Supplementary Material for
Re-Identification of Individuals in Genomic Data-Sharing
Beacons via Allele Inference

Nora von Thenen, Erman Ayday, A. Ercument Cicek

A  SNP Network

When two SNPs A and B are in LD, the probability of two major or two minor alleles occurring together increases or decreases by D. As shown in Table A.1 this leads to the formula: \( \Pr(ab) = p_2q_2 + D \), where \( p_2 \) is the minor allele frequency of SNP A (with minor allele a) and \( q_2 \) is the minor allele frequency of SNP B (with minor allele b). D is calculated as follows: \( D = \sqrt{r^2(q_1q_2p_1p_2)} \), where \( q_1 \) and \( p_1 \) are the major allele frequencies and \( r^2 \) is a common measure of LD. To determine whether the LD correlation increases or decreases the probability of the two loci occurring together \( D' \) is needed. \( D' > 0.5 \) implies \( D \) is added, whereas \( D' < 0.5 \) leads to a subtraction of \( D \).

Table A.1: Relationship between Linkage Disequilibrium (LD) measured by \( D \) between the SNPs A and B and their allele frequencies.

<table>
<thead>
<tr>
<th>( \Pr(A) = p_1 )</th>
<th>( \Pr(a) = p_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Pr(B) = q_1 )</td>
<td>( p_1q_1 + D )</td>
</tr>
<tr>
<td>( \Pr(b) = q_2 )</td>
<td>( p_1q_2 - D )</td>
</tr>
</tbody>
</table>

The SNP network that is used to determine correlated SNPs is build as a directed graph. The edges are labeled with the probability of two minor alleles occurring together in the considered SNP positions as shown in the lower right field of Table A.1. In order to ensure high correlation between the SNPs, only LD relationships between SNP pairs with an \( r^2 \) value of more than 0.7 were considered. The average of the correlation between SNP pairs and therefore the labels on the edges of the SNP network is 0.9511. An example of a SNP network with 5 nodes is shown in Figure A.1.

B  LRT - Query Inference Attack

The likelihood of the null hypothesis is calculated as follows:

\[
L_{H_0}(R) = \sum_{i=1}^{n} \left( x_i \log(1-D_N^i) + (1-x_i) \log(D_N^i) + \sum_{j=1}^{m} \gamma x_i \log(1-D_N^j) + \gamma (1-x_i) \log(D_N^j) \right)
\]  

(1)

where, \( x_i \) is the response of queried SNP \( i \), \( n \) the number of posed queries, \( m \) the number of queries that were inferred with query SNP \( i \), and \( \gamma \) the confidence of the inferred responses.

Accordingly, the likelihood of the alternative hypothesis is found as follows:

\[
L_{H_1}(R) = \sum_{i=1}^{n} \left( x_i \log(1-\delta D_{N-1}^i) + (1-x_i) \log(\delta D_{N-1}^i) + \sum_{j=1}^{m} \gamma x_i \log(1-\delta D_{N-1}^j) + \gamma (1-x_i) \log(\delta D_{N-1}^j) \right)
\]  

(2)
Figure A.1: A SNP network that contains 5 nodes (i.e., SNPs). The SNP network is a directed graph, where the weight of edges correspond to the correlation between SNPs. No edge between a pair of SNPs means the correlation is less than the $r^2$ threshold. The correlation is in the most cases not symmetric, since it depends on the minor allele frequencies of the SNP pair. This example shows a fully connected graph, which is not necessarily the case for all SNP networks.

Using (1) and (2), $\Lambda$ can be calculated as shown in Equation 3.

$$\Lambda = L_{H_0}(R) - L_{H_1}(R)$$

$$= \sum_{i=1}^{n} \left( x_i \log(1 - D^i_N) + (1 - x_i) \log(D^i_N) \right)$$

$$+ \sum_{j=1}^{m} \gamma x_i \log(1 - D^j_N) + \gamma (1 - x_i) \log(D^j_N)$$

$$- \left[ \sum_{i=1}^{n} \left( x_i \log(1 - \delta D^i_{N-1}) + (1 - x_i) \log(\delta D^i_{N-1}) \right) \right]$$

$$+ \sum_{j=1}^{m} \gamma x_i \log(1 - \delta D^j_{N-1}) + \gamma (1 - x_i) \log(\delta D^j_{N-1})$$

$$= \sum_{i=1}^{n} \left( x_i \log \left( \frac{1 - D^i_N}{1 - \delta D^i_{N-1}} \right) + (1 - x_i) \log \left( \frac{D^i_N}{\delta D^i_{N-1}} \right) \right)$$

