Privacy and Security in Genomic Information

Sara Rodríguez Cubillas

Advised by

Jaime Delgado

MIRI Master Thesis

Barcelona 2016

Acknowledgements

I would like to thank my advisor, Jaime Delgado, for all his support, guidance words of encouragement and unwavering patience.

I would also like to thanks Silvia Llorente for her valuable feedback on the project, as well as Daniel Naro for the crash course genomic privacy.

Special thanks also go to the Secure GENomic information COMpression research group for providing additional technical guidance for this project.

Contents

Abstract

Today, whole human genome sequencing is a reality affordable for many individuals. Technology advances in this arena are going fast, in fact it is possible that we are living the beginning of a “genomic revolution”. However, human genome contains highly sensitive information about individuals. Concerns regarding privacy and security are getting wider as technology advances ever more rapidly. The present work focuses on contributing and support the development of new security and privacy mechanisms in genomic information formats. An intended software prototype has been developed and it is presented in the current report. The tool permits reading and handling genomic information after evaluating the access requirements according to defined rules.

1 Introduction

1.1 Introduction and structure

The development of Next Generation Sequencing (NGS) technologies opens the usage of genomic information. However appropriate representation and an efficient compression of genomic data is widely recognized as a critical element limiting its application potential [1] [2] [3]. In addition to compression other requirements has been identified [4], as those regarding privacy and security.

An international standardization work has been initiated, see the Joint Call for Proposals described in [1], in which this project is involved actively since his inception. This work aims at contributing to the definition of a new file format for genomic data, see the proposal at [5], focusing on the privacy and security aspects.

Personal genomic information has a number of characteristics, which are described in detail in section 2 “The genome”, that make it important to protect. Technology should make it possible for individuals to exercise their legal rights concerning their genomic information. This rights includes access to the content, who should be able to read or do operations on it and under what circumstances. The focus on this report is in the study on how to achieve this goal.

The rest of the report is structured as follows: First, background on the context of this “genomic revolution” is given as well as some basic terminology to have a better understanding; Section 2 "The genome" includes a more in detail explanation of what the Deoxyribonucleic acid (DNA) is, which are his characteristics and the main applications that it has. Moreover, in this section some ethical aspects and controversy are pointed out which are important to understand why this information is so critical; This is followed by “Privacy and security in genomic information”, Section 3, where it is explained relevant privacy and security aspects that for the nature of the project act as requirements. Related with this aspect also it is presented a list of use cases that will serve us for developing the solution proposed thus justifying this work; After that in Section 4 a complete analysis of standards and tools is done. It introduces the currently used formats and tools as well as the new proposal on which the Moving Picture Experts Group (MPEG) is working, as they are the ones that will be used on the development of the proposal; Then the contribution is carefully exposed, explaining where the proposal fits within the actual genomic information generation and manipulation flow. This is the main section, it gives details of the implementation of the software developed as proof of concept of the proposal and at the end explains how you can test it. Conclusion and further work is presented in the last section.

1.2 Background

To place the DNA in its context, genetics is a science without history. Before Mendel, less than a century and a half ago, there was nothing. Genes were discovered through crosses between mice, flies or mushrooms, and observing the distribution of traits among their offspring. DNA seemed to be a very elementary thing. It had only four chemical subunits -repeated many times in a long chain.

The idea that something so simple could be the instrument of the inheritance had to wait until 1944. Then it became possible to change the appearance of certain bacteria colonies by treating them with DNA extracted from other colonies with different form. The astonishing thing was that this modification was inherited. Information was transmitted from one generation to another through DNA. Although nobody understood how.

In the early 1950s, American biologist James Watson and British physicist Francis Crick proposed their famous double-helix DNA model. Some of his most important clues about DNA structure were the work of Rosalind Franklin, a chemist who works in the laboratory of physicist Maurice Wilkins. The model and the discoveries that made it form the foundation of much of today's cutting-edge research in biology and biomedicine.

Figure 1. The International Human Genome Project (IGHP) published its draft sequence in Nature on Feb .15, 2001 and Celera Genomics published its draft sequence in Science on Feb. 16, 2001

At the very beginning of the 21st century, a complete sequence was published at 90 percent of the three billion base pairs in the human genome.

The whole genome sequencing (WGS) contained unexpected findings. Men and chimpanzees share 99 percent of their genes. Hereditary human diseases are found in mice, cats and dogs. Genes that control the fundamental processes of life, such as cell division, are similar even in creatures as different from each other as the human being and yeast. What makes a human embryo develop as a man, rather than as a woman, have only a couple of hundred bases. Other equally spectacular effects (which, for example, make a fly grow a couple more wings) are also very simple.

The advances on this field have led to substantial reductions in the cost of genome sequencing and the WGS will soon become affordable for many individuals.

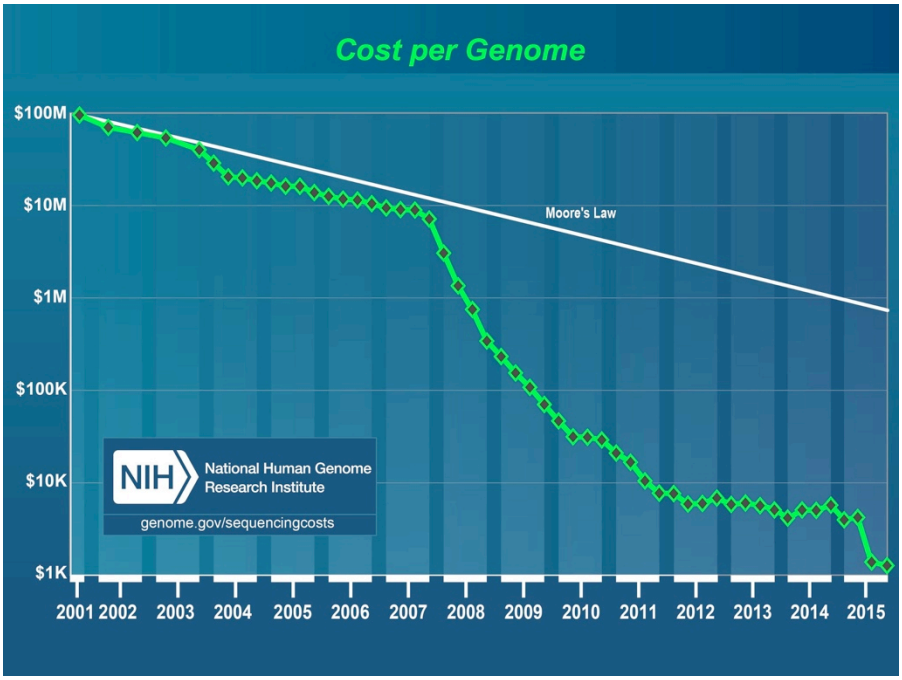

Figure 2. Cost per genome by year chart, by the National Human Genome Research Institute

In this situation in which individuals become consumers of the technology and can generate their own data, WGS stops being for a research oriented only environment. This type of data can be used in a wide range of application for different aspect as ancestry investigations, specific treatment and its relations to diseases. This progress prompts to new important challenges concerning privacy and ethical issues.

1.3 Terminology

A human being has 46 chromosomes in the nucleus of each cell, coming in 23 homologous pairs (22 pairs of autosomes and one pair of sex chromosomes (X/Y)). Of each pair one chromosome comes from the mother and one from the father.

All genetic information of an individual constitutes his/her genome.

A sequence read is the readout, by a specific technology prone to errors, of a continuous part of a segment of DNA extracted from an organic sample.

> 3,000,0000,000 **base pairs** (G, C, A, T). Humans share 99.9 % of DNA sequences.

A sequence alignment is sequence read mapped on a reference DNA sequence.

Human genome has more than 20,000 genes.

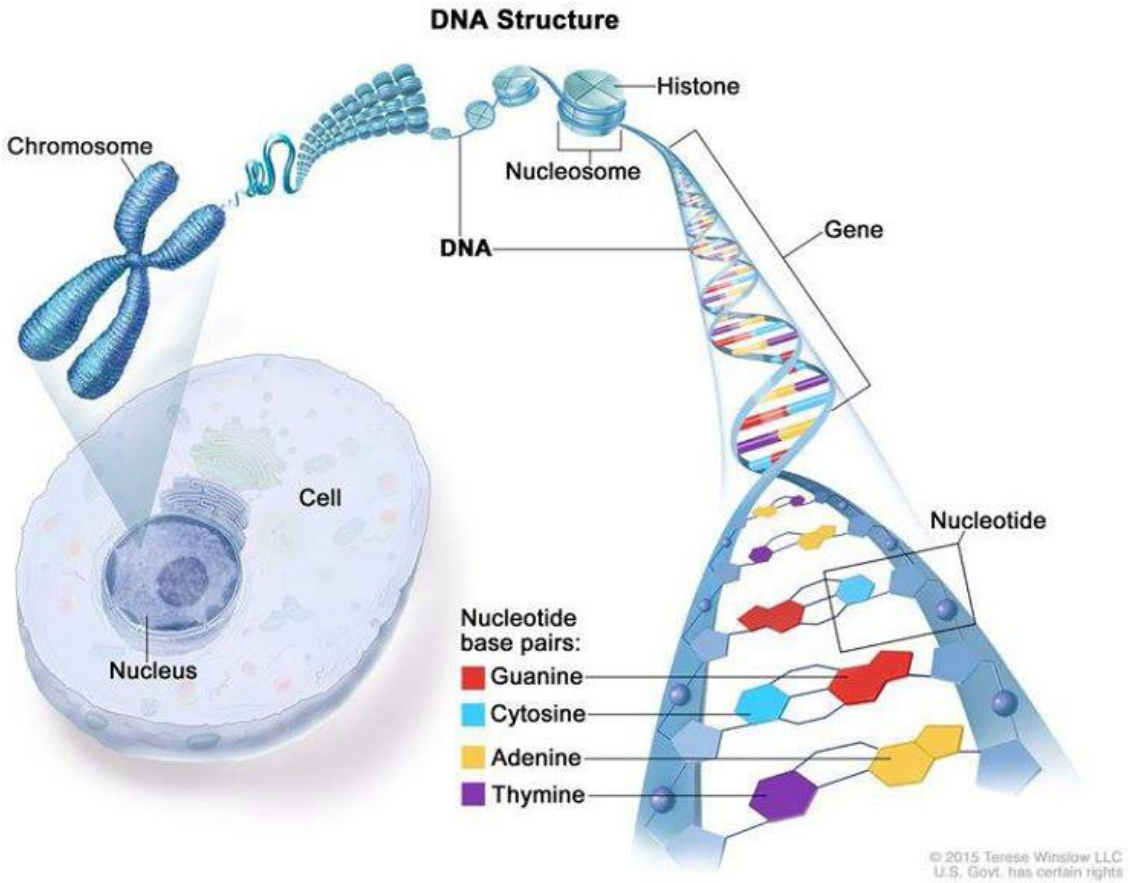

Figure 3. DNA Structure. Image from https://www.cancer.gov/about-cancer/causes-prevention/genetics Credit: Terese Winslow

A **gene** corresponds to a piece of a chromosome, a stretch of DNA.

