# Supplementary Materials:

# Addressing Beacon Re-Identification Attacks: Quantification and Mitigation of Privacy Risks

*Table S1. Table of symbols and abbreviations*

| Notation | Description |
|---|---|
| $N$ | Total number of genomes in the beacon. |
| $Q = \{q_1, \ldots, q_n\}$ | Set of $n$ queries. |
| $R = \{x_1, \ldots, x_n\}$ | Set of $n$ responses returned by the beacon. |
| $H_0$ | Null hypothesis: query genome is not in the beacon. |
| $H_1$ | Alternative hypothesis: query genome is in the beacon. |
| $f_i$ | Alternate allele frequency at the SNP corresponding to query $q_i$. |
| $p_i$ | Reference allele frequency at the SNP corresponding to query $q_i$, $(p_i = 1 - f_i)$. |
| $L(R)$ | Log-likelihood of a response set $R = \{x_1, \ldots, x_n\}$. |
| $L_{H_0}(R)$, $L_{H_1}(R)$ | Log-likelihood under the null/alternative hypothesis. |
| $beta(a, b)$ | Alternate allele frequency distribution assumed in the original by Shringarpure and Bustamante [1]. |
| $D^i_{N-1}$ | Probability that none of the $N-1$ genomes in the beacon has an alternate allele for query $q_i$ under $H_1$ |
| $D^i_N$ | Probability that none of the $N$ genomes in the beacon has an alternate allele for query $q_i$ under $H_0$ |
| $\delta$ | Probability of mismatch between the query genome and its copy in the beacon due to sequencing errors. |
| $j \in 1, \ldots, N$ | Index of individuals in the beacon. |
| $i \in 1, \ldots, n$ | Index of queries. |
| $\alpha$ | Type I error: $P(\text{reject } H_0 \mid H_0 \text{is true})$. False positives. |
| $\beta$ | Type II error: $P(\text{accept } H_0 \mid H_1 \text{is true})$. False negatives. |
| $power$ | $P(\text{reject} H_0 \mid H_1 \text{istrue}) = 1 - \beta$. |
| $r_i$ | Risk of query $i$. |
| $b_j$ | Budget of patient $j$. |
| $LRD_{H_1}$, $LRD_{H_0}$ | Likelihood ratio distribution under the alternative/null hypothesis. |
| $\Lambda$ | LRT statistic. |
| $t$ | Cut-off for the LRT statistic $\Lambda$ (the null hypothesis is rejected if $\Lambda < t$). |
| $Q^j$ | Set of queries answered by individual $j$. |
| $k$ | Threshold on the number of individuals carrying an alternate allele at the queried SNP (used in strategy $S1$). |
| $\varepsilon$ | Probability of adding noise on unique alleles (used in strategy $S2$). |
| $b_j$ | Budget for individual $j$. Initially $b_j = -log(p)$ for every $j$ (used in strategy $S3$). |
| $r_i$ | Risk for query $i$ (i.e., how much budget for every individual $j$ is deducted from $B_j$ if the beacon answers query $i$). |
| LRT | Likelihood Ratio Test. |
| SNP | Single Nucleotide Polymorphism. |
| VCF | Variant Call Format. |

**APPENDIX A: LRT UNDER BEACON ALTERATION STRATEGY (*S1*)**

The first strategy ($S1$) is based on the observation that most of the statistical power in the re-identification attack comes from queries targeting unique alleles in a beacon database. In particular, the proposed algorithm alters the beacon by answering a query with *"Yes"* only if there are at least $k > 1$ individuals sharing the queried allele. We assume the value of $k$ is made public, hence the attacker will modify the attack to accommodate this change.

Formally, the attacker knows the allele frequencies for the SNPs in the victim's genome, and these frequencies can be ordered randomly or sequentially. In this setting, Equation (1) still holds, but Equations (2) and (3) needs to be modified as they now depend on $k$.