$$+ \sum_{j=1}^{m} \gamma x_i \log \left( \frac{1 - D^j_N}{1 - \delta D^j_{N-1}} \right) + \gamma (1 - x_i) \log \left( \frac{D^j_N}{\delta D^j_{N-1}} \right)$$

$$= \sum_{i=1}^{n} \left( \log \left( \frac{D^i_N}{\delta D^i_{N-1}} \right) + \log \left( \frac{\delta D^i_{N-1}(1 - D^i_N)}{D^i_N(1 - \delta D^i_{N-1})} \right) x_i \right)$$

$$+ \sum_{j=1}^{m} \log \left( \frac{D^j_N}{\delta D^j_{N-1}} \right) + \log \left( \frac{\delta D^j_{N-1}(1 - D^j_N)}{D^j_N(1 - \delta D^j_{N-1})} \right) \gamma x_i$$

C High-Order Markov Chain

Individuals may hide certain loci on their genome before publishing their VCF files. It is possible to infer these hidden positions by applying a high-order Markov chain as introduced by Samani et al.. The probability of a certain allele occurring at a specific position can then be determined as

$$P_k(SNP_i) = P(SNP_i|SNP_{i-1}, SNP_{i-2}, ..., SNP_{i-k}),$$

(4)
Table C.1: Comparison of different values for \( k \) (order of the high-order Markov chain). \# of same markers shows how many markers that were inferred by the Markov chain were also asked in the Optimal attack. Distance to real response shows the amount of queries the inferred response differs from the Optimal attack’s response (on average). \# of people not inferred shows the amount of people that could not be inferred for that \( k \).

<table>
<thead>
<tr>
<th>( k ) (order)</th>
<th># of same markers</th>
<th>distance to real response</th>
<th># of people not inferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>13</td>
<td>0.6</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>0.62</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>0.79</td>
<td>5</td>
</tr>
</tbody>
</table>

where \( k \) is the order of the Markov chain. Accordingly, Samani et al., 2015 define the \( k \)-th-order model by:

\[
P_k(SNP_i) = \begin{cases} 0 & \text{if } F(SNP_{i-k,i-1}) = 0 \\ \frac{F(SNP_{i-k,i-1})}{F(SNP_{i-k,i-1})} & \text{if } F(SNP_{i-k,i-1}) > 0 , \end{cases}
\]

where \( F(SNP_{i,j}) \) is the frequency of occurrence of the sequence that contains \( SNP_i \) to \( SNP_j \).

To build a high-order Markov chain to infer hidden SNPs, genome sequences from public sources such as the 1000 Genomes project or HapMap can be used to train the model. Our Markov chain is build of 100 individuals of the CEU population HapMap dataset. The SNPs are represented as 0, 1, or 2 depending on the number of minor alleles at the specific position of the genome. That is, major homozygous, heterozygous, and minor homozygous, respectively. If the VCF file of the victim has too much missing data we cannot infer beacon membership of the victim. That is, there is not enough SNP information to build the necessary \( k \)-order Markov chain to infer the hidden SNPs with low MAFs. Missing data can occur due to sequencing errors or can be intentionally hidden by the victim.

In order to select \( k \), we considered three criteria: (i) the number of markers that are inferred by the GI-attack and also asked by the Optimal attack, (ii) the euclidean distance between the number of queries needed by the Optimal attack and the GI-attack for all tested individuals, and (iii) the number of people whose SNPs could not be inferred due to missing data.

As shown in Table C.1, we picked \( k = 4 \), which provides the largest number of markers, with minimum number of missed people. The distance to the real response is similar to \( k = 3 \) and is less than the case when \( k = 5 \). As our Markov chain is build on 100 individuals, we determined \( k = 3 \) as the maximum order to be tested to prevent over-fitting.

### D LRT - Power Calculation

The power \( (1 - \beta) \) of the LRT is determined as the proportion of control individuals (that are in the beacon) for which we can reject the null hypothesis when \( \Lambda < t_\alpha \). The threshold \( t_\alpha \) is found by building the null hypothesis with the 40 case individuals (that are not in the beacon), where \( \alpha = 0.05 \) (corresponding to 5% false positive rate).

For each individual and query \( x_i \), we calculate the value of \( \Lambda \), where \( \Lambda \) changes according to the attack being performed. As \( t \) increases, the power of the QI-attack shows a zig-zag behavior unlike the Optimal attack and the GI-attack. That is because as \( t \) increases, more queries are needed to determine beacon membership, and more SNPs are in inferred in the QI-attack. The more neighbors a posed query can infer from the SNP network, the more extreme the value of \( \Lambda \) changes.

