UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Analysis and improvement of security and privacy techniques for genomic information

Master Thesis

José Antonio García
Master's degree in Cybersecurity

19/10/2021

Director

Jaime Delgado
(Computer Architecture Department)

UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC)
ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA DE TELECOMUNICACIÓ DE BARCELONA (ETSETB)
FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

ANALYSIS AND IMPROVEMENT OF SECURITY AND PRIVACY TECHNIQUES FOR GENOMIC INFORMATION

José Antonio García
Director: Jaime Delgado

Master Thesis in Cybersecurity
October 2021

UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC)
ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA DE TELECOMUNICACIÓ DE BARCELONA (ETSETB)
FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

Abstract

In recent years we have witnessed a major progress in the genomic field. Genomic sequencing and understanding is now easier than ever, and genomic research has become useful for a very large span of applications, especially health related. But our genome is very complex, and it is still complicated to link the DNA code to observable traits, such as eye color or the presence of a disease. Genomic studies try to relate specific patterns or similarities in the DNA to those traits, but to be able to perform them a lot of genomic data is required to be collected and shared to researchers. Due to the singular properties of genomic data, and how sensitive and private it is to everyone, the process of sharing this information to legit researchers is a current security problem of our times, widely discussed by the experts in the last few years.

The first purpose of this work is to perform a systematic review of the current literature of all those privacy measures for genomic data currently present. We aim to perform and develop an analysis of the current situation of the most present techniques nowadays. Reviewing the state of those techniques will help us to properly understand and categorize them, allowing us to propose a classification scheme to add to our contribution. This scheme, based on the performance of each of those techniques, is intended to be a long-term classification system.

The Beacon system, from the GA4GH (Global Alliance for Genomics and Health), is a remarkable technique among the reviewed. It acts as the interface the researchers use to access stored genomic data but, instead of handling the data in any way, it only answers queries from them. Researchers, with the use of queries, ask the Beacon if some particular mutation is found on a genomic dataset and, the Beacon, in response, will answer a "Yes" or a "No". Avoiding the disclosure of the donor's genomic data makes the Beacon a very extraordinary system to allow genomic studies and protect the data of our donors but it is not a perfect environment. As geneticists Shringarpure and Bustamante demonstrated, attackers in possession of the genome of a victim could, through the use of queries, find out if the victim is present or not on the dataset.

Apparently innocuous, this re-identification attack could lead attackers to infer some traits from the victim. Discovering the presence of an individual in a dataset of cancer study could mean this individual actually suffers from cancer, for example. And while a lot of different measures have been contemplated to make the Beacon a more secure environment, it is essential to ensure proper genomic research can still be performed.

The second objective of this thesis is to develop a security improvement proposal for the Beacon system, ensuring it respects the required trade-off between privacy and utility. Our contribution aims to be a list of measures that can work conjunctly in order to improve the privacy of the genomic data of the donors, while still maintaining the utility for potential research.

List of Acronyms

AAI: Authorization and Authentication Infrastructure

DP: DIfferential Privacy

DUO: Data Use Ontology

GA4GH: Global Alliance for Genomics & Health

HE: Homomorphic Encryption

MPC: Multi-secure Party Computation

SECRAM: Selective retrieval on Encrypted and Compressed Reference-oriented Alignment Map

SNP: Single Nucleotide Polymorphism

XACML: eXtensible Access Control Markup Language

Figures and Tables

Contents

Chapter 1
Introduction

1.1 Background

1.1.1 Current State

Since the human genome was discovered the knowledge we have over it has never stopped growing. This is especially true within the last few years, when major advances have been made in practically all its regards, including reading, understanding or storage. Sequencing systems to read and print genome data are more and more cheaper every day, meaning the overall quantity of genomic data collected around the world has increased exponentially in recent times.

Keeping up with this, and partially as a consequence, the knowledge in understanding genome's implications have grown at par. There are now multiple traits, better known as phenotypes, that have been linked to certain sequences of the genome that might differ from the most common forms, which are usually known as mutations. When the chain of nucleotides that form our DNA, composed by a combination of Adenines, Citosines, Thymines and Guanines, has an uncommon variation of its parts, we call this particular section a mutation. When a deviation from the universal standard genetic code is fairly common, we call it a polymorphism. Both mutations and polymorphisms are, essentially, variations of the universally accepted most ordinary DNA sequencing. If genes are sections of our DNA, composed by the concatenation of the nucleotides, alleles are what the variations of those genes, produced by either mutations or polymorphisms, are called. Recently, more and more lines can be drawn between alleles and phenotypes, being a phenotype any possible trait, like the color of the eyes and hair, some temperament characteristic, or even being more prone to cancer development.

This potential inference of what clearly should be private and intimate information has raised obvious concerns about privacy. Even when being able to infer the color of someone's eyes, by lecturing their genome, can appear to be harmless and safe, the disclosure of that genome can also reveal a person is more prone to be addicted to any kind of substance, which surely can lead to discrimination in a lot of different environments. In a similar thought process, if medical history is accepted as private information in order to avoid any

type of discrimination, a genome that reveals being prone to develop brain tumors should also be treated in the same way, as it may not be factual medical history but potential one.

While giant steps have been made towards the understanding of the genome, it is more clear than ever that its total implications are yet so far to be completely unfolded. From health to investigation, the applications this knowledge has are being expanded more and more, and further researchment is still required now for both scientific knowledge and privacy consequences.
Typically, genomic researchment would be done by providing the genomic data of different people to the researchers which, if they would be studying a particular phenotype, they would know which of those genomes would expose the phenotypic trait. By a rigorous examination, they would ideally be able to tell if a specific mutation or allele is correlated to the evaluated phenotype. But genomic data of a person carries an immeasurable load of sensitive data that will probably be not related to the research. And even assuming strict professionality and only good intentions from the researchers, there is still the possibility the data can be accessed by another third person. An attacker could steal the data from a valid researcher, or even impersonate one and access the data directly. For this reason, even for trusting parties, it is advisable to restrict the level of genomic information shared to the least compromising possible.
But regarding the sensitive nature of the genome, it is hard to ensure no compromising data is being given. If we disclose too much information, privacy can be seriously affected to the owner of the genome, and if we restrict too much of the information shared, we might be truncating the researcher's possibilities to make any advancements, which is the primary goal of genomic studies. A line must be drawn to ensure the most possible is done for the two sides, this sort of equilibrium is often referred as the trade-off between privacy and utility, and it means finding the correct balance between protecting too much or disclosing too much genomic information.

1.1.2 Genomic data

In order to correctly ensure a good trade-off between privacy and utility, first it is important to understand the privacy dangers related to the genomic data. As Daniel Naro and Jaime Delgado explained in their thesis called Critical analysis and comparison of data protection techniques for genomics data sets [A], DNA data holds some very special features that make them differ from regular sensitive data. In the thesis, the authors comment

how the authors of *Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider* [2] establish those special features as:

- *Uniqueness:* one of the most special features of DNA is how unique it is. There are certainly not two different individuals that share the genomic information. Blood relatives, or even identical twins, will share a big part of the DNA but in any case two of them are exactly the same. This holds the power to be able to always identify a person.

- *Predictive Capability:* some DNA variations have been proven to be linked with being more prone to having a certain disease. Typical studies will try to discover these sequences, thanks to that, the risks of having a disease can be predicted long before it shows any symptoms.

- *Immutability:* our DNA code is immutable over time, meaning it will never change through our lives. The most immediate concern about this is that it would be irresponsible to not consider it very sensitive data just because we do not know how to properly read it yet. If an attacker were able to catch someone's genome 10 years ago, he could wait for another 10 years when there might be easier ways to read it, and the victim's genome would have remained equal for all of that time. This property of never losing relevance over time is what makes it paramount to correctly ensure its privacy even when a lot of the possible inferred conclusions are still unknown.

- *Requirement of testing:* genetic diseases, explain the authors, frequently need genetical tests aside from the usual clinical tests.

- *Historical misuse:* with this the authors mean to maximize caution measures, as there have already been examples in the past of people who tried to use DNA knowledge for eugenics and discriminatory purposes.

- *Impact on family:* DNA variations are widely shared among blood relatives, which is the reason for hereditary traits within a family, for example physical likelihoods. Using those similarities, it is possible not only to assume two different individuals are relative but even to be able to potentially infer some conditions of people through reading their relative's DNA.

- *Evolving perception:* the authors consider that, as it has been during the recent years, it is possible that the perception of DNA evolves during the following times, meaning, for example, to reclassify some traits or diseases at discovering new DNA characteristics.

- *Ubiquity:* as our whole body is a result of our DNA, this is present in a lot of different parts of us from where it can be extracted (as blood, hair or saliva).

1.1.3 Security in Genomics

For the reasons aforementioned, security in genomics has been a widely accepted issue of fundamental importance to address. Many cryptographic systems have proven to be a consistent option to keep data secured, but genomics need to be able to perform some special actions. Besides regular communication, researchers will need to apply some operations into the data in order to find resemblances or discrepancies. With this in mind, different approaches have appeared offering very different solutions. Homomorphic Encryption, for example, is an encryption system that allows researchers to perform their operations into the encrypted data directly. Decrypting this data will produce the same exact result as if the operations were performed to clear data. Differential Privacy, another different technique, relies on the fact of adding a minimal noise in order to make genomes more indistinguishable and, thus, making it difficult to identify a particular genome's owner.

Another system, which has the major weight on this thesis, is the Global Alliance for Genomics Health (GA4GH) Beacon System. This tool acts as a mediator between the researchers and the proper data owners. The main difference with regular security systems is that Beacon never provides the researchers the actual genome of the participants, rather, they simply ask questions about the data they protect, which makes them able to perform their research. Not giving the actual data to anybody will surely ensure our genomic data will not end in malicious hands. In further chapters, at the Systematic Review, we also see other privacy breaches still exist. This technology is the one of our choice to develop, in the chapter 5 Security Improvement Proposal for the Beacon System, a proposal of privacy enhancement to the system.

1.2 Further Background

The previous introduction, even brief, means to englobe all the genomic concepts needed to understand this thesis. This project is concerned about the security of genomics and is focused only on security matters. However, we include a selection of literature that serves as a more extensive lecture on the biology of the genomic, which we insist is not required to understand this thesis but could be used to expand the knowledge of the related terms.

- *Security strategies in genomic files, Daniel Naro et al. (2020), from Universitat Politècnica de Catalunya* [3].

The first chapters of the Phd thesis from Daniel Naro includes a good base of concepts that dive deeper in what we have seen here, but are still practical for the security context.

- *The Sequence of the Human Genome, J. Craig Venter et al. (2001), from SCIENCE* [4].

Done by a very large conjunction of authors, this work is focused in describing the particularities of sequencing the genome in order to retrieve it.

- *Human Molecular Genetics, 4th Edition, Tom Strachan and Andrew Read (2011), from NHBS* [5].

A very extensive lecture about genomics by two authors who were recipients of the European Society of Human Genetics Education Award in 2007. It is aimed for any kind of student of human genetics and contains a lot of genomic knowledge, far beyond the security aspect.

1.3 Objectives

As it has been explained, this is a relevant topic and current issue of our times. Hence, in order to try to contribute in some manner to the advancement of the field we set two different objectives for this thesis.

The proper treatment of genomic data security has today the importance it deserves and, in order to contribute to the field, our first objective is to perform an investigation of the state of the art. From this, we select a technology we think we can improve its security. This divides our work, therefore, into two objectives with a common purpose.

1.3.1 Systematic Review and Classification Proposal of current privacy Genomic Techniques

The first of our objectives is to perform a Systematic Review of the literature of the currently used techniques to achieve genomic privacy. Each section of this review is dedicated to a single privacy technique exclusively, and includes a collection of selected works that help us to not only define the technique by itself but, in some of the cases (Beacon essentially), also provides a timeline of its evolution in recent years.

A successful Systematic Review is crucial to perform the second part of this objective, a Classification Proposal of the current privacy genomic techniques. Based on how those techniques approach the privacy issue in the genomic field, this classification aims to illustrate all the possible paths that can be taken. Ideally, any future technique will be classifiable in some part of its structure (if it is not some sort of hybrid technique that would use two or more of the features used to classify).
The complexity of this part does not only rely on being able to accurately understand and define each of those techniques, but also being able to discern the similarities and disparities from one to each other.

If done correctly, this can become practical to help people understand what different types of techniques are used, which ones are the most relevant at this moment, how does current literature uses them, and how do they approach security. The final schema could also be used as a long-term classification system, where future techniques that are still unknown and were not considered could be added as a part of it.

The development of these objectives is in chapter 2 Systematic Review of the Literature of Current Genomic Privacy-preserving Techniques and chapter 3 Brief Classification of Privacy-preserving Techniques for Genomic Information.

1.3.2 GA4GH Beacon's Improvement Proposal

The second objective is to develop an improvement proposal for one of the techniques that have been reviewed. We have chosen the Global Alliance 4 Genomic Health Beacon System to be this technique, as we see, through the Systematic Review, that it is a very recent standard that has had constant adjustments since then. We believe the Beacon System is one

of the most promising solutions to genomic privacy, if not the most, and we try to make a thoughtful proposal with different measures that we think could lead to improved privacy. The ideal final form of this idea is not a conjunction of different suggestions that could only work separately but to use a combination of propositions that help each other, aiming for a shared purpose, discarding in the process any security measures that could not work conjunctively.
If the final proposal meets our expectations, it is safe to say it will be a significant contribution in the Beacon environment. Submitting a list of possible adjustments to improve the Beacon system will not only enrich the environment but prove that, even where it actually is a valid and viable option to secure genomic data, there is still room for improvement. In any case, finding possible solutions that improve security and accommodate a fair trade-off between privacy and utility at the same time has proven to be a challenging task to do.

The chapter 5 Security Improvement Proposal for the Beacon System develops our contribution to this purpose.

1.4 Structure

This thesis presents a basic structure formed by the following sections:

1. Introduction: this precise introduction this section is a part of. It has the purpose to provide basic background and explain the importance of our objectives.
2. Systematic Review of the Literature of Current Genomic Privacy-preserving Techniques: the systematic review of the current literature, where the tools are analyzed by using scientific papers.
3. Brief Classification of Privacy-preserving Techniques for Genomic Information: a proposal of a scheme to classify all the reviewed measures by their security approaches.
4. Beacon Evolution: some background of Beacon's history through its standardization and evolution. This chapter does also contain the added features of the second version of the Beacon, the Beacon v2.
5. Security Improvement Proposal for the Beacon System: a conjunction of different measures we think have the power to improve security in the Beacon system that could work as a united proposal, where none of them interfere with each other.
6. Conclusions: a dedicated chapter for the conclusions this thesis has led us to.

7. Future Work: we hope to be able to provide some guidelines that will mark what we think the nexts steps should be aimed at.
8. Glossary: with the idea of assisting the lecture of this thesis, a glossary defining its more specific terms is attached.
9. References: all the references used in the production of this thesis are finally included here.

Chapter 2
Systematic review of the Literature of Cybersecurity in Genomics

In order to correctly analyze the state of cybersecurity in genomics, a correct classification of the techniques is mandatory. Before that, however, a previous evaluation of each of those techniques serves not only as a strong guideline for further chapters but also as reliable analysis on its own. The strategy of our choice, which we are going to use to pursue this purpose, is a systematic review.

Systematic reviews are a schematic type of review that are generally based on study of the current literature, with the idea of answering a predefined question, either narrow or wide, with the intention of assessing and helping in an unbiased decision-making process.

To elaborate this systematic review, we will not use any established methodology. Rather, we will adapt the concept to perform an *ad hoc* review of the literature that fits our objectives. For this reason, short guidelines summarizing our methodology are included, helping to understand our approach.

2.1 Defining research question

Correctly specifying the research questions of a systematic review is a crucial step in defining the objectives of the analysis. In our case, we do not assume any knowledge of these techniques by the reader's part and, as we practically have not talked about them yet, not only a small introduction precedes each of those techniques but also our first question to answer is "How do the authors define this particular technique?". Conjunctly, a second question is also done, this time focused on the contributions made by the particular study. This question can be written as "What does this study do in order to improve this technique?", helping us to determine how that technique has been evolving for the last few years. While this second question is not exclusive to the Beacon part, its relevancy is notoriously empowered for the studies related to the Beacon, as it serves as stepping stone for the future sections of this work dedicated to the Beacon system (chapter 4 Beacon Evolution and chapter 5 Security Improvement Proposal for the Beacon System).