Under the alternative hypothesis, the beacon responds "*No*" if either of the following two conditions is met.

- A sequencing error $\delta$ occurred and less than $k$ other individuals have a copy of the allele
- No sequencing error occurred but less than $k-1$ other individuals have a copy of the allele

Hence, we have

$$L_{H_1}(R,k) = \sum_{i=1}^{n} x_i \log(\Pr(x_i = 1|H_1,k)) + (1-x_i)\log(\Pr(x_i = 0|H_1,k))$$
$$= \sum_{i=1}^{n} x_i \log(\delta(1 - D^i_{N-1}(k)) + (1-\delta)(1 - D^i_{N-1}(k-1)))$$
$$+(1-x_i)\log(\delta D^i_{N-1}(k) + (1-\delta)D^i_{N-1}(k-1)) \quad \text{(S1)}$$

where $D^i_{N-1}(k)$ denotes the probability that fewer than $k$ out of $N-1$ individuals have an alternate allele (for query $q_i$). Let $X_{N-1,s_i}$ be a random variable following a binomial

distribution with $N-1$ trials and success probability $s_i = 1-(1-f_i)^2$, where $s_i$ represents the probability that a given individual (other than the victim) has at least one copy of an alternate allele (for query $q_i$) with frequency $f_i$. Then,

$$D_{N-1}^{i}(k) = \Pr(\text{less than k out of N} - 1 \text{ genomes have an alternate allele at position i})$$
$$= \Pr(X_{N-1,s_i} < k) = \sum_{j=0}^{k-1} \binom{N-1}{j} (1-(1-f_i)^2)^j ((1-f_i)^2)^{N-1-j}. \quad \text{(S2)}$$

Similarly, under the null hypothesis, the probability that the beacon responds "*No*" to a query $q_i$ for an allele with frequency $f_i$ is the probability that at most $k-1$ individuals have a copy of the query allele. Hence, we have

$$L_{H_0}(R,k) = \sum_{i=1}^{n} x_i \log(\Pr(x_i = 1|H_0,k)) + (1-x_i)\log(\Pr(x_i = 0|H_0,k))$$
$$= \sum_{i=1}^{n} x_i \log(1-D_N^i(k)) + (1-x_i)\log(D_N^i(k)) \quad \text{(S3)}$$

Therefore, the likelihood ratio test statistic $\Lambda(k)$ when $k \geq 2$ can be computed by

$$\Lambda(k) = L_{H_0}(R,k) - L_{H_1}(R,k)$$
$$= \sum_{i=1}^{n} \log\left(\frac{D_N^i(k)}{\delta D_{N-1}^i(k)+(1-\delta)D_{N-1}^i(k-1)}\right)$$
$$+\log\left(\frac{(1-D_N^i(k))(\delta D_{N-1}^i(k)+(1-\delta)D_{N-1}^i(k-1))}{D_N^i(k)(\delta(1-D_{N-1}^i(k))+(1-\delta)(1-D_{N-1}^i(k-1)))}\right) x_i \ . \quad \text{(S4)}$$

Note that if $k = 1$, from Equations (S1) and (S3) we can obtain Equations (2) and (3), respectively.

An alternative approach is to hide the precise number of individuals within a beacon database and instead provide an approximate database size (e.g., the reported database size is 100 although the actual database size is 1000). In this case, let the approximate size of a beacon database that the attacker knows be $N_a$; thus, the LRT statistic $\Lambda$ can be calculated according to Equation (5), where $N = N_a$.

$$\Lambda = \sum_{i=1}^{n} \log(\delta^{-1}(1-f_i)^2) + \log\left(\frac{\delta}{(1-f_i)^2} \cdot \frac{1-(1-f_i)^{2N_a}}{1-\delta(1-f_i)^{2N_a-2}}\right) x_i \ . \quad \text{(S5)}$$

**APPENDIX B: LRT UNDER RANDOM FLIPPING STRATEGY (*S2*)**

The second strategy (*S2*) relies on the same observation of *S1* but instead of altering the beacon response, it introduces noise into the original data. *S2* improves the usability of the beacon over *S1* as it hides only a portion $\varepsilon$ of unique alleles, but not all. In other words, a beacon with *S2* will add noise with probability $\varepsilon$ only to unique alleles in the database and provide false answers (e.g., "*No*" instead of "*Yes*") to queries targeting these unique alleles. Without loss of generality, we assume the value of $\varepsilon$ is public. As for *S1* the attacker will adapt the LRT statistic to take it into account.

