Paper accepted for publication in the Journal of the American Medical Informatics Association (JAMIA)

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

Jean Louis Raisaro[1,*], Florian Tramèr[1,*], Zhanglong Ji[2,*], Diyue Bu[3,*], Yongan Zhao[3], Knox Carey[4], David Lloyd[5,9], Heidi Sofia[6], Dixie Baker[8,9], Paul Flicek[5,9], Suyash Shringarpure[7], Carlos Bustamante[7], Shuang Wang[2], Xiaoqian Jiang[2], Lucila Ohno-Machado[2], Haixu Tang[3], XiaoFeng Wang[3], Jean-Pierre Hubaux[1]

1 School of IC, EPFL, Lausanne, Switzerland

2 Department of Biomedical Informatics, UC San Diego, CA, US

3 School of Informatics and Computing, IU Bloomington, IN, US

4 GeneCloud, Intertrust, CA, US

5 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

6 Division of Genomic Medicine, NIH, MD, US

7 Department of Genetics, Stanford University, Stanford, CA, US

8 Martin, Blanck and Associates, Redondo Beach CA, US

9 Global Alliance for Genomics and Health

*These authors contributed equally to this work

Corresponding author:

Jean-Pierre Hubaux

Professor, School of Computer and Communication Sciences

École Polytechnique Fédérale de Lausanne

Station 14, EPFL IC ISC LCA1, BC 207

1015 Lausanne, Switzerland

Ph: +41 21 693 2627

Email: jean-pierre.hubaux@epfl.ch

Key words: genomic privacy, ga4gh, beacon, re-identification, genomic data sharing

Main text word count: 4154

ABSTRACT

The Global Alliance for Genomics and Health (GA4GH) created the Beacon Project as a means of testing the willingness of data holders to share genetic data in the simplest technical context – query for the presence of a specified nucleotide at a given position within a chromosome. Each participating site (or “beacon”) is responsible for assuring that genomic data are exposed through the Beacon service only with the permission of the individual to whom the data pertains, and in accordance with the GA4GH policy and standards.

While recognizing the inference risks associated with large-scale data aggregation, and the fact that some beacons contain sensitive phenotypic associations that increase privacy risk, the GA4GH adjudged the risk of re-identification based on the binary yes/no allele-presence query responses as acceptable. However, recent work demonstrated that, given a beacon with specific characteristics (including relatively small sample size, and an adversary who possesses an individual’s whole genome sequence), the individual’s membership in a beacon can be inferred through repeated queries for variants present in the individual’s genome.

In this paper, we propose three practical strategies for reducing re-identification risks in beacons. The first two strategies manipulate the beacon such that the presence of rare alleles is obscured; the third strategy budgets the number of accesses per user for each individual genome. Using a beacon containing data from the 1000 Genomes Project, we demonstrate that the proposed strategies can effectively reduce re-identification risk in beacon-like datasets.

INTRODUCTION

The Global Alliance for Genomics and Health (GA4GH) [1] conceived the Beacon Project as a means of testing the willingness of international sites to share genomic data in the simplest of all technical contexts: A public web service that any data holder could implement to enable users to submit queries of the form, “Do you have any genomes with nucleotide A at position 100,735 on chromosome 3?”, to which the service would respond with “Yes” or “No”. A site offering this service is called a beacon and is responsible for assuring that genomic data are exposed through the Beacon service only with the permission of the individual to whom the data pertain, and in accordance with the GA4GH ethical *Framework* [2] and *Privacy and Security Policy* [3]. Thus the Beacon service is designed to be technically simple, easy to implement, and privacy protective.

The availability of vast quantities of high-quality genomic and health data is essential to the advancement of biomedical knowledge. Yet, privacy concerns often limit researchers’ ability to access potentially identifiable health data. Indeed, in some cases, privacy laws and regulations actually impede individuals’ ability to make their own data available to researchers [4]. This problem is particularly acute in the field of genomics, where the vast majority of variants predicted to be functionally important are extremely rare, occurring in less than 0.5% of the population [5]. As a result, it is unlikely that any single institution will hold enough data to achieve sufficient statistical power in studying any particular condition. Recognizing the urgent need for federation across organizations, the GA4GH was formed in 2013 to enable responsible sharing of genomic and health-related data by establishing consistent policy, and interoperable standards and protocols.

From its inception, the GA4GH has been committed to achieving a responsible and effective balance between data sharing and individual privacy, a challenge that has been extensively explored in the literature [6–9]. In 2008, Homer et al. [6] showed that statistical techniques can reveal the presence or absence of an individual in a genomic data set, even when the targeted individual's genome accounts for less than 0.1% of the total data. The publication of this paper had a significant impact, prompting several major institutions including the Wellcome Trust and the US National Institutes of Health to limit public access to data formerly adjudged to be safely anonymous [10]. As this scenario demonstrates, privacy concerns can undermine the ability of researchers to publish and access genomic data.

At the outset, the GA4GH recognized that the Beacon approach can reveal information about the individuals in a data set. However, in performing the risk assessment, the GA4GH recognized several conditions that served to mitigate the risk that any individual would be identified based on Beacon search. First, the Beacon user-interface is extremely restrictive, enabling query only for the presence or absence of the four nucleotides (A, C, T, G) that comprise every individual's genome. Second, the number of individual genomes aggregated in each beacon is very large. Third, for a data seeker to be able to identify an individual through Beacon queries would require as a pre-condition that the data seeker possess a significant amount of genomic data associated with the targeted individual, such as a variant call format (VCF) file of the individual's whole genome sequence. In such case, a potential adversary would know all variants in the individual's genome, and would have much more efficient means of discovering a disease association than persistent beacon queries. Thus GA4GH concluded that the risk of a data seeker identifying an individual through Beacon queries was acceptably low, even for the case of a data seeker willing to violate GA4GHs ethical standards.

However, Shringarpure and Bustamante [11] describe an attack in which an anonymous adversary, even with knowledge of only a small portion of a target's genome can successfully launch a re-identification attack: In a beacon comprising 1,000 individuals for instance, 5,000 queries suffice. Such an attack relies on a likelihood ratio test whose power is a function of the responses returned by the beacon, the size of the data set, the allele-frequency spectrum, and the sequencing error rate. Their paper demonstrates that under certain conditions, the anonymous-access model implemented by the Beacon Project does not prevent identification of individuals whose genomes could be exposed through a Beacon interface.

The goal of this paper is to further examine the potential vulnerabilities and risks associated with the Beacon model, and to explore ways of mitigating re-identification risks – thus enhancing Beacon privacy protections. Re-identification is the process by which anonymized personal data is matched with its true owner [12]. We first analyze the re-identification threat described by Shringarpure and Bustamante and the vulnerability the attack exploited. We then propose an optimized version of the attack that considers an adversary with some background knowledge about the allele frequencies in the targeted beacon. We describe three potential strategies for mitigating the risk of re-identification and assess their effectiveness through several experiments with data obtained from the 1000 Genomes Project [13]. We conclude the paper by discussing the strengths and weaknesses of the proposed strategies and by providing some recommendations for strengthening Beacon privacy protections.

