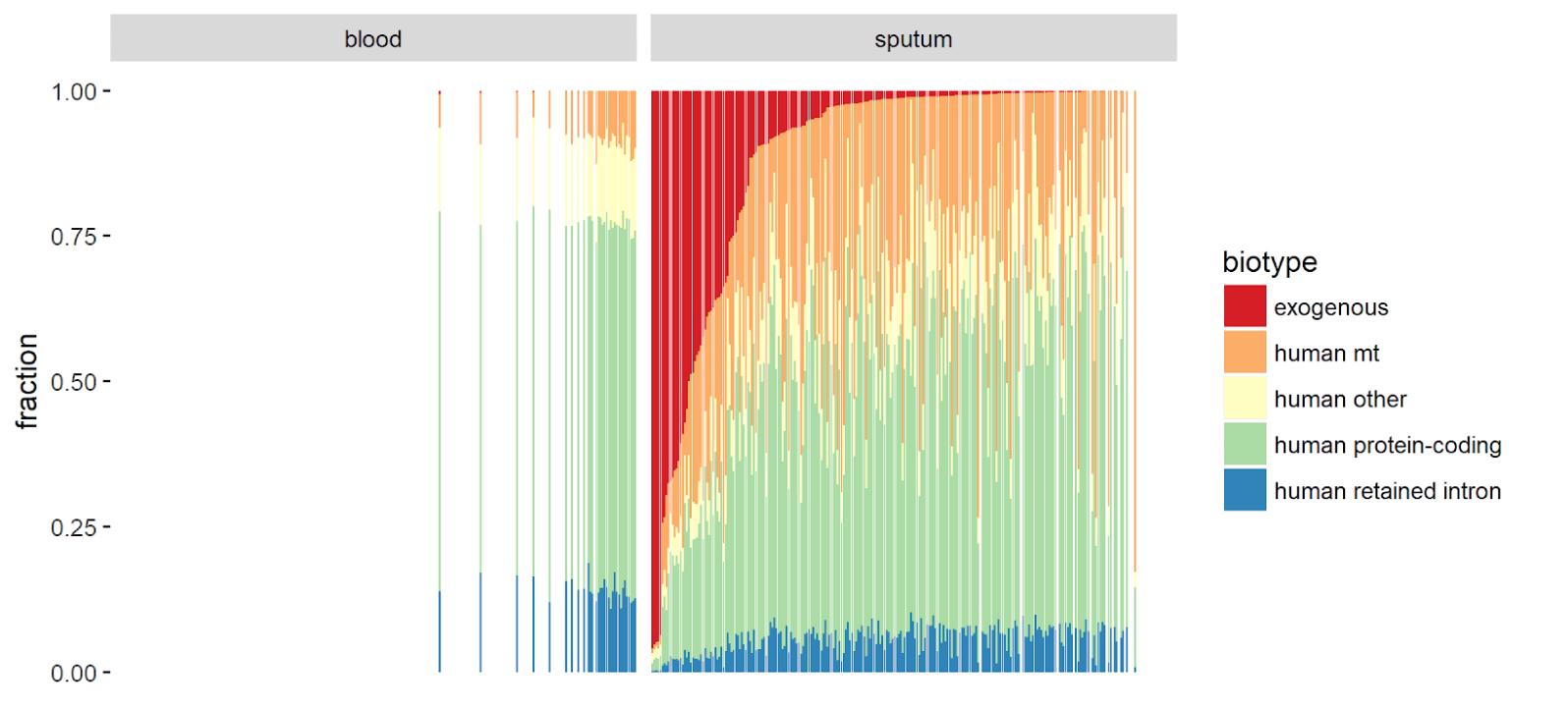
Exogenous Alignment:

