Real-time Protection of Genomic Data Sharing in Beacon Services

Diyue Bu, Xiaofeng Wang, Haixu Tang
School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA

Abstract *The acquisition of human genomic sequences is of increasing convenience and reduced expense. The sharing of these data is critical for biomedical researchers to study genomic loci or variants that are potentially associated with human diseases*[1]*. However, sharing genomic data broadly is impeded by privacy concerns. The statistical inference techniques for the re-identification of genomic data donors have been extensively investigated in the literature*[2–5]*. The Beacon services project is recently brought into view, aiming to test the willingness of data holders to share genomic data in a simple technical context: a query to ask a specified nucleotide at a given position within a chromosome*[6]*, also suffering from being compromised*[7,8]*. In this paper, we introduce a real-time mitigation method to protect Beacon services from re-identification attacks*[7]*, and show that it performs favorably in comparison with previous approaches on mitigation efficiency, i.e., with lower re-identification risks and higher utility of Beacon database.*

Introduction

With the help of high-throughput DNA sequencing techniques[9], human genomes can now be sequenced with low cost. Genomic datasets from large-scale projects such as 1000 Genomes Project[10] and HapMap Project[11] provided valuable resources for biomedical researchers to study genetic basis of human diseases. In addition, many medical institutes started to collect human genomic data from patients of various diseases, in particular complex diseases such as cancer. However, access and sharing of these valuable data in the biomedical community are impeded by potential privacy risks to the participants. Homer *et al.*[2] showed that inference techniques could be utilized to identify the presence/absence of an individual in a genomic dataset from aggregate statistics (*e.g.*, allele frequencies), even when the dataset contains thousands of human genomes. Numerous follow-up studies showed the existence of privacy risks in other types of genomic data[12–14], resulting in further reluctance of sharing human genomic data broadly for research uses. Nonetheless, sharing human genomic data for research purpose is not commonly regulated by privacy laws and policies[15,16], which is different from the scenario in clinical settings (*e.g.*, biomedical data including genomic data is regulated by the HIPPA compliance). Notably, the most sensitive genomic data largely overlap with those most useful in biomedical research, *e.g.*, human diseases are often associated with *rare alleles* that are carried by only a small fraction of individuals in the population, which poses great challenges to protect the privacy of data donors while preserving the utility of genomic data in biomedical research[17,18].

The Global Alliance for Genomics and Health (GA4GH), formed in 2013, serves as a platform for responsible genomic and health data sharing with consistent policy and interoperable standards and protocols[19]. The GA4GH has been taking the position of striking the balance between effective data sharing and responsible protection of individual privacy. Hence, it is important to explore and understand potential privacy risks in human genomic data[2–5]. The Beacon services[6] is a demonstration project led by GA4GH, aiming to provide a general public web service for disseminating human genomic data while ensuring sensitive information about participants is not leaked. A biomedical researcher may query the presence of a genetic variant in a set of genomic databases, each is owned by an independent institute, through the unified Beacon web service platform. The owner of each database can choose to register and host a Beacon service without releasing the whole dataset, which significantly reduces the privacy concerns for the data owners to share their data. The queries accepted by the Beacon services follow the forms like "Do you have any genomes with nucleotide A at position $114,235$ on chromosome 2?", while the responses from Beacon would be "Yes" or "No" (True/False answer). With such information, the users would learn whether the queried variant is present in any database registered at Beacon. No additional information (such as summary statistics of variants) is exposed to queriers. At present, the variants supported by Beacon services are only single nucleotide polymorphisms (SNPs); but the Beacon consortium plans to extend its services to other types of variants, including structural variations (SVs). The genomic information shared by Beacon is limited, but still useful: users may take further steps to get access to the whole genomic dataset (*e.g.*, by signing a user agreement with the data owner) if they know some genomes in a dataset carrying the variations of their interests. Overall, the Beacon services set an example of responsible and effective genomic data sharing with technical simplicity.

Even though Beacon services merely return the True/False answer for each query, genomic information leakage from the target genomic database still exists. A *re-identification* approach devised by Shringarpure and Bustamante (*SB* attack)[7] assumes that a malicious user (*attacker*) has obtained the genomic sequence of a victim, and attempts to determine if she is present in a *target* genomic database through a series of Beacon queries to the database. The *SB* attack is an inference attack based on a *log-likelihood ratio test* (LRT) to assess if the probabilities of obtaining the specific set of Beacon answers under two hypotheses (null hypothesis and alternative hypothesis) shown as in Formula 1 are significantly different. The power of the test, *i.e.*, $Pr(\textit{reject}H_0|H_1\textit{is true})$, indicates the confidence of the attackers can conclude that the victim (with queried variants) is present in the target database, which also measures the re-identification risk of an individual genome in a genomic database.

