

RESEARCH

Aftermath of bustamante attack on genomic beacon service

Md Momin Al Aziz^{1*}, Reza Ghasemi², Md Waliullah¹ and Noman Mohammed¹

Abstract

Background: With the enormous need for federated eco-system for holding global genomic and clinical data, Global Alliance for Genomic and Health (GA4GH) has created an international website called beacon service which allows a researcher to find out whether a specific dataset can be utilized to his or her research beforehand [1]. This simple webservice is quite useful as it allows queries like whether a certain position of a target chromosome has a specific nucleotide. However, the increased integration of individuals genomic data into clinical practice and research raised serious privacy concern. Though the answer of such queries are yes or no in Bacon network, it results in serious privacy implication as demonstrated in a recent work from Shringarpure and Bustamante [2]. In their attack model [2], the authors demonstrated that with a limited number of queries, presence of an individual in any dataset can be determined.

Contributions: In this paper, we further elaborate the strength and weakness of the attack by explaining some of their statistical and mathematical models using real world genomic database. We also extend their experimental simulations for different adversarial assumptions and parameters. This helps the data owners to calculate certain risks involved while sharing their genomic dataset in a beacon service. Also, we provide mechanisms to measure data privacy and utility tradeoffs that can be delivered in the tiered access model before any individual is identified in beacon network. The main contribution of this article is the proposal of two simple lightweight solutions that will make the attack very difficult to achieve while maintaining the fundamental motivation of beacon database network.

Methods: Our proposed algorithms capture the efficacy of beacon service while preserving the privacy of the participants in a beacon service. We also experimentally evaluated our solution on the original attack model [2] to better understand the privacy and utility tradeoffs provided by these two methods.

Keywords: Bustamante attack; Genomic beacon service; GA4GH genomic beacon; Bustamante attack mitigation; human genomic data privacy; re-identification attack

1 Background

Recent improvements on Genomic data sharing efforts have led researchers and clinicians gaining access and make comparisons across data from millions of individuals. Such development made it easier for genetic variant interpretation and in some cases treatment of rare diseases such as some special cancer types [1]. Most of the big organisations i.e., Broad institute in the U.S., BGI in china, Wellcome Trust Sanger in the UK etc. have an interest of making DNA data easier to access in order for their researchers to treat patients one on one. However, after twelve years of completion Human Genome project, the tremendous

growth of genomic data has exceeded the containers build to hold such data. Genomic and clinical data are generally still collected in either by disease, institution or by country. More importantly, current data sharing privacy requirements do not necessarily protect individuals identity within and across institutions and countries. Furthermore, data often stored in incompatible file format and there are no standardized tools and analytical methods are in place [1, 3, 4].

With such tremendous needs for global genomic and clinical data repository system, Global Alliance for Genomic and Health (GA4GH) has created a federated data ecosystems called Beacon data network, a way for searching genomic data as simple as World Wide Web. Since the project's launch in the middle of 2015, the beacon network has currently 23 different organizations covering over 250 genomic datasets.

*Correspondence: azizmma@cs.umanitoba.ca

¹Department of Computer Science, University of Manitoba, Winnipeg Canada

Full list of author information is available at the end of the article

The data sets served through beacons can be queried individually or in aggregate via the Beacon Network, a federated search engine (www.beacon-network.org) [1]. Thus, the Beacon Project aims to simplify data sharing through a web service (beacon) that provides only allele-presence information. Users can query institutional beacons for information about genomic data available at the institution. For example, an individual could ask the beacon web server about a genome that has a specific nucleotide and the beacon would respond either yes or no [5]. By providing only allele-presence information, beacons were assumed safe from attacks that require allele frequencies.

Although the beacon network has set up to share data and protect patient privacy simultaneously, it could potentially leak phenotype and membership information of an individual [5]. There is currently no cap on the number of queries a user can make in the Beacon database. Recently, Shringarpure and Bustamante showed that anonymous patients whose DNA data is shared via beacon network can be re-identified [2]. If an attacker has access to victims DNA, s/he can query different beacons to see whether the victim is in the dataset. They further demonstrated that it is possible to infer whether or not the victim is affected by a certain condition or disease [2]. Therefore, the anonymous beacons are inherently insecure and are open to re-identification attacks.

Very recently, some solutions [6, 7] have been proposed based on different policies around the access of the beacon service. However, these solutions will disrupt the quintessential feature of the proposed beacon service: that is to provide faster access to genomic data and to give open access to the research community. Different access controls are highly necessary for human genomic data access where phenotype or sensitive information about the disease is disclosed. However, the beacon service only provides us aggregate results of yes or no leading the researcher to a decision regarding the dataset's relatedness to his or her research. Therefore, we propose two solutions based on privacy-preserving techniques, which fit well with the beacon service and mitigate the possibility of identifying an individual from the dataset.

