GENOME RECONSTRUCTION ATTACKS ON GENOMIC DATA-SHARING BEACONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Kerem Ayöz
July 2021

GENOME RECONSTRUCTION ATTACKS ON GENOMIC DATA-SHARING BEACONS

By Kerem Ayöz

July 2021

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Abdullah Ercüment Çiçek(Advisor)

Can Alkan

Tolga Can

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

GENOME RECONSTRUCTION ATTACKS ON GENOMIC DATA-SHARING BEACONS

Kerem Ayöz
M.S. in Computer Engineering
Advisor: Abdullah Ercüment Çiçek
July 2021

Sharing genome data in a privacy-preserving way stands as a major bottleneck in front of the scientific progress promised by the big data era in genomics. A community-driven protocol named *genomic data-sharing beacon protocol* has been widely adopted for sharing genomic data. The system aims to provide a secure, easy to implement, and standardized interface for data sharing by only allowing yes/no queries on the presence of specific alleles in the dataset. However, beacon protocol was recently shown to be vulnerable against membership inference attacks. In this thesis, we show that privacy threats against genomic data sharing beacons are not limited to membership inference. We identify and analyze a novel vulnerability of genomic data-sharing beacons: genome reconstruction. We show that it is possible to successfully reconstruct a substantial part of the genome of a victim when the attacker knows the victim has been added to the beacon in a recent update. In particular, we show how an attacker can use the inherent correlations in the genome and clustering techniques to run such an attack in an efficient and accurate way. We also show that even if multiple individuals are added to the beacon during the same update, it is possible to identify the victim's genome with high confidence using traits that are easily accessible by the attacker (e.g., eye color or hair type). Moreover, we show how a reconstructed genome using a beacon that is not associated with a sensitive phenotype can be used for membership inference attacks to beacons with sensitive phenotypes (e.g., HIV+). The outcome of this work will guide beacon operators on when and how to update the content of the beacon and help them (along with the beacon participants) make informed decisions.

Keywords: Privacy, Genome Reconstruction Attack, Genomic Data-Sharing Beacons, Genomics.

ÖZET

GENOMİK VERİ PAYLAŞAN BEACON SİSTEMLERİNE GENOM YENİDEN İNŞA SALDIRILARI

Kerem Ayöz
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Abdullah Ercüment Çiçek
July 2021

Genom verilerinin gizliliğini koruyarak paylaşılması, büyük verinin genomik alanına getirdiği bilimsel ilerlemenin önünde büyük bir engel olarak duruyor. Genomik verilerin paylaşımı için *genomik veri paylaşım beacon protokolü* adlı topluluk odaklı bir protokol yaygın olarak benimsenmiştir. Sistem, veri kümesindeki belirli alellerin varlığına ilişkin yalnızca evet/hayır sorgularına izin vererek, veri paylaşımı için güvenli, uygulaması kolay ve standartlaştırılmış bir arayüz sağlamayı amaçlamaktadır. Ancak yakın zamanda beacon protokolünün kimlik tespiti saldırılarına karşı savunmasız olduğu gösterildi. Bu tezde, genomik veri paylaşan beacon sistemlerine yönelik tehditlerin yalnızca kimlik tespiti saldırıları ile sınırlı olmadığını gösteriyoruz. Genomik veri paylaşan beacon sistemlerinin yeni bir güvenlik açığını tanımlıyor ve analiz ediyoruz: genom yeniden inşası. Saldırgan, kurbanın son güncelleme beacon sistemine eklendiğini bildiğinde, kurbanın genomunun önemli bir bölümünü başarılı bir şekilde yeniden inşa etmenin mümkün olduğunu gösteriyoruz. Özellikle bir saldırganın böyle bir saldırıyı verimli ve doğru bir şekilde yürütmek için genomdaki doğal korelasyonları ve kümeleme tekniklerini nasıl kullanabileceğini gösteriyoruz. Ayrıca aynı güncelleme sırasında beacon sistemine birden fazla kişi eklense bile, kurbanın saldırgan tarafından kolayca erişilebilinecek özelliklerini (ör. göz rengi veya saç tipi) kullanarak kurbanın genomunu yüksek doğrulukla tanımanın mümkün olduğunu gösteriyoruz. Ayrıca hassas bir fenotiple ilişkili olmayan bir beacon sistemi kullanılarak inşa edilmiş bir genomun, hassas fenotiplere sahip beacon sistemlerine (ör. HIV+) kimlik tespiti saldırıları için nasıl kullanılabileceğini gösteriyoruz. Bu çalışmanın sonucu, beacon sistemi operatörlerine (ve beacon katılımcılarına) beacon içeriğinin ne zaman ve nasıl güncelleneceği konusunda rehberlik edecek ve bilinçli kararlar vermelerine yardımcı olacaktır.

Anahtar sözcükler: Gizlilik, Genom Yeniden İnşa Saldırısı, Genomik Veri Paylaşan Beacon Sistemleri, Genomik.

Acknowledgement

I would like to say a special thank you to my thesis advisor Asst. Prof. A. Ercument Cicek for inspiring me to pursue my dreams, for guiding me to become a curious, fast-paced and honest researcher and for his precious support, encouragement and patience. His faith in me helped me to overcome my problems in research projects and most importantly in my life. I am proud of meeting with such an elite person and working under his supervision.

I also want to thank to Asst. Prof. Erman Ayday whose guidance, support and encouragement has been invaluable throughout this study. I enjoyed every minute of working with him in research projects that we have worked together. His enthusiasm and hard work have made me a better researcher.

Finally, I would like to express my gratitude and appreciation to my family and friends for their continuous support and trusting in my progress and skills. I would not be who I am and where I am today without each one of you. Thank you.

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM013429.

Contents

List of Figures

Chapter 1

Introduction

With plummeting sequencing costs, we look forward reaching a capacity of sequencing one billion individuals over the next 15-20 years, resulting in availability of very large genomic datasets [1, 2, 3]. Although such large datasets are promising a revolution in medicine, it has been shown in numerous studies that it is not straightforward to ensure anonymity of the participants in such datasets [4, 5, 6, 7, 8].

Human genome is the utmost personal identifier and sharing genomic data for research while preserving the privacy of the individuals have been challenging many different fields (e.g., medicine, bioinformatics, computer science, law, and ethics) for long, due to possibly dire ethical, monetary, and legal consequences. To address this challenge and create frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic data, the Global Alliance for Genomics and Health (GA4GH) was formed by the community [9]. The current genomic data sharing standard of the GA4GH is called the *genomic data-sharing beacons*. Beacons are the gateways that let users (researchers) and data owners exchange information without -in theory- disclosing any personal information. A user who wants to apply for access to a dataset can learn whether individuals with specific alleles (nucleotides) of interest are present in the beacon through an online interface. That is, a user can submit a query, asking whether a genome

exists in the beacon with a certain nucleotide at a certain position, and the beacon answers as "yes" or "no". If the dataset does not contain the desired genome, genomic data is not shared and distributed unnecessarily. In addition, researchers do not have to go through the paperwork to obtain a dataset which will not be helpful for their research. The GA4GH provides a shared beacon interface [10] that as of December 2020 provides access to 81 beacons and acts as a hub where researchers and data owners meet.

Beacons are typically associated with a particular sensitive phenotype (e.g., the SFARI beacon that host individuals with autism). Therefore, presence of an individual in a particular beacon is considered as privacy-sensitive information and the main aim of the beacons is to protect this information. An attacker, using the responses of a beacon and genomic data of a victim, may try to infer the membership of the victim in a particular beacon by running a membership inference attack. Beacon framework sets a barrier against membership inference attacks by allowing only presence/absence queries for variants and not tying any response to any specific individual. In that sense, beacons are considered to have stronger privacy measures compared to other statistical genomic databases. Despite these barriers, several works have proven that beacons are not bulletproof and they are vulnerable to membership inference attacks [11, 12, 13].

However, threats against genomic data-sharing beacons are not limited to membership inference attacks. In this thesis, for the first time, we identify and analyze the vulnerability of genomic data-sharing beacons for the "genome reconstruction" attack. We consider a scenario, in which the attacker knows the membership of a victim to a beacon that may not be associated with a sensitive phenotype. Therefore, we consider a targeted attack, in which either (i) the attacker knows that the victim donated their genome to take part in a study or (ii) infer the membership of the victim from beacon's metadata (as done in [11]). Then, we show how the attacker can accurately infer the genome of the victim by using the beacon responses. Such an attack may result in serious consequences if the attacker uses the reconstructed genome to infer sensitive information (e.g., disease diagnosis) about the victim or to infer the victim's membership to another statistical genomic database of interest (e.g., another beacon that is associated

with a sensitive phenotype). In particular, we show how the attacker can use the inherent correlations in the genome to run such an attack in an efficient and accurate way compared to a baseline approach. We also show how clustering techniques can be used to further improve the accuracy of such an attack.

Previous works in the literature assume beacons are static and do not change over time. However, beacons are dynamic datasets (donors join and leave) and this results in an increased risk for the genome reconstruction attack. An attacker can monitor the number of newly added donors to the beacon and the number of donors leaving the beacon from the meta-information of the beacon. With this information, newly joined donors (or donor leaving the beacon) become more vulnerable for genome reconstruction attacks. Thus, for the first time, we consider the beacons as dynamic databases and formulate the genome reconstruction attack accordingly. Privacy vulnerabilities due to dynamic changes in a system has been recently explored in the context of dynamic model changes in machine learning models [14]. It has been shown that different model outputs can constitute a new attack surface for an adversary to infer information of the dataset used to perform a model update [14]. Here, rather than model updates, we focus on the changes in the query responses to a dynamic database.

In a genome reconstruction attack, the attacker reconstructs all or a subset of the genomes in the beacon. Among the reconstructed genomes, it is not trivial to infer which one belongs to the victim. Therefore, we also show how the attacker can identify the victim's genome among the set of reconstructed genomes using moderate auxiliary information about the victim (i.e., a set of visible physical characteristics of the victim, which is public information). Finally, to show one of the consequences of the identified genome reconstruction attack, we show how the attacker can utilize the outcome of this attack to initiate a membership inference attack against the same victim in another beacon, which can be associated with a sensitive phenotype. To do this, we combine the identified genome reconstruction attack with the membership inference attacks against beacons from the literature.

We implement and evaluate the identified vulnerability using real genome data obtained from OpenSNP [15] and HapMap [16] datasets. We particularly evaluate the success of the attacker to reconstruct a victim's point mutations that include at least one rare nucleotide (i.e., minor allele) since minor alleles (i) reveal sensitive attributes of individuals (e.g., predispositions to privacy-sensitive diseases); and (ii) provide rich information to the attacker for membership inference attacks [12, 13]. We show that for a beacon with 50 individuals, precision and recall of the reconstruction reach up to 0.9 (each) when 3 individuals are added to the beacon and the victim is one of the newcomers. Even when 10 new participants are added to the beacon (causing a 20% increase in beacon size), we show that the attacker has a precision of 0.7 and a recall of 0.8. Furthermore, our results show that when more than one individual is added to the beacon, the attacker can accurately pinpoint the victim's reconstructed genome by using moderate (and publicly available) auxiliary information about the victim. For this, we show how the attacker can match the victim's phenotypical characteristics to the reconstructed genomes using machine learning algorithms. We also show via experiments that the outcome of the genome reconstruction attack can be accurately used for the membership inference attack on another beacon and it helps an attacker infer the membership of a victim only with a few queries.

Overall, we identify an important vulnerability and show how it can be exploited. We notably show how dependencies between point mutations can be used in a clustering algorithm to have high accuracy in a genome reconstruction attack. Furthermore, our methodology consists of a complete pipeline, showing how an attacker use the information it infers in the genome reconstruction attack in a subsequent membership inference attack. Therefore, this study clearly shows that privacy risks for genomic data-sharing beacons are much severe than perceived. This is particularly important since the number of beacon participants, and hence the privacy risk of individuals increase rapidly.