Formerly, and also in this case, the attacker knows the allele frequencies for the SNPs in the victim's genome and performs queries by following the *rare-allele-first* model. Similarly to *S1*, Equation (1) still holds, but Equations (2) and (3) needs to be modified again as they now depend on $\varepsilon$.

Under the alternative hypothesis, the beacon responds "*No*" if either of the following two conditions is met.

- A sequencing error $\delta$ occurred and none of the other $N-1$ participants has a copy of the allele.
- An artificial error $\varepsilon$ occurred and the allele is unique. Note that an allele is unique if a sequencing error occurred and another participant has a copy of the allele or if no sequencing error occurred and none of the other $N-1$ participants has a copy of the allele.

Hence, we have

$$L_{H_1}(R,\varepsilon) = \sum_{i=1}^{n} x_i \log(\Pr(x_i = 1|H_1,\varepsilon)) + (1-x_i)\log(\Pr(x_i = 0|H_1,\varepsilon)), \quad \text{(S6)}$$

where the probability of a "*No*" answer is

$$\begin{aligned}\Pr(x_i = 0|H_1,\varepsilon) &= \Pr(\text{none of N} - 1 \text{ genomes have an alternate allele at position } i) \\ &+\varepsilon\Pr(\text{allele at position i is unique}) \\ &= \delta D_{N-1}^{i} + \varepsilon(\delta\Pr(X_{N-1,s_i} = 1) + (1-\delta)D_{N-1}^{i}) \\ &= \varepsilon\delta\Pr(X_{N-1,s_i} = 1) + (\delta + \varepsilon - \varepsilon\delta)D_{N-1}^{i}. \quad \text{(S7)}\end{aligned}$$

Note that $\Pr(X_{N-1,s_i} = 1)$ denotes the probability that another participant has a copy of the allele at position $i$. As in Appendix A, we can derive such a probability as

$$\begin{aligned}\Pr(X_{N-1,s_i} = 1) &= \binom{N-1}{1}(1-(1-f_i)^2)((1-f_i)^2)^{N-1} \\ &= (N-1)(1-(1-f_i)^2)((1-f_i)^2)^{N-1}. \quad \text{(S8)}\end{aligned}$$

Similarly, under the null hypothesis we have

$$L_{H_0}(R,\varepsilon) = \sum_{i=1}^{n} x_i \log(\Pr(x_i = 1|H_0,\varepsilon)) + (1-x_i)\log(\Pr(x_i = 0|H_0,\varepsilon)), \quad \text{(S9)}$$

where the probability of receiving a "*No*" answer from the beacon is

$$\begin{aligned}\Pr(x_i = 0|H_0,\varepsilon) &= \Pr(\text{none of N genomes have an alternate allele at position i}) \\ &+\varepsilon\Pr(\text{allele at position i is unique}) \\ &= D_N^i + \varepsilon\Pr(X_{N,s_i} = 1)\end{aligned} \quad \text{(S10)}$$

Finally, the likelihood ratio test statistic $\Lambda(\varepsilon)$ can be easily derived from Equations (S6) and (S9) as in Appendix A.