MATERIALS AND METHODS

Original Re-Identification Attack

We begin by describing the re-identification attack proposed by Shringarpure and Bustamante [11]. In the following we refer to it as the "SB attack."

As noted earlier, the setting of the SB attack is similar to that of previous works such as that of Homer et al. [9]. The attacker is assumed to have access to the VCF file of a target victim's genome and queries the beacon at heterozygous positions to determine whether the victim is in the beacon or not. The SB attack relies on a likelihood-ratio test (LRT) that evaluates the likelihood of the beacon's responses under two possible hypotheses:

- The null hypothesis H_0: The queried victim's genome is not in the beacon.
- The alternative hypothesis H_1: The queried victim's genome is in the beacon.

The re-identification risk is measured by the power of such a test, i.e., $Pr(reject\ H_0 \mid H_1\ true)$. To make their test as general as possible, Shringarpure and Bustamante assume only that the attacker knows the beacon size N, as well as the site frequency spectrum of the beacon population. Formally, the alternate allele frequency f_i of a heterozygous SNP observed in the population is assumed to be distributed as $f_i \sim \mathrm{beta}(a, b)$ for population parameters a, b. Their LRT further allows for a probability δ of *sequencing errors*, resulting in a mismatch between the attacker's copy of a genome and the copy in the beacon.

Given a set of beacon responses $R = \{x_1, \dots, x_n\}$, the log-likelihood of the sequence is

$$L(R) = \sum_{i=1}^{n} x_i \log\Pr(x_i = 1) + (1 - x_i) \log\Pr(x_i = 0) \ . \tag{1}$$

Under H_1, let D_{N-1}^{i} denote the probability that none of the $N-1$ other genomes in the beacon have an alternate allele at position i. Similarly, under H_0 we denote by D_N^i the probability that none of the N genomes in the beacon have an alternate allele at i. Then, under the two hypotheses, we have

$$L_{H_1}(R) = \sum_{i=1}^{n} x_i \log(1 - \delta D_{N-1}^{i}) + (1 - x_i)\log(\delta D_{N-1}^{i}) \ , \quad (2)$$

$$L_{H_0}(R) = \sum_{i=1}^{n} x_i \log(1 - D_N^i) + (1 - x_i)\log(D_N^i) \ . \quad (3)$$

Shringarpure and Bustamante show that under their assumptions, for any position i we have $D_{N-1}^{i} = \mathbb{E}[p_i^{2N-2}]$ and $D_N^i = \mathbb{E}[p_i^{2N}]$, where $p_i \sim \text{beta}(b, a)$. The log of the LRT is given by

$$\Lambda = L_{H_0}(R) - L_{H_1}(R) = nB + C \sum_{i=1}^{n} x_i \ , \quad (4)$$

where B and C are constant for N, δ, a, b fixed. Thus, $\sum_{i=1}^{n} x_i$ (the number of "*Yes*" responses from the beacon) is a sufficient statistic for the LRT.

"Optimal" Attack with Real Allele Frequencies

The SB attack removes direct dependency on allele frequencies and sets conservative bounds for the number of queries required for successful re-identification. We consider here a more capable and determined attacker who has access to some background knowledge on allele frequencies and optimizes his attack by querying the rarest alleles in the victim's genome first. In other words, similarly to best practices in forensics, the attacker makes use of alleles with maximum re-identification power instead of performing random requests. This assumption appears reasonable in practice, as allele frequency information for different ancestries is already publicly available on the Web (e.g., 1000 Genomes Project [14], HapMap Project [15], etc.) and easily accessible even by non-expert attackers. We show through several experiments (see Results Section) that this new attack is significantly more powerful than the original SB attack, even when the attacker has incomplete knowledge on allele frequencies in the beacon.

Formally, the attacker assumes allele frequencies $f_1, f_2, \dots, f_M$ for the M SNPs in the victim's genome. Without loss of generality, we assume the frequencies are already ordered (i.e, $f_1 \leq f_2 \leq \cdots \leq f_M$). Then, the attacker will maximize his re-identification power by first querying those SNPs which are least likely to appear in the beacon under H_0, specifically those with lowest frequency. In this setting, Equations (2) and (3) still hold, but the computation of D_{N-1}^{i} and D_N^i is different. Under the alternative hypothesis, we have

$$\begin{aligned} D_{N-1}^{i} &= \Pr(\text{none of the other N} - 1 \text{ genomes have an alternate allele at position i}) \\ &= ((1-f_i)^2)^{N-1} \\ &= (1-f_i)^{2N-2} \ . \end{aligned}$$

Similarly, under H_0 we have $D_N^i = (1-f_i)^{2N}$.

As the probabilities D_{N-1}^{i} and D_N^i now directly depend on the position i, we have that the following LRT

$$\begin{aligned} \Lambda &= L_{H_0}(R) - L_{H_1}(R) \\ &= \sum_{i=1}^{n} \log\left(\frac{D_N^i}{\delta D_{N-1}^i}\right) + \log\left(\frac{\delta D_{N-1}^i (1-D_N^i)}{D_N^i (1-\delta D_{N-1}^i)}\right) x_i \\ &= \sum_{i=1}^{n} \log(\delta^{-1}(1-f_i)^2) + \log\left(\frac{\delta}{(1-f_i)^2} \cdot \frac{1-(1-f_i)^{2N}}{1-\delta(1-f_i)^{2N-2}}\right) x_i \ . \end{aligned} \quad (5)$$

We will evaluate the power of this test empirically, through experiments in a variety of settings with real data and different levels of adversarial background knowledge. We will estimate the null distribution of the LRT by computing Equation (5) for a number of control individuals known not to be in the beacon. The null hypothesis is rejected if $\Lambda < t$ for some threshold t. We then let t_α be such that $\Pr[\Lambda < t_\alpha | H_0] = \alpha$. The power of the test is computed as $1 - \beta = \Pr[\Lambda < t_\alpha | H_1]$, where the distribution of Λ given H_1 is estimated by querying individuals in the experimental beacon.

Risk Mitigation Strategies

Based on the "optimal" version of the re-identification attack, we propose three different practical strategies to mitigate the risk. Without loss of generality, we can assume that any defense mechanism that effectively mitigates the "optimal" re-identification attack also effectively mitigates the original SB attack. Our experimental results (see Results section) show the validity of this assumption.

Beacon Alteration Strategy

The first strategy (*S1*) relies on the observation that most of the statistical power in the re-identification attack comes from queries targeting *unique* alleles in the beacon. In particular, *S1* alters the beacon by answering a query with *"Yes"* only if there are at least $k > 1$ individuals sharing the queried allele. In other words, k is the minimum number of individuals in the beacon sharing the queried allele when returning *"Yes"*. Current beacons set $k = 1$, i.e., when there are one or more individuals in the population with the queried allele, the answer will be *"Yes"*. We assume the value of k is made public, hence the attacker will modify the attack to accommodate this change (see Appendix A for LRT under *S1*). Yet, already for $k = 2$ we found that in practice what the attacker can infer is limited (see Results Section).