In this article, we explain the 'Bustamante Attack' [2] on genomic beacon services and propose two privacy preserving solutions. The contributions of this article can be summarized in two folds: *a*) understanding the statistical formulations and soundness of the attack, *b*) analyze lightweight privacy preserving solutions to mitigate the attack.

The main contributions of our work are as follows:

- We present the statistical and the mathematical model of the attack in a simplified form. This

helps us to analyze different and more realistic parameters on the original attack framework to exploit some weakness and justify our solutions accordingly.

- We show the required steps for any data owner to calculate the risk involved in sharing their genomic data in a beacon service.
- We propose two easy to implement and lightweight privacy preserving solutions which ensure the applicability of the beacon service as well as the privacy of the participants.
- We provide extensive experiments over synthetic data (according to [2]) to show the privacy-utility evaluation of our proposed methods which will help the development of different privacy preserving techniques on such attack model later on.

1.1 Beacon service for genomic data

A beacon is an online web search engine developed by the Global Alliance for Genomic and Health (GA4GH), which provides a way for genomic data owners and research institutes to easily share genomic data while maintaining patients privacy (Figure 1). It is a genetic mutation sharing platform that allows any user to query an institution's databases to determine whether these databases contain a genetic variant of interest while keeping all other sequence data obscured. A query in this search engine is defined by three parameters: chromosome number, position in that chromosome, and target nucleotide (A/T/G/C). A beacon query answer is either true or false, denoting the presence of that nucleotide in that specific position and target chromosome. In other words, it will only answer yes/no for the questions like: Do you have any genomes with an 'A/T/G/C' at some position 'Y', on specific chromosome 'Z'. This allows a researcher to target some specific dataset, which is relevant to his or her research. This service also helps a clinician to check whether a mutation found in one of her patients is also present in others without actually having access to their genomes [8].

Beacons are easy-to-implement techniques for several large-scale organizations when it comes to sharing genomic data. It also saves researchers a tremendous amount of time for tracking down useful data for their work as well [9, 10]. Unlike large centralized data repositories, a beacon network is distributed across many databases around the world and is virtually connected through software interfaces allowing continuous authorised access. This federated data ecosystem allows each organization to control their legal data within their jurisdiction [1]. The shared Genomics API in the beacon framework makes it easy to query all at once and ensures that GA4GH team can quickly add new beacons to the network.

1.2 Bustamante attack on beacon service

A recent study done by Shringarpure and Bustamante [2], developed a likelihood-ratio test that uses only allele presence information to predict if the genome of an individual is present or not in the beacon database. This study suggested that beacons are susceptible to re-identification attacks and thus can be subjugated to invade genetic privacy. Since a beacon database includes data with known phenotypes information such as cancer, autism or other diseases, this re-identification also potentially disclose phenotype information about an individual whose genomic data is present in the beacon [11]. Through simulations, they demonstrated that by making just *5,000 queries*, it was possible to identify someone and *even their relatives* in a beacon consisting *1,000 individuals*. They found that re-identification of an individual is possible even with the sequencing errors and variant-calling differences. They also demonstrated that a beacon constructed with 65 European individuals from the 1000 genome projects, it is possible to detect membership in the beacon with just 250 SNPs [2].

In this section, we briefly introduce the Bustamante attack and analyze its statistical methods. The goal of this attack is to know whether a genomic sequence g belongs to a specific database with the help of the beacon service. To answer this question they considered two hypothesis:

- 1 Null hypothesis H_0 : the query individual is not in the beacon service
- 2 Alternative hypothesis H_1 : the query individual is in the beacon service.

To determine the correct one, the adversary is allowed to query the beacon service with unlimited amount of queries. The adversary queries specific locations where the query individual has alternative allele to see whether the beacon server also contains an individual with the same allele values. Therefore, the responses of the beacon service are a sequence $x_1, \dots, x_n$ of yes or no. If we consider yes and no with '1' and '0' respectively, the the answer sequence, R will be a binary vector. For example, if the query individual is in the database, we will get yes (or 1) in each query. However if there are some genome sequencing error, we might get some wrong answers as well. This error is denoted by δ and also considered by the attack [2].

There is also another considerable case where multiple individual have the same allele in the database. This is why the attacker needs to leverage the likelihood ratio of both the assumptions whether the user is in the dataset or not. For a database of N genome, the log of this likelihood ratio can be computed for the response series R regarding the hypothe-

ses H_i as follows:

$$L_{H_i}(R) = \sum_{i=1}^n x_i \log P(x_i = 1|H_i) + (1 - x_i) \log P(x_i = 0|H_i)$$

where, n is the number of queries and x_i is the result from the beacon. $x_i = 1$ denotes the query is present in the database which can come either from the target genome or any of the other $N - 1$ genomes. x_i is only 0 when the query is not present in any of the N genomes.