The RNA-Seq pipeline will be designed to enable confident assessment of non-human sequences in biofluids after careful, explicit removal of as many known or likely contaminants as possible. Before interrogating non-contaminant and non-endogenous reads for potential exogenous sequences, the pipeline will include a second-pass alignment against the host genome and known repetitive sequences using significantly relaxed mapping criteria. This step serves to remove additional sequences that most likely derive from the host genome and will make the pipeline conservative in the identification of exogenous sequences. Reads are then aligned to curated libraries exogenous rRNA sequences in the Ribosomal Database Project (RDP), followed by alignment to the full genomes of all sequenced bacteria, viruses, plants, fungi, protists, metazoa, and selected vertebrates that are potentially part of the host diet. Existing approaches for exogenous sequence alignment remove degenerate sequences (i.e. those that co-occur across multiple species), which results in a loss of potentially valuable data as reads frequently align across multiple species/strains. By characterizing exogenous genome alignments generated by the pipeline in terms of the NCBI taxonomy tree and assigning reads to the most specific possible node in the phylogenetic tree (many reads can only be assigned to nodes higher up in the phylogenetic tree due to not uniquely mapping to a specific genome of a sub-species), users may obtain valuable information regarding the contribution of the flora to various RNA-Seq samples and can generate phylogenies for cross-sample comparison. For example, in Figures XX and XX we present the phylogenetic trees of the reads that we assign to bacterial ribosomal and genome sequences for a specific saliva sample. In both trees, we find an abundance of reads assigned to the node corresponding to the genus *Streptococcus.* We have a high degree of confidence in these results given that the sets of short reads used for constructing these two trees are disjoint.



2- show preliminary run bacterial run - figures from DS&SKL paper [SKL] - paragraph





Sputum from 115 patients with heterogenous asthma phenotypes was collected and sequenced by RNAseq. The median read depth per sample was 47.5 million, which meets recommendations for dataset and analyses of this type. These reads were processed through the exceRpt pipeline, which first aligns to human reference and then conservatively matches to exogenous reference databases such as rRNA databases and complete genomes. The percent of reads mapping to different biotypes was highly heterogeneous; a median of 60% of the reads aligned to the human reference genome and 50% to annotated transcripts (green bars), which is consistent with other RNA sequencing efforts on samples of this type. A median of 0.7% of the input reads aligned to exogenous sources, with some samples containing as much as 28.1% exogenous reads.

entex [LS&GG&FN] - paragraph

To examine whether we can identify exogenous RNA using sequencing and alignment techniques, we applied our pipeline on 4 different samples from aorta, stomach, colon, and skin samples from the same individual. While *bacillus subtillis* was the most prevalent across all tissues, we were able to identify clear qualitative and quantitative differences between the 4 samples. As expected, the aorta tissue returned the lowest number of unmapped reads, number of detected species, and exogenous ribosomal RNA. In the stomach sample, were able to detect sequences of *Gallus gallus* (chicken), *Sus scrofa* (wild boar), *Triticum aestivum* (common wheat), *Hordeum vulgare* (Barley), *Meleagris gallopavo* (Wild Turkey), and *Bos Taurus* (Cattle) in the top 15 most prevalent organisms. These results suggest that we are detecting true exogenous RNA signals, while the content can also reveal dietary and lifestyle habits. Similarly, we were able to detect a high prevalence of different *Bacteroides* *sp*. and *E. coli* in the colon, both strongly associated with gastrointestinal microbiota; again indicating true exogenous signals. Finally, skin tissue contained a diverse mixture of organisms, including bacteria, cattle, and barley.

[[GG Update 09/18

Andrew said he is fine with it

He liked the control experiment

He said it could use a bit editing - didn’t say anything specific]]

**3.3.2 Sub-aim 2: Control human genome sequencing experiments to determine the reproducibility and variance of exogenous reads**

Unlike human genome variation data, there is no established measures to quantify the expected amount of reads belong to exogenous species from the standard human genome sequencing data. In addition, there is no control to determine to what extent these reads are reproducible when the experiments are replicated.For this, we plan to collect blood samples from 10 consented individuals. Each sample will be divided into total of 10 biological replicas. The aim is to determine the amount of microbiome we can infer from the standard sequencing of blood tissues and to determine the variation and reproducibility of the inferred microbiome.

