The Effect of Kinship in Re-identification Attacks Against Genomic Data Sharing Beacons

Miray Aysen[1], Erman Ayday[2,*], and A. Ercument Cicek[1,3,*]

[1] Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800
[2] Computer and Data Sciences Department, Case Western Reserve university, Cleveland, OH 44106
[3] Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract. Big data era in genomics promises a breakthrough in medicine, but sharing data in a private manner limits the pace of field. Widely accepted "genomic data-sharing beacon" protocol provides a standardized and secure interface for querying the genomic datasets. The data is only shared if the desired information (e.g., a certain variant) exists in the dataset. Various studies showed that beacons are vulnerable to re-identification (or membership inference) attacks. As beacons are generally associated with sensitive phenotype information, re-identification creates a significant risk for the participants. Unfortunately, proposed countermeasures against such attacks have failed to be effective, as they do not consider the utility of beacon protocol. In this study, for the first time, we analyze the mitigation effect of the kinship relationships among beacon participants against re-identification attacks. We argue that having multiple individuals from the same family in a beacon can garble the information for re-identification attacks since a substantial number of single nucleotide polymorphisms (SNPs) are expected to be shared among kin-related people. We show that having at least one of the parents of a victim in the beacon causes (i) significant decrease in the power of attacks and (ii) a substantial increase in the number of queries needed to confirm an individual's beacon membership. We also show how the protection effect attenuates when more distant relatives, such as grandparents are included alongside the victim. Furthermore, we quantify the utility loss due adding relatives and show that the usability of the system is still preserved using this technique.
Contact: *exa208@case.edu or *cicek@cs.bilkent.edu.tr

1 Introduction

In the last two decades, emerging sequencing technologies have been providing researchers with larger genomic datasets which creates new opportunities for understanding the genetic architectures of diseases and have been providing insights for new therapies [9]. This was further fueled by the exponential growth of the personal genomics industry in the last five years which attracted consumers that want to (i) familiarize themselves with their genetic origins or (ii) take precautions against possible health risks [8]. Growing size of genomic datasets promises new opportunities for research through data sharing. However, data inherently contains highly sensitive information and privacy-preserving and secure sharing of data comes up as a major challenge. Anonymization of the genomes is a straightforward solution. However, the genome is the utmost personal identifier and it can reveal the identity of an individual. Such a scenario can dire ethical consequences, such as discrimination (e.g., on the basis of employment or insurance [9], [2], [10]). Leakage of genomic information of an individual not only jeopardises their privacy but also the privacy of their relatives since genomic information of an individual can be used to infer genomes (and hence genetic predisposition to a diseases) of other family members [7]. For instance, Deznabi *et al.* demonstrate that the single nucleotide polymorphisms (SNPs) of relatives can be reconstructed with high confidence using (i) Mendel's law, (ii) high-order correlations between SNPs, and (iii) minor allele frequencies (MAFs) of the SNPs in a population [5]. Thus, researchers face a trade-off between (i) sharing data to empower genetic research, which puts the participants under risk and legally binds them for possible repercussion and (ii) not sharing the data, which potentially bars the advances in life sciences.

In 2016, the Global Alliance for Genomics and Health (GA4GH) introduced the Beacon Project, a system constructed with the aim of providing a secure and systematic way of sharing genomic data. Beacons provide an interface, in which a user can query the existence of a specific nucleotide at a given position in a particular chromosome. For instance, "is there a participant carrying nucleotide C at the $100{,}000^{th}$ position of chromosome 1?" is a valid query. The beacon responds to such a query only with a simple "yes" or "no". Therefore, the beacon protocol is considered safer (compared to other statistical databases), as the query responses are binary and they do not include any information about allele frequencies. Moreover, a "yes" answer cannot be tied to a specific individual in the beacon. The beacon protocol also encourages cross site collaborations because the users do not have to go through the rigorous paperwork unless they identify a useful dataset for their research.

Nonetheless, previous studies showed that beacons are vulnerable to re-identification attacks [13],[**?**],[14]. Shringarpure and Bustamante showed that a likelihood ratio test (LRT) can be used to infer the membership of an individual to a beacon by querying the beacon for a couple of hundred SNPs of that individual (SB Attack). This study clearly showed

that the beacons indeed leak information which potentially leads to the disclosure of sensitive information if the beacon is associated with a sensitive trait (e.g., SFARI beacon which contains participants with autism). Raisaro *et al.* advanced the SB Attack by assuming the attacker has information regarding the MAFs in the population (Optimal Attack). By asking SNPs with low MAF values first, they showed that an attacker actually needs only a handful of queries to achieve the same power as the SB Attack. Finally, von Thenen *et al.* introduced two new attacks. First, they showed that the attacker can infer beacon responses using the responses of previously asked queries (Query Inference - QI Attack). Second, they showed that the attacker can still launch an attack even if the victim has concealed their SNPs with low MAF values (Genome Inference - GI Attack). Both attacks utilize the correlations among SNPs and they further decrease the number of required queries for confident inference.

Several countermeasures have been proposed in the literature to protect the privacy of the beacon participants against re-identification attacks. Shringarpure and Bustamante considered: (i) having larger beacon sizes, (ii) sharing only small genomic regions (e.g., genes of interest) instead of full genome, (iii) having a uniform ancestry composition in the beacon, and (iv) not publishing the metadata (e.g., dataset size). However, as also stated by the authors, these techniques reduce the utility of the beacon. Raisaro *et al.* proposed a query budget per participant, which expires if many SNPs of an individual is queried and the participant is taken out of the system (i.e., queries including them are not answered). Yet, von Thenen *et al.* showed that inference of beacon answers via SNP correlations can get around such budget based countermeasures. Al Aziz *et al.* proposed two algorithms that randomize the beacon responses. However, such noise-based techniques reduce the utility of the users and substantially affect the usability of the system.

