The Author(s) *BMC Medical Genomics* 2017, **10**(Suppl 2):39
DOI 10.1186/s12920-017-0282-1

BMC Medical Genomics

RESEARCH Open Access

Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services

Zhiyu Wan[1*], Yevgeniy Vorobeychik[1,2], Murat Kantarcioglu[3] and Bradley Malin[1,2,4]

From iDASH Privacy and Security Workshop 2016
Chicago, IL, USA. 11/11/2016

Abstract

Background: Genomic data is increasingly collected by a wide array of organizations. As such, there is a growing demand to make summary information about such collections available more widely. However, over the past decade, a series of investigations have shown that attacks, rooted in statistical inference methods, can be applied to discern the presence of a known individual's DNA sequence in the pool of subjects. Recently, it was shown that the Beacon Project of the Global Alliance for Genomics and Health, a web service for querying about the presence (or absence) of a specific allele, was vulnerable. The Integrating Data for Analysis, Anonymization, and Sharing (iDASH) Center modeled a track in their third Privacy Protection Challenge on how to mitigate the Beacon vulnerability. We developed the winning solution for this track.

Methods: This paper describes our computational method to optimize the tradeoff between the utility and the privacy of the Beacon service. We generalize the genomic data sharing problem beyond that which was introduced in the iDASH Challenge to be more representative of real world scenarios to allow for a more comprehensive evaluation. We then conduct a sensitivity analysis of our method with respect to several state-of-the-art methods using a dataset of 400,000 positions in Chromosome 10 for 500 individuals from Phase 3 of the 1000 Genomes Project. All methods are evaluated for utility, privacy and efficiency.

Results: Our method achieves better performance than all state-of-the-art methods, irrespective of how key factors (e.g., the allele frequency in the population, the size of the pool and utility weights) change from the original parameters of the problem. We further illustrate that it is possible for our method to exhibit subpar performance under special cases of allele query sequences. However, we show our method can be extended to address this issue when the query sequence is fixed and known a priori to the data custodian, so that they may plan stage their responses accordingly.

Conclusions: This research shows that it is possible to thwart the attack on Beacon services, without substantially altering the utility of the system, using computational methods. The method we initially developed is limited by the design of the scenario and evaluation protocol for the iDASH Challenge; however, it can be improved by allowing the data custodian to act in a staged manner.

Keywords: Genomic databases, Beacon service, Data sharing, Privacy, Perturbation, iDASH challenge

* Correspondence: zhiyu.wan@vanderbilt.edu
[1]Department of Electrical Engineering and Computer Science, Vanderbilt University, 2525 West End Avenue, Suite 800, 37203 Nashville, TN, USA
Full list of author information is available at the end of the article

Background

Genomic data is increasingly collected by a wide array of organizations [1], ranging from direct-to-consumer genomics companies [2] to clinical institutions [3, 4]. This data serves as the basis of discovery-driven research [5, 6] and, more recently, for personalized medicine programs [7, 8]. However, as the quantity and coverage of genomic data grow, so too does the chance for the discovery and reporting of rare alleles [9, 10]. This is challenging for researchers and clinicians who aim to discern if such an allele (or combination of alleles across the genome) is meaningful with respect to an individual's phenotypic status or should influence the design of a personalized treatment regimen. To mitigate uncertainty, there is a desire to open data held by one organization to those who may need it elsewhere [11, 12]. While there are some initiatives, like the Personal Genome Project [13], that freely and publicly share genomic data linked to phenomic data, the existence of such systems does not necessarily translate into a large number of participants [14]. There are numerous reasons why individuals may not contribute their data to such programs, one of which is a privacy concern that the data would be misused or abused in some way [15, 16]. To mitigate privacy risks, data custodians have turned towards sharing summary level data about the pool of individuals who were in a study or were treated in a clinical setting.

The practice of summary data sharing began on a large scale in the mid-2000's, with programs like the Database of Genotypes and Phenotypes (dbGaP) at the National Institutes of Health [17], which aimed to standardize and centralize genomic data, making it easier to access. Summary statistics about the allele rates were made publicly accessible over the Internet because it was assumed that the privacy risks for such data were minimal. Yet in 2008, Homer and colleagues [18] demonstrated that an adversary could apply a statistical inference attack to discern the presence of a known individual's DNA sequence in the pool of subjects. This was specifically accomplished by measuring the distance between an individual's sequence to the allele rates exhibited by the pool versus some reference population, such as the International Haplotype Mapping Program [19] or 1000 Genomes [20]. When the target was deemed to be sufficiently biased towards the pool, the adversary could assign the hallmarks of the pool, such as membership in a specific group for case-control study (e.g., individuals positively-diagnosed with a sexually transmitted disease). As an artifact of this demonstration, the NIH, Wellcome Trust, and other genomic data custodians restricted access to summary-level genomic data [21, 22]. Since the initial attack, there have been a number of advancements in pool detection methodology (e.g., [23–27]).

As such inference methods evolved, the Global Alliance for Genomics and Health (GA4GH) formed to facilitate the sharing of genomic and health data in a federated manner [28]. In light of the known attacks, GA4GH created the Beacon Project, which enables data custodians to respond to queries through a web service (i.e., a beacon) about the presence/absence of a specific allele [29]. For instance, the Beacon service could respond yes/no to a question like, "Do you have any genomes with nucleotide A at position 121,212,028 on chromosome 10?" When the answer was affirmative, the requesting system user would learn that the variant in question may not be unique (i.e., because it was observed in a genome collected elsewhere) and that it might be worth pursuing further investigation into its meaning (possibly with the assistance of the answering data custodian).

Though it obscures allele rates, in late 2015, the Beacon service was also shown to be vulnerable to a statistical inference attack. Specifically, Shringarpure and Bustamante (SB) described the statistical theory behind the attack and illustrated how it might require no more than 5,000 responses to infer an individual's, or their relatives', membership in the pool [30].

Given the increasing adoption of Beacon, the Integrating Data for Analysis, Anonymization, and Sharing (iDASH) National Center for Biomedical Computing allocated one of the three tracks of their 2016 Genomic Privacy Protection Challenge to explicitly focus on this vulnerability. The organizers formulated the problem as, "Given a sample Beacon database, we will ask [the] participating team to develop solutions to mitigate the Bustamante attack. We will evaluate each algorithm based on the maximum number of correct queries that it can respond [to] before any individual can be re-identified by the Bustamante attack." [31] A subset of the authors of this paper developed the winning solution to this challenge. While this paper provides the details behind this solution, we have further extended our initial analysis to illustrate its limits as well as introduce alternative formulations of the problem to evolve the investigation into a setting closer to the real world.

The specific contributions of this paper include:

(1) We introduce a method that simultaneously optimizes the privacy (based on the SB attack as augmented by the iDASH Challenge organizers) and the utility of the system;
(2) We generalize the genomic data sharing problem to be more representative of scenarios in which beacons will actually be deployed; and
(3) We provide a sensitivity and robustness analysis of our method under various parameterizations of the variables relied upon by the iDASH Challenge.

Methods

The iDASH challenge

The goal of the first track of the 2016 iDASH Challenge was to mitigate an augmented version of the SB attack. The problem was how to find such a strategy for the genomic data custodian. This section begins with a description of the attack model and then models the data custodian's strategy as an optimization problem. In this setting, the attacker is defined as a malicious user launching the SB attack. The defender, by contrast, is defined as the data custodian sharing the genomic data while mitigating the SB attack.

iDASH variation of the SB attack

Given the binary genomic summary statistics of a pool of genomes (i.e., the beacon), the attacker relies upon a likelihood ratio test (LRT) to infer whether a targeted genome is in the pool or not. The null hypothesis, H_0, is that the targeted genome is not in the beacon, While the alternative hypothesis, H_1, is that the targeted genome is in the beacon. The attack model used in the iDASH Challenge is based on the SB attack, but it is amended to allow the attacker to know the alternative allele frequency (AAF) of all single nucleotide variants (SNVs) in the underlying population of the beacon. Here, AAF is the frequency at which the alternative allele occurs in a given population. The alternative allele is defined as the second most common allele in a commonly recognized global population (e.g., 1000 Genomes Project). We refer to this scenario as the augmented SB attack (ASBA).