**APPENDIX C: QUERY BUDGET PER INDIVIDUAL STRATEGY (*S3*)**

The third strategy (*S3*) aims at mitigating the re-identification risk by assigning a budget to each individual in the database, which is applied to each authenticated Beacon user. With respect to strategies *S1* and *S2* described above, *S3* leverages two additional assumptions:

- Each Beacon user has been identity proofed, holds a single account, is authenticated, does not collude. If users are allowed to collude, then *S3*, to be effective, will have a dramatic impact on the utility of the system.
- The attacker has accurate genomic information, which means $\delta = 0$. This assumption is necessary to simplify the mathematics of the problem and is a worst-case assumption, as if we can prevent re-identification under this condition, we can prevent against the optimal attack.

Let $R$ be the set of responses of the beacon, the basic idea of *S3* is to keep track of the power of the attack which is based on the log likelihood-ratio test $\Lambda = L_{H_0}(R) - L_{H_1}(R)$, in order to prevent any individual genome from contributing to a query response that can leak identity information with high confidence.

More formally, we define a cut-off threshold $t_\alpha$ on the value of $\Lambda$ to determine which hypothesis to accept (i.e., the null hypothesis is rejected if $\Lambda < t_\alpha$). Then the false-positive rate is $\alpha = \Pr[\Lambda < t_\alpha | H_0]$ and the power of the test is $1 - \beta = \Pr[\Lambda < t_\alpha | H_1]$.

So to validate that the original attack is thwarted by *S3*, we first need to know the distribution of $\Lambda$ under $H_0$ and $H_1$. In the analysis by Shringarpure and Bustamante , it is shown that $\Lambda$ is asymptotically Gaussian under both hypotheses (with different parameters). In our case, this result does not hold because we set $\delta = 0$ and assume fixed allele frequencies $f_i$ for each allele.

The crucial observation here is that since $\delta = 0$, if the queried individual is in the beacon it must be that the beacon responds "Yes" to all queries $q_i \in Q$ made by the adversary for a query individual. Let $R_{\text{yes}}$ denote the sequence of all "Yes" responses. We consider two cases:

- $R = R_{\text{yes}}$. One then easily obtains:

  $$L_{H_1}(R) = 0, \quad L_{H_0}(R) = \sum_{i=1}^{n} \log(1 - D_N^i), \quad \Lambda = \sum_{i=1}^{n} \log(1 - D_N^i) \ . \tag{S11}$$

- $R \neq R_{\text{yes}}$. Then, we have:

  $$L_{H_1}(R) = -\infty, \quad L_{H_0}(R) \in \mathbb{R}, \quad \Lambda = \infty \ . \tag{S12}$$

So we see that in any case, the random variable $\Lambda$ can only take on two values, either $\sum_{i=1}^{n} \log(1 - D_N^i)$ or $\infty$. Now, if $H_1$ is true, $R$ must be $R_{\text{yes}}$. Thus, we have that the distribution of $\Lambda$ under $H_1$ reduces to the constant $\sum_{i=1}^{n} \log(1 - D_N^i)$. If $H_1$ is true, the beacon responds "Yes" to query $q_i$ with probability $1 - D_N^i$. Thus, $\Pr[R = R_{\text{yes}}|H_0] = \prod_{i=1}^{n} (1 - D_N^i)$. Then, under $H_0$, $\Lambda$ is a random variable that takes value $\sum_{i=1}^{n} \log(1 - D_N^i)$ with probability $\prod_{i=1}^{n} (1 - D_N^i)$, and value $\infty$ otherwise. In summary:

$$\Lambda|H_1 = \sum_{i=1}^{n} \log\left(1 - D_N^i\right) \quad \text{with probability } 1, \tag{S13}$$

$$\Lambda|H_0 = \begin{cases} \sum_{i=1}^{n} \log\left(1 - D_N^i\right) & \text{with probability } \prod_{i=1}^{n} \left(1 - D_N^i\right), \\ \infty & \text{otherwise.} \end{cases} \tag{S14}$$

