# ANALYZING THE EFFECT OF KINSHIP FOR RE-IDENTIFICATION ATTACKS IN 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
Miray Ayşen
August 2019

Analyzing the Effect of Kinship for Re-identification Attacks in Genomic Data Sharing Beacons
By Miray Ayşen
August 2019

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.

Erman Ayday(Advisor)

Cevdet Aykanat

Hüsrev Taha Sencar

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

# ABSTRACT

## ANALYZING THE EFFECT OF KINSHIP FOR RE-IDENTIFICATION ATTACKS IN GENOMIC DATA SHARING BEACONS

Miray Ayşen
M.S. in Computer Engineering
Advisor: Erman Ayday
August 2019

Genomic data contains sensitive information about an individual. Family members' genome sequence can be re-constructed with high confidence or individuals' may face discrimination because of predisposition of a disease if genome sequence of a person is obtained. To protect the genomic information and provide a standardize and secure way for using this data the "Beacon project" initiated. Studies show that the genomic data sharing beacons are vulnerable to re-identification attacks. Since beacons generally constructed based on types of diseases, re-identification creates a significant risk for individuals. On the other hand, genomic data enables researchers to find the cause of diseases and improves personalized medicine. Previously proposed counter measures against re-identification attacks proved to be not effective as earlier researches show. In this thesis, we analyze the kin relationships' effect on the genomic data sharing beacons. Our study is based on the fact that kinship may be misleading for re-identification attacks since same SNPs can be appear in multiple family members. We showed that adding at least one of the parents to the beacon (i) cause significant decrease in the power of attacks and (ii) increase in the number of queries needed to confirm an individual's beacon membership. To investigate the suitability of using kinship as a counter measure for beacons we also calculate the utility decrease. We further show the effects of adding more distant relatives to the beacon such as grandparents.

*Keywords:* beacon, genome privacy, counter measure.

# ÖZET

# AKRABALIK İLİŞKİLERİNİN İSTATİSTİKSEL VERİTABANLARINA YAPILAN KİMLİK TESPİTİ ATAKLARINA KARŞI ETKİLERİNİN İNCELENMESİ

Miray Ayşen
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Erman Ayday
Ağustos 2019

Gen verisi kişiye özel hassas bilgiler içerir. Kişinin gen verisi kullanılarak aile üyelerinin gen dizilimleri yüksek olasılıkla doğru olacak şekilde tahmin edilebilir veya kişi hastalık yatkınlıkları nedeniyle ayrımcılığa maruz kalabilir. Gen verilerini korumak ve bu verilerin kullanımını standart ve güvenli bir hale getirmek için "Beacon" sistemleri geliştirildi. Yapılan araştırmalar, beacon sistemlerinin bir bireyin veri setinde olup olmadığını anlamak için yapılan kimlik tespiti ataklarına karşı yetersiz olduğunu göstermiştir. Beacon sistemleri genelde belirli bir hastalığı içeren bireylerden oluştuğu için, bir bireyin veri bankasında olup olmadığının tespit edilmesi bireyin kişisel gizliliğini tehdit eden önemli bir unsurdur. Bunun yanı sıra gen verileri hastalıkların nedeni olan genlerin tespit edilmesi ve kişiselleştirilmiş tıp alanlarında yapılan çalışmalar için önemli bilgi kaynaklarıdır. Beacon sistemlerini kimlik tespiti ataklarına karşı korumak adına şu ana kadar alınan önlemlerin yetersiz kaldığı araştırmalar sonucunda gösterilmiştir. Bu tezde akrabalık ilişkilerinin gen verisi paylaşan beacon sistemleri üzerindeki etkisi araştırıldı. Çalışmamız, aynı nokta mutasyonları birden fazla aile üyesinde bulunabileceği için, akrabalığın kimlik tespiti atakları için yanıltıcı olabileceği gerçeğine dayanmaktadır. Çalışmamız sonucunda ebeveynlerden en az birinin beacon sistemine eklenmesinin (i) atakların başarı oranında önemli bir düşüşe yol açtığını ve (ii) bir bireyin veri setinde olup olmadığını anlamak için gereken sorgu sayısında artışa sebep olduğunu gösterdik. Akrabalık ilişkilerinin beacon sistemlerinde bir savunma mekanizması olarak kullanılmasının beacon sistemini nasıl etkileyeceğini araştırmak adına sistemdeki fayda azalması hesaplamaları da yapıldı. Son olarak büyük anne ve büyük baba gibi bireye daha uzak akrabalar eklemenin beacon sistemleri üzerindeki etkilerini analiz edildi.

*Anahtar sözcükler*: beacon, gen verisi mahremiyeti, savunma mekanizması.

# Acknowledgement

Firstly I would like to express my gratitude to my advisor Asst. Prof. Erman Ayday for accepting me into his research group as a master student and giving me the opportunity to work under his supervision. I am so grateful for him for introducing me how to make research, broaden my perspective, answer my questions whenever I need help and sharing his invaluable knowledge. I also would like to thank Asst. Prof. A. Ercüment Çiçek for his support and valuable feedback whenever I need it.

I would also like to thank the "The Office" group that includes Alper Eroğlu, Cihan Eryonucu, Çağlar Öksüz, Gizem Çaylak, Onur Karakaşlar and Ömer Gözüaçık for all their supports and sharing their knowledge and helping me during the master's.

Finally, I would like to thank you my dearest friends, my sisters from other mothers Begüm Köktürk and Elif Deniz Haberal for being there for me whenever I need support, giving me motivation and renewing my belief in myself. I would like to thank my family who gave me the chances that led me where I am today. I am absolutely lucky to have that kind of support.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Researchers are excited regarding the increase of genomic datasets as a result of developments in the sequencing technologies as it creates new opportunities for detecting the cause of diseases and enhances the treatments based on these genetic information [1]. Also people have the chance to benefit from these studies to become aware of their genetic roots and to take precautions against possible diseases and health risks. Genomic data contains sensitive information which is private for an individual and as a result of that there is a need for genetic data sharing with anonymization. While sharing genetic data with anonymization, there is always a risk of leakage and revealing the real identities of people. The reveal of identity can cause discrimination such that an employer refuse to hire a person who has carried a genetic mutation of Gaucher's disease even though the person does not show any symptoms for the disease [1]. Several studies show the genetic discrimination in different areas such as employment, insurance, etc [1, 2, 3].

The genomic information about a person does not affect only the privacy of the individual but also the privacy of his/her relatives since genomic information can be used to infer kin relations. In other words, when a person decides to share his/her genetic data there is a risk of revealing genetic predisposition to a disease or jeopardizing the privacy of the other family members [4]. Genomic

information of a person can be used to infer kin relations. As a recent study shows the Single-nucleotide polymorphism (SNPs) of relatives can be reconstructed with high confidence by using Mendel Laws, correlation between SNPs and minor allele frequencies (MAFs) of the population [5]. So it is clear that there is a trade off between the risk of re-identification and the significant studies that can be done with the genomic data.