$$\begin{aligned} &H_0 : \textit{The queried victim's genome is not in the target database.} \\ &H_1 : \textit{The queried victim's genome is in the target database.} \end{aligned} \tag{1}$$

The log-likelihood statistic of R is calculated by Equation 2[7,8], where R is the set of Beacon answers $R = \{x_1, x_2, ..., x_n\}$ on a set of variants (SNPs). The equation could be further specified according to the hypothetical conditions H_0 and H_1 (Equation 3[7,8]), where D_N^i denotes the probability that none of the N individuals in the database carries the queried SNP. Under H_1, a genotyping error of δ is considered that allows a small probability of mismatches between the SNPs in the target database and those known by the attacker. The LRT statistic Λ (Equation 4[7,8]) is computed by the difference between $L_{H_0}(R)$ and $L_{H_1}(R)$.

$$L(R) = \sum_{i=1}^{n} x_i \log(Pr(x_i = 1)) + (1 - x_i)\log(Pr(x_i = 0)) \tag{2}$$

$$\begin{aligned} L_{H_0}(R) &= \sum_{i=1}^{n} x_i \log(1 - D_N^i) + (1 - x_i)\log(D_N^i) \\ L_{H_1}(R) &= \sum_{i=1}^{n} x_i \log(1 - \delta D_N^i) + (1 - x_i)\log(\delta D_N^i) \end{aligned} \tag{3}$$

$$\Lambda = L_{H_0}(R) - L_{H_1}(R) \tag{4}$$

The *SB* attack can be further strengthened by approximating the allele frequencies of SNPs in the target database using those in a public genomic dataset, which is available in real world (*e.g.*, through the 1000 Genomes Project[10]). Two attack models based on this strengthened LRT attack were proposed by performing different query orders of SNPs from the victim: the *rare-first* attack queries *rare* SNPs (the SNPs contained by only one genome in the database) first[8], while in the *discriminative-first* attack, the SNPs with higher discriminative power, *i.e.*, a SNP's capability to distinguish the records in the target database from those in a reference dataset, were queried first[20].

Mitigation methods were proposed to protect privacy risks in Beacon services from *SB* attack and other inference attacks (*e.g.*, the *rare-first* attack and the *discriminative-first* attack)[8,20,21]. The idea of these methods is to *flip* the answers to the queries of some variants: for some queried SNPs present in the genome database, the Beacon will answer "no" instead of "yes", and thus hide the truth about the presence of such variant in the database from the querier; on the other hand, for any variant absent in the database, the answer always remains "no" (*unflipped*). Note that these methods will always give consistent answers for the same query, because Beacon queriers are anonymous. In general, all these mitigation methods attempted to flip a small subset of queries to reduce the attack power measured by the LRT test on the entire set of queries, assuming all the answers can be obtained by a potential malicious user; but each method adopts a different approach to selecting the subset of queries. The performance of these methods can be compared based on the number of queries with flipped answers (better methods flip the smaller number of answers) and the remaining power of re-identification attack.

A common pitfall of existing mitigation methods is that they do not distinguish the queries targeting different individuals in the database, and thus select the queries with flipped answers solely based on their potential re-identification power. In a realistic application scenario of Beacon, most queries are submitted by biomedical researchers with research interests on different genomic variants. As a result, these variants are equally likely carried by each individual in the database. The queries for the re-identification purpose, however, always target the same individual. Intuitively, more protection should be made on the variants carried by the *vulnerable individuals*, whose variants have been queried more than the others in the database. Only one previous mitigation method considers the vulnerability of individuals in the database, i.e., the query budget per individual strategy[8]. However, this method removes all variants in an individual genome from the database permanently, whenever it is detected vulnerable, which would significantly reduce the data utility for open-access Beacon services. In this paper, we propose a Real-Time Flipping (RTF) method, which monitors the *vulnerability* (*i.e.*, the re-identification risk based on the query history) of each individual in the database with increasingly numbers of queries and decides if the answer to a new query should be flipped (if the queried variant is present) based on the entire query history. We compared the performance of the RTF method with other mitigation methods using genomic database consisting of the 1000 Genome Project data. The results showed that, when various query models (including the attack models such as rare-first and discriminative-first models and the model mimicking the real-world Beacon queries[22]) are adopted, RTF flips answers for fewer than 10% rare SNPs (carried by only one genome) or 4% of all SNPs in the database after millions of queries are made, while ensuring the re-identification confidence for every individual is below a threshold. In contrast, the other methods need to flip answers for much more queries, while the re-identification risks become high after thousands of queries are made. We implement the RTF method along with two other mitigation methods in the *secure-Beacon* system[23] extending the Beacon source code, which is readily used for mitigating re-identification risks in sharing genomic data, in particular for those acquired from vulnerable disease populations (*e.g.*, Autism patients).