The location of a gene in the genome is called its **locus**.

Everyone inherits two copies of each gene, one from the father and one from the mother.

The different forms of a gene are called **alleles**, if there are two often indicated by A and a, A1 and A2, or A and B.

The genetic makeup of an individual is called his/her **genotype**. For a gene with two alleles, this can be AA, Aa or aa.

Biomarker is a genetic variable that shows variation over individuals and has a known locus, it can be use as an indicator of the presence of a disease state or any other physiological state.

Single nucleotide polymorphism **SNP** is a position in the genome sequence with a nucleotide that varies between individuals.

2 The genome

2.1 Characteristics

In [6], the authors established some characteristics of the DNA that are relevant to consider:

Uniqueness

The DNA is unique to each individual. Genetic/genomic information could be easily be used for individual identification purposes.

Predictive capability

By combining the genetic information with other factors as environmental conditions and lifestyle of an individual it is possible to infer the individual's phenotype, including the predisposition to developing a given disease or the response to a specific treatment.

Moreover, as scientific understanding of the relationship between genotype and phenotype increases, genetic/genomic information may be used more accurately to predict an individual's physical characteristics from his/her DNA sequence information.

Immutability

Genetic information is also immutable; an individual's inherited information does not change throughout life. As such, public disclosure of personal genetic/genomic test information could create long lasting and unpredictable effects, given unforeseen technological and interpretive advances.

Requirement of testing

According to the authors, test for disease predisposition and drug response must be derived from a genetic/ genomic test and cannot be ascertained in the normal course of clinical care.

Historical misuse

There are already examples of the use of genetic information to promote eugenics initiatives, discriminate in insurance and the workplace, and obtain information about individuals' medical histories.

Impact on family

An individual's decision to share his genomic information may reveal blood-relatives genomic information as well. This is a new type of privacy threat.

Temporality

The scope and ability to interpret test results is evolving rapidly, which means that with the same information in the future we will be able to prove more detailed conditions and/or therapeutic responses. Changes on social perspectives on how to oversee this genetic information will also evolve over time.

Ubiquity and ease of procurement

The whole genome of an organism can be revealed from tissue samples (saliva, blood, hair, etc). Thus, genetic material can be easily procured, also without the individual's permission.

2.2 Applications

Diagnostics and Personalized Medicine

WGS has the potential to bring about a new era of predictive, preventive, participatory, and personalized (P4) medicine. A medicine that focuses on the integrated diagnosis, treatment and prevention of disease in individual patients [7].

Genome-based diagnostic tests have been recently developed to make personalized treatment possible thanks to discovered links between specific genetic variants and diseases. Such tests have the potential to predict risk and drive preliminary therapeutic interventions, to detect onset of disease, or detect residual disease. Preventive efforts and improved health is a major goal of genomic research. [8]

Pharmacogenomics

The combination of the science of how drugs work, called pharmacology, with the science of the human genome, called genomics, for the clinical development of new and existing drugs is known as pharmacogenomics.

Patients and doctors with the genetic information needed could predict diseases, help prescribe optimized treatment that work better and prevent harmful side effects avoiding adverse drug interactions of the tri- al-and-error period characteristic of traditional drug evaluation.

Examples of pharmacogenomics include testing for a genetic variant that makes people infected with the human immunodeficiency virus (HIV) more likely to have a bad reaction the antiviral drug abacavir (Ziagen).

In the case of leukemia some people have a genetic variant that interferes with their ability to process the medicine. This processing problem can cause severe side effects and increase risk of infection, unless the standard dose is adjusted according to the patient's genetic makeup.

Another example is the breast cancer drug trastuzumab (Herceptin). This therapy works only for women whose tumors have a particular genetic profile that leads to overproduction of a protein called HER2 [9, 10].

Because each person responds differently to medicines pharmacogenomics is soon expected to lead to better ways of using drugs to manage heart disease, cancer, asthma, depression and many other common diseases according to genetic and specific individual markers.

Testing for hereditary diseases risk

One of the most direct applications to know the sequence of genes that make up the human genome is that you can know the molecular basis of many genetic diseases and can make an appropriate diagnosis. These types of diagnoses can be performed in presymptomatic and even in a prenatal way.

As a consequence of the development of in vitro fertilization techniques, today it is possible to perform the so-called preimplantation genetic diagnosis (DGPI). This allows to test the embryos from a genetic and chromosomal point of view in order to choose the one that is healthy and to implant it in the mother's uterus. The DGPI avoids the gestation of a genetically or chromosomally affected child, and implies the parents' decision to perform, if necessary, a therapeutic abortion.

Paternity and ancestry testing

By comparing the DNA of different people it is possible to check their relationships. The Y chromosome is passed almost unchanged from father to son.

There are already commercial ancestry and genealogical testing products in which software compares an individual’s genomic information to other strands and identify whether or not it is related to another person or ethnic group.

2.3 Ethical aspect and controversy

An important gap is thus perceived between the diagnostic and predictive capability that WGS brings and the lack of preventive and therapeutic interventions on the other, leading to ethical conflicts.

The preventive diagnosis of a disease in a way could be an emotional burden that the patient should endure in the best possible way, coexisting with the impotence and anxiety that can mean to a patient knowing that in a certain period of time is possible to suffer a disease.

With the prenatal diagnosis is possible to know both, diseases that develop from the first day of life of the individual and diseases that can appear in his advanced age, as Alzheimer, for

example. In that case, would we abort a fetus that can present Alzheimer's disease almost at the end of its life? Or that he had the disease of color blindness?

The idea of genetic discrimination is not new in 1997 Gattaca, an American science fiction film written and directed by Andrew Niccol, presents a society driven by eugenics where potential children are conceived through genetic manipulation to ensure they possess the best hereditary traits of their parents [11]. This sterilized world has led to severe discrimination against the genetically unmodified that can be compared with racism.

Another aspect to consider is the 'gene patent'. Can genes be patented or is it considered a World Heritage?

In both the United States and the European Union, programs have been developed to address the ethical and social consequences of scientific research and to avoid conflicts. In the United States is the ELSI (Ethical, Legal and Social Implications Research Program), meanwhile outside of them is the Universal Declaration on the Human Genome and Human Rights, promoted by UNESCO.

3 Privacy and security in genomic information and Identification Project

The special characteristics of the DNA that has been previously mentioned as uniqueness, his predictive nature, immutability, the requirement of testing and so on are critical yet complex considerations make security and privacy elements a key aspect when providing mechanisms for generating, processing, storing and transmitting it. For this reason, it has been identified relevant privacy aspects that will be converted into requirement at the time developing this work.

3.1 Requirements for the project as privacy aspects

Granularity

Required granularity for genomic information. Inside this category, we can find different levels of granularity: Complete file, one chromosome, range of regions or even one specific position. It also depends on the kind of genomic information being accessed: sequenced or aligned information. Other genomic information formats, like variant/genotype information stored in a variant calling file could be also considered in this category.

Variability over time, territory and situation

Social perspectives about the variety of genomic utilities will likely evolve over time as well as it will differ among different cultures. In the other hand the ability to interpret genomic information will increase with the research advance, meaning that the same genetic data may reveal new information in the future.

There is another aspect related to the moment of asking for permission, as is the case of an emergency, this could lead to a “break glass scenario”.

Variety of roles

There are different types of roles accessing the information, patient, healthcare-provider, genetic researcher, laboratory, employee, physician, health insurance provider, etc.

Variety of usages

Usage of genomic information. Inside this category, we foresee the definition of different purposes: Full control, research, characterization of individual for a given biological feature, genetic analysis, lineage search, commercial, forensics, etc.

Access notification

Provide information to genomic information owner (data sharer). Here, we identify the possibility to inform of the result of the study to the data sharer.

3.2 About securitiy

Concerning security, multiple strategies have been devised in academia for protecting genomic information. For example, one way that people handle sensitive data currently is to encrypt the entire file. This then forbids any random access on an encrypted stream and it requires a local unencrypted file, which inherently becomes a security risk. One solution is to divide the information into blocks and encrypt them independently, presumably with some salt to stop some of the attack forms. This could make it possible to do random access on the encrypted files, perhaps even permitting remote browsing, although it could be a whole new can of worms with the authorization. There are many more aspect to have into account regarding security the index for example either needs to be stored unencrypted (so coverage data can be gleaned) or for it to be encrypted but entirely decrypted prior to processing. There are almost certainly timing attacks too. However, based on the nature of this project, we propose only focus on encrypting the file.

3.3 Use cases

This section presents a list of use cases where the privacy aspects concerning granularity of information, persons with granted access, genomic information usage and if it is required to inform of the results to the data owner are combined.
The use cases are organized into five different types based on the initiator role involved, grouping several aspects. So, first type of use cases groups those where the permission to access the genomic information is given by an individual. The second type includes those use cases where permission is requested by a data analyst. The third type involves healthcare professionals, the fourth type describes those use cases where researchers are involved and there is a fifth type for other kind of use cases. It is worth noting that other groupings are possible, this is just a way of organizing them.