2.2 Searching for relevant data sources

From the moment a research question is defined, our search strategy is to look for relevant papers on the topic of each technique, including also some other works dedicated to make some sort of classification of the techniques, which helps us to answer the first of the questions. This said relevancy is judged primarily upon two things, its publication site and its publication year. We prioritize, whenever it is possible, using studies from renowned and reliable institutions as well as trying to select the most recent papers possible. If a study feels with a small dedication on defining the technique, we follow their references to reach a root study that we feel correctly defines the technique.

The main reason we prioritize recent papers, aside from being more suitable for an up to date analysis, is that literature review of more antique documents has already been done. In [1], by Daniel Naro and Jaime Delgado, in 2016, the authors provided a similar review of the literature for the current genomic privacy preserving techniques. It is also our goal to focus on the changes some of the reviewed techniques have been since the release date of the document, including some other techniques that were not originally present.

2.3 Extraction of relevant data

When the studies have been selected, the extraction of the data is done by rigorous examination of what contributions the studies make in answering each of the questions, especially the first one if it is the first study presented. For more particular cases as it is the Beacon part, we increase the focus on the reading into trying to answer the second of the questions, with the intention to correlate the evolution of the system from one study to another.

2.4 Literature Review

This section presents the information we consider relevant from the previously commented works. Technologies are reviewed one by one by selecting the information found on the different papers, where recent texts have been prioritized for a more accurate analysis of the state-of-the-art techniques. These documents often rely and point to older literature.

2.4.1 Homomorphic Encryption (HE)

Homomorphic Encryption (HE) is an encryption scheme that has the advantage of allowing users to perform calculations on its encrypted state and, upon decrypting the result, the obtained would be equal as if the calculations were done to the initial unencrypted data. A fully homomorphic encryption system was first described back in 1978, but it was not after 30 years that Fully Homomorphic Encryption was proved to exist. While being an encryption technique on its own, completely unrelated to the medical world, HE has been frequently linked to health care research schemes due to its capacities. The next papers will talk about that, defining HE as a technique itself and describing its advantages and difficulties in the genomic field.

[6] *Privacy-preserving techniques of genomic data—a survey, Md Momin Al Aziz et al. (2019), published in Briefings in Bioinformatics.*

This paper defines HE as a technique to allow one party to compute and perform predefined operations over encrypted data without the need to decrypt it. It classifies HE as Fully, Leveled and Somewhat HE depending mostly on compactness, correctness or the functions they can compute, at the time of the article different successful approaches have been made to use this technology on the genomic field.

It also points to the fact that HE solutions are still in their first steps, and it has yet to be at the level of efficiency required for a realistic large-scale generic genomic data computation. «For example, a multiplication or bootstrapping operation takes some time making it unrealistic to use on complex functions such as training machine learning» (Md Momin Al Aziz et al. 2019, p.891), they comment. The paper clearly states that real-world scenarios demand machine learning proof systems, which would require much faster HE schemes.

[7] *Private genome analysis through homomorphic encryption, Miran Kim et al. (2015), published in BMC Medical Informatics and Decision Making.*

An early approach of this technology for genomic purposes, the authors showed that full-scale privacy-preserving GWAS is practical, as long as the statistics can be computed by low degree polynomials. Two different models were examined both theoretically and practically, concluding both as efficient approaches, highlighting the trade-off between security and performance, where empowering one side could hinder the other.

This paper presents the computation of minor allele frequencies, and the $\chi 2$ statistic with the use of the homomorphic BGV and YASHE encryption schemes. They use a specific encoding technique to improve on the work of Kristin Lauter et al. at Progress in Cryptology, published in 2015 [8].

[9] *Towards practical privacy-preserving genome-wide association study, Charlotte Bonte et al. (2018), published in BMC Bioinformatics.*

The authors define HE as «a set of cryptographic tools that allow certain computations to take place in the encrypted domain, while the resulting ciphertext, when decrypted, is the expected (correct) result of operations on the plaintext data» (Charlotte Bonte et al,. 2018, p. 3). This paper presents a new approach of HE. They claim that as the work on [7] and [8] only compute the allele counts homomorphically, leaving the other operations to be performed upon decrypted clear data, the solutions are not resistant to the Nils Homer et al. [10] attack.

The $\chi 2$ computations will, according to the HE schemes, be performed on an encrypted state, but it is only revealed whether or not the $\chi 2$ is relevant for the case. This behaviour will invalidate the previously commented type of attacks and, therefore, the authors consider it as an improved version of already existing solutions. They conclude the work presenting their solution as efficient and reliable, even in large and down to grow groups, however, they find their HE solution outperformed by the MPC solution they are also proposing, which will be commented on in the MPC section of this Systematic Review.

[11] *Privacy challenges and research opportunities for genomic data sharing, Bonomi L. et al. (2020), published in Nature Genetics.*

In posterior study about the challenges of sharing genomic information the authors dedicated a part to a technique classification. There, they describe HE procedure as after the data has been encrypted into a ciphertext, some simple operations (such as addition or multiplication) can be performed on it and produce the same results we would obtain by using the original unencrypted data. The study points at the fact as how this is a solution generally involved in systems where the data is shared in the cloud and in federated environments.

[12] *Ultra-Fast Homomorphic Encryption Models enable Secure Outsourcing of Genotype Imputation, Miran Kim et al. (2021), published in Cell Systems.*

Lastly, this recent paper presents the first fully secure genotype imputation by utilizing ultra-fast homomorphic encryption techniques which enables the evaluations of millions of imputations in a few seconds. Among the arguments they provide to support their implementation they say highlight the HE characteristic feature of allowing the genotype data to be encrypted from end to end. That means the data is encrypted while in transit, at rest, and most importantly, in analysis. Compared to other current techniques, they argue «HE-based methods provide full genetic data security with comparable or slightly lower accuracy. In addition, HE-based methods have time and memory requirements that are comparable and even lower than the non-secure methods» (Miran Kim et al., 2021, p.1*)*. Along the paper, they provide five different HE-based implementations, based on the schemes developed by the iDASH19 Genome Privacy Challenge contestants. Finally, they comment «our results provide strong evidence that HE-based methods can practically perform resource-intensive computations for high throughput genetic data analysis» (Miran Kim et al., 2020, p.1*)*.

2.4.2 Multi-secure Party Computation (MPC)

Multi-secure Party Computation (MPC) is another type of encryption scheme which, in this case, is aimed to protect privacy among the participants. The result obtained would be a function computed by all of the inputs while keeping those inputs private. MPC is neither related to health care by itself but has fitting advantages for the genomic research situation, which will be discussed in the next papers.

[9] *Towards practical privacy-preserving genome-wide association study, Charlotte Bonte et al. (2018), published in BMC Bioinformatics.*

This paper, which has been already commented on by its Homomorphic Encryption proposal, presents also a Multi-secure Party Computation solution. In their words: «Secure multiparty computation aims at allowing a similar functionality, amongst several mutually distrusting parties, who wish to compute a function without revealing their private inputs» (Charlotte Bonte et al,. 2018, p.3). This approach necessarily means the communication between the computing parties.

The proposal the researchers do here is based on a previous setting done by Kamm et al. [13], both cases were based on large data collections from different genetic repositories. In this previous setting a lot of sensible information was required to be added to the database (i.e. all raw genotype, phenotype and clinical data). In [9], the authors say: «To the contrary, our setting assumes that only the aggregate values, necessary to identify the significance of a gene-disease relationship (i.e., the contingency tables recording the counts of genotypes vs. phenotypes), are contributed by each biobank. This is a simpler, and more realistic setting, which not only is likely to be implemented in the near future, but also alleviates the computational cost of the proposed solutions» (Charlotte Bonte et al,. 2018, p.4).
In addition to this, unlike [13] and other previous alternatives, [9] solution achieves security with a dishonest majority, which means that the protocol tolerates some dishonest behaviour by the majority of the involved computing parties without losing any privacy or correctness. [13] and alternatives just assume the computing parties (i.e. the genetic repositories) can not be dishonest or corrupted in any way, which can be seen as a strong assumption.

[14] *Recent Advances in Practical Secure Multi-Party Computation, Satsuya Ohata (2020), published in IEICE TRANSACTIONS.*

This more recent paper from October of 2020 tries to show and discuss the latest situation of this technique. Similarly described as our previous paper, here the authors define MPC as to «allows a set of parties to compute a function jointly while keeping their inputs private» (Satsuya Ohata, 2020, p.1134). The recent advances commented are the higher-level secure protocols, the privacy-preserving data analysis and its low-cost implementation. They conclude the text talking about how the integration with other fields is progressing, how MPC is more needed of specific algorithms and how many organizations are doing efforts for improving the usability (e.g. MPC compilers). On a last note, the authors consider the performance of MPC will be improved by accelerating those aspects.

2.4.3 Differential Privacy (DP)

Differential Privacy (DP) is another cryptographic technique, in this case has the purpose to protect the privacy of all the participants of a dataset by, essentially, adding noise to the data. This noise would be, ideally, the minimum to ensure that no one can be identified through queries.

Different authors present DP as a valid option and, through our literature review, we will remain unbiased in either way. In further sections (specially on Beacon's Proposal) we will express our concerns relating DP and genomic field. Those concerns are mostly based on the fact that while in theory it does achieve privacy enough to prevent Homer's attack, it obviously affects the accuracy of the results, which some people have argued adding noise to the aggregation hinders too many potential results. We will later develop our opinion regarding DP but, even when it looks impractical or counterproductive to the genomic field, it is an efficient working technique which has its flaws and its advantages, and we truly believe there are other fields in which DP can prove to be a viable option.

The next papers present DP solutions and talk about its situation in the genomic field.

[15] *A community assessment of privacy preserving techniques for human genomes, Xiaoqian Jiang et al. (2014), published in BMC medical informatics and decision making.*

Aimed to answer the need for reliable protection on biomedical data, this paper discusses mainly what DP is and how we can perform it. They describe DP's basic fundamentals as adding the noise required to the allele frequencies in order to transform our mixture, in a way an attacker will not be able to tell the presence or absence of an individual. To achieve this they show the Laplacian Mechanism [16] and the exponential mechanism presented by McSherry [17] which differ in the maths that lie behind.

While this is a work focused only on the technical aspects of privacy, it clearly states that some new privacy-preserving techniques need to be developed as DP data perturbation techniques have limitations in sharing large volumes of human genomic data.

[18] *Privacy in the Genomic Era, Muhammad Naveed et al. (2015), published in ACM computing surveys.*

In this paper the authors explain DP as a «well-known technique for answering statistical queries in a privacy preserving manner (...). In simple words, if we compute a function on a database with and without a single individual and the answer in both cases is approximately the same, then we say that the function is differentially private» (Muhammad Naveed et al., et al., 2015, p.19). That means the answer does not change significantly if the genomic data of a particular donor is in the database or not, ensuring the privacy of this donor has not been compromised.

But when simulating a choice procedure of a technology for a GWAS it warns DP is highly flawed in its basis, as it makes the data more noisy, therefore modified data. This addresses the trade-off between utility and privacy where it is important to say the paper considers using DP to be detrimental, because biomedical researchers often demand the most accurate data possible.

[19] *Protecting genomic data privacy with probabilistic modeling, Simmons S. et al. (2019), published in Pacific Symposium on Biocomputing.*

This more recent paper is not about using a DP solution but another one that does not share its flaws. It is the comparative they do against DP where they comment this technique as «by adding noise to a released statistics, one is able to achieve a level of plausible deniability that a particular individual was in your dataset. This level of deniability is measured by a privacy parameter. The larger the parameter, the less plausible deniability is preserved» (Simmons et al., 2019, p.411). They express, however, the concern of the price DP carries, which they see as an alteration of the most interesting data. From this comparison they finally show how their model, based on Probabilistic Modeling, allows the disclosure of much more data, agreeing with the commonly extended opinion DP reduces too much the potential data sharing.

2.4.4 Anonymity (k-anonymity and pseudonymity)

Anonymizing the data, i.e., converting your data into another data where it is not possible to re-identify any of its integrants is a commonly used technique in this type of research but, as we have seen, it is important we keep the information in a way the studies don't lose accuracy. Two proposed ways to perform this on genomic research are k-anonymity and pseudonymity. K-anonymity property is defined as where each participant cannot be distinguished from at least k - 1 with the others. Pseudonymity is a different approach in which the identifiable fields of every participant are replaced with other ones, which act as pseudonyms. The next papers cover recent proposals for the genomic privacy problem.

[11] *Privacy challenges and research opportunities for genomic data sharing. Bonomi L. et al. (2020), published in Nature Genetics.*

One of the papers we used at the HE section does also define other techniques, as k-anonymity. Essentially identical to what we already commented, they say: «k-anonymity ensures

that, for each record, there are at least k-1 records with the same quasi-identifiers (e.g., Zipcode) and therefore any record is hidden in a group» (Bonomi L et al., 2020, p.7). This is accomplished by applying some transformations to the original data, which are based on the suppression and generalization of the identifiers, regardless of their type. An example of this could be to use the decade of birth instead of the year or, for commenting the one they use, a 3 digit representation of the Zipcode.

This method has not only been used with those quasi-identifiers values (that we can see on different works, as [20] and [21]) but also on a genomic level, performing this type of operations at the degree of the SNPs, as authors on [22] and [23] exposed.

[24] *Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis, Scheibner J et al. (2021), published in Journal of Medical Internet Research.*

This recent paper, which has a more legal and ethical aspect compared to others, exposes how a somewhat combination of HE and MSC, commonly referred as multiparty homomorphic encryption, fulfills legal requirements for medical data sharing under the European Union's GDPR (which is used as a benchmark for data protection). While this is a very fresh and new approach, it is still soon to accept it as a new stepstone rather than an interested party trying to sell an idea. In any case, among the procedures described, some data anonymization is required, in which the paper comments typical researcher's preferences in this regard with this words: «Researchers prefer to rely upon simple pseudonymization techniques (such as replacing direct identifiers with pseudonymous codes) combined with legal measures defining each party's responsibilities regarding data transfer, access, and use» (Scheibner J et al., 2021, p.3). This combination of more traditional methods of pseudonymization with governance strategies does meet the GDPR pseudonymization legal requirements, but it fails to meet its anonymization requirements.

Lastly, the authors comment that there are a lot of different techniques that allow pseudonymization, for example, encrypting data with a secret key implies the data can be decrypted for anyone possessing the key but not for any other party. For this key holder the data would be pseudonymised, but for any other party the data would be therefore anonymized.

The authors comment then how under the GDPR view, the separation between the pseudonymised data and the identifiers is analogous to the separation between the decryption keys and the encrypted data. And regarding pseudonymised data, any party that possesses a way to access the data (legally or not) will have access to personal data (which is the data that allows the identification of a person). In the same way, it is considered by the authors that any encrypted data is legally personal data for any party or entity with lawful ways to obtain the decryption keys (or lawfully in possession of them).

2.4.5 Hardware based techniques

Defined as Hardware based techniques, we reference the using of a secure hardware to perform secure computations. It is mainly represented by Intel Software Guard Extensions (SGX) and allows the definition of private regions memory of a device. This approach was also considered for the genomic system and the next papers talk about how this works and how the technique has evolved over the last few years.

[6] *Privacy-preserving techniques of genomic data—a survey, Md Momin Al Aziz et al. (2019), published in Briefings in Bioinformatics.*

This paper, already commented on the HE section, talks about other cryptographic techniques, being Hardware based techniques one of them. In their words, «using secure hardware for a secure computation is considered a seminal contribution from Intel when they introduced SGX (...). It allows a user to separate their confidential data and code from the regular ones and allows him/her to do the secure computation in a secure enclave inside the processor» (Md Momin Al Aziz et al., 2019, p.891).

At the history line they draw for this technology it stands out as the first work using secure hardware on genomic data, by Canim et al. [24], where they leveraged a trusted hardware inside an untrusted cloud to ensure privacy back in 2012. This proposal relied on the use of symmetrical cryptography to achieve security on the count queries needed on the genomic data set.

A more recent attempt in 2016, from the work of Chen et al. [25], (and after the arrival of Intel SGX), introduced a new solution based on SGX named PREMIX, which guaranteed privacy preservation along with efficiency.

A different solution also reviewed by the authors is PRINCESS, which, in their words, «introduced an international collaboration framework (Federated) for privacy-preserving analysis of rare disease genetic data that is distributed around the world» (Md Momin Al Aziz et al., 2017, p.892). They evaluated this system in a study of the genetic architecture of the Kawasaki disease.