In this study, we consider using the kinship of beacon participants as a countermeasure against re-identification attacks. We show that the power of the state-of-the-art attacks substantially decrease when at least one of the parents of a victim is added to the beacon. The key idea is that kinship garbles the information returned to the attacker since family members share many SNPs and the re-identification attack algorithm cannot conclude weather the "yes" answer coming from the beacon originates due to the victim or their relatives.

Using a beacon constructed from the CEU population of the HapMap dataset, we show that the number of queries to infer beacon membership of a victim increase when at least one of their family members is added to the beacon. We also show how the power loss for the state-of-the-art re-identification attacks changes with different degrees of relatives in the beacon. Finally, we quantify the utility loss of the beacon due to this proposed mitigation technique. We define the utility as the proportion of the flipped beacon responses (due to the proposed mitigation technique) and we show that the proposed mitigation technique does not cause a significant decrease in utility (especially for SNPs with low MAF values).

The rest of the paper is organized as follows. In Section 2, we provide technical details on methods and datasets we use. In Sections 3 and 4, we provide and discuss the results about the effect of kinship on re-identification attacks under various settings. Finally, in Section 5, we conclude the paper.

2 Materials and Methodology

In this section, we present the technical details on the state-of-the-art re-identification attacks against genomic data sharing beacons. We also describe our techniques to quantify the power of the attacker and the family simulation procedure.

2.1 Re-identification Attacks Against Beacons

Shringarpure and Bustamante introduced the first re-identification attack against beacons. The algorithm repeatedly queries for a victim's heterozygous SNPs and a likelihood ratio test (LRT) is performed to choose between a null hypothesis (H_0, in which the queried genome is not in the beacon) and an alternative hypothesis (H_1, in which the queried genome is a member of the beacon). The log-likelihood (L) under the null and alternate hypothesis are shown as follows:

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

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

where R is the set of responses and x_i is the binary response for the ith query. δ term in the alternate hypothesis indicates a small probability that attacker's copy of the victim's genome does not match the beacon's copy (e.g., due to differences in the sequencing pipelines). n is the number of queried SNPs, D_N is the probability that none of the N individuals in the beacon has the corresponding allele for the queried SNP, and D_{N-1} is the probability that no individual except for the victim having the corresponding allele for the queried SNP. The LRT statistic Λ is calculated as follows:

$$\Lambda = \sum_{i=1}^{n} \log\left(\frac{D_N}{\delta D_{N-1}}\right) + \log\left(\frac{\delta D_{N-1}(1-D_N)}{D_N(1-\delta D_{N-1})}\right) x_i. \tag{3}$$

H_0 is rejected if it is less than a threshold and this threshold can be found theoretically under the assumption that the queried SNPs are i.i.d.

Raisaro *et al.* introduced the Optimal Attack, which assumes that the attacker has access to the minor allele frequencies (MAF) of a population representing the beacon participants. Then, the SNPs are queried in the ascending MAF order. The formulation is identical to the SB Attack, but in Optimal Attack, the computations of D_{N-1} and D_N depend on the query i since each query has a different effect on the LRT statistic. Thus, in Optimal Attack, D^i_{N-1} and D^i_N are calculated as follows: $D^i_{N-1} = (1-f_i)^{2N-2}$ and $D^i_N = (1-f_i)^{2N}$, where f_i represents the MAF of SNP i. The likelihood-ratio test (LRT) statistic, Λ is then computed as follows:

$$\Lambda = \sum_{i=1}^{n} \log\left(\frac{D^i_N}{\delta D^i_{N-1}}\right) + \log\left(\frac{\delta D^i_{N-1}(1-D^i_N)}{D^i_N(1-\delta D^i_{N-1})}\right) x_i. \tag{4}$$

The Λ threshold (t_α) for rejection of the null hypothesis is determined empirically for every query since the i.i.d. assumption in Shringarpure and Bustamante no longer holds. That is, for every query, the distribution of Λ under the null hypothesis is found using k individuals that are not in the beacon. When Λ value of a victim is less than t_α, the alternative hypothesis is chosen, where α represents the false positive rate.

von Thenen *et al.* introduced the Query Inference (QI) Attack, which extends the Optimal Attack by showing that in addition to the MAF information of population, the attacker can also utilize the correlations between the SNPs. The correlation between the SNPs are calculated based on their LD values and a SNP-SNP network is generated, in which the vertices are the SNPs and the weights on directed edges represent the LD values. When a SNP is queried, the beacon responses of the neighboring SNPs in the SNP-SNP network are inferred, and hence the required number of queries is significantly decreased. In the QI Attack, the null and the alternative hypothesis formulations (and the corresponding Λ definition) are changed, so that the new calculation also reflects the information obtained from (m) inferred queries. The inferred queries are weighted by an inference confidence (γ) and the new log-likelihoods and LRT statistics are computed as follows:

$$\begin{aligned} L_{H_0}(R) = \sum_{i=1}^{n} \Big(& x_i \log(1-D^i_N) + (1-x_i)\log(D^i_N) \\ & + \sum_{j=1}^{s} \gamma x_i \log(1-D^j_N) + \gamma(1-x_i)\log(D^j_N) \Big) \end{aligned} \tag{5}$$