In article [2], the authors using some simplifying assumptions proved that if the query individual is in the beacon database, $R = x_1, \dots, x_n$ follows a $Binomial(n, 1 - D_N)$ distribution, otherwise R has a $Binomial(n, 1 - \delta D_{N-1})$ distribution. Therefore, the hypothesis can be rewritten as follows:

- 1 Null hypothesis $H_0: \theta = \theta_0 = n(1 - D_N)$
- 2 Alternative hypothesis $H_1: \theta = \theta_1 = n(1 - \delta D_{N-1})$

Therefore, we have:

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

and for alternative hypothesis,

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

where D_{N-1} is the probability that other $N - 1$ individuals (all individual except the query individual) have not the specified allele in the determined location.

Basically, the $L_{H_i}(R)$ will maximize if the H_i hypothesis is correct. Therefore, we compute $\Lambda = L_{H_0}(R) - L_{H_1}(R)$ and the Λ will declare which hypothesis is true.

The log of the likelihood-ratio statistics can be rewritten from equation 1, 2 as,

$$\begin{aligned} \Lambda &= L_{H_0}(R) - L_{H_1}(R) \\ &= n \log\left(\frac{D_N}{\delta D_{N-1}}\right) + \log\left(\frac{\delta D_{N-1}(1 - D_N)}{D_N(1 - \delta D_{N-1})}\right) \sum_{i=1}^n x_i \\ &= nB + C \sum_{i=1}^n x_i \end{aligned} \quad (3)$$

In any distribution, a threshold t can be fixed where the null hypothesis will be rejected if $\Lambda < t$ and accepted otherwise. The attacker need to decide an appropriate threshold for a specific beacon dataset before

launching the attack. Suppose a false positive error α is given. Regarding this value and the beacon statistical properties, the threshold t_α is determined such that $Pr(\Lambda < t_\alpha | H_0) = \alpha$. From equation 3,

$$\begin{aligned} Pr(nB + C \sum_{i=1}^n x_i < t_\alpha | H_0) &< \alpha \\ Pr(\sum_{i=1}^n x_i > \frac{t_\alpha - nB}{C} | H_0) &< \alpha \quad (C \text{ is negative}) \\ Pr(\sum_{i=1}^n x_i > t'_\alpha | H_0) &< \alpha \end{aligned} \quad (4)$$

In the attack instead of calculating Λ and comparing it to the threshold t_α , $\sum_{i=1}^n x_i$ is computed and compared with t'_α to make the decision. This threshold t'_α is used to decide whether the null or the alternative hypothesis is correct. In other words whether the individual is present in the beacon database or not will be dictated by this t'_α . To calculate this, the adversary sums the responses from the beacon x_i and retrieves $\sum x_i$. The null hypothesis is rejected simply if $\sum x_i > t'_\alpha$ which leads to a conclusion that the query individual is present in the beacon and the attack is successful.

To calculate the D_N , the authors assumed that the adversary has an idea about the distribution of the allele frequencies on those query positions. Specifically, alternate allele frequencies, f for all SNPs observed in the population are claimed to be distributed as a β distribution according to [2]. Here, $f \sim \beta(a', b')$, where $a = a' + 1$ and $b = b' + 1$, and (a', b') can be precomputed from the genomic dataset in which the beacon service is running. Thus, the adversary needs $n \sim N^{a'+1}$ queries to make his or her decision whether the target individual is present in the database. The value D_N can be approximated as,

$$D_N \approx \frac{\Gamma(a+b)}{\Gamma(b)(2N+a+b)^a} \quad (5)$$

To see the details of deriving and proving the above formula see [2]. We will need this t'_α and D_N for further analysis in the upcoming section as these parameters dictate the attack.

2 Methods

In this section, we provide an analysis of equations 4 and 5 to calculate t'_α before describing our privacy preserving solutions and experimental results. This analysis allows the beacon service providers and data owners to calculate the risk involved while sharing their beacon data.

2.1 Risk analysis of a beacon service

In this section, we evaluated t'_α in greater depth for analyzing the risk involved for a specific beacon dataset.

Given N samples and n number of queries, this analysis will help us to determine the number of correct answers that can be returned without identifying the victim.

In other words, as t'_α directly effects the decision boundary of the correctness of null or alternative hypothesis, its better to theoretically ratify its value on a specific setting. We simulated this on some real life human genomic databases like 1000 Genomes Project, SSMP [12] and GoNL [13] for better understanding.