Formally, the log-likelihood of a set of beacon responses $x = \{x_1, \cdots, x_m\}$ and a set of SNVs $d_i = \{d_{i,1}, \cdots, d_{i,m}\}$ for target i is:

$$L(d_i, x) = \sum\nolimits_{j=1}^{m} d_{ij}\left(x_j \ \log \ P(x_j = 1) + (1-x_j) \ \log \ P(x_j = 0)\right) \tag{1}$$

where d_{ij} and x_j are binary variables. Specifically, $d_{ij} = 1$ when SNV j of target i has at least one alternative allele and $d_{ij} = 0$ when SNV j of target i has no alternative alleles. Additionally, $x_j = 1$ when SNV j has at least one alternative allele in the beacon; $x_j = 0$ if SNV j has no alternative alleles in the beacon.

The null hypothesis can be formulated as:

$$L_{H_0}(d_i, x) = \sum_{j=1}^{m} d_{ij}\left(x_j log\left(1-D_n^j\right) + (1-x_j) \ log\left(D_n^j\right)\right) \tag{2}$$

And the alternative hypothesis can be formulated as:

$$L_{H_1}(d_i, x) = \sum\nolimits_{j=1}^{m} d_{ij}\left(x_j \ log \ \left(1-\delta D_{n-1}^j\right) + (1-x_j) \ log \ \left(\delta D_{n-1}^j\right)\right) \tag{3}$$

where δ is the sequencing error rate, D_n^j represents the probability that none of the n genomes in the beacon have an alternative allele at SNV j:

$$D_n^j = \left(1-f_j\right)^{2n} \tag{4}$$

Here, f_j is the AAF of SNV j in the population.

The LRT statistic for target i can thus be stated as:

$$\begin{aligned}\Lambda(d_i, x) &= L_{H_0}(d_i, x) - L_{H_1}(d_i, x) \\ &= \sum_{j=1}^{m} d_{ij}\left(x_j \ log\frac{1-D_n^j}{1-\delta D_{n-1}^j} + (1-x_j) \ log\frac{D_n^j}{\delta D_{n-1}^j}\right)\end{aligned} \tag{5}$$

Given this statistic, a threshold is selected, such that only targeted genomes with a test statistic below the threshold are regarded as being in the beacon. We assume that the attacker will always select the threshold according to a maximal allowable false positive rate (FPR).

To illustrate the ASBA, we present an example of the entire attack and defense process of the iDASH challenge in Fig. 1. For this example, we selected eight SNVs from chromosome 10 and populated a pool of 100 records in a beacon repository according to their global AAFs. Now, let us say the attacker has access to a set of genomes with known identities, which we refer to as a target set. The target set can be divided into two mutually exclusive subsets: i) a set of targets that are actually in the pool and ii) a set of other targets. The attacker will query the Beacon service about whether each alternative allele is in the pool behind the beacon and make their attack decision based on all of the returned answers. If the defender in control of the beacon server answers truthfully, the attacker is likely to achieve a high detection rate. However, if the defender invokes some data protection method (as we introduce later on), then the risk will be mitigated. A flipping action of "T" and "F" for a particular SNV position represents answering the query regarding this position truthfully and untruthfully, respectively. In this example, as shown in the bottom right corner of Fig. 1, it can be seen that the risk has been mitigated substantially.

Optimization problem

Strategies available to the defender

The above problem description implies that the only action the defender can take is to lie in their answers to the attacker's queries.

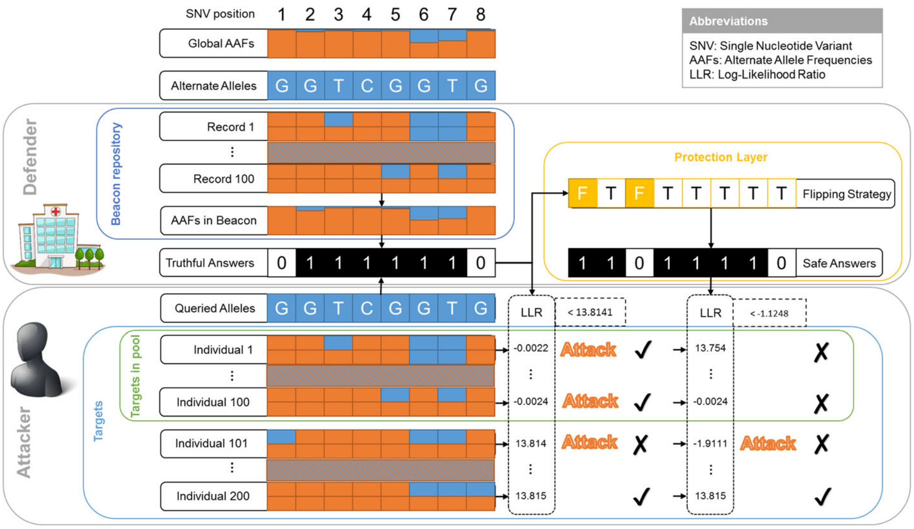


Fig. 1 An illustration of the ASBA attack and defense process of the iDASH challenge

In practice, the answer to a query from the attacker for a particular allele in the beacon can be 1, 0 and *null*, the latter of which means that the answer to the query is not applicable (e.g., when the defender does not have any records that cover the SNV of interest to the attacker). Henceforth, for simplicity, and alignment with the data analyzed in the iDASH Challenge, we assume that all data accessible through the beacon is a single nucleotide variant (SNV). In the attack model, the *zero* and *null* answers can be treated differently. The contributions from the SNV j to the final LRT statistics according to Equation (5) for answers 1, 0 and *null* are $log\frac{1-D_n^j}{1-\delta D_{n-1}^j}$, $log\frac{D_n^j}{\delta D_{n-1}^j}$, and zero, respectively. Lying about the answer to a query means two alternatives: 1) flipping or 2) masking. We define flipping as changing the answer from 1 → 0 or 0 → 1. We define masking as changing the answer from 1 → *null*, or from 0 → *null*. It should be recognized that we only consider the former type of lies for simplicity in presentation. In other words, we choose to flip each SNV or not. Now, let us say that $S_d = \{s_d\}$ is the set of the strategies available to the defender, where each strategy represents a set of SNVs. Then, the number of all available strategies is $|S_d| = 2^m$.

Query sequences available to the attacker

The effectiveness of the defender's strategy is influenced by the attacker's query sequence. We further assume the attacker has a pre-determined query sequence and has the potential to query all SNVs. Let us say that $S_a = \{s_a\}$ is the set of query sequences over all of the SNVs available to the attacker. Then the number of all possible query sequences is $|S_a| = m\,!$.

We also assume that the only uncertain action raised by the attacker for the defender is the SNV query sequence. All of the other parameters are fixed and known to the defender before he chooses the best strategy.

Objective function

Given this formulation, the iDASH Challenge scenario can be modeled as an optimization problem for the defender. Specifically, we wish to find a set of SNVs to flip that maximizes the utility and the privacy of the data simultaneously.