So the cut-off threshold $t$ must be chosen somewhere in $\left]\sum_{i=1}^{n} \log(1 - D_N^i), +\infty\right[$. According to the above, the power of the adversary will always be 1 (the adversary will never conclude that the victim is not in the beacon when she actually is). So our only control is over the false-positive rate $\alpha = \prod_{i=1}^{n} (1 - D_N^i)$. The goal of our strategy here is to dismiss an individual from consideration for any further query responses as soon as including her data would enable the adversary to construct a powerful re-identification test for that individual. By this, we mean a test with power 1 and false positive rate $\alpha \leq p$, for some chosen $p$. Our budget method sets $b_j = -\log(p)$ at first and then each time a query is made for an allele that an individual possesses, we first check whether the budget of the individual is larger than $-\log(1 - D_N^i)$, then reduce his/her budget by $-\log(1 - D_N^i)$. In this way we ensure that for each individual $j$, $\prod_{i \in Q^j} (1 - D_N^i) > p$, where $Q_j$ represents the subset of queries made for alleles that individual $j$ possesses, and for which individual $j$ was considered when constructing the response.

For simplicity, we consider here that an adversary that wishes to re-identify individual $j$ will only query SNPs for which $j$ possesses the alternate allele (assuming $\delta = 0$). Indeed, for a query for a variant that $j$ does not possess, we have $\Pr[x_i = 1|H_1] = 1 - D^i_{N-1}$ and $\Pr[x_i = 1|H_0] = 1 - D^i_N$, which are negligibly close for large $N$. Thus, such queries can simply be considered as useless for distinguishing $H_0$ from $H_1$.

**APPENDIX D: RESULTS ON OPTIMAL RE-IDENTIFICATION ATTACK IN MULTI-POPULATION BEACON**

Beacons often contain individuals coming from different ancestry groups. As a consequence, we further evaluated the attack based on real allele frequencies on a multi-population beacon and considered the case where an attacker might have only partial information about the different ancestries in the beacon. We set up a different beacon by removing individuals with European (EUR) ancestry from phase 3 data set of the 1000 Genomes Project, and by selecting $1{,}235$ random individuals from the remaining ones. The resulting population is composed by individuals with African (AFR), Ad Mixed American (AMR), East Asian (EAS) or South Asian (SAS) ancestries. We picked 100 random samples from the beacon and 100 random samples not in the beacon and not of EUR ancestry to compose the query set.

As expected, results in Figure S1 show that also in the multi-population beacon the new re-identification attack based on allele frequencies is more effective than the one by Shringarpure and Bustamante. Especially, when the attacker knows the allele frequencies for a population with the same mix of ancestries of the individuals in the beacon (blue curve), 5

queries on average[1] are enough to obtain 100% of statistical power with 5% false-positive rate. As expected, with the same background knowledge but by querying alleles in random order, the attacker needs 750 more queries (azure curve) to obtain the same statistical power. A more realistic scenario is represented by the attacker knowing partial (e.g., allele frequencies from a population with AFR ancestry) or unrelated information (e.g. allele frequencies from a population with EUR ancestry) about the ancestries in the beacon. In these cases, 100% of statistical power with 5% false-positive rate can be obtained with 20 (green curve) or 37 (red curve) queries, respectively.