Methods

In this section, we describe the workflow of the *secure-Beacon* system, and the real-time flipping method in comparison with existing mitigation methods that are also implemented in the system.

Existing Mitigation Methods

As discussed in the Introduction section, several mitigation methods were proposed to mitigate the re-identification risks in Beacon services. Here, we briefly review two methods, the Random Flipping[8] (RF) method and the Strategic Flipping[20] (SF) method which were implemented in the secure-Beacon system and were compared with our method. The other methods were not considered here for various reasons: the Query Budget method[8] and the Greedy Accountable method[20] assume that each user utilizing Beacon service has an account, which is not adopted by the current Beacon with the open-access mode (*i.e.*, queries are submitted by anonymous users); on the other hand, the Random Positions Elimination method[21] and the Biased Randomized Response method[21] were shown to perform worse than the Strategic Flipping method in terms of the genomic data utility, *i.e.*, the number of wrong answers ("yes" flipped to "no") to be returned under the same *SB* re-identification power of 0.6.[24] Therefore, we focus on these two mitigation methods (RF and SF), and compare them with our new method.

Random Flipping (RF) Method. The mechanism of the RF method is to randomly flip ϵ of rare SNPs in the beacon database, where ϵ is a constant, representing the proportion of incorrectly answered cases among all queried rare SNPs[8]. Previous studies showed that with $\epsilon \geq 0.15$, the power of the rare-first attack would not exceed 0.35 after querying $10,000$ SNPs in the database.

Strategic Flipping (SF) Method. As proposed by Wan *et al.*[20], the SF method flips the k percent of SNPs with the greatest *differential discriminative power*, which measures the discriminative power before and after flipping a SNP. As shown in the previous study, SF can keep the power of *SB* attack[7] below 0.1 after querying $400,000$ SNPs in the database, if k is set to be 5 or higher. We note that the discriminative power is calculated for each SNP in the database; hence, the flipped answers by SF may not be limited to the queries of rare SNPs. The answers to the queries of some common SNPs (shared by more than one genomes in the database) may also be flipped.

Real-time Flipping (RTF) Method

We propose a real-time flipping (RTF) method to mitigate the re-identification risks. Similar as the previous RF approach[8], RTF attempts to hide some rare variants in the database: when these variants are queried, the answer "no" instead of "yes" is returned. As shown in previous studies, by flipping a small fraction (15%) of rare SNPs, the RF method can reduce the re-identification power to an insignificant level[8]. The RTF method aims at further improving it by flipping the rare variants from more vulnerable individuals, whose variants were queried more frequently than other individuals. To achieve this goal, the RTF method monitors and records internally the *vulnerability* of each individual in the database by conducting the log-likelihood ratio test (LRT) based on the answers (including some flipped) to previously queried variants carried by the individual [*]. For each new query of a rare variant in the database, RTF will decide if the answer should be flipped depending on the vulnerability of the individual carrying the variant: the answer will be flipped if the vulnerability is sufficiently high after releasing the presence of this variant. After the answer is returned, the vulnerability of this individual would not be increased.

The workflow of the RTF method is laid out in Figure 1a. When a new query to a rare variant in the database is received, we start from assessing the privacy risk of the individual carrying the variant, and the decision whether or not the answer should be flipped will depend on the p-value of LRT (denoted as p_e) using previously queried variants from the individual. If the p-value is greater than 0.05 (*i.e.*, the re-identification risk below the significance of 0.05), the answer will not be flipped, which implies that in this case, the individual has low vulnerability, and thus no mitigation action is taken to protect her. In addition, if the p-value has not changed (within variations of 0.001) for 50 consecutive queries of rare SNPs from the target individual, the answer will not be flipped. In this case, the target individual's vulnerability is not expected to change by additional queries of her rare variants. In contrast, if the p-value of LRT is smaller than 0.05 and varies in each step, implying the vulnerability of the target individual is high, the answer will be flipped with a probability proportional to an *extreme function* that models the difference between the LRT scores of the target individual (ll_{target}) and a group of control individuals ($ll_{control}$). Specifically, we compute p_e as the percentage of LRT scores in the control group ($ll_{control}$) equal to or smaller than that of the target individual (ll_{target}), and the probability of flipping the answer is set to be $1 - p_e$ (rounded to 1 decimal).

Experiments And Results

In this section, we describe the experimental settings and the comparison results from two evaluation criteria: the re-identification risk and the data utility.