Use cases where the permission is given by an individual:

1.1 An individual wants to share the variant/alignment/sequence information of specific genomic regions and experiment with an analyst, so that he/she performs a genetic analysis.

1.2 Analogously to 1, but combining data of several experiments.

1.3 An individual wants to share the non-aligned reads from a given experiment with an

analyst, so that he/she performs a genetic analysis.

1.4 An individual wants to share variant/alignment/sequence information genome wide for a given experiment with an analyst, so that he/she performs a genetic analysis.

1.5 Analogously to 4 but combining data of several experiments.

1.6 An individual wants to share his variant/alignment/sequence genome wide data with a researcher for a given project. He/she wants to be re-contacted back if necessary.

1.7 Analogously to 6, but he/she wants to do this anonymously.

1.8 An individual wants to donate his/her genome for research when he/she dies.

1.9 An individual wants to share access of his whole genome with his trusted healthcare professional for an undetermined amount of time, being able to cancel that access whenever he/she likes.

Use cases where the permission is requested by an analyst:

2.1 An analyst wants to access the variant/alignment/sequence information of specific genomic regions from a given individual and experiment.

2.2 Analogously to 2, but combining data of several experiments.

2.3 An analyst requires access to non-aligned reads from a given individual.

2.4 An analyst wants to access the variant/alignment/sequence information genome wide from a given individual and experiment.

2.5 Analogously to 4 but combining data of several experiments.

Use cases where a healthcare professional is involved:

3.1 An analyst wants to share with a healthcare professional the outcome of a genetic analysis.

3.2 A healthcare professional wants to access the genetic analysis of an individual.

Use cases where the permission is requested by a researcher:

4.1 A researcher wants to access anonymized genomic data for a given project.

4.2 A researcher wants to ask an individual for more experiments to be used in a research study.

Other use cases:

5.1 Paternity test

5.2 Forensic use

Table 1. Use case summary

Use Case	Role giving permission	Role receiving permission	Role Initiative	Object of the permission	Permission	Experiments number	Inform owner?
1.1	Individual	Analyst	Giving	Regions of Variant / Alignment / Sequence	Genetic analysis	One	Yes, results
1.2	Individual	Analyst	Giving	Regions of Variant / Alignment / Sequence	Genetic analysis	Several	Yes, results
1.3	Individual	Analyst	Giving	Non-aligned reads	Genetic analysis	One	Yes, results
1.4	Individual	Analyst	Giving	Complete genome Variant / Alignment / Sequence	Genetic analysis	One	Yes, results
1.5	Individual	Analyst	Giving	Complete genome Variant / Alignment / Sequence	Genetic analysis	Several	Yes, results
1.6	Individual	Researcher	Giving	Complete genome Variant / Alignment / Sequence	Research project	One	Yes, re-contact
1.7	Individual	Researcher	Giving	Complete genome Variant / Alignment / Sequence	Research project	One	No, anonymous
1.8	Individual	Everyone	Giving	Complete genome Variant / Alignment / Sequence	Donation	-	No, it is donated after individual's death
1.9	Individual	Healthcare	Giving	Complete genome Variant / Alignment / Sequence	View	-	-
2.1	Individual	Analyst	Receiving	Regions of Variant / Alignment / Sequence	View	One	-
2.2	Individual	Analyst	Receiving	Regions of Variant / Alignment / Sequence	View	Several	-
2.3	Individual	Analyst	Receiving	Non-aligned reads	View	One	-
2.4	Individual	Analyst	Receiving	Complete genome Variant / Alignment / Sequence	View	One	-
2.5	Individual	Analyst	Receiving	Complete genome Variant / Alignment / Sequence	View	Several	-
3.1	Analyst	Healthcare	Giving	Outcome genetic analysis	Share	One	-
3.2	Individual	Healthcare	Receiving	Genetic analysis	View	One	-
4.1	Owner/Custodian	Researcher	Receiving	Anonymized genomic data	Research project	One/Several	No, anonymous
4.2	Individual	Researcher	Receiving	More experiments	Research project	Several	Yes, request more data
5.1	Individual	Analyst	Giving	Perform paternity test	Paternity test	One	Yes, results
5.2	Owner	Forensics	Receiving	Complete genome	Forensic	One/Several	Unknown

Table 1 Summarizes the information contained in each use case, indicating which is the role that should give the permission, the role receiving it, who has the initiative when requesting permission (giving or receiving), the object of the permission to be given (regions, complete genome, etc.), which permission is requested, the number of experiments involved and if the owner of the genome should be informed. For the use case 1.9, there is a condition that is that the permission is given for an undetermined amount of time, but the individual can cancel access at any time. It has not been added for clarity.

There is no standard yet defined on how assure privacy and security in the uses cases citated above. Clearly, privacy concerns represent a formidable obstacle to assembling large human genomic databases and can delay (or derail) genome-wide association studies, which in turn could thwart advances in medicine and subsequent healthcare improvements. [12]

This project is a collaboration to solve the challenges prompted by the conflict between the need to facilitate genomics research and that to protect privacy of individuals.

4 Analysis of standards and tools

4.1 File Formats

Nowadays the largest majority of public repositories of sequence data provide data formatted in two - very similar - textual file formats named FastA and FastQ. FastQ exists in a few different flavours defined by different sequencing machine vendors. FastA and FastQ have been adopted in the recent past when the amount of generated information was not so important to create any issue of storage space. [8]

The simplicity this format makes it easy to manipulate and parse sequences using text-processing tools and scripting languages like Python, Ruby, and Perl.

Unmapped data. FastA/FastQ

fastA

It only contains two lines the header starting with ">"(greater-than) and followed by the sequence names optional comments and in the second line the raw sequence should be not longer than 120 and usually shorter than 80 characters in length. Sequence may be protein sequences or nucleic acid sequences are represented using single-letter codes.

The symbols used to represent nucleotides are [10]:

A	Adenine	K	T, U, or G
C	Cytosine	W	T, U, or A
G	Guanine	S	C or G
T	Thymine	B	C, T, U, or G (not A)
U	Uracil	D	A, T, U, or G (not C)
R	Purine (A or G)	H	A, T, U, or C (not G)
Y	Pyrimidine (C, T, or U)	V	A, C, or G (not T, not U)
M	C or A	N	Any base (A, C, G, T, or U)

Table 2. Nucleic acid symbols

The amino acid codes are:

A	Alanine	H	Histidine	T	Threonine
R	Arginine	I	Isoleucine	W	Tryptophan
N	Asparagine	L	Leucine	Y	Tyrosine
D	Aspartic acid	K	Lysine	V	Valine
C	Cysteine	M	Methionine	B	Aspartic acid or Asparagine
Q	Glutamine	F	Phenylalanine	Z	Glutamine or Glutamic acid
E	Glutamic acid	P	Proline	X	Any amino acid
G	Glycine	S	Serine		

Table 3. Amino acid symbols

```
>gi|528476511|ref|NW_004929286.1| Homo sapiens chromosome 1 genomic scaffold,
alternate assembly CHM1_1.1, whole genome shotgun sequence
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCAAC
CCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGGTCTG
ACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTCCGCC
CTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCACCTCCGCCGGCGCAGGCGCAGAGAGGCGCGCC
GCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAG
```

Figure 4. Head of a fastA document containing the chromosome 1 of the Homo sapiens

```
>gi|164682561|ref|NC_010247.1| Amapari virus segment S, complete sequence
CGCACAGTGGATCCTAGGCGCAATTGTTCACGCAATTTTGCATATCTATTAACAATTGATCATGGGTCAA
CTTGTTAGTTTCTTTCAGGACATACAACTCTTTTTCCAAGAGGCTCTCAATGTGGCTCTAGCTGTTGTCA
CCGTCCTGGCTATCATTAAGGGGCTTGTGAACCTTTGGAAATCAGGTCTCTTCCAGTTCCTCTTTTTCTT
AATCCTGGCAGGAAGGTCTTGCTCCTTCAGAATTGGTCATCATACAACTTTTGAATCTGTCACAATGTCA
GTTGGGGGAGTCTTTCACGAACTTCCTGCTTTGTGCAGGATCAATAACAGCCACAGTCTAATTCAACTGT
CTCACAACAGTAGTTTAGCTTTGTCTGTCGAATATGTGGATTTGTGTGTCGTTCTAGAGTCAGACCAGTA
TTTAGTGGCTGGGGATTACTCCAACTGTACTGGGGAAGCGACAGGATACAACTGGGTTATTGACTGGACG
CTAAAGGGTCTGGGTCACGGTCTTGAGGGCGACCCCAAGCTGCACTGTCAGCCAAAGAGATCTACCAATG
CTGAATTCACCCTCCAACTTAATATATCTCGGAGGCATACAAATGACCACTACAGGGAAAGAATTGAAAC
```

Figure 5. Head of a fastA document containig the segment S of Amapari virus

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Figure 6. Head of a fastA document containing ovax chick gene as proteins

fastQ

It provides an extension to the fastA format. The header line of FastQ begins with "@ instead of ">" and it add two additional fields. One third line starting with "+" and optionally containing additional metadata. Last line encodes the quality scores expressing the level of confidence for each nucleotide in Line 2. This quality scores are usually machine dependent.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

Figure 7. short fastq document using the metadata field

```
@IL31_4368:1:1:996:8507/1
NTGATAAAGTAATGACAAAATAATGACATTATTGTTACTATGGTTACTGTGGGA
+
(94**0-)*7=06>>><<<<<<22@>6;;;5;6:;63:4?-622647..-.5.%
@IL31_4368:1:1:996:21421/1
NAAGTTAATTCTTCATTGTCCATTCCTCTGAAATGATTCAGAAATACTGGTAGT
+
(***+2396,@<+<:@@@;;5)<0)69606>4;5>;>6&<102)0*+8:&137;
@IL31_4368:1:1:997:10572/1
NAATGTATGTAGACCCTTCACATTCAAAGGCAAATACAATATCATCATGTCTTC
+
(/9**-0032>:>>9>4@@=>??@@:-66,;>;<;6+;255,1;7>>>>3676'
```

Figure 8. Multiple fastq document containing three different reads

Aligned data. SAM/BAM and CRAM

By mapping FastA/FastQ reads against a reference genome it can be compute similarities and differences between the two sequences. This aligned reads are produced by alignment tools which the most used output format are SAM/BAM and CRAM. Although this popular formatas also could be used for simply encapsulate and compress the unmapped reads produced by sequencing machines.