The Global Alliance for Genomics and Health (GA4GH) introduces the Beacon Project and beacons are constructed with the aim of providing a secure and systematic way for using genomic data. Genomic beacons provide an interface that only gives yes/no answers to the user queries about whether a specific nucleotide is at a specific position in a particular chromosome. Beacons are considered as secure and privacy provided since query results are just "yes" or "no" answers without including any information about allele frequencies. By this way it is believed that genomic data can be used without any disclosure of individuals. Nonetheless, previous studies show that beacons are vulnerable to re-identification attacks [6, 7, 8] In 2015, Shringarpure and Bustamante develop a likelihood ratio test (LRT) that predicts the membership of an individual to the beacon with high confidence by recurrently querying the beacon with the SNPs of person. The attack that introduced by Shringarpure and Bustamante is called SB Attack [6]. This study clearly shows that the beacons conceal the identity of individuals even though the allele frequencies are not provided. Another work on that area, moves SB Attack one step forward and assume that the attacker also have knowledge about MAFs of the population. Before querying the beacon, SNPs of the victim sorted according to MAFs in ascending order. Since the SNPs with lower MAFs is more elucidator, asking the SNPs with lower allele frequencies decrease the number of queries needed to determine the membership of an individual to the beacon [7]. This attack developed by Raisaro *et al.* called Optimal Attack.

After the demonstration of vulnerabilities of beacons to re-identification attacks counter measures are proposed. Counter measures to make beacons resistant to these attacks include solutions such as hiding the single nucleotide polymorphisms(SNPs) that have minor allele frequency lower than a pre-defined

threshold, sharing only some regions of the genome, giving wrong answers for some queries or limit the number of queries that can be asked by a single user [6, 7, 8]. These counter measures are considered as inefficient for different reasons; re-identification attacks can still be achieved or the utility of the beacon can significantly decrease hence useful results cannot be obtained as researches on the subject have proved. In short, proposed counter measure may decrease the utility of beacon and may inhibit the beneficial results that can be reached by researches. In this study we aim to find a solution such that the privacy of the individuals in the beacon can be protected against re-identification attacks without inhibiting the beneficial results that can be achieved by the researchers.

In this thesis, kinship is used as a counter measure for the re-identification attacks and it is showed that the power of re-identification attacks significantly decrease when at least one of the parents of an individual is added to the beacon. The basis of the research is that kin relationships are misleading while re-identifying an individual in the beacon since the same single nucleotide polymorphisms may occur in multiple family members. Our results shows that adding relatives to the beacon increases the data that is available for researches while protecting the beacon member's privacy. The contributions of this thesis can be summed up as follows:

- The number of queries that is needed to confirm the beacon membership of individual when at least one family member is in the beacon is at least the number of queries in the case of no relative is in the beacon. In short, the number of queries needed to re-identify person in the beacon tends to increase.
- The power of the re-identification attacks significantly decrease when at least one of the parent is added to the beacon.
- Adding relatives to the beacon increase the data that is available for researches while protecting the beacon member's privacy without a significant decrease the utility of beacon.

In this thesis, we will firstly give a brief background information in Chapter 2. Then we will continue by presenting literature review and present the related works in Chapter 3. In Chapter 4, we present the datasets that are used in experiments and explain the family generation algorithm. In Chapter 5, our evaluation strategy and the re-identification attacks on genomic data sharing beacons with family members are explained and the results are discussed. In Chapter 6, we discuss how our proposed counter measure effects the utility of beacon. Finally, in Chapter 7, we give a brief conclusion to sum up the thesis.

# Chapter 2

# Background Knowledge About Genomics

In this chapter definitions of the terms that used throughout the thesis can be found.

- **Allele** For each characteristic, humans inherit genes one from each of their parents. These genes may vary and the individual can show different characteristics from their parents. The variations in the characteristics are called allele. In other words, the different forms of gene are called allele. A basic example for this can be the eye color of people.
- **Minor Allele Frequency (MAF)** The allele that commonly occurs in a population called major allele. The allele that is less common in a population called minor allele. The minor allele frequency indicates the incidence rate of minor allele in a particular population.
- **Single Nucleotide Polymorphisms (SNPs)** SNPs are the variations on the DNA sequence that occurs when a single nucleotide is different in at least 1% of the population. SNP positions can be observed as any of three: major homozygous, minor homozygous or heterozygous as can be seen in Figure 2.1. Major homozygous SNP means that major allele for a specific

gene is appeared on both homologous chromosomes and similarly, minor homozygous SNP means that minor allele for a specific gene is appeared on both homologous chromosomes. When two different alleles appear on homologous chromosomes for a particular gene it is called heterozygous SNP position [1].

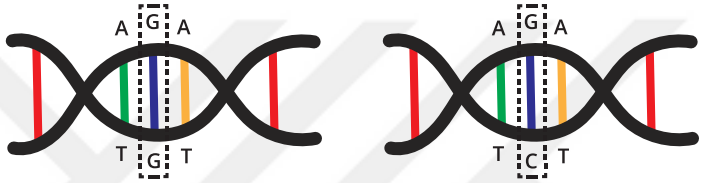


Figure 2.1: Left side of the figure shows the homozygous SNP with two different nucleotide and an example of heterozygous SNP can be found on the right side of figure.

- **Linkage Frequencies (LD)** Show the correlations between the SNP pairs. The correlation demonstrates the possibility that SNPs occur together. In addition to that, LD can also indicate the SNPs do not occur together. In short, it statistically shows the dependence of SNPs to each other. Generally LD occur based on the position of the SNPs and observed on the alleles on the same chromosome.

- **Beacons** A project introduced by The Global Alliance for Genomics and Health (GA4GH) and aimed the usage of genomic data in a secure and standardize way without any privacy risk. Beacons are basically databases that contain genomic data and do not allow direct access to the information it contains. Beacons generally build in a way such that the individuals who have a particular disease gathered in one place. Beacons provide an interface that just gives "yes" or "no" answers for the queries about whether a specific nucleotide is at a specific position in a particular chromosome.

[1] https://en.wikipedia.org/wiki/ZygosityHomozygous

# Chapter 3

# Related Work

In this chapter a summary of the previous works in the area of genomic privacy and the proposed counter measures for re-identification attacks in genomic data sharing beacons are introduced. In addition to that, the explanation of re-identification attacks which are simulated to show the effectiveness of our proposed counter measure are given. In Section 2.1, the attack developed by Shringarpure and Bustamante, also called SB Attack, is represented. Optimal Attack that proposed by Raisaro *et al.* and takes SB Attack a step further is introduced in Section 2.2. A recently presented attack by von Thenen *et al.* that is an extended version of Optimal Attack that is introduced in Section 3.3 (referred to as Query Inference Attack, also known as QI Attack). Another attack approach that recently developed by von Thenen *et al.* is introduced in Section 3.4 (referred to as Genome Inference Attack, also known as GI Attack).

The evolvement in the sequencing technologies make the genomic data become widely available. While this development contributes the researches to achieve beneficial results such as personalized treatment, it also brings the risks for privacy as a result of re-identification of individuals. Survey conducted by Naveed *et al.* showed that researches in the biomedical domain believe that genomic privacy is an issue that should be given importance and revealed that approximately 75% the researchers believe that the beneficial results that can be obtain

by genome based researches do not justify the privacy risks that an individual may face by sharing [9]. Goodrich showed even though comparison protocols are cryptographically guaranteed, usage of these protocols on genomic data cause information leaks [10]. In [4], it is showed that the genome sequence of relatives can be inferred with high confidence so sharing genomic data also carries a privacy risk for other family members. In [11], discussed the privacy risks and proposed solutions to this issue. Malin showed that systems have deficits in terms of protecting the genomic data [12]. In short, various works show that privacy of genomic data is a concern that should be taken seriously so solutions to solve the problems related to this subject are searched.