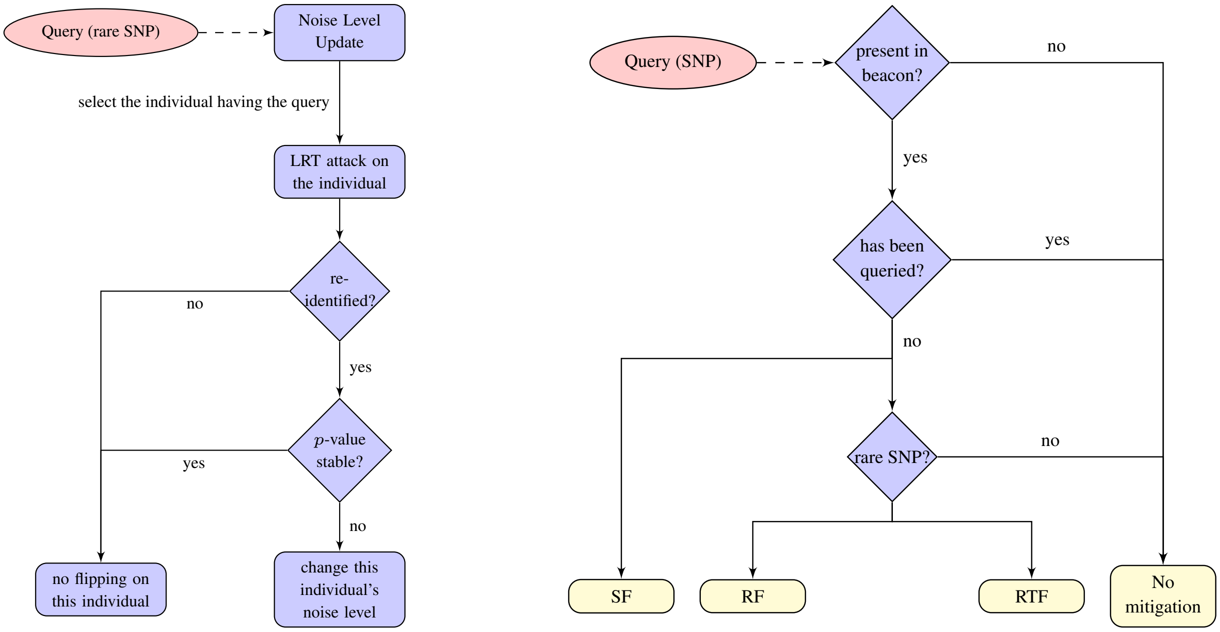
Experiments

To investigate the performance of the RTF mitigation methods and to align with the experimental settings of the alternative methods, we simulate a Beacon services database using a cohort of $1,235$ individuals, which are randomly selected from $2,470$ non-relative individuals in Phase 3 of 1000 Genomes Project[10]. The database contains a total of $3,992,219$ variants from Chromosome 10, in which $1,588,903$ (39.8%) are rare variants. The control cohort (individuals not included in the database) harbors 300 genomes randomly selected from the remaining $1,235$ individuals. We choose the moderate control group size to reduce the pre-processing time and the potential information leakage of the control cohort. The parameters in the other mitigation methods are set as their defaults: $\epsilon = 0.15$ in RF; $k = 5$ in SF.[8,20]

Figure 1b shows the workflow implemented in *secure-Beacon* and used in the experiments. When a new query is received, it is first to be checked if the target variant is present in the database. If it is not in the database, no mitigation method would be applied. If the query is present in the database but has been queried previously, the previous answer to this query will be returned without further alteration (and no mitigation method would be applied). Otherwise, all three mitigation methods will be applied, among which RTF and RF are applied only to rare variants, and SF is applied to both the rare and common variants.

As discussed previously[20], the query order has a strong impact on the performance of mitigation methods. To evaluate this impact on RTF in comparison with the SF and RF methods, we simulated the queries following four patterns: (1)

[*] Note that the same answer will be returned for the same query when it is submitted again (potentially by a different anonymous user).



(a) RTF Mitigation Method **(b)** Secure-Beacon Workflow

Figure 1: (a) RTF mitigation method. The method flips the answers to some rare SNPs. The vulnerability of each individual in the database is monitored using the LRT scores on previously queried variants. The decision of flipping the answer to the query of a rare variant is made based on the vulnerability of the target individual who carries the rare variant. The vulnerability of the target individual is updated and recorded internally after the answer is returned. (b) The secure-Beacon workflow. Three methods (RF, SF and RTF) were implemented for mitigating re-identification risks in an open-query Beacon system. For a newly-received query, the true answer ("yes" or "no") is first obtained. If query has been previously answered, the same answer will be returned. Otherwise, the mitigation procedure is activated. Note that RF and RTF are applied to rare variants while SF may be applied to both rare and common variants (see Text for details).