According to central limit theorem the value $R_e = \frac{1}{n} \sum_{i=1}^n x_i$ follows normal distribution $\mathcal{N}(1 - D_N, \frac{D_N(1 - D_N)}{n})$. The threshold t'_α can then be calculated from equation 4 as follows:

$$\begin{aligned} Pr(\sum_{i=1}^n x_i > t'_\alpha | H_0) &= \alpha \\ \Rightarrow Pr(\frac{1}{n} \sum_{i=1}^n x_i > \frac{t'_\alpha}{n} | H_0) &= \alpha \\ \Rightarrow Pr(R_e > t''_\alpha | H_0) &= \alpha \end{aligned}$$

where $t''_\alpha = \frac{t'_\alpha}{n}$. We know that R_e follows the $\mathcal{N}(1 - D_N, \frac{D_N(1 - D_N)}{n})$ distribution where $\theta_0 = 1 - D_N$ and variance $\sigma_0^2 = \frac{D_N(1 - D_N)}{n}$. Therefore, $\frac{R_e - \theta_0}{\sigma_0}$ has standard normal distribution, $\mathcal{N}(0, 1)$. Suppose we want to have a $\alpha = 0.05$ false positive probability then,

$$Pr(\frac{R_e - \theta_0}{\sigma_0} > \frac{t''_\alpha - \theta_0}{\sigma_0}) = 0.05 \quad (6)$$

According to the normal cumulative table and given the fact that $\frac{R_e - \theta_0}{\sigma_0}$ follows standard normal distribution, we have,

$$\begin{aligned} \frac{t''_\alpha - \theta_0}{\sigma_0} &= 1.65 \Rightarrow t''_\alpha = 1.65\sigma_0 + \theta_0 \\ \Rightarrow t'_\alpha &= n(1.65\sigma_0 + \theta_0) = n(1.65 \frac{\sqrt{D_N(1 - D_N)}}{\sqrt{n}} + \theta_0) \end{aligned}$$

Table 1 shows the computation of the t'_α for $\alpha = 0.05$ in different beacon databases constructed from real life genomic datasets according to [2].

The threshold t'_α indicates the number of *yes* that an adversary requires to conclude that the query individual is within the dataset. For example in table 1, the experimental results show that for '1k Genomes Phase 1' dataset with 1092 individuals, the adversary needs 1826 *yes* answers of queries to infer that the victim is present in the dataset. Any quantity less than 1826 *yes* answers (with mismatch rate $\delta = 10^{-3}$) will conclude that the individual is not present.

The relationship between null and alternative hypothesis along with the threshold are showed in Figure 2. In Figure 2, the black and the green lines represent the outputs distributions. If the null hypothesis is true, then the outputs follows black line and the outputs follows the green line if the alternative hypothesis holds. Three other real world datasets were tested and depicted in the supplementary document.

However, regardless of the risk analysis of a specific dataset with the outlined equations, the beacon service still needs privacy preserving mechanisms. The necessity of such methods are amplified due to the fact that these beacons are designed to support thousands of public queries. Though the simplified equations above can enlighten a data owner about the sensitivity of the underlying data, the data owner still needs some methods, which we will present in the next section, to protect the privacy of the individuals.

2.2 Proposed methods

In this section, we propose two privacy preserving methods which are similar in nature. However the probability of different outputs from these two methods are different. The methods are:

- 1 Method 1: Eliminating random positions
- 2 Method 2: Biased randomized response

Both of these methods introduce inaccurate results to hide the presence of any individual in a beacon web service. Due to this inaccuracy, these methods destroy some utility of the beacon service as the underlying data will be perturbed. Hence, we need to devise methods that give incorrect answers (false positives or negatives) as less as possible. A false positive is answering *yes* when the query result is *no*. Similarly, a false negative is answering *no* when the query result is *yes*. In Section 3, we experimentally evaluate these two methods in terms of data privacy and utility.

2.2.1 Eliminating random positions

In this method, the data owners apply algorithm 1 to output their data to the beacon service provider. The output of algorithm 1 will infuse inaccuracies in the result from the original data with the help of a driving factor *bias*. For example, if data owner has a dataset $\mathcal{D}$ then this method will transform it to a $\mathcal{D}'$ where there will be some false positives and negatives with respect to bias, b .

In algorithm 1, for higher bias value, we will get higher accuracy and for lower bias value, we will get lower accuracy. For example, if the bias value is 50, we will obtain answers with a probability of 50% false positives and negatives. We further analyzed different bias value for this algorithm in Section 3.

Algorithm 1: Eliminating Random Positions

```

Data: query, dataset, bias b
Result: Yes/No
random_value  $\leftarrow$  RandomValue(0,1);
true_result  $\leftarrow$  CheckPresence(dataset,query);
if random_value  $\leq$  b/100 then
|   return true_result;
else
|   return ! true_result;           // wrong answer

```

2.2.2 Biased randomized response

Randomized response [14] was proposed in 1965 by Warner as a statistical tool to remove potential bias and add a probabilistic noise to the answers. For example, data owners transform $\mathcal{D} \rightarrow \mathcal{D}'$ with respect to certain probability. In the original method, the person who has been asked a private question flips a coin. If it is tail then s/he answers truthfully. Otherwise, for head, s/he flips the coin again and responds truthfully for tail and provides opposite answer for head.