Random Flipping Strategy

The second strategy (*S2*) relies on the same observation but instead of altering the beacon response, it introduces noise into the original data. The disadvantage of *S1* is that only a subset of variations (e.g., the *non-unique* SNPs when $k = 2$) in the beacon population can be queried. In practice, *unique* alleles that are likely to be the most useful in human genetics research, are completely hidden. *S2* improves the usability of the beacon over *S1* as it hides only a portion ε of unique alleles, but not all. In other words, a beacon with *S2* will add noise by sampling from a binomial distribution with probability ε only to unique alleles in the database and provide false answers (e.g., "*No*" instead of "*Yes*") to queries targeting these unique alleles. The main goal of *S2* is to share as many unique alleles as possible while reducing the likelihood that the information released will be sufficient to re-identify an individual in the database. We assume the value of ε is public. As for *S1*, the attacker will adapt the LRT statistic to take it into account (see Appendix B for LRT under *S2*).

Query Budget per Individual Strategy

The third strategy (*S3*) mitigates the re-identification risk by assigning a budget to every individual in the database; this budget is applied to each authenticated Beacon user. With respect to strategies *S1* and *S2*, *S3* leverages two additional assumptions:

- Each Beacon user has been identity proofed, holds a single account, is authenticated, and does not collude. If users are allowed to collude, then to be effective, *S3*, will have a dramatic impact on the utility of the system. This assumption appears reasonable in practice as, in order to collude, a user needs by definition to involve someone else. We assume that each user holds a single Beacon account to eliminate the possibility of a single user simulating multiple profiles in collusion, which carries higher risk than either collusion among multiple users or a re-identification attack that can be undertaken at an individual scale. This is because an attack involving multiple accounts, all working on behalf of a single attacker, does not require exchanging files with other users, and could be conducted more quickly than a single-threaded attack.

- The attacker has accurate genomic information, which means $\delta = 0$. This is a worst-case assumption because, if we can prevent re-identification under this condition, we can prevent against the proposed "optimal" attack, too. Note that in practice, as there are some sequencing errors (i.e., $\delta > 0$), the attacker will actually have less power. Hence, this approach is conservative from a re-identification point of view. Moreover, by assuming $\delta = 0$, we can significantly simplify the analytical treatment of the problem.

The basic idea is that each time an individual's genome contributes to a *"Yes"* answer for a given query (i.e., the individual has the queried allele), her corresponding budget for that Beacon user is reduced by an amount that depends on the frequency of the queried allele. If her budget is less than this amount, her information will not be used to answer that query and the individual will be removed from the dataset, as shown in Algorithm1 in Table 1. In this way, the privacy of the individual will be always preserved at a cost of a slight decrease of utility.

Table 1. Algorithm describing mitigation strategy S3

```
Algorithm1
Requires: upper bound on test errors $p$
1. Set all $b_j = -\log(p)$
2. Receive $i$-th query and check whether it has been asked before. If yes, go to Step 3. If no, go to Step 4
3. Return the previous answer, then go to Step 2.
4. Compute the risk $r_i = -\log(1 - D_i^N)$.
5. Check whether there are any records with the asked variant and $b_j > r_i$. If no, return no and go to Step 2.
6. For all the individuals with such variant and $b_j > r_i$, reduce their budgets by $r_i$. Then return yes.
7. Go back to Step 2 and wait for the next query.
```

Let R be the set of responses of the beacon, the goal of *S3* is to keep track of the power of the attack which is based on the LRT $\Lambda = L_{H_0}(R) - L_{H_1}(R)$, in order to prevent any individual genome from contributing to a query response that can leak identity information with high confidence (see Appendix C for formal description of *S3*).

Experiments with Real Data

To evaluate the effectiveness of the proposed strategies in reducing risk under the "optimal" attack with real allele frequencies, we designed and ran several experiments on real data with the following setup. We created a beacon composed of 1,235 samples of chromosome 10 randomly chosen from the 2,504 individuals in phase 3 of the 1000 Genomes Project [13]. A total of 31 relatives were removed. The resulting data set consists of individuals with either European, African, admixed American, East Asian or South Asian ancestries. Among these samples, 100 were selected as the control set. Similarly, from the remaining individuals not in the beacon, 100 were selected as the test set.

The null distribution of the LRT statistic was obtained through the exact-test computation on the 100 individuals in the test set (i.e., not in the beacon). With a false positive rate of $\alpha = 5\%$ we computed the power $(1 - \beta)$ as the proportion of test rejected (i.e., when $\Lambda < t_{\alpha}$) for the control set (i.e., how many individuals in the control set, hence in the beacon, were successfully re-identified).

RESULTS

"Optimal" Re-Identification Attack in Single-Population Beacon

We evaluated the re-identification power of our attack on a beacon composed by individuals coming from the same ancestry group. From phase 3 of the 1000 Genomes Project, we selected 502 samples of European (EUR) ancestry and we randomly picked half of them to set up the beacon. The remaining half was used to compute the EUR population allele frequencies. We considered several scenarios where the attacker has different types of background information.

As expected, results in Fig.1 show that the worst case scenario is represented by an attacker knowing the exact ancestry of the population in the beacon. With only 3 SNPs, beacon membership could be re-identified with 100% power and 5% false positive rate. Yet, as the beacon ancestry information is not always public, a more realistic scenario is to consider an attacker that only knows the allele frequencies of a random population possibly from a different ancestry than the one of the beacon. Even with the least precise background information (in this case the allele frequencies from EAS ancestry), 36 SNPs are sufficient to re-identify an individual. Fig.2 shows the Kendall rank correlation coefficient [16] between the actual allele frequencies in the beacon and the allele frequencies from different ancestry groups. By combining the information in Fig.1 and Fig.2 it is easy to observe that the higher the ordinal association is between the beacon allele frequencies and the allele frequencies known by the attacker, the fewer queries are needed to re-identify with 100% power and 5% false positive rate (see Appendix D for results on the "optimal" attack in multi-population beacon).

"Optimal" Re-Identification Attack in Beacon with *S1*

We evaluated the proposed solution *S1* by considering an attacker who knows the allele frequencies of the 1000 Genomes Project and the value of threshold parameter k. As such, we set up a beacon as described in Section *Materials and Methods* and computed the LRT statistic as described in Appendix A. Fig.3 shows that, under such an attack, no individual in the beacon can be re-identified if a "*Yes*" answer is provided only when the queried allele appears at least $k = 2$ times in the database. Yet, the downside of this method is that only a fraction of the alleles that are in the beacon can be shared. For example, in our experimental beacon, only 60% of the alleles are shared by two or more individuals and thus can be shared; the queries to the remaining rare alleles ($\approx 40\%$) will receive a "*No*" answer even though they are actually present in the Beacon database.