Chapter 2

Genomics Background

Approximately 99.9% of the all individuals' DNA are identical and the remaining 0.1% is responsible for our differences. Single nucleotide polymorphism (SNP) is the most common source of variation in the human genome. SNP is a point mutation (e.g., substitution of a single nucleotide in the genome - A,T,C, or G) and there are around 50 million known SNPs in the human genome [17]. The alternative nucleotides for each locus (SNP position) are called alleles and each allele of a SNP can be either the major or the minor allele for that SNP. The major allele is the most frequently observed nucleotide for a SNP position and the minor allele is the rare nucleotide (i.e., the second most common). The frequency (or probability) of observing the minor allele at a SNP position is called the minor allele frequency (MAF) of that SNP. Human genome has two copies for each locus (one per chromosome) and a SNP can be represented in terms of the number of its minor alleles (i.e., 0 for homozygous major, 1 for heterozygous, or 2 for homozygous minor).

Particular SNPs in human population are inherently correlated and this correlation model may change for different populations. Linkage disequilibrium (LD) is the non-random association of alleles at two or more loci. If two SNPs are in LD, they are correlated and co-occur more frequently than expected. Some SNPs are pathogenic and cause genetic diseases [18] and hence, they may carry sensitive

information regarding individuals' health conditions. As discussed in Chapter 3, most existing works in genomic privacy literature focus on the protection of the SNPs to prevent the risk of genetic discrimination.

Chapter 3

Related Work

Genomic privacy has recently been explored by many studies [19, 20, 21]. In the following subsections, we will summarize existing work on privacy in statistical genomic databases, inference attacks, and privacy of genomic data-sharing beacons.

3.1 Privacy in Statistical Genomic Databases and Inference Attacks on Genomic Privacy

Several works have shown that anonymization does not effectively protect the privacy of genomic data [22, 23, 24, 25, 26, 27, 28]. It has been shown that the identity of a participant of a genomic study can be revealed by using a second sample (e.g., part of the DNA information from the individual) and the results of the clinical study [29, 30, 31, 8, 32]. Differential privacy (DP) [33] concept has been frequently used to mitigate membership inference attacks when releasing summary statistics from genomic databases [34, 35, 36, 37]. Compared to statistical databases, genomic data-sharing beacons have stronger privacy measures since they only allow presence/absence (or yes/no) queries for variants.

Humbert *et al.* proposed an inference attack on kin genomic privacy using the family ties between individuals, pairwise correlations between the SNPs, and publicly available statistics about DNA [38]. Then, Deznabi *et al.* demonstrated that stronger inference techniques can be generated by combining high-order correlations and family ties [39]. Furthermore, several studies have examined phenotype prediction from genomic data, as a means of tracing identity [40, 41, 42, 43, 44, 45, 46, 47, 48, 49]. To mitigate such attribute inference attacks, cryptographic solutions has been proposed for privacy-preserving processing and sharing of genomic data (e.g., to outsource the computation to a public cloud or to conduct collaborative association studies). Existing cryptographic solutions mainly focus on (i) private pattern-matching and the comparison of genomic sequences [50, 51, 52, 53, 54] and (ii) privacy-preserving personalized medicine [55, 56]. In this work, we identify and analyze a different type of attribute inference attack particularly against genomic data-sharing beacons.

3.2 Privacy in Genomic Data Sharing Beacons

Researchers showed that presence (membership) of an individual in a genome sharing beacon can be inferred by repeatedly querying the beacon. Here, the attacker is assumed to be an active (or authorized) user of the beacon, in practice, it can ask as many queries as it wishes to the beacon (there is no limitations and cost for this in the current beacon protocol), and it can decide which queries to ask to the beacon. Furthermore, the attacker is assumed to have access to the set of SNPs of the victim. Shringarpure and Bustamante introduced a likelihood-ratio test (LRT) that can predict whether an individual is in the beacon by querying the beacon for multiple SNPs of a victim [11]. Note that inferring the membership of an individual in a beacon that is associated with a sensitive phenotype is equivalent to uncovering the sensitive phenotype about the victim. Then, Raisaro *et al.* showed that if the attacker first queries the SNPs with low minor allele frequency (MAF) values, it needs fewer queries for a successful attack [12]. In Section 6.5, we use this attack when we show how the proposed genome reconstruction attack can be combined with the membership inference attack. We

provide further background information about this attack in Appendix A. Later, von Thenen *et al.* showed that even if the attacker does not have victim's low-MAF SNPs, it is still possible to infer membership by exploiting the correlations in the genome [13]. Furthermore, they showed that beacon responses can also be inferred using such correlations (via a query inference, or QI-attack). In an orthogonal work, Hagestedt *et al.* have hypothesized that while current beacons systems are limited to genomic data, in the near future, the community is going to need a similar system for other biomedical data types. They proposed a beacon system for sharing DNA methylation data (an epigenetic mechanism to regulate transcriptional activity) and then showed that it is possible to successfully launch a membership inference attack against this system. They proposed a DP-based solution in their proposed *MBeacon* [57] system. The approach retains utility by adjusting the noise level for high risk methylation regions that might leak phenotypic information (i.e., regions which are related to disease).

Contribution of this thesis. In this thesis, we identify and analyze a genome reconstruction attack against genomic data-sharing beacons by particularly exploiting the information leaked due to beacon updates and the correlations between the point mutations. So far, all works in the literature have focused on membership inference attacks against genomic data-sharing beacons. To the best of our knowledge, this is the first work that identifies, thoroughly analyzes, and shows the consequences of the genome reconstruction attack against the beacons. Furthermore, as opposed to existing work (that only consider a snapshot of the beacon), we show the privacy risk in dynamic beacons, in which new donors may join or existing donors may leave.

Chapter 4

System Model

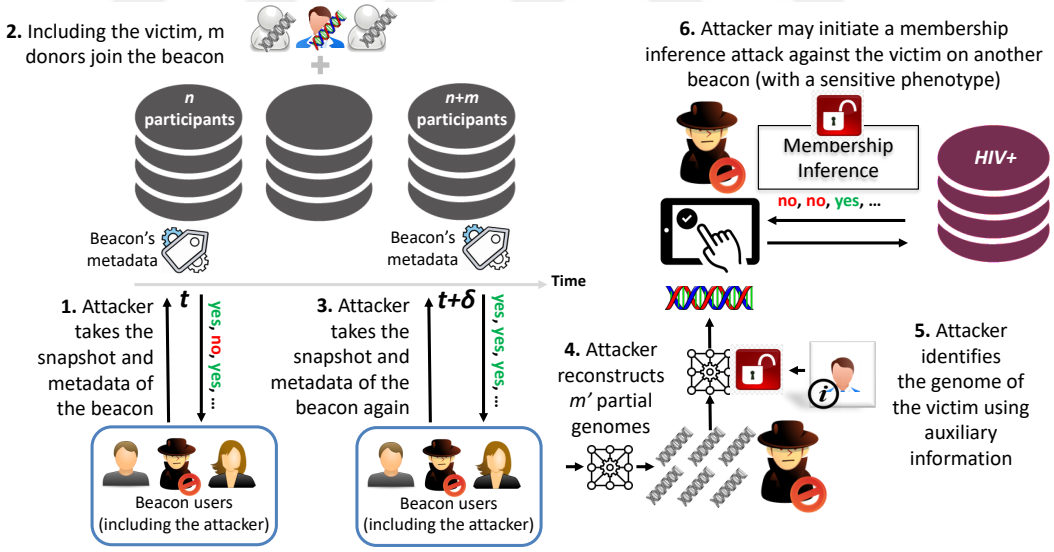


Figure 4.1: Proposed system model.

As shown in Figure 4.1, we consider a system between the beacon participants (e.g., donors), the beacon, and the beacon users (which may include the attacker). The donor shares their genome with the beacon. It is possible that the donor may share their genome with multiple beacons that may or may not be associated with sensitive traits. Genome donor is not active during the protocol after they share their data with the beacon. Also, beacon never publicly shares its dataset, but some beacons may share metadata about (i) their content (e.g., size) or (ii)

their donors (e.g., their gender, age, or ethnicity). In general, we consider the beacon as a dynamic dataset, in which new donors may join and existing donors may leave over time. Beacon users issue queries to the beacon. As discussed, the beacon user can only ask the presence of a genome with a particular allele (nucleotide) at a particular position of a given chromosome and the beacon only responds as "yes" or "no". In this work, we assume beacon honestly reports the result of each query to the user (e.g., without introducing intentional noise to the query results) and we do not consider a query limit for the users, as it is usually trivial to overcome such limits (e.g., by registering several times with different accounts).

Chapter 5

Threat Model

Depending on the attacker's objective, two attacks that can be launched against genomic data-sharing beacons are: (i) membership inference attack and (ii) genome reconstruction attack. In both attacks (including this work), the attacker is assumed to be a registered beacon user who can send unlimited number of queries to the beacon. In this work, for the first time, we identify and study the genome reconstruction attack. We assume that the attacker knows the membership of an individual to a beacon. Thus, we consider a targeted attack, in which the attacker knows that the victim donated their genome (to take part in a study). Given the current rise in personal genomics (people uploading their genomes to public sites), this is feasible. Also, beacons with no sensitive-phenotype report metadata about their donors. For instance, Shringapure and Bustamante [11] verified a specific person being in PGP and Kaviar [58] beacons via metadata, and hence the attacker can also identify the membership of the victim using such metadata. Using the membership information, the goal of the attacker is to reconstruct the victim's genome by issuing queries to the corresponding beacon. Genome inference attack can be considered both for static and dynamic beacons. In static beacons, knowing that the victim is a member of the beacon, only the "no" responses would provide certain information about the victim's genome to the attacker. "Yes" responses may be due to any other participant of the beacon and as the size of the beacon increases, "yes" responses do not provide much

information to the attacker. However, in dynamic beacons, when the beacon is updated, using the change in the responses of the beacon, the attacker can learn more about the genomes of new participants. Thus, in this thesis, we analyze this vulnerability for dynamic beacons and we assume that the victim is added between times t and $t+\delta$ along with other $(m-1)$ newly added donors to the beacon. As discussed before, the attacker can monitor the number of newly added donors to the beacon and the number of donors leaving the beacon from the metadata of the beacon. We assume that, along with the fact that the victim is among the newly joined participants to the beacon, the attacker also knows (i) the number of other newly joined individuals that are added to the beacon along with the victim; (ii) a snapshot of the beacon before the victim is added (at time t). That is, responses to all queries before the victim joins to the beacon. The beacon protocol does not bar someone from taking a complete snapshot. Thus, querying a beacon to take a complete snapshot only requires a high-bandwidth internet connection. Economic cost of such an internet service is around 79$ per month [59] and there is no other economic cost, as the system is publicly available at [10]. Even though the number of SNPs in a complete snapshot is large, typically, only low-MAF SNPs are useful for the attacker (as they are typically the sensitive ones); (iii) auxiliary information about the victim to identify victim's genome among the reconstructed ones. For this we assume the attacker has moderate information, such as a set of victim's visible characteristics (phenotype); and (iv) publicly available information about genomics, such as minor allele frequencies (MAF values) of SNPs and correlation between the SNPs in the population of interest. Finally, we assume that the attacker does not collude with the beacon.

In genome reconstruction attack, due to the nature of beacon responses, the attacker can infer if a victim has at least one minor allele at every SNP position. This is because the response of the beacon only tells if there is an individual in the beacon with at least one minor allele at a given SNP position. Thus, for each SNP j of victim v (S_j^v), the goal of the attacker is to infer $Pr(S_j^v = 0)$ and $Pr(S_j^v \neq 0)$ (i.e., $Pr(S_j^v = 1)$ or $Pr(S_j^v = 2)$). For simplicity, we define the event $\hat{S}_j^v = \mathbf{1}_{S_j^v=1 \vee S_j^v=2}$. Thus, $\hat{S}_j^v = 0$ if $S_j^v = 0$, and $\hat{S}_j^v = 1$, otherwise. Note that inferring this information for a victim results in a serious privacy concern. As we will discuss and show later, using this information, an attacker can associate

the genotype of the victim to related phenotypes (e.g., diseases) and initiate a membership inference attack for the victim by targeting another beacon that is associated with a sensitive phenotype (e.g., cancer or HIV+).