The effectiveness of a defender's strategy, $Y(s_d, s_a)$, considering both the utility and the privacy, is a function of both the defender's strategies and the attacker's query sequences, which is defined below. This creates a dependency on the definitions of the utility and privacy measures, the manner by which utility and privacy are combined, and the attack model.

As the attacker proceeds through the ordered set of SNVs, he runs a hypothesis test based on the responses for the subset of SNVs queried so far. Now, we assume the defender does not know the ground truth of the attacker's query sequence. As a consequence, the defender's best strategy is the one that maximizes his or her own expected effectiveness:

$$\begin{aligned} s^*{}_d &= \operatorname{argmax}_{s_d} E\Big(Y(s_d)\Big) \\ &= \operatorname{argmax}_{s_d} \frac{1}{|S_a|} \sum_{s_a \in S_a} Y(s_d, s_a) \end{aligned} \tag{6}$$

This implies that the attacker will choose any of the query sequences with equal probability. So long as this effectiveness function is defined and known to the defender, the simplest solution for the defender is to examine all available strategies. However, such a brute force approach is computationally challenging because the size of the defender's strategy space, $|S_d|$, increases exponentially with the number of SNVs.

Similarly, calculating the exact valuation of a defender's strategy, $E(Y(s_d))$, is difficult because of the large number of query sequences available to the attacker, $|S_a|$. However, the valuation of a defender's strategy can be estimated based on a subset of the attacker's strategies $S_a^{'}$, under a random selection model.

Evaluation criteria

There are many alternative definitions of effectiveness that could be invoked to evaluate the defender's strategies. Here, we first introduce the approach that was relied upon in the iDASH Challenge and then consider several alternatives. As will become evident, this is important because it influences how the defender will search for their best strategy.

In the description of the iDASH Challenge task, the effectiveness of a strategy was defined as the number of answers that can be served correctly before any targeted individual's presence is successfully detected. However, no matter what the defender's strategy may be, the detection power of the LLR test is always larger than 0 for the first several SNVs unless queries for these SNVs are not answered truthfully. This implies that at least one individual will be successfully detected if the first several SNVs are answered truthfully, which is highly likely because we assume the defender does not know the position of a specific query in the entire query sequence.

However, such a definition dictates towards a worst-case privacy scenario. Specifically, it assumes that a system is considered vulnerable if any one record can be breached. As noted earlier, there are alternative definitions that could be applied. For instance, a more pragmatic definition of the effectiveness of a protection model may be the proportion of correct answers that are returned before the presence of a certain number of individuals is detected. Under this formulation, when a method is evaluated, the iDASH Challenge organizers stated that 60%, instead of 0%, would be applied as the threshold for the detection power.

To define several alternative evaluation criteria, let us limit the utility and privacy measures in the [0, 1] range. For the purposes of the iDASH Challenge, utility can be regarded as the proportion of queries that are answered truthfully. By contrast, privacy is defined as a binary variable that is one if, and only if, a certain portion of the targeted individuals are never detected in the beacon, and 0 otherwise. Now, there are numerous ways to combine the utility and privacy measures into an effectiveness measure. In the iDASH Challenge, the effectiveness of the defense was defined as the utility for the proportion of SNVs shared before a certain portion of targeted individuals are re-identified.

Alternatively, the utility can be defined as a weighted sum of correct answers. In this scenario, each SNV can be weighted according to its importance (e.g., correlation with some phenotype). On the other hand, privacy can be defined as the expected false negative rate when the number of used SNVs is uncertain. The effectiveness of the defense can thus be defined as a weighted sum of the utility and the privacy.

Protection method

In this section, we start with a description of the solution we submitted to the iDASH Challenge. To perform a comprehensive empirical analysis, we then provide a description of alternative methods that could be applied to this problem.

Our iDASH submission

The solution we submitted to the iDASH Challenge entails searching through a collection of possible strategies that the defender can invoke to protect the system. Each of these strategies utilizes the same method, in the form of flipping some SNV query answers. In this section, we illustrate the principles by which such answers are flipped and how the strategy space is prioritized and searched.

Flipping responses

In the iDASH Challenge, it was assumed that the defender is not aware of the attacker's query sequence a priori and does not keep track of the queries. As a result, we need to find a defender's strategy that is independent of the attacker's query sequence. Since utility is basically measured as the number of queries that the defender responds to truthfully, we introduce the notion of discriminative power for each SNV in the pool. The *discriminative power* represents an SNV's ability to distinguish the records in the pool of individuals behind the beacon from a reference dataset. We define a *differential discriminative power* for each SNV in the pool, which represents the difference between its discriminative power before and after a flip. The top k percent of the SNVs in the pool, ranked by their differential discriminative power, will have their query responses flipped.

Here, we take a moment to formally define the discriminative power. We assume that the defender knows the attacker's target set (because, otherwise, the defender's strategy could not be directly measured). If this is not the case, then there are various ways for the defender to estimate the target set (the details of which are deferred to below). Let us say that the targeted pool with

n individuals and m SNVs is represented as a binary matrix with n rows and m columns $D_{n \times m} = \{d_{ij}\}$ in which d_{ij} is one when individual i in the pool has the alternative allele for SNV j and zero otherwise. Let us further say that the targeted reference dataset with n' individuals and m SNVs (the same as the SNVs in the pool) is represented by a binary matrix with n' rows and m columns $R_{n' \times m} = \{r_{ij}\}$ in which r_{ij} is one when individual i in the reference has the alternative allele in SNV j and zero otherwise.

Given the truthful answer x_j for the query regarding SNV j, the ability for SNV j to indicate if an individual is behind the beacon, according to the pool data, is defined as the average LLR for all the individuals in the pool if only SNV j is queried:

$$A_j(x_j) = \frac{1}{n}\sum_{i=1}^{n} d_{ij}\left(L'_{H_1}(x_j) - L'_{H_0}(x_j)\right) \tag{7}$$

where

$$L'_{H_0}(x_j) = x_j \ \ log\left(1-D_n^j\right) \ + \ \left(1-x_j\right) \ \ log\left(D_n^j\right) \tag{8}$$

$$L'_{H_1}(x_j) = x_j \ \ log\left(1-\delta D_{n-1}^j\right) \ + \ \left(1-x_j\right) \ \ log\left(\delta D_{n-1}^j\right) \tag{9}$$

Similarly, the ability for an SNV to indicate if an individual is behind the beacon, according to the reference dataset, is defined as the average LLR for all the individuals in the reference if only SNV j is queried:

$$A'_j(x_j) = \frac{1}{n'}\sum_{i=1}^{n'} r_{ij}\left(L'_{H_1}(x_j) - L'_{H_0}(x_j)\right) \tag{10}$$

The more similar these two values, the less powerful the LLR test will be. Based on this formulation, the discriminative power for SNV j becomes:

$$\begin{aligned} D_j(x_j) &= A_j(x_j) - A'_j(x_j) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} d_{ij} \ - \frac{1}{n'}\sum_{i=1}^{n'} r_{ij}\right)\left(L'_{H_1}(x_j) - L'_{H_0}(x_j)\right) \end{aligned} \tag{11}$$

The difference of the discriminative powers before and after flipping the SNV is:

$$\Delta D_j(x_j) = D_j(x_j) - D_j(1-x_j) \tag{12}$$

As a result, the first step of our Strategic Flipping method will flip the top k percent of the SNVs sorted according to the differential discriminative power. Note that, when a set of SNVs have the same differential discriminative power, the SNVs with the highest discriminative powers are selected first. When a set of SNVs have the same discriminative power, the SNVs with the lowest AAFs are selected first. A random selection is applied to break AAF ties.