Ayday *et al.* developed a schema that guaranteed privacy in usage of genomic data in medical tests by using homomorphic encryption and proxy re-encryption [13]. Raisaro *et al.* propose various counter measures for protecting the privacy in the genomic data sharing beacons [7]. Firstly, the beacons give "yes" answer even if only one person in the beacon contains queried allele. So they proposed to put a restraint such that the beacon give "yes" only if the asked allele occurs multiple times. The second solution they proposed is adding noise to alleles that are unique and giving wrong answers, give "no" instead of "yes", to the queries that targeted these unique alleles. Finally, they proposed a counter measure called as query budget to each individual in the beacon. The query budget is determined by the risk factor based on allele frequency and when beacon gives "yes" response to a query, query budget of every individual who has the queried position is decreased. Shringarpure and Bustamante also proposed various counter measures for protecting privacy in beacons [6]. Their proposed solutions includes, increasing the beacon size, sharing only small genomic regions in beacons, create beacons only contains people from the same population, not sharing metadata of beacon since the information such as population of the beacon members can reduce the security. In[14], a schema is introduced to protect privacy of beacons without a significant decrease in utility assuming a scenario that the attacker has knowledge about minor allele frequencies. In [15], a tool SecureGenome introduced to determine the SNPs that can be shared in without creating any risk for privacy using LD correlations. Aziz *et al.* discussed the

possible usage of homomorphic encryption, garbled circuit, secure hardware and differential privacy in genomic data privacy domain [16]. In another study of Aziz *et al.* two methods are proposed to preserve the privacy in beacons such as, (i) creating inaccurate results by adding bias to the beacon responses randomly and (ii) similar to the first solution result set is randomized based on bias factor [17]. The previously proposed solutions to solve the privacy problem in beacons either cause decrease in utility of beacon or it is showed that these counter measures are not effective. As a result of this a new solution is needed to achieve the privacy of the individuals in the beacon against re-identification attacks without inhibiting the beneficial results that can be achieved by the researchers.

## 3.1 Shringarpure and Bustamante's Attack

In 2015, Shringarpure and Bustamante introduce the SB Attack that shows membership of an individual to a genomic data sharing beacon may be detected with high confidence by repeatedly querying the beacon with minor alleles from the randomly selected heterozygous SNP positions [6]. In this attack scenario, it is assumed that the attacker has the VCF file of victim. The methodology is based on a likelihood ratio test (LRT) which gives information about the membership of an individual to the beacon. This test is based on the "yes" answers from the query results. In an ideal case if the victim is in the beacon all the query results should me 1, although in real there may be differences between the victim's the genome sequence that attacker has and the one that is in the beacon. To compensate these situations an error factor is included to the method. In SB Attack null hypothesis ($H_0$) defined as the case where queried genome is not in the beacon and the alternative hypothesis ($H_1$) defined as the case where queried genome is in the beacon. The log-likelihood for both hypothesis are defined as shown in Equations 3.1 and 3.2.

$$L_{H_0}(R) = \sum_{i=1}^{n} x_i \log(1 - D_N) + (1 - x_i) \log(D_N) \tag{3.1}$$

In the equation above, R represents the set of beacon responses, $D_N$ is the probability that no individual in the beacon has the queried allele at that position, $x_i$ stands for the beacon answer for queried position $i$ (value of $x_i$ may be either 1 if the beacon response is "yes" and 0, otherwise.) and $n$ gives the total number of queries asked to the beacon.

$$L_{H_1}(R) = \sum_{i=1}^{n} x_i \log(1 - \delta D_{N-1}) + (1 - x_i) \log(\delta D_{N-1}) \tag{3.2}$$

In the equation above, $D_{N-1}$ represents the probability that no individual in the beacon except the victim has the queried allele at that position and $\delta$ shows the possibility of error that is caused by the differences in sequencing the individuals' genome.

For this attack, Shringarpure and Bustamante defined the LRT as the difference between $L_{H_0}(R)$ and $L_{H_1}(R)$ as seen in 3.3.

$$\Lambda = L_{H_0}(R) - L_{H_1}(R) \tag{3.3}$$

## 3.2 Optimal Attack

In 2016, the Optimal Attack is introduced by Raisaro *et al.* [7] based on the previous work of Shringarpure and Bustamante [6]. The SB attack shows that by repeatedly querying beacon with the minor alleles from randomly selected heterozygous SNP positions, the beacon membership of an individual may be detected with high confidence. The Optimal attack take this attack scenario a step further and assume that the attacker also known the MAF values of the victim's population. As a result, in this scenario, first the attacker sorts the SNPs according to their MAF values and then querying the beacon starting from the minor allele that has the lowest MAF value.

The difference of this attack from the SB Attack is the computation method of variables $D_{N-1}$ and $D_N$ which shows the probability of no individual except from the queried person having the same SNP and the probability that no individual in the beacon has the queried allele at that position, respectively. When calculating the $D_{N-1}$ and $D_N$, MAF value of the SNP at the queried position $i$ also take into consideration. Since the computation depends on position $i$ at each iteration the value of these factors are changed and the calculation of these probabilities is shown in Equations 3.4 and 3.5.

$$D^i_{N-1} = (1 - f_i)^{2N-2}, \tag{3.4}$$

$$D^i_N = (1 - f_i)^{2N}, \tag{3.5}$$

where $f_i$ represents the MAF value of the SNP at that specific position $i$. The change in the computation of effects the calculation of $\Lambda$ as shown in Equations 3.6.

$$\begin{aligned} \Lambda &= \sum_{i=1}^{n} \log\left(\frac{D^i_N}{\delta D^i_{N-1}}\right) + \log\left(\frac{\delta D^i_{N-1}(1 - D^i_N)}{D^i_N(1 - \delta D^i_{N-1})}\right) x_i \\ &= \sum_{i=1}^{n} \log(\delta^{-1}(1 - f_i)^2) + \log\left(\frac{\delta}{(1 - f_i)^2} \frac{(1 - f_i)^{2N}}{1 - \delta(1 - f_i)^{2N-2}}\right) x_i \end{aligned} \tag{3.6}$$

## 3.3 Query Inference Attack

The QI Attack extends the Optimal attack further by making the assumption that beside from MAF information of population the attacker also has knowledge about linkage disequilibrium(LD) values [8]. The correlation between the SNPs are calculated based on these LD values. In this scenario, the attacker make a graph network with directed and weighted edges. The weights of edges correspond

to the possibility of two minor alleles occurring together. After an SNP is queried the beacon responses of the correlated SNPs to queried SNP are inferred. As a result of this, it is showed that the number of queries needed to ask to beacon for re-identification is decreased. The null hypothesis ($H_0$) and the alternative hypothesis ($H_1$) formulation is changed so that the new calculation reflect the inferred queries and take into consideration the inference error in beacon response. The equations for null hypothesis and the alternative hypothesis are shown in Equations 3.7 and 3.8.