Our methodology consists of a complete pipeline, showing how an attacker uses the information it infers in the genome reconstruction attack in a subsequent membership inference attack. Therefore, we evaluate the success of the attacker using different metrics in different parts of the pipeline as follows. For genome reconstruction (in Chapter 6.3), we use precision and recall to quantify this inference power of the attacker. As we will show in Chapter 7, the success of genome reconstruction mainly depends on the size of the beacon, the number of newly added donors to the beacon between times t and $t + \delta$, and the fraction of attacker's snapshot at time t. In real life, sizes of beacons show a large variation. The size of a beacon can be as small as 100, such as NBDC Human Database [60] or as large as 100K, such as The Genome Aggregation Database (gnomAD) [61]. As discussed, these numbers can be monitored from the metadata of such beacons. Thus, as we will we show, for small-size beacons, even if the size of the beacon is significantly increased (compared to its original size), the attacker's success may be high. For large-size beacons, on the other hand, the number of newly added donors should be a small fraction of the original size for a successful attack. As a result of the genome reconstruction, the attacker potentially reconstructs multiple genomes and among these, one belongs to the victim. For this part, we show how the attacker can utilize machine learning techniques to identify the victim's genome among the reconstructed ones (in Chapter 6.4) and we use the classification accuracy of the attacker as its success metric. Finally, to quantify the success of the membership inference (Chapter 6.5), we use a power analysis as the success metric. To evaluate the success of the attacker in the membership inference attack, we first let the attacker run the genome reconstruction attack and then use the proposed machine learning technique to identify the victim's genome among the reconstructed ones. Thus, the success metric for the membership inference considers the attacker's success in the entire pipeline.

Chapter 6

Genome Reconstruction Attack on Genomic Data-Sharing Beacons

As discussed, we define the genome reconstruction attack as inferring genomic data of a genome donor (i.e., victim) given their membership information to the beacon. To show the effect of genome reconstruction attack more clearly, we consider dynamic beacons and we assume the victim is among the newly joined donors to the beacon. For clarity of the discussion, we present the identified attack only considering newly joined donors. Considering the donors that leave the beacon is symmetrical and trivial. We discuss this case in Chapter 8.2. We consider a scenario, in which the attacker has no information about the victim's genome, but it knows that the victim is added to the beacon between times t and $t + \delta$. Let n and $(n + m)$ represent the number of individuals in the beacon at times t and $t + \delta$, respectively. As discussed, for most real-life beacons, the attacker knows m (by monitoring the changes in beacon using the metadata of the beacon). In all attack scenarios, we assume that the attacker reconstructs m' genomes (m' can be different than m and the selection of m' effects the precision and recall of the attacker). Our goal is to evaluate the performance for different m' values to show the attack is robust even if the attacker does not know how many

people are added. When metadata of the beacon, and hence m is not available, the attacker can determine a potential upper bound (k) for the number of newly added donors (m) by examining the number of flipped responses (from "no" to "yes"). Then, for each i from 1 to k, it can reconstruct genomes using $R_{N \rightarrow Y}$ assuming $m = i$, and hence instead of m, the attacker ends up having $\frac{k(k+1)}{2}$ potential genomes to identify the victim's best matching reconstructed genome.

Using its auxiliary information (as discussed in Chapter **??**), the attacker can probabilistically infer the genome of the victim by utilizing the changes in beacon's responses (at times t and $t + \delta$) as follows: (i) if the previous response (at time t) was "no" and the current response (at time $t + \delta$) is "yes", the probability that the victim having a minor allele at the corresponding query position increases depending on how many new individuals are added to the beacon in this time interval; (ii) if the previous response was "yes" and the current response is also "yes", attacker cannot infer much about the victim's genome, especially if the total size of the beacon is large; and (iii) if both the previous and the current responses are "no", the attacker understands that the victim does not have a minor allele at the corresponding query position.

Here, the most important (or the most sensitive) information for the attacker can be considered as the "no" responses at time t that turn to "yes" at time $t+\delta$. Because, such responses let the attacker infer the positions that the victim has at least one minor allele with a high probability (depending on how many new individuals are added to the beacon in this time interval). Since minor alleles of individuals are typically the indicators for privacy-sensitive information about them, in this work, we focus on the success of the attacker based on its success in inferring the minor alleles of a victim using the beacon responses that turn to "yes". Exhaustively generating all potential solutions of this problem would result in a total of $2^{\beta * m'}$ genomes, where β is the total number of responses that turn to "yes" at time $t + \delta$ (which can be on the order of tens of thousands), and hence it is intractable. In the following, we first describe a baseline method that provides a tractable solution to this problem. Next, we present a greedy approach to run such an attack more accurately, and then we will detail a more sophisticated, clustering-based approach for the genome reconstruction attack.

6.1 Baseline Approach for Genome Reconstruction

Here, we describe a baseline approach, in which the attacker, using the responses of the beacon, reconstructs the genomes (of the newly joined donors) by assigning them to m' bins according to MAF values of the SNPs. Genome reconstruction attack using the baseline algorithm for a particular victim v at time $t + \delta$ can be described as follows. The input of the attacker is (i) responses of the beacon to all possible queries at time t (i.e., complete snapshot of the beacon at time t); (ii) the fact that m new donors are added to the beacon between times t and $t + \delta$; (iii) the fact that the victim is among the newly added donors; and (iv) publicly available MAF values of the SNPs.

First, the attacker identifies the set of SNPs for which the response of the beacon was "no" at time t and it becomes "yes" at time $t+\delta$. Thus, the attacker constructs a set $R_{N\rightarrow Y}$, consisting of these SNPs. Then, the attacker creates m' empty bins representing SNP sets of newcomer donors. For each SNP j in set $R_{N\rightarrow Y}$, the attacker retrieves its MAF value, MAF_j. Next, the attacker assigns the value of SNP j for each individual i (in each bin) consistent with the SNP's MAF value as follows: (i) $\hat{S}^i_j = 0$ with probability $(1 - MAF_j)^2$ and (ii) $\hat{S}^v_j = 1$ with probability $MAF_j^2 + 2MAF_j(1 - MAF_j)$. Since the beacon's response for SNPs in $R_{N\rightarrow Y}$ has flipped from "no" to "yes", for all SNPs in $R_{N\rightarrow Y}$, there should be at least one bin (among m' bins) with at least one mutation (i.e., homozygous minor or heterozygous SNP). Thus, once the values of the SNPs in $R_{N\rightarrow Y}$ for all m' bins are determined, the attacker checks if there is any SNP in set $R_{N\rightarrow Y}$ that is not assigned to any bin. If there is such a SNP, the attacker randomly picks a bin and assigns the value of the corresponding SNP as $\hat{S}^i_j = 1$ for the corresponding bin. The details of this baseline approach are also shown in Algorithm 2 (in Appendix B).

6.2 Greedy Algorithm for Genome Reconstruction

The above-mentioned baseline algorithm assumes every SNP is independent and the correlations among them are disregarded. However, SNPs are inherently correlated and considering such correlations in the genome reconstruction attack may result in significantly more accurate results. In the greedy algorithm discussed here, the attacker forms the bins considering the correlations between the SNPs in set $R_{N \to Y}$. Using an iterative approach, the attacker assigns each SNP (minor allele) to an individual such that the probability of assignment is proportional to the average correlation of the new SNP with the already assigned SNPs of the individual (i.e., bin i). If no assignment is made this way, a random individual is selected to make sure there is at least one person with the corresponding new SNP.

Genome reconstruction attack using the greedy algorithm for a particular victim v at time $t+\delta$ can be described as follows. The input of the attacker includes everything in the baseline approach and also a correlation model between the SNPs that is consistent with the population structure of the beacon (that can be computed using publicly available genomic datasets). Different correlation models have been explored for genomic data before. In [62], authors showed how the correlations in the genome can be modelled using a Markov chain model. We create our correlation model by considering the pairwise correlations between all the SNPs in the beacon (which results in richer information for the attacker). The attacker calculates the likelihood of the victim v having at least one minor allele at a SNP position j as $P_k(\hat{S}_j^v) = P(\hat{S}_j^v | \hat{S}_k^v)$, where k may be any other position in the genome. We use Sokal-Michener distance to compute correlations between SNPs as follows:

$$A = 2(n_{\hat{S}_j^v=1, \hat{S}_k^v=0} + n_{\hat{S}_j^v=0, \hat{S}_k^v=1})$$

$$B = n_{\hat{S}_j^v=1, \hat{S}_k^v=1} + n_{\hat{S}_j^v=0, \hat{S}_k^v=0}$$

$$D_{Sokal-Michener}(\hat{S}_j^v, \hat{S}_k^v) = \frac{A}{A+B}$$

In the greedy approach, first, the attacker constructs set $R_{N \to Y}$. Then, it creates m' empty bins (m' does not have to be equal to m) representing the number of rare SNPs in $R_{N \to Y}$. We assume that the SNPs with an MAF value below a threshold τ are categorized as rare SNPs. Observing rare SNPs do not have correlations among each other, assigning the rare SNPs in $R_{N \to Y}$ to different bins as seeds is assumed to result in an accurate initial separation of individuals. Next, for each remaining SNP j in $R_{N \to Y}$, the attacker computes the average correlation between that and all the previously assigned SNPs in bin i using the aforementioned correlation model. This is done for each bin i. Let $\hat{S}_j^i$ be a binary random variable for SNP j and bin i. The attacker assigns $\hat{S}_j^c = 1$ for bin c which has the highest average correlation value and $\hat{S}_j^i = 0, \forall i \in [1, m']$ and $i \neq c$. Eventually, the attacker constructs m' potential genomes (in m' bins) belonging to m newcomer donors.

6.3 Clustering-Based Algorithm for Genome Reconstruction

Greedy algorithm (in Chapter 6.2) reconstructs genomes by following a particular order (determined based on the MAFs of the SNPs). Different orders may provide different solutions. Thus, to consider all query responses together in a collective way, we propose clustering-based approaches for the genome reconstruction attack that cluster the identified minor alleles for the newly joined donors to the beacon. The proposed clustering techniques essentially use the correlations between the SNPs (that are computed using the aforementioned correlation model) to distribute SNPs into different bins. We use two types of clustering techniques: (i) hard clustering to create non-overlapping bins and (ii) soft or fuzzy clusterin to assign a SNP into multiple bins.

For (i), we employ spectral clustering, in which a standard clustering method (such as k-means clustering) is applied on certain eigenvectors of the Laplacian matrix of a graph [63]. In this graph, the SNPs correspond to vertices and

correlations between the SNPs correspond to weights of edges. Spectral clustering is our method of choice as it has been shown to provide favorable results in many high dimensional feature spaces like ours [64]. And, for (ii) we employ the fuzzy c-means clustering (FCM) algorithm [65], which is a common choice for these types of tasks. The algorithm is similar to k-means clustering, but it also allows probabilistic assignments of samples to multiple clusters. Different from k-means clustering, FCM assigns a membership value $u_{ij} = P(\hat{S}_j^i = 1)$ for each element j and for each cluster i. This membership values are used as weights in the objective function. After convergence, these membership values are used as the probability of assignments of elements to each cluster. The description of both clustering methods are similar except for the clustering steps. Thus, in the following, we describe both methods together.

