1 Genomic Privacy

1.1 GWAS Identification

- Common SNPs can be used to identify individuals and predict phenotype if their GWAS information is known
- On the other hand, wide-scale sharing of GWAS data for large numbers of individuals is extremely useful for the advancement of science
- How can we balance privacy and progress? Can useful research be done in a way that preserves privacy?

1.2 Privacy and Information

- Different information can identify a person to different degrees
- DOB, ZIP, and sex uniquely identify about 60% of people
- Y-Chromosome mapping can easily identify some males
- One idea to hide individual identities while still using the data is presenting only summary statistics, but these can actually be used to identify individuals

1.3 Limits of Individual Detection

One way to formalize these ideas is to treat identification as a hypothesis test. The null hypothesis is that the individual of interest is not in the study, unless there is sufficient evidence to identify that person as a study member. A Log-Likelihood ratio for this hypothesis can be written as

\[
\bar{L} = \sum_{i=1}^{n} \left( x_i \log \frac{\hat{p}_i}{p_i} + (1 - x_i) \log \frac{1 - \hat{p}_i}{1 - p_i} \right)
\]

which can be used to give the relation

\[
Z_{\alpha} + Z_{1-\beta} = \sqrt{\frac{m}{n}}
\]

where \(\alpha\) is the significance level, \(\beta\) is the power, \(m\) is the numbers if SNPs, and \(n\) is the number of individuals.

1.4 Completion Attacks

- How can you protect privacy?
  1. You can’t it’s too hard
  2. You should restrict access to information
3. You should make sure that a person isn’t identifiable beyond a certain level

- One idea for achieving this last goal is $k$-anonymity. Each study should have at least $k$ people with the same genotype, so that no individual is identifiable beyond a group of size $k$ from gene data

Differential privacy requires that

$$\frac{P(Z \in S|X = x)}{P(Z \in S|X = y)} \leq e^\alpha$$

where $Z$ is a test statistic for a hypothesis test distinguishing two individuals and $\alpha$ is a security level. If DP holds and the threshold of test statistic $T$ is $\gamma$, then the power of the test statistic is at most $\gamma e^\alpha$. 