We will device two independent experiments with 5 replicas for each sample. (1) We will perform standard DNA extraction, whole genome amplification followed by Illumina sequencing on 5 replicas of 10 samples. The sequences will then be mapped to human reference genome. Unmapped reads will be processed with the pipeline described in Sub Aim 1. The exogenous sequences that are reproduced in all of the replicas will set the baseline for the amount of reads that is required to quantify the microbiome from human genome sequencing data. We filter any species that are not supported by this baseline amount of reads. (2) We will mix a few exogenous species with known amounts and contaminate the samples by adding this mixture to the blood samples. We will then perform standard DNA extraction, whole genome amplification followed by Illumina sequencing on 5 replicas of 10 samples. The sequences will then be mapped to human reference genome. Unmapped reads will be processed with the pipeline described in Sub Aim 1. The expected microbial communities that are observed in (1) will be subtracted from the reads that are mapped to exogenous species. The remaining fraction of the reads of exogenous species will then be compared to the original amount of the bacterial DNA that was in the mixture. The difference between the original and detected amounts as well as the variation of the difference will be regarded as the baseline for the technical reproducibility and variation of capturing microbiome from human sequencing data.

3- simple ideas for deconvolution - 2 ideas

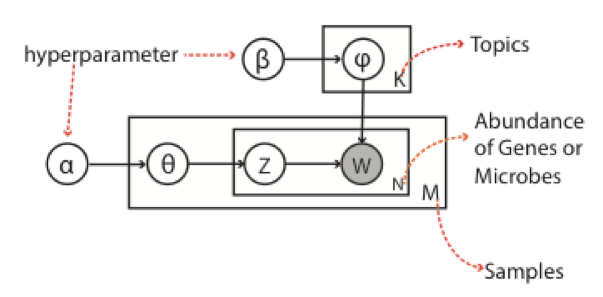
SKL - LDA method 1 or 2 paragraph , deconvolving the microbiome reads relate to phenotype

The raw read counts we get from RNA-seq is high-dimensional data with noisy, especially for microbes. In this grant, we are going to develop a method to infer a low dimensional representation or pattern of microbe abundance from RNA-Seq samples, and then we correlate these patterns of different occurrences and different microbes with various environmental and phenotypic conditions. The linkages between the microbes and environmental conditions might be, for instance, exposure to dusts with allergens, people may be sick . After develop, we will use latent Dirichlet allocation, or LDA, for dimension reduction and feature extraction.

MEG edit: [[[

The raw read counts we get from RNA-sequencing (RNA-Seq) data is noisy and high-dimensional, particularly for microbes. Using this grant, we are going to develop a method to infer a low-dimensional representation, or pattern, of microbe abundance from RNA-Seq samples, and then correlate those patterns with various environmental and phenotypic conditions. One such correlation, for instance, could be between a population's exposure to dust and allergens, and prevalence of sickness in that population. After development, we will use latent Dirichlet allocation, or LDA, for dimension reduction and feature extraction.

]]]



Given each patient sample M as a document, all the genes and microbes are treated as corpus (V) respectively. The word (w) is gene or microbe, and the word count is gene expression or microbe abundances. As shown in the diagram, M is the number of patient, N is the number of words in all the document, K is the number of topics. \alpha, \beta is the hyper parameters of dirichlet distribution for \phi and \theta. Z is the topic of word in the document, and is chosen by a multinomial distribution with parameter \theta of a document. w denotes a word in the document, and each word can be selected by a multinomial distribution with \phi of the word in the document from the corpus(V).

MEG edit[[[

Given each patient sample M as a document, the collection of genes and the collection of microbes are each treated as separate corpora (V). Depending on the corpus, the word (W) is a gene or microbe, and the word count is gene expression or microbe abundances, respectively. As shown in the diagram, M is the number of patients, N is the number of words in all the documents, and K is the number of topics. \alpha and \beta are the hyper parameters of the Dirichlet distribution for \phi and \theta. Z is the topic of word in the document, and is chosen by a multinomial distribution with parameter \theta of the document. W denotes a word in the document, and each word can be selected by a multinomial distribution with \phi of the word in the document from the corpus (V).