Searching the strategy space

We use a greedy algorithm to search the defender's strategy space for a local optimum. To do so, we begin by randomly selecting q query sequences. Next, we traverse the l nearest neighbors of the current best strategy and find the neighbor with the best average measure of effectiveness, which is averaged across the q query sequences. Two strategies are regarded as neighbors if they only have one answer that is different. The distance between two neighbors is calculated as the absolute difference of the average number of answers provided truthfully by the two strategies and the rank of the different SNVs of these two strategies, in descending order, sorted by the differential discriminative power. In other words, if the number of the answers provided truthfully by these two strategy is t and t' and the rank of the different SNV is τ, then the distance between two neighbors is $|\tau - (t + t')/2|$. When l equals to two, only the top SNVs, in terms of differential discriminative power, are flipped.

We start from the result of the aforementioned Top-K Flipping step and keep searching until no strategy with better effectiveness can be found. In the case where two measures need to be optimized simultaneously (such as utility and privacy measures), we search the neighborhoods for a Pareto-optimal strategy. In a multiple-objective optimization problem, a strategy is Pareto optimal if no other strategies dominate it (i.e., is better than it in terms of both measures in consideration).

Alternative methods

We selected five alternative methods beyond the one we proposed above. The first three are baseline methods that set the upper and lower bounds on the measures. The last two are state-of-the-art methods in the literature that address the beacon detection problem. For reference purposes, we name our iDASH solution as the Strategic Flipping Method or M_{SF}.

(1) **Truthful Method (marked as M_T).** The defender simply responds to all queries truthfully. This sets the lower bound for the privacy measure and the upper bound for the utility measure.

(2) **Baseline Method (marked as M_B).** The defender flips the k percent of the SNVs with the lowest AAF in the underlying population from which the pool is sampled. This method was used by the organizers of the iDASH Challenge to establish a lower bound for the effectiveness measure.

(3) **Greedy Accountable Method (marked as M_{GA}).** We assume the users' queries are accountable, whereby each user has an account such that the defender documents the attacker's queries and results as they are processed. Upon submission of the next SNV query, the defender will flip the answer for this SNV if, and only if, the power of the resulting LLR test would be smaller than when no flip is applied.

(4) **Random Flips (marked as M_{RF}).** This method was recently proposed by Raisaro et al. [32]. In this method, the defender flips ε portion of SNVs that exhibit unique alleles in the beacon.

(5) **Query Budget (marked as M_{QB}).** This method was also proposed by Raisaro et al. [32]. In this method, a privacy budget is assigned to each individual in the beacon. Each time a record contains the queried allele, the budget for that user is reduced by a certain amount. Once a record's budget is exhausted, their genome will no longer contribute to responses provided by the beacon. Similar to M_R, this method requires the users' queries to be documented.

Results

In this section, we introduce the evaluation measures, the experimental setup, and the results of the empirical analysis.

Evaluation measures

In addition to the measure of effectiveness provided by the organizers of the iDASH Challenge, we propose several alternatives to assess the effectiveness of each method in a more comprehensive manner.

Utility

We measure utility according to the proportion of queries responded to truthfully (U).

Privacy

We measure privacy according to two different criteria, which we refer to as P_1 and P_2. First, we measure P_1 as a binary variable that is set to one if, and only if, a certain portion of the targeted individuals were never detected in the beacon, and 0 otherwise. We select 60% as the threshold because this is the definition used in the iDASH Challenge. It should be noted that our method generalizes to any threshold, but the results we present are limited to this parameterization. Second, we measure P_2 as the expected false negative rate when the number of SNVs to be queried is uncertain. We assume that the total number of SNVs about which the attacker has already queried, when they stop, is a random integer number uniformly distributed in the range [0, m].

Effectiveness

The effectiveness, which considers both utility and privacy, is measured according to two criteria. First, we measure E_1 as the proportion (in terms of all SNVs) of queries that are responded to truthfully before 60% of the individuals are successfully detected. Second, we measure E_2 as the proportion of truthful answers plus the expected false negative rate.

Computational efficiency

Computational efficiency is an important factor to be considered for deployment of solutions in a working system of beacons. As such, we also present the running time of each method. To measure the running time, we use a machine with Intel Core i7 quad-core 3.00GHz CPU and 8 GB memory.

Experimental design

To evaluate the effectiveness of the methods, we created a pool based on the first 400,000 SNVs in Chromosome 10. The pool is composed of 250 individuals randomly selected from the 2,504 individuals in Phase 3 of the 1000 Genomes Project [12]. The reference includes 250 individuals randomly selected from the remaining individuals in Phase 3 of the project. Note that the higher the association between the allele frequencies in the pool and the allele frequencies estimated by the attacker (as demonstrated by a sensitivity analysis), the higher the detection power. In our experiments, we assume a scenario where the allele frequency estimates available to the attacker are the allele frequencies in the entire population of 1000 Genomes. All of the other parameters are set as shown in Table 1. The maximal allowable false positive rate (α) is set to 0.05 and the sequencing error rate (δ or the mismatch rate) is set to 0.000001 according to Shringarpure and Bustamante [30]. The number of sampled query sequences is set to 10, and the number of examined neighbors is set to two for

Table 1 Parameter settings for the experiments

Parameter	Notation	Setting
Size of pool	n	250
Number of SNVs	m	400,000
Maximal allowable false positive rate	α	0.05
Sequencing error rate	δ	0.000001
Number of sampled query sequences	q	5
Number of examined neighbors	l	2
Percentage of flipped answers	k	5
Noise level in the Random Flip method	ε	0.75
Maximal allowable power in the query budget method	β	0.9

simplicity. The percentage of flipped answers is set to five, but this setting will be discussed further in the section of sensitivity analysis. The noise level in the Random Flip method is set to 0.75, and the maximal allowable power in the Query Budget method is set to 0.9 according to Raisaro et al. [32].

Findings

We compare all methods using the average results across 10 randomly generated query sequences. Fig. 2 shows how the detection power and the proportion of flipped answers (i.e., lies to the attacker) every 1000 queries change as a function of the number of queried SNVs for one of the query sequences.

By inspecting Fig. 2(a) and (b), it can be seen that the detection power does not increase monotonically when the defender is flipping answers. From Fig. 2(a), it can be seen that the three methods with the lowest (on average) detection power are M_{GA}, M_{SF}, and M_{QB}. Notably, none of these methods exceed the threshold of 60%. From Fig. 2(b), it can be seen that the three methods with fewest (on average) induced lies are M_{GA}, M_B, and M_T. Considering the intersection of these results, it appears that M_{RF} and M_{GA} are likely the best options.

Inspecting the result of only one query sequence provides some intuition into the trends of the utility and privacy measures, but it may be biased by a single run of the experiment. Fig. 3 summarizes the mean and +/-1 standard deviation for each of the performance measures across 10 query sequences.

Figure 3 reveals several notable findings. First, it can be seen that, according to the evaluation measures defined in the iDASH Challenge (E_1), our proposed method (M_{SF}) is the second best. However, since the best method (M_{GA}) assumes the users are accountable - which does not exist in the current system - it cannot be regarded as a practical solution for the iDASH Challenge.

With respect to effectiveness, we find that the two measures (E_1 and E_2) are in complete agreement regarding the rank order of the best methods. With respect to privacy, the second measure (P_2) does a better job of distinguishing between the methods than the first measure.

With respect to running time, we find that the solution we proposed (M_{SF}) exhibited the longest running time - about 56% longer than the quickest method. However, we assume that such a difference efficiency may not be critical because the defender's strategy for each beacon is determined once the beacon is published rather than calculated on-the-fly.