$$\begin{aligned} L_{H_1}(R) = \sum_{i=1}^{n} \Big(& x_i \log(1-\delta D^i_{N-1}) + (1-x_i)\log(\delta D^i_{N-1}) \\ & + \sum_{j=1}^{s} \gamma x_i \log(1-\delta D^j_{N-1}) + \gamma(1-x_i)\log(\delta D^j_{N-1}) \Big), \end{aligned} \tag{6}$$

$$\begin{aligned} \Lambda = \sum_{i=1}^{n} \Bigg(& \log\left(\frac{D^i_N}{\delta D^i_{N-1}}\right) + \log\left(\frac{\delta D^i_{N-1}(1-D^i_N)}{D^i_N(1-\delta D^i_{N-1})}\right) x_i \\ & + \sum_{j=1}^{s} \log\left(\frac{D^j_N}{\delta D^j_{N-1}}\right) + \log\left(\frac{\delta D^j_{N-1}(1-D^j_N)}{D^j_N(1-\delta D^j_{N-1})}\right) \gamma x_i \Bigg). \end{aligned} \tag{7}$$

Finally, von Thenen *et al.* also introduced the Genome Inference (GI) Attack, in which the attacker can still launch an attack even if it does not have some informative (low-MAF) SNPs of the victim (e.g., in case they are hidden by the victim as a mitigation). To infer the existence of such hidden SNPs, authors calculate the probability of a victim v having a minor allele at a locus j using k neighboring SNPs on the genome. This is calculated as follows:

$$P_k(SNP^v_j) = P(SNP^v_j | SNP^v_{j-1}, ..., SNP^v_{j-k}), \tag{8}$$

where k is the order of the Markov chain and also the number of neighbors considered. The Markov chain is built as described in [12] as follows:

$$P_k(SNP_j) = \begin{cases} 0 & \text{if F(SNP}_{j-k,j-1}) = 0 \\ \frac{F(SNP_{j-k,j})}{F(SNP_{j-k,j-1})} & \text{if F(SNP}_{j-k,j-1}) > 0, \end{cases} \tag{9}$$

where $F(SNP_{j,i})$ is the frequency of the sequence [SNP_i, SNP_j] on the genome. Once the hidden SNPs of the victim are inferred, authors proposed launching either the Optimal Attack or the QI Attack with the (inferred) genome of the victim.

2.2 Power Calculation

We perform a power analysis to quantify the success of a re-identification attack [11],[14]. All Optimal, QI and GI Attacks query SNPs in the ascending MAF order. In this scheme, for every query i, a t^i_α value is determined which is the Λ threshold to reject the NULL hypothesis. α represents the desired false positive rate. We pick k people (controls) who are not in the beacon. We assume these k people has a similar population structure as the beacon participants. For each of the k controls, i^{th} query is posed and a Λ value set is obtained: $\vec{\Lambda^i_{control}} = (\Lambda^i_1, \Lambda^i_2, ..., \Lambda^i_k)$. $\vec{\Lambda^i_{control}}$ is sorted (in ascending order) and the Λ^i value, for which α percent of k people have smaller Λ^is, is picked as the t^i_α. For instance, if k is 20 and α is 0.05, then the second smallest Λ^i is picked as t^i_α as 1 person is below that threshold. This represents the false positive threshold as for that person, the NULL hypothesis would have been rejected. Given n queries, the $\vec{t_\alpha}$ list is generated which contains the t_α values of all n queries: $\vec{t_\alpha} = (t^1_\alpha, t^2_\alpha, ..., t^n_\alpha)$.
To measure the power per each query i, first, l people from the beacon (cases) are obtained. Then, for everyone in this set, the ith query is posed and a Λ value set is obtained: $\vec{\Lambda^i_{case}} = (\Lambda^i_1, \Lambda^i_2, ..., \Lambda^i_l)$. Then, the power for the ith query is calculated as follows: $P^i = \sum_{\Lambda^i_j \in \Lambda^i_{case}} \mathbb{1}_{\Lambda^i_j < t^i_\alpha} \Big/ k$. This is the fraction of the cases who have Λ^i_j value that is less than t^i_α. For that fraction of the l people, the NULL hypothesis is (correctly) rejected. For instance, if l is 20 and 5 people have Λ^i_j that is less than t^i_α, this means for the ith query, the power of the attack is 25% at the α false positive rate. The vector $\vec{P^i} = (P^1, P^2, ..., P^n)$ is then plotted to see the power change with respect to increasing number of queries. An attack which reaches to 100% power earlier than others is considered more powerful. Our goal in using kinship as a countermeasure aims at decreasing the power of the state-of-the-art attacks, and thus either increasing the number of queries to reach 100% power or preventing the attack to reach 100% power at all.