$$\begin{aligned} L_{H_0}(R) = \sum_{i=1}^{n} \Big( & x_i \log(1 - D_N^i) + (1 - x_i) \log(D_N^i) \\ & + \sum_{j=1}^{m} \gamma x_i \log(1 - D_N^j) + \gamma(1 - x_i) \log(D_N^j) \Big), \end{aligned} \tag{3.7}$$

$$\begin{aligned} L_{H_1}(R) = \sum_{i=1}^{n} \Big( & x_i \log(1 - \delta D_{N-1}^i) + (1 - x_i) \log(\delta D_{N-1}^i) \\ & + \sum_{j=1}^{m} \gamma x_i \log(1 - \delta D_{N-1}^j) + \gamma(1 - x_i) \log(\delta D_{N-1}^j) \Big), \end{aligned} \tag{3.8}$$

where m represents the number of queries that can be inferred, $\gamma$ shows the assurance of the inferred beacon response. The $\Lambda$ value calculated with the same logic as can be seen in 3.9.

$$\Lambda = L_{H_0}(R) - L_{H_1}(R) \tag{3.9}$$

## 3.4 Genome Inference Attack

In GI Attack, the attacker, besides the information about MAF of the victim's population and LD values between SNPs, have knowledge about the high order

correlations between SNP pairs [8]. In this attack scenario the attacker firstly inferred the hidden positions in the victim's genome, the SNPs with a MAF lower than a pre-defined threshold, by using a $4^{th}$ order Markov chain. Then the minor allele from the inferred heterozygous SNP positions are queried to the beacon. One of the proposed counter measure for the re-identification attacks in genomic data sharing beacons is hiding the SNPs that have a lower MAF than a pre-defined threshold. GI Attack shows that this method is not effective to protect the beacon members.

# Chapter 4

# Datasets

In this chapter first we explain datasets that are used in experiments in Section 4.1. In Section 4.2, algorithm that generates SNPs of a person's relatives is explained.

## 4.1 Datasets for Experiments

To show that our counter measure is effective against existing re-identification attacks, we make the environment of experiments as about the same with the previous attacks. In other words, the datasets and parameters used in experiments are selected based on the previous studies. As a result, we can evaluate our work according to their results.

The beacon used in the experiments is consist of 65 individuals from the Utah Residents with Northern and Western European Ancestry (CEU) population of the HapMap project. In the paper, term "old beacon" is referred to the initial beacon set up meaning a beacon without containing any relatives of the individual. For some of the attack set up individuals mother or father or both of them are added on to that initial beacon setup. Furthermore, in some experiments person's grandparents on both sides of their parents are added. LD values, the

frequency distribution of major/minor homozygous or heterozygous SNPs and MAFs are calculated based on data obtained from 1000 Genome Project and HapMap.

The individuals who are used for attacks divided into two as case and control group. Individuals who are in the case group are selected as the first 20 people that are in the beacon and individuals who are selected as control group consist of 20 individuals who are outside of the beacon.

For each individual in case and control groups family members are generated. The generated family members include mother, father and grandparents. After generating family members for each individual in the case and control groups, these relatives added to the initial beacon set up systematically.

## 4.2 Family Generation

The SNPs of the mother and father of the individuals are generated based on the frequency distribution of major homozygous, minor homozygous or heterozygous SNP positions in the population (FD) and Mendel's first law. The Law of Segregation, Mendel first law, indicates for each SNP of offspring, one allele is taken from mother and the other allele taken from the father. In our algorithm, the genome of both parents are generated at the same time. Based on the genome of child, occurrence possibilities of all possible variations for parents' genome are calculated. Possible variations are determined based on Mendel's first law and the possibilities calculated based on distribution of alleles throughout the population. To preserve the randomness of the generation process, the genome couple for parents selected randomly. By doing this we do not eliminate the possibility that the genome of the parents can show a characteristic that does not frequently observed in the population. The possible variations for parents genome for different cases can be seen Table 4.1 and the algorithm for the family generation can be seen in the following algorithm.

**Algorithm 1** Generation of Parents

**Require:** SNP file of individual (SF), frequency distribution of SNP positions in the population (FD)

**for** each individual in case and control groups **do**

Open and read the SF

**for** each $SNP_i$ in SF **do** ▷ $i$ represents the SNP position at each iteration

According to $SNP_i$ all possibilities for all variations calculated.

Random value between [0 1] generated

The combination that has the closest possibility to that random value is selected

**end for**

**end for**

**Result:** SNPs of the parents are generated

In process of generating grandparents the same logic is applied. The only difference is the input parameter of the family generation algorithm. Instead of giving SNPs of individuals inside case and control groups, their mother/father's previously generated SNPs are given to the method.

For family generation we also try a different approach such as creating the genome of mother only based on frequency distribution of SNPs and generate the genome of father's by combining the Mendel's first law and the frequency distribution of SNPs. The drawback of this approach is that the variations on the child's genome can only be observed in father's genome. In our approach we also guarantee that the variations of the individual's genome can be caused by either of parents.

| | | Mother | Father |
|---|---|---|---|
| Child | AA | AA | AA |
| | | AA | Aa |
| | | Aa | AA |
| | | Aa | Aa |

(a)

| | | Mother | Father |
|---|---|---|---|
| Child | aa | aa | aa |
| | | Aa | aa |
| | | aa | Aa |
| | | Aa | Aa |

(b)

| | | Mother | Father |
|---|---|---|---|
| | | AA | Aa |
| | | AA | aa |
| Child | Aa | Aa | AA |
| | | Aa | Aa |
| | | Aa | aa |
| | | aa | AA |
| | | aa | Aa |

(c)

Table 4.1: A basic example for showing the possible SNPs for the parents for the three cases where the child's SNP is (a) major homozygous, (b) minor homozygous or (c) heterozygous.

# Chapter 5

# Effect of Kinship for Re-identification Attacks in Genomic Data Sharing Beacons

In this chapter first we explain the re-identification attacks on genomic data sharing beacons that contain family members in Section 5.1. Then in Section 5.2, the strategy to evaluate the success of our proposed counter measure against re-identification attacks is explained. In Section 5.3, the results are presented and discussed.

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

We claim that adding family members to the beacon will increase the privacy of beacon members. The origin of the idea of counter measure is the inheritance, the fact that an individual's genome is constructed based on their parents genomic information. We also claim that addition of relatives to the beacon does not cause a significant utility change, discussed further in Chapter 6. To show that our

proposed counter measure is effective, the re-identification attacks are simulated on artificially created beacons.

To show how the results of attacks change with the presence of family members in the beacon, same experiment parameters are used with the previous re-identification attacks. Detailed information about the used datasets can be found in Section 4.

| | **Known Information for Attacker** |
|---|---|
| **Optimal Attack** | MAF of Victim's Population |
| **Query Inference Attack** | MAF of Victim's Population<br>LD Values |
| **Genome Inference Attack** | MAF of Victim's Population<br>LD Values<br>High Order Correlation between SNPs |

Table 5.1: The attacker's knowledge for each of the attack scenarios that are simulated.

In our attack scenario, the attacker's aim is to learn whether the targeted individual is in the beacon or not. We assumed that besides the knowledge of attacker in the attacks as listed in Table 5.1, the attacker has only the VCF file of the victim. The attacker does not have access to VCF files of relatives of the victim. The attacker may or may not know if any family members of target is inside the beacon since it makes no difference. If attacker knows that a relative is in the beacon he cannot be sure the reason of the "yes" answers of beacon. In the opposite case, if the attacker does not have any knowledge about relatives beacon membership, the attacker thinks that he decided the membership of target which may lead to a wrong conclusion since a relative of victim can be the cause of "yes" responses of beacon.