Figure D.1 shows, for three example case and three example control individuals, how \( \Lambda \) steadily decreases for control individuals and clearly increases for “no” responses of case individuals (i.e., at queries 24, 26 and 84) for the Optimal attack. Here, \( \Lambda \) decreases by a similar value for all individuals that receive a “yes” response, as only one query is asked and the queries have similar MAFs. Therefore, if Control 1 had a lower \( \Lambda \) value at query \( x_{10} \) than Case 2, Case 2 will not have a lower \( \Lambda \) value than Control 1 for the following queries, unless Control 1 receives a “no” response (which leads to a significant increase in \( \Lambda \) but is highly unlikely for an individual in the control set).

On the contrary, Figure D.2 shows an irregular behavior of \( \Lambda \), that is \( \Lambda \) does not steadily decrease unlike the Optimal attack in Figure D.1. This can be explained by the different amount of neighbors in the SNP network that can be inferred at the different loci. Considering Control 1 and Case 2 again, Control 1 can have a lower \( \Lambda \) value than Case 2 for query position \( x_{10} \). Nevertheless, if query \( x_{11} \) of Case 2 has a high number of neighbors to be inferred from \( x_{11} \) and the inferred responses are all “yes” responses, the \( \Lambda \) value of Case 2 decreases significantly and is
now lower than the \( \Lambda \) value of Control 1 for \( x_{11} \), as \( x_{11} \) of Control 1 has fewer neighbors in the SNP network.

![Figure D.1: Example \( \Lambda \) distributions for 3 of the 40 case and 3 of the 20 control individuals of the experiments with a simulated beacon in Section for the Optimal attack.](image)

![Figure D.2: Example \( \Lambda \) distributions for 3 of the 40 case and 3 of the 20 control individuals of the experiments with a simulated beacon in Section 3.1 for the QI-attack.](image)

**E Ground Truth for the Tests on Existing Beacons**

In order to determine if the selected PGP(Personal Genomes Project) individual (PGP180/hu2D53F2) is in a beacon or not, we applied the SB attack on all the beacons used in Section 3.2. Thus, the decision made by the SB attack is independently used as the ground truth. The null hypothesis (the individual is not in the beacon) is rejected if \( p \) value is smaller than 0.05. The \( p \) value is calculated as \( P(x \geq k; x \text{binomial}(n, 1-D_N))) \). Here, \( N \) is the size of the beacon, \( k \) is the number of “yes” responses to \( n \) asked queries, \( x \) is the response and \( D_N \) is the probability of no individual in the beacon having the queried allele. The tested individual had \( p \) value = 1 for all beacons and we concluded that s/he is not a member of any of the beacons we tested. In addition, the meta-data of the Kaviar beacon does not list this person as a member.

**F Experiments on Existing Beacons**

Here, we show the evaluation results of our tests on the beacons of the beacon-network when an empty answer is not treated as a “no”. The results are shown in Table F.1. In summary, We could not detect the correct membership status for only 1 of the 9 beacons. The main difference of this case from the case in which an empty answer is treated as a “no” is that if empty answers are considered as a “no” response, the number of queries needed to determine beacon membership decreases significantly for some beacons. Nevertheless, if the empty answers are a result of two different genome copies (one in the beacon and one at hand) this conclusion would be incorrect.

**G GI-Attack without Population Information**

In this section, we used a different training dataset to train the high-order Markov chain. The case and control individuals are the same as in the results shown in Section 3.1, that is from the CEU population. The high-order Markov chain was trained on the 77 individuals from the HapMap dataset “Mexican ancestry in Los Angeles” (MEX).

We observed that the GI-attack still performs very well, even when the population of the training dataset for the high-order Markov chain does not match the victim’s population as shown in Figure G.1.
Table F.1: Number of queries required to receive a no within 1000 queries to existing beacons using an individual from PGP when \( t = \{0, 0.03, 0.05\} \) for the Optimal attack, the QI-attack, and the GI-attack. Here, empty answers (i.e., the beacon has no information about the queried locus in the underlying dataset and returns neither a no nor a yes) are considered as a no response. - means no no was found in 1000 queries.

<table>
<thead>
<tr>
<th>Beacon Name</th>
<th>Optimal attack</th>
<th>QI-attack</th>
<th>GI-attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known VAriants</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Broad Institute</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1000 Genomes Project</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Cafe CardioKit</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger Institute</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCBI</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ICGC</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AMPLab</td>
<td>20</td>
<td>45</td>
<td>73</td>
</tr>
<tr>
<td>1000 Genomes Project phase 3</td>
<td>20</td>
<td>130</td>
<td>250</td>
</tr>
</tbody>
</table>

Figure G.1: The GI-attack for \( t = 0.03 \) with the high-order Markov chain trained on the victim’s population (CEU) in comparison to the high-order Markov chain trained on a different population (here MEX) from the HapMap dataset.

References