]]]

The collapsed Gibbs sampling are used for parameter inferences. K modules(topics) will be identified and the high dimension gene expression and microbes abundance data will be represented by module/topics distribution \theta (\theta\_g,k and \theta\_m,k), and gene or microbes to module distribution \phi (\phi\_k,g or \phi\_k,m). Then \phi\_g and \phi\_m is used to learn the pattern of interactions between gene and microbe and then to infer the linkages.

MEG edit [[[

A collapsed Gibbs sampler is used for parameter inferences. K modules (topics) will be identified, and high-dimensional gene expression and microbe abundance data will be represented by a topic distribution \theta (\theta\_g,k and \theta\_m,k), and gene expression or microbe abundance with a topic distribution \phi (\phi\_k,g or \phi\_k,m). Then \phi\_g and \phi\_m are used to learn patterns of interactions between gene expression and microbial abundance, which are then used to infer potential relationships.

]]]

**References**

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

**[2] Rodriguez LL, Brooks LD, Greenberg JH, Green ED. 2013. The Complexities of Genomic Identifiability.Science 339:275–276**

**[3] Erlich Y, Narayanan A. 2014. Routes for breaching and protecting genetic privacy.Nat Rev Genet 15:409–421.**

**[4]** [**Franzosa EA**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Franzosa%20EA%5BAuthor%5D&cauthor=true&cauthor_uid=25964341)**,** [**Huang K**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Huang%20K%5BAuthor%5D&cauthor=true&cauthor_uid=25964341)**,** [**Meadow JF**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Meadow%20JF%5BAuthor%5D&cauthor=true&cauthor_uid=25964341)**, Gevers D,** [**Lemon KP**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lemon%20KP%5BAuthor%5D&cauthor=true&cauthor_uid=25964341)**,** [**Bohannan BJ**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Bohannan%20BJ%5BAuthor%5D&cauthor=true&cauthor_uid=25964341)**,** [**Huttenhower C**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Huttenhower%20C%5BAuthor%5D&cauthor=true&cauthor_uid=25964341)**. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A. 2015 Jun 2;112(22):E2930-8.**

**[5] Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. 2013. Highly Recurrent TERT Promoter Mutations in Human Melanoma.Science 339:957–959.**

**[6] Sotiriou C, Pusztai L. 2009. Gene-expression signatures in breast cancer.New England Journal of Medicine 360:790–800.**

**[7] GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project 45(6):580-5**

**[8] The ENCODE Project Consortium. 2012. An Integrated Encyclopedia of DNA Elements in the Human Genome.Nature 489:57–74.**

**[9] The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation.Nature 526:68–74.**

**[10] The Cancer Genome Atlas (TCGA). The Cancer Genome Atlas Research Network. National Institutes of Health;** [**http://cancergenome.nih.gov/**](http://cancergenome.nih.gov/)

**[11] Harmanci A, Gerstein M. 2016. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nature Methods. 13(3):251-6**

**[12] Harmanci A, Gerstein M. 2018. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat Commun. 9(1):2453**

**[13]** [Schadt EE](https://www.ncbi.nlm.nih.gov/pubmed/?term=Schadt%20EE%5BAuthor%5D&cauthor=true&cauthor_uid=22484626), [Woo S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Woo%20S%5BAuthor%5D&cauthor=true&cauthor_uid=22484626), [Hao K](https://www.ncbi.nlm.nih.gov/pubmed/?term=Hao%20K%5BAuthor%5D&cauthor=true&cauthor_uid=22484626). 2012. **Bayesian method to predict individual SNP genotypes from gene expression data.** [Nat Genet.](https://www.ncbi.nlm.nih.gov/pubmed/?term=ature+Genetics%2C+vol.+44.+pp.+603%E2%80%93608%2C+2012#)44(5):603-8.