While there is no perfect way to combine the utility and privacy measures, we can compare the methods directly on these dimensions. In Fig. 4, for each method, utility (U) is shown by the x-axis, privacy (P_2) is shown by the y-axis, and effectiveness (E_1) is shown in text. It can be seen that, after dismissing the two impractical methods ($M_T$and M_{GA}), the method we proposed for the iDASH Challenge dominates all other solutions.

Sensitivity Analysis

To gain a deeper appreciation for the stability of the results of the iDASH Challenge, we assessed the performance of the proposed methods when certain key parameters are varied.

Tunable parameters

All of the protection methods, except M_T and M_{GA}, include a tunable parameter. Here, we systematically investigate how changes to the value of this parameter influences their performance. For brevity in presentation,

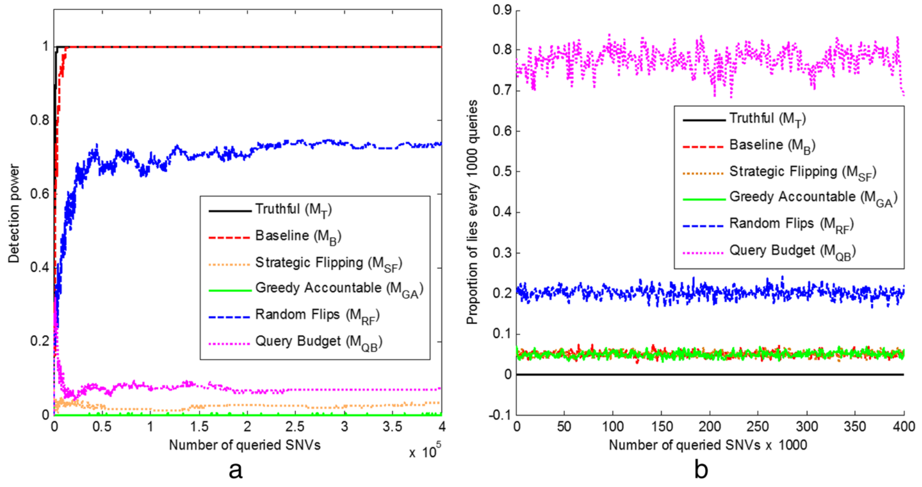
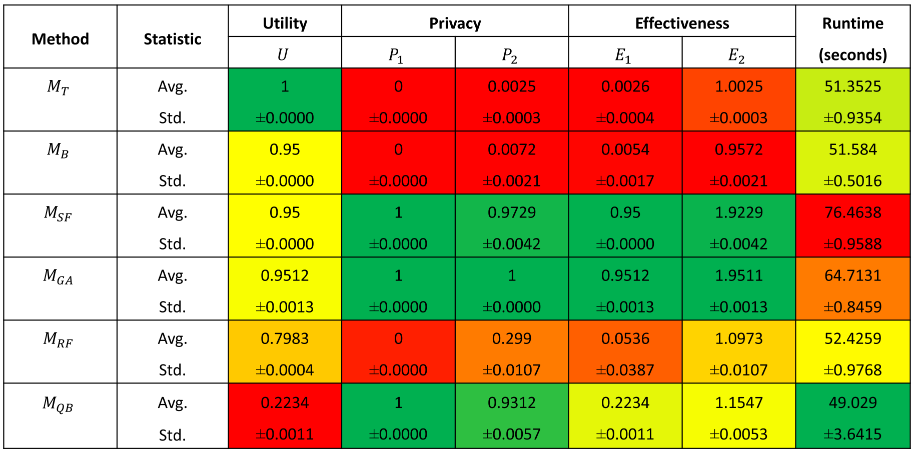


Fig. 2 How the number of SNVs queried influences the (**a**) detection power and (**b**) proportion of lies



Method	Statistic	Utility	Privacy		Effectiveness		Runtime
		U	P_1	P_2	E_1	E_2	(seconds)
M_T	Avg.	1	0	0.0025	0.0026	1.0025	51.3525
	Std.	±0.0000	±0.0000	±0.0003	±0.0004	±0.0003	±0.9354
M_B	Avg.	0.95	0	0.0072	0.0054	0.9572	51.584
	Std.	±0.0000	±0.0000	±0.0021	±0.0017	±0.0021	±0.5016
M_{SF}	Avg.	0.95	1	0.9729	0.95	1.9229	76.4638
	Std.	±0.0000	±0.0000	±0.0042	±0.0000	±0.0042	±0.9588
M_{GA}	Avg.	0.9512	1	1	0.9512	1.9511	64.7131
	Std.	±0.0013	±0.0000	±0.0000	±0.0013	±0.0013	±0.8459
M_{RF}	Avg.	0.7983	0	0.299	0.0536	1.0973	52.4259
	Std.	±0.0004	±0.0000	±0.0107	±0.0387	±0.0107	±0.9768
M_{QB}	Avg.	0.2234	1	0.9312	0.2234	1.1547	49.029
	Std.	±0.0011	±0.0000	±0.0057	±0.0011	±0.0053	±3.6415

Fig. 3 The performance of the genomic data protection methods across 10 runs

we designed four cases: two with values smaller and two with values larger than the value applied in the scenario investigated above. The specific values for the sensitivity analysis are detailed in Table 2. Only the most representative values are chosen for each parameter. The value of each parameter in Case three (Mid) is the default value as we used in the above experiment that simulated the iDASH Challenge. The values for each parameter in Case one (Low) and Case five (High) are the smallest and largest values, respectively, while Case two (Low-Mid) and Case four (Mid-High) provide gradations between these case to provide a more complete view.

Figure 5(b) displays the results of the sensitivity analysis with respect to the utility, privacy and effectiveness measures. The series of numbers near the series of circles represent the effectiveness (E_1) of the methods in different cases (from the Low Case to the High Case). There are several notable findings from

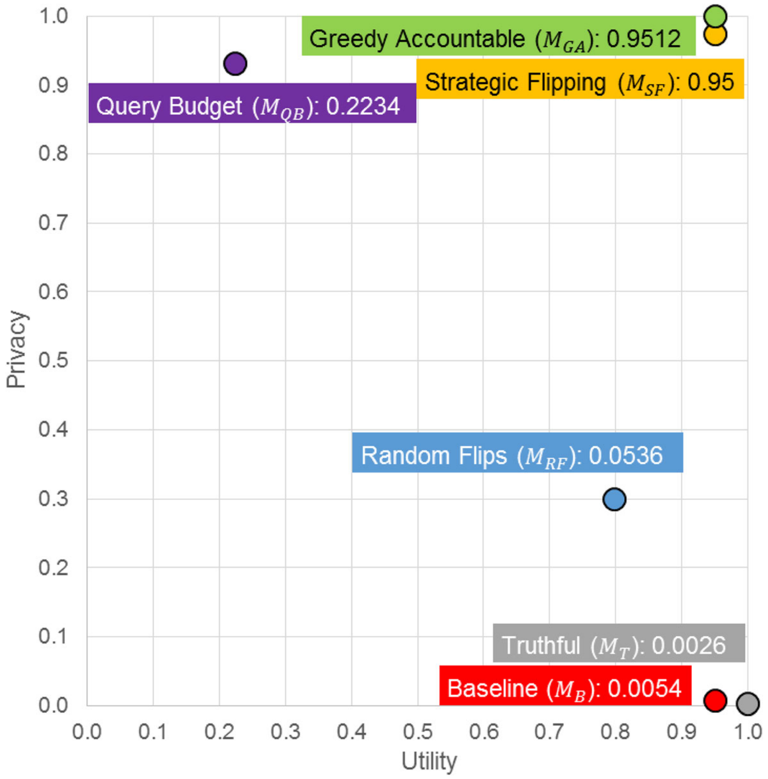


Fig. 4 A comparison of the genomic data protection methods with respect to utility and privacy