We simulated the Optimal, QI and GI Attacks for different scenarios such as (i) the old beacon, (ii) the beacon that contains individual's mother, (iii) the beacon that contains individual's father, (iv) the beacon that contains both of individual's parents, (v) the beacon that contains individual's grandparents from mother's side and (vi) the beacon that contains individual's grandparents from father's side. These different scenarios can be seen in Figure 5.1.

One of the attack parameter $t$ represents a threshold such that the SNPs with a lower MAF value than $t$ are hidden. For Optimal Attack and QI Attack there are three scenarios such as (i) $t = 0$ (meaning no part of the genome is hidden), (ii) $t = 0.03$ and (iii) $t = 0.05$. For the GI Attack the scenario where $t = 0$ is not feasible since if hidden part is non-existing then the genome inference is meaningless.

Minor alleles from heterozygous SNP positions of the targeted individual are queried. The reason behind querying only these alleles is to have the same experiment set-up with previous re-identification attacks.

For GI Attack a $4^{\text{th}}$ order Markov chain is used for the inference process for hidden parts of the genome.

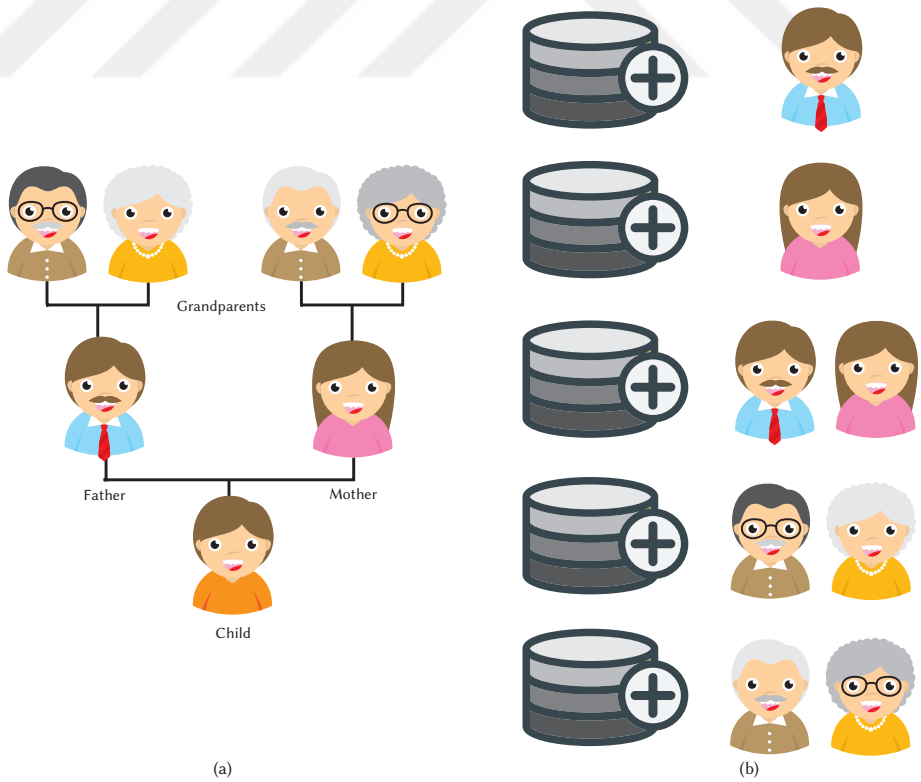


Figure 5.1: Experiment set up for analyzing the effects of kinship for the re-identification attacks (a) Family tree and (b) Beacons with relatives in for different attack scenarios. For each scenario different values of parameter $t$ is applied.

## 5.2 Evaluation Strategy

To show the effectiveness of our method same evaluation strategy is followed with previous studies. In earlier works, the success of the re-identification attacks measure by power. For each individual in case and control groups and for each query a $\Lambda$ value is calculated. (Details about calculation of $\Lambda$ can be found in Section 3) By using the $\Lambda$ values obtained from control group (people who are not in the beacon) a threshold, $t_\alpha$, is determined such that $\Lambda$ values of least 95% of control group (with 5% false positive rate) is greater than the determined $t_\alpha$. Power curve is obtained by applying this threshold to case group (people who are in the beacon). Basically, power curve presents the proportion of case individuals where we can reject null hypothesis.

## 5.3 Results and Discussion

When one of the family members, mother or father, presents in the beacon the number of queries needed to decide the membership of person to the beacon is at least the number of queries that are asked a beacon that does not contain any relatives. Therefore, it can be said that the number of queries that is needed to be confirm the membership of the individual in the beacon with a high confidence is most likely to increase. In Figure 5.3, it can be observed that power curve shifted to the left when one of the parents added into the beacon. The same behavior can also observed in the power curves of the other attacks (Figure 5.5 and Figure 5.4).

In [8], it is shown that the individual's membership to the beacon can be decided with a few queries. For instance, for GI attack it is showed that the power reaches to 1 with two queries. We showed that when relatives are in the beacon by only 2 queries, power is reached to 0.1. Meaning that the re-identification of an individual cannot be done with two queries(as shown in previous study) even if there is only one parent is in the beacon.

In the simulated attacks it is observed that the power of the attack decreases when a family member is in the beacon. As can be seen in Figure 5.4 when mother of person is in the beacon the power reaches approximately 0.55. Similarly, adding father of person to the beacon cause power to reach approximately 0.7. From these results we can conclude that the confidence decreases when a family member is in the beacon as can be seen in the power curves. In other words, by adding relatives to the beacon an attacker statistically cannot be determine the victim's beacon membership with high confidence.

Minor alleles of child either come from mother or father's genome so addition of parents inside the beacon cause a more significant decrease in power of the attacks.

We further investigate the behavior of power curve after converges and stabilize at a particular power value and observed that once power is converges it does not change.

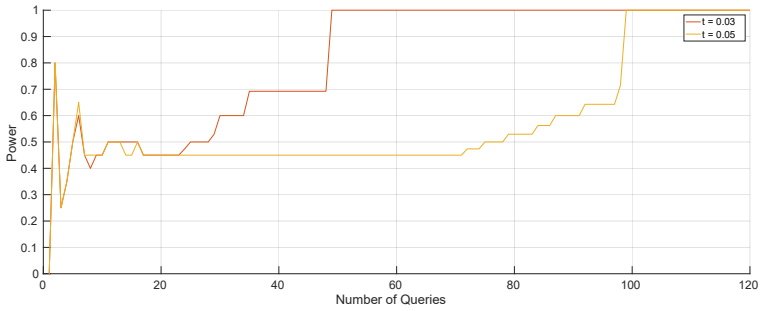


Figure 5.2: The case where the genome of mother is specifically created different than child. As a result the power of attack reaches one, meaning the attacker identified the membership of individual to the beacon with high confidence.

In order to investigate the reason why power does not reach the %100 in GI Attack, the attack is simulated again. For this scenario one of the parent is generate such that she does not carry the minor allele variations that the child has. As can be seen in Figure 5.2 in this case power reaches to %100. From this result we can derive that the more minor alleles in the genome of parents that are similar to the child, a more significant decrease can be observed in the power of the attacks.

The intersecting behavior of power curves is depends on the genome of parents. When mother's genome is generated with a different algorithm for Optimal Attack, the intersecting curves is observed (different from Figure 5.3). For different values for parameter $t$, SNPs are asked in different order so a "no" response from beacon can cause a increase in power.