2.3 Generation of Relatives' Genomes

The power calculation procedure for all considered attacks, which is detailed in Section 2.2, requires a large number of people (i) to determine α_t (e.g., at least 20 people that are not in the beacon as applied in the literature); and (ii) to determine the power (e.g., at least another 20 people that are in the beacon as applied in the literature) [11],[14]. To perform such an experiment, one needs sequencing data of at least 40 families, which, to the best of our knowledge, does not publicly exist. Thus, we opted for a simulation procedure to generate parents and grandparents of a person (i.e., the victim), which also fits the binary response scheme of the beacon protocol.
For each allele of the victim, the genotype of the parents are determined. Table 1 shows the possible genotypes of the mother and the father given the genotype of the victim. For instance, if the victim is homozygous minor (aa) for a given locus, then we know that for that locus, both parents have at least one minor allele (a). These options are not equally likely to occur though. The options presented in Table 1 are generated with respect to the allele frequencies calculated from the corresponding population. For instance, if the victim is heterozygous for a rare allele, then both parents having the minor allele is less likely than one being homozygous major (AA) and the other being heterozygous. Note that we do not consider *de novo* mutations and one of the parents have to have the minor allele given the child (i.e., the victim) is heterozygous at that locus. We also assume an allele is equally likely to come from the father or the mother.
We use the same algorithm to generate the genomes of the victim's grandparents, using the generated parents as the child. We also assume that parents and grandparents are not related.

3 Results

In this section, we provide the results on the reduction in the attacker's power once a relative (or a set of relatives) is added to the beacon. We also quantify the utility loss in beacon responses once a relative of the victim is added to the beacon for privacy protection.

3.1 Experimental Setup

The beacon we used in the experiments contains 65 individuals from the Utah residents with northern and western European ancestry (CEU) population of the HapMap project [4]. The same population and beacon size are also used in all previous re-identification attacks [13],[11],[14]. We generated artificial genomes for the parents and grandparents of the victim as described in Section 2.3 and added them to this beacon to observe the effect of kinship on re-identification attacks.
We calculated the MAF values (needed for the Optimal, QI, and GI Attacks) from the HapMap dataset using 100 individuals that are not in the beacon. The QI and GI Attacks require the LD values for the considered SNPs to (i) create the SNP-SNP network for query inference and (ii) build the high-order Markov model for the genome inference of the hidden informative SNPs. We used the same 100 individuals from the HapMap dataset to create these models.

Table 1: A toy example showing the possible SNPs of the parents for three cases, in which the child's SNP is (a) major homozygous, (b) minor homozygous, and (c) heterozygous.

		Mother	Father
Child	AA	AA	AA
		AA	Aa
		Aa	AA
		Aa	Aa

(a)

		Mother	Father
Child	aa	aa	aa
		Aa	aa
		aa	Aa
		Aa	Aa

(b)

		Mother	Father
Child	Aa	AA	Aa
		AA	aa
		Aa	AA
		Aa	Aa
		Aa	aa
		aa	AA
		aa	Aa

(c)

We divided the individuals that are used to evaluate the attacks into two as case and control groups. We selected the individuals in the case group (i.e., the ones that are in the beacon) as the first 20 individuals in the beacon. We selected 20 individuals in the control group (i.e., the ones that are not in the beacon) from the HapMap dataset (that are not used to calculate the MAF values nor the correlation models). For each individual in case and control groups, we generated synthetic family members. The generated family members include mother, father, and grandparents. We added these synthetically generated family members into the beacon in a systematic way. We quantified the power (as in Section 2.2) with and without the relatives and measured the utility loss when relatives are added.

3.2 Re-identification Attacks on Genomic Data Sharing Beacons with Family Members

We claim that adding family members to the beacon will improve the privacy of beacon participants, and hence it will be a natural mitigation technique. The origin of this idea is the inheritance, the fact that an individual's genome is constructed based on their parents genomic information. We also claim that addition of family members to the beacon does not cause a significant utility loss. We further discuss this in Section 3.3. To show how the results of attacks change with the presence of family members in the beacon, we used the same experiment parameters as the previous re-identification attacks (detailed information about datasets and experimental settings are in Section 3.1).

Table 2: Attacker's auxiliary information for each of the considered attack scenarios.

	Attacker's Auxiliary Information
Optimal Attack	MAF of victim's population
Query Inference Attack	MAF of victim's population LD values
Genome Inference Attack	MAF of victim's population LD values high-order correlations between SNPs

The attacker's goal is to infer whether the targeted individual (victim) is in the beacon or not. Besides the auxiliary knowledge of the attacker shown in Table 2, the attacker also has the VCF file of the victim. We let t be an evaluation parameter representing the threshold for the hidden SNPs of the victim (e.g., as a countermeasure against the re-identification attack). That is, we assume the victim hides their SNPs with MAF values less than t. Note that for the GI Attack, we do not consider the case when $t = 0$, since when there is no hidden SNPs, the GI attacks becomes the same as the QI Attack. For the GI Attack, to infer the hidden parts of the victim's genome, we used a 4th order Markov chain model (that is consistent with [14]). We also let the attacker query the beacon for the heterozygous SNP positions of the victim (to have the same settings with previous re-identification attacks).

We assume that the attacker does not have access to VCF files of victim's family members. The attacker may or may not know the existence of victim's family members in the beacon since this knowledge does not provide an advantage to the attacker to infer the membership of the victim. If attacker knows that (at least) a family member is in the beacon, it cannot be sure about the reason of the "yes" responses (e.g., whether they are due to the victim or other family members). If the attacker does not know the membership information of victim's family members, it will possibly come to a wrong conclusion about the membership of the victim to the beacon. Thus, in both cases, attacker's inference power for the victim's membership will be low (due to the existence of family members in the beacon).
We performed the Optimal, QI, and GI Attacks for different scenarios for the individuals in the beacon: (i) the original beacon that does not involve victim's family members, (ii) beacon that contains victim's mother, (iii) beacon that contains victim's father, (iv) beacon that contains victim's both parents, (v) beacon that contains victim's grandparents from mother's side, and (vi) beacon that contains victim's grandparents from father's side. We also show these different scenarios in Figure 1.