SAM/BAM

Sequence Alignment/Mapping (SAM) format is a generic alignment format for storing read alignments against reference sequences, supporting short and long reads (up to 128 Mbp) produced by different sequencing platforms. It is flexible in style, compact in size, efficient in random access and is the format in which alignments from the 1000 Genomes Project are released. [91]

BAM format is the compressed, indexed, binary version of SAM. Compression in BAM is implemented as a block-based zip.

CRAM

CRAM is a format designed by the European Bioinformatics Institute (EBI). It uses reference-based compression, new sequences are aligned to a reference genome and only the differences are encoded for storage

Typically CRAM achieves 40-50% space saving over the alternative BAM format. [13]

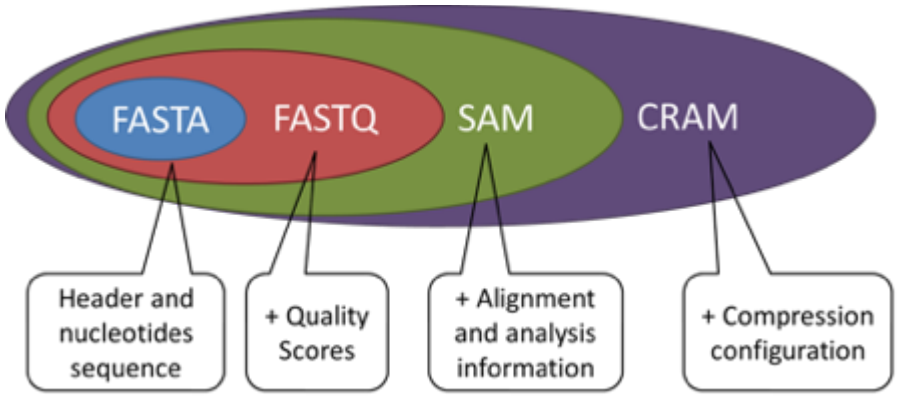

Figure 9. Relationship among some of the most popular file formats [14]

SAM/BAM and CRAM and the related tools are both managed by the Data Working Group File Formats Task Team of the Global Alliance for Genomics and Health.

Even though CRAM is more efficient that BAM most organization and international repositories are using regular bam files.

4.2 SAM

The Sequence Alignment/Map (SAM) format is a text format consisting of an optional header section and a alignment section storing the sequence data in a TAB-limited ASCII columns.

SAM is human readable and is efficient in random access. Currently is the output format of different sequencing platforms and is the format in which alignments from the 1000 Genomes Project are released. The current specification of the format is at [15, 16]

An example

Once the sequences read are aligned with the references genome a file similar but far longer that the following is obtained:

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

Figure 10. Read alignment

The previous alignment can be stored to memory in SAM format:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Figure 11. SAM document

SAM Terminology

Template	A DNA/RNA sequence part of which is sequenced on a sequencing machine or assembled from raw sequences.
Segment	A contiguous sequence or subsequence.
Read	A raw sequence that comes off a sequencing machine. A read may consist of multiple segments. For sequencing data, reads are indexed by the order in which they are sequenced.
Linear alignment	An alignment of a read to a single reference sequence that may include insertions, deletions, skips and clipping, but may not include direction changes (i.e. one portion of the alignment on forward strand and another portion of alignment on reverse strand). A linear alignment can be represented in a single SAM record (e.g. r002 and r004 in the example above).
Chimerica alignment	An alignment of a read that cannot be represented as a linear alignment. A chimeric alignment is represented as a set of linear alignments that do not have large overlaps (e.g. r003 in the example above is composed by two linear alignments). Typically, one of the linear alignments in a chimeric alignment is considered the "representative" alignment and the others are called "supplementary" and are distinguished by the supplementary alignment flag.
Read alignment	A linear alignment (1 SAM record) or a chimeric alignment (several SAM records) that is the complete representation of the alignment of the read.
Multiple mapping	The correct placement of a read may be ambiguous, e.g. due to repeats. In this case, there may be multiple read alignments for the same read. One of these alignments is considered primary. All the other alignments are considered "secondary". Typically, the alignment designated primary is the best alignment, but the decision may be arbitrary.
Phred scale	Given a probability $0 < p \leq 1$, the phred scale of p equals -10 log10p, rounded to the closest integer.

Table 4. SAM Basic Terminology

The SAM header

In the header, each line is TAB-delimited and except the @CO lines, each data field follows a format `TAG:VALUE' where TAG is a two-letter string that define the content and the format of "VALUE."

The SAM header start with an '@' followed by a two-letter record type code. It is not required, but if it is there, it contains some mandatory fields. The header may contain the version information, information about the reference sequences like the length, if sorted or not and the method used, the processing and the programs that was used to generate the various reads and also may contain free-form text comments.

A group of reads may be assigned to the same reference sequence, the information of this reference is only in the header instead of be repeated in every read. The alignment "points" to the reference sequence are in the header too via the RNAME field.

The alignment section: mandatory fields

In SAM format, each alignment line typically represents the linear alignment of a segment and have 11 mandatory fields and variable number of optional fields. Their values can be `0' or `*' if the corresponding information is unavailable.

The optional fields are presented as key-value pairs in the format of TAG:TYPE:VALUE.

The following table gives an overview of the mandatory fields in the SAM format:

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference

11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Table 5. Mandatory fields in the SAM format

For example, in Figure 11. SAM document , the value for the mandatory fields for the two first alignments is depicted in the following table:

Fields	**Alignment 1**	**Alignment 2**
QNAME	r001	r002
FLAG	99	0
RNAME	ref	ref
POS	7	9
MAPQ	30	30
CIGAR	8M2I4M1D3M	3S6M1P1I4M
MRNM	=	*
MPOS	37	0
ISIZE	39	0
SEQ	TTAGATAAAGGATACTG	AAAAGATAAGGATA
QUAL	*	*

Table 6. Example of SAM fields

4.3 SAMtools

Once the reads are aligned and expressed as SAM or BAM files, they can be manipulated by the SAMtools toolkit [17].

SAMtools is an open source software library maintained by a community of software developers and originally written by Heng Li, a bioinformatics research scientist working at the Broad Institute.

Both simple and advanced tools are provided, supporting complex tasks like variant calling and alignment viewing as well as sorting, indexing, data extraction and format conversion [18].

4.4 MPEG Genome Information Storage and Compression

Compression of the genomic information is necessary due to the high volume of data that is being generated by the NGS thus and standardized compression format should increase the application potential of genomics.

ISO/TC 276 works on standardization in the field of biotechnology and ISO/IEC JTC 1/SC 29/WG 11 (MPEG) has the mission to develop standards for coded representation and compression of digital audio and video and related data. In its 28 years of activity MPEG has developed many generations of video and audio compression standards. By combining their respective expertise, ISO/TC 276 and MPEG are working for develop a file format for the storage and transmission of genomic information as a compression standard capable of providing new effective solutions to the stated problem [1].

GENomic Information File Format, GENIFF

GENIFF, GENomic Information File Format is one of the proposal for the new standard that is being developed, see it in [5]. The organizations presenting this proposal are Distributed Multimedia Applications Group from the Universitat Politècnica de Catalunya (DMAG-UPC), Barcelona Supercomputing Center (BSC), Centro Nacional de Análisis Genómico - Centre de Regulació Genòmica (CNAG-CRG), the Spanish companies Made of Genes and DAPCOM Data Services, The Pirbright Institute and Stanford University.

GENIFF is based on the ISO Base Media File Format and adapted according to Compressed ARchiving for GenOmics (CARGO) [5] with specific new boxes for genomic information in order to define a Genomic Information Transport Layer (GITL) to support the Transport requirements defined in [19].

4.5 XACML, eXtensible Access Control Markup Language

From uses cases in section 3.3 Use cases, several privacy rules can be defined, including information about:

- who is giving the permission to access to some piece of genomic information
- to whom the permission is given
- the time frame for the permission

- the operations permitted
- the purpose
- even if the data sharer should be informed of the result of the analysis performed.

The eXtensible Access Control Markup Language (XACML) [20] defined by OASIS defines a core XML schema for representing authorization and entitlement policies. The standard defines a declarative fine-grained policy language and an access control decision request/response protocol (both written in XML), and an architecture reference.

XACML is flexible for changes in policies because the modules that conform his architecture are decoupled, it also permits the granularity that is required. Using such a language, it is possible to create rules with the required level of detail.

The XACML policy language is used to express access control requirements. It has the elements PolicySet, Policy and Rule. A Policy is expressed through a set of Rules. It has standard extension points for defining new functions, data types, combining logic, etc

The request/response protocol is used to issue an access request against the XACML policies, the response will be Permit, Deny, Indeterminate or Not Applicable.

Architecture in XACML consist in a series of modules: the request go through the Policy Enforcement Points (PEP) that protect the resources which will evaluate it against the Policy Decision Point (PDP). The PDP or PEP may also need to query a Policy Information Point (PIP) to gather descriptive attributes about the user or the information asset to which access is requested. Policies are maintained via a Policy Administration Point (PAP).