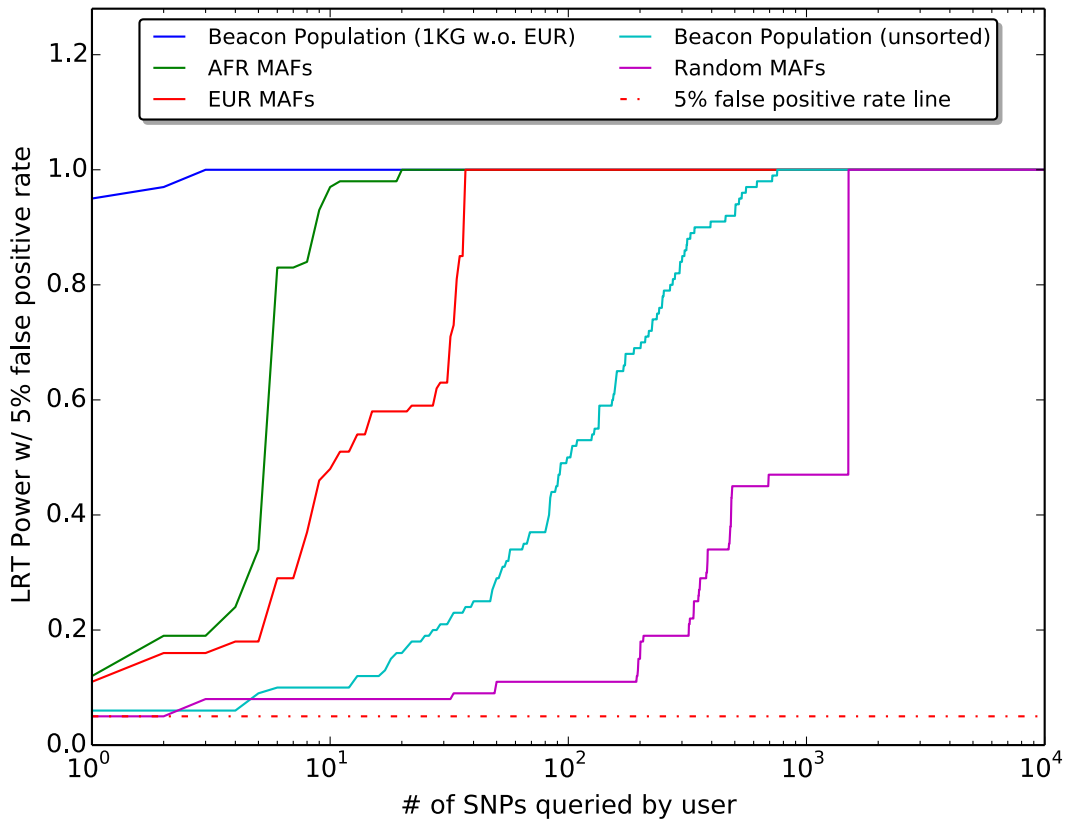


*Figure S1. Different power rates per number of SNPs queried from an unprotected multi-population beacon (the beacon contains individuals from all ancestry in the 1000 Genomes Project but the European ancestry) by an adversary with background knowledge on allele frequencies. Different colors represent different types of background knowledge.*

[1] The attack is repeated on 100 different individuals.

**APPENDIX E: MITIGATION STRATEGIES COMPUTATIONAL COMPLEXITY EVALUATION**

The first and second strategies induce very little overhead. The allele frequencies can be pre-calculated, which takes only linear time in the size of the database, and kept as a table in the database. Once $k$ or $\varepsilon$ is pre-determined, the beacon will just need to check if the query allele's frequency is smaller than $k$ (for strategies *S1* and *S2*) and to generate a random number (for *S2*) before composing a response of "*Yes*" or "*No*".

For mitigation strategy *S3*, we can easily compute the complexity of the proposed algorithm (see Algorithm 1 in the paper). Suppose there are *N* individuals in the dataset, then for a given query, we need to:

. Compute the risk of the query, which can be done in constant time *O(1)*
. Check whether there are individuals who have the queried allele and a budget higher than the risk, which can be done in linear time *O(N)*
. If there is no such person, answer "*No*", which can be done in constant time *O(1)*
. If there is at least one, answer "*Yes*", then reduce those people's budget by the risk. This can be done in linear time *O(N).*

So in total, the computational time required for each query on a beacon with mitigation strategy *S3* is linear with respect to the number of individuals in the beacon. We note that the order of the required time for *S3* is the same as if no-privacy-preserving mechanisms were imposed.

**REFERENCES**

1 Shringarpure SS, Bustamante CD. Privacy risks from genomic data-sharing beacons. The American Journal of Human Genetics. 2015 Nov 5;97(5):631-46.