If the attacker know at least one of the individual's parent is in the beacon and the beacon answers are come as "yes" then the attacker cannot be confident about the membership of the individual since the case where both of the parents are in the beacon and the individual is not in the beacon will give the same results as the case where the individual is in the beacon. In the other situation, where the attacker doesn't know anything about whether or not the parent(s) of the individual is in the beacon and the beacon answers come "yes" again the reason behind the beacon answers can be the individual's parents existence in the beacon.

Adding mother's parent to the beacon is almost equivalent to adding mother. Similarly, adding father's parent to the beacon is almost equivalent to adding father. (Figure 5.6) The "yes" answers from the beacon do not change but adding grandparents to the beacon may change some of the "no" answers of the beacon comparing to the case where only parent is in the beacon. So, adding grandparents to the beacon may result more decrease in power than adding only one parent.

Adding only one of the grandparents cause a power decrease less than adding mother/father, since degree of kinship is decrease. In other words, power decrease is reversely proportional with the distance between victim and his relatives added to the beacon as can be seen in 5.7.

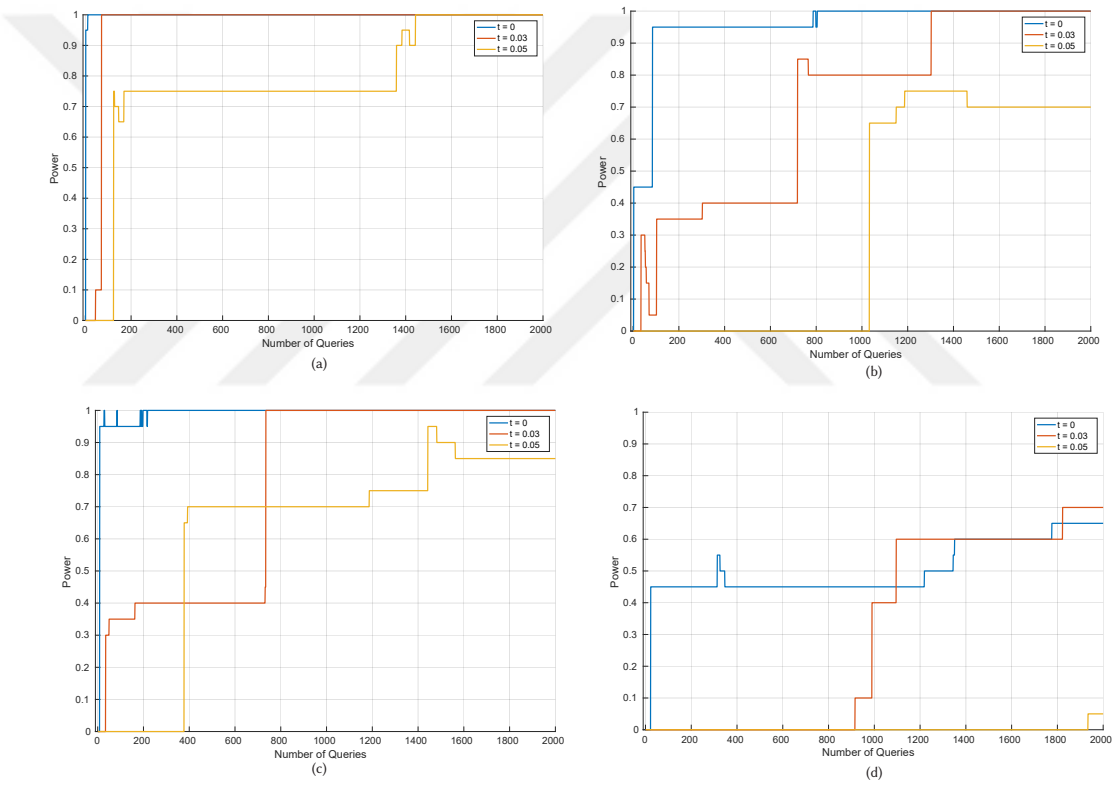


Figure 5.3: Power curves of the Optimal Attack on (a) old beacon, (b) the beacon that includes individual's mother, (c) the beacon that includes individual's father and (d) the beacon that includes individual's parents for different thresholds of $t$. The SNPs with MAF $< t$ is hidden.

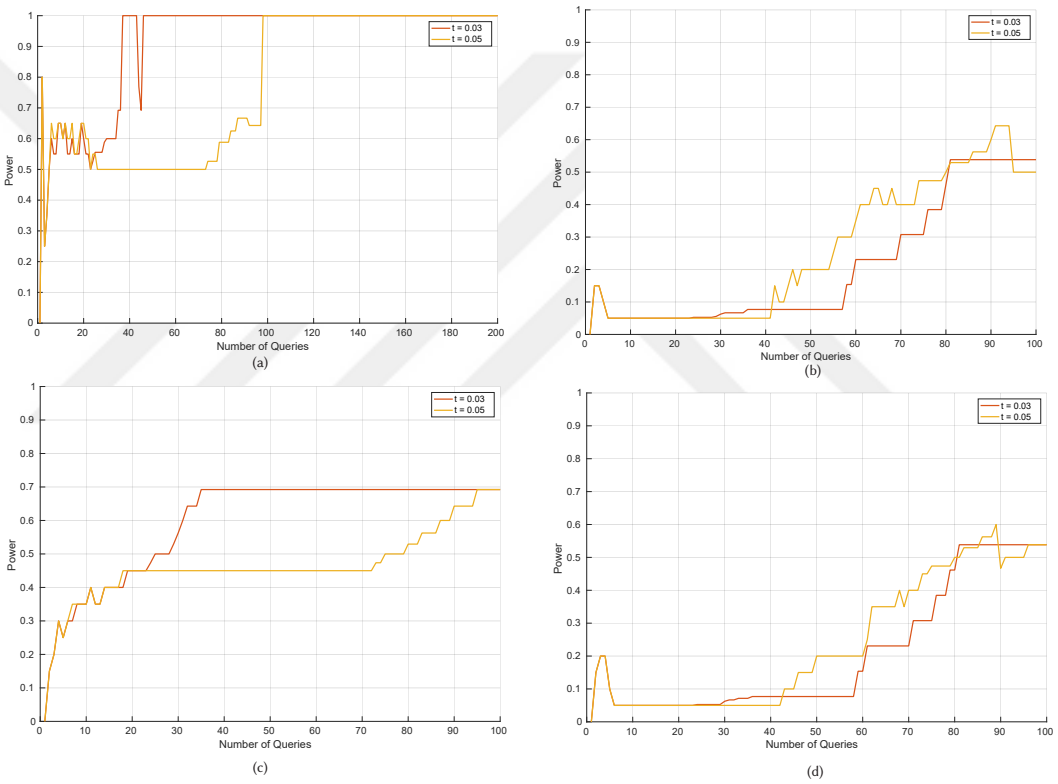


Figure 5.4: Power curves of the Genome Inference Attack on (a) old beacon, (b) the beacon that includes individual's mother, (c) the beacon that includes individual's father and (d) the beacon that includes individual's parents for different thresholds of $t$. $t = 0$ scenario is not applicable here since if none of the SNPs are hidden, the genome inference will not be necessary. The SNPs with MAF $< t$ is hidden.