[14] [Gymrek M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gymrek%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23329047), [McGuire AL](https://www.ncbi.nlm.nih.gov/pubmed/?term=McGuire%20AL%5BAuthor%5D&cauthor=true&cauthor_uid=23329047), [Golan D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Golan%20D%5BAuthor%5D&cauthor=true&cauthor_uid=23329047), [Halperin E](https://www.ncbi.nlm.nih.gov/pubmed/?term=Halperin%20E%5BAuthor%5D&cauthor=true&cauthor_uid=23329047), [Erlich Y](https://www.ncbi.nlm.nih.gov/pubmed/?term=Erlich%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=23329047). 2013. **Identifying personal genomes by surname inference.** [Science.](https://www.ncbi.nlm.nih.gov/pubmed/23329047#) 339(6117):321-4

[15] [Homer N](https://www.ncbi.nlm.nih.gov/pubmed/?term=Homer%20N%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Szelinger S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Szelinger%20S%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Redman M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Redman%20M%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Duggan D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Duggan%20D%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Tembe W](https://www.ncbi.nlm.nih.gov/pubmed/?term=Tembe%20W%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Muehling J](https://www.ncbi.nlm.nih.gov/pubmed/?term=Muehling%20J%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Pearson JV](https://www.ncbi.nlm.nih.gov/pubmed/?term=Pearson%20JV%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Stephan DA](https://www.ncbi.nlm.nih.gov/pubmed/?term=Stephan%20DA%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Nelson SF](https://www.ncbi.nlm.nih.gov/pubmed/?term=Nelson%20SF%5BAuthor%5D&cauthor=true&cauthor_uid=18769715), [Craig DW](https://www.ncbi.nlm.nih.gov/pubmed/?term=Craig%20DW%5BAuthor%5D&cauthor=true&cauthor_uid=18769715). 2008. **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.** 4(8):e1000167

[16] [Im HK](https://www.ncbi.nlm.nih.gov/pubmed/?term=Im%20HK%5BAuthor%5D&cauthor=true&cauthor_uid=22463877), [Gamazon ER](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gamazon%20ER%5BAuthor%5D&cauthor=true&cauthor_uid=22463877), [Nicolae DL](https://www.ncbi.nlm.nih.gov/pubmed/?term=Nicolae%20DL%5BAuthor%5D&cauthor=true&cauthor_uid=22463877), [Cox NJ](https://www.ncbi.nlm.nih.gov/pubmed/?term=Cox%20NJ%5BAuthor%5D&cauthor=true&cauthor_uid=22463877). 2012. **On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy.** Am J Hum Genet. 90(4):591-8

[17] **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 20:1297-303**

**[18] 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. 2013. Current Protocols in Bioinformatics. 43:11.10.1-11.10.33**

[19] Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. [CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research.](https://www.ncbi.nlm.nih.gov/pubmed/21324876) 21(6):974-84.

**[20]** [**Harewood L**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Harewood%20L%5BAuthor%5D&cauthor=true&cauthor_uid=28655341)**,** [**Kishore K**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kishore%20K%5BAuthor%5D&cauthor=true&cauthor_uid=28655341)**,** [**Eldridge MD**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Eldridge%20MD%5BAuthor%5D&cauthor=true&cauthor_uid=28655341)**,** [**Wingett S**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Wingett%20S%5BAuthor%5D&cauthor=true&cauthor_uid=28655341)**,** [**Pearson D**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Pearson%20D%5BAuthor%5D&cauthor=true&cauthor_uid=28655341)**,** [**Schoenfelder S**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Schoenfelder%20S%5BAuthor%5D&cauthor=true&cauthor_uid=28655341)**,** [**Collins VP**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Collins%20VP%5BAuthor%5D&cauthor=true&cauthor_uid=28655341)**,** [**Fraser P**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Fraser%20P%5BAuthor%5D&cauthor=true&cauthor_uid=28655341)**. 2017. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangementsand copy number variation in human tumours. Genome Biology. 18(1):125**