Random Order, where the variants are queried in a random order; (2) *Rare-first Order*, where the variants are queried in the increasing order of their allele frequencies; (3) *Discriminative-first Order*, where the variants are queried in the decreasing order of their discriminative level[20]; (4) *Typical User Order*, where the variants were queried in the order emulating a typical honest beacon user behavior. All four query patterns were emulated with the $3,992,219$ SNPs from chromosome 10 in Phase 3 reported in the 1000 Genomes Project[10]. To simulate query pattern (4), we generate queries based on the frequency distribution of the queries to the variants with various allele frequencies, which was obtained from the queries to the Beacon browser of ExAC[22] logs over a period of 12 weeks. $1,345,291$ queries in total are asked on $934,680$ SNPs in ExAC under Open Database License (ODbL). Table 1 shows the distribution of queried SNPs in terms of their allele frequencies. We note that in the ExAC statistics, the second category "<0.001 not singleton" is specified as "0.0001 - 0.001". This range overlaps with the allele frequency of singletons (rare variants) for the size of the simulated database. Alternatively, we emulate queries for this category by defining the modified range.

To further investigate the behavior of RTF, RF and SF under query patterns that are more frequently used in real world (*Random Order* and *Typical User Order*), we included more variants (in the genomes from the same Beacon cohort) in the database and performed additional queries under *Random Order* and *Typical User Order*: a total of $1,054,447$ variants from Chromosome 21 are imported into the database, among which 39.2% are rare variants (similar to that of Chromosome 10). The re-identification risk and the data utility of three mitigation methods were evaluated respectively. The results indicate that the behavior of the mitigation methods remains stable after a large number of queries ($>$ 1 million) were performed (see Results section for details).

The log-likelihood statistics of control group are required for computing the LRT score. In RTF, for the sake of computing efficiency, these statistics were computed in advance before performing query experiments, where they

Table 1: Proportions of queries in each allele frequency range

Allele frequency	**Singleton**	**$<$ 0.001 not singleton**	**0.001 - 0.01**	**0.01 - 0.05**	**0.05 - 0.5**	**0.5 - 1**
Queries in ExAC	0.434	0.418	0.0076	0.023	0.033	0.014

are recorded individually in a table when every new query is performed based on each of the four query patterns, respectively. During the experiment, for each new query (to a variant from the target genome in the database), the corresponding control group log-likelihood statistics (with the same set of queried SNPs) are extracted from the pre-computed and used in the LRT assessment. In addition, the location of SNPs (chromosome and positions) and the final returned answers ("yes" or "no", flipped or not) of the queried variants are recorded in a reference table for further queries to the same SNPs, as mentioned in the Methods section.

Evaluations

To evaluate the mitigation methods, we assume the size of Beacon services database (the total number of genomes in the database) and the minor allele frequencies in the beacon database are known to potential malicious users. In reality, we note that the minor allele frequencies in beacon services database are not publicly accessible, but could be approximated by using public genomic datasets (*e.g.*, 1000 Genomes Project[10]).

The performance of RTF in comparison to RF and SF is evaluated in terms of the re-identification risk (measured by the power of LRT) and the data utility (measured by the total number of flipped answers). A method is preferred if it has lower risk (*i.e.*, lower power of LRT) and higher data utility (fewer number of flipped answers).

Re-identification Risk. We evaluate the power of LRT by the percentage of re-identified genomes at a 5% false positive rate[8], which also represents the confidence of re-identification. The greater power indicates the higher re-identification risk.

Utility. To serve the purpose of Beacon services which is to share as much genomic information, the mitigation method should hide only the minimum amount of information (*i.e.*, the variants present in the database), and thus returning the un-modified answers to the queries by most users to variants of their interests. To achieve this goal, we define the utility of the database as Equation 5. To better evaluate the data utility with increasing number of queries, we computed the utility of each mitigation method with the number of flipped answers (Equation 6) .

$$utility = 1 - \frac{\textit{the number of flipped answers}}{\textit{the total number of queries}} \tag{5}$$

$$\textit{percentage of flipped answers} = \frac{\textit{the number of flipped answers}}{\textit{the total number of queries}} \tag{6}$$

Computational Environment. For a mitigation method to be practical, it should be very efficient to make the flipping decision so that the Beacon platform can return the answers to queriers instantly. Therefore, we evaluate the running time of RTF on a 2.60GHz Intel Xeon CPU and 8GB memory, except for pre-processing steps, which can be conducted only once in advance. The running time of the alternative flipping methods were not measured, because their mitigation step could be calculated in advance before any query is submitted.