The input of both clustering-based algorithms is the same as the input of the greedy algorithm. First, the attacker identifies the set of SNP positions for which the response of the beacon was "no" at time t and it becomes "yes" at time $t + \delta$ and constructs set $R_{N \to Y}$. Then, the attacker builds a graph of SNPs using the correlation model, in which the vertices are the SNPs in $R_{N \to Y}$ and undirected edges are weighted by the correlation values between these SNPs. This graph represents a pairwise similarity model for the SNPs and is used for a quantitative assessment of the correlation of each SNP pair in $R_{N \to Y}$.

Next, the attacker applies either the spectral or fuzzy clustering algorithms on the constructed graph. The outcome of spectral clustering is a set of disjoint clusters. Fuzzy clustering results in groups of SNPs that maximizes the similarity in a group while allowing a SNP to be shared by multiple individuals. Thus, in fuzzy clustering, each SNP i is assigned to clusters for which the algorithm returns a relatively high probability of association. After clustering, the attacker obtains m' different clusters which corresponds to m' reconstructed genomes. The details are shown in Algorithm 1.

6.4 Identifying the Victim Using Genotype-Phenotype Associations

In previous sections, for genome reconstruction, we assumed that the attacker can correctly identify the victim's genome among several reconstructed bins. Assuming the attacker has some moderate auxiliary information about the victim, here, we study and show how accurately the attacker can identify the victim's genome among other candidates. For this, we assume the attacker uses information about some phenotypic characteristics of the victim and it relies upon the fact that SNPs are intrinsically linked to phenotypic traits (such as eye color, hair color, etc.) This provides a complete methodology for the genome reconstruction attack against beacons in real-life. As we will discuss later, the success of the attacker to correctly identify the victim's genome among the reconstructed ones increases if the attacker has access to more auxiliary information about the victim.

Assume victim v is among the m new additions to the beacon (it is trivial to extend the methodology if there are more than one victim). The attacker is assumed to have access to two distinct sets: (i) a set $\mathcal{S} = \{\vec{S}_1, \vec{S}_2, \ldots, \vec{S}_{m'}\}$ of m' reconstructed genotypes as a result of the genome reconstruction attack, where $\vec{S}_i = (\hat{S}_1^i, \ldots, \hat{S}_k^i)$ is a vector containing the SNP values of genotype i (or bin i); and (ii) a set $\mathcal{P}_v = (p_1^v, \ldots, p_t^v)$ containing the values of t phenotypic traits of victim v. Such phenotype information can be obtained from publicly available resources or using the physical traits of the victim. For instance, the attacker can obtain such information from victim's social media accounts. The goal of the attacker is to correctly match the victim's phenotype to the correct reconstructed genome (that is the most similar to the victim's) among all candidate reconstructed genome sequences. In the test phase, the attacker has m newly added donors and m' reconstructed genomes. Attacker's task is to match each donor with the best matching reconstructed genome. Thus, for each newly added donor, the attacker calculates the likelihood scores of matching with all m' reconstructed genomes.

In [66], Humbert *et al.* focused on the deanoymization risk and modelled genotype-phenotype association as an assignment problem. They showed this risk by using the Hungarian algorithm [67]. Different from [66], here, we rely on machine learning for maximizing the matching likelihood and genotype-phenotype associations. We observe that such a formulation provides more accurate results. Also, rather than using SNP values (0, 1 or 2), due to the nature of the proposed attack, we represent the state of each SNP j of individual i as $\hat{S}_j^i$, which can be either 0 or 1, as discussed before.

For phenotype inference, we train a separate model for each of the considered phenotypes, where SNPs with flipped responses (from "no" to "yes") are used as features. Since phenotype datasets are highly imbalanced, we apply Synthetic Minority Oversampling Technique (SMOTE) [68] for each of these datasets to resolve this problem. In SMOTE, a minority class instance is selected along with its nearest neighbors at random. Then, a new sample is generated as a combination of the original instance and a random neighbor. Next, we train a random forest model for each phenotype. We use repeated stratified 5-fold cross validation to tune the hyperparameters. After training the phenotype models, we form the ensemble classifier using the ones that have better validation F1-macro score than random guess. We discard the other models.

Ensemble classifier calculates the matching likelihood of given genome and set of phenotypic traits. Softmax output of each phenotype model corresponding to a given phenotypic trait of the victim (i.e., probability that a reconstructed genome having blue eye) are summed to calculate the matching likelihood. For single victim, this calculation is done for each reconstructed genome and the victim is matched with the reconstructed genome with the highest matching likelihood score. Note that this matching does not need to be one-to-one; a single reconstructed genome might match with different set of phenotypic traits. We discuss the performance of identification of victim's reconstructed genome under different settings in Chapter 7.4.

6.5 Using Genome Reconstruction in Membership Inference Attack

To show one consequence of the proposed genome reconstruction attack, we also model and analyze how the proposed attack can be utilized for membership inference attack (introduced in Appendix A). We consider a scenario in which the attacker knows the membership of an individual to a beacon with which no sensitive associated phenotype (e.g., phenotype neutral). The attacker first utilizes the responses of this beacon to infer specific parts of a victim's genome (i.e., SNPs). Then, it uses these inferred SNPs to infer the membership of the victim to a beacon with a sensitive phenotype. This attack is important and realistic, because knowing the membership of an individual to a phenotype neutral beacon (e.g., Kaviar Beacon) may not seem to pose a privacy issue. However, using the proposed genome reconstruction attack and the membership information of the victim to the beacon with non-sensitive phenotype, the attacker can first infer the SNPs of the victim and then, infer the membership of the victim to another beacon which is associated a sensitive phenotype (e.g., SFARI beacon which is associated with autism phenotype).

To show this, first, we run the proposed genome reconstruction attack that is explained in Chapter 6.3 and infer the SNPs of the victim with at least one minor allele on a beacon B_1. Using these inferred SNPs, we then run the membership inference attack to infer the membership of the victim in another beacon B_2. For membership inference attack, we use the optimal attack in [12] (described in Appendix A), which is shown to be an effective attack for membership inference (for our scenario, optimal attack in [12] and the QI-attack in [13] perform similarly, so we choose to use the optimal attack due to its simplicity). However, in contrast to the original optimal attack, in the null and alternate hypothesis equations in (A.1) and (A.2), there is an additional error due to the inference error of the genome reconstruction attack. This is because the attacker queries the alleles of the victim that it infers as a result of the genome reconstruction attack and there is a degree of uncertainty. Thus, we first experimentally compute the error rate

of the genome reconstruction attack for a particular scenario (e.g., for particular m and n values). We then include this additional error on the γ parameter in (A.2), which represents the probability that the attacker's copy of the victim's genome does not match the beacon's copy for a SNP. Furthermore, as opposed to original optimal attack, here the attacker may not have access to the SNPs of the victim with the lowest MAF values; instead the attacker only knows the SNPs that are inferred as a result of the genome reconstruction attack.

We evaluate the success of this attack in terms of the power of the attacker in Chapter 7.5. Similar to [12] and [13], we plot the power curve of the membership inference attack at 5% false positive rate. We empirically build the null hypothesis (H_0 in Appendix A). For every query, we determine the distribution of Λ under the null hypothesis using 20 individuals that are not in B_2. In this work, in order to model the uncertainty of correctly matching the victim (using phenotype inference as in Chapter 6.4), we first experimentally compute the error rate of the overall process. For instance, if the accuracy of correctly matching the phenotype of the victim to their reconstructed genome is p%, then p% of the 20 individuals are selected from correctly identified reconstructions and remaining individuals are selected from other new people added to the beacon along with the victim (incorrect identifications).

When Λ is less than a threshold t_α, the null hypothesis is rejected and we find t_α from the null hypothesis with $\alpha = 0.05$ (corresponding to 5% false positive rate). Then, we computed the power as proportion of the individuals in the alternate hypothesis (including 20 different individuals in B_2) having a Λ value that is less than t_α. As before, p% of the 20 individuals are selected from correctly identified reconstructions and remaining people are selected from other new people added to the beacon along with the victim.

Algorithm 1: Clustering-Based Algorithm for Genome Reconstruction Attack

Input: b: beacon; m: Number of added people to b; Population P that represents the composition in b

Output: m' reconstructed genomes

```
// Step 1: Query Beacon
snapshot1 ← queryBeacon(b, t)
// Including victim, m donors join Beacon between time t and t + δ
snapshot2 ← queryBeacon(b, t + δ)

// Step 2: Obtain No-Yes SNPs
NoYesResponses ← []
for i ← 0 to snapshot1.length do
    if snapshot1[i] = "No" and snapshot2[i] = "Yes" then
        NoYesResponses.append(i)
    end
end

// Step 3: Cluster No-Yes SNPs
G ← Graph()
for i ← 0 to NoYesResponses.length − 1 do
    for j ← i + 1 to NoYesResponses.length do
        ri ← NoYesResponses[i]
        rj ← NoYesResponses[j]
        c ← corr(P,ri,rj)
        G.addEdge(ri,rj,c)
    end
end
clusters ← graphClustering(G, m')

// Step 4: Reconstruct genomes
S ← []
for i ← 0 to m' do
    S[i] ← getReferenceGenome(P)
    foreach s in clusters[i] do
        S[i][s] ← getMinorAllele(P, s)
    end
end
return S
```

Chapter 7

Evaluation

To evaluate the identified vulnerabilities, we evaluated our methods using real-life genomic datasets. Here, we describe the datasets and present the evaluation results.

7.1 Datasets

We used two different genome datasets for evaluation: (i) genome dataset of CEU population from the HapMap dataset [69] and (ii) OpenSNP genome dataset [70]. Using the HapMap dataset, we created the beacons and victims from CEU population which contains 164 donors and around 4 million SNPs for each donor. We created the correlation model (i.e., SNP-SNP relation network or similarity model) for this beacon using individuals from the same HapMap dataset that are not in the constructed beacon and set of victims. Using the OpenSNP dataset, we created the beacons and victims from a random population which contains 2980 donors and around 2 million SNPs for each donor. We created the correlation model using the rest of the OpenSNP dataset.

For the OpenSNP dataset, we also collected the reported phenotypes of individuals. Since sample sizes are small, we used the reported phenotypes in a

binary form. From OpenSNP, we used the following commonly reported phenotypes: (i) eye color, 967 samples, (ii) hair type, 371 samples, (iii) hair color, 468 samples, (iv) tan ability, 287 samples, (v) asthma, 226 samples, (vi) lactose intolerance, 347 samples, (vii) earwax, 244 samples, (viii) tongue rolling, 434 samples, (ix) intolerance to soy, 136 samples, (x) freckling, 277 samples, (xi) ring finger being longer than index finger, 268 samples, (xii) widow peak, 176 samples, (xiii) ADHD, 154 samples, (xiv) acrophobia, 155 samples, (xv) finger hair, 155 samples, (xvi) myopia, 152 samples, (xvii) irritable bowel syndrome, 142 samples, (xviii) index finger being longer than big thumb, 131 samples, (xix) photoptarmis, 133 samples, (xx) migraine, 129 samples, and (xxi) Rh protein, 311 samples. We used 1320 genomes which are associated with at least one of the listed phenotypes while training the models. Newly added donors are chosen from the individuals who have reported at least 10 out of 21 considered phenotypes. We repeated each experiment for 10 times with different sets of newly added donors. For each experiment, remaining samples (except for the beacon participants and newly added donors) are used to train and validate phenotype models.

7.2 Evaluation Metrics

We evaluated the precision and recall for the reconstruction of a victim's SNPs based on the changes in beacon responses. For precision and recall, we defined the success as correctly inferring the SNPs of the victim with at least one minor allele. Thus, for the calculation of precision and recall, we defined (i) *true positive* as correctly inferring a SNP j of victim v with $\hat{S}_j^v = 1$ (with at least one minor allele); (ii) *false positive* as incorrectly assigning $\hat{S}_j^v = 1$ for v who is homozygous major at that locus; (iii) *true negative* as correctly inferring a SNP j of victim v with $\hat{S}_j^v = 0$ (with no minor allele, homozygous major); and (iv) *false negative* as incorrectly assigning $\hat{S}_j^v = 0$ for v who has at least one minor allele at that locus (i.e., heterozygous or homozygous minor).