"Optimal" Re-Identification Attack in Beacon with *S2*

To evaluate the effectiveness of *S2* against an attack with background knowledge on allele frequencies, we consider an attacker who knows the allele frequencies of the 1000 Genomes Project and the value of the parameter ε. Fig.4 shows how the statistical power of the attacker decreases when different portions (ε) of unique alleles are hidden. When ε is set to be 0.001, the attacker has to query around 10^4 unique alleles to obtain a strong power of re-identification, compared to 200 queries for 100% re-identification when no random flipping on unique alleles (of $S2$ strategy) is applied. When ε is set to be equal or greater 0.15, the re-identification power will not increase above $\approx 30\%$, which will keep the power at an acceptable risk level (i.e., relatively low confidence of re-identification).

Budget Evaluation in Beacon with *S3*

We evaluated strategy *S3* with the same experimental setting as for *S1* and *S2*. By default, we set $p = 0.05$, which means the statistical power of attack cannot exceed 0.95. Differently from experiments performed on solutions *S1* and *S2*, which show an increase in re-identification risk given certain levels of utility of the beacon, we evaluate the efficacy of *S3* by computing the decrease of utility across queries for a certain level of re-identification risk. To this purpose, we emulate the query behavior of a typical honest beacon user by generating queries based on the distribution of query frequency per allele frequency extracted from ExAC browser [17] logs over a period of 12 weeks.[1] During this time frame, a total of $1{,}345{,}291$ queries were asked on $934{,}680$ variants present in ExAC. Table 2 shows the proportion of queries and allele per range of allele frequencies (AF).

Fig.5 shows how the number of individuals with enough budget decreases with respect to the number of queries answered by the beacon. Note that the beacon's utility is completely preserved for the first 2,000 queries.

Table 2. Proportions of queries (over a period of 12 weeks) for each range of allele frequency.

Allele frequency	<0.001	0.001~0.01	0.01~0.05	0.05~0.5	>0.5
Queries in ExAC	0.853	0.0.076	0.023	0.033	0.014

[1] Data on beacon query frequencies were not available at the time of this work.

DISCUSSION

In this paper, we have analyzed in detail the beacon re-identification attack originally proposed by Shringarpure and Bustamante and a new and “optimal” version of it by considering a smarter adversary who makes use of public information on allele frequencies. We evaluated the power of our new attack through several experiments on real data by considering different conditions of adversarial background knowledge. Our results show that our attack always outperforms the original SB attack. As one might expect, we have observed that the power of an adversary’s re-identification attack is directly related to the completeness and accuracy of the adversary’s knowledge of the allele frequencies of the targeted beacon. As already analyzed by Shringarpure and Bustamante the underlying LRT test can be extremely harmful when a beacon is linked to sensitive phenotypes. Yet, it is important to emphasize that, although our attack further reinforces SB’s concern, the re-identification risk is relative to each beacon. These attacks fundamentally rely on the assumption that the attacker already has access to the genome of the victim.

Despite such a strong assumption, several research efforts in genomic privacy have studied the problem of re-identification of membership in genetic databases and have shown that this is extremely hard to prevent and sometimes even impossible [18].

Based on the “optimal” re-identification attack we have proposed three different strategies aimed at effectively thwarting beacon membership re-identification. As the accuracy of the beacon re-identification attack depends on the power and false positive rate of the LRT test, the probability that a test behaves correctly (rejecting the null hypothesis when it is false and failing to reject when it is true) is given by *Power*(Probability of alternative hypothesis)+(1-False positive rate)*(Probability of null hypothesis)*. From the perspective of a beacon administrator, the attacker's test should be incorrect most of time, i.e., power should be low and/or false positive rate should be high.

The three proposed strategies all address the mitigation problem by controlling the power or the false positive rate. The first (*S1*) and the second (*S2*) strategies reduce power to nearly zero when the LRT must have a small false positive rate, whereas in the third solution (*S3*), the test always has 100% power but a high false positive rate. In particular, *S1* and *S2* directly alter the beacon to reduce the inference power of the attacker whereas *S3* introduces a new idea of personal budget that decreases when the genome of the individual is used to positively answer a query.

Results of our experiments have shown that all proposed mitigation strategies have advantages and disadvantages, as summarized in Table 3. *S1* effectively mitigates the attack by keeping the power of the LRT to 0.2 if all unique alleles are flipped. Yet, it generates a significant loss in utility of the beacon, as the majority of the queries of a typical user of beacon usually target rare alleles. We define the utility of a beacon as the proportion of true answers it can provide. *S2* can be considered as a more sophisticated version of *S1* because it only flips a portion of unique alleles affording a more fine-grained control over the utility vs. privacy trade-off. The attack inference power can be confined to a secure level by flipping only 15% of unique alleles (which means a drop in utility of 6% against 40% of *S1*). Note that the utility of a beacon adopting either *S1* or *S2* is fixed a priori and does not change along with the power of the attack.

Finally, results of experiments on *S3* show that, given a certain assurance level ($p = 0.05$), the beacon utility is completely preserved for the first $2{,}000$ queries. Yet, *S3* relies on the assumption that the beacon system is not anonymous and has a controlled level of access with user authentication and identity proofing. Based on data collected from the ExAC browser logs, a budget of 2,000 query per beacon user seems a reasonable compromise between re-identification risk and utility.

Table 3. Summary of advantages and disadvantages of the three proposed mitigation strategies.

Risk Mitigation Strategy	Disadvantages	Advantages
S1: Beacon Alteration	Eliminates possibility of querying for unique alleles highly likely to be most useful in genetic research	Protects privacy of individuals possessing variants most likely to be targeted by attackers
S2: Random Flipping	Decreases rate of true answers returned from querying unique alleles likely to be useful in genetic research	Permits some unique alleles to be discoverable and to fine-tune the privacy-utility trade-off
S3: Query Budget per Individual	Requires the assumption of Beacon user being non-anonymous and holding no more than one Beacon account; may require complicated accounting scheme	Enables all alleles to be discoverable until budget is exceeded

Preventing inference attacks on large databases is widely known to be one of the most daunting of database security challenges [19]. This fact has been a major consideration in the development of GA4GH's *Framework for Responsible Sharing of Genomic and Health-Related Data*, *Privacy and Security Policy*, and *Security Infrastructure*. Effective risk management must leverage policy, technology, and community governance to address re-identification risks. Effective risk management is fundamental to facilitating and promoting data sharing across the GA4GH global community. We emphasize that security and privacy are components of risk management. Technical risk-management strategies such as those proposed in this paper, are practical and can be adapted according to the context of each beacon. Therefore, they represent a valuable set of options for assessing and mitigating risk within the GA4GH community.

FUNDING

This work was supported by NIH/NHLBI grant number [U54HL108460], NIH/NHGRI grant numbers [R01HG007078, K99HG008175, R00HG008175)] and NIH/NLM grant numbers [T15LM011271, R00LM011392, R21LM012060].