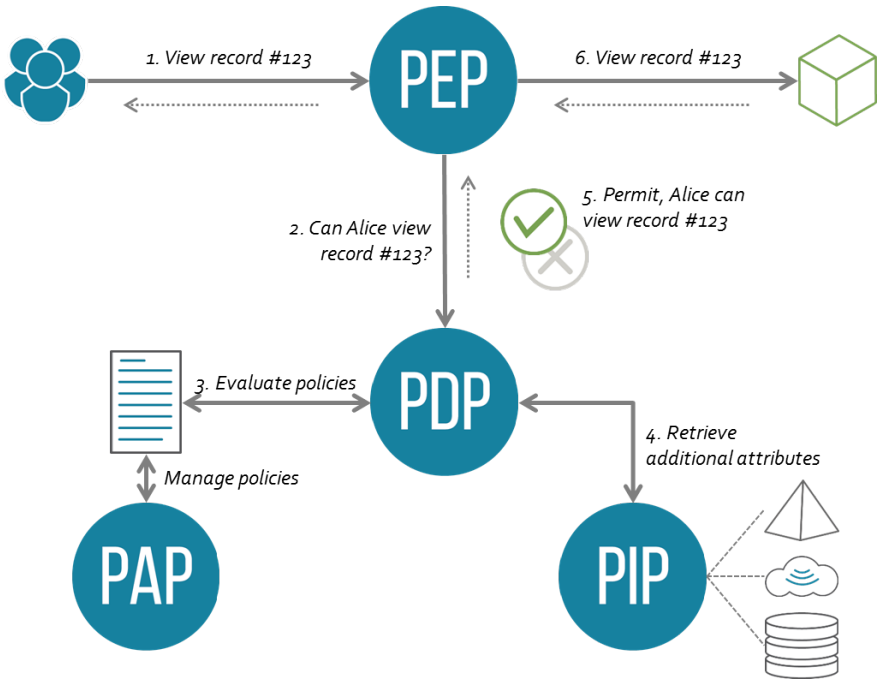

Figure 12. XACML architecture. Source Wikipedia [21]

XACML is primarily an Attribute-Based Access Control system (ABAC), where attributes are subjects, resources, environments, and actions. Role-based access control (RBAC) can also be implemented in XACML as a specialization of ABAC [22].

5 Contribution

5.1 Goals

Concerning privacy, one of the mechanisms foreseen to fulfil with the requirements defined in section 3 is the definition of privacy and security rules. These rules can be used to both describe security (encryption techniques, access from external systems, etc.) and privacy (who can access, when, for what purpose, who should be informed, etc.) features associated to the genome. In this way, the security and privacy requirements already identified could be fulfilled by an application/system accessing to the genome file, which should be aware of them.

The addition of rules should not affect the genomic information processing, although it may govern its usage. They should be included into the formats allowing governance checking before further processing.

The main goal of this project is to propose and test the use of this type of rules with the genomic information, developing a proof of concept on the current formats. This idea will be use in the creation of new standardized formats. For that a tool, that allows you to read and manipulate genomic information files only after evaluate access requests according to the rules defined, is developed.

To formalize policies, we propose the use of the XACML language. As already explained in 4.5 XACML, eXtensible Access Control Markup Language, this language gives the possibility of defining rules with a detail. The tool developed is a Java module as a lightweight wrapper of the samtools C-API, which perform authorizations, depending on user requests, using the open source XACML implementation Balana [23] for access control.

5.2 Genomic information generation and manipulation

Figure 13 shows the main stages of genomic information manipulation in existing bioinformatics applications as indicated in [8] plus one extra step added. The steps depicted include:

1. Sequencing: expression of genomic information as strings (a.k.a. sequences or reads) of nucleotides identifiers.
2. Alignment/mapping: sequences arrangement to identify regions of similarity among them (*de-novo* assembly) or with respect to an external reference (a pre-constructed genome). Sequences are encoded in the form of SAM files and its binary dual named BAM.
3. Compression: data encoding to use less bit.
4. Storage: compressed data is stored and made available via database interfaces or files.
5. Decompression/access: access to data to perform analysis.
6. Update: previously sequenced genomic information might be updated by means of new alignment techniques or new sequencing (a.k.a. re-sequencing).

The new step has been added after 5. Decompressed data and before performing the analysis:

- Authorised data/access: access only to the data you have permission to perform analysis.

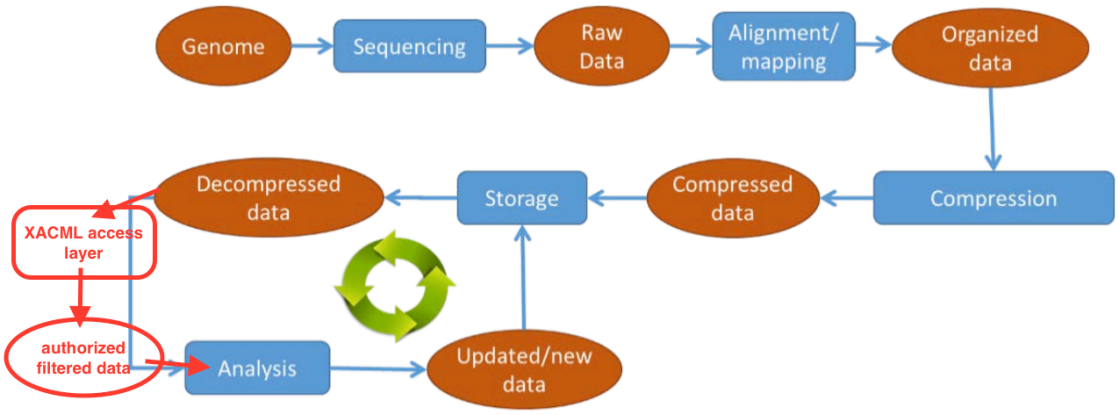

Figure 13. Genomic information generation and manipulation stages, plus one more new stage added in red

5.3 Integrating XACML with Samtools

Figure 14 shows the process of evaluation of access rights:

1. A user sends a request, which is intercepted by XACML engine.

2. The XACML engine converts the request into a XACML authorization request.

3. The XACML engine evaluates the authorization request against the policies that the file requested is configured with.

4. The XACML engine reaches a decision (Permit / Deny / NotApplicable / Indeterminate).

5. In the case that the decision result is equal to “Permit” the file is decoded an analyzed by the user.

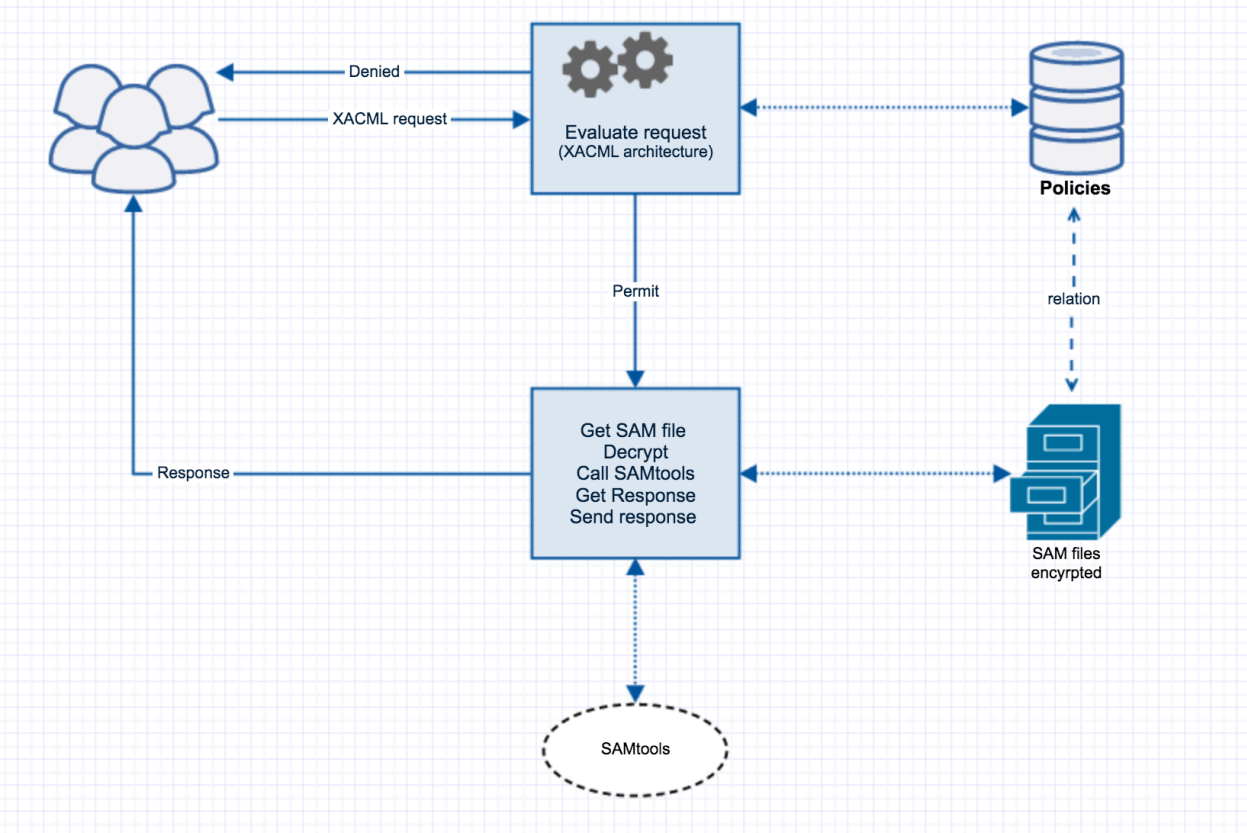

Figure 14. Authorization flow

Rules could be stored in a repository of rules associated with the genomic resources or they could be stored inside the genomic information file. The size of the rules is negligible when compared with the size of genomic information. In this way, the inclusion of rules inside the genome file format should not be a problem, even if several rules are defined.