Results

Re-identification Risk. We performed LRT on each individual in the database ($1,235$ cases in total) and measured the power of LRT with increasing number of queries. For RTF, the power trend does not vary significantly under different query patterns, because the re-identification risks are updated based on the p-values of LRT. In contrast, RF and SF perform differently given various query patterns, which agrees with the previous findings[20]: RF yields dominant result (less power) under *Rare-first Order* while SF performs better under the other three query patterns. Accordingly, we compare RTF to RF under *Rare-first Order* and to SF under the remaining three query orders. Under *Rare-first Order*, RF does not yield great power (< 0.1) until $1,000$ rare SNPs are queried when the power increases to 1.0. Under *Random Order*, *Discriminative-first Order* and *Typical User Order*, SF does not yield great power (< 0.1) until $1,000$ rare SNPs are queried when the power increases to 1.0. For RTF, it does not yield great power (< 0.1) across the process when $\sim 120,000$ rare SNPs are queried: the power of LRT remains < 0.05 (false positive rate) until $1,000$ rare SNPs are queried when the power increases to ~ 0.3, which is still an acceptable risk level[8]. Based on Raisaro *et al.*[8], a power larger than 0.6 indicates an existing re-identification risk and the greater power suggests the increasing re-identification risk. There is a 100% re-identification risk when the power reaches 1.0. We note that in the simulated database, the rate of rare SNPs on each individual is ~ 0.004, which means the total number of SNPs queried on an individual is $\sim 250,000$ (the scale used in Raisaro *et al.*[8]), and the total number of SNPs queried among the whole database is $\sim 3,600,000$ (the scale used in Wan *et al.*[20]). For both cases, RTF allows for much more queries to be made than two previous methods (RF and SF) while still keeping the re-identification risks low.

Utility. The percentages of flipped rare SNPs and total flipped SNPs are shown in Figure 2 and Figure 3, respectively. In Figure 2, we notice that the trends are almost the same for *Random Order*, *Rare-first Order* and *Discriminative-first Order*, when the number of queried rare SNPs are about the same. Under *Typical User Order*, the trend is slightly different because the proportion of queries to rare SNPs among all queries is different. This observation is consistent with the formulation of LRT, where each rare SNP contributes the same amount to the overall re-identification risk because only its allele frequency and the answer are used[7]. For the total number of flipped queries in the database, the figures vary across different query order because the rare SNPs are queried at various steps in different query patterns. The trends of the number of flipped answers to rare SNPs and that of total number of flipped SNPs agree with each other. The slope of each line indicates proportion of flipped answers. RTF flips more SNPs at the beginning of the query sequence. However, in terms of rare SNPs, RTF flips fewer than RF after $\sim 250,000$ queries, and fewer than SF after $\sim 400,000$ queries; in terms of all the SNPs in the database, RTF flips fewer than RF after $\sim 600,000$ queries, and fewer than SF after $800,000 - 1,000,000$ queries. Under *Discriminative-first Order*, RTF always flips much fewer than RF and SF. We note that the RF and SF lines are computed based on their mechanisms, rather than the experimental results that may contain some artificial fluctuations at the beginning of query sequences.

Computational Efficiency. For the implementation of RTF, each mitigation step is performed at back end with no impact on answering subsequent queries. Furthermore, this step only takes 0.15 seconds in average for each query. Although it is slower than the operation time of returning an answer (< 0.001 seconds in average), it is still practical in a real-world Beacon system.

Conclusion

In this paper, we propose a real-time mitigation method that enables the Beacon services for sharing human genomic data while protecting the participants' privacy from re-identification attempts, including *SB* attack and other strengthened attacks. Our method monitors privacy risks of each individual in the database and decides if answers should be flipped for the queries to variants of each individual respectively, to achieve high efficacy of risk mitigation *i.e.*, lower re-identification risk and higher data utility. We also implement efficient algorithms that enable the mechanism to be embedded into a Beacon web platform, resulting in a practical *secure-Beacon* system for providing strong privacy protection in Beacon services when it is used for sharing highly sensitive human genomic data.

Acknowledgements

This work was funded in part by the NIH (Grant No: R01HG007078 and U01EB023685), NSF (Grant No: CNS-1408874) and the Indiana University Initiative of Precision Health.

References

1. Ching Lee Koo, Mei Jing Liew, Mohd Saberi Mohamad, and Abdul Hakim Mohamed Salleh. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed research international*, 2013, 2013.

2. Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of

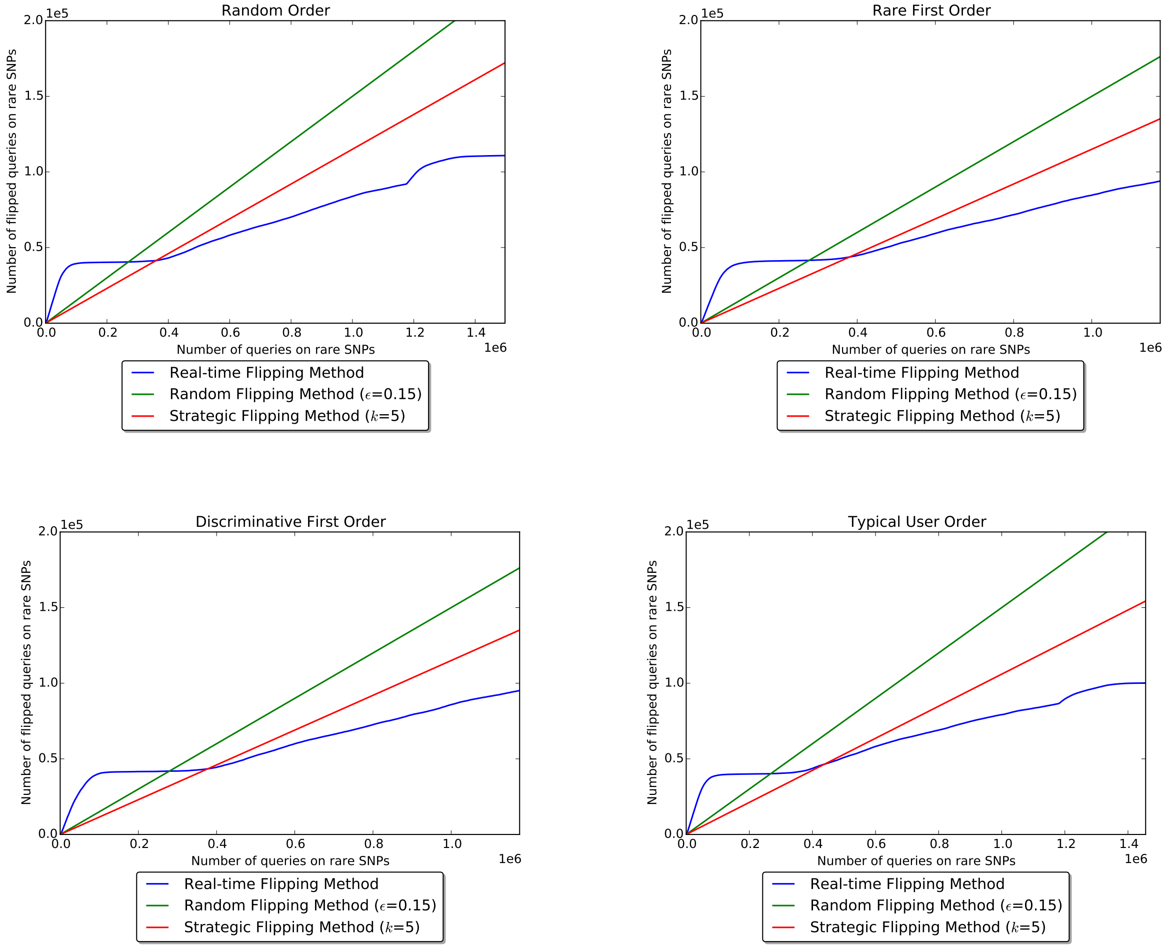


Figure 2: The percentage of flipped rare SNPs under different query patterns

dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

3. Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

4. Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.

5. Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

6. Miroslav Cupak. Beacon network: A system for global genomic data sharing.

7. Suyash S Shringarpure and Carlos D Bustamante. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5):631–646, 2015.

8. Jean Louis Raisaro, Florian Tramèr, Zhanglong Ji, Diyue Bu, Yongan Zhao, Knox Carey, David Lloyd, Heidi Sofia, Dixie Baker, Paul Flicek, et al. Addressing beacon re-identification attacks: quantification and mitigation of privacy risks. *Journal of the American Medical Informatics Association*, page ocw167, 2017.