COMPETING INTERESTS

L. O-M. is editor-in-chief of JAMIA. The other co-authors have no competing interests to declare.

CONTRIBUTORS

Conception and design: JLR, FT, ZJ, DB, YZ, KC, SW, XJ, HT, XFW and JPH. Algorithms implementation: JLR, FT, ZJ and DB. Data analysis and interpretation, writing of manuscript, final approval of manuscript: JLR, FT, ZJ, DB, YZ, KC, DL, HS, DB, PF, SS, CB, SW, XJ, OML, HT, XFW and JPH. JLR, FT, ZJ, DB share first co-authorship.

ACKNOWLEDGMENTS

The authors would like to thank the GA4GH for its continuous support, Konrad Karczewski and Marc Duby from the Broad Institute for their valuable feedback and for having provided the data on query access patterns for the ExAC browser.

REFERENCES

1 Global Alliance for Genomics and Health. [Online]. Available: https://genomicsandhealth.org. Access date: May 2016

2 Framework for responsible sharing of genomic and health-related data. 2014. [Online]. Available: https://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data . Access date: May 2016

3 GA4GH privacy and security policy. 2015. [Online]. Available: https://genomicsandhealth.org/work-products-demonstration-projects/privacy-and-security-policy. Access date: May 2016

4 Terry SF, Shelton R, Biggers G, Baker D, Edwards K. The haystack is made of needles. Genetic testing and molecular biomarkers. 2013 Mar 1;17(3):175-7.

5 Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. science. 2012 Jul 6;337(6090):64-69.

6 Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008 Aug 29;4(8):e1000167.

7 Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. Nature genetics. 2009 Sep 1;41(9):965-7.

8 El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PloS one. 2011 Dec 2;6(12):e28071.

9 Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013 Jan 18;339(6117):321-4.

10 Greenbaum D, Sboner A, Mu XJ, Gerstein M. Genomics and privacy: Implications of the new reality of closed data for the field. PLoS Comput Biol. 2011 Dec 1;7(12):e1002278.

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

12 EPIC, https://epic.org/privacy/reidentification/

13 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65.

14 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68-74.

15 Gibbs RA, Belmont JW, Hardenbol P, et al. The international HapMap project. Nature. 2003 Dec 18;426(6968):789-96.

16 Kendall MG. Rank correlation methods. 1948

17 Consortium, Exome Aggregation and Others. ExAC Browser. 2015. [Online]. http://exac.broadinstitute.org/. Access date: May 2016

18 Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nature Reviews Genetics. 2014 Jun 1;15(6):409-21.

19 Adam NR, Worthmann JC. Security-control methods for statistical databases: a comparative study. ACM Computing Surveys (CSUR). 1989 Dec 1;21(4):515-56.

FIGURES

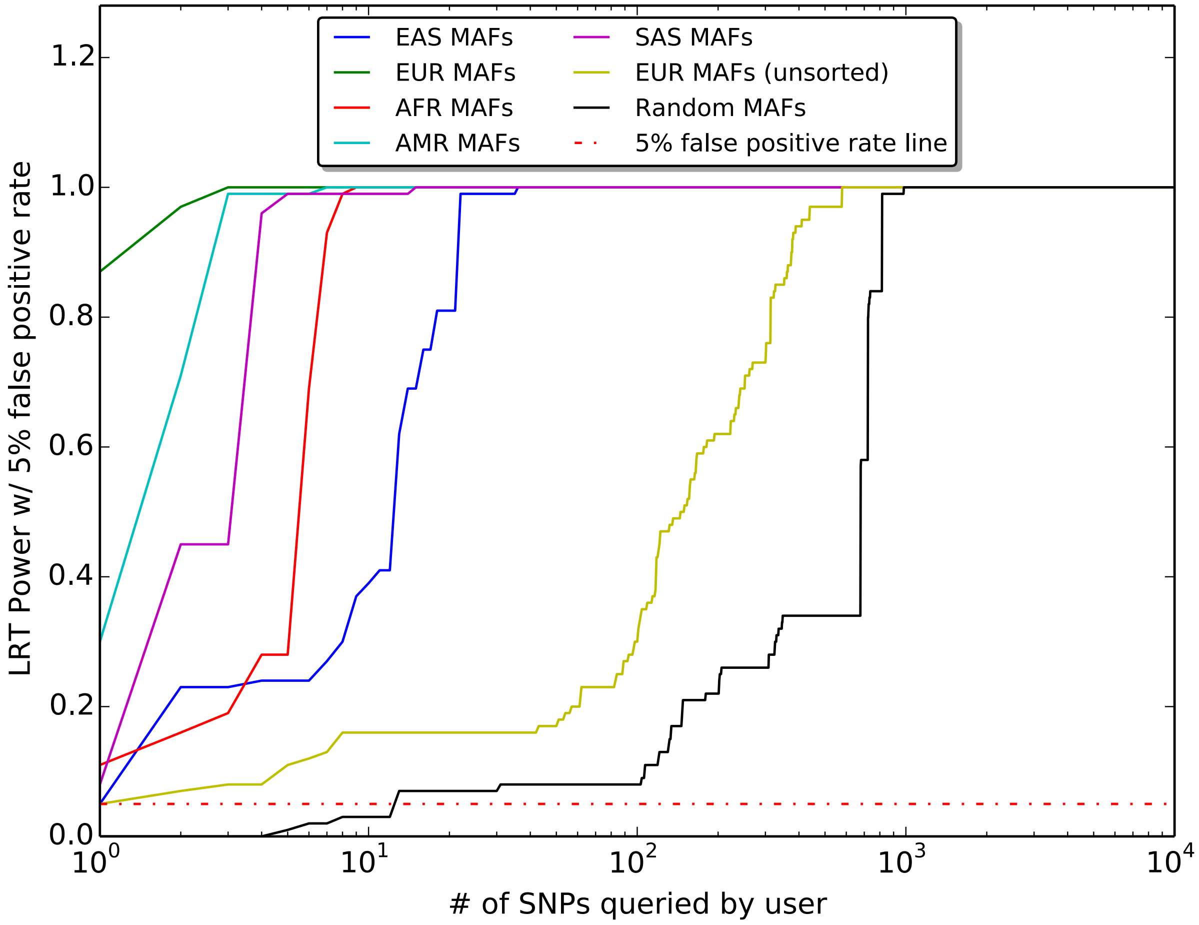

Figure 1: "Optimal" Re-Identification Attack in Single-Population Beacon. Different power rates per number of SNPs queried from an unprotected beacon with a single population (EUR) by an adversary with different types of background knowledge: (Green) The attacker knows the allele frequencies of a population from the same ancestry (EUR) as the one in the beacon and performs queries following the rare-allele-first logic; (Red, Cyan, Blue and Purple) The attacker knows the allele frequencies of a population from an ancestry different from the one in the beacon and performs queries following the rare-allele-first logic (African (AFR), admixed American (AMR), East Asian (EAS) or South Asian (SAS), respectively); (Yellow) The attacker knows the allele frequencies of a distinct population with the same ancestry (EUR) other than the one in the beacon but performs queries in random order; (Black) The attacker

does not have any information on allele frequencies (i.e., the original attack by Shringarpure and Bustamante [11]).

