Information theory-based measures and privacy-preserving file formats for sensitive information leakage from raw functional genomics data

Functional genomics experiments on human subjects present a privacy conundrum. On one hand, many of the conclusions we infer from these experiments are not tied to the identity of individuals but represent universal statements about disease and developmental stages. On the other hand, by virtue of the experimental procedures, the reads from them are tagged with small bits of patients' variant information, which presents privacy challenges in terms of data sharing. There is great desire to share the data as broadly as possible. Therefore measuring the amount of variant information leaked in a variety of experiments, particularly in relation to the amount of sequencing will allow us to uncover ways of reducing the information leakage, and determine an appropriate set point for sharing information with minimal leakage. To this end, we aimed to derive information theoretic measures for the private information leaked in experiments and develop various file format manipulations to reduce much of the leaked variants. We showed that high depth experiments such as Hi-C provide accurate genotyping that can lead to large privacy leaks. Counter intuitively, noisy and partial genotypes from low-depth experiments such as ChIP-Seq and single-cell RNA-Seq, although not useful genotypes, can be used as strong quasi-identifiers for re-identification purposes through linking attacks. We showed that these incomplete genotypes can further be used to construct an individual's complete variant set and inference of individual identifying phenotypes when combined with imputation. We then provide a proof-of-concept theoretical framework, in which the amount of leaked information can be estimated from the depth and breadth of the coverage as well as the sequencing bias of the functional genomics experiments. In order to solve the dilemma between data sharing and privacy leakage, we propose a file formatting system that enables the sharing data while protecting individuals' sensitive information and preserving the utility of the data. The proposed file format can achieve different levels of privacy and utility balance. At the highest level of privacy, our file format masks all the variant information leaked from reads, which can be used to calculate signal profiles with 99% recovery of the original profiles and 100% recovery of the original gene expression levels.