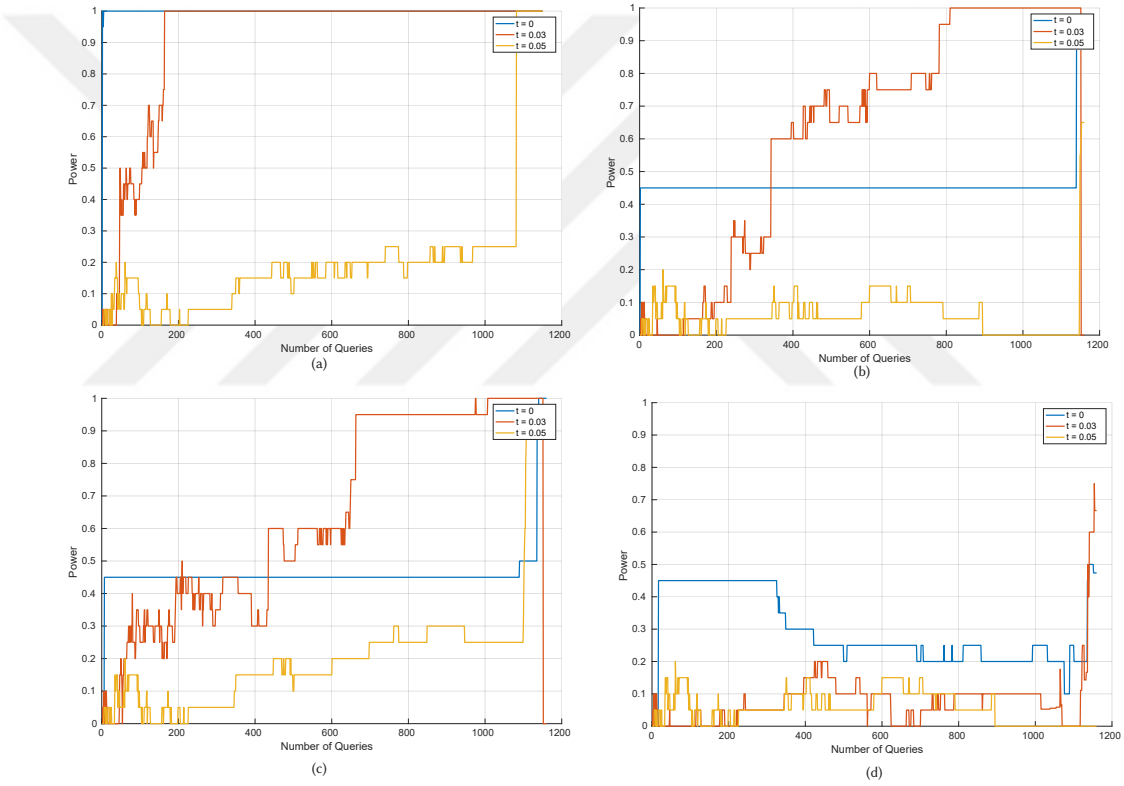


Figure 5.5: Power curves of the Query Inference Attack on (a) old beacon, (b) the beacon that includes individual's mother, (c) the beacon that includes individual's father and (d) the beacon that includes individual's parents for different thresholds of $t$. The SNPs with MAF $< t$ is hidden.

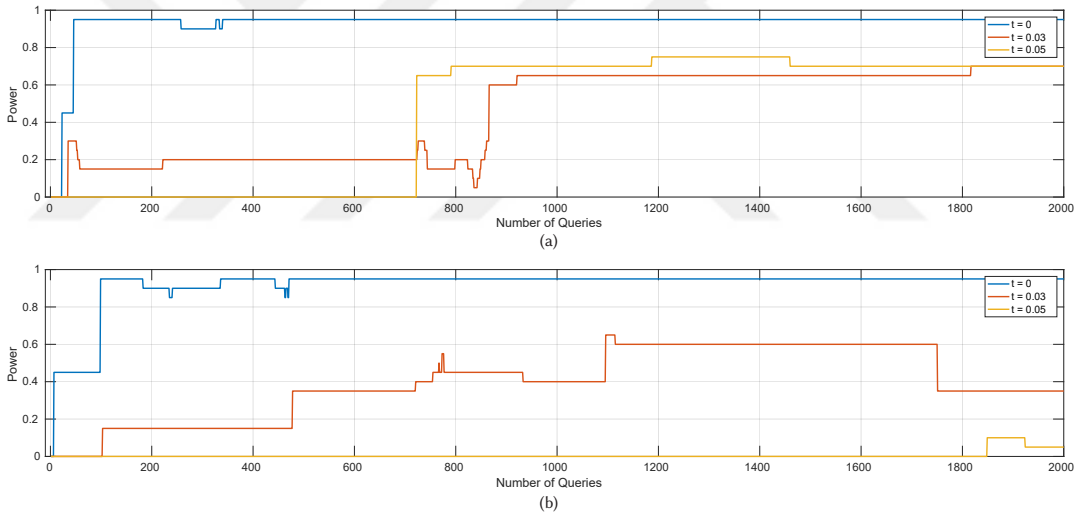


Figure 5.6: Power curves of the Optimal Attack on (a) the beacon that includes individual's father's parents and (b) the beacon that includes individual's father's parents.