The document points out that hardware solutions such as Intel SGX, or AMD memory encryption were still unknown and could potentially solve a lot of problems genomic privacy has. On a final note, they comment that techniques such as SGX come with limited low-level memory attached only to the processor.

2.4.6 Beacon

A Beacon is a platform developed by the Global Alliance for the Genomics and Health (GA4GH) which serves as a repository for the genomic information in a way researchers can access its data by the use of questions such as "Is there at least one genome with a mutation on this position?", where the Beacon would simply answer "Yes" or "No". The whole concept of the Beacon resides in avoiding sharing the genomic information directly to whoever is asking for them. Instead the answer to these questions, called queries, has to be enough for the researchers, therefore avoiding the privacy problem of owning someone else's genome, discussed at the introduction.

We will see, however, that the biggest found vulnerability assumes the attackers already own the victim's genome and, using Beacon's datasets linked to disease, they could potentially deduct if this victim suffers from this disease. The following papers address this problem and its potential solutions mainly.

The Beacon project is actually developing a second version of it which will be discussed later, this second version would allow a lot of more different possibilities for the researchers to ask for the information witholded but, in exchange, could present some other vulnerabilities too.

[26] *Privacy Risks from Genomic Data-Sharing Beacons, Shringarpure, S. S., & Bustamante, C. D. (2015), published in American Journal from Human Genetics.*

This famous paper by Shringarpure and Bustamante exposed a vulnerability of Beacons of that time, but before getting deeper on it, they described GA4GH Beacons as

«web servers which can be asked allele-presence queries» (Shringarpure, S. S., & Bustamante, C. D., 2015, p. 637). Those queries can be very different from each other but share in common the fact that they return a simple answer that is not an actual genomic information but a "Yes" or a "No", which is done in order to avoid re-identification. But they pointed out that re-identification can still be done, even if all information the Beacon gives is the presence or absence of alleles. Through their document they presented a specific way to counter Beacons defenses by using just queries. A likelihood-ratio test was proposed to find out if a particular individual was or not in a certain genetic beacon. In a beacon with 1000 individuals, they showed that re-identifications was factible by performing just 5000 queries, identifying also relatives in the beacon. The final goal was to demonstrate how the basic scheme of a Beacon was not enough to protect privacy and phenotypic information.

Among the approaches they give to improve the security of the Beacon against such types of attacks, they highlight the importance of the Beacon's size, advising to set up a required minimum size. They also comment how dangerous it is to also publish metadata as, for instance, the ethnicity of the samples, which Beacons usually do, and limiting the number of queries per researcher (or even not allowing anonymous pings of genetic beacons). These topics will be further discussed later on in the Proposal Section of this work.

Finally, they comment that Breaking into Beacons data has a particular danger, Beacons are usually designed to share samples with a certain phenotype (a disease, for example), which also discloses phenotype information about the individual who is detected to be part of the beacon. Essentially, confirming the presence of an individual in a disease related Beacon, where we know there are donors who suffer the disease, makes it possible for the attackers to deduct if the re-identified person suffers the disease. Although knowing if a donor is a member of a certain Beacon or not does not compromise its security, the potential phenotype deductions that can be made are a severe privacy breach of sensitive information.

[27] *Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks, Raisaro JL et al. (2017), published in Journal of the American Medical Informatics Association.*

In this paper, Raisaro et al. propose some measures to face the vulnerability Shringarpure et al. exposed. This was addressed by proposing three different measures, two of them had the purpose to manipulate the beacon in some way by using different DP approaches in order to obscure the presence of rare alleles. While the third one is about not allowing infinite accesses per user for each individual genome, the three ideas will be explained and

developed at the Proposal section of this work, where we will try to qualify them as suitable or not for the genomic field.
A Beacon with data from the 1000 Genomes Project was used to run a successful test that showed those measures effectively reduce re-identification risk. A counterpart of those measurements is that they potentially eliminate the possibility of querying for unique alleles highly likely to be most useful for the genomic researchers.

On a final note the study highlights the fact that all known attacks, as the one Shringarpure et al. demonstrated, fundamentally rely on the assumption that the attacker already has access to the genome of the victim, which is a very strong assumption.

[28] *Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services, Zhiyu Wan et al. (2017), published in BMC Med Genomics.*

Aimed also to patch the security hole Shringarpure and Bustamante detected, this study describes a new computational method to enhance the balance between utility and privacy of the GA4GH's Beacon system. As a proposal for the 2016 iDASH Challenge, the solution they present is a tailored version made with Shringarpure and Bustamante attack in mind. The scientific paper they wrote was a formalization of their solution after it won the iDASH Challenge, where they included their general design of security and an evaluation that demonstrated its capacities.

While the strategy submitted was complex and relied on different concepts to work correctly, the most important two factors are: first, the term they call "discriminative power", which is a value they assign to each of the SNPs on the dataset and is related to how much information that can cause re-identification they are giving. The second, the measure applied when a certain threshold is surpassed to protect the information, which is flipping different query responses. This can be seen as a particular system of DP adapted to specifically counter Shringarpure and Bustamante attacks.

[29] *Genome Reconstruction Attacks Against Genomic Data-Sharing Beacons, Kerem Ayoz et al. (2020), published in Proceedings on Privacy Enhancing Technologies.*

This more recent text about genomic Beacons is headed also to prevent the inference attack Shringarpure and Bustamante discovered, but also unfolds a new type of vulnerability, genome reconstruction. They showed «that it was 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» (Kerem Ayoz et al., 2020, p.28). Specifically, they explain how attackers could use the genomic information and clustering techniques to successfully obtain a reconstruction. Even when multiple donors are added at the same time, they explain, reconstruction is still feasible. Easily accessible phenotypic traits an attacker could know, as victim's eye color or hair type, can be used to identify the victim among the donors added on the update.

This is especially dangerous as it could serve as a previous step into those attacks that basically relied on knowing the victim's genome. An attacker could, by using a not associated with a sensitive phenotype Beacon, reconstruct the genome of a victim and therefore obtaining it, and later perform an inference attack on an associated with a sensitive phenotype Beacon to identify and draw dangerous conclusions. Author's recommendations to avoid this new danger that are given to the Beacons systems is to always do big updates and add large numbers of new donor's genomes rather than just a few of them each time (and never a single one).

And finally, after making a balance about the inference attack solutions proposed in previous works, some 3 new considerations are pointed to add protection in this regard. The first one would be updating the beacon content when $m > 1$ (being m the number of newly added donors); the second one would be to add or remove the genomic information of donors after having measured how prone they are to reconstructions; and the third one would be to regulate ethnic diversity of the beacon in order to have mixed datasets (as it was observed that beacons with mixed ethnicity donors presented a harder challenge when constructing a correlation model, unless, obviously, ethnicities were disclosed as metadata).

[30] *Federated discovery and sharing of genomic data using Beacons, Marc Fiume et al. (2019), published in Nature Biotechnology.*

In this moderately recent paper, the authors talk about the paper of the Beacon system on genomic security. Among the different sections they talk about, the concept of Beacon Network stands out as a clever system that acts as some sort of intermediary between the user accessing the beacon and the Beacon information. The advantage it offers to the user is that it queries across multiple Beacon instances at once, using various types of data such as VCF files or patient records. Then, an aggregated answer made by the Beacon Network will be handed to the user.

This paper also praises the standardization of the Beacon system, «the first version of the Beacon Project has validated the feasibility of a globally federated system for genomic data sharing» (Marc Fiume et al., 2019, p.223). The simplicity of its basis, the question "Have you observed this allele?", has played a crucial role on this, which has led the system to a fast widespread acceptance. «However, the narrow focus of the initial Beacon question limits its utility to support other closely related use cases, and successive iterations of the protocol are planned to enable coverage of these» (Marc Fiume et al., 2019, p.223) makes clear the authors were already waiting for different upgrades from the already announced version 2 (Beacon v2). Different characteristics were listed by Marc Fiume et al. as expected and desirable advances of the first version. Beacon v2 and its capabilities will be, however, commented at a further part of this work, at the Beacon's Evolution section.

2.4.7 Encryption Solutions

This final section is for specific technologies that simply rely on plain encryption to protect data. These solutions are meant to be the safer options to store and check genomic information but not for sharing purposes. These solutions are Crypt4GH, Cryfa and SECRAM, and the following papers explain their advantages and their actual place in the genomic field.

2.4.7.1 Crypt4GH

[31] *Crypt4GH: a file format standard enabling native access to encrypted data, Senf A. et al. (2021), published in Bioinformatics.*

This paper published this year is dedicated to present the Crypt4GH solution. Here is described the security problem of having a lot of different files that often need to be checked and analyzed. Decrypting those files into clear text is needed, but keeping the decrypted file compromises the information it contains. The main idea of Crypt4GHis to prevent this issue, ensuring a permanent encrypted state of the file. Senf A et al. described how Crypt4GH «allows for reading encrypted data from file or remote APIs and performing in-memory decryption of just the byte ranges needed» (Senf A. et al., 2021, p.1). Doing so reduces in great measure the possible attacks, mitigating the inherent risk related to the decryption of the whole file and its storage. On a final note, they point to the fact that Crypt-4GH is not intended to address all aspects of genomic data security as it would need more features (e.g. key management), and should form part of an overall data security strategy.

2.4.7.2 Cryfa

[32] *Cryfa a secure encryption tool for genomic data, Morteza Hosseini et al. (2019), published in Bioinformatics.*

In this paper, Cryfa is presented as the best alternative for securing genomic data, due to its fast speed and reliability. In their words «Cryfa is a fast secure encryption tool for genomic data, namely in Fasta, Fastq, VCF, SAM and BAM formats, which is also capable of reducing the storage size of Fasta and Fastq files» (Morteza Hosseini et al., 2019, p.146). It is based on AES encryption but, compared to a regular AES encryption, which is a general-purpose tool, Cryfa is an industry-oriented tool, while being not only more reliable but performing the encryption at 4 times more speed and, also, compressing the files to ⅓ of its previous size. In the results, the authors comment how a straightforward shuffling mechanism and not exploring the complexity of the files encrypted ensure the safety of the genomic data.

2.4.7.3 SECRAM

[33] *A privacy-preserving solution for compressed storage and selective retrieval of genomic data, Zhicong Huang et al. (2016), published in Genome Research.*

This paper of 2016 presented SECRAM (Selective retrieval on Encrypted and Compressed Reference-oriented Alignment Map) addressing the problem of a reliable and secure storage for the genomic data in a compressed state. The authors compared SECRAM with BAM, which they refer to as the de facto standard for storage, and found that SECRAM used 18% less storage. Compared with CRAM, a format for compression that does not encrypt, «SECRAM maintains efficient compression and downstream data processing, while allowing for unprecedented levels of security in genomic data storage» (Zhicong Huang et al., 2016, p.1687) they say.

The three steps of preparing a SECRAM file would involve transposition, compressions and encryption. The inverse procedure can always be done where, through decryption, decompression and inverse transposition, we would obtain the original file. Thanks to this propriety, any BAM file transposed to a SECRAM format can still be converted to the original BAM format (with no further data requirement aside SECRAM metadata), which makes it suitable for a standard option. In this regard, the authors wrote:

«If necessary, our format can be inversely transposed to BAM without losing information; this functionality makes our format compatible with several other applications designed for a read-based format» (Zhicong Huang et al., 2016, p.1693).

2.4.8 Data Use Ontology (DUO)

Data Use Ontology, often referred as DUO, is a language of tags that allows semantic labelling to our datasets. While it would clearly not be suitable as a standalone privacy measure, it offers, through its tagging capabilities, not only to enhance the use we can make of our datasets but to improve our privacy options. The DUO system has to be seen as a restriction method that will allow data proprietaries to set up their boundaries, granting or not the permissions of use of their data in a more malleable way.
In February of 2019, the GA4GH announced DUO as the genomic and health data sharing standard [34]. DUO's properties were able to fit a previous necessity to allow better communication between data holders and data researchers. Thanks to a consensus system of developing the terms, each term is part of a shared system of terms, allowing any restriction to be understood by any researcher. Data queries will need to be accompanied with the DUO restrictions it comprehends, which the algorithm will match to the datasets that fit the query. DUO code is public and can be found at its online repository [35].

The following paper is not based on DUO but on Library Cards, which would be a prototype of some sort of identification that would include allowing policies. For this specific purpose, the DUO system would be used as a part of those Library Cards.

[36] *Simplifying research access to genomics and health data with Library Cards, Moran N. Cabili et al. (2018), from Scientific Data, Nature.*

Alluding to the growth of genomic data volume, the authors point out the great problem of using fragmented datasets. This is a very inefficient and time-consuming step for both the data proprietors, ensuring their boundaries are clear, and the researchers, ensuring they are permitted to use a specific dataset (for each one they intend to use). For this and other purposes, more related to user's authentication, Cabili et al. propose the use of Library Cards, as some identification system that, among the information that would hold of any dataset, would also include the use policy. This use policy would be based on attributes, and those attributes would be translated into DUO tags.

While it does not explain further how DUO works, a big part of this work is aimed to make clear that a standard system to set up permissions is highly needed. And DUO is, by what was its current trajectory, the most promising system to be developed as a standard. In the words the authors dedicate to DUO, they expose how close it is to be announced as a standard, explaining how it is not only a GA4GH's work as it was already included in EGA's new and old datasets. EGA, they finally say, was also part of DUO development.

2.4.9 GA4GH's Passport

In a similar perspective, the GA4GH's Passport is another tool that is used as a complementary help when ensuring genomic privacy. The GA4GH formally announced the Passports specification on the December of 2019 [37], along with the Authentication & Authorization Infrastructure (AAI), as one complemented each other.

Both GA4GH Passports and AAI have public repositories that can be accessed via web [38] [39].

«*GA4GH Passports and the Authorization and Authentication Infrastructure*»[1].

Where AAI is the layer on duty to authenticate users with the intention to grant or not access to a particular dataset, user's GA4GH Passports are what make this possible. Passports would not only carry the information about if a user should have or not access to a dataset but also if the user did accept the related terms of service. The access token AAI uses for authentication purposes is later used by the Passport to transport a researcher identity, and hence its permissions, to different environments.

Inside each Passport, a Passport Visa carries not only the identity of the holder but also one or more Passport Visa Objects. These objects will carry a lot of information, which includes the consents linked to the identity. The article also explains how the Visas can be used as an international visa, so a single visa can be used different times before it expires, granting different types of temporary access.

1 As far as we know there is no paper yet about the GA4GH Passports. In this case, the literature review is performed upon the presentation article of the Passport by the GA4GH [37].

Finally, it is also stated that both GA4GH's Passports and the AAI are a step towards automation. Different types of data owners will have different ways to check the Passports, but they should always be able to tell if consent is given to a particular Passport or not. When illustrating this, the article says «for example, a researcher can combine data between two or more datasets spanning multiple clouds, provided they have been authorized to access each» [37].

2.4.10 XACML

The last technology reviewed will be XACML, which stands for eXtensible Access Control Markup Language. It is an access control policy language based on attributes and a clear example of a security system not linked to genomics in particular.

XACML is, like the last two previous technologies reviewed, another way to set up rules for access control. To accomplish this it relies on the Attribute Based Access Control system for it (ABAC), which allows it to assign different attributes, either related to the user, environment or resources, granting or not permits depending on them. This is a complex but potentially very flexible way to condition permits to age of the user, time of the day or location of the file the user is trying to access, for example. In the next paper XACML's role as a genomic privacy preserving tool will be analyzed.

[40] *Protecting Privacy of Genomic Information, J. Delgado et al. (2017), published in Studies in Health Technology and Informatics.*

In this paper, the authors proposed a privacy preserving scheme for genomic information to the MPEG Committee (Moving Picture Experts Group). On the proposal, XACML was the tool of choice to ensure proper access control. With the help of XACML expressions, the authors propose to limit the access control based on user's roles, specific data files, day, purpose of the research or consent of the data provider. Those expressions, that would act as rules, would be stored directly into the genomic containing files (e.g. SAM or BAM formats). «The inclusion of the rules inside the genomic files allows us to extract them and authorize access to the file according to the permissions defined in the rules» (J. Delgado et al., 2017, p.320), they explain.

Under this premise, when a genomic repository receives some user request to access a genomic file, it will simply extract the rules on the file and read them. An Authorization point is then asked by the repository, which will respond if the access is granted or not depending on the rules. If the Authorization Point has permitted the access through its reply, the repository will then reach the user with the decrypted data that was requested.

