Functional genomics experiments, such as RNA sequencing, reveal important information about

dynamic changes in gene expression under different conditions. There is great desire to share these databecause they are extremely valuable for biomedical and disease research. The raw reads are not shared because of privacy concerns. To enable safe sharing, aggregated formats such as read depth signal profiles and gene expression quantifications are used. Projects such as GTEx and ENCODE publicly share these data because they ostensibly do not leak much identifying information. We study the leakage from genome-wide signal profiles when signal profile is used for genotyping genomic deletions. We present information theoretic measures for genotypability of deletions and leakage therefrom. We develop practical genotyping methods and demonstrate that an individual can be identified within a large sample in the context of linking attacks. We finally present an anonymization method that removes much of the leakage from signal profiles.

Functional genomics experiments, such as RNA-seq, provide generic (i.e. non-individual specific) information about gene expression under different conditions (e.g. disease v. normal). There is great desire to share these data. However, privacy concerns often preclude sharing of the raw reads. To enable safe sharing, aggregated summaries such as read-depth signal profiles and levels of gene expression are used. Projects such as GTEx and ENCODE share these because they ostensibly do not leak much identifying information. Here, we attempt to quantify the validity of this statement, measuring the leakage of genomic deletions from signal profiles. We present information theoretic measures for the degree to which one can genotype these deletions. We then develop practical genotyping approaches and demonstrate how to use these to identify an individual within a large cohort in the context of linking attacks. Finally, we present an anonymization method removing much of the leakage from signal profiles.

Functional genomics experiments such as RNA sequencing are performed to reveal gene expression changes under different conditions and diseases. Although the main purpose of these approaches is to understand dynamic changes in gene expression levels, the data also contain a large number of genetic variants in the raw reads. Therefore, the raw reads cannot be shared publicly because of privacy concerns. There is, however, great desire to publicly share as much of the data as possible since they are extremely valuable for biomedical and disease research. To enable safe data sharing, researchers often use aggregated representations computed from raw reads, such as read depth signal profiles and gene expression quantifications. These representations do not explicitly reveal variant information and are generally assumed to be safe to share. Here, we study the privacy aspects of genome-wide signal profiles of functional genomics experiments, which represent measurement of activity at each genomic position. We show that the signal profiles, which are often publicly shared, can be used to genotype small and large deletions, which can then be used to breach privacy. We first present information-theoretic measures for predictability of genotypes from signal profiles and information leakage from signal profiles. We then present practical methods for detecting and genotyping small and large genomic deletions, and demonstrate that the genotyped deletions can accurately identify an individual from a large sample in the context of linking attacks. We also present an effective anonymization procedure for the protection of signal profiles against the presented genotype prediction based attacks. Given that several consortia, such as the GTEx and ENCODE, publicly share signal profiles, these results point to a potential source of sensitive information leakage.