In a beacon service, we incorporated this method as beacon queries are considered private and their answers are in binary (yes or no) form. For example, a typical query inquires about the presence of a major and minor allele in a specific position of a chromosome. Algorithm 2 can transform the raw data according to the randomized response method and this transformed data can be used further to answer queries.

However, answering queries in this fashion will induce some error and the utility of the beacon services will be at question. Thus, we experimented on a biased randomized response where this 1/2 probability is modified for better utility or true results in algorithm 2.

In algorithm 2, we changed the dichotomous behaviour of general randomized response with a control variable named bias. Similar to algorithm 1, a higher bias will give more accurate result and will provide less privacy on the data. We showed the analysis for different bias values in Section 3.

However, there is a similarity between both the algorithms. Algorithm 1 returns true answer with probability b given $b \in [0, 1]$, while Algorithm 2 returns true answer with probability $1 - (1 - b)^2$. Therefore, Algorithm 2 with bias b_2 will be same as Algorithm 1 having bias $b_1 = 2b_2 - b_2^2$ where $b_1, b_2 \in [0, 1]$.

Theorem 1 *The response from Algorithm 2 is $|\ln(\frac{1}{(1-b)^2} - 1)|$ differentially private.*

Proof Lets fix a respondent and a randomized device (i.e., coin flip) with bias b and range $[0, 1]$. For a ‘Yes’

Algorithm 2: Biased Randomized Response

Data: *query, dataset, bias b*
Result: Yes/No
random_value $\leftarrow$ **RandomValue**(0,1);
true_result $\leftarrow$ **CheckPresence**(*dataset*,*query*);
if *random_value* $\leq$ *b*/100 **then**
 | **return** *true_result*;
else
 | *random_value* $\leftarrow$ **RandomValue**(0,1);
 | **if** *random_value* $\leq$ *b* **then**
 | **return** *true_result*;
 | **else**
 | **return** ! *true_result*; // wrong answer

answer from this respondent, we get

$$P(\text{Response} = \text{Yes} | \text{Truth} = \text{Yes}) = b + (1 - b)b$$

$$P(\text{Response} = \text{Yes} | \text{Truth} = \text{No}) = (1 - b)^2$$

Thus for ‘Yes’ answer we have,

$$\frac{P(\text{Response} = \text{Yes} | \text{Truth} = \text{Yes})}{P(\text{Response} = \text{Yes} | \text{Truth} = \text{No})} = \frac{b + (1 - b)b}{(1 - b)^2}$$

$$= \frac{1 - (1 - b)^2}{(1 - b)^2} = \frac{1}{(1 - b)^2} - 1$$

Similarly for a ‘No’ answer we have,

$$\frac{P(\text{Response} = \text{No} | \text{Truth} = \text{No})}{P(\text{Response} = \text{No} | \text{Truth} = \text{Yes})} = \frac{1}{(1 - b)^2} - 1$$

Since, both the probabilities are bounded by $\frac{1}{(1 - b)^2} - 1$, Algorithm 2 satisfies $|\ln(\frac{1}{(1 - b)^2} - 1)|$ differentially privacy. $\square$

For example, if the randomized device is a regular coin then we have bias $b = \frac{1}{2}$ in range $[0, 1]$. Thus the mechanism would be $|\ln(\frac{1}{(1 - \frac{1}{2})^2} - 1)| = |\ln(3)|$ differentially private [15].

3 Results

In this section, we evaluated both the methods according to the proposed attack [2]. We used similar experimental setup according to the original paper to directly benchmark our solution to the attack scenario. We also changed some of their population size and other parameters in order to do further analysis.

3.1 Original results

The original simulation [2] were experimented on a sample of 1,000 individuals containing 500,000 SNPs

which we doubled to 1,000,000 SNPs. Alternate allele frequencies of these SNPs were sampled from binomial distribution for a standard neutral model under the assumption of a population size of 20,000 individuals. Then the query beacon was constructed having 1,000 individuals (from 20,000). We also considered higher beacon size with 1200, 1500, 2000 individuals. Then the log likelihood ratio tests (LRT) to confirm the hypothesis were done assuming,

- 400 individuals from the beacon
- 400 individuals not from the beacon

The comparison between both setups are also shown in Table 2.

The outcome of the attack in our setting is depicted in Figure 4 where the power of the log-likelihood ratio tests (LRT) are on Y axis while the X axis shows the number of SNPs queried by any adversary. The figure demonstrates that the proposed attack has more than 95% power to detect whether an individual is present in the beacon of 1,000 individuals with just 5,000 SNP queries. This result also supports the claim of the original paper [2].

As Equation 5 shows the dependency between the LRT outputs and the beacon size (number of individuals, N), we further analyzed the attack for a different number of N . We show the power of the attack for different beacon size $N = \{1000, 1500, 2000\}$ in Figure 3.