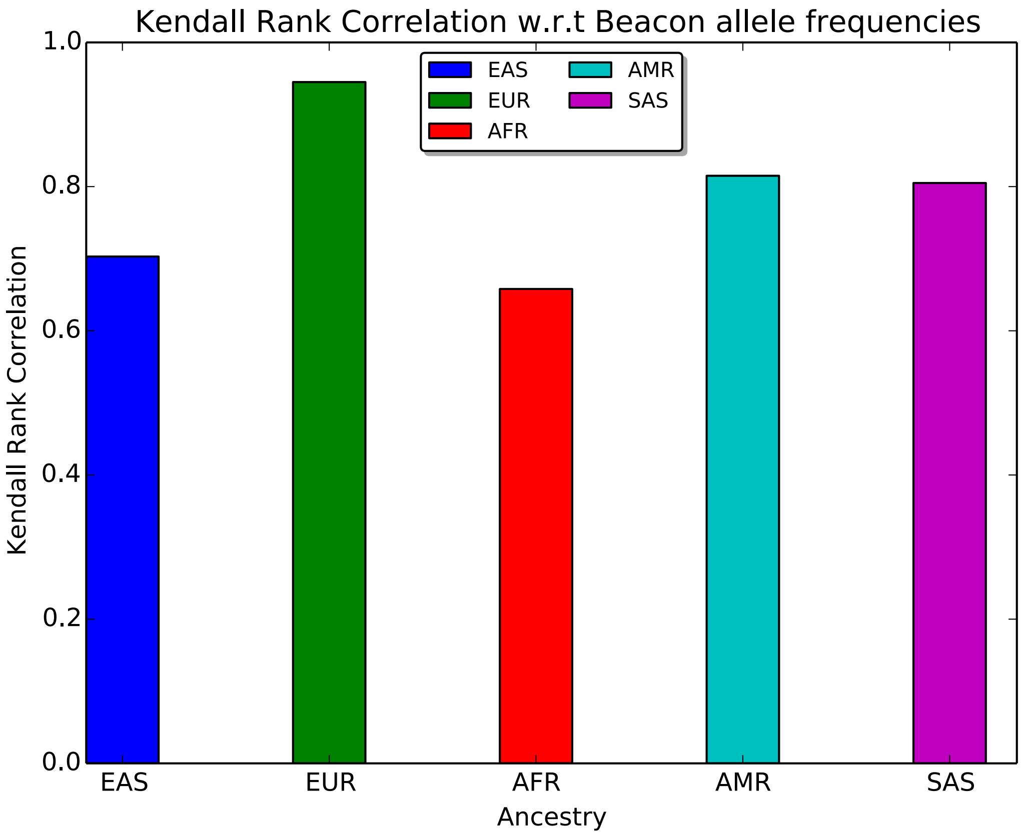

Figure 2: Kendall Rank Correlation Coefficient with respect to true beacon allele frequencies. Kendall rank correlation coefficient between the actual allele frequencies of the single-population beacon of Fig.1 and the allele frequencies of populations with different ancestries. Values closer to 1 represent higher correlation. Colors mapping as in Fig.1

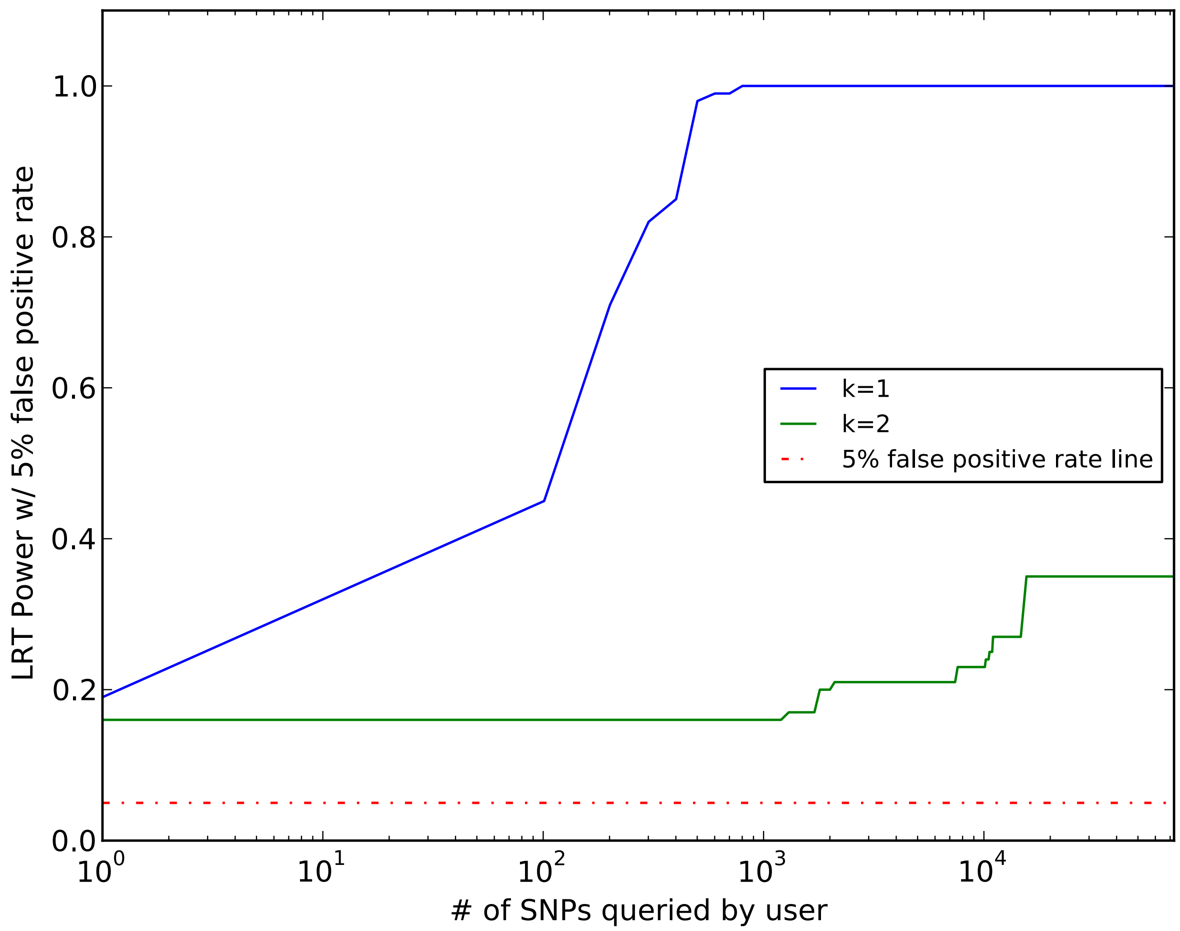

Figure 3: "Optimal" Re-Identification Attack in Beacon with *S1*. Different power rates per number of SNPs randomly queried from a beacon with mitigation *S1* by an adversary with knowledge on k and on allele frequencies from the 1000 genomes project: (Blue) $k = 1$; (Green) $k = 2$.