Chapter 3
Brief classification of privacy preserving techniques for genomic information

3.1 Documentation

The next step on the literature reviewing phase of this thesis consists in grouping all the privacy preserving techniques we have seen on different papers and trying to schematize them into a comprehensible classification that could serve us on further steps. All articles of the actual bibliography have had something to do with these techniques but some were especially useful in this regard as they provided their own classification or list of techniques.

On *Systematizing Genome Privacy Research: APrivacy-Enhancing Technologies Perspective* [41], by Mittos A et al., they did a comparison of the representative genomic privacy methodologies (Mittos A et al., 2019, p.95). It will be hard for us to benefit from the actual comparison as they compare a lot of different papers rather than the method itself, but this also has the advantage that it is a proven work method. The list is the following:

- SWHE: Somewhat Homomorphic Encryption
- LHE: Leveled Homomorphic Encryption
- Fuzzy: Fuzzy Encryption
- PSI-CA: Private SetIntersection Cardinality
- A-PSI: Authorized Private Set Intersection
- C-FE: Controlled Functional Encryption
- HoneyEncr: Honey Encryption
- OPE:Order-Preserving Encryption
- MPC: Secure Multiparty Computation
- PIR: Private Information Retrieval
- SGX: Software Guard Extensions

The *Privacy challenges and research opportunities for genomic data sharing* [11], by Bonomi L et al., written in 2020, includes different explanations on how this sector works and finally adds an overview of the current known techniques. While not as extensive as the previous list, this one is classified, detailed, and more up to date. Summarized, the original table differentiates two approaches (Bonomi L et al., 2020, p.20):

- Data Security, where you can find:
    - Access Control
    - Homomorphic Encryption
    - Secure Multiparty Computation
- Data Anonymization
    - k-anonymity
    - Differential Privacy

On *Privacy-preserving techniques of genomic data—a survey* [6], by Md Momin Al Aziz et al., published in 2017, the evolution of the methodologies is studied and they draw a timeline of the evolution of genomic data studies and seminal development of different privacy-preserving techniques (Md Momin Al Aziz et al., 2019, p. 891). They will later talk mainly about Homomorphic Encryption, Garbled Circuits, Secure Hardware (with Software Guard Extensions) and Differential Privacy.

Finally, Francesco Marino and Jean-Pierre Hubaux held this year a presentation, *CWS APIs for supporting next generation secure workflows, Gap analysis* [42]. This presentation revolved around a joint work done by James Scheibner et al., including Jean-Pierre Hubaux, called *Revolutionizing Medical Data Sharing Using Advanced Privacy Enhancing Technologies: Technical, Legal and Ethical Synthesis* [24]. Among its slides, one is used to classify the techniques. While this has to be considered completely biased (their purpose is to present and promote their idea of Multi-party homomorphic encryption, which has been left apart in the Systematic Review as it is not a working method yet), it does provide a very simple branched tree of some of the current methodologies.

3.2 Classification Proposition

Based mainly in the previous Systematic Review we have performed, we present at this chapter a Classification Proposition of the Technique, framing the methodologies with the actual technique when necessary. This will be helpful given the situation of security in the genomics field, as it is easy to understand Homomorphic Encryption as a technique without providing any of its many examples of implementations but, conversely, it is hard to classify a very specific system like the Beacon without naming directly the GA4GH Beacon standard.

3.2.1 Level I. Based on Software and Based on Hardware

As we have seen the vast majority of the reviewed techniques are software-based, with the sole exception of the section of hardware-based. For this reason the first division among our techniques is in this regard, differentiating the ones based on software from the ones based on hardware. On the former, we will find further divisions and, essentially, the rest of the techniques, while on the latter we find it is a technique on its own, where Intel SGX or ARM TrustZone are probably the two most relevant representatives.

3.2.2 Level II. Based on Software: Data Security and Data Manipulation

At the software-based techniques we face the next division, which is about how the data is treated. We propose to differentiate by Data Security and Data Manipulation. The idea of Data Security is to comprehend all those techniques whose data protection measures avoid any kind of data alteration. Data Manipulation includes those techniques that use the manipulation of the data to ensure privacy. The purpose of this division is no other than separate what can be seen as regular encryption from what is based on some type of data manipulation.
Data Security includes HE, MSC and all the other techniques that rely on Access Control, which divisions will be further discussed. Data Manipulation includes anonymity techniques, such as k-anonymity or Pseudo Anonymity and the well-spread technique of Diffe-

rential Privacy. Data Manipulation techniques are fundamentally based on the alteration the original data suffers, ensuring privacy at the cost of providing false information, as it is not possible to recover the original data.

Data Security measures are, instead, related to avoid any attacker to access or decrypt the data while maintaining its integrity. As we already know from the Systematic Review, some of the techniques we have classified as Data Security actually may involve some techniques that do not grant all the information to the users (e.g. Crypt4GH) or even use Data Manipulation techniques (i.e. some implementations of the Beacon that use DIfferential Privacy). The first case is actually a matter of Access Control and partial encryption but does not manipulate the data, while the second case is actually the conjunction of two separate techniques, as the Beacon system itself is just an interface users interact with in order to control the information that is given.

Homomorphic Encryption and Multi Secure Party Computation are, on their basis, different ways to deal with sharing encrypted data, thus fall in the scope of what we define as Data Security. Implementations of Homomorphic Encryption include Fully Homomorphic Encryption and Leveled Homomorphic Encryption in the same way Multi Secure Party Computation include Garbled Circuits and Point Set Interjections implementations.

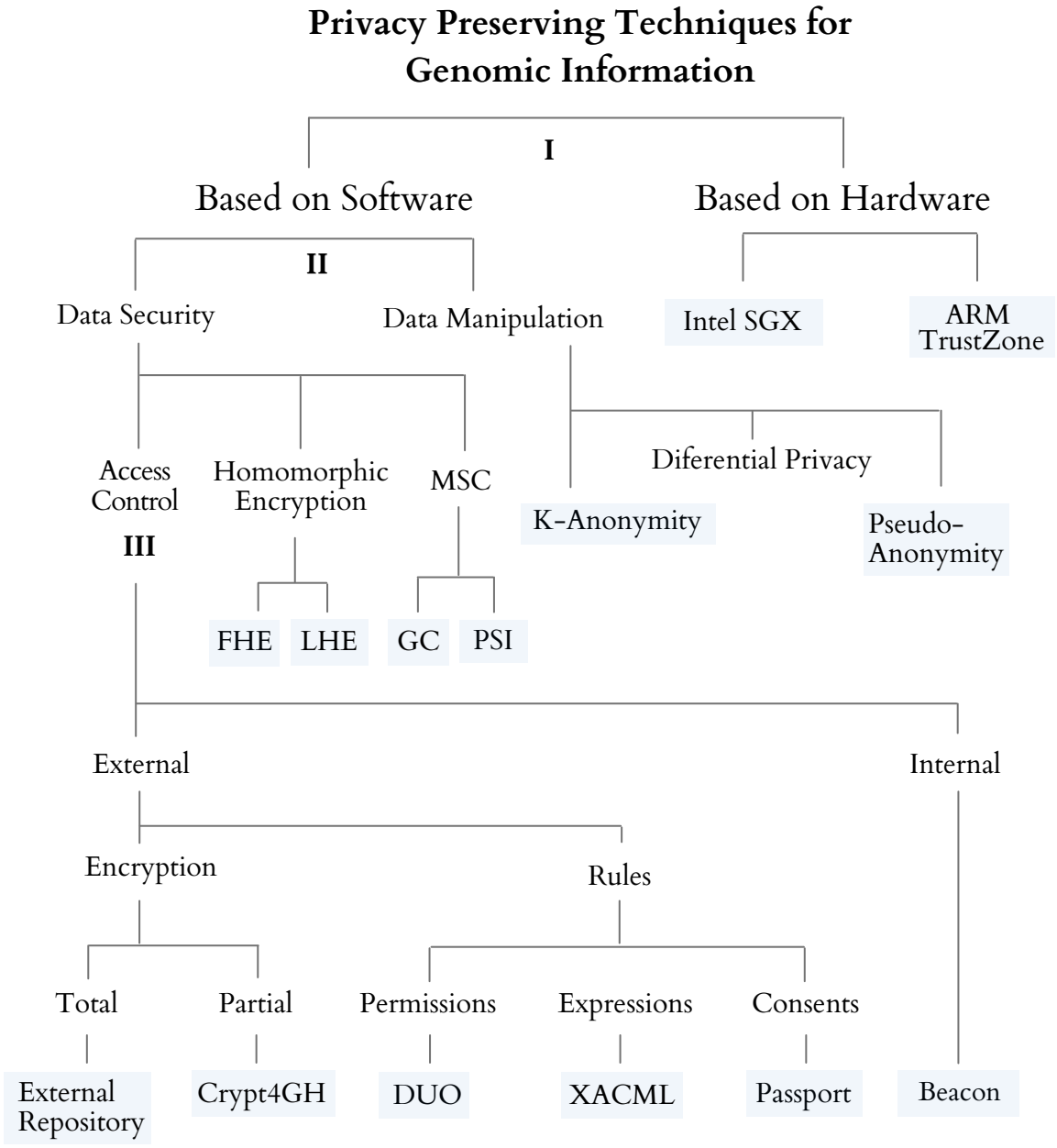
3.2.3 Level III. Data Security: Access Control

The third Data Security implementation listed is Access Control which is the most differentiated one. It still has the purpose to avoid undesired parties to get sensible information but, instead of using encryption techniques, this is achieved primarily by controlling who gets access to the data.
This idea is commonly found as, by performing this Access Control (i.e. a login), we know who we are handling the data to. On our scheme, this is represented as External, as the data is actually sent over to external parties. Internally, we found further divisions can be made, separating those that use Encryption techniques to discern if an access is trustworthy or not, and those that use more special rules. The former group we can find the simple idea of external repositories that require certain credentials to be accessed or Crypt4GH,which as we have seen credentials may unlock access to specific parts of a file (both are, then, differentiated as Total or Partial).

The latter is a more complex group of specific technologies which work differently from each other but share this common factor that requires some conditions to meet. In one case, as DUO, by granting permissions to use some data, in other case, as Passport, by se-ttling up some specific authorizations linked to a user and, in other cases, as XACML, by using specific expressions that will allow us to grant or not the access depending on the role of the user, the date, or some characteristic of the data.

But, perhaps, the most unique implementation of an Access Control system is the one presented by the Beacon System. The Beacon System by itself does not control its access in any way, meaning everyone can access it, but it does perform a control over the data shared. Data is never directly shared as it was sent to the user, data is always stored in some dataset or repository from which the Beacon acts as a facade, an interface between the users and the data. Previously described in the Systematic Review, Beacon's working mechanism is based on answering precise questions the researchers will ask it. By this peculiar but effective system, the GA4GH Beacon standard has been more and more accepted between the community but, as we reviewed its recent literature, we are aware it is not a perfect system.

At least in some specific and unlikable situations it has been proved the Beacon is not flawless, but its procedures, and hence its results, can be yet improved. Further sections will describe Beacon processes and its capabilities with the purpose to define a solid proposal of measures to enhance Beacon's actual security potential (chapter 4 Beacon Evolution and chapter 5 Security Improvement Proposal for the Beacon System respectively).



Legend:
DUO = Data Use Ontology
DP = Differential Privacy
FHE = Fully Homomorphic Encryption
GC = Garbled Circuits
LHE = Leveled Homomorphic Encryption
MSC = Multi-Party Secure Computation
PSI = Point Set Intersection
SGX = Software Guard Extensions
XACML= eXtensible Access Control Markup Language

I = Level I
II = level II
III = level III

Figure 3.1. Proposal of classification of privacy preserving techniques for genomic information

Chapter 4
Beacon Evolution

We have until now presented the Beacon system as a way to ensure genomic privacy and classified it among the other techniques generally used for this same purpose. We have performed a systematic review of the related literature and also described the Beacon Project as the GA4GH Standard.

4.1 Beacon Standardization process

Since its conception the Beacon project has evolved significatively. In 2015 the GA4GH answered Shringarpure and Bustamante's work [26] with the common observation of how an attacker should previously own the genome of a specific person (or a very close relative one) in order to identify them among the Beacon database, potentially guessing phenotypes of this person as a result [43]. Peter Goodhand, the executive director of the GA4GH, welcomed the vulnerability exposition as a security concern and Bustamante himself pointed out how subtle were the problems they found on the Beacon project. It is indeed a very special scenario to try to hide things to an attacker that already knows someone's genome but, if there is some room for irony, it is the existence of a phenotype associated Beacon that allows the attacker to perform Shringarpure and Bustamante's deductions. The mitigation efforts announced were the aggregation of data among multiple Beacons to increase database size and obscuring its origin, the creation of a budgeting system to control the rate the information is revealed to perform some actions if the threshold is exceeded, and making use of different tiers of secured access, where users should be authorised for data access and agree to not attempt some specific risky scenarios.

The collaboration between the GA4GH and ELIXIR, the European infrastructure for life-science data, is announced in 2017 with the objective to conjunctually make the Beacon project progress [44][45]. This partnership was built on an existing collaboration between the two organizations that already led to six ELIXIR Beacons in Sweden, Finland, France, Switzerland, Belgium and Barcelona, an additional one was launched later in the Netherlands.

Serena Scollen, Co-Lead of the Beacon Project and Head of Human Genomics and Translational Data at ELIXIR, highlighted the importance of the partnership with the GA4GH in the matter of a correct alignment on the standard development. The establishing of a network of ELIXIR Beacons was therefore the first objective, and ensuring data generated in Europe could be used by researchers across the world is its ultimate purpose.

Later on the same year, the GA4GH announced some new strategies for securing Beacon datasets [46]. Those measures were based on the Raisaro et al. very recent work [27] as a response to the high number or cyber-attacks health data was receiving (in 2015 health data was the number one target for cyber-attackers) [47]. JP Hubaux commented that «each mitigation strategy tackles the problem from a different angle, and together they provide Beacon owners a collection of solutions to choose from when determining how best to secure their particular datasets» [46]. One of the measures, for instance, could be to only respond "yes" to a question about the presence of an allele in the Beacon only if the allele is present in at least two of the genomes the Beacon contains. Another strategy, aimed for Beacons with rare variants, might rely on the privacy budget strategy.

In 2018, the GA4GH announced a new security measure for data access termed "registered access" [48]. Based on the proposition of the paper *Registered access: authorizing data access* [49], done recently by GA4GH members, a new tier between open access and controlled access was introduced. If the previous state where data was whether in an open access tier (public for everyone) or in a controlled access tier (private for everyone except for some qualified researchers and only for a specific project that went through a review process), the new registered acces would allow users to access the data after verifying their identities and roles (e.g., researchers), obtaining in the process and attestation of their agreement to the standard responsibilities. In the words of Dixie Baker, co-chair of the GA4GH Data Security Work Stream and an author of the paper, at the GA4GH article, «this model supports role-based access control, which is well established in other fields, including government and industry and can be implemented using existing technology standards widely used around the globe» [48]. The approach was from the very beginning piloted within the ELIXIR Beacon network.

In late 2018 GA4GH and ELIXIR announced the release of Beacon API v1 (the version 1.0) with increased security measures[50][51]. This Beacon API extended the functionality by adding support for additional types of genomic variants and improved metadata support.

Also, the ELIXIR Beacon reference implementation used the ELIXIR Authorization and Authentication Infrastructure (AAI). The implementation, primarily funded by ELIXIR but still an open access GA4GH standard, includes support for anyone wishing to light a Beacon around the globe, and uses the three different tiers that control access already commented (open, registered and private).

Future extensions of the Beacon API are presented in a letter to the editor of Nature Biotechnology published on 4 March 2019 by members of the GA4GH [52]. Through the announcement, the GA4GH makes public some interesting facts as there are how there were more than 200 datasets, which conjunctly contain the genomic data from more than 100,000 individuals, as well as the fact that the Beacon Network [53], a distributed search engine across all the world's public Beacons, was already queried 1.5 million times. Is in this time where the GA4GH announces its plan to expand further the Beacon project, intending to add real-world clinical value beyond research. «Previous versions of the Beacon API only provided a yes or no answer to the question, 'does this dataset contain X allele at Y genomic position?'», explains the article. «Today, that information can also be served alongside additional metadata, including allele frequencies, pathogenicity scores, and phenotypic information associated with the queried allele» [52].

4.2 Beacon v2

Beacon v2, the second version of the project, is the conclusion to the commented intentions to improve the original Beacon.