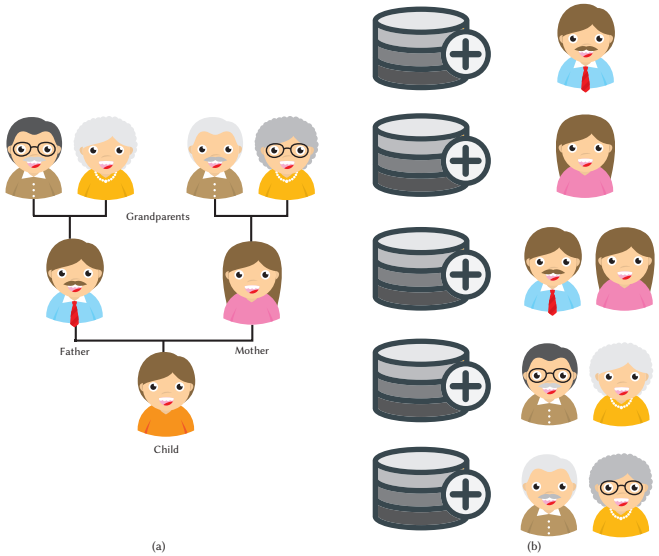

Fig. 1: (a) Family tree of the victim (child in the figure). (b) Experimental setup for different scenarios for the individuals in the beacon.

First, we show how the power of the attack changes for (i) the original beacon (that does not include any family members of the victim), (ii) the beacon that only includes the mother of the victim, (iii) the beacon that only includes the father of the victim, and (iv) the beacon that includes both parents of the victim. We show the results of this experiment for Optimal, QI, and GI Attacks in Figure 2.
In [14], using the QI or GI Attacks, authors show that the individual's membership to the beacon can be inferred with high power with only a few queries. For instance, for the GI attack, they show that the power reaches to 1 after only two queries. Our results for the original beacon (that does not include any family members of the victim) are also consistent with the results of [14]. We also observed that when at least one family member of the victim is in the beacon, the power curves shift to right, meaning that the attacker needs more queries to infer the membership of the victim to the beacon. For instance, when at least one family member of the victim is in the beacon, the power only reaches to 0.1 after two queries (for which the QI Attack's power reaches to 1 for the original beacon).
In the Optimal Attack, when $t = 0$, including only the mother or father of the victim to the beacon increases the number of queries for the attacker (to have a high power) to hundreds. In the GI Attack, we observed that adding at least one family member to the beacon prevents the attacker's power from reaching to 1. When only the mother of the victim is in the beacon, the power reaches approximately to 0.55. Similarly, adding only the father of the victim to the beacon causes the power to reach approximately to 0.7. In the QI Attack, we observed that when at least one family member of the victim is in the beacon, the attacker's power reaches to 1 in hundreds of queries for only smaller values of t. Furthermore, for all attacks, when both parents of the victim are in the beacon, attacker's power never reaches to 1, and it is always low. This is expected since the minor alleles of the child (victim) either come from the mother or the father. Thus, when both parents of the victim are in the beacon, there is no way for the attacker to make inference about the membership of the victim. We also observed that once the power converges to a value, it does not change even if the attacker keeps asking for more queries.

In order to investigate why power does not reach to 100% in the GI Attack when either the mother or the father is in the beacon, we synthetically generated a parent that does not share any minor alleles with the victim. We observed that when that parent is in the beacon, power reaches to 100%. Thus, we conclude that as the parent (that is included in the beacon) shares more minor alleles with the victim, the decrease in the power of the attacker is more.