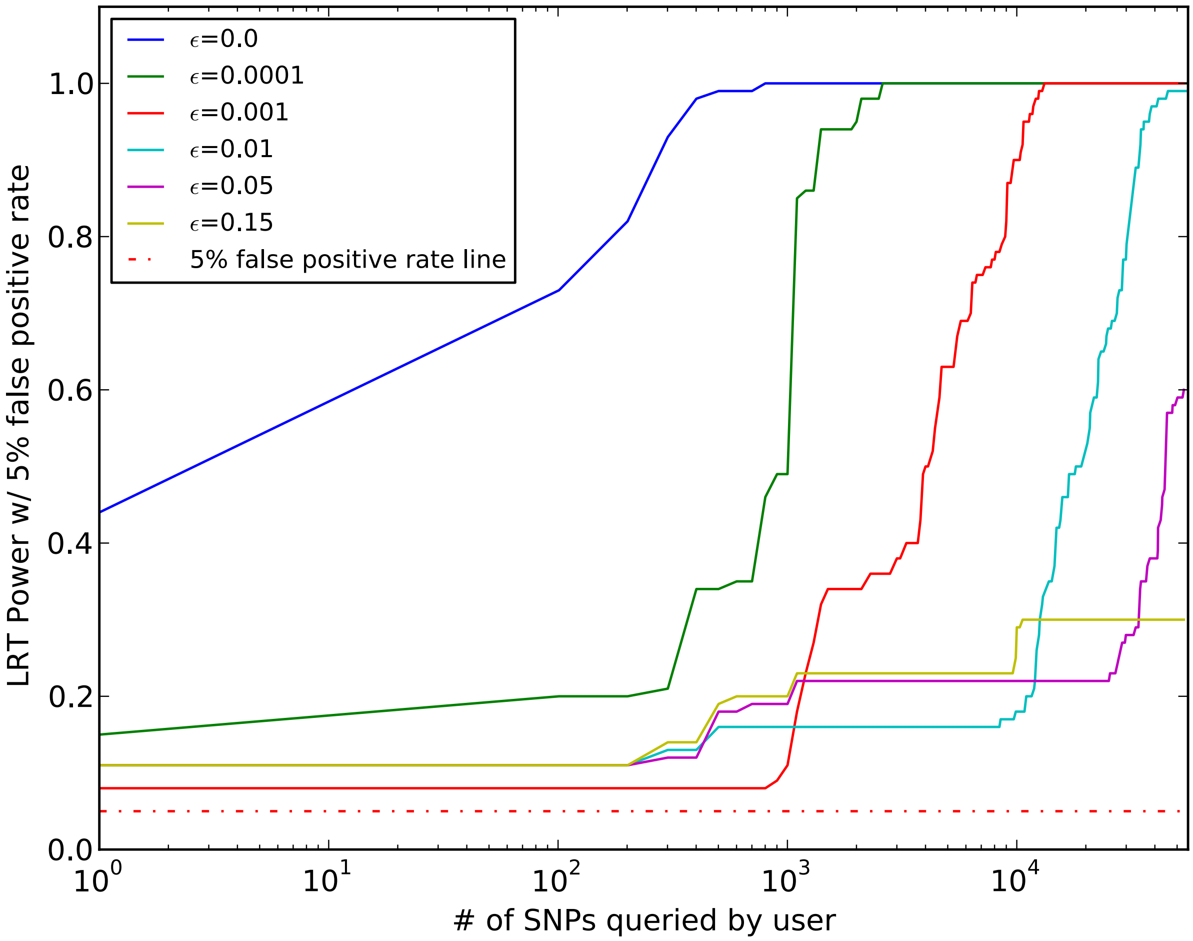

Figure 4: "Optimal" Re-Identification Attack in Beacon with *S2.* Different power rates per number of SNPs queried (with rare-first logic) from a beacon with mitigation *S2* by an adversary with knowledge on ε and on allele frequencies from the 1000 genomes project. Different colors for different values of ε.

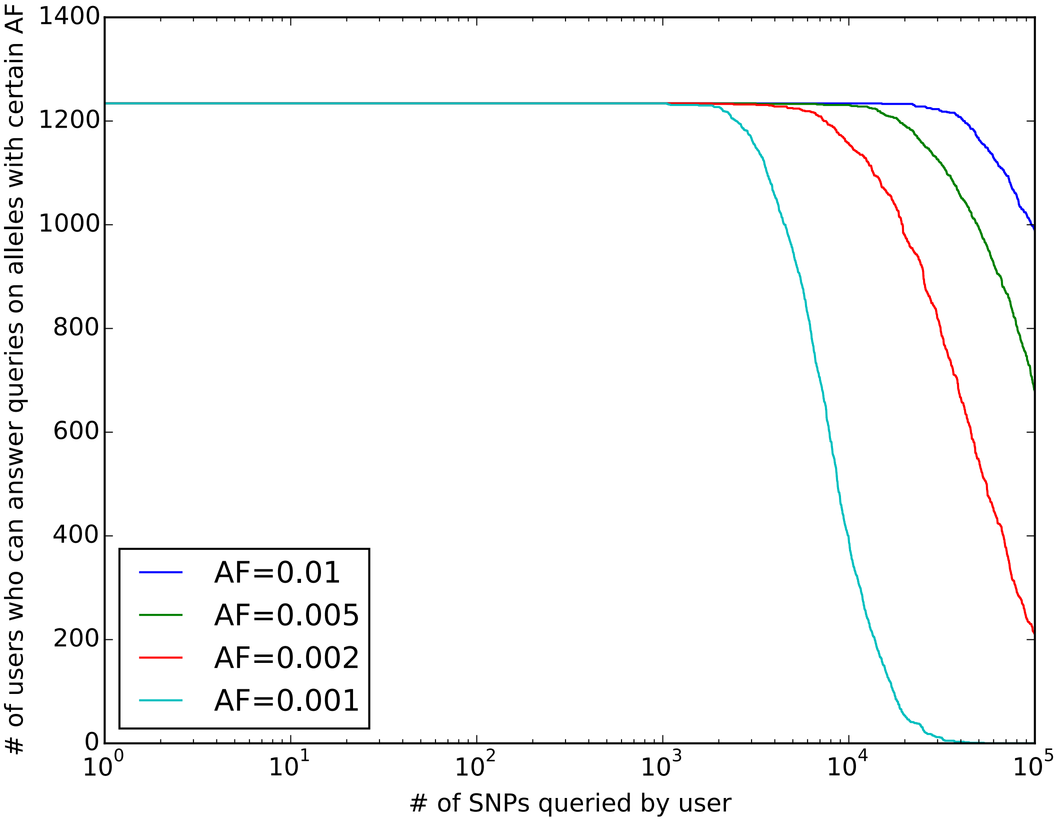

Figure 5: Budget Evaluation in Beacon with *S3*. Behaviors of individual budgets per number of SNPs queried according to the typical user's query profile obtained from ExAC log data. The cyan curve represents the number of individuals with enough budget to answer "*Yes*" to queries targeting alleles with AF=0.001. Red, Green and Blue curves correspond to 0.002, 0.005, 0.01, respectively.