Michael Baudis et al. presented at the University of Zurich, in 2020, this Beacon v2 through their exposition *Beacon v2 - Towards flexible use and clinical applications for a reference genomic data sharing protocol* [54]. The presentation pointed out the ELIXIR Beacon Project roadmap, clearly stating the improvements the Beacon v1.1 achieved: more structural variations (as DUP or DEL) in addition to SNV, more structural queries (translocations/fusions), Beacon queries as entry for data handover (outside Beacon protocol), and the layered authenticated system the use of ELIXIR AAI provides. Leaving then the filters (both for phenotypic data and technical metadata), and the extended quantitative responses the first version of the Beacon lacks.

The improvements each version brings to the table are summarized into the Table. The first of them, in the second column and already implemented by Beacons, being the separate query types for different genomic variants allows users to correctly reference their queries as SNPs, structural variants, or by region, for example. And while access levels were already described, filters allow researchers to precise their queries (e.g. by filtering by gender or age).

CINECA, another third-party beacon implementation, explains more characteristics of the Beacon v2 [55] as allowing to reach further in the data access process (e.g. who to contact or what are the data use conditions), jumping to another system where the data could be accessed (e.g. if the Beacon is tied to a hospital, it could provide the ID of the Electronic Health Record of the patients with the searched mutation, while preserving confidentiality), or including annotations such as some conclusions about a mutation role on a phenotype.

The last of the characteristics listed belong to the version 2.n, which is nothing else than being a still in progress characteristics a new version of the Beacon will hold. There are the new types of queries (allowing querying by sample or patient, or even more complex queries, and schema and service info, which would translate into negotiated queries based on individual Beacon capabilities.

Beacon v1.1	**Beacon v2.0**	**Beacon v2.n**
Structural Variations	Separate query types for different genomic variants*	New type of queries**
Structural Queries	Access Levels	Schema versions & Service Info
Beacon queries as entry for data handover	Filters	
Layered authentication by using ELIXIR AAI		

SNPs, Structural Variants, Region, etc

** By sample or patient, for example. Also more complex queries.

Table 4.1: Introduced features of Beacon versions

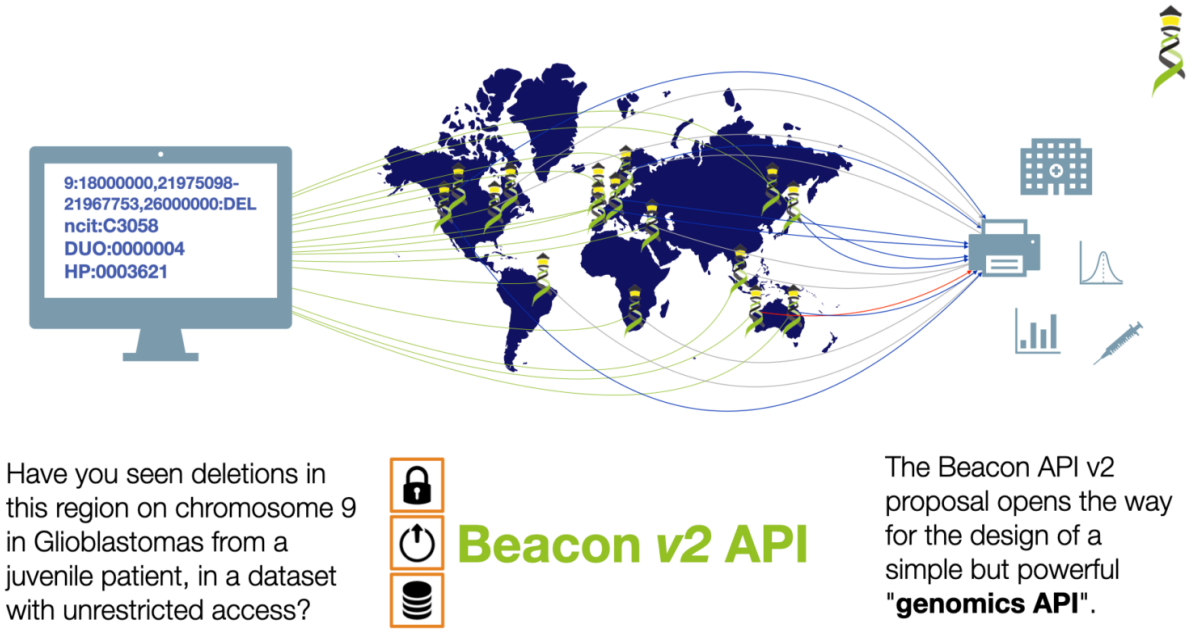


Figure 4.1: Beacon v2 API.
Beacon v2 is able to answer more complex queries, allowing researchers to specifically ask for certain regions of the genomic code in a more flexible way, filtering by donors characteristics (such as their age) and ensuring, with the use of DUO, we have legit access to obtain information from a particular dataset. [54]

4.3 Beacon Network

As we briefly commented earlier the Beacon Network is a search engine that allows researchers to query all world's public beacons at once. This is a very useful advantage for obvious reasons such as having a huge amount of samples for the study or not having to search for a lot of different beacons and ask them separately.

The Beacon Network was developed under the directions of ELIXIR Finland, and integrated the authentication process from ELIXIR AAI. Michael Baudis comments the importance of the Beacon Network in his presentation of the Beacon v2 [54] and, although it is a characteristic, he defines in the Beacon v1.n roadmap it is considered not completed, eventually including "networking of v1.n Beacons with AAI integration as demonstrated by the ELIXIR Beacon Network" among the features and possibilities of the Beacon v1.n.

Through the presentation, Baudis states how the use of a Beacon API made possible the Beacon Network and how this improved the scope of the queries performed. Beacon v2 API is even a further step in the same direction thanks to its simple but powerful "genomics API", and what Baudis calls "Internet of Genomics", which would be a common interface for a dynamic compendium of different genomic related data sources and services the GA-4GH plans to launch [56].

When Shringarpure and Bustamante exposed Beacon's vulnerabilities [26] they also made some notes on how to make the re-identification attacks harder for the attacker. As the success rate of their attack was heavily based on the proportion of disease related genomic samples on a dataset compared to what we could call control samples, it is clear that increasing the number of those control samples hinders the attacker's possibilities to reach a phenotypic conclusion. The Beacon Network could not only contribute to increasing the scope of the queries then but also in terms of security by adding a huge number of genomic samples from different datasets, which can act as supplementary control samples from one to another. «By including non-case samples, these solutions reduce the phenotype information that can be obtained from a Beacon while keeping the reduction in the utility of the Beacon to a minimum» (Shringarpure, S. S., & Bustamante, C. D., 2015, p. 638), they commented.

But this only limits the conclusions an attacker would be able to draw about the phenotype of a victim, which is indeed very important, but does not address the problem from its root, as the re-identification risk still persists. Md Momin Al Aziz et al. commented on [57] that the attackers could just query different beacons thanks to the Beacon Network until finding the re-identification of some particular genome they were looking for. Even when Shringarpure and Bustamante found the Beacon Network useful to limit the impact upon a successful attack of re-identification, the attack per se is still feasible to perform.

Chapter 5
Security Improvement Proposals for the Beacon System

Since vulnerabilities were discovered on the Beacon system, different proposals have been presented in order to ensure people's genomic privacy. This following chapter contains a review of some already known methods plus some other new proposals we think could lead to a safer Beacon environment. The final aim of this text is, therefore, to present a compendium of different solutions that could work together towards this direction.

5.1 Noise-based solutions

A common solution widely spread among the existing proposals is to simply protect those genetic oddities that only happen once in our repository. In the scenario where a Beacon is asked about the presence of some particular SNP that is only found once in our repository, a Beacon that effectively uses the commented feature would answer "No" instead of "Yes". This misinformation adds a huge layer of protection against re-identification attacks to the point that it is nearly impossible to successfully perform it anymore, but this is no other thing than adding noise for protection, which falls into the spectrum of differential privacy. The previous Systematic Review exposed how differential privacy was a studied method for some authors and, while safe enough, it pushed too much the trade-off between utility and security too much to the latter [15][18][19].

It is also our opinion that this imbalance should not be simply accepted in the medical field since some consequences it could carry could potentially be, literally, a matter of life or death. On top of this, there are two more relevant reasons against this noise adding practice. The first is that our genomes are huge extensions of base pairs and oddities occur approximately once for every 1000 nucleotides, meaning there are 4 to 5 million of SNPs in every person's genetic code [58], because of that a lot of them are always unique. And the second reason is that those uncommon particularities are what the researchers are actually looking for, meaning not only we are hiding a huge amount of information but the most interesting one.

For a numeric reference, we will use the first proposal of Raisaro's et al. paper [27] we briefly reviewed at the Systematic Review. The authors used an experimental Beacon with 1000 individuals and, at the moment to implement the mechanism that would answer "No" every time there was only a single individual with the queried characteristic, they found that roughly 40% of the alleles stored presented at least one unique trait. Beyond that, as we commented, those are potentially the most interesting ones for the researchers. The authors recognize the infeasibility of this proposal they call Beacon alteration and find a solution for it. They called their second proposal random flipping, and with it they will not hide the entirety of the alleles but a portion ε of them. Through testing, they discovered hiding 15% of the alleles (at their most unique part) hindered re-identification possibilities. Adding the noise now in a more accurate way than before, only a 15% part of the 40% of the alleles are obfuscated, which is 6% of the total information our repository contains. The trade-off between utility and security is now more acceptable indeed, but in our opinion, the fact that this particular 6% is among the most interesting information for the researchers and that medical research should not accept any kind of misinformation still outweigh the advantages, which is enough to not propose this kind of solutions for the Beacon system. In the next *Figure* we can see the results of both of these proposals Raisaro et al. presented.

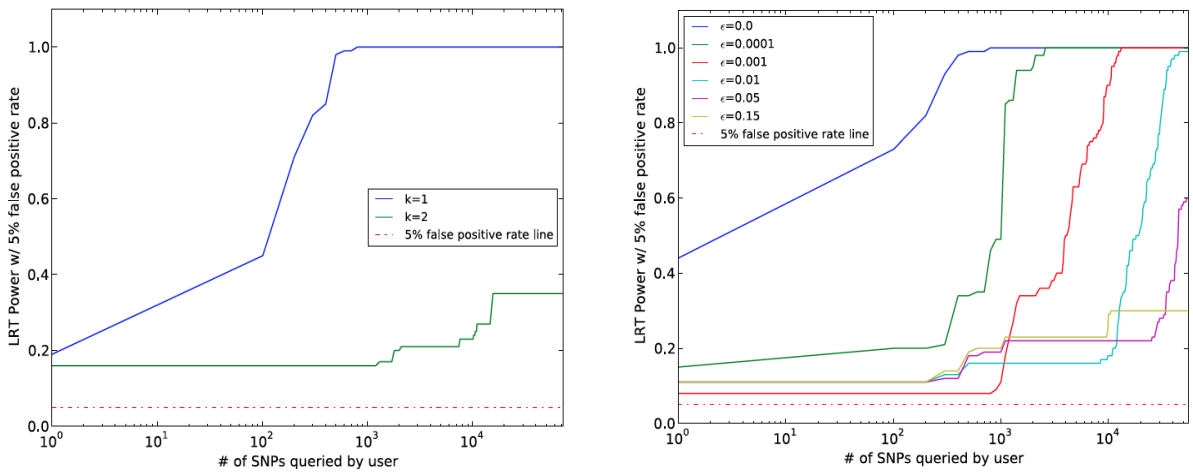


Figure 5.1: Comparison of Differential Privacy with the use of random flipping
Left. Beacon alteration, with k being the minimum number of times a genetic trait has to appear in order to not be obfuscated, k=1 does nothing and re-identification occurs with strong confidence before 103 queries. **Right**. At ε=0.001, only a 0.1% of a unique allele is hidden, and it takes around 104 queries to reach strong confidence re-identification, at the red line. Yellow line is hiding 15% of a unique allele, and is enough to not allow strong level reidentification.

Source: [27]

5.2 Increase Beacon's datasets size

The problem of unique traits in the genomic code is a complicated one and, in order to partially address it, we present the first of our proposals, which is to make our Beacons bigger. This is an easy in theory but hard in practice solution, something as simple as to make our repository bigger is, obviously, achieved by adding more genomic samples (more donors) to our system. This approach has been suggested by several researchers, and was even advised by Shringarpure and Bustamante at their vulnerability exposure [26]. In their scenario of a 1000 donors they could reach a re-identification within 5000 queries, and besides the obvious fact that the more genome samples we have, more difficult the re-identification will be, it is also worth pointing out that the number queries will grow not proportionally but exponentially, as we are less and less sure about the identity of the information we can get. In their words «Increasing beacon size can make detection harder, but protection against genome-wide re-identification attacks will require tens of thousands of individuals» (Shringarpure, S. S., & Bustamante, C. D., 2015, p. 637), they clearly state the number to ensure security would be immense, which invalidates this solution as a unique solution.

5.3 Use of control samples

However, increasing the size does not only address the re-identification problem but also the phenotypic linkage possibility. We can do this by adding samples not related to a disease a repository is linked to, samples we can refer to as control samples. Upon the re-identification of those control samples an attacker will not be able to tell for sure if that donor does suffer the repository-related illness or not. This is commented on in the previous section of this work dedicated to the Beacon Network, at the chapter 4 Beacon Evolution. There, this platform is proposed as a tool to easily achieve this control sample strategy, effectively obstructing the attacker's power to make phenotypic assumptions of the re-identified donors. The downside this carries is that researchers will need to perform a bigger number of queries to extract the information they need, but instead of relying on delivering false information (as Differential Privacy did), this always discloses good and reliable data. For this reason, a part of our proposal is to always increase the size of our datasets, both with disease related donors (if possible) and control samples, in which case the Beacon Network can be used.

5.4 Protecting most vulnerable genomes by not sharing their data

Ideally, through the process of making our Beacon's repository bigger we have increased the number of times a particular SNP appears, therefore protecting those that were previously present in a single donor and now are, at least, also in some other donor. If every single mutation now has multiple appearances we are severely limiting the re-identification potential of the attackers, but this ideal outcome would be an extremely unrealistic assumption. As commented earlier we humans carry so many unique particularities from one to another so there is indeed the possibility of actually adding more people with some SNP we only had once on our repository. In regular circumstances this could only happen for the more common mutations (some SNPs for example), but it is extremely unlikely we can find it in more complex polymorphisms or actual mutations (which differences will be addressed later). We can safely admit then that while some mutations will be more present, new unique ones will appear.

Increasing the size of our datasets is still one of the most obvious and easy options to make our Beacon more secure, and beyond the difficulty of performing it, it has no drawbacks, but to ensure the security of our donors we think this measure is not enough. Differential Privacy, we just said, comes at the cost of misinformation, because we are effectively lying about the presence of some alleles, which in a greater or lesser measure, will impact the medical conclusions. This happened by answering "No" instead of a "Yes" when asked about the most vulnerable (and unique!) parts of a genome. Instead, we propose answering "*You have no permissions to see this information*" to protect those alleles, which added to some other measures we are about to explain (basically not protecting every single unique variation and a very strong system of permissions) we expect to improve the overall security of our donors.

Although answering "You have no permissions to see this information" when the researcher actually has permission is also lying, this lie has no repercussions on the data integrity, and no false genetic information will be given to the researchers. The basic purpose of this lie lies in the fact that it does not influence the medical conclusions the researchers will extract. The problem we face, if we just do this expression replacing, is a new one, answering "No" using Differential Privacy to protect mutations only worked because it was indistiguishable from a genuine "No" the Beacon would

answer if a specific mutation would have not been present. We now have a new expression which, if we do not find a solid purpose for, it will be detrimental to our objective.

Let us use an example to illustrate this issue: if we simply say "*This information cannot be disclosed as it could lead to re-identification*" everytime we have only one donor with the queried genetic particularity, we are actually not only confirming the presence of the allele in our dataset, but also letting an attacker know we only have one donor with it (or, at least, that we consider it vulnerable enough to be protected, so it is relevant). To prevent anyone from being able to make these deductions, we propose the aforementioned expression, "*You have no permission to see this information*", in conjunction with a very strong system of permissions, which is then not only a reliable feature on its own but also a strong support for our new proposed expression. If it is possible they actually lack some permissions, attackers will not be able to tell if this expression is telling the truth or not, even if they know the existence of this security mechanism.