5.4 Development

GitHub

A repository on GitHub has been created making a fork of the original Samtools repository which is also available on GitHub, the code can be seen in the following link:

https://github.com/saracubillas/samtools/tree/develop/privacy

Balana

To not reinvent the wheel developing a specific XACML engine, we decided to use one of the already available Open Source implementations. Balana [23] is coded in Java and supports XACML 3.0.

To create a Balana instance it is needed a configuration file from where the instance may read, in this configuration file we have specified that policies will be defined in on a role-based access control model (RBAC). In addiction, we need to specify where to find the authorization polices, this polices my be in a policy repository or emmbebed in the genomic information format as is explained below in section GNNIFFormat

```
try {
    policyLocation = (new File(".")).getCanonicalPath() + File.separator + "policy";
    System.setProperty(FileBasedPolicyFinderModule.POLICY_DIR_PROPERTY, policyLocation);
} catch (IOException e) {
    System.err.println("Can not locate policy repository");
}

try {
    configPath = (new File(".")).getCanonicalPath() + File.separator + "config/config_rbac.xml";
    System.setProperty(ConfigurationStore.PDP_CONFIG_PROPERTY, configPath);
} catch (IOException e) {
    System.err.println("Can not locate configuration repository");
}


balana = org.wso2.balana.Balana.getInstance();
```

Figure 15. Creating a Balana instance

5.5 Enforcement policies

It has been defined a number of rules contained inside the Policy and this forms the basic unit that can be acted upon by the Policy Decision Point (PDP) of the XACML engine. The Rules are the basic building block of a XACML policy: it contains the desired Effect - either Permit or Deny.

Because there are more than one Rule, it becomes necessary to provide a means to evaluate the decision of each one of these subsets in relation to the others. In this case, it has been chosen the combining algorithm option “first-applicable” which means that each rule is evaluated in the order that it appears in the policy. For a rule, if the target evaluates to "True" and the rule evaluates to a determinate value of "Permit" or "Deny", then the evaluation SHALL halt and the policy SHALL evaluate to the effect value of that rule. For a particual rule, if the target evaluates to "False", or the rule evaluates to "NotApplicable", then the next rule in the order SHALL be evaluated. For the case that no further rule match an additional rule has been added at the end of the policy that makes any other request regardless of user, profile, resource, action and condition denied.

Further down in figure 15 it can be seen the XACML police containing two rules:

1) A physician may view the genomic information file for which he or she is the designated primary care physician, provided an email is sent to the patient.

2) A researcher may view chromosome 20 of a genomic information file if he is the responsible of the study, provided an email is sent to the data sharer.

```
<Policy
    xmlns="urn:oasis:names:tc:xacml:3.0:core:schema:wd-17"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="urn:oasis:names:tc:xacml:3.0:core:schema:wd-17
    http://docs.oasis-open.org/xacml/3.0/xacml-core-v3-schema-wd-17.xsd"
    PolicyId="urn:isdcm:policyid:2"
    RuleCombiningAlgId="urn:oasis:names:tc:xacml:1.0:rule-combining-algorithm:first-applicable"
    Version="1.0">
        <Description> Policy rules sample</Description>
        <PolicyDefaults>
            <XPathVersion>http://www.w3.org/TR/1999/REC-xpath-19991116</XPathVersion>
        </PolicyDefaults>
        <Target/>
        <Rule RuleId="urn:oasis:names:tc:xacml:3.0:ejemplo:RuleSAM" Effect="Permit">
            <Description> A physician may view the genomic information file
            for which he or she is the designated primary care
            physician, provided an email is sent to the patient</Description>
            <Target>
            <AnyOf>
                <AllOf>
                    <!-- Which kind of user: physician -->
                    <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                        <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                            physician
                        </AttributeValue>
                        <AttributeDesignator MustBePresent="false"
                            Category="urn:oasis:names:tc:xacml:3.0:role" AttributeId="role"
```

```
                        DataType="http://www.w3.org/2001/XMLSchema#string"/>
                </Match>

                <!-- Which resource -->
                <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:regexp-string-match">
                    <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                        examples/toy.sam
                    </AttributeValue>
                    <AttributeDesignator MustBePresent="false"
                        Category="urn:oasis:names:tc:xacml:3.0:attribute-category:resource"
                        AttributeId="urn:oasis:names:tc:xacml:1.0:resource:resource-id"
                        DataType="http://www.w3.org/2001/XMLSchema#string"/>
                </Match>

                <!-- Which action  -->
                <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                    <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                        VIEW
                    </AttributeValue>
                    <AttributeDesignator MustBePresent="false"
                        Category="urn:oasis:names:tc:xacml:3.0:attribute-category:action"
                        AttributeId="urn:oasis:names:tc:xacml:1.0:action:action-id"
                        DataType="http://www.w3.org/2001/XMLSchema#string"/>
                </Match>
            </AllOf>
        </AnyOf>
        </Target>
        <Condition>
            <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:and">
                <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:integer-less-than">
                  <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:integer-one-and-only">
                    <AttributeDesignator MustBePresent="false"
                        Category="urn:oasis:names:tc:xacml:3.0:count"
                        AttributeId="countView"
                        DataType="http://www.w3.org/2001/XMLSchema#integer"/>
                  </Apply>
                  <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#integer">
                    4
                  </AttributeValue>
                </Apply>
            </Apply>
        </Condition>
    </Rule>
    <Rule RuleId="urn:oasis:names:tc:xacml:3.0:ejemplo:RuleSAMChromosome" Effect="Permit">
        <Description>A researcher may view chromosome 20 of a genomic information
        file if he is the responsible of the study,
        provided an email is sent to the data sharer </Description>
        <Target>
        <AnyOf>
            <AllOf>
                <!-- Which kind of user: researcher -->
                <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                    <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                        researcher
                    </AttributeValue>
                    <AttributeDesignator MustBePresent="false"
                        Category="urn:oasis:names:tc:xacml:3.0:role" AttributeId="role"
                        DataType="http://www.w3.org/2001/XMLSchema#string"/>
                </Match>

                <!-- Which resource -->
                <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:regexp-string-match">
                    <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                        examples/toy.sam#ref2
                    </AttributeValue>
                    <AttributeDesignator MustBePresent="false"
                        Category="urn:oasis:names:tc:xacml:3.0:attribute-category:resource"
                        AttributeId="urn:oasis:names:tc:xacml:1.0:resource:resource-id"
                        DataType="http://www.w3.org/2001/XMLSchema#string"/>
                </Match>
```

```
                    <!-- Which action  -->
                    <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                        <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                            VIEWCHROMOSOME
                        </AttributeValue>
                        <AttributeDesignator MustBePresent="false"
                             Category="urn:oasis:names:tc:xacml:3.0:attribute-category:action"
                             AttributeId="urn:oasis:names:tc:xacml:1.0:action:action-id"
                             DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </Match>
                </AllOf>
            </AnyOf>
            </Target>
            <Condition>
                <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:and">

                  <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:integer-less-than">
                    <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:integer-one-and-only">
                      <AttributeDesignator MustBePresent="false"
                          Category="urn:oasis:names:tc:xacml:3.0:count" AttributeId="countView"
                          DataType="http://www.w3.org/2001/XMLSchema#integer"/>
                    </Apply>
                    <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#integer">
                        4
                    </AttributeValue>
                    </Apply>
                </Apply>
            </Condition>
        </Rule>
        <Rule RuleId="urn:oasis:names:tc:xacml:3.0:lab6:FinalRule" Effect="Deny"/>
            <ObligationExpressions>
                <ObligationExpression
                    ObligationId="urn:oasis:names:tc:xacml:example:obligation:email"
                    FulfillOn="Permit">
                    <AttributeAssignmentExpression
                        AttributeId="urn:oasis:names:tc:xacml:3.0:example:attribute:mailto">
                        <AttributeSelector
                            MustBePresent="true"
                            Category="urn:oasis:names:tc:xacml:3.0:attribute-category:resource"
                            Path="patient-email"
                            DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </AttributeAssignmentExpression>
                    <AttributeAssignmentExpression
                        AttributeId="urn:oasis:names:tc:xacml:3.0:example:attribute:text">
                        <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                            Your genomic information has been accessed by:
                        </AttributeValue>
                    </AttributeAssignmentExpression>
                    <AttributeAssignmentExpression
                        AttributeId="urn:oasis:names:tc:xacml:3.0:example:attribute:text">
                        <AttributeDesignator
                         MustBePresent="false"
                         Category="urn:oasis:names:tc:xacml:1.0:subject-category:access-subject"
                         AttributeId="urn:oasis:names:tc:xacml:1.0:subject:subject-id"
                         DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </AttributeAssignmentExpression>
             </ObligationExpression>
        </ObligationExpressions>
    </Rule>
</Policy>
```

Figure 16. XACML Policy

In the rules described above everything is prohibited when there are no permits, but also the need to include rules with the opposite philosophy was also seen. A rule that includes access prohibitions and allows access to what is not prohibited.