Supplementary Materials:

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

Table S4. 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 beacon.
H_1	Alternative hypothesis: query genome is in 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_{N-1}^i	Probability that none of the $N - 1$ genomes in the beacon has an alternate allele for query q_i.
D_N^i	Probability that none of the N genomes in the beacon has an alternate allele for query q_i.
δ	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.
α	Type I error: $P(\text{reject } H_0 \mid H_0 \text{is true})$. False positives.
β	Type II error: $P(\text{accept } H_0 \mid H_1 \text{is true})$. True positives.
$power$	$P(\text{reject} H_0 \mid H_1 \text{istrue}) = 1 - \beta$.
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.
Λ	LRT statistic.
t	Cut-off for the LRT statistic Λ (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 defense $S1$).
ε	Probability of adding noise on unique alleles (used in defense $S2$).
B_j	Budget for individual j. Initially $B_j = -log(p)$ for every j (used in defense $S3$).
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 δ 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

$$\begin{aligned} L_{H_1}(R,k) &= \textstyle\sum_{i=1}^{n} x_i \log(\Pr(x_i = 1|H_1,k)) + (1-x_i)\log(\Pr(x_i = 0|H_1,k)) \\ &= \textstyle\sum_{i=1}^{n} x_i \log(\delta(1-D_{N-1}^{i}(k)) + (1-\delta)(1-D_{N-1}^{i}(k-1))) \\ &+ (1-x_i)\log(\delta D_{N-1}^{i}(k) + (1-\delta)D_{N-1}^{i}(k-1)) \end{aligned} \qquad \text{(S1)}$$

where $D_{N-1}^{i}(k)$ 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,

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

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

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

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

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

Note that if $k=1$, from Equations (A1) and (A3) 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 Λ can be calculated according to Equation (5), where $N = N_a$.

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

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 ε of unique alleles, but not all. In other words, a beacon with *S2* will add noise with probability ε 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 ε 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 ε.

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

- A sequencing error δ occurred and none of the other $N-1$ participants has a copy of the allele.
- An artificial error ε 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

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

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

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

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

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

Similarly, under the null hypothesis we have

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

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

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

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 every 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_α on the value of Λ 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 Λ under H_0 and H_1. In the analysis by Shringarpure and Bustamante , it is shown that Λ 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_{yes} denote the sequence of all "Yes" responses. We consider two cases:

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

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

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

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

So we see that in any case, the random variable Λ can only take on two values, either $\sum_{i=1}^{n} \log(1 - D_N^i)$ or ∞. Now, if H_1 is true, R must be R_{yes}. Thus, we have that the distribution of Λ 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, Λ 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 ∞ otherwise. In summary:

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

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

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 a individual possesses, we first check whether the budget of the individual is larger than $-\log(1 - D_N^i)$, then reduce his 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 $\Pr[x_i = 1|H_1] = 1 - D_{N-1}^i$ and $\Pr[x_i = 1|H_0] = 1 - D_N^i$, 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 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 Fig.S1 show that also in the multi-population beacon the new re-identification attack based on allele frequencies is more effective than the one of 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[2] 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.

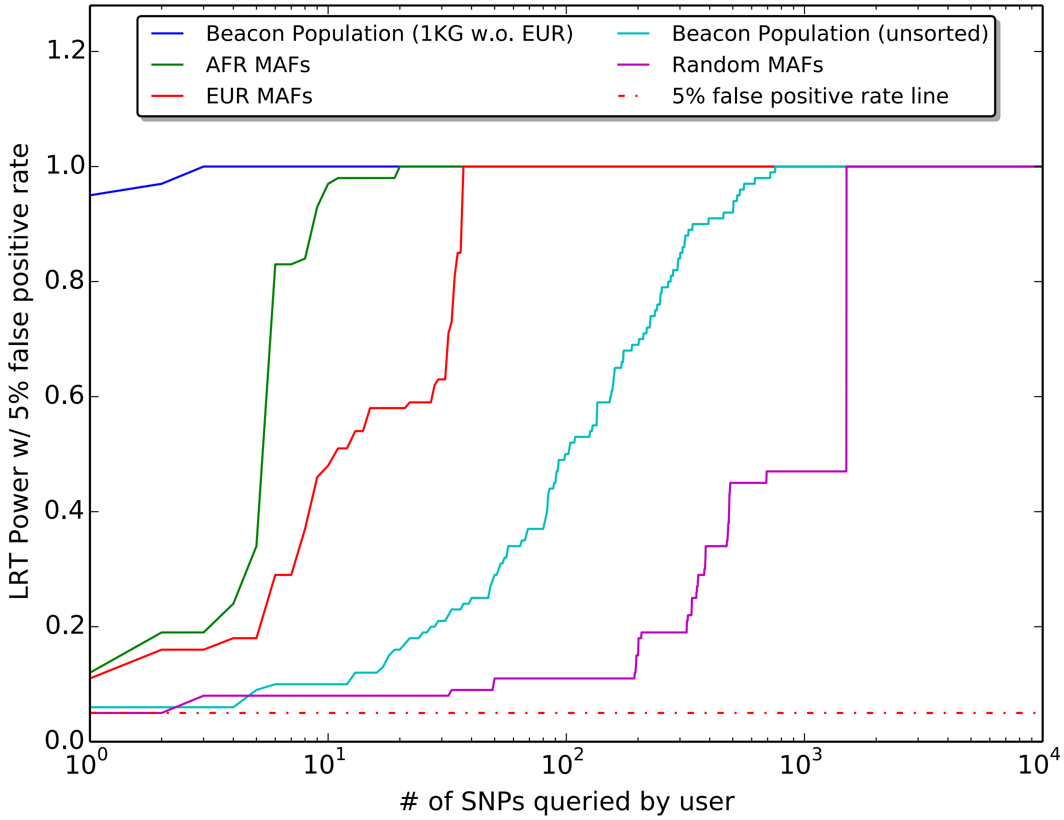

Figure S1. Different power rates per number of SNPs queried from an unprotected multi-population beacon (the beacon contains individuals from all ancestry in the 1000 Genomes Project but the European ancestry) by an adversary with background knowledge on allele frequencies. Different colors represent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

[2] 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 to the size of the database, and kept as a table in the database. Once k or ε 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 Algorithm1 in the paper). Suppose there are *N* individuals in the dataset, then for given a query, we need to:

- Compute the risk of a query, which can be done in constant time *O(1)*
- Check whether there are individuals that have the queried allele and a budget greater than the risk, *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 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.