5.5 DUO based strong system of permissions

Permissions are then a vital part of our proposal, which are made possible thanks to the use of DUO tags [35]. With them, permission can be set by the donors in very specific ways, as only allowing, for example, the use of their genomes for public research (by author), or for cancer oriented research (by purpose), and not just granting or denying the information by default. Encouraging the use of the DUO system is primordial, as by its flexible way to set up boundaries, data owners are capable of defining under what means they are comfortable with sharing genomic information.
The use of these DUO permission tags at querying is already part of Beacon v2 [54], at which point we would want to add two extra features. The first of them will be allowing the donors to set the permissions of their genomes by parts, in other words, not granting access to the whole totality of their genome but using part of it instead. This was another of the measures advised by Shringarpure and Bustamante on [26] and later, in 2018, proposed and tested by Demmler et al. at [59]. Specific Beacons that only hold and share small genomic regions are a possible outcome of this, and those are, by default, a lot more secure than the ones that share full genome information.

For the subdivision of the genomes, which we consider escapes the scope of this project, a system similar to Daniel Naro's proposal at [1] can be used in order to make differences among the parts of a genome.

He presented three different possible subdivisions, the first one based on a fixed length to separate blocks from one to another. The second one would separate the blocks by using genomic information, as per known genes or until STOP codons are found (which are known to end the protein constructions). Finally, the third one would be based on the pre-vious option, but also limiting the size of each block with a fixed length when there is no knowledge of a gene in the zone.

More complex subdivisions can be made if we instead use the randomly distributed mar-kers related to the genome mapping. Following the explanations of Tom Stratchan and Andrew Read found at [5], we learn how during the genome process, identical DNA se-quences coming from a single individual are then only cut in a fraction of the possibilities where this could happen. This leads to different cuts or sections whose joint is how a geno-me is expressed, and as those cuts are randomly performed several times against the same DNA, those cuts do not start when the previous one ends nor do they stop where the next one starts. Instead, they partially overlap one to another; a visual example of this is repre-sented in the *Figure*. Using these random subdivisions (specially in conjunction) could lead to a complex way to adapt the partial encryption we are talking about, but this is only to demonstrate that complex subdivisions can be made in a very easy way, as they are already done when mapping occurs. However, this genomic subdivision example has not to be considered a strong proposal on the same level as our other security related ideas that we are expressing. Ultimately, we leave this decision to the judgment of genomic researchers.

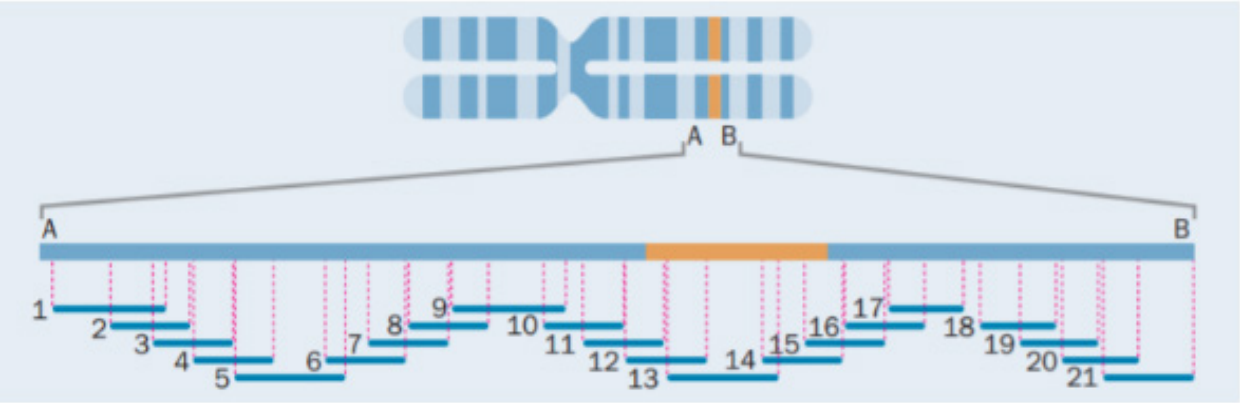


Figure 5.2: Subdivisions of a chromosome
The particular section of a chromosome, in this case from point A to point B, is expressed as the complete series of the subdivisions, in this case from 1 to 21. As those subdivisions have been done randomly against the same DNA, separations will hardly meet, and some overlap will take place from to the adjacent ones. Those separations are unique in each mapping process.
Source: [5]

With the particular setup of allowing our donors to set the grade of privacy they want, we think we are making possible a better environment for both our researchers and our donors involved in our Beacon. Ideally, donors could express their consent whenever they want using some sort of dedicated application but typically this is done when donors are actually giving their data. There, they should be able to give or not give their consent, or even apply some conditions, as the ones we previously mentioned like depending on the author or the purpose of the investigation. The granularity of this process has yet to be discussed and while we do not address this problem in this work we assume a somewhat wide span of possibilities is offered to the donor, which is what will offer the most security. It is possible they use this feature to make their data almost (or completely) unaccessible, but we think it will lean towards the opposite pole. If a patient suffers from a strange disease that causes severe health conditions, or even death, as for example a cancer would do, it is easily understandable the patient will accept the risks of being re-identified in order to help any possible favorable investigation (specially considering how some strong assumptions need to be made, as the attacker being in possession of the victim's genome from the beginning).

As security designers, we should not disclose information about a donor if we think it can lead to a re-identification, but if the donor himself has expressed his desire to share this information, regardless of the consequences, we reach a point that is beneficial for both the scientist and the patient. This also contributes to our previous measure of answering "Y*ou have no permission to see this information*" because we are now not answering this even when we have just a single donor matching the query, eliminating now the assumptions an attacker could make (now even when a response is made, it does not mean that is present in more than one donors). Following this idea we have what we differentiate as the second permission related measure, which is to create what we call "trusted environments".

5.6 Concept of "Trusted environments"

"Trusted environments" are like safe zones where the genome information may be disclosed without any security measures or needing any extra permission. These zones do not make any effort to try to avoid re-identification and by definition will only be available to the most trusted parties. The evident danger they carry is high, but do not forget something similar to these "trusted environments" already exists, as the owner of a certain Beacon, the holder of the repository, has inherently total access to it. What we propose is to carefully extend this ideal environment to the most trusted parties under some specific circumstances, e.g. to prevent some strong disease in which time could be fundamental.

5.7 Avoiding sensible metadata disclosure

If until now measures were focused on how the information should be given, the next measures, although related, are more focused into distinguishing which information should be given and which one should not. And if we are talking about data disclosure, we need first to draw a line between the two types of data we have, metadata and actual genomic data.

The metadata of a Beacon's repository is the information related to that repository. How many donors are in a dataset, their gender, age, nationality or residence would all be metadata. With the use of filters in the Beacon v2 queries, some metadata is particularly dangerous to share. Specific knowledge of someone known in real life could be used to re-identify a patient and, if a Beacon allows us to filter by street of residence, or the facility where the patient has been hospitalized, re-identification could become easier to perform. Shringarpure and Bustamante already commented on [26] how harmful publishing metadata can be: «Publishing metadata—such as the ethnicity of samples, beacon size, or the names of datasets included—reduces beacon security» (Shringarpure, S. S., & Bustamante, C. D., 2015, p. 637).
However, the biggest concern publishing metadata brings to the table is far more threatful, with enough knowledge about someone's life, and if the Beacon allows us to filter by this knowledge, an attacker could re-identify someone without being in possession of the victim's genome. This would allow this attacker to retrieve the genomic information of the victim, which would permit this attacker to perform later a regular re-identification attack of this precise victim upon other datasets, even if they were not publishing any metadata at all. This is called a genome reconstruction attack which Kerem Ayoz et al. recently proved its feasibility in [29]. In their work, they used the metadata published to successfully identify someone without having his genome, therefore acquiring his genomical information in the process. Their system was mainly based on knowing the size of the Beacon, especially before and after any update where some genomes were added or removed. In essence, if some attacker has the certain knowledge that an individual has been added (or removed) to a dataset, even if multiple individuals are added at the same time, they can easily infer which of the genomes pertains to the victim by using some basic medatadata knowledge (as they comment: eye color or hair type).

While forbidding any type of metadata disclosure can be tempting, we understand some of them are heavily linked to disease presence. We think that those traits that can seem disease

unrelated, such as street of residence, should never be shared, while also trying to avoid sharing those that feel less relevant (as eye color could be). More relevant traits should only be disclosed when some important part of the repository does exhibit them or, ideally, the whole dataset does. As we feel some metadata such as the age of the patients are indeed important, we should ensure nobody is the only person in the dataset with a specific age or, even better, concentrate people from 50 to 60 years in a single dataset. In other words, if a Beacon repository contains only people from Barcelona, for example, it is safe to share this information.

This should be the way to proceed if something that appears to be disease unrelated is actually the object of study. Street of residence, to keep our previous example, should only be disclosed if the purpose of the study is to determine if a disease is more present in that street, for example, and always if a considerable part of the dataset is conformed by donors with that particularity. It is highly unlikely the patient's street of residence has anything to do with cancer but researchers could want to know if some particular street, due being in a zone with special medioambiental characteristics, could be related to the appearance of respiratory issues. Those security measures will minimize the deduction an attacker could draw. Our proposal in this regard can be summarized as to avoid sharing metadata if possible but, if not, ensure a minimum part of the dataset shares the same metadata value. Additionally, under any circumstance an update should ever publish the number of donors that are being added or removed, and always doing so when a significant number of them will be added or removed will disable genome reconstruction shown at [29].

5.8 Identification of most vulnerable donors with the "Risk value" concept

Then we have the genomic data the Beacon protocol is designed to share, which is shared as answers to the received queries, in regular conditions, as "Yes" or a "No". This is a measure applied to avoid providing the sensible genomic information of an individual to anybody which, as commented, is a potential threat to our privacy. In general conditions owning someone else's genome does tell a very little information of that individual and, in a quite ironic way, attackers can use Beacons to help them reach phenotypic assumptions through re-identification attacks (Shringarpure and Bustamante [26]). Reading the genomic data is also still a very hard task but, as we said in the introduction, the real consequences cannot be predicted yet. Your genetic information will remain exactly

the same for your whole life, meaning its validity will not be reduced over time, as a regular message would have (a message you sent 30 years ago, for example, has more debatable validity now), and maybe in the future extracting conclusions about your genome becomes really easy.

This said, while Beacons can be used for someone to deduct something about your genome, a few things have to be noted. First of all, it does not provide directly your genome, which as we said can be the biggest threat to our privacy that can be done, second, an attacker has to be in previous possession of your genome in order to perform re-identification, which is a very strong assumption, and third, it has to be seen as price in the trade-off between utility and privacy. What this does mean is that security mechanisms have to be implemented in order to make this re-identification as difficult as possible.

Part of our proposal involves a new term, "Risk value", which gives numeric references for when is necessary to answer "You have no permission to see this information" instead of a "Yes" (or a "No", if seen under the premise of differential privacy). As we have seen in [27], unique genomic particularities represent not only a very important part of a dataset, but also the most relevant one. In our opinion, hiding all those particularities hinders too much the conclusions a scientific research party can extract, which is the main objective of this protocol. We propose the use of a certain value, called "Risk value", to help us ponder how vulnerable a donor is to a re-identification in a single Beacon dataset. A donor with a bigger assigned "Risk value" compared to another donor, is more susceptible to be re-identified than the second. Upon a certain threshold, this susceptibility would reach a security concern and, when a particular mutation of the genome is queried, and is only present in a single donor with a high "Risk value", the Beacon will answer "*You have no permission to see this information*", effectively protecting the identity of the donor without either providing false information nor letting anyone know if the answer is true or not (because some other parts will deny permission per se, as commented earlier. The proper assignment of "Risk values" is fundamental and should be based on the number of uniqueness a donor has. If compared to the other donors on the set a single individual has 20% of unique alleles, his "Risk value" will be far greater than another donor which happens to only have 5% of unique alleles. This basic idea should be the fundamental basis when assigning these values. In our opinion, protecting only the most vulnerable fraction of the donors, ideally a very small part, is actually enough, as not only re-identification attacks rely on strong previous assumptions but we also believe the united use of all the other measures proposed here are enough to ensure safety without withdrawing most of its utility.

5.9 Mutation and polymorphism distinction

We also encourage to not only validate the oddities of a donor's genome by comparing them to the rest of the dataset but comparing them worldwide too. While mutations and polymorphisms are generally treated as equal in cybersecurity works, they are separate things from the perspective of biology. Essentially, both of them are DNA variations, but they differ in how common they are compared to the rest of the population. The widely accepted threshold for these variations to be one thing or the other is 1%. A DNA variation is considered a polymorphism of an allele if it has been detected in more than 1% of the population, conversely, it is considered a mutation if it is below this number [60] [61]. We consider, then, that if we have two different DNA variations each one only found once in our dataset, and we know one corresponds to a mutation and the other one to a polymorphism, the donor that has the mutation should have a greater "Risk value". This is because he can appear to be equally identifiable in our particular dataset than the second donor, but this is not the case because an attacker who has knowledge of this fact would understand this is more revealing than the other case. Also, we should not forget our donors are probably present in more Beacon datasets than ours, and being especially unique among the entire population will make a donor more identifiable using other Beacon systems than if it is just unique in our dataset.

This proposal can be summarized as trying to distinguish between mutations and polymorphisms, providing extra protection for the former cases (in our case through an increased "Risk value"). The mechanism to make this differentiation possible falls beyond the scope of this project but could be based in some available genetic database.

The next two proposals are partially related to themes just commented, as we will try to reduce the possibilities a genetic particularity is unique in our dataset which will be a strong security feature by itself and will give us the opportunity to reduce its "Risk value", therefore having more chances to be able to share this information.

5.10 Adding donor's relatives to the dataset

One of them is based on the fact that DNA is heavily similar among our relatives, specially with first level order relatives, parents, children and siblings, being everyone way more alike with their respective relatives than with unrelated people [62] [63].

This is fulfilled with sufficient certainty to allow the search of missing persons or criminals by using their relatives' DNA [64] [65]. Based on this principle Miray Aysen et al. proposed at [66] to benefit from this by adding donor's relatives as new donors in the same dataset. In their work, they demonstrated this by experimenting with the inclusion of one or more relatives from different levels of order to the dataset, demonstrating its effectiveness. When the relatives are introduced into the dataset, all the particularities they are highly likely to share (SNPs or mutations) are garbled into the mixture of genomic information, making every related one less unique. In their words: «We show that having at least one of the parents of a victim in the beacon causes a significant decrease in the power of attacks and a substantial increase in the number of queries needed to confirm an individual's beacon membership» (Miray Aysen et al, 2020, p.1).

Lastly, they comment on how this effect is weakened when more distant relatives are used instead of closer ones, like using the victim's grandparents instead of the victim's parents. The authors compare this measure as adding control samples, which we already commented on, and point out how this would have less impact on its utility. In any case, control samples do not lose their purpose, as their advantage is that they do not share the phenotypic trait (e.g. a disease), which for some heritable traits the victim's relatives will probably do. In the special cases where relatives who exhibit the phenotypic trait the Beacon is linked to are added, we are improving both the utility and security of our system. The most inmediat drawback is that this is highly dependant on those relatives allowing and delivering their genomic data.

5.11 Ensure a minimal ratio for every continental region present

Highly related with this last one proposal, we would like to introduce the new one which aims to perform a similar effect using geographic groups. Based on the likelihood ratio typically found among geographic groups [67], the main idea will be to ensure a minimum ratio of every present geographic zone in our database. Basically, if we only have people from Europe on a Beacon, adding a single donor from Asia will make him so much more vulnerable than the others. Our proposal is to ensure a minimum ratio from Asia now that we have a donor from there. For this purpose, control samples or samples from the victim's relatives would also bring their own advantages. The *Figure* illustrates the portion of genomic information that people from each region have in common with their continents or populations. As it can be seen, the information shared through a continent but not with other continents is still significant in most of the cases.

From this, we conclude that if a reasonable ratio of presence is ensured, we can simplify the divisions at a continent level, as in most of the cases the information private to population represents a very narrow part, plus ensuring a presence ratio of every population present would highly increase the difficulty of the measure. Thus, the proposal could be summarized as, when having at least a single individual from a specific continent, a minimum presence of that continent in the dataset should be ensured while also increasing their "Risk values" as necessary. If possible, ethnicity of the samples should be considered sensible metadata, which means its disclosure should be avoided.