Furthermore, we quantified the success of identifying the victim's genome among the reconstructed genomes in terms of the accuracy of the developed

genotype-phenotype inference mechanism. We evaluated the accuracy of the ensemble classifier (to identify victim's genome from phenotype) using the reconstructed genomes of newly added donors. Given ensemble classifier f, set of indices $\vec{J} = (j_1, ..., j_v)$ that represent the indices of best matching clusters for each newly added donors, vector containing SNP values of the i^{th} cluster $\vec{S}_i = (\hat{S}_1^i, \ldots, \hat{S}_k^i)$, and set of phenotypic traits of victim v, $\mathcal{P}_v = (p_1^v, \ldots, p_t^v)$, we computed the accuracy as $(\sum_{v=1}^{m} \mathbf{1}_{(\underset{1 \le i \le m'}{\operatorname{argmax}} f(\vec{S}_i, \mathcal{P}_v)) = j_v})/m$. Finally, we used power analysis for the membership inference to show how the outcome of the genome reconstruction attack can be used for membership inference attack. Power for the i^{th} query is calculated from given set of l case people as $P^i = (\sum_{\Lambda_j^i \in \Lambda_{case}^i} \mathbf{1}_{\Lambda_j^i < t_\alpha^i})/l$, which is defined as the fraction of the cases who have Λ_j^i value that is less than t_α^i as described in Chapter 6.5. Then, the vector $\vec{P}^n = (P^1, ..., P^n)$ is plotted to see the power change with respect to a total of n queries. Higher power value represents a more successful attack.

7.3 Evaluation of Genome Reconstruction

First, using both OpenSNP and HapMap beacons and only focusing on genome reconstruction, we evaluated and compared the baseline method (in Chapter 6.1) and the proposed clustering-based approach (in Chapter 6.3) when the size of the beacon (n) is 50 and $m = m'$. Here, we assume that the attacker can identify the victim's reconstructed genome among the other candidates. Later, we will also show that attacker can indeed identify this genome with high accuracy using public (i.e., not sensitive) phenotype information about the victim.

Figures 7.1 and C.1 (in Appendix C) show the precision and recall of the reconstruction for various number of newly added donors (m) for OpenSNP and HapMap beacons, respectively. Overall, we observed that the success of the attack to be higher for OpenSNP beacon. The reason of this is the limited data we had to build the correlation model for HapMap dataset (we used 945 donors to build the correlation model in OpenSNP beacon, while we could only use 110 donors

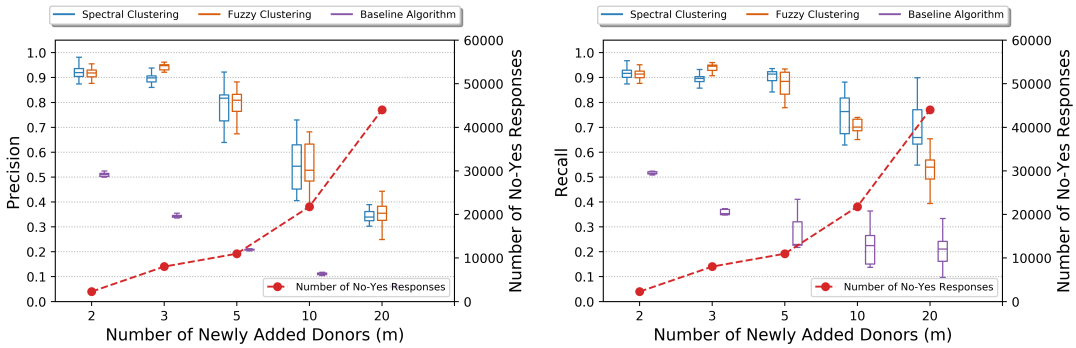


Figure 7.1: Precision and recall for the genome reconstruction of a newly added donor to OpenSNP beacon with varying number of newly added donors.

to build it for the HapMap beacon). For both datasets, we used individuals that are not in the beacon to construct the correlation models. When we compared the correlation models that are constructed using the individuals that (i) are not in the beacon and (ii) are in the beacon, we observed that correlation model constructed for the OpenSNP beacon is significantly more accurate (i.e., it is very close to the correlation model of the individuals that are in the OpenSNP beacon) mainly due to the number of individuals we used to create the model. Therefore, in the following, we mostly discuss the results we obtained from the OpenSNP beacon.

The results show that on average, the identified attack using spectral clustering can reconstruct the victim's genome with a precision close to 0.9 when the size of the beacon is increased by adding 3 people (i.e., a 6% increase in beacon size). We also obtained more than 0.7 precision and 0.8 recall even when the size of the beacon is increased by adding 10 people (i.e., a 20% increase in beacon size). This indicates a substantial privacy risk, especially if the reconstructed SNPs are tied to sensitive phenotypes. Also, the baseline algorithm (in Chapter 6.1) performs substantially worse than the proposed clustering-based approach. The results also show that spectral clustering-based genome reconstruction is slightly better than the fuzzy clustering-based approach. We observed that allowing a SNP (that includes at least one minor allele) to be in multiple bins results in high false positives. Therefore, in the remaining of this chapter, we use spectral

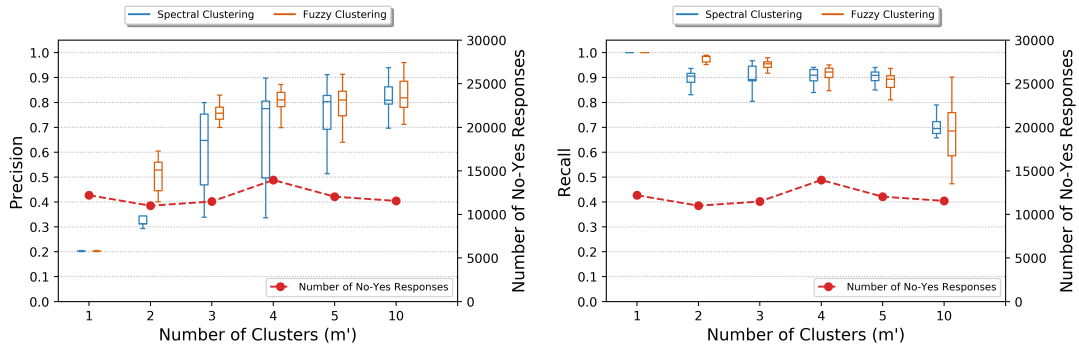


Figure 7.2: Precision and recall for the genome reconstruction of a newly added donor to OpenSNP beacon with varying number of bins/clusters (m') in the genome reconstruction attack. Number of newly added donors (m) is 5.

clustering-based genome reconstruction for the evaluations.

To show the benefit of utilizing a beacon (and beacon update) in its genome reconstruction attack, we also computed the reconstruction accuracy of an attacker when it only uses publicly available information (e.g., population statistics and victim's phenotype). As discussed, each victim we consider has a subset of 21 phenotypes listed in Chapter 7.1. Using the associations of victim's phenotypes with the corresponding SNPs (extracted from SNPedia [71]), we assigned some SNP values of the victim. We observed that, on the average, such a reconstruction achieves a precision of 18% and a recall of 47% on total of 232 SNPs. Therefore, we conclude that having access to a beacon and knowing the membership of a victim to a beacon significantly increases the success of the genome reconstruction attack.

To show the effect of varying number of bins (m') in the genome reconstruction attack, in Figures 7.2 and C.2 (in Appendix C), we show the attacker's success when the number of newly added donors $m = 5$ and beacon size $n = 50$ for OpenSNP and HapMap beacons, respectively. We observed that for both beacons, precision increases and recall decreases with increasing m'. Also, as expected, precision and recall becomes balanced when $m' = m$.

Next, in Figures 7.3 and C.3 (in Appendix C), we show the effect of the

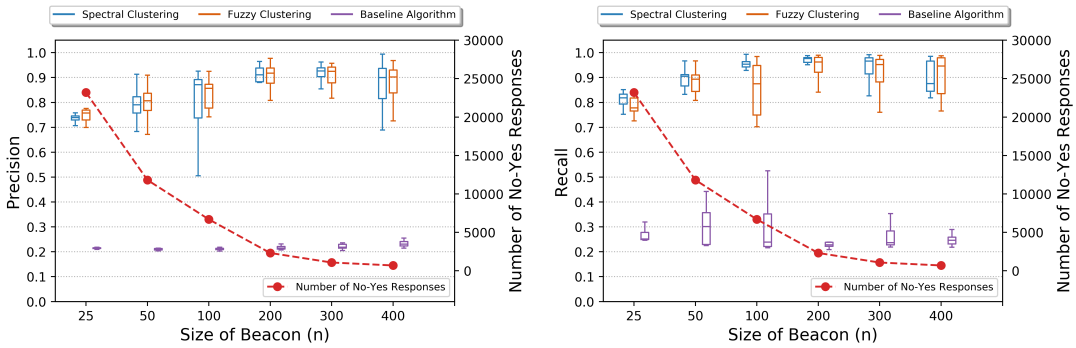


Figure 7.3: Precision and recall for the genome reconstruction of a newly added donor to OpenSNP beacon with varying number of beacon size (n). Number of newly added donors m is 5 and $m' = m$ for all plots.

beacon size (n) at time t when 5 new donors are added between times t and $t + \delta$ for OpenSNP and HapMap beacons, respectively. Here, we assume that the number of bins (m') is equal to the number of newly added donors (m). We observed that as the size of the beacon increases, both the precision and recall of the reconstruction attack almost remains the same (for a fixed number of newly added donors).

Even if the success of the genome reconstruction remains high, the number of flipped responses (from "no" to "yes") may decrease when beacon size is increased (as shown in Figure 7.3). In other words, the number of vulnerable SNPs (the ones that can be inferred using the change in the beacon responses) of a victim decreases and this might result in lower performance in phenotype inference and membership inference parts of the attack. However, with high probability, as the beacon size increase, low-MAF SNPs of the victim (which typically provide the most valuable information for the membership inference attack) still remain vulnerable, since with high probability, such SNPs are not observed in other donors in the beacon. For example, in the previous experiment (in Figure 7.3), when the size of the beacon is increased from 50 to 400, total number of vulnerable SNPs of a victim reduces by 94%, however, number of vulnerable SNPs of a victim with MAF value smaller than 0.01 only reduces by 52%.

Keeping the ratio of newly added donors fixed (to 5%), we also observed the

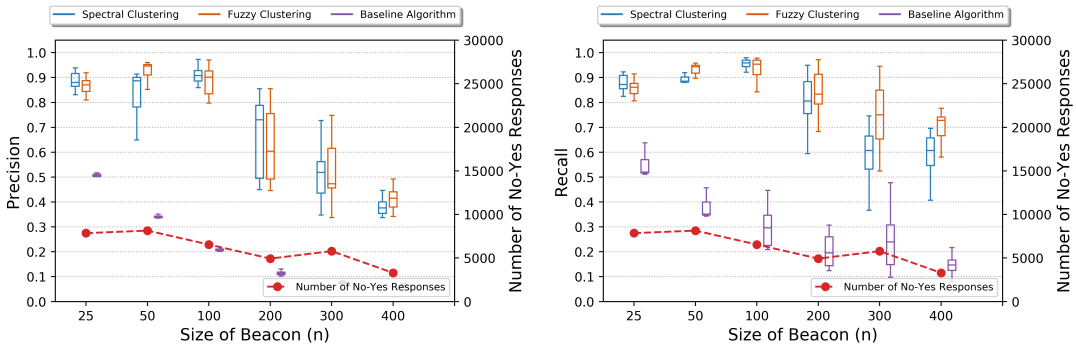


Figure 7.4: Precision and recall for the genome reconstruction of a newly added donor to OpenSNP beacon with varying number of beacon size (n). Number of newly added donors m is always 5% of the beacon size and $m' = m$ for all plots.

change in the success of the attack with increasing beacon size when $m' = m$ in Figure 7.4 (we did this evaluation only for the OpenSNP beacon since HapMap beacon did not have more than 100 donors). We observed that, when the beacon size increases beyond 100, although the recall of the attacker still remains high, its precision starts decreasing. This shows that the success of the identified attack mainly relies on the number of clusters the attackers needs to generate (in the proposed clustering-based algorithm). For small or mid-size beacons (e.g., NBDC Human Database [60] with slightly more than 100 individuals), even if the beacon update significantly increases beacon's size, the identified attack is still effective. On the other hand, for large size beacons (e.g., gnomAD [61], with more than 100K individuals), the update size should be small to have a vulnerability.