Table 2 Parameterizations for the sensitivity analysis of the genomic data protection methods

Method	Parameter	Case 1 (Low)	Case 2 (Low-Mid)	Case 3 (Mid)	Case 4 (Mid-High)	Case 5 (High)
M_B M_{SF}	Percentage of flipped answers (k)	1	2	5	10	20
M_{RF}	Noise level (ε)	0.1	0.5	0.75	0.9	1
M_{QB}	Maximal allowable power (β)	0.1	0.5	0.9	0.95	1

this analysis to highlight. First, it should be recognized that utility is negatively correlated with the parameter in methods M_B, M_{SF} and M_{RF} and positively correlated with the parameter in the method M_{QB}. Second, privacy (P_2) is positively correlated with the parameter in M_B, M_{SF} and M_{RF} and negatively correlated with the parameter in M_{QB}. Third, effectiveness (E_1) is positively correlated with the parameter in M_B, M_{RF} and M_{QB} and negatively correlated with the parameter in the M_{SF}. The effectiveness of our proposed M_{SF} method is larger than all of the alternative methods, including M_{GA}, when the value for the k parameter is smaller than five. Most importantly, we found that the method we proposed for the iDASH Challenge dominates all of the alternative methods - except M_T and M_{GA} when the value for the k parameter is two.

Allele frequency in the population

In the iDASH Challenge, it was assumed that the attacker's estimate of the allele frequencies is similar to those in the 1000 Genomes population of around 2500 individuals. However, in the real world, the attacker is likely to have a stronger capability. For instance, the attacker may know the allele frequencies are from a smaller population.

Thus, we investigated how an enhancement of the attacker's capability influences the performance of the methods. Specifically, we assess the performance of the protection methods when the attacker has a more accurate estimate of the allele frequencies by gaining access to a smaller population of only 500 individuals. We further examine the scenario where the attacker relies on allele

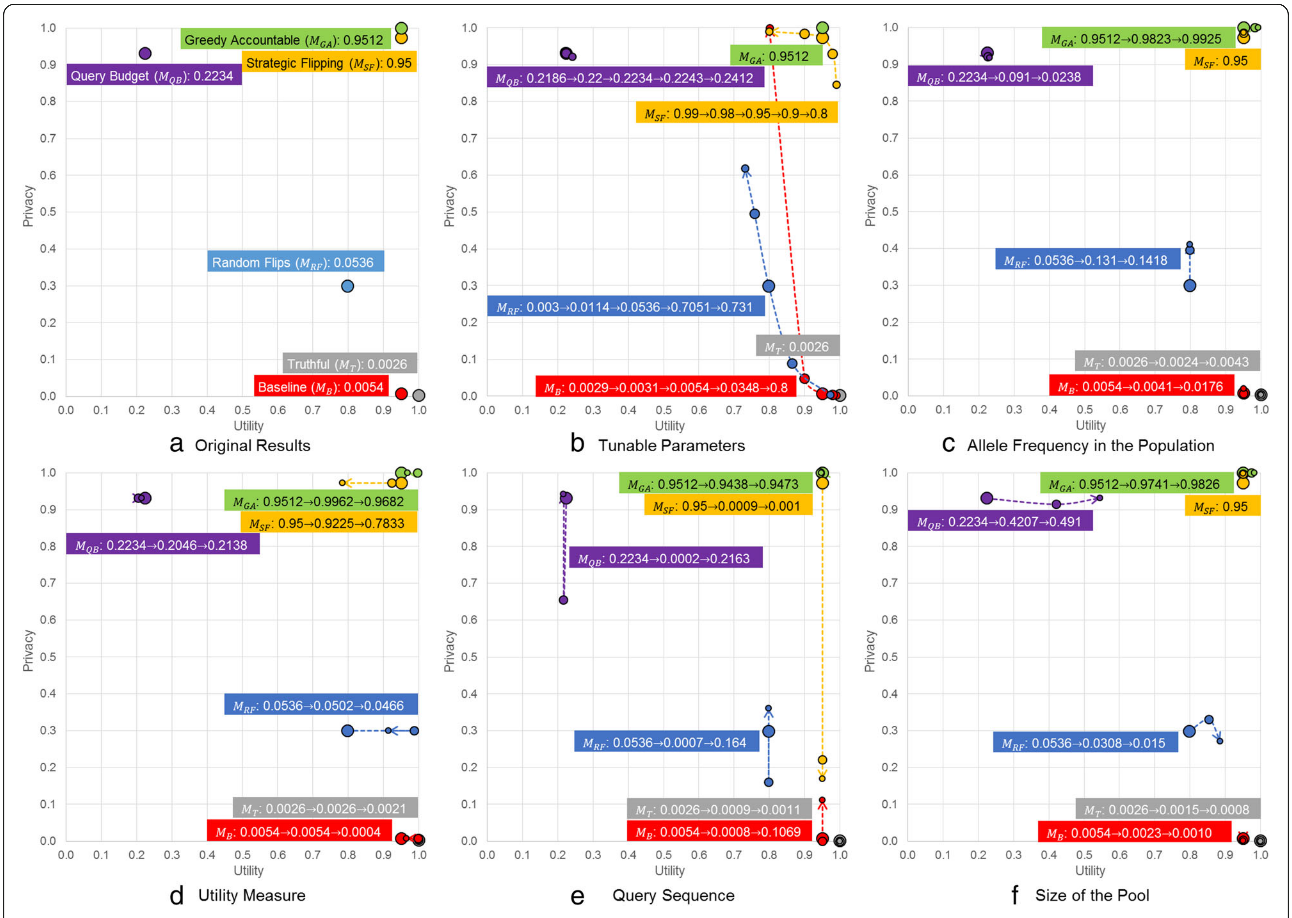


Fig. 5 Influence of the key parameters and factors on the utility and privacy measures. **a** Original Results. **b** Tunable parameters. **c** Attacker's knowledge about allele frequencies in the population. **d** The utility measure. **e** Query sequence. **f** Size of the Pool

frequencies that are exactly the same as the pool behind the beacon.

Figure 5(c) depicts the utility, privacy and effectiveness measures of different methods. The series of numbers near the series of circles (from larger to smaller circles) represent the effectiveness (E_1) of the methods in different cases (from larger to smaller populations). There are several important take-away messages from this analysis. First, both the effectiveness of protection (in terms of both E_1 and E_2) increases for M_B, M_{SF}, M_{GA}, M_{RF} and decreases for M_{QB} when the attacker's allele frequency estimate becomes more accurate. Second, the solution we proposed for the iDASH Challenge remains the best method for current application (noting that M_{GA} still remains the best overall).

Utility measure

In the iDASH Challenge, we assumed the utility for each SNV was equivalent. Here, we assume the utility is measured as a weighted sum of the truthful answers. The weight for each SNV can be (1) the absolute difference between the allele frequency in the pool and population or (2) a worst-case scenario for our proposed method, where the utility is equal to the absolute differential discriminative power the defender used for each SNV.

The results of this analysis are shown in Fig. 5(d). The series of numbers near the series of circles (from larger to smaller circles) represent the effectiveness (E_1) of methods in different cases (from the original setting to the Low Case, and from the Low Case to the Low-Mid Case). In these scenarios, we find that our proposed M_{SF} for the iDASH Challenge performs substantially worse than it did in the challenge because utility is measured differently. Nonetheless, even in such an extreme case, our method is never dominated by others, except M_{GA}, and dominates M_{QB}. However, it should be recognized when method A fails to dominate method B, it does not imply that method B dominates method A.