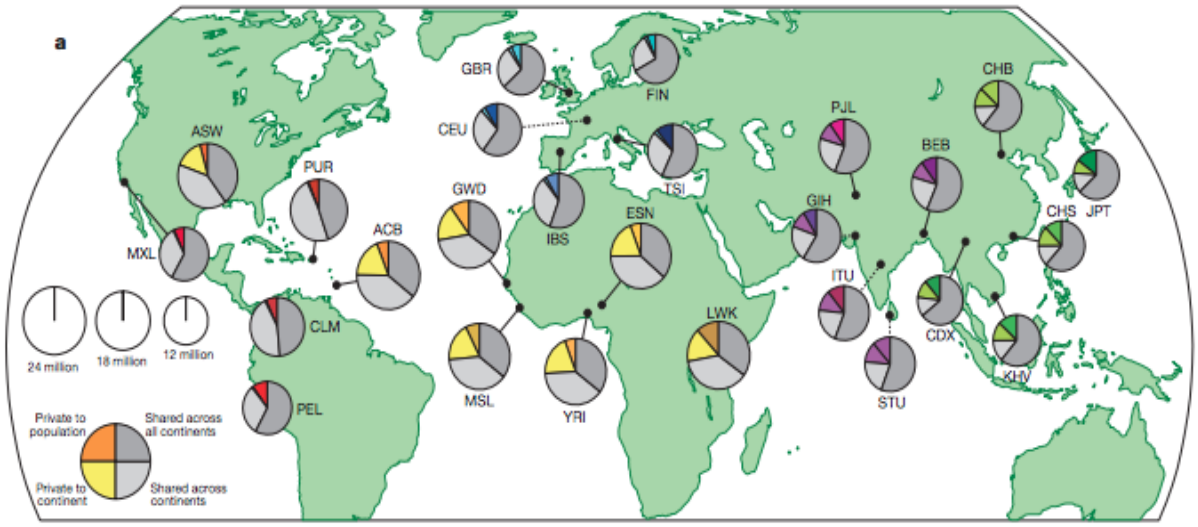


Figure 5.3: Genomic similarities by ethnicty
Here we can see, as pie charts, how much genomic information is shared among geographic distributions. Each of the four pieces every chart is made of represents the following: 1. private to population (darker colour), 2. private to continent (lighter colour), 3. shared across continental areas (lighter grey), and 4. shared across all continents (darker grey). A discontinuous line indicates the samples proceed from outside of the ancestral continental zone.
Source: [67]

5.12 Access Control and Query Budget

Finally, our last proposal is to follow the third measure Raisaro et al. explained at [27], which is to implement a query budget for the researchers. Of course, this would also bring to the table implementing a control access system, in which we also agree it would be a good improvement. Nowadays, the Beacon Network permits anonymous queries to be processed, which clearly is how any attacker would act. Adding a control access system, attackers should first bypass this measure or, if they for any case have credentials, expose their identity in the process.

The authors linked this measure to their other two proposals, which were based on differential privacy, in which case it would indirectly cause noise too. The authors used an algorithm to calculate the weight of a query and its cost, which is based on the frequency of the allele, if the researcher has not enough budget the query will not be answered and this particular donor's genome will be removed from the dataset for the future questions of this researcher (but not from the repository).

For our proposal, we will use the concept of "Risk value" to ponder this cost, making the budget decrease in a higher rate when donors with a high "Risk value" are queried and, when budget is not enough to make a query, the system would simply letting the user know he has run out of budget and no more queries can be asked. To be consistent with the general aim of our proposals of trying to preserve all the utility that is possible, we recommend avoiding a very strict policy with the queries, permitting always more than what researchers supposedly need. Re-identification attacks usually need a high number of queries and we estimate that, with the help of the other measures aforementioned, their success rate would be severely limited, meaning even a bigger number of queries is needed, which we expect surpasses easily what a normal legit research would need. Ultimately, we feel control access will discourage most of the possible attacks normally performed from anonymous access. As commented by Raisaro et al., this measure obviously carries the implementation of a complicated accounting scheme.

The initial proposal (5.1) in which we do not agree is listed here in the *Table* along with the proposals we found suitable for being a Beacon security improvement, which from (5.2) to (5.12) conjuntly form what it is our proposal. For every entry, its security advantage and utility drawback are commented on, as well as its method to achieve it. The source column indicates where this particular idea has been proposed for the first time.

Proposal	Source	Method	Security Advantage	Utility Impact	Obstacle to Performance
5.1 **Avoid** Noise-based Solutions	Analyzed proposal is from [27]	Differential Privacy, occults less common mutations.	Re-identification is harder.	Utility is greatly reduced. Most relevant information is now hidden and false information is disclosed.	Negligible.
5.2 Increase Beacon's size	From [X]	Adding more samples.	Re-identification is harder.	Utility is slightly reduced. No information is hidden.	The inherent difficulty of finding donors.
5.3 Adding control samples	From [X]	Adding more samples not related to the Beacon's linked disease, Beacon Network can be used.	Phenotypic assumptions are harder.	Utility is slightly reduced. No information is hidden.	The inherent difficulty of finding donors, in this case can be avoided by the use of the Beacon Network.
5.4 Answer "*You have no permission to see this information*" to protect rarest genomes	Original idea	Simply replacing some answers.	Protects most rare genomics data without giving false information.	Utility is slightly reduced. A very small part of the information is hidden but no false information is disclosed.	Negligible.
5.5 Strong Permission System (DUO)	Beacon v2	Using the DUO System, already in use.	Partially visible genomes are harder to re-identify.	Utility is slightly reduced. If a researcher does not have enough permissions, the information will be hidden, but no false information is disclosed.	It needs a working complex permission system, which already exists.
5.6 Use of "Trusted Environments"	Original idea	Providing total access.	No security advantage is involved, which is the reason this can only be offered to the most secure parties.	Maximum utility is ensured.	The cautious selection of a safe environment.

Table 5.1: Proposals for Beacon privacy improvement

Proposal	Source	Method	Security Advantage	Utility Impact	Obstacle to Performance
5.7 Avoiding Metadata Disclosure	From [26]	Not revealing sensible metadata.	Avoids using real-world knowledge about the victim.	Utility could be reduced. No genomic information is hidden.	The judgment of metadata's relevancy, mostly negligible.
5.8 "Risk Value" to ponderate genome's potential vulnerability	Adapted from Differential Privacy Solutions	Computing a numeric value by comparing samples to each other.	Finds the most vulnerable data to protect it.	Utility is not directly affected by this, but by the protection measure that uses this value, (5.5) in our case.	Setting up a complex and balanced system.
5.9 Distinguish mutations from polymorphisms	Original idea	Comparing DNA variations to an universal database.	Helps to ponderate worldwide vulnerability.	Utility is slightly reduced for the cases this means a high "Risk Value".	Needs a powerful tool or genomic database to work.
5.10 Adding Relatives	From [66]	Adding the closest possible donor's relatives.	Re-identification is harder.	Relatives can be relevant or not to the research and, if they are not, utility is slightly reduced.	Consent of the donor's relatives is required.
5.11 Ensure a minimum of continental region presence	Original idea	Controlling a ratio of each present ethnic group.	Re-identification is harder.	Utility is slightly reduced.	The inherent difficulty of finding donors. If control samples are accepted (5.3), the Beacon Network can be used.
5.12 Access Control & Budget Query	From [27]	Compulsory identification and setting a maximum threshold for queries.	Disable infinite anonymous queries.	Budget Query can reduce utility if researchers perform a very high number of queries.	A complex access control system needs to be implemented.

Table 5.1: Proposals for Beacon privacy improvement (continuation)

Chapter 6
Conclusions

Along this thesis we have developed two related but distinct objectives. Each one of the objectives have had different development of the ideas and has led us to different conclusions.

6.1 Systematic Review of the literature and Classification of current techniques

For this topic, we showed how critical the proper management of privacy and security is in genomic data. However, the importance of this data does not only apply in security, genomic data is greatly requested by researchers (and every day in bigger terms) for a lot of different purposes, especially health related. For this reason a secure way to share genomic information, allowing researchers to do their labor while ensuring the privacy of the donors, has become a crucial necessity of the field.

We have also shown, through a selection of studies for the Systematic Review, that it is a very present concern, widely discussed by a great number of authors. And that is also a subject with a lot of different approaches which are still progressing. By how recent were the papers or how a clear evolution could be seen from one to another, it is clear the topic is not only recent but alive, which explains the growth some of the techniques have perceived.

The purpose of our Systematic Review of defining and categorizing the techniques, alongside of stating an evolution over time is fulfilled, which serves as a key step on the posterior classification. Briefly, we introduce a scheme system for all the techniques aforementioned in the Systematic Review, and state clear differentiations between the most relevant divisions. The divisions include three different levels, being the Level I discerning techniques based on software from those based on hardware. Hardware techniques are shown to have a rather small representation which, yet not obsolete, lacks the evolution other software based techniques are receiving. Level II divides software based techniques in either Data Manipulation or Data Security.

Techniques classified as Data Manipulation, like Anonymity or Differential Privacy, will alter the data in order to ensure privacy. Data Security, otherwise, consists of those techniques meant to share sensitive data that do not modify it, keeping its integrity. Data Security techniques include Homomorphic Encryption, Multi-secure Party Computation and Access Control approaches. Access Control applications are divided as the Level III, which explains the differences among them. Some implementations will solely rely on its encryption, and some others in very different systems of rules, needing some requirements or conditions to be fulfilled before decrypting the data. Lastly, we mark how a very particular system falls in the scope of Access Control but with enough differences to need a private branch of the scheme. The GA4GH Beacon system stands out from other techniques by protecting the data by not directly sharing it, and only answering questions about the data.

Backed by the systematic review conclusions, our classification has provided consistent and clear divisions based on the behaviour of the techniques. We expect the proposed scheme to be capable of including any technique or implementation the future can uncover, and if some new characteristics appear, we are sure new branches could be drawn without altering the basic structure.

6.2 Improvement Proposals for the Beacon System

The second objective imposed was to create a conjunction of propositions that could improve Beacon's security and privacy. Those ideas should not only be able to complement each other but also allow a fair trade-off between privacy and utility, meaning the potential results a researcher could extract from the data should not be ripped off. United in a single proposal list, we have presented twelve different suggestions in chapter 5 Security Improvement Proposal for the Beacon System, one to avoid (5.1), and eleven to implement (5.2-5.12).

We realized, while performing the Beacon's review of the literature, that it is among the most discussed solutions in the community. Posterior analysis during the chapter 4 Beacon Evolution also reasserted its currentness, being a highly debated topic among researchers. By observing its evolution, standardization and upgrading process, alongside the particularity of the Beacon system operation, we decided to study and develop a proposal list of enhancement ideas to contribute to this particular technology.

We have argued that data manipulation (and by extension, Differential Privacy) should not be acceptable in health related data analysis. We think we should avoid medical false diagnosis at any cost and, as demonstrated in (5.1), it is not only a matter of ethics but a matter of utility, where the information most unique and relevant is the one most likely to be hidden, seriously hindering potential conclusions of a research. Instead of a lie that might lead to incorrect conclusions we proposed a new expression, "*You have no permission to see this information*", in (5.4), and shown how, in conjunction with a strong and granular permissions system (5.5), this cannot lead to false data disclosure. The use of this expression, assuming an attacker does not know if it is true, will avoid untruthful results that could appear using Differential Privacy. Separately, we think a flexible use of permissions gives donors a fair control over their data.

Ideally, this expression should only be used in the cases it is true the researcher lacks permits to access the data, or in the cases it is estimated it would give too much information about an individual and thus allowing a re-identification. To ponder if it is necessary or not, we propose the use of a "Risk value" (5.8), which would only protect the most sensible genomes. Instead of a regular computation of uniqueness of a certain mutation in the dataset (as Differential Privacy cases), this value could compute how compromised is a particular donor by assigning a numeric value, based on the number of uniquenesses a donor would have. Some uniquenesses of a donor can be shared if the donor has no other major identifying traits, meaning more data would be shared.

In order to assist an efficient computation of a "Risk value", we also proposed to distinguish mutations from polymorphisms (5.9), which differ in how common a genetic pattern alteration can be found in a human's genome.

In order to make it harder for a possible re-identification an attacker could perform (and therefore reducing its vulnerability and its "Risk value", which allows us to securely share more information), we have shown different propositions. Increasing Beacon's dataset size (5.2), by including more samples, is an efficient method to accomplish it. With a bigger dataset, attackers will need even larger sets of questions to ensure a re-identification, if still possible. We have also shown that, as our genetic code is greatly related to ethnicity, any donor of different ethnicity from the majority is more likely to stand out, making it more vulnerable to re-identification attacks.

We have shown how continental traits in the genome are different enough from other continents to make them more unique, and similar enough inside the continent to propose the reasonable task of ensuring a minimum representation of continental donors per dataset (5.9). This would only apply for the ethnicities already present in a dataset. Finally, we have shown that adding the data of the relatives of already existing donors (5.10), as a more precise version of the last proposition, will furtherly hinder re-identification possibilities. However, this needs careful treatment, as it could lead to the re-identification of the family.

We have also shown how adding control samples (5.3), i.e. samples that do not present the phenotype a certain dataset is linked to, we make harder a phenotype inference. In a situation where an attacker has performed a successful re-identification, we can severely limit the conclusions he could draw if he is not sure if the victim does indeed present the phenotype or not. A disadvantageous outcome of this measure is that researchers will therefore need a greater amount of questions to reach correct conclusions. Avoiding metadata disclosure (5.7) of our dataset when possible, as stated in the report, will avoid using real-life knowledge of a victim to help a re-identification attack. If not relevant for the researchers, data such as the zipcode of donors should remain hidden and inaccessible. We have also shown how this, with a careful use of additions and deletions of donor's data (never in small amounts), also helps us prevent reconstruction attacks, where an attacker that knows that someone is going to be added to a dataset can not only re-identificate victims but retrieve their genome in the process (a very dangerous possibility that could enable posterior attacks).

Implementing some sort of access control and budget query (5.7) is another part of our proposal. Even setting a generous limit of questions per user, as we do not want to potentially interfere with the research of legit users, attackers will find a harder time when accessing the system, as their attempts of re-identification will be logged and no more infinite questions are able to be made anonymously.

Finally, another significant conclusion of our thesis is that major efforts need to be made in order to make the genomic data of our donors less vulnerable, as a lot of our proposals do. But not only for security reasons but also to allow us to safely share the information with authorized users. The purpose of genomic research is often related to health conditions, which could lead to premature diagnosis of some diseases, even when no symptoms are yet present. In any case it is acceptable to lose sight of the real purpose of privacy preserving

genomic techniques, which is not to safely store genomic data but to safely share it, allowing future discoveries and conclusions. If leaning too much, in the trade-off between privacy and utility, towards utility is dangerous for the privacy of our donors, leaning too much towards privacy can potentially disable proper research, making all the process completely useless. For this reason, we also propose to create, only for the most reliable parties, the "Trusted Environments" (5.6), a safe space where no privacy is ensured and the whole information is disclosed. The purpose, as we have shown, is to ensure that, on some level, there is a very trusted party that could always be sure the best decision is made regarding both privacy and utility. This would only be made under the most faithful environment, and would not represent a common case but a very exceptional minority.

Lastly, as a final note, the re-identification attack concept as we know it was firstly introduced by Shringarpure and Bustamante in [26], and required the attacker to be in presence of the genome of the victim. For this reason is a very expanded opinion that is a very unlikely situation, and is for that reason we think it is reasonable to try to share the most information possible. Even in this case, however, the attack should never be considered impossible or unfeasible, and protection measures should always come as a part of sharing genomic data. As a last proposal, if not a thought, as the standardization of the Beacon system goes further and further, the less actual genomic data will be shared outside of the answers of questions, and more and more difficult will be for attackers to be in possession of a victim's genome.

Chapter 7
Future Work

After reviewing our efforts and results, we would like to set up what we think could be the next steps of this thesis. As we understand our project, we expect it to be able to serve as a stepping stone for different possible studies. We have separated these as ideas in the following list:

- A very useful study could be to deepen and expand the Beacon system security, ideally not only for improving genomic data treatment but also in order to adapt the system, making it viable to protect other types of data in addition to genomic data. Future im-plementations of the Beacon could be used to protect all kinds of sensitive data, even beyond the health field.

- Another possibility would be to choose another of the reviewed techniques in the section 2 Systematic Review. A similar analysis and improvement proposal could also be written, using our literature review to move forward into another direction. Almost every technique is still discussed nowadays and could benefit from a security enhancement.