**[21]** Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, Yardımcı GG, Chakraborty A, Bann DV, Wang Y, Clark R, Zhang L, Yang H, Liu T, Iyyanki S, An L, Pool C, Sasaki T, Rivera-Mulia JC, Ozadam H, Lajoie BR, Kaul R, Buckley M, Lee K, Diegel M, Pezic D, Ernst C, Hadjur S, Odom DT, Stamatoyannopoulos JA, Broach JR, Hardison RC, Ay F, Noble WS, Dekker J, Gilbert DM, Yue F. 2018. Nature Genetics. [Epub ahead of print]

[22] Chakraborty A, Ay F. 2017. [Identification of copy number variations and translocations in cancer cells from Hi-C data.](https://www.ncbi.nlm.nih.gov/pubmed/29048467) Bioinformatics. [Epub ahead of print]

[23] [Zhao Z](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhao%20Z%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Tavoosidana G](https://www.ncbi.nlm.nih.gov/pubmed/?term=Tavoosidana%20G%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Sjölinder M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sj%C3%B6linder%20M%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Göndör A](https://www.ncbi.nlm.nih.gov/pubmed/?term=G%C3%B6nd%C3%B6r%20A%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Mariano P](https://www.ncbi.nlm.nih.gov/pubmed/?term=Mariano%20P%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Wang S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Wang%20S%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Kanduri C](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kanduri%20C%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Lezcano M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lezcano%20M%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Sandhu KS](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sandhu%20KS%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Singh U](https://www.ncbi.nlm.nih.gov/pubmed/?term=Singh%20U%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Pant V](https://www.ncbi.nlm.nih.gov/pubmed/?term=Pant%20V%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Tiwari V](https://www.ncbi.nlm.nih.gov/pubmed/?term=Tiwari%20V%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Kurukuti S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kurukuti%20S%5BAuthor%5D&cauthor=true&cauthor_uid=17033624), [Ohlsson R](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ohlsson%20R%5BAuthor%5D&cauthor=true&cauthor_uid=17033624). 2006. **Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics.** 38(11):1341-7

[24] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak M, Gaffney D, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology. 17:13-14

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

**[26] Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016.** [**uicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.**](https://www.ncbi.nlm.nih.gov/pubmed/27467249) **Cell Syst. 3(1):95-8.**

**[27] Ay F, Bailey TL, Noble WS. 2014.** [**Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts.**](https://www.ncbi.nlm.nih.gov/pubmed/24501021) **Genome Research. 24(6):999-1011**

**[28]** [**Yan KK**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Yan%20KK%5BAuthor%5D&cauthor=true&cauthor_uid=28742097)**,** [**Lou S**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lou%20S%5BAuthor%5D&cauthor=true&cauthor_uid=28742097)**,** [**Gerstein M**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gerstein%20M%5BAuthor%5D&cauthor=true&cauthor_uid=28742097)**. 2017. MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions.** [**PLoS Comput Biol.**](https://www.ncbi.nlm.nih.gov/pubmed/?term=mrtadfinder#) **13(7):e1005647**

**[29]** [**Yan KK**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Yan%20KK%5BAuthor%5D&cauthor=true&cauthor_uid=28369339)**,** [**Yardimci GG**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Yardimci%20GG%5BAuthor%5D&cauthor=true&cauthor_uid=28369339)**,** [**Yan C**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Yan%20C%5BAuthor%5D&cauthor=true&cauthor_uid=28369339)**,** [**Noble WS**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Noble%20WS%5BAuthor%5D&cauthor=true&cauthor_uid=28369339)**,** [**Gerstein M**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gerstein%20M%5BAuthor%5D&cauthor=true&cauthor_uid=28369339)**. 2017. HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. Bioinformatics. 33(14):2199-2201**

**[30] Sweeney L. 2002. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 0(5):557-570**

**Page breakdowns**

**Significance = 1 page**

**Innovation = 0.5 page**

**Research Strategy prelude = 0.5 page**

**Aim 1 = 3.5 pages**

**Aim 2 = 3 pages**

**All are within target.**

**For Aim 3, 0.5 page for Andrew's part so we have 3 pages left for the rest.**