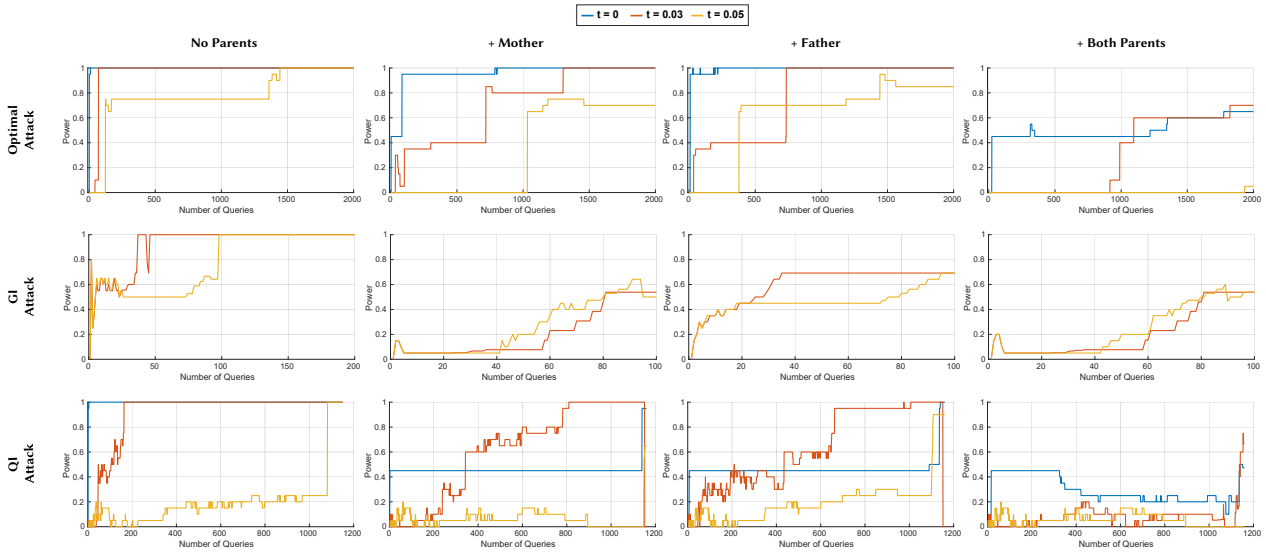

Fig. 2: The power curves of the Optimal Attack, GI Attack, and QI Attack with different beacon setups. For all attacks: (i) the first plot is when the beacon does not include any family members of the victim, (ii) the second plot is when only the mother of the victim is in the beacon, (iii) the third plot is when only the father of the victim is in the beacon, and (iv) the fourth plot is when both parents of the victim are in the beacon. SNPs of the victim with MAF values smaller than t are hidden from the attacker.

Next, we show how adding more distant relatives of the victim (i.e., grandparents) effect the attacker's power for the re-identification using the Optimal Attack. As we show in Figure 3, adding only one of the grandparents to the beacon (mother's father as in Figure 3(a) or mother's mother as in Figure 3(b)) causes the attacker's power decrease less than adding the mother (in Figure 2) since degree of kinship decreases. In other words, as expected, the decrease in attacker's power is inversely proportional with the distance between the victim and their relatives. We also obtained similar results when we added father's mother and father's mother separately. Furthermore, we observed that adding mother's both parents (i.e., victims grandparents from mother's side as in Figure 3(c)) to the beacon is almost equivalent to adding the mother. Similarly, adding father's parents to the beacon (as in Figure 3(d)) is almost equivalent to adding the father. Note however that adding mother's (or father's) both parents provide a slightly stronger mitigation compared to adding only the mother (or father). This is because adding mother's both parents introduce more diversity to the beacon compared to adding just the mother. For instance, comparing the beacon including only victim's mother and mother's parents, the beacon including victim's grandparents may include more "yes" responses (due to heterozygous SNPs of the grandparents that may not occur in the mother).

3.3 Utility Analysis of the Proposed Mitigation Technique

We showed that adding victim's family members to the beacon significantly increases the number of queries needed for the attacker to have a high power. However, as discussed, beacons are typically associated with a particular phenotype (i.e., all participants of the beacon has the corresponding phenotype). Thus, adding a family member of the victim to the beacon may result in a utility decrease in beacon's responses since (i) the added family member(s) may not have the corresponding phenotype of the beacon and (ii) the added family member(s) may result in a change in beacon's original responses.

In particular, if the original beacon response (before adding any family members as a mitigation technique) is "no" for a query and adding a family member changes that beacon response to "yes" (due to heterozygous SNPs of the added family member), utility of beacon decreases. Therefore, we define utility loss of the beacon as the fraction of additional "yes" responses that arise due to the addition of one or more extra individuals (family members of the victim) as a result

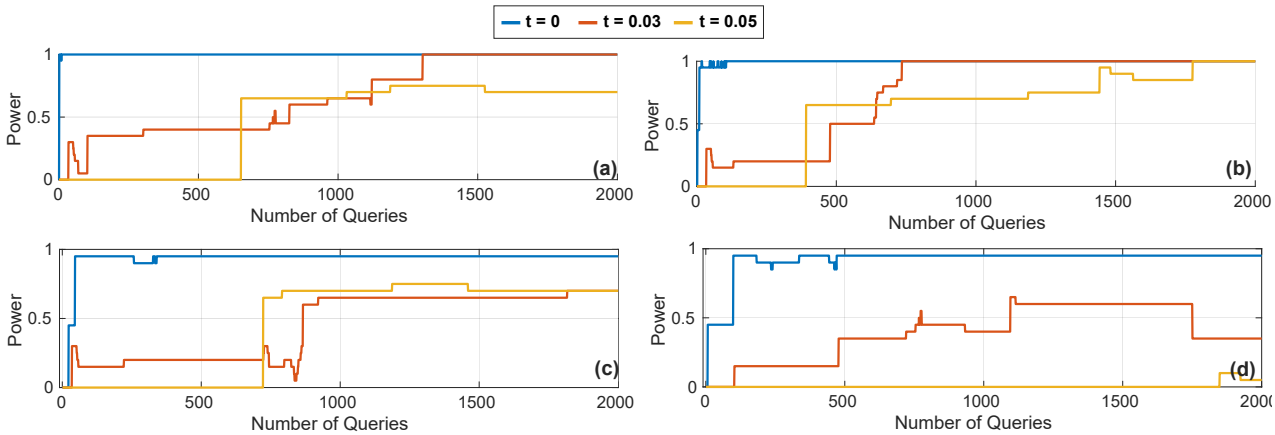

Fig. 3: The power curves of the Optimal Attack with different beacon setups. The beacon includes (a) victim's mother's father, (b) victim's mother's mother, (c) victim's both grandparents from mother's side, and (d) victim's both grandparents from father's side.

of the proposed mitigation technique. In Table 3, we show the decrease in utility of beacon's responses for both case and control groups (that we used in our experiments) due to the addition of the family member(s) as a mitigation technique. We observed that the utility loss is less than 10% even when both parents of the victim are included in the beacon.