Finally, we explored the scenario, in which the attacker only has a partial snapshot of the beacon (instead of a full snapshot). In Figure 7.5, we show the success of the reconstruction attack when $m = 5$ donors are added (at time $t + \delta$) into the OpenSNP beacon with size 50 when the attacker has varying snapshots of the beacon at time t and when $m = m'$. We observed that the success (precision and recall) of reconstruction do not change with varying snapshots. However, the number of inferred SNPs (as a result of the genome reconstruction attack) decreases linearly with the decreasing snapshot that is known by the attacker at time t.

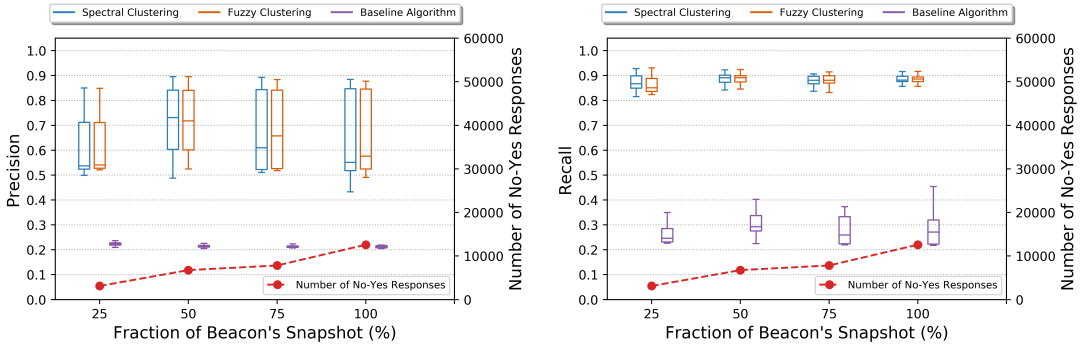


Figure 7.5: Precision and recall for the genome reconstruction of a newly added donor to OpenSNP beacon when the attacker knows varying fractions of beacon's snapshot. Number of newly added donors m is 5, beacon size n is 50 and $m' = m$ for all plots.

7.4 Identifying the Victim's Genome Using Phenotype Inference

Here, we evaluate the success of the attacker in identifying the reconstructed genome of the victim among all reconstructed genomes using the algorithm in Chapter 6.4. Since HapMap dataset does not include phenotype information about the genome donors, we only use the OpenSNP beacon for this evaluation.

We employed and compared several machine learning models for genotype-phenotype associations, including: Logistic Regression [72], SVM [73], Multi-layer Perceptron [74], Random Forest [75], and XGBoost [76]. Among these, we obtained the highest classifier accuracy with the Random Forest, and hence all reported results are based on this model.

In Figure 7.6, we show the ensemble classifier accuracy for varying number of newly added donors to the beacon (here, we assumed $m' = m$ and we observed similar patterns when $m' \neq m$ as well). We used the original genomes of individuals in the training dataset when building the model. For test, we used reconstructed genomes of the victims (that may have noise due to reconstruction error). Beacon size is 50 in these experiments (i.e., $n = 50$).

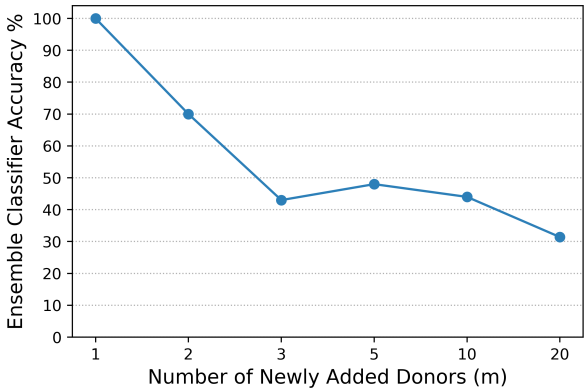


Figure 7.6: Classification accuracy of genotype inference from phenotype for varying number of newly added donors (m) to the beacon.

We observed that the proposed algorithm provides 70% accuracy when the size of the beacon is increased by adding 2 individuals in the update, and the accuracy slightly decreases with increasing number of newly added donors. These results show that the attacker can identify the reconstructed genome of the victim among all m' reconstructed genomes with high accuracy. As discussed before, in this experiment, we assumed the attacker has moderate auxiliary knowledge about the victim (i.e., phenotypic-traits, which can be easily learnt from social network profiles of the victim). However, since genotype-phenotype associations are not strong yet, there is an accuracy bottleneck in the overall process due to this step. A stronger attacker (that has access to richer auxiliary information about the victim) may utilize victim's known mutations or genomes of family members. Then, the phenotype inference part is not required and accuracy loss would not happen.

7.5 Using Genome Reconstruction in Membership Inference

In Chapter 7.3, we evaluated the success of the reconstruction and in Chapter 7.4, we showed that the attacker is able to identify the victim among many

added donors with high accuracy. Here, we show a severe consequence of the proposed genome reconstruction attack, in which the outcome of the previous steps can be utilized in a membership inference attack. By doing so, we also explicitly explore the impacts of (i) incorrect inference of some SNPs during reconstruction and (ii) imperfect choice of the reconstructed genotype due to the use of genotype-phenotype associations in terms of the success of this membership inference attack.

We randomly constructed two non-overlapping beacons from the OpenSNP dataset: (i) B_1 includes 50, and (ii) B_2 includes 60 individuals. We assume that B_2 is associated with a privacy-sensitive phenotype and the goal of the attacker is to infer the membership of the victim to B_2. We also assume that m new individuals are added to B_1 at time $t + \delta$ and the victim is among these newly joined donors. The attacker only knows that the victim is among these m individuals that are added to B_1 at time $t + \delta$ along with a snapshot of B_1 at time t.

First, we applied the spectral clustering-based genome reconstruction (that provides the best performance in Chapter 7.3) to reconstruct the genomes of newly joined m donors to B_1. Then, we identified the reconstructed genome of the victim using phenotype information about the victim (as in Chapter 6.4). Finally, using the reconstructed genome of the victim, we conducted the membership inference attack on B_2 using the optimal attack (as described in Appendix A).

We used the identification accuracy in Chapter 6.4 to construct and infer victims' genomes for alternate and null hypotheses. For instance, when $m = 2$ we have 70% identification accuracy. In this scenario, 14 genomes are chosen from correctly reconstructed genomes, while the remaining 6 genomes are chosen from incorrectly reconstructed genomes for corresponding victims.

In Figure 7.7, we show the power plots of this attack with varying number of newly added donors (m) to beacon B_1. As expected, with decreasing values of m, the power increases faster since the accuracy of genome reconstruction increases (and hence the error rate of the membership inference attack decreases). For

instance, when the victim is the only newly added donor to beacon B_1 ($m = 1$), the attacker can reconstruct their genome and then infer the victim's membership to beacon B_2 with a very high confidence (100% power) in just slightly more than 15 queries. We also observed that when m is increased, the power decreases, yet still reaches to 0.8 with approximately 80 queries when 2 individuals are added. These results show that the attacker may confidently conduct membership inference attacks as a result of genome reconstruction even though it has many sources of uncertainties in its input for membership inference.

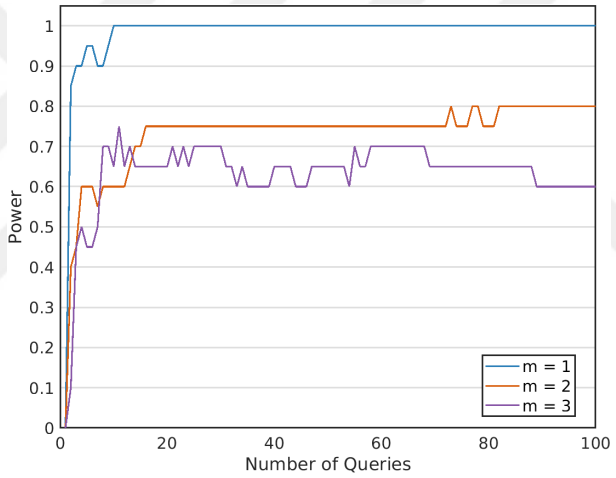


Figure 7.7: Power of membership inference attack on beacon B_2 with varying number of newly added donors (m) to beacon B_1.

Chapter 8

Discussion

This work pinpoints a new information leak and identifies beacon updates as a new risk, which leads to genome reconstruction attacks. We show that an attacker can efficiently and accurately link this new vulnerability to a membership inference attack. Furthermore, recently, we observed that some beacons even report the number of occurrences for a "yes" response (e.g., Sinai Health System Beacon in [10]). Using such information in the identified attack would further improve the accuracy of the proposed clustering-based algorithm (in Chapter **??**). We will explore this in future work.

8.1 Extension of the Proposed Attack

For the proposed genome reconstruction algorithm in Chapter **??**, we only focused on the "no" responses of the beacon at time t that turn to "yes" at time $t+\delta$ (i.e., no-yes responses) since such responses reveal the SNPs of the victim with minor alleles and minor alleles are typically the indicators for privacy-sensitive information about individuals. As a result, we also only considered the correlations between such SNPs of a victim. As briefly discussed before, no-no responses also provide deterministic information to the attacker (about the victim certainly not

having a minor allele in such SNP positions). Furthermore, using the information from no-no responses, the attacker can utilize the correlations between such SNPs (with no minor alleles) and others. In this work, we did not consider the no-no responses in the attack since (i) typically there are excessive number of no-no responses in a beacon and this creates a computational burden to compute all the pairwise correlations between such SNPs and (ii) SNPs with no minor alleles (learned from no-no responses) typically are not highly correlated with the SNPs with minor alleles, which are instrumental to the attacker for the membership inference attack. We will further consider the impact of such no-no responses in future work.

8.2 Donors Leaving the Beacon

In Chapters 6 and 7, we presented and evaluated the identified vulnerability by only considering the newly joined donors to the beacons. It is also possible that existing donors may leave the beacon. However, such a scenario can be easily addressed by using the identified attack mechanism. Considering the donors that leave the beacon brings up two different scenarios: (i) victim is among the newly joined donors (while there are also donors leaving the beacon between times t and $t+\delta$) and (ii) victim is among the donors that leave the beacon (while there may be other donors leaving or joining the beacon between times t and $t+\delta$).

Scenario in (i) is no different than what we discussed in Chapter 6. The number of "no" responses at time t that turn to "yes" at time $t+\delta$ does not change due to the donors leaving the beacon. On the other hand, some "yes" responses at time t may turn to "no" at time $t+\delta$ due to the donors leaving the beacon. However, such responses do not provide information about the minor alleles of the victim, and hence we do not consider such responses in this work. In scenario (ii), "yes" responses at time t that turn to "no" at time $t+\delta$ will provide information about the minor alleles of the victim (and other donors that leave the beacon during that time interval). Using such responses, one will need to run the algorithms proposed in Chapter 6 to reconstruct the genome of the victim.