In the supplementary document, we include the analysis for different genome sequencing error rate and re-identifying the relatives of those 400 individuals. For example, the relatedness (ϕ) can be defined as twins, parent-offspring, siblings, cousins etc., where $\phi = \{1, 0.5, 0.25, 0.125\}$. As twins share the same genomic sequence, the LRT tests should be similar and conclusive after 5,000 queries.

3.2 Our results

According to the test framework, we evaluated our proposed methods. We employed our privacy preserving mechanisms to perturb the original answer and then evaluated the performance of our techniques.

Figure 5 and 6, show the results of algorithm 1 and 2. As expected from the privacy-utility relation, we see that more accurate answer results in less privacy as the LRT powers keep rising for bias 90 (90% accuracy) after 300,000 queries. That is even with only 10% errors, the adversary needs more than 300,000 queries to determine the presence of an individual in the beacon database.

3.3 Accuracy Analysis

As mentioned previously, there is a need for a method which will induce errors to provide the privacy of the

individuals' present in a beacon service. Both of our methods add random errors to the beacon database where these errors can be defined as false positives and negatives. In this context, *false positives* are those where the beacon service answered yes regardless of the fact that there was no existence of that data. *False negatives* are those where the beacon answered false to a true answer. Accuracy is defined as,

$$accuracy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}$$

Figure 7 shows the calculation of the accuracy for both of our methods. It is clear from the figure that both methods with a higher bias provide more accurate result. This allows the corresponding data owner to decide the amount of utility they want to provide with respect to the privacy of the individuals.

We also show multiple levels of privacy achieved for different accuracy of the beacon data. Figure 8 shows the different LRT powers for different accuracy of the data in 300,000 queries. As higher LRT powers defines better assumptions from the adversary, we can model it as the privacy loss where utility can be defined as the accuracy aforementioned. It is noteworthy that, higher utility results in higher privacy loss as we can see with 98% accuracy we have LRT power as 1 where 75% accuracy has 0.22.

4 Discussion

In this section, we discuss few issues regarding the original attack and the applicability of our solutions.

4.1 Different bias on tiered access control

One clear indication from GA4GH and the research community on this privacy issue of genomic beacon service is implementing an access control over this sensitive information [1, 7]. Multiple layers of access control have been proposed where a different level of users will have different privileges over the beacon service. This kind of hierarchy in accessing a service is often named as 'tiered access control'. The applicability of this model in beacon service is already proposed in a recent study [6].

Our solution methods fit the tiered access control as we have different levels of privacy guarantee for different bias value. Higher bias leading to higher accuracy might be granted to a more trusted user where a public user might only get the lowest utility with high privacy over the beacon data. This will ensure the utility that the beacon promises while not revealing the presence of an individual.

4.2 Statistical inference attack

There are two different ways we can incorporate our methods on a beacon service. First, by using them while answering queries in real time and secondly, using them to preprocess the database beforehand to answer queries. In our analysis, we use the algorithms to preprocess the database due to the statistical inference attack. If we use the algorithms in real time, then an adversary might average the outputs of a specific query and obtain the original output.

4.3 Different allele frequency assumptions

The original attack scenario assumes that the allele frequency of the dataset follows a beta distribution [2]. However, in real life, an adversary can find the specific allele frequency of any position from a public database. This enables the adversary to launch more powerful attacks against the beacon service. It is noteworthy that iDASH 2016 competition [16] presented the problem under this formulation [17, 18].

However, in this paper we also assume that the adversary has limited background knowledge and s/he does not have access to specific frequencies of each position. More rigorous privacy guarantee like *differential privacy* [15] can be provided against a stronger adversary.

5 Related Work

Genomic privacy has recently gained significant concern among the general public and research community. De-identification is a common practice in research and clinical practice to protect genomic privacy and confidentiality of the participants. Normally, privacy is achieved by anonymizing a person's identity while sharing genomic related data. Since the de-identified genomic data are typically published with additional metadata and anonymized quasi-identifiers information, these pieces of information can be used to re-identify an unknown genome and thus disclosing the identity of the participant. Significant research has been done so far in this area. Below are some of the recent works related to re-identification attacks in genomic and health-related data.

In the recent study, Sweeney et. al [19] showed that participants in the Personal Genome Project (PGP) can be easily identified based on their demographics without even using any genomic information. They also stressed that 84% to 97% of the participants are correctly identified by linking the demographics to publicly available records such voter list and the name hidden in the attached documents.

Gymrek et al. [20], showed that a person's identity can be exposed via surname inference by profiling

short tandem repeat on the Y-chromosome and querying recreational genomic genealogy databases. In their study, they showed that by scanning two largest Y-chromosome genealogical websites, 10-14% US white male individuals are subject to surname inference attack. Moreover, when the attacker gains access to that target DNA sample, they can simply search available genomic databases with sensitive attributes (e.g., drug abuse). Hence, the person's identity with attributes can be easily found.