In Figure 17. XACML rules combining algorithm, there is:

- A rule that forbids access to chromosome 2 of "file.sam" for doctor´s role
- A rule that gives access to all the "file.sam" chromosomes for doctor´s role

Using rule-combining-algorithm: first-applicable, having the deny rule first makes that the access to this part of the file is denied whereas access for any other part of the same file is allowed.

```
<RuleRuleId="urn:oasis:names:tc:xacml:3.0:ejemplo:RuleSAMChromosomeDeny" Effect="Deny">
<Description> A doctor can not view chromosome 2 </Description>
<Target>
           <AnyOf>
                <AllOf>
                    <!-- Which kind of user: researcher -->
                    <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                        <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#string">doctor</AttributeValue>
                        <AttributeDesignator MustBePresent="false"
Category="urn:oasis:names:tc:xacml:3.0:role" AttributeId="role"
DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </Match>
                    <!-- Which resource -->
                    <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                        <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#string">file.sam#ref2</AttributeValue>
                        <AttributeDesignator MustBePresent="false"
Category="urn:oasis:names:tc:xacml:3.0:attribute-category:resource"
AttributeId="urn:oasis:names:tc:xacml:1.0:resource:resource-id"
DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </Match>
                    <!-- Which action  -->
                    <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                        <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#string">VIEWCHROMOSOME</AttributeValue>
                        <AttributeDesignator MustBePresent="false"
        Category="urn:oasis:names:tc:xacml:3.0:attribute-category:action"
AttributeId="urn:oasis:names:tc:xacml:1.0:action:action-id"
                                    DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </Match>
                </AllOf>
            </AnyOf>
        </Target>

        <Condition>
            <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:and">
```

```
                <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:integer-less-
than">
                    <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:integer-
one-and-only">
                        <AttributeDesignator MustBePresent="false"
Category="urn:oasis:names:tc:xacml:3.0:count" AttributeId="countView"
DataType="http://www.w3.org/2001/XMLSchema#integer"/>
                    </Apply>
                    <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#integer">4</AttributeValue>
                </Apply>
            </Apply>
        </Condition>
    </Rule>
    <Rule RuleId="urn:oasis:names:tc:xacml:3.0:ejemplo:RuleSAMChromosomeALL"
Effect="Permit">
        <Description>A doctor may view all genomic information,
            provided an email is sent to the data sharer </Description>
        <Target>
            <AnyOf>
                <AllOf>
                    <!-- Which kind of user: doctor -->
                    <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                        <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#string">doctor</AttributeValue>
                        <AttributeDesignator MustBePresent="false"
Category="urn:oasis:names:tc:xacml:3.0:role" AttributeId="role"
DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </Match>
                    <!-- Which resource -->
                    <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:regexp-string-
match">
                        <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#string">file.sam*</AttributeValue>
                        <AttributeDesignator MustBePresent="false"
                                Category="urn:oasis:names:tc:xacml:3.0:attribute-
category:resource"
AttributeId="urn:oasis:names:tc:xacml:1.0:resource:resource-id"
DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </Match>
                    <!-- Which action  -->
                    <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                        <AttributeValue
```

```
DataType="http://www.w3.org/2001/XMLSchema#string">VIEWCHROMOSOME</AttributeValue>
                        <AttributeDesignator MustBePresent="false"
Category="urn:oasis:names:tc:xacml:3.0:attribute-category:action"
AttributeId="urn:oasis:names:tc:xacml:1.0:action:action-id"
DataType="http://www.w3.org/2001/XMLSchema#string"/>
                    </Match>
                </AllOf>
            </AnyOf>
        </Target>
        <Condition>
            <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:and">
                <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:integer-less-
than">
                    <Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:integer-
one-and-only">
                        <AttributeDesignator MustBePresent="false"
Category="urn:oasis:names:tc:xacml:3.0:count" AttributeId="countView"
DataType="http://www.w3.org/2001/XMLSchema#integer"/>
                    </Apply>
                    <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#integer">4</AttributeValue>
                </Apply>
            </Apply>
        </Condition>
    </Rule>
```

Figure 17. XACML rules combining algorithm

The XACML policy used that is shown Figure 16 and Figure 17, can be found in GitHub in [15].

5.6 XACML request and request evaluation

The XACML model supports and encourages the separation of the access decision from the point of use. Policy Enforcement Points, abbreviated as PEPs, are the endpoints in the XACML reference architecture where authorization questions are formulated and their resulting decisions are enforced. An authorization question comprises of attributes grouped into subject, resource, action, environment and the corresponding values for each such attribute. This request is created on the fly when the data is inserted by the user into the application. Below in Figure 18 a request example is shown.

```
<?xml version="1.0" encoding="UTF-8"?>
<Request xmlns="urn:oasis:names:tc:xacml:3.0:core:schema:wd-17"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="urn:oasis:names:tc:xacml:3.0:core:schema:wd-17 http://docs.oasis-
open.org/xacml/3.0/xacml-core-v3-schema-wd-17.xsd"
  ReturnPolicyIdList="true" CombinedDecision="true">
  <Attributes Category="urn:oasis:names:tc:xacml:1.0:subject-category:access-subject">
    <Attribute AttributeId="urn:oasis:names:tc:xacml:1.0:subject:subject-id"
IncludeInResult="false">
      <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">John
Doe</AttributeValue>
    </Attribute>
  </Attributes>
  <Attributes Category="urn:oasis:names:tc:xacml:3.0:role">
    <Attribute AttributeId="role" IncludeInResult="true">
      <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#string">doctor</AttributeValue>
    </Attribute>
  </Attributes>
  <Attributes Category="urn:oasis:names:tc:xacml:3.0:count">
    <Attribute AttributeId="countView" IncludeInResult="true">
      <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#integer">1</AttributeValue>
    </Attribute>
  </Attributes>
  <Attributes Category="urn:oasis:names:tc:xacml:3.0:attribute-category:resource">
    <Attribute AttributeId="urn:oasis:names:tc:xacml:1.0:resource:resource-id"
      IncludeInResult="false">
      <AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#string">toy.sam</AttributeValue>
    </Attribute>
  </Attributes>
  <Attributes Category="urn:oasis:names:tc:xacml:3.0:attribute-category:action">
    <Attribute AttributeId="urn:oasis:names:tc:xacml:1.0:action:action-id"
IncludeInResult="false">
      <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">VIEW</AttributeValue>
    </Attribute>
  </Attributes>
</Request>
```

Figure 18. XACML request example

The code to evaluate the request using Balana is shown in Figure 19.

```
private static String evaluateRequest(String request) {
    Balana balana = new Balana();
    String response = balana.evaluateRequest(request);
    return XACMLparser.getDecision(response);
}

if (decision.equals("Permit")){
    System.out.println("\n" + userName + " is authorized to perform this action\n\n");
    executeAction(resource, actionName);
} else {
    System.out.println("\n" + userName + " is NOT authorized to perform this action\n");
}
```

Figure 19. Construct and show evaluation request

In Figure 20 is shown a response example.

```
<Response xmlns="urn:oasis:names:tc:xacml:3.0:core:schema:wd-
17"><Result><Decision>Permit</Decision><Status><StatusCode
Value="urn:oasis:names:tc:xacml:1.0:status:ok"/></Status><Obligations><Obligation
ObligationId="urn:oasis:names:tc:xacml:example:obligation:email"><AttributeAssignment
AttributeId="urn:oasis:names:tc:xacml:3.0:example:attribute:text"
DataType="http://www.w3.org/2001/XMLSchema#string">
Your genomic information has been accessed by:</AttributeAssignment>
<AttributeAssignment  AttributeId="urn:oasis:names:tc:xacml:3.0:example:attribute:text"
DataType="http://www.w3.org/2001/XMLSchema#string">
JohnDoe</AttributeAssignment>
</Obligation></Obligations><PolicyIdentifierList><PolicyIdReference>urn:genomicaccescontrol:po
licyid:2</PolicyIdReference></PolicyIdentifierList><Attributes
Category="urn:oasis:names:tc:xacml:3.0:role"><Attribute AttributeId="role"
IncludeInResult="true">
<AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#string">researcher</AttributeValue></Attribute>
</Attributes><Attributes Category="urn:oasis:names:tc:xacml:3.0:count"><Attribute
AttributeId="countView" IncludeInResult="true">
<AttributeValue
DataType="http://www.w3.org/2001/XMLSchema#integer">1</AttributeValue></Attribute>
</Attributes></Result></Response>
```

Figure 20. Response example

Upon receiving the response, a message is prompted to the user and if the decision is “Permit” an action is executed.

5.7 Command Factory

Different actions can be executed over a genomic information file. The action that the user wants to perform is indicated in the XACML request in the attribute action, see Figure 18. Each action is translated into several Samtool commands for the element Command factory of the application developed. This commands manipulate the SAM file also indicated in the XACML request as attribute resource.

It has been developed two commands: the view command and the view chromosome command.

View command

The view command convert the SAM format file given as a resource in the request in to a BAM format file. The order of extracted reads is preserved.

View Chromosome command

This command extract the reads corresponding to a specific chromosome from the SAM file given in the request, the reference of the chromosome also is included as an attribute resource in the XACMLrequest. For that in Samtools the following action are performed:

- First is needed to convert the file into BAM format.
- Then an index BAI for the BAM file need to be created
- Finally, the data for specific region, for example chromosome 20, is extracted.

The code snippet to construct the view chromosome command is showed in Figure 21:

```java
package application.service;

public class ViewChromosome extends Command {

    @Override
    public void execute(String file) {
        String fileWithoutRef = file.substring(0, file.lastIndexOf('#'));
        String fileWithoutExtn = fileWithoutRef.substring(0, fileWithoutRef.lastIndexOf('.'));
        String chromosomeRef = file.substring(file.lastIndexOf("#") + 1);

        viewSAM(fileWithoutRef);
        IndexBam(fileWithoutExtn);
        String command = "samtools view -h " + fileWithoutExtn + ".bam " + chromosomeRef;

        String output = executeCommand(command);
        System.out.println(output);
        System.out.println("\n========OUTPUT======");
    }

    private void IndexBam(String fileWithoutExtn) {
        Command Index = new Index();
        Index.execute(fileWithoutExtn + ".bam");
    }

    private void viewSAM(String file) {
        Command ViewSAM = new ViewSAM();
        ViewSAM.execute(file);
    }
}
```

Figure 21. Create samtools command

The fact that the tool utilizes a domain driven design architectural pattern makes it easier to add or change commands. This approach certainly helps to fix fine-grained rules which allow to publish just small portions of the DNA, ensuring thus both privacy and utility.

5.8 Fitting the work with the GENIFF format

As explained in Section 4.4 MPEG Genome Information Storage and Compression, GENIFF is the proposal based on the ISO Base Media that is being developed and in which this project collaborates.

GENIFF provides a way to store the XACML definitions within the dataset and it has been implemented a parameter option in order to specify which element of the dataset should be extracted. In this case, this specific element is the XACML file. Then, by extracting this region of the file, we have enough information to evaluate the rules and decide if the request should be granted or not. The software developed is extensible to perform this task.