8.3 Risk Quantification for the Genome Reconstruction Attack

The identified vulnerability and the proposed attack algorithm can be used as a privacy risk quantification tool by the beacon operator. For this, we foresee a simulation-based technique to quantify the risk and show it to the beacon operator. This will be a customized technique for each donor in the beacon and the following discussion is for one particular donor. Assume that a total of m new donors are gathered by the beacon between times t and $t + \delta$. To quantify the genome reconstruction risk, one may run the attack we introduced in Chapter 6, pretending the donor is added to the beacon along with the other $(m - 1)$ newcomer donors and compute the fraction of the SNPs that can be reconstructed. Then, using public sources (such as HapMap), one can gather a small number (e.g., s) of genomes belonging to individuals from the same population as the donor. Then, the same attack can be run for the selected s people (i.e., adding each random individual along with the other $(m - 1)$ newcomer donors), their reconstruction rates can be set as the baseline, and eventually, a privacy risk percentile can be provided for the donor. Moreover, for all correctly inferred SNPs, one can perform a pathogenic scan on ClinVar [77] to inform the donor about what traits they might be linked should their genome is put onto the beacon. Using this information and based on the privacy risk of the donor, either the donor or the beacon operator will decide whether or not to add the donor to the beacon at time $t + \delta$. This process can be repeated for all the newcomer donors.

We foresee that using such a quantification algorithm, a potential beacon participant can provide informed consent about how (and what portion of) their data can be used by the beacons (e.g., when the beacon can start using their data in its responses or when the beacon should stop using their data). Similarly, such a tool can guide a beacon operator on the number of participants to include in a batch to update the beacon.

8.4 Mitigation Techniques

To mitigate membership inference attacks against beacons, several countermeasures have been proposed [11, 12, 78]. However, most of such techniques directly reduce the utility of the beacon without carefully analyzing a balance between privacy (of beacon participants) and utility (of beacon responses). Thus, we believe that existing countermeasures proposed for membership inference are not directly applicable to mitigate genome reconstruction attack. To mitigate genome reconstruction, here we suggest three simple methods: (i) updating the beacon content considering the beacon size and the size of the update. For instance, as we showed in Chapter **??**, for small and mid-size beacons, even large-sized updates create a vulnerability, while for large-size beacons, only small-sized updates pose a threat; (ii) adding (or removing) donors after quantifying their risks against genome reconstruction (as discussed in Chapter **??**); and (iii) adjusting diversity of the beacon to have beacons with mixed ethnicity genome donors. For beacons with mixed ethnicity donors, it is hard to construct the correlation model (unless the beacon discloses the ethnicities of the donors as metadata), and hence it is hard to conduct the proposed correlation-based genome reconstruction attacks. It is worth noting that the OpenSNP beacon in our evaluations was a diverse one, however we also created the correlation model from the same diverse population (i.e., in our settings, the attacker had access to a very similar population to the target population). We will further work on more sophisticated countermeasures in future work.

Chapter 9

Conclusion

Thus far, the only privacy vulnerability that has been identified for beacons was membership inference. We have identified and, via extensive analysis, showed the impact of another serious privacy concern for beacons: genome reconstruction. We showed the practicality of the identified privacy concern in real-life by showing the whole attack strategy including genotype-phenotype inference. Furthermore, we showed how genome reconstruction attack can be used together with the membership inference to identify privacy-sensitive phenotypes of individuals.

Bibliography

[1] M. C. Schatz, "Biological data sciences in genome research," *Genome Research*, vol. 25, no. 10, pp. 1417–1422, 2015.

[2] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[3] H. Ledford, "Astrazeneca launches project to sequence 2 million genomes.," *Nature*, vol. 532, no. 7600, p. 427, 2016.

[4] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, vol. 4, no. 8, 2008.

[5] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic privacy and limits of individual detection in a pool," *Nature Genetics*, vol. 41, no. 9, pp. 965–967, 2009.

[6] K. B. Jacobs, M. Yeager, S. Wacholder, D. Craig, P. Kraft, D. J. Hunter, J. Paschal, T. A. Manolio, M. Tucker, R. N. Hoover, *et al.*, "A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies," *Nature genetics*, vol. 41, no. 11, pp. 1253–1257, 2009.

[7] P. M. Visscher and W. G. Hill, "The limits of individual identification from sample allele frequencies: theory and statistical analysis," *PLoS Genet*, vol. 5, no. 10, 2009.

[8] D. Clayton, "On inferring presence of an individual in a mixture: a Bayesian approach," *Biostatistics*, vol. 11, no. 4, pp. 661–673, 2010.

[9] `https://www.ga4gh.org/about-us/`, 2020. [Online; accessed 10-January-2020].

[10] `http://beacon-network.org`, 2020. [Online; accessed 10-January-2020].

[11] S. S. Shringarpure and C. D. Bustamante, "Privacy risks from genomic data-sharing beacons," *The American Journal of Human Genetics*, vol. 97, no. 5, pp. 631–646, 2015.

[12] J. L. Raisaro, F. Tramer, J. Zhanglong, D. Bu, Y. Zhao, K. Carey, D. Lloyd, H. Sofia, D. Baker, P. Flicek, S. S. Shringarpure, C. D. Bustamante, S. Wang, X. Jiang, L. Ohno-Machado, H. Tang, X. Wang, and J.-P. Hubaux, "Addressing beacon re-identification attacks: Quantification and mitigation of privacy risks," *The Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 799–805, 2016.

[13] N. von Thenen, E. Ayday, and A. E. Cicek, "Re-identification of individuals in genomic data-sharing beacons via allele inference," *Bioinformatics*, vol. 35, no. 3, pp. 365–371, 2018.

[14] A. Salem, A. Bhattacharyya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," *ArXiv*, vol. abs/1904.01067, 2020.

[15] B. Greshake, P. E. Bayer, H. Rausch, and J. Reda, "Opensnp–a crowdsourced web resource for personal genomics," *PLoS One*, vol. 9, no. 3, p. e89204, 2014.

[16] I. H. Consortium *et al.*, "The international hapmap project," *Nature*, vol. 426, no. 6968, p. 789, 2003.

[17] `https://ghr.nlm.nih.gov/primer/genomicresearch/snp`, 2020. [Online; accessed 10-January-2020].

[18] "Disease risk," 2020. [Online; accessed 10-January-2020].

[19] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 409–421, 2014.

[20] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang, "Privacy in the genomic era," *ACM Computing Surveys (CSUR)*, vol. 48, no. 1, p. 6, 2015.

[21] E. Ayday, E. De Cristofaro, J.-P. Hubaux, and G. Tsudik, "The chills and thrills of whole genome sequencing," 2013.

[22] J. Gitschier, "Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project," *American Journal of Human Genetics*, vol. 84, no. 2, pp. 251–258, 2009.

[23] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[24] E. C. Hayden, "Privacy protections: The genome hacker," *Nature*, vol. 497, pp. 172–174, 2013.

[25] B. A. Malin and L. Sweeney, "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems," *Journal of Biomedical Informatics*, vol. 37, no. 3, pp. 179–192, 2004.

[26] L. Sweeney, A. Abu, and J. Winn, "Identifying participants in the personal genome project by name," *arXiv preprint arXiv:1304.7605*, 2013.

[27] Z. Lin, A. B. Owen, and R. B. Altman, "Genomic research and human subject privacy," *Science*, vol. 305, p. 183, Jul 2004.

[28] G. Kale, E. Ayday, and Ö. Tastan, "A utility maximizing and privacy preserving approach for protecting kinship in genomic databases," *Bioinformatics*, 2017.

[29] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving

individuals contributing trace amounts of dna to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, vol. 4, no. 8, 2008.

[30] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou, "Learning your identity and disease from research papers: Information leaks in genome wide association study," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09, (New York, NY, USA), p. 534–544, Association for Computing Machinery, 2009.

[31] H. K. Im, E. R. Gamazon, D. L. Nicolae, and N. J. Cox, "On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy," *American Journal of Human Genetics*, vol. 90, no. 4, pp. 591–598, 2012.

[32] X. Zhou, B. Peng, Y. F. Li, Y. Chen, H. Tang, and X. Wang, "To release or not to release: Evaluating information leaks in aggregate human-genome data," *ESORICS'11: Proc. of the 16th European Conf. on Research in Computer Security*, pp. 607–627, 2011.

[33] C. Dwork, "Differential privacy," *Proceedings of the 33rd International Conference on Automata, Languages and Programming*, 2006.

[34] S. E. Fienberg, A. Slavkovic, and C. Uhler, "Privacy preserving GWAS data sharing," in *IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, pp. 628–635, 2011.

[35] F. Yu, S. E. Fienberg, A. B. Slavković, and C. Uhler, "Scalable privacy-preserving data sharing methodology for genome-wide association studies," *Journal of Biomedical Informatics*, vol. 50, pp. 133–141, 2014.

[36] A. Johnson and V. Shmatikov, "Privacy-preserving data exploration in genome-wide association studies," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1079–1087, ACM, 2013.

[37] F. Tramer, Z. Huang, J.-P. Hubaux, and E. Ayday, "Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies," in *Proceedings of ACM Conference on Computer and Communications Security (CCS)*, pp. 1286–1297, 2015.

[38] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the Lacks family: quantification of kin genomic privacy," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1141–1152, ACM, 2013.

[39] I. Deznabi, M. Mobayen, N. Jafari, O. Tastan, and E. Ayday, "An inference attack on genomic data using kinship, complex correlations, and phenotype information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 15, no. 4, pp. 1333–1343, 2018.

[40] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, and J.-P. Hubaux, "Deanonymizing genomic databases using phenotypic traits," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, pp. 99–114, 2015.

[41] C. Lippert, R. Sabatini, M. C. Maher, E. Y. Kang, S. Lee, O. Arikan, A. Harley, A. Bernal, P. Garst, V. Lavrenko, K. Yocum, T. Wong, M. Zhu, W.-Y. Yang, C. Chang, T. Lu, C. W. H. Lee, B. Hicks, S. Ramakrishnan, H. Tang, C. Xie, J. Piper, S. Brewerton, Y. Turpaz, A. Telenti, R. K. Roby, F. J. Och, and J. C. Venter, "Identification of individuals by trait prediction using whole-genome sequencing data," *Proceedings of the National Academy of Sciences*, 2017.

[42] M. Kayser and P. de Knijff, "Improving human forensics through advances in genetics, genomics and molecular biology," *Nature Reviews Genetics*, vol. 12, no. 3, pp. 179–192, 2011.

[43] D. Zubakov, F. Liu, M. Van Zelm, J. Vermeulen, B. Oostra, C. Van Duijn, G. Driessen, J. Van Dongen, M. Kayser, and A. Langerak, "Estimating human age from T-cell DNA rearrangements," *Current Biology*, vol. 20, no. 22, pp. R970–R971, 2010.

[44] X.-l. Ou, J. Gao, H. Wang, H.-s. Wang, H.-l. Lu, and H.-y. Sun, "Predicting human age with bloodstains by sjTREC quantification," *PloS ONE*, vol. 7, no. 8, 2012.

[45] H. L. Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, *et al.*, "Hundreds of variants clustered in genomic loci and biological pathways affect human height," *Nature*, vol. 467, no. 7317, pp. 832–838, 2010.

[46] A. K. Manning, M.-F. Hivert, R. A. Scott, J. L. Grimsby, N. Bouatia-Naji, H. Chen, D. Rybin, C.-T. Liu, L. F. Bielak, I. Prokopenko, *et al.*, "A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance," *Nature Genetics*, vol. 44, no. 6, pp. 659–669, 2012.

[47] S. Walsh, F. Liu, K. N. Ballantyne, M. van Oven, O. Lao, and M. Kayser, "IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information," *Forensic Science International: Genetics*, vol. 5, no. 3, pp. 170–180, 2011.

[48] P. Claes, D. K. Liberton, K. Daniels, K. M. Rosana, E. E. Quillen, L. N. Pearson, B. McEvoy, M. Bauchet, A. A. Zaidi, W. Yao, *et al.*, "Modeling 3D facial shape from DNA," *PLoS Genetics*, vol. 10, no. 3, 2014.