Query sequence

Different sequences of the queries have the potential to yield different detection results. As such, we need to consider how well the attacker can perform if he chooses the sequence with the highest possible detection power. Let us consider two scenarios: (1) the attacker always queries the most discriminative set of SNVs first; (2) the attacker always queries the rarest SNVs first.

The performance of the protection methods with respect to these two scenarios are depicted in Fig. 5(e). The series of numbers near the series of circles (from larger to smaller circles) represent effectiveness (E_1) of the methods in different cases (from the original setting to case 1, and from case 1 to case 2). As anticipated, it can be seen that the defender tends to lose privacy no matter what protection method is invoked in the first scenario. This is because the SNVs queried first have very strong discriminative power. However, in the second scenario, the defender loses privacy only when our proposed method is invoked and gains privacy when other methods are invoked.

These results primarily stem from three reasons. First, the SNVs with high differential discriminative power are not the same as the SNVs with high discriminative power or the SNVs with low alternative allele frequency. As a result, in the face of these two query sequences, the defender does not flip any SNVs until the very end of the query sequence. This leads to a high detection power quickly. Second, a flipping strategy works best in the scenario where all SNVs that need to be flipped are also queried first. In other words, our proposed method works best when the top k percent SNVs with highest differential discriminative power are queried first. By contrast, the baseline method works best when the top k percent SNVs with lowest alternative allele frequency are queried first. Third, Raisaro et al. [32] assumed that the rarest SNVs are queried first. Thus, the Random Flips method (M_{RF}) and the Query Budget method (M_{QB}) perform well in the scenario where this assumption holds true. Still, the solution we proposed for the iDASH Challenge is not dominated by any of the other methods except for the Greedy Accountable method (M_{GA}) and dominates the baseline method (M_B). Still, if the attacker's query sequence is fixed and known by the defender, the Greedy Accountable method (M_{GA}) becomes practical and the defender will end up with nearly perfect scores.

Size of the pool

In the iDASH Challenge, we used a dataset where there were 250 individuals in the pool behind the beacon. Here, we consider scenarios where there are fewer individuals in the beacon. Specifically, we assess the performance of the methods when there are only 1) 100 individuals and 2) 50 individuals in the pool. The results are shown in Fig. 5(f). The series of numbers near the series of circles (from larger to smaller circles) represent the effectiveness (E_1) of the protection methods in different cases (from larger to smaller sized pools).

It can be seen that the effectiveness of protection is positively correlated with the pool size in methods M_T, M_B and M_{RF}. And, once again, the method we submitted to the iDASH Challenge remains the best practical method in terms of all evaluation measures.

Discussion

The findings illustrate that the ASBA attack against a Beacon service can be sufficiently mitigated through a strategy that intelligently prioritizes which genomic variants to provide truthful answers about. However, we wish to point out that addressing the iDASH Challenge task is only the first step to solving the more general problem of mitigating the SB attack against the Beacon service in the real world. To create a more realistic solution, we wish to highlight two sets of limitations. The first set is an artifact of the design of the iDASH Challenge, while the second set is an artifact of the design of our model.

Limitations of the iDASH Challenge

There are, at least, three key limitations to the design of the iDASH challenge that hinder its practicality: 1) the evaluation measure, 2) the available strategy space, and 3) the construction of the attack model.

First, the iDASH Challenge regarded each SNV as having the same utility. This is critical to recognize because all of the defender's strategies trade between utility and privacy. In our proposed method, we first flip answers for the SNVs that will lead to the greatest loss of privacy. Yet, if these SNVs happen to have a more severe influence on utility than other SNVs, the best strategy for the defender will likely change.

Second, the Beacon service in the iDASH Challenge allows the defender to flip an answer from zero to one or from one to zero when answering the queries from a user. However, in the real world, the defender could have access to either a smaller or a larger strategy space. For instance, a smaller strategy space might be realized if a data custodian only allows for flipping from one (presence of a variant) to zero (absence of a variant) - in other words, the data custodian chooses to never lie about presence. By contrast, a larger strategy space might be created if an data custodian can play out their actions over a series of stages to query response. For example, as Raisaro et al. mentioned [24], if each user must create an account to query the beacons, the data custodian can distinguish between them and document their query history. As a consequence, the Query Budget (M_{QB}) and Greedy Accountable (M_{GA}) and methods will become practical, while the latter could become the optimal solution.

Still, the strategy space for the defender is likely to be much larger than what we have alluded to because they may have access to more than technical methods to lean on for protection. For instance, the defender may institute legal and/or economic methods to change the utility of the system for the attacker. For example, if each user must sign a data use agreement before querying a Beacon service, the malicious users might be detected as attempting to re-identify records and could be pursued as violators of a contract and penalized for liquidated damages (e.g., sued for a monetary loss). Note that such approaches will add minimal burden on typical law abiding users, but could come at a substantial cost for the defender. Still, in this scenario, a cost-benefit analysis becomes necessary for an economically-motivated defender, such that the defender will need to choose the strategy that optimizes his monetary payoff while satisfying the privacy requirements. And, when the attacker makes a decision based upon the defender's strategy, a game theoretic analysis should be invoked to solve this problem as shown by Wan et al. [33].

Third, both the original and augmented SB attack in the iDASH Challenge consider an adversarial model that tends to overestimate the privacy risk. This happens for several reasons. First, the attacker does not know the ground truth about who is in the pool. As such, he does not know the exact false positive rate. When he overestimates the false positive rate, fewer individuals will be attacked. Second, as pointed out by Craig and colleagues [25] the prior probability that a targeted individual is actually in the pool is likely to be much smaller than 50% as was the case for the iDASH Challenge, which can substantially influence the statistical inference power [33]. Third, it should be recognized that a possible attack is not the same as a probable attack. When an attacker is economically-motivated, he is unlikely to attack when the cost exceeds the benefit of an attack.

Limitations of the protection method

There are three primary aspects of our protection method that should be addressed before it is instituted in practice: 1) the deterministic approach to flips, 2) an estimation of the attacker's target set, and 3) a grounded approach to select k (i.e., the percentage of SNVs to flip).

First, our solution invokes a deterministic approach to selecting which SNVs to flip. This is potentially problematic because, if the attacker was able to ascertain some of the allele frequencies in the pool behind the beacon, then they could mimic the strategy of the defender. In other words, the attacker would be able to determine which SNVs the defender would choose to lie about. As a consequence, each query response for such SNVs could then be flipped back to the correct answer about the underlying pool, thus rescinding all of the protection. Therefore, in the event that there is a concern about such exposure, our model could incorporate a randomization component, where the answers provided to the adversary are non-deterministic. If such a feature were to be incorporated, it is critical to minimize the level of randomization to achieve the desired level of security.

Second, in the iDASH Challenge, the target set (i.e., the set of genomic records for presence/absence testing) was provided to the competition teams. However, in the real world, there may be multiple attackers, each of which may harbor a different target set. In such a scenario, the computation of the discriminative power for each SNV in the pool should be dependent on the underlying population of the beacon instead of a particular target set. In other words, when the defender is uncertain about the number of attackers and their targets, the entire population from which the beacon is sampled from should be used as the target set in our model. However, the mismatch in target sets may affect the performance of our method.

Finally, the parameter k in our method determines the starting point of the search for local optimal strategy. A well-specified value of k increases the probability that the local optimal strategy is also globally optimal. The best choice for k is dependent upon a number of factors, including 1) the size of the pool, 2) the number of SNVs, 3) the maximal allowable false-positive rate, 4) the specific data in the pool, and 5) the target set. In practice, when the defender needs to determine k, he could simulate an attacker with an estimated maximal allowable false-positive rate (which is often set to 5%), as well as a target set, and then select the best choice empirically. For example, according to the results in Fig. 4, the best choice of k is five in terms of maximizing the effectiveness (E_2).