We show in Figure 22 a placeholder request to obtain the policy contents from a GENIFF file, instead of taking the policy from a policy repository as explained in before.

```
private static void initBalana() {
    // using file based policy repository. so set the policy location as system property
    String policyLocation = null;
    String configPath = null;
    try {
       // policyLocation = getPolicyFromGniff(resourceName);
```

Figure 22. Extracting XACML rules from GENIFF file

As previously mentioned, the rules are evaluated using the Balana library. If the response is positive, then the original file is extracted from the GENIFF container and decrypted if necessary, allowing further access through another tool such as Samtools, as we have seen before. The specific command which is issued is the one for which we evaluated the policy as the first step (see Figure 20).

```
        if (decision.equals("Permit")){
            System.out.println("\n" + userName + " is authorized to perform this action\n\n");
            //resource = decode(resource);
            executeAction(resource, actionName);
        } else {
            System.out.println("\n" + userName + " is NOT authorized to perform this action\n");
        }

    }
```

Figure 23. Show evaluation result

Finally, Figure 24 shows a revised authorization flow in the case of the GENIFF format.

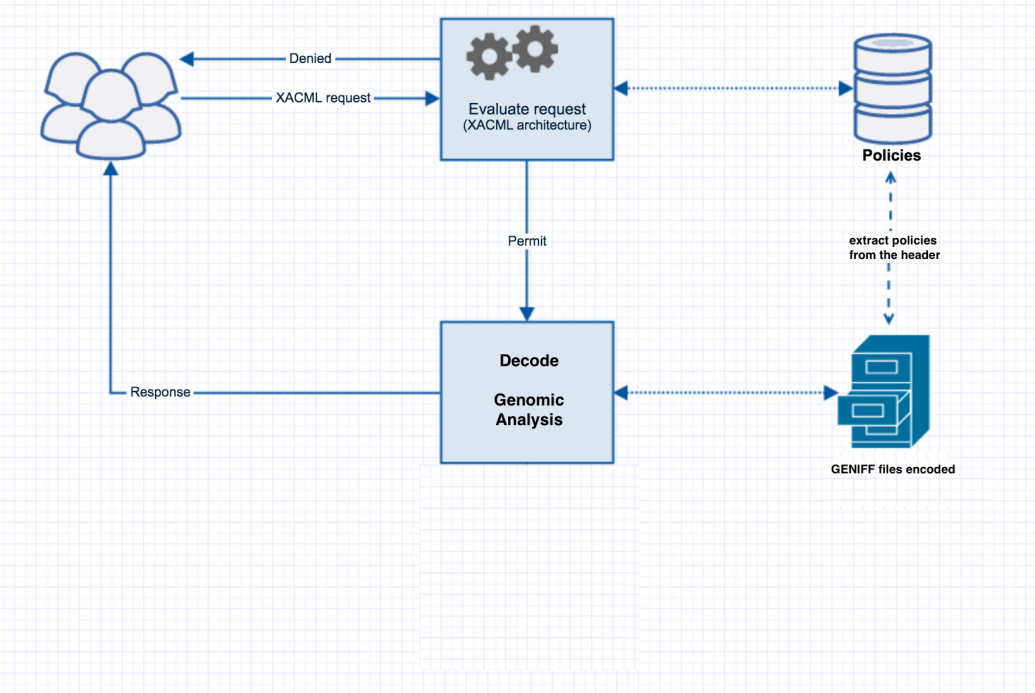

Figure 24. Revised authorization flow

5.9 How to try out the tool

To try out the tool follow these instructions:

- First it is needed to have installed Samtools [24]
- Download the executable file ”privacy.jar” on GitHub [16]
- Run: java -jar privacy.jar

As explained before, there are two operations allowed:

1) View a SAM file. It generates a BAM file.

2) View a chromosome of a SAM file. It generates a BAM file, creating the index file BAI and then it extracts the data for the specific region requested.

The file used in this example, “toy.sam”, provided by Samtools, can be downloaded from the folder “examples” in the same repository [16].

Figure 25 and Figure 26 below shows two different scenarios where the request is authorized:

```
Enter User name : John Doe
Enter role : physician

You can select one of following action:

[1] View - Convert SAM to BAM   [2] View Chromosome - extract the data for specific region
Enter action Id : 1
Enter file path : /Users/sara/Documents/MIRI/thesis/samtools-1.3.1/examples/toy.sam
```

Figure 25 . Example of execution 1

```
Enter User name : JaneDoe
Enter role : researcher

You can select one of following action:

[1] View - Convert SAM to BAM   [2] View Chromosome - extract the data for specific region
Enter action Id : 2
Enter file path : /Users/sara/Documents/MIRI/thesis/samtools-1.3.1/examples/toy.sam
Enter chromosome reference : ref2
```

Figure 26. Example of execution 2

6 Conclusions and future work

Understanding the implications of WGS of the human genome has been both daunting and exhilarating. The opportunities and threats that WGS enable described here justify the importance on research collaboration to address privacy issues and empower genomics research for the benefiting humankind.

In this project, it has been proposed a way to endure privacy and security in genomic information files. The participation in the project GENCOM (Secure GENomic information COMpression) allowed us to engage in intense debates about best ways of face the requirements given, adding value and richness to the outcome that was not fully anticipated when the planning process began.

The insights gained in this work have assisted understanding better the potential of using XACML as a security/privacy standard for new genomic compression standardized formats. In addition, it has been demonstrated that the approach is valid for currents formats and we have showed that such privacy measures can be integrated in specifications which are currently under development as it is the GENIFF case.

The first aspect of future work is to extend and consolidate the proposed model by signing the rules embedding them in Security Assertion Markup Language (SAML) requests and responses. Another point to consider is the management of the temporary files that need to be created for access some particular parts of DNA.

There is still much work to be done, many technical challenges to resolve for this kind of data protection where we cannot rely only on legislation. However, we hope this work will inspire further research in formally designing a standard for DNA compression.

Bibliography

[1] ISO/IEC JTC 1/SC 29/WG 11 - ISO/TC 276/WG 5 MPEG2016/N16320, «Joint Call for Proposals for Genomic Information Compression and Storage,» Geneva, June 2016.

[2] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron y G. E. Robinson, «Big Data: Astronomical or Genomical?,» *PLOS Biology,* 2015.

[3] S. D. Kahn, «On the Future of Genomic Data,» *Science,* vol. 331, pp. 728-729, 2011.

[4] ISO/IEC JTC 1/SC 29/WG 11 - ISO/TC 276/WG 5, «N16323/N97 - Requirements for Genomic Information Compression and Storage,» Geneva, 2016.

[5] Jaime Delgado, Silvia Llorente, Daniel Naro, Sara Rodríguez Cubillas, Francesc Tarrés, Luis Torres (Distributed Multimedia Applications Group – Universitat Politècnica de Catalunya), Łukasz Roguski (Centro Nacional de Análisis Genómico – Centre de Regulac, «GENIFF (GENomic Information File Format), a proposal for a Secure Genomic Information Transport Layer (GITL) based on the ISO Base Media File Format Response to the Joint Call for Proposals for Genomic Information Compression and Storage,» de *ISO/IEC JTC 1/SC 29/WG 11 MPEG2016/m39175*, Chendu, October 2016.

[6] Amy L McGuire, Rebecca Fisher, Paul Cusenza, Kathy Hudson, Mark A Rothstein, Deven McGraw, Stephen Matteson, John Glaser, and Douglas E Henley., «Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider,» *Genetics in medicine : official journal of the American College of Medical Genetics.,* vol. 10, nº 7, p. 495, 9 July 2008.

[7] L. H. a. D. Galas, «P4 Medicine: Personalized, Predictive, Preventive, Participatory: A Change of View That Changes Everything,» *Computing Research Association,* 2009.

[8] ISO/IEC JTC 1/SC 29/WG, «Investigation on genomic information compression and storage,» de *11 N15346*, Geneva, February 2015.

[9] USA, The National Human Genome Research Institute., [On line]. Available: https://www.genome.gov/27530645/faq-about-pharmacogenomics/.

[10] «IUPAC code table,» 11 08 2011. [On line].

[11] Wikipedia. [On line]. Available: https://en.wikipedia.org/wiki/Gattaca.

[12] Ayday, Erman; De Cristofaro, Emiliano; Hubaux, Jean-Pierre; Tsudik, «Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare,» *IEEE Computer,* vol. 48, pp. 58-66, 2015.

[13] [On line]. Available: http://www.sanger.ac.uk/science/tools/cram.

[14] INTERNATIONAL ORGANISATION FOR STANDARDISATION ORGANISATION INTERNATIONALE DE NORMALISATION, «Database for Evaluation of Genomic Information Compression and Storage,» *ISO/IEC JTC 1/SC 29/WG 11 CODING OF MOVING PICTURES AND AUDIO ISO/IEC JTC 1/SC 29/WG 11 N16145,* Febreary 2016.

[15] [On line]. Available: https://github.com/saracubillas/samtools/blob/develop/privacy/policy/XACMLPolicy.xml .

[16] [On line]. Available: https://github.com/saracubillas/samtools/tree/develop/privacy.

[17] [On line]. Available: http://www.htslib.org/.

[18] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis y R. Durbin, «The Sequence Alignment/Map format and SAMtools,» *1000 Genome Project Data Processing Subgroup,* 2009.

[19] «ISO/IEC JTC 1/SC 29/WG 11 - ISO/TC 276/WG 5 MPEG2016/N16323 - Requirements on Genomic Information Compression and Storage, Geneva, June 2016.».

[20] [On line]. Available: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml#CURRENT.

[21] «Wikipedia,» [On line]. Available: https://en.wikipedia.org/wiki/XACML.

[22] [On line]. Available: https://en.wikipedia.org/wiki/XACML.

[23] [On line]. Available: https://github.com/wso2/balana.

[24] [On line]. Available: http://www.htslib.org/ .