In recent study [21], Gitschier showed that a surname of an individual participating in HapMap database can be inferred by the combination of information from genealogical registries and a haplotype analysis of the Y-chromosome collected for the HapMap Project. In [22], the authors presented an attack that involves the association of DNA sequences to personal names, through diagnosis codes.

Zhou et al. [23] studied the privacy risks of releasing aggregate genomic data and showed that individuals participating in such research study can be easily identified and for some cases, their DNA sequences can be fully recovered. They have proposed a risk-scale system to classify aggregate data and a guide for their release.

Homer et al. [24] proved it is possible to detect the presence of an individual in a complex genomic DNA mixture even when the mixture contains only trace quantities of his or her DNA. They showed that an individual participating publicly released Genome Wide Association Study (GWAS) can be easily identified by his/her known genotypes and analysing the allele frequencies of a large number of SNPs.

Wang et al. [25] showed a higher risk that individuals can actually be identified from a relatively small set of statistics such as those routinely published in GWAS papers. Their first attack is the extension of Homer's attack and showed that the presence of an individual in the case group can be determined based upon the pairwise correlation among as few as a couple of hundred SNPs. The second attack can lead to a complete disclosure of hundreds of the participants SNPs, by analyzing the information derived from the published statistics.

In another study, Malin and Sweeney in [26] introduced re-Identification of Data in Trails (REIDIT) algorithms which link individuals genomic data to the publicly available records. They showed that it is possible to identify a person by looking at the unique features in patient-location visit in a distributed healthcare environment.

Other than these, there are multiple surveys available which summarizes and demonstrates some other attacks [27, 28].

6 Conclusion

Bustamante attack on beacon service presents a privacy problem of sharing genomic data publicly and demonstrates the need for further research to achieve genomic data privacy. In this paper, we analyzed Bustamante attack and provided a method to calculate the risk involved in sharing the genomic data. We proposed two simple privacy preserving solutions: eliminating random positions and biased randomized response. Our lightweight privacy preserving solutions ensure a good trade-off between data privacy and utility. Experimental results demonstrate that given higher bias, both the methods are able to provide high data utility.

Declarations

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MMA came up with experiments and detailed analysis of the techniques. RG provided the statistical interpretation of the original attack and helped in the drafts. MW and NM provided written drafts, detailed edits and assisted in analyzing and interpreting the methods. All authors have given approval for the final version of the manuscript.

Funding

The research is supported in part by the NSERC Discovery Grants (RGPIN-2015-04147) and University of Manitoba Startup Grant.

Acknowledgements

The iDash competition 2016 evaluation team for Track 1 greatly helped in evaluating and benchmarking the solution. We also thank the reviewers for their insightful comments.

Availability of data and material

Our implementations are readily available on github [29] which was pipelined in the original attack simulation [2].

Consent for publication

Not applicable

Ethics approval

Not applicable

Consent to participate

Not applicable

Author details

¹Department of Computer Science, University of Manitoba, Winnipeg Canada. ²Department of Mathematics, Faculty of Sciences, Bu-Ali Sina University, Hamedan Iran.

References

1. GA4GH: A federated ecosystem for sharing genomic, clinical data. *Science* **352**(6291), 1278–1280 (2016)
2. Shringarpure, S.S., Bustamante, C.D.: Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics* **97**(5), 631–646 (2015)
3. Krol, A.: Beacon Projects Cracks the Door for Genomic Data Sharing. <http://goo.gl/3fo1VU>. [Online; accessed 15-August-2016] (2015)
4. Al Aziz, M.M., Hasan, M.Z., Mohammed, N., Alhadidi, D.: Secure and efficient multiparty computation on genomic data. In: *Proceedings of the 20th International Database Engineering & Applications Symposium*, pp. 278–283 (2016). ACM
5. Privacy Risks in Genome Sharing Network. <http://goo.gl/rQqHkz>. [Online; accessed 19-August-2016] (2015)
6. Dyke, S.O., Dove, E.S., Knoppers, B.M.: Sharing health-related data: a privacy test? *NPJ Genomic Medicine* **1**, 16024 (2016)