Table 3: The fraction of additional "yes" responses that arise due to the addition of family members of the victim (as a result of the proposed mitigation technique) is shown as a measure of utility loss for the case and control groups. Each individual in the case and control groups are selected as the victim and are added to the CEU beacon with 65 individuals (note that cases are already in the beacon). The utility loss is calculated when the parents are also added to the beacon one-by-one. That is, parents of only one victim is in the beacon at a time.

	Mother in Beacon	Father in Beacon	Both Parents in Beacon
Control Group	4.50%	7.52%	9.78%
Case Group	2.67%	6.78%	9.13%

SNPs with a lower MAF values are particularly important for the researchers since there is an inverse relationship between the a variant's disease odds ratio and its frequency [3]. Thus, we also quantified the utility loss of the beacon responses considering the SNPs with low MAF values. One by one, we added the mothers of 20 case individuals to the beacon and observed the utility loss for various MAF thresholds. In Figure 4, we show the utility loss in beacon responses (y-axis) for all SNPs with an MAF value less than a threshold (x-axis; cumulative). We observed that utility loss is substantially smaller for SNPs with lower MAF values. For instance, for SNPs with an MAF less than 0.01, the utility loss is less than 0.05%. In other words, adding a family member of the victim to the beacon does not cause much change in the results of queries that involve low-MAF SNPs, which is expected as such SNPs are rare and are not frequently observed.

4 Discussion

Genomic data sharing beacon protocol has been widely accepted by the community as the golden standard for secure and privacy preserving data sharing. The Beacon Network[4], providing a central querying mechanism to 80 beacons, lit all over the world for various phenotypes ranging from autism to cancer (accessed on January 28, 2020). However, the information leaks identified by several re-identification attack algorithms and by a recently introduced genome reconstruction attack [1] had eyebrows raised on the usability of the system. Currently, setting up a beacon is a risk

[4] https://beacon-network.org/

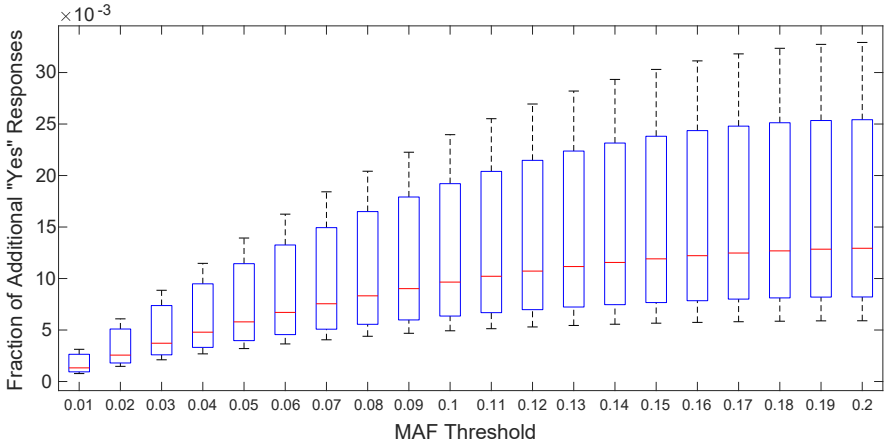

Fig. 4: The utility loss of the beacon responses considering the SNPs with low MAF values. The box plots show the fraction of additional "yes" responses that arise due to the addition of family members for 20 cases when the mother of each victim is added to the beacon one-by-one independently. The x-axis shows various MAF thresholds. For each x value, all SNPs with an MAF value less than that threshold are considered.

for all parties including genome donors, data owners, and even for the beacon system operators due to possible ethical, legal, and monetary repercussions.

A correspondence published in 2019 by the GA4GH acknowledges possible re-identification risks and offers possible mitigation strategies [6]. One strategy is using aggregate beacons. Aggregation process involves querying multiple beacons and joining their responses. A "yes" answer means at least one beacon contains the queried variant; a "no" answer means none has the desired allele. Such an approach leads to having more data points than the individual beacons, which, as also suggested by Shringarpure and Bustamante, makes it harder for the attacker to pinpoint the origin of a "yes" answer. One example of such is the Conglomerate Beacon. However, this strategy also results in a substantial utility loss for the users (researchers) as they might have to apply for access to all individual datasets if they find out that at least one of the beacons have the variant they are interested in. The second suggestion is the usage of participant budgets as suggested by Raisaro *et al.*. This strategy assigns a personal budget to each participant and if many rare SNPs (i.e., relatively more informative and identifying SNPs) of a participant are queried, the algorithm takes them out of the system (i.e., it does not provide a "yes" response if that person is the only carrier of that SNP in the beacon). This seems sensible, yet, in [14], authors show that an attacker, by inferring the responses of a beacon via linkage disequilibrium between the SNPs in a population, can get around these budgets. Considering the individual NA12272 from the HapMap project in a beacon of 65 CEU individuals (constructed from the HapMap project), they show that while the Optimal Attack requires 7 queries for re-identification, the QI attack can identify this person with only 5 queries, before the budget expires.