Conclusions

This paper introduced a technical solution for mitigating the Shringapure and Bustamante (SB) attack on the Beacon service of the Global Alliance for Genomics and Health. This solution was specifically tailored to address an augmented version of this attack as posed in the 2016 iDASH Challenge. This solution was the winning entry and in this paper, we provided a formalization of the iDASH Challenge problem, a general design of the protection model, and an empirical evaluation of our solution to demonstrate its potential for protecting privacy with minimal influence on the utility of the system within a practical computational runtime.

We further showed, via an extensive experimental evaluation, that our proposed method outperforms all posited baseline and state-of-the-art methods (that were applicable to real world scenarios) regardless of how key parameters that drive the attack (e.g., the effectiveness measure, the number of records behind the beacon, and the attacker's estimate of allele frequency) vary. In most scenarios, the advantages of our method over other alternative methods are substantial. Still, it should be recognized that our method is limited by the design of the iDASH Challenge scenario (e.g., a strategy space limited to changing query answers) and the evaluation protocol (e.g., adversarial knowledge of minor allele frequencies).

Abbreviations

AAF: Alternative allele frequency; ASBA: Augmented SB attack; dbGaP: Database of Genotypes and Phenotypes; FPR: False positive rate; GA4GH: Global Alliance for Genomics and Health; iDASH: Integrating Data for Analysis, Anonymization, and Sharing; LLR: Log-likelihood ratio; LRT: Likelihood ratio test; SB: Shringarpure and Bustamante; SNV: Single nucleotide variant

Acknowledgements

The authors would like to thank the organizers of the iDASH Challenge for the design, data, and evaluation in the competition.

Funding

Publication charges for this article have been funded by the NIH. This research was sponsored, in part, by NIH grants R01HG006844, RM1HG009034, U01HG006385, and U01HG006378.

Availability of data and materials

The data that support the findings of this study are available from the 2016 iDASH Challenge. [http://www.humangenomeprivacy.org/2016/competition-tasks.html]. The computer programs that support the findings of this study are available on request from the corresponding author, or at the GitHub repository. [https://github.com/vanderbiltheads/SFM]

Author's contributions

ZW and BM designed the solutions. ZW developed the algorithms, conducted the experiments, and analyzed the results. ZW, YV, MK, and BM wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests in this research.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

About this supplement

This article has been published as part of BMC Medical Genomics Volume 10 Supplement 2, 2017: Proceedings of the 5th iDASH Privacy and Security Workshop 2016. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-10-supplement-2.

Publisher's Note

Author details

[1]Department of Electrical Engineering and Computer Science, Vanderbilt University, 2525 West End Avenue, Suite 800, 37203 Nashville, TN, USA. [2]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. [3]Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA. [4]Department of Biostatistics, Vanderbilt University, Nashville, TN, USA.

Published: 26 July 2017

References

1. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big data: astronomical or genomical? PLoS Biol. 2015;13(7):e1002195.
2. Philips AM. Only a click away – DTC genetics for ancestry, health, love… more: a view of the business and regulatory landscape. Appl Transl Genomics. 2016;8:16–22.
3. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet. 2013;14:295–300.
4. Taber KAJ, Dickinson BD, Wilson M. The promise and challenges of next-generation genome sequencing for clinical care. JAMA Intern Med. 2014; 174(2):275–80.

5. Green ED, Guyer MS, and the National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011;470:204–13.
6. Gottesman O, Kuivaniemi H, Tromp G, et al. The electronic medical records and genomics (eMERGE) network: past, present and future. Genet Med. 2013;15:761–71.
7. Collins F, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372:793–5.
8. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature. 2015;526:336–42.
9. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet. 2013;14:681–91.
10. Kobalt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013; 155(1):27–38.
11. ACMG Board of Directors. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. Genetics in Medicine. 2017; doi:10.1038/gim.2016.196.
12. Hayden EC. Geneticists push for global data-sharing. Nature. 2013;498(7452): 16–7.
13. Ball MP, Bobe JR, Chou MF, Clegg T, Estep P, Lunshof JE, Vandewege W, Zaranek AW, Church GM. Harvard personal genome project: lessons from participatory public research. Genome Med. 2014;6:10.
14. Sanderson SC, Linderman MD, Suckiel SA, Diaz GA, Zinberg RE, Ferryman K, Wasserstein M, Kasarskis A, Schadt EE. Motivations, concerns and preferences of personal genome sequencing research participants: baseline findings from the HealthSeq project. Eur J Hum Genet. 2016;24:14–20.
15. Hull SC, Sharp RR, Botkin JR, Brown M, Hughes M, Sugarman J, Bolcic-Jankovic D, Clarridge BR, Wilfond BS. Patients views on identifiability of samples and informed consent for genetic research. Am J Bioeth. 2008; 8(10):62–70.
16. Kaufman DJ, Muphy-Bollinger J, Scott J, Hudson K. Public opinion about the importance of privacy in biobank research. Am J Hum Genet. 2009;85(5): 643–54.
17. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. The NCBI dbGaP database of genotype and phenotypes. Nat Genet. 2007;39:1181–6.
18. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNV genotyping microarrays. PLoS Genet. 2008;4(8): e1000167.
19. International HalMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. A second generation human haplotype map of over 3.1 million SNVs. Nature. 2007;449(7164):851–61.
20. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR, And 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:68–74.
21. Felch J. DNA profiles blocked from public access. Los Angeles Times. August 29, 2008. URL: http://articles.latimes.com/2008/aug/29/local/me-dna29. Accessed 4 June 2017.
22. Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science. 2008; 322:44.
23. Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. Nat Genet. 2009;41:965–7.
24. Wang R, Li YF, Wang XF, Tang H, Zhou W. Learning your identity and disease from research papers: information leaks in genome wide association study. Proceedings of the 16th ACM Conference on Computer and Communications Security. 2009: pp. 534-44. http://dx.doi.org/10.1145/1653662.1653726.
25. Craig D, Goor RM, Wang Z, Paschall J, Ostell J, Feolo M, Sherry ST, Manolio T. Assessing and managing risk when sharing aggregate genetic variant data. Nat Rev Genet. 2011;12:730–6.
26. Gymrek M, MCGuire A, Golan D, Halperin EE. Identifying personal genomics by surname inference. Science. 2013;339(6117):321–4.
27. Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. Robust traceability from trace amounts. Proceedings of the 56th Annual Symposium on Foundations of Computer Science. 2015. pp. 650-69. https://doi.org/10.1109/FOCS.2015.46.
28. Knoppers B. International ethics harmonization and the Global Alliance for Genomics and Health. Genome Med. 2014;6(2):13.
29. Torres-Espanol M, Anvar SY, Sobrido MJ. Variations in the genome: the mutation detection 2015 meeting on detection, genome sequencing, and interpretation. Hum Mutat. 2016;37(1):1106–9.
30. Shringarpure SS, Bustamante CD. Privacy risks from genomic data-sharing beacons. Am J Hum Genet. 2015;97:631–46.
31. iDASH Privacy and Security Workshop. URL: http://www.humangenomeprivacy.org/2016/. Accessed 4 June 2017.
32. Raisaro JL, Tramer F, Ji Z, Bu D, Zhao Y, Carey K, Lloyd D; Sofia H, Baker D, Flicek P, Shringarpure SS, Bustamante CD, Wang S, Jiang X, Ohno-Machado L, Tang H, Wang X, Hubaux JP. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. Journal of the American Medical Informatics Association. 2017; doi:10.1093/jamia/ocw167.
33. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Malin B. Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach. Am J Hum Genet. 2017;100(2):316–22.