7. Beacon Project Mitigates Privacy Risks While Maximizing Value of Responsible Data Sharing. goo.gl/1Uy0d1. [Online; accessed 1-November-2016] (2015)
8. Williams, R.: Toward Protecting Participants Privacy. <http://goo.gl/vwk7ok>. [Online; accessed 20-August-2016] (2015)
9. Beacon Network. <https://www.beacon-network.org/#/about>. [Online; accessed 14-August-2016] (2015)
10. Beacon FAQs. <https://genomicsandhealth.org/files/public/Beacon-FAQ.pdf>. [Online; accessed 14-August-2016] (2015)
11. Shi, X., Wu, X.: An overview of human genetic privacy. *Annals of the New York Academy of Sciences* **1387**(1), 61–72 (2017)
12. Wong, L.-P., Ong, R.T.-H., Poh, W.-T., Liu, X., Chen, P., Li, R., Lam, K.K.-Y., Pillai, N.E., Sim, K.-S., Xu, H., et al.: Deep whole-genome sequencing of 100 southeast asian malays. *The American Journal of Human Genetics* **92**(1), 52–66 (2013)
13. of the Netherlands Consortium, T.G.: Whole-genome sequence variation, population structure and demographic history of the dutch population. *Nature Genetics* **46**(8), 818–825 (2014)
14. Warner, S.L.: Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60**(309), 63–69 (1965)
15. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* **9**(3-4), 211–407 (2014)
16. iDASH Privacy security workshop 2016. <http://www.humangenomeprivacy.org/2016/competition-tasks.html>. [Online; accessed 23 December 2016]
17. Raisaro, J.L., Tramer, 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. Technical report (2016)
18. Genomic Beacon Privacy IDASH 2016. <https://github.com/mominbuet/GenomicBeaconPrivacyIDASH>. [Online; accessed 14 January 2017]
19. Sweeney, L., Abu, A., Winn, J.: Identifying participants in the personal genome project by name. Available at SSRN 2257732 (2013)
20. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. *Science* **339**(6117), 321–324 (2013)
21. Gitschier, J.: Inferential genotyping of y chromosomes in latter-day saints founders and comparison to utah samples in the hapmap project. *The American Journal of Human Genetics* **84**(2), 251–258 (2009)
22. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences* **107**(17), 7898–7903 (2010)
23. Zhou, X., Peng, B., Li, Y.F., Chen, Y., Tang, H., Wang, X.: To release or not to release: evaluating information leaks in aggregate human-genome data. In: *European Symposium on Research in Computer Security*, pp. 607–627 (2011). Springer
24. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.: Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet* **4**(8), 1000167 (2008)
25. Wang, R., Li, Y.F., Wang, X., Tang, H., Zhou, X.: Learning your identity and disease from research papers: information leaks in genome wide association study. In: *Proceedings of the 16th ACM Conference on Computer and Communications Security*, pp. 534–544 (2009). ACM
26. Malin, B., Sweeney, L.: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of biomedical informatics* **37**(3), 179–192 (2004)
27. Wang, S., Jiang, X., Singh, S., Marmor, R., Bonomi, L., Fox, D., Dow, M., Ohno-Machado, L.: Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the united states. *Annals of the New York Academy of Sciences* **1387**(1), 73–83 (2017)
28. Naveed, M., Ayday, E., Clayton, E.W., Fellay, J., Gunter, C.A., Hubaux, J.-P., Malin, B.A., Wang, X.: Privacy and security in the

genomic era. arXiv preprint arXiv:1405.1891 (2014)

29. Genomic Beacon Privacy Shringarpure. <https://github.com/mominbuet/GenomicBeaconPrivacyShringarpure>. [Online; accessed 14 January 2017]

Figures

Figure 1: **Beacon architecture** where researcher and data owners are connected to the central beacon service

Figure 2: **Risk analysis of data in any genomic dataset.** The green line represents the responds of the beacon service when the query individual being in beacon database while the black one represents them not being there. The red line denotes the t'_α .

Figure 3: **Power(LRT) of re-identification attacks** on beacons constructed with different number of individuals. We show the results of the attack for 3 different beacon database size (1000, 1200, 1500)

Figure 4: **Power(LRT) of re-identification attacks** of individuals on beacons constructed with 1,000 individuals on our experimental setting without any privacy preserving mechanism

Figure 5: **Effect of method 1** on the power of the attack on beacon database considering different bias

Figure 6: **Effect of method 2** on the power of the attack on beacon database considering different bias

Figure 7: **Accuracy** of both methods with different bias (50, 75, 90) consideration

Tables

Additional Files

Additional File 1 contains further analysis on the attack and the solution mechanisms.

Table 1: Evaluating the threshold $t'_{0.05}$ for three different databases with $\delta = 10^{-3}$ mismatch error.

Beacon Database	Number of Records N	a'	b'	D_N	t'_α
1k Genomes Phase 1	1092	0.0735	1.0096	0.0005594974767507827	1826
1k Genomes Phase 1 Affymetrix	1074	0.6483	1.2876	1.5352703647724165e-05	99084
GoNL	498	0.1131	0.8574	0.0009412979457329326	1005
SSMP	100	0.1848	0.8500	0.00403048895537907	234
Simulation	2000	0.1178793	1.1188360	0.00022374264418961542	4900

Figure 8: **Privacy-Utility curve** for different accuracy on X-axis and their corresponding LRT powers for 300,000 queries

Table 2: Parameter consideration in our experiment and the original paper [2] (1k=1000)

Parameter Name	Original paper [2]	Our setup
Population Size	10k	20k
SNPs considered	500k	1,000k
Beacon Size	1k	1k, 1.2k, 1.5k, 2k