# Experience on Genomic Privacy Analysis focused on functional genomics

Given the growing amount of available human genome data, a lot of which may be personally identifying, we investigated the privacy risks associated with the proliferation of functional genomics data. We also looked into possible mechanisms for mitigating those risks. Specifically, we considered options for balancing the requirement of individual privacy with the need for sufficient data to perform useful analyses. The large scale of functional genome data also lends itself naturally to data science approaches, including machine learning, knowledgebase design, and modeling by simulation.

# Previous experience in RNA-seq and ChIP-seq private information quantification

We previously developed a formalism for eQTL leakage from gene expression values and SV leakage from signal profiles of functional genomics data [1,2]. We showed that through various linking attack scenarios [14,15] we were able to connect individuals to databases of personal information. In the first scenario, we used the inference of 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 the gEUVADIS dataset. In this scenario, we assumed that we have a public phenotype panel, where personal information such as anonymized individuals’ HIV status and gene expression levels are released. For example, if we assume we have an individual’s SNVs and we would like to learn about his/her HIV status, we could infer the eQTLs from the expression values in the released phenotype panel and overlap them with the known SNVs, which then compromise the individual’s HIV status [1] .In the second attack scenario, we focused on the signal profiles from functional genomics data. The attack involves cross-referencing the individuals in a signal profile dataset, *S*, against the individuals in a genotype dataset, *G*. The signal profile dataset is publicly available for research purposes and contains a genome-wide signal profile and an anonymized identifier for each individual such as RNA-Seq data from TCGA. The signal profile for an individual represents the measurements of activity at each genomic position. This dataset stores a genome-wide signal profile for each individual, for example containing RNA-Seq or ChIP-Seq data. In addition, the signal profile dataset may contain sensitive characterizing information about each individual (e.g., HIV status). The genotype dataset, *G*, contains the genotypes of a panel of SV. *G* also contains the identities of the individuals and its access is restricted. The adversary obtains access to *G* by lawful or unlawful means (e.g., the adversary might have stolen it or might be allowed to access it but violated the terms of accession, known as stealing “variants from a glass”). The main objective of the adversary is to link *G* and *S* by first predicting the genotypes using signal profiles in *S*, and then matching the predicted genotypes to the genotypes in *G*. For any matching individuals in *G* and *S*, the name and sensitive information are revealed to the adversary. Our preliminary analysis on ChIP-Seq and RNA-Seq signal profiles showed that RNA-Seq is concentrated on exonic regions and has high depth but low breadth of coverage and therefore can be used to genotype small deletions; by contrast, ChIP-Seq signal profiles generally have high breadth but low depth of coverage and can be used to detect large deletions.

# Previous experience in RNA-seq and ChIP-seq data anonymization

We have developed a number of tools and data formats to handle the increasingly large quantities for data generated by RNA-Seq experiments. For example, we have developed the Mapped Read Format (MRF), a compact data summary format for short, long and paired-end read alignments that enables the anonymization of confidential sequence information. We also developed privacy-aware binary alignment mapping (pBAM) file format to convert the sequence alignment files for functional genomics experiments such as RNA-Seq and ChIP-Seq to a format, which we can calculate the commonly calculated properties such as gene expression, peak calling but we can also remove the sensitive information [3].

We also developed tools to deal with the leakage from signal tracks [2]. The most effective way to protect against a linking attack scenario is to ensure that deletion genotypes cannot be inferred from signal profiles. Deletions are a major source of leakage of genetic information from functional genomics signal profiles. We proposed solutions to the signal profiles where we can mask the sensitive information leakage, while providing high utility. Our proposed solution systematically removes the dips in signal profiles as a way to anonymize the profiles against the prediction of deletions. To remove these dips systematically, we used 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.

# REFERENCES

1. Harmanci and Gerstein, Quantification of private information leakage from phenotype-genotype data: linking attacks. Nature Methods 13 (3), 251

2. Harmanci and Gerstein, Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nature Communications 9 (1), 2453

3. Gursoy et al., Sensitive information leakage from functional genomics data: Theoretical quanbtifications & practical file formats for privacy preservations