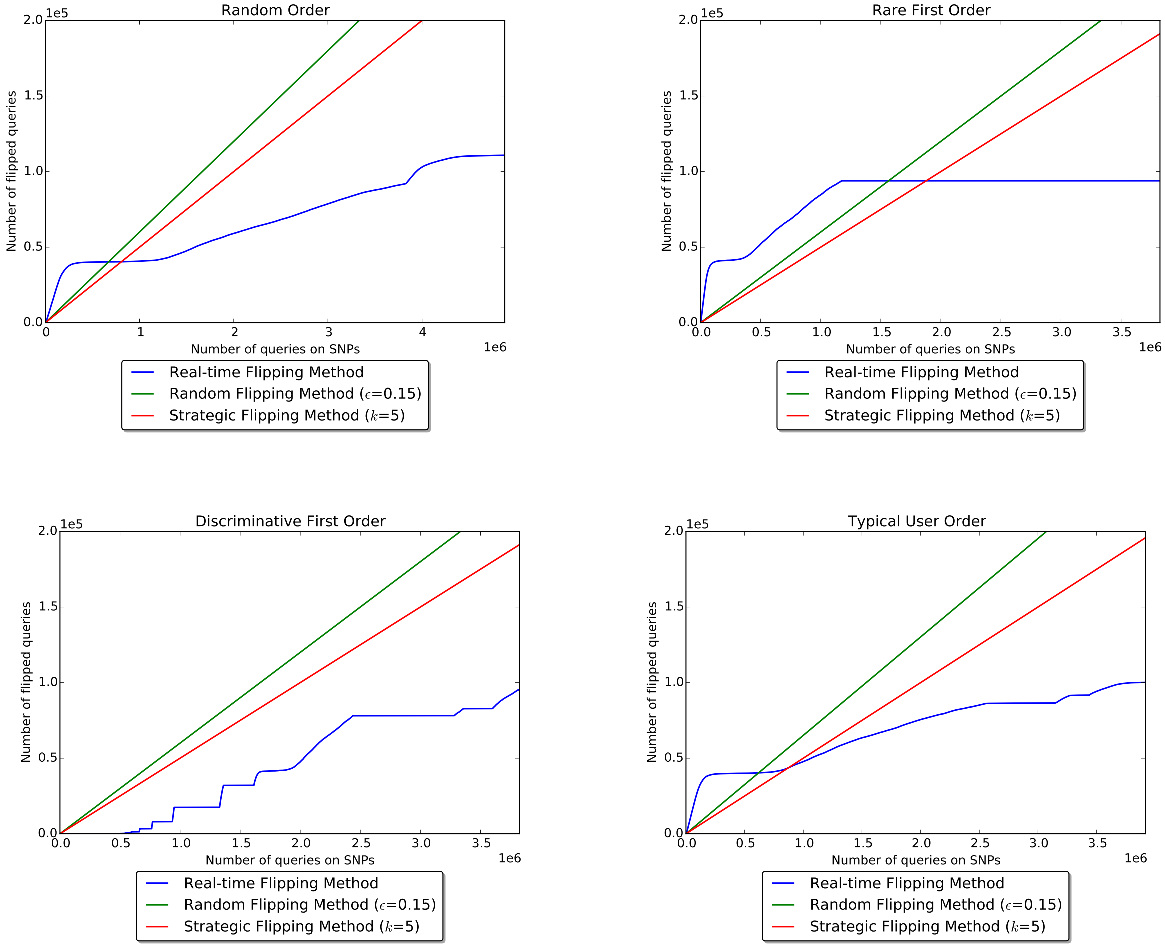


Figure 3: The percentage of flipped SNPs in total under different query patterns

9. Wilhelm J Ansorge. Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203, 2009.

10. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

11. Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, FL Yu, HM Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. 2003.

12. Michael T Goodrich. The mastermind attack on genomic data. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 204–218. IEEE, 2009.

13. Arif Harmanci and Mark Gerstein. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature methods*, 13(3):251, 2016.

14. Iman Deznabi, Mohammad Mobayen, Nazanin Jafari, Oznur Tastan, and Erman Ayday. An inference attack on genomic data using kinship, complex correlations, and phenotype information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017.

15. Dov Greenbaum, Andrea Sboner, Xinmeng Jasmine Mu, and Mark Gerstein. Genomics and privacy: implications of the new reality of closed data for the field. *PLoS Computational Biology*, 7(12):e1002278, 2011.

16. Bartha Maria Knoppers. Framework for responsible sharing of genomic and health-related data. *The HUGO journal*, 8(1):3, 2014.

17. Sharon F Terry, Robert Shelton, Greg Biggers, Dixie Baker, and Kelly Edwards. The haystack is made of needles. *Genetic testing and molecular biomarkers*, 17(3):175–177, 2013.

18. Jacob A Tennessen, Abigail W Bigham, Timothy D OConnor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *science*, 337(6090):64–69, 2012.

19. A Page, D Baker, M Bobrow, K Boycott, J Burn, S Chanock, et al. Genomics. a federated ecosystem for sharing genomic, clinical data. global alliance for genomics and health. *Science*, 352(6291):1278–1280, 2016.

20. Zhiyu Wan, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley Malin. Controlling the signal: Practical privacy protection of genomic data sharing through beacon services. *BMC medical genomics*, 10(2):39, 2017.

21. Md Momin Al Aziz, Reza Ghasemi, Md Waliullah, and Noman Mohammed. Aftermath of bustamante attack on genomic beacon service. *BMC medical genomics*, 10(2):43, 2017.

22. Monkol Lek, Konrad Karczewski, Eric Minikel, Kaitlin Samocha, Eric Banks, Timothy Fennell, Anne O'Donnell-Luria, James Ware, Andrew Hill, Beryl Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *BioRxiv*, page 030338, 2016.

23. Diyue Bu. *Secure-Beacon*, 2017. http://darwin.informatics.indiana.edu/diybu/secureBeacon/main.htm.

24. Shuang Wang, Xiaoqian Jiang, Haixu Tang, Xiaofeng Wang, Diyue Bu, Knox Carey, Stephanie OM Dyke, Dov Fox, Chao Jiang, Kristin Lauter, et al. A community effort to protect genomic data sharing, collaboration and outsourcing. *npj Genomic Medicine*, 2(1):33, 2017.