Private information leakage from functional genomics data: Theoretical quantifications & practical file formats

Sensitive information leakage from functional genomics data: Theoretical quantifications & practical file formats for privacy preservation

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 biology and disease. On the other hand, by virtue of the experimental procedure, the sequencing 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 is a key first step in reducing the information leakage and determining an appropriate "set point" for sharing, with minimal leakage. To this end, we derive information-theoretic measures for the private information leaked in experiments and develop various file formats to reduce this in sharing. We show that high depth experiments such as Hi-C provide accurate genotyping that can lead to large privacy leaks. Counterintuitively, low-depth experiments such as ChIP-Seq and single-cell RNA-Seq, although not useful for genotyping, can be create strong quasi-identifiers for re-identification through linking attacks. We show that partial and incomplete genotypes from many of these experiments can further be combined to construct an individual's complete variant set and identifying phenotypes. We provide a proof-of-concept analytic framework, in which the amount of leaked information can be estimated from the depth and breadth of the coverage as well as sequencing biases of a given functional genomics experiment. Finally, as a practical instantiation of our framework, we propose file formats that maximize the potential sharing of data while protecting individuals’ sensitive information. Depending on the desired sharing set point, our proposed format can achieve differential tradeoffs in the privacy-utility balance. At the highest level of privacy, we mask all the variants leaked from reads, but still can create useable signal profiles that give complete recovery of the original gene expression levels.