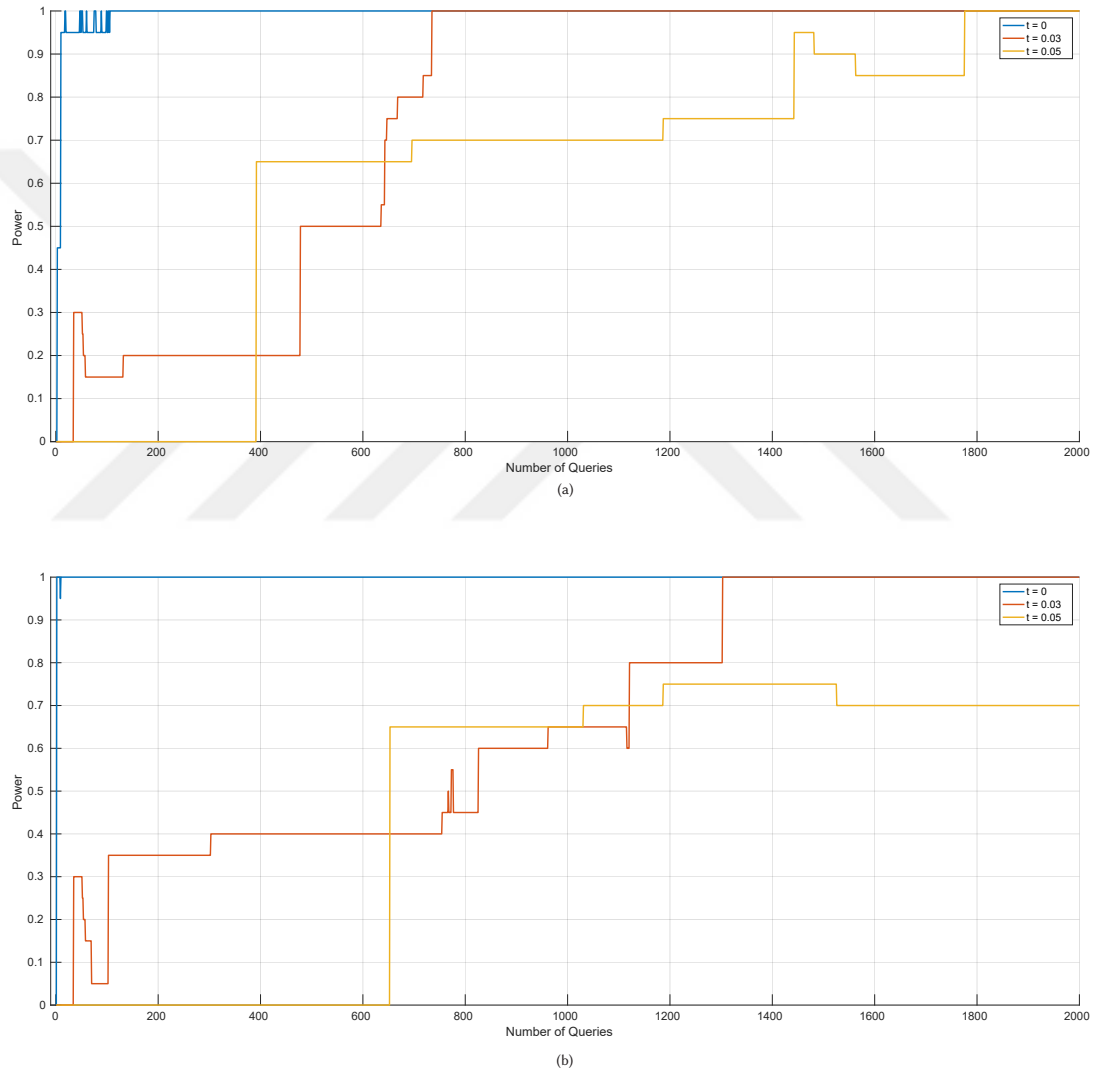


Figure 5.7: Power curves of the Optimal Attack on (a) the beacon that includes individual's mother's father and (b) the beacon that includes individual's mother's mother.

# Chapter 6

# Utility Analysis of the Proposed Solution

We showed that adding relatives to the beacon increase the number of queries needed to confirm beacon membership. Additionally, we also showed that our method cause decrease in the power of attacks. To consider our approach as a counter-measure we investigate the results of adding new people to the beacon. There are four scenarios that we consider:

- If the beacon response is "yes" for a query, after addition of any individual(s) does not change the answer of beacon since the already existing members of the beacon always give "yes" as response for particular query.
- If the beacon response is "no" for a query and adding a person does not change this response, it means that the new member of beacon does not contain a different variation of the queried allele.
- If the beacon response is "no" for a query and adding a person change that beacon response to "yes", it means that the added person has a different variation of the queried allele compared to other beacon members.

The last scenario is the one that should give importance. Because a researcher

that query a particular beacon before can notice the changing response of the beacon for that specific query. As a result of this we define utility as the number of beacon responses that changes from "no" to "yes", meaning a genome that contains a different allele variation added to the beacon.

| Average (%) | Mother in Beacon | Father in Beacon | Both Parents in Beacon |
|---|---|---|---|
| Control Group | 0.71 | 1.35 | 1.77 |
| Case Group | 0.40 | 1.08 | 1.48 |

Table 6.1: The average percentage change of utility in the beacons. The maximum utility change observed when parents are added to the beacon, since parents genome are independent from each other and can bring different variations.

The SNPs with a lower MAF should effect the privacy more since it shows a less frequently observed characteristic in the population. As a result of this, we also investigate the utility by taking into account the MAF values of the SNPs. We calculate the changing beacon responses for SNPs that a lower MAF than 0.2. The same calculation is repeated for MAF values between 0.2 to 0.01 with a 0.01 drop rate. As can be seen in Figure 6.1 while MAF values decrease, the utility increase.

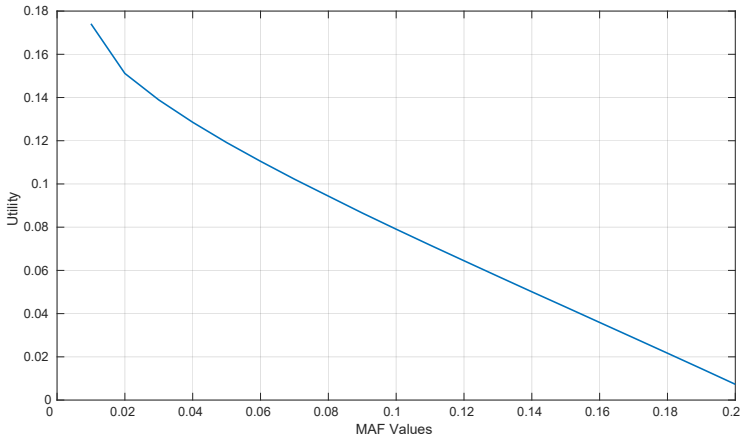


Figure 6.1: MAF values changes starting from 0.01 to by increasing 0.01 each time. MAF values and utility are inversely proportional.

# Chapter 7

# Conclusion

In this thesis, we explain the vulnerability of beacons to re-identification attacks. We mentioned the reasons of why the existing counter measures are insufficient, they either proved not to be effective against attacks or most importantly they decrease a significant decrease of beacon's utility. We propose a method as a counter measure for re-identification attacks in genomic data sharing beacons by using kinship. We showed that adding family members to the beacon cause a significant decrease in the power of these attacks since the same SNPs shared throughout the relatives. We argued that attacker cannot be sure whether the "yes" answers from beacon to queries are provided by the victim in the beacon or the victim's relatives. Besides, our work does not cost a substantial utility decrease in the beacon. Hence, adding relatives to the beacon prevents the privacy risk of people while increasing the data that is available for research in genomics.

# Bibliography


[1] P. T. Kim, "Genetic discrimination, genetic privacy: rethinking employee protections for a brave new workplace," *Nw. UL Rev.*, vol. 96, p. 1497, 2001.

[2] P. R. Billings, M. A. Kohn, M. De Cuevas, J. Beckwith, J. S. Alper, and M. R. Natowicz, "Discrimination as a consequence of genetic testing.," *American journal of human genetics*, vol. 50, no. 3, p. 476, 1992.

[3] E. V. Lapham, C. Kozma, and J. O. Weiss, "Genetic discrimination: perspectives of consumers," *Science*, vol. 274, no. 5287, pp. 621–624, 1996.

[4] 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 & communications security*, pp. 1141–1152, ACM, 2013.

[5] 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.

[6] 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.

[7] J. L. Raisaro, F. Tramèr, Z. Ji, D. Bu, Y. Zhao, K. Carey, D. Lloyd, H. Sofia, D. Baker, P. Flicek, *et al.*, "Addressing beacon re-identification attacks: quantification and mitigation of privacy risks," *Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 799–805, 2017.

[8] 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.

[9] 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.

[10] M. T. Goodrich, "The mastermind attack on genomic data," in *2009 30th IEEE Symposium on Security and Privacy*, pp. 204–218, IEEE, 2009.

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

[12] B. A. Malin, "An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future," *Journal of the American Medical Informatics Association*, vol. 12, no. 1, pp. 28–34, 2005.

[13] 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 Workshop on privacy in the electronic society*, pp. 95–106, ACM, 2013.

[14] Z. Wan, Y. Vorobeychik, M. Kantarcioglu, and B. Malin, "Controlling the signal: Practical privacy protection of genomic data sharing through beacon services," *BMC medical genomics*, vol. 10, no. 2, p. 39, 2017.

[15] 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, p. 965, 2009.

[16] M. M. A. Aziz, M. N. Sadat, D. Alhadidi, S. Wang, X. Jiang, C. L. Brown, and N. Mohammed, "Privacy-preserving techniques of genomic data—a survey," *Briefings in bioinformatics*, 2017.

[17] 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.