[49] F. Liu, F. van der Lijn, C. Schurmann, G. Zhu, M. M. Chakravarty, P. G. Hysi, A. Wollstein, O. Lao, M. de Bruijne, M. A. Ikram, *et al.*, "A genome-wide association study identifies five loci influencing facial morphology in europeans," *PLoS Genetics*, vol. 8, no. 9, 2012.

[50] J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik, "Privacy preserving error resilient DNA searching through oblivious automata," *Proceedings of ACM CCS '07*, 2007.

[51] E. De Cristofaro, S. Faber, and G. Tsudik, "Secure genomic testing with size- and position-hiding private substring matching," in *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*, 2013.

[52] S. Jha, L. Kruger, and V. Shmatikov, "Towards practical privacy for genomic computation," in *Proceedings of IEEE Symposium on Security and Privacy*, pp. 216–230, 2008.

[53] M. Blanton, M. J. Atallah, K. B. Frikken, and Q. Malluhi, "Secure and efficient outsourcing of sequence comparisons," in *Proceedings of European Symposium on Research in Computer Security*, pp. 505–522, 2012.

[54] M. Naveed, S. Agrawal, M. Prabhakaran, X. Wang, E. Ayday, J.-P. Hubaux, and C. Gunter, "Controlled functional encryption," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014.

[55] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik, "Countering GATTACA: efficient and secure testing of fully-sequenced human genomes," in *Proceedings of the 18th ACM conference on Computer and communications security*, pp. 691–702, 2011.

[56] E. Ayday, J. L. Raisaro, J.-P. Hubaux, and J. Rougemont, "Protecting and evaluating genomic privacy in medical tests and personalized medicine," in *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*, pp. 95–106, 2013.

[57] I. Hagestedt, Y. Zhang, M. Humbert, P. Berrang, H. Tang, X. Wang, and M. Backes, "Mbeacon: Privacy-preserving beacons for DNA methylation data," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*, 2019.

[58] G. Glusman, J. Caballero, D. E. Mauldin, L. Hood, and J. C. Roach, "Kaviar: an accessible system for testing snv novelty," *Bioinformatics*, vol. 27, no. 22, pp. 3216–3217, 2011.

[59] Verizon, "Verizon fios home internet," 2021.

[60] `https://humandbs.biosciencedbc.jp/en/hum0029-v1`, 2020. [Online; accessed 03-December-2020].

[61] https://gnomad.broadinstitute.org/, 2020. [Online; accessed 03-December-2020].

[62] S. S. Samani, Z. Huang, E. Ayday, M. Elliot, J. Fellay, J.-P. Hubaux, and Z. Kutalik, "Quantifying genomic privacy via inference attack with high-order SNV correlations," in *Security and Privacy Workshops (SPW), 2015 IEEE*, pp. 32–40, 2015.

[63] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856, 2002.

[64] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. d. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PloS one*, vol. 14, no. 1, p. e0210236, 2019.

[65] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[66] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, and J.-P. Hubaux, "De-anonymizing genomic databases using phenotypic traits," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 2, pp. 99–114, 2015.

[67] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[68] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *CoRR*, vol. abs/1106.1813, 2011.

[69] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, *et al.*, "The international HapMap project," *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[70] "OpenSNP." http://opensnp.org, 2020. [Online; accessed 10-January-2020].

[71] "SNPedia." https://www.snpedia.com/, 2020. [Online; accessed 10-January-2020].

[72] J. Cramer, "The origins of logistic regression," *Tinbergen Institute, Tinbergen Institute Discussion Papers*, 01 2002.

[73] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[74] C. von der Malsburg, "Frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms," *Brain Theory*, pp. 245–248, 01 1986.

[75] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282 vol.1, 1995.

[76] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016.

[77] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, *et al.*, "Clinvar: improving access to variant interpretations and supporting evidence," *Nucleic acids research*, vol. 46, no. D1, pp. D1062–D1067, 2017.

[78] M. M. Al Aziz, R. Ghasemi, M. Waliullah, and N. Mohammed, "Aftermath of Bustamante attack on genomic beacon service," *BMC Medical Genomics*, vol. 10, no. 2, p. 43, 2017.

Appendices

Appendix A

Membership Inference Attack Against Genomic Data-Sharing Beacons

In [12], Raisaro *et al.* introduced the "optimal attack" using the same attacker assumptions discussed in Section 3.2. In optimal attack, the attacker constructs a set of candidate SNPs S to be queried and submits queries starting from the lowest MAF SNP_i. Let the null hypothesis (H_0) refer to the case in which the queried genome is not in the beacon and alternative hypothesis (H_1) be the case in which the queried genome is a member of the beacon. In [12], the log-likelihood (L) under the null and alternate hypothesis are shown as follows:

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

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

where R is the response set, x_i is the answer of the beacon to the query at position i (1 for "yes", 0 for "no"), and γ represents a small probability where the attacker's copy of the victim's genome does not match the beacon's copy for a locus (e.g., due to difference in variant calling pipeline). n is the number

of posed queries. D_N^i is the probability that none of the N individuals in the beacon has the queried allele at position i and D_{N-1}^i represents the probability of no individual except for the queried person having the queried allele at position i. The computations of D_{N-1}^i and D_N^i depend on the queried position i and they change at each query as follows: $D_{N-1}^i = (1-f_i)^{2N-2}$ and $D_N^i = (1-f_i)^{2N}$, where f_i represents the MAF of the SNP at position i. The likelihood-ratio test (LRT) statistic, Λ, is then determined as

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

Appendix B

Baseline Approach for Genome Reconstruction

The details of this baseline approach for genome reconstruction (described in Section 6.1) are shown in Algorithm 2.

Algorithm 2: Baseline Algorithm for Genome Reconstruction Attack

Input: b: beacon; m: Number of added people to b; Population P that represent the composition in b

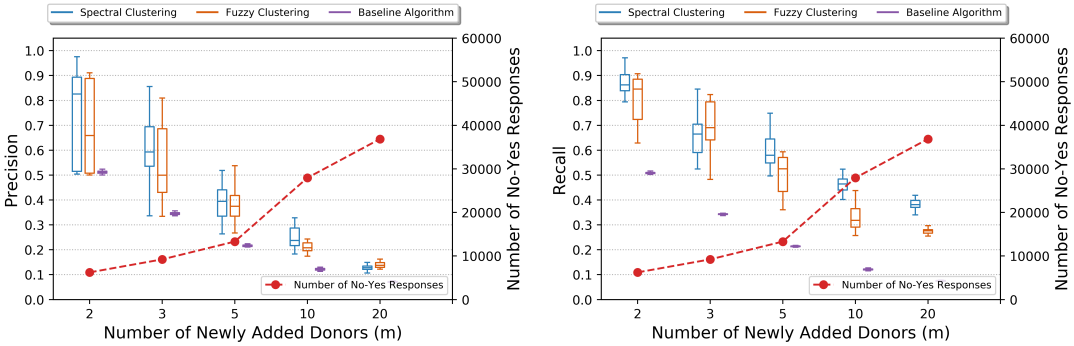
Output: m' reconstructed genomes

```
// Step 1: Query Beacon
snapshot1 ← queryBeacon(b, t)
// Including victim, m donors join Beacon between time t and t + δ
snapshot2 ← queryBeacon(b, t + δ)

// Step 2: Obtain No-Yes SNPs
NoYesResponses ← []
for i ← 0 to snapshot1.length do
    if snapshot1[i] = "No" and snapshot2[i] = "Yes" then
        NoYesResponses.append(i)
    end
end

// Step 3: Reconstruct genomes
S ← []
for i ← 0 to NoYesResponses.length do
    s ← NoYesResponses[i]
    assigned ← False
    for j ← 0 to m' do
        S[j][s] ← getMajorAllele(P, s)
        randnum ← getRandomFloat(0, 1)
        if randnum ¡ getMAF(P,s) then
            S[j][s] ← getMinorAllele(P, s)
            assigned ← True
        end
    end

    // Step 4:If a SNP is unassigned, randomly assign it to a reconstruction
    if !assigned then
        randnum ← getRandomInteger(0, m')
        S[randnum][s] ← getMinorAllele(P, s)
    end
end

return S
```

Appendix C

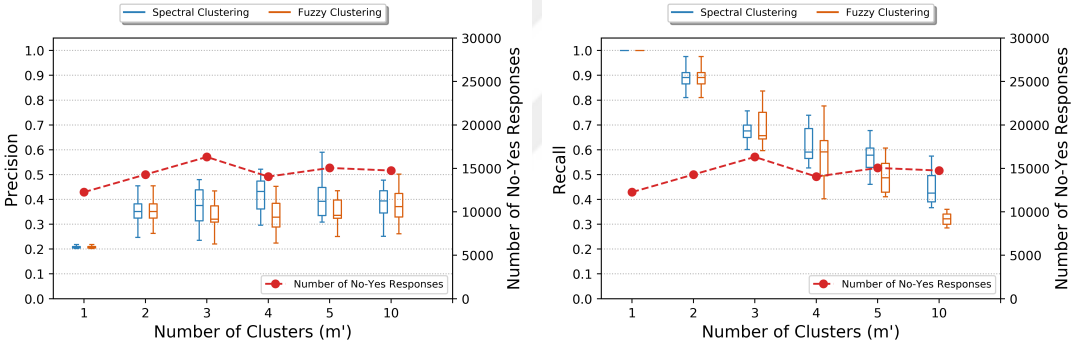
Evaluation of Genome Reconstruction on the HapMap Beacon

In Figure C.1 we show the success (precision, recall, and accuracy) of the reconstruction for various number of newly added donors (m) in HapMap beacon. In Figure C.2, we show the effect of varying number of bins (m') in the genome reconstruction attack when the number of newly added donors (m) is 5 for HapMap beacon. Next, in Figure C.3, we show the effect of the beacon size (n) at time t when 5 new donors are added between times t and $t + \delta$ for HapMap beacon.



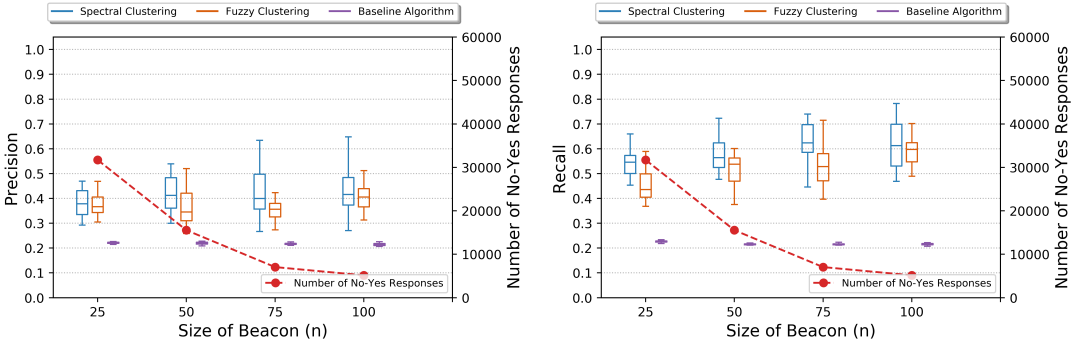
(a) Precision. (b) Recall.

Figure C.1: Precision and recall for the genome reconstruction of a newly added donor to HapMap beacon with varying number of newly added donors.



(a) Precision. (b) Recall.

Figure C.2: Precision and recall for the genome reconstruction of a newly added donor to HapMap beacon with varying number of bins/clusters (m') in the genome reconstruction attack. Number of newly added donors (m) is 5.



(a) Precision. (b) Recall.

Figure C.3: Precision and recall for the genome reconstruction of a newly added donor to HapMap beacon with varying number of beacon size (n). Number of newly added donors m is 5 and $m' = m$ for all plots.