Shringarpure and Bustamante suggested inclusion of control samples in a beacon. Similar to the aggregate beacon strategy, this decreases the usability and utility of the system since controls, who do not carry the phenotype which the beacon is associated with and who are not relatives of the people in the beacon, would result in flipping of many irrelevant "no" answers to "yes". In this work, we investigate the feasibility of adding relatives of individuals to a beacon as a countermeasure. Adding relatives still results in a utility loss, however, as shown in Section 3 the loss is not significant given the fact that most SNPs are shared between the victim and their relatives. Moreover, in beacons of heritable diseases, a relative is more likely to be related to the trait than a random control individual. Thus, the utility loss caused by the proposed approach will be less compared to adding random controls. Yet, we show that this creates a major confusion for the attacker. As clearly shown in various settings, the power curves for the state-of-the-art attacks shift right, which indicates that the number of required queries substantially increase to achieve the same re-identification power. In many cases, the power does not even reach to 100%, which means the attacker cannot have high confidence about the success of the attack.

We also investigated the scenario, in which the attacker also has knowledge about the genomes of the victim's relatives. For example, we assumed that the attacker has the SNPs of both victim and their mother (and/or father). The initial idea is, if the attacker applies the attacks by using the SNPs that differentiates between relatives and the victim, the power of the attack will reach to 100%. So, the power decrease we achieve by existence of a relative in the beacon will be ineffective. However, in practice, launching the QI and GI attacks by using these differentiated SNPs will not be effective since (i) these SNPs are sparse and are not enough for inference of the hidden parts of the genome, as such inference requires many SNPs sequentially positioned on the genome (GI Attack); and (ii) similarly, these SNPs are less likely to be in linkage disequilibrium, and thus less likely to be correlated to enable inference of the beacon

answers (QI Attack). To evaluate the power of this new scenario, a new power calculation approach is needed which we will consider as a future work.
One drawback of this mitigation strategy is the additional sequencing cost of the relatives. Moreover, the technique depends on relatives giving consent to sharing their data, which also puts them under re-identification risk. However, the protection effect is symmetric for the victim and the relatives. To circumvent these problems, one can opt for simulating relative data, as we did in this work due to the unavailability of large family studies required to quantify the power of re-identification attacks.

5 Conclusion

In this paper, we have proposed a mitigation technique against re-identification attacks for genomic data sharing beacons. The existing countermeasures to prevent re-identification attacks in beacons are shown to be ineffective since they either proved to be vulnerable against the attacks or they cause a significant decrease in beacon's utility. Our proposed technique relies on inheritance and it is based on adding genomes of a victim's family members to the beacon in order to mitigate the re-identification attacks. We have shown via experiments that adding at least one family member of the victim to the beacon results in a significant decrease in the power of the re-identification attacks. We have also shown the effect of adding different family members to the beacon to the power of the attacker. Furthermore, the proposed technique does not cause a substantial utility loss in beacon's responses. In particular, we have shown that the utility loss is significantly smaller for SNPs with low MAF values (which are of high importance for the researchers due to their associations with complex diseases).

References

1. Ayoz, K., Ayday, E., Cicek, A.E.: Genome reconstruction attacks against genomic data-sharing beacons. arXiv preprint arXiv:2001.08852 (2020)
2. Billings, P.R., Kohn, M.A., De Cuevas, M., Beckwith, J., Alper, J.S., Natowicz, M.R.: Discrimination as a consequence of genetic testing. American journal of human genetics **50**(3), 476 (1992)
3. Bomba, L., Walter, K., Soranzo, N.: The impact of rare and low-frequency genetic variants in common disease. Genome biology **18**(1), 77 (2017)
4. Consortium, I.H., et al.: The international hapmap project. Nature **426**(6968), 789 (2003)
5. Deznabi, I., Mobayen, M., Jafari, N., Tastan, O., Ayday, E.: An inference attack on genomic data using kinship, complex correlations, and phenotype information. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **15**(4), 1333–1343 (2018)
6. Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S.O., Brookes, A.J., Carey, K., Lloyd, D., Goodhand, P., et al.: Federated discovery and sharing of genomic data using beacons. Nature biotechnology **37**(3), 220–224 (2019)
7. Humbert, M., Ayday, E., Hubaux, J.P., Telenti, A.: Addressing the concerns of the lacks family: quantification of kin genomic privacy. In: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. pp. 1141–1152. ACM (2013)
8. Khan, R., Mittelman, D.: Consumer genomics will change your life, whether you get tested or not. Genome biology **19**(1), 120 (2018)
9. Kim, P.T.: Genetic discrimination, genetic privacy: rethinking employee protections for a brave new workplace. Nw. UL Rev. **96**, 1497 (2001)
10. Lapham, E.V., Kozma, C., Weiss, J.O.: Genetic discrimination: perspectives of consumers. Science **274**(5287), 621–624 (1996)
11. Raisaro, J.L., Tramèr, F., Ji, Z., Bu, D., Zhao, Y., Carey, K., Lloyd, D., Sofia, H., Baker, D., Flicek, P., et al.: Addressing beacon re-identification attacks: quantification and mitigation of privacy risks. Journal of the American Medical Informatics Association **24**(4), 799–805 (2017)
12. Samani, S.S., Huang, Z., Ayday, E., Elliot, M., Fellay, J., Hubaux, J.P., Kutalik, Z.: Quantifying genomic privacy via inference attack with high-order SNV correlations. In: Security and Privacy Workshops (SPW), 2015 IEEE. pp. 32–40 (2015)
13. Shringarpure, S.S., Bustamante, C.D.: Privacy risks from genomic data-sharing beacons. The American Journal of Human Genetics **97**(5), 631–646 (2015)
14. von Thenen, N., Ayday, E., Cicek, A.E.: Re-identification of individuals in genomic data-sharing beacons via allele inference. Bioinformatics **35**(3), 365–371 (2018)