- Lastly, probably the most practical possibility, and the one we would like to pursue, is to test the aforementioned propositions in the section 5 Beacon Proposal. A practical test would recognize them as helpful beyond theoretical concept, and it could also reveal which of them would have contributed the most. With this information, we could know in what aspects our efforts might be more constructive. An ideal part of this solution would also incorporate accurate and precise numbers into our proposals, e.g. as how big the size of a dataset should be in order to ensure a significant improvement (proposition 5.2), or developing an exact procedure to correctly compute the "Risk value" of our genomic data (proposition 5.8).

Finally, we would like to express, as part of the future work, what our next immediate steps will be. As a subsequent part of this thesis, we plan to release two articles from the sections 2 Systematic Review of the Literature, and 5 Beacon's Improvement Proposal. For the former, we hope our contributions can aid future efforts as a guidance in understanding, identifying and classifying privacy preserving genomic techniques. For the latter, we expect it to contribute to the privacy measures of the Beacon system, hoping some of our ideas to be implemented at some point. For this reason, besides the article that is still in process, we will present our proposal to the team at GA4GH dealing with the specification of the Beacon standard, which may lead to future improvements.

Glossary

Allele: an allele is each version of a gene. We inherit two alleles per gene, one from each of our parents, meaning we can have up to two different alleles per gene. As genes are formed by DNA code the difference between alleles is produced by the presence of mutations or polymorphisms.

Chromosome: the structures formed by the DNA code are called chromosomes. We humans carry 23 pairs of chromosomes, one from each parent.

DNA (desoxyribonucleic acid): DNA code is our unique chain of nucleotides that identify us, encoding all our biological processes that occur in our bodies. DNA is inherited from our parents and perpetuated in our descendants.

Gene: a gene is a section of the DNA code that encodes instructions for making a specific molecule, often a protein. Humans have between 20.000 and 25.000 genes, which widely vary in size, from a few hundreds nucleotides to more than a million.

Genome: the genome is the complete set of genetic instructions of an organism. The human genome is the complete DNA code of a person.

Mutation: mutations are DNA deviations away from the standard genome. Each mutation implies the existence of a different and rare allele that is less common than the standard and most common one.

Nucleotide: nucleotides are the basic code of the DNA. DNA has four different types of bases of nucleotides, guanine, adenine, cytosine and thymine.

Phenotype: the observable traits of an organism. Facial features, behaviour or a presentation of a disease in a certain individual are all phenotypes.

Polymorphism: polymorphisms are DNA variations that are commonly present in the po-pulation. Polymorphisms differ from mutations in the sense that they do appear frequently, when a variation is found in more than the 1% of the population it is called a polymor-phism.

SNP (Single Nucleotide Polymorphism): the most common type of genetic variation, SNPs are polymorphisms that apply only to a single nucleotide.

Bibliography

[1] Naro, D., Delgado, J. Security strategies in genomic files. (2020). *UPC, Departament d'Arquitectura de Computadors*. Master Thesis. http://hdl.handle.net/2117/105973

[2] McGuire, A., Fisher, R., Cusenza, P. et al. (2008). Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider. *Genetics in Medicine* 10, 495–499 https://doi.org/10.1097/GIM.0b013e31817a8aaa

[3] Naro, D. Security strategies in genomic files. (2020). *UPC, Departament d'Arquitectura de Computadors*. Doctoral Thesis. http://hdl.handle.net/2117/190249.

[4] Venter, J.C., et al. (2001). The sequence of the human genome. *Science 291(5507)*, 1304–1351. https://doi.org/10.1126/science.1058040

[5] Strachan, T., Read, A. P., & Strachan, T. (2011). *Human molecular genetics*. New York: Garland Science.

[6] Aziz, M., Sadat, M. N., Alhadidi, D., Wang, S., Jiang, X., Brown, C. L., & Mohammed, N. (2019). Privacy-preserving techniques of genomic data-a survey. *Briefings in bioinformatics, 20(3)*, 887–895. https://doi.org/10.1093/bib/bbx139

[7] Kim, M., & Lauter, K. (2015). Private genome analysis through homomorphic encryption. *BMC medical informatics and decision making*. https://doi.org/10.1186/1472-6947-15-S5-S3

[8] Lauter, Kristin & Rodríguez-Henríquez, Francisco. (2015). *Progress in Cryptology -- LATINCRYPT 2015: 4th International Conference on Cryptology and Information Security in Latin America, Guadalajara, Mexico*. Springer.

[9] Bonte, C., Makri, E., Ardeshirdavani, A. et al. (2018). Towards practical privacy-preserving genome-wide association study. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-018-2541-3

[10] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genet,* Volume 4, Issue 8. https://doi.org/10.1371/journal.pgen.1000167

[11] Bonomi, L., Huang, Y. & Ohno-Machado, L. (2020) Privacy challenges and research opportunities for genomic data sharing. Nat Genet 52, 646–654. https://doi.org/10.1038/s41588-020-0651-0

[12] Miran K., Arif H., et al. (2021). Ultra-Fast Homomorphic Encryption Models enable Secure Outsourcing of Genotype Imputation. Cell Systems12, 1–13. https://doi.org/10.1016/j.cels.2021.07.010

[13] Kamm, L., Bogdanov, D., et al. (2013). A new way to protect privacy in large-scale genome-wide association studies, Bioinformatics, Volume 29, Issue 7, pages 886–893, https://doi.org/10.1093/bioinformatics/btt066

[14] Ohata, S. (2020). Recent Advances in Practical Secure Multi-Party Computation. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences. Vol.E103-A No.10 pp.1134-1141. https://search.ieice.org/bin/summary.php?id=e103-a_10_1134

[15] Jiang, X., Zhao, Y., Wang, X., Malin, B., Wang, S., Ohno-Machado, L., & Tang, H. (2014). A community assessment of privacy preserving techniques for human genomes. BMC medical informatics and decision making, 14 Suppl 1(Suppl 1), S1. https://doi.org/10.1186/1472-6947-14-S1-S1

[16] Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006) Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S., Rabin T. (eds) Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, vol 3876. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11681878_14

[17] McSherry, F. et al. (2007). Mechanism Design via Differential Privacy. 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), pages 94-103. https://ieeexplore.ieee.org/abstract/document/4389483

[18] Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J. P., Malin, B. A., & Wang, X. (2015). Privacy in the Genomic Era. ACM computing surveys, 48(1), 6. https://doi.org/10.1145/2767007

[19] Simmons, S., Berger, B., & Sahinalp, S.C. (2019). Protecting Genomic Data Privacy with Probabilistic Modeling. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 24, 403-414. https://doi.org/10.1142/9789813279827_0037

[20] Malin, B., & Sweeney, L. (2000). Determining the identifiability of DNA database entries. Proceedings. AMIA Symposium, 537–541.

[21] Malin B. A. (2005). 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 : JAMIA, 12(1), 28–34. https://doi.org/10.1197/jamia.M1603

[22] Li, N., Qardaji, W., Su, D. (2012). On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. https://doi.org/10.1145/2414456.2414474

[23] Malin B. A. (2005). Protecting genomic sequence anonymity with generalization lattices. *Methods of information in medicine*, 44(5), 687–692.

[24] Scheibner, J., Raisaro, J. L., Troncoso-Pastoriza, J. R., Ienca, M., Fellay, J., Vayena, E., & Hubaux, J. P. (2021). Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis. *Journal of medical Internet research, 23(2*). https://doi.org/10.2196/25120

[25] Chen, F., Dow, M., Ding, S., Lu, Y., Jiang, X., Tang, H., & Wang, S. (2017). PREMIX: PRivacy-preserving EstiMation of Individual admiXture. AMIA ... Annual Symposium proceedings. AMIA Symposium, pages1747–1755.

[26] Shringarpure, S. S., & Bustamante, C. D. (2015). Privacy Risks from Genomic Data-Sharing Beacons. *American journal of human genetics, 97(5*), 631–646. https://doi.org/10.1016/j.ajhg.2015.09.010

[27] Raisaro, JL., Tramèr, F., Ji, Z., et al. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks, *Journal of the American Medical Informatics Association*, Volume 24, Issue 4, July 2017, Pages 799–805, https://doi.org/10.1093/jamia/ocw167

[28] Wan, Z., Vorobeychik, Y., Kantarcioglu, M. et al. (2017) Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services. *BMC Med Genomics 10, 39*. https://doi.org/10.1186/s12920-017-0282-1

[29] Ayoz, K., Ayday, E., Cicek, A.(2021).Genome Reconstruction Attacks Against Genomic Data-Sharing Beacons. *Proceedings on Privacy Enhancing Technologies*, volume 202, issue 3, p. 28-48. https://doi.org/10.2478/popets-2021-0036

[30] Fiume, M., Cupak, M., Keenan, S. et al. (2019). Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol 37*, 220–224. https://doi.org/10.1038/s41587-019-0046-x

[31] Senf, A., Davies R., Haziza, F., et al.(2021). Crypt4GH: a file format standard enabling native access to encrypted data, *Bioinformatics*, Volume 37, Issue 17, p. 2753–2754, https://doi.org/10.1093/bioinformatics/btab087

[32] Morteza, H., Pratas, D., et al. (2019) Cryfa: a secure encryption tool for genomic data, *Bioinformatics*, Volume 35, Issue 1, p.146–148, https://doi.org/10.1093/bioinformatics/bty645

[33] Huang, Z., Ayday, E., Lin, H., Aiyar, R. S., Molyneaux, A., Xu, Z., Fellay, J., Steinmetz, L. M., & Hubaux, J. P. (2016). A privacy-preserving solution for compressed storage and selective retrieval of genomic data. *Genome research,* 26(12), 1687–1696. https://doi.org/10.1101/gr.206870.116

[34] Global Alliance for Genomics and Health. (2019). Data Use Ontology approved as a GA4GH technical standard. *Ga4gh.org*. https://www.ga4gh.org/news/data-use-ontology-approved-as-a-ga4gh-technical-standard/

[35] *EBISPOT/DUO*. GitHub. https://github.com/EBISPOT/DUO#what-is-duo

[36] Cabili, M., Carey, K., Dyke, S. et al. (2018). Simplifying research access to genomics and health data. *Scientific Data*. https://doi.org/10.1038/sdata.2018.39

[37] Global Alliance for Genomics and Health. (2019). GA4GH Passports and the Authorization and Authentication Infrastructure.Ga4gh.org.https://www.ga4gh.org/news/ga4gh-passports-and-the-authorization-and-authentication-infrastructure/

[38] GitHub. ga4gh-duri/ga4gh-duri.github.io. https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md#passport

[39] GitHub. ga4gh/data-security. https://github.com/ga4gh/data-security/blob/master/AAI/AAIConnectProfile.md

[40] Delgado, J., Llorente, S., Naro, D. (2017) Protecting privacy of genomic information. "Studies in health technology and informatics", vol. 235, p. 318-322. http://hdl.handle.net/2117/106545

[41] Mittos, Alexandros & Malin, Bradley & De Cristofaro, Emiliano. (2019). Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective. Proceedings on Privacy Enhancing Technologies. https://doi.org/10.2478/popets-2019-0006

[42] Marino F. and Hubaux JP. (February 19, 2021). CWS APIs for supporting next generation secure workflows [Slide 11].

[43] Global Alliance for Genomics and Health. (2015). Beacon project mitigates privacy risks while maximizing value of responsible data sharing.Ga4gh.org.https://www.ga4gh.org/news/beacon-project-mitigates-privacy-risks-while-maximizing-value-of-responsible-data-sharing/

[44] Global Alliance for Genomics and Health. (2017). ELIXIR and GA4GH beacon team up to advance genomic data sharing.Ga4gh.org. https://www.ga4gh.org/news/elixir-and-ga4gh-beacon-team-up-to-advance-genomic-data-sharing/

[45] The ELIXIR Beacon-2017.(2017). Elixir-europe.org https://elixir-europe.org/about-us/commissioned-services/beacons

[46] Global Alliance for Genomics and Health. (2017). New strategies for securing Beacon datasets.Ga4gh.org. https://www.ga4gh.org/news/new-strategies-for-securing-beacon-datasets/

[47] Morgan, S. (2021). Top 5 Industries At Risk Of Cyber-Attacks. Forbes. https://www.forbes.com/sites/stevemorgan/2016/05/13/list-of-the-5-most-cyber-attacked-industries/?sh=21c627f1715e

[48] Global Alliance for Genomics and Health. (2017). A new access tier for genomic and health-related data. Ga4gh.org. https://www.ga4gh.org/news/a-new-access-tier-for-genomic-and-health-related-data/

[49] Dyke, S.O.M., Linden, M., Lappalainen, I. et al. (2018). Registered access: authorizing data access. European Journal of Human Genetics 26, 1721–1731. https://doi.org/10.1038/s41431-018-0219-y

[50] Global Alliance for Genomics and Health. (2018). GA4GH and ELIXIR Release Beacon API v1 with increased security measures. Ga4gh.org. https://www.ga4gh.org/news/ga4gh-and-elixir-release-beacon-api-v1-with-increased-security-measures/

[51] Elixir. (2018). GA4GH and ELIXIR Release V1.0.0 of Beacon API with increased security measures. Elixir-europe.org. https://elixir-europe.org/news/beacon-API-release

[52] Global Alliance for Genomics and Health. (2019). Extensions to the GA4GH Beacon API will enable a more powerful community resource. Ga4gh.org. https://www.ga4gh.org/news/extensions-to-the-ga4gh-beacon-api-will-enable-a-more-powerful-community-resource/

[53] Beacon Network. https://beacon-network.org

[54] Baudis, M. Beacon v2 - Towards flexible use and clinical applications for a reference genomic data sharing protocol [PDF Presentation]. Department of Molecular Life Sciences, University of Zurich. https://info.baudisgroup.org/pdf/2020-06-30___Michael-Baudis__Beacon-evolution__NEXUS-PHRT2020-slides.pdf

[55] Laurent, F. (2021). Beacon cohorts: A model for cohort discovery in CINECA and beyond. CINECA. https://www.cineca-project.eu/blog-all/beacon-cohorts-a-model-for-cohort-discovery-in-cineca-and-beyond

[56] Global Alliance for Genomics and Health. (2018). GA4GH Connect: A 5-Year Strategic Plan. Ga4gh.org. https://www.ga4gh.org/wp-content/uploads/GA4GH-Connect-A-5-year-Strategic-Plan.pdf

[57] Aziz, M., Ghasemi, R., Waliullah, M. et al. (2017). Aftermath of bustamante attack on genomic beacon service. BMC Med Genomics. https://doi.org/10.1186/s12920-017-0278-x

[58] Medline Plus. (2020). What are single nucleotide polymorphisms (SNPs)?. https://medlineplus.gov/genetics/understanding/genomicresearch/snp/

[59] Demmler, D., Hamacher, K., Schneider, T., & Stammler, S. (2017). Privacy-Preserving Whole-Genome Variant Queries. CANS. pp 71-92. https://doi.org/10.1007/978-3-030-02641-7_4

[60] Karki, R., Pandya, D., Elston, R. C., & Ferlini, C. (2015). Defining “mutation” and “polymorphism” in the era of personal genomics. BMC medical genomics, 8, 37. https://doi.org/10.1186/s12920-015-0115-z

[61] Twyman, R. (2003). Mutation or polymorphism?. Western Washington University, Biology Department. https://fire.biol.wwu.edu//trent/trent/polymorphism.pdf

[62] C C Li et al. (1993). A: Similarity of DNA Fingerprints Due to Chance and Relatedness. Human Heredity. https://doi.org/10.1159/000154113

[63] Chakraborty, R., & Jin, L. (1993). Determination of Relatedness between Individuals Using DNA Fingerprinting. Human Biology, 65(6), 875–895. http://www.jstor.org/stable/41464929

[64] Ge, J., Budowle, B. and Chakraborty, R. (2011), Choosing Relatives for DNA Identification of Missing Persons. Journal of Forensic Sciences, 56: S23-S28. https://doi.org/10.1111/j.1556-4029.2010.01631.x

[65] Bieber F. R. et al. (2006). Finding Criminals Through DNA of Their Relatives. SCIENCE, Vol 312, Issue 5778, pp.1315-1316. http://dx.doi.org/10.1126/science.1122655

[66] Huang, Z., Ayday, E., Lin, H., Aiyar, R. S., Molyneaux, A., Xu, Z., Fellay, J., Steinmetz, L. M., & Hubaux, J. P. (2016). A privacy-preserving solution for compressed storage and selective retrieval of genomic data. *Genome research, 26(12)*, 1687–1696. https://doi.org/10.1101/gr.206870.116

[67] Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393