POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Graduado en Matemáticas e Informática

Universidad Politécnica de Madrid

Facultad de Informática

TRABAJO FIN DE GRADO

Estimación del coste de almacenamiento de una solución de preservación de la privacidad en archivos BAM reales.

Autor: Vanina Veselinova Dimitrova

Director: Erman Ayday, Zhicong Huang

27 MADRID, JULIO DE 2015

Contents

1 INTRODUCTION

A menudo los científicos secuencian el ADN de un gran número de personas con el objetivo de determinar qué genes se asocian con determinadas enfermedades. Esto permite mejorar el cuidado de la salud moderno y aporta una mejor visión del genoma humano. El precio de un perfil genómico completo se ha posicionado por debajo de los 200 dólares y este servicio lo ofrecen muchas compañías, la mayor parte localizadas en EEUU. Como consecuencia, en unos pocos años la mayoría de las personas procedentes de los países desarrollados tendrán los medios para tener su ADN secuenciado.

Alrededor del 0.5% del ADN de cada persona (que corresponde a varios millones de nucleótidos) es diferente del genoma de referencia debido a variaciones genéticas. Así que el genoma contiene información altamente sensible y personal y representa la identidad biológica del individuo. Al combinar datos genómicos con información sobre el entorno o estilo de vida de uno (a menudo facilmente obtenible de las redes sociales), sería posible inferir el fenotipo del individuo.

Multiples GWAS (Genome Wide Association Studies) realizados en los últimos años muestran que la susceptibilidad de un paciente a tener una enfermedad en particular, como el Alzheimer, cáncer o esquizofrenia, puede ser predicha parcialmente a partir de conjuntos de sus SNP (Single Nucleotide Polimorphism). Estos resultados pueden ser usados para medicina genómica personalizada (facilitando los tratamientos preventivos y diagnósticos), tests de paternidad genéticos y tests de compatibilidad genética para averiguar a qué enfermedades pueden ser susceptibles los descendientes.

Estos son algunos de los beneficios que podemos obtener usando la información genética, pero si esta información no es protegida puede ser usada para investigaciones criminales y por compañías aseguradoras. Este hecho podría llevar a discriminación genética.

Por lo que podemos concluir que la privacidad genómica es fundamental por el hecho de que contiene información sobre nuestra herencia étnica, nuestra predisposición a múltiples condiciones físicas y mentales, al igual que otras características fenotípicas, ancestros, hermanos y progenitores, pues los genomas de cualquier par de individuos relacionados son idénticos al 99.9%, contrastando con el 99.5% de dos personas aleatorias.

La legislación actual no proporciona suficiente información técnica sobre como almacenar y procesar de forma segura los genomas digitalizados, por lo tanto, es necesaria una legislación mas restrictiva.

Scientists typically sequence DNA from large numbers of people in order to determine genes associated with particular diseases. This allows to improve the modern healthcare and to provide a better understanding of the human genome. The price of a complete genome profile has plummeted below $200 and this service is offered by a number of companies, most of them located in the USA. Therefore, in a few years, most individuals in developed countries will have the means of having their genomes sequenced.

Around 0.5% of each person's DNA (which corresponds to several millions of nucleotides) is different from the reference genome, owing to genetic variations. Thus, the genome contains highly personal and sensitive information, and it represents our ultimate biological identity. By combining genomic data with information about one's environment or lifestyle (often easily obtainable from social networks), could make it possible to infer the individual's phenotype.

Multiple Genome Wide Association Studies (GWAS) performed in recent years have shown that a patient's susceptibility to particular diseases, such as Alzheimer's, cancer, or schizophrenia, can be partially predicted from sets of his SNPs. This results can be used for personalized genomic medicine (facilitating preventive treatment and diagnosis), genetic paternity tests, ancestry and genealogical testing, and genetic compatibility tests in order to have knowledge about which deseases would the descendant be susceptible to.

These are some of the betefits we can obtain using genoma information, but if this information is not protected it can be used for criminal investigations and insurance purposes. Such issues could lead to genetic discrimination.

So we can conclude that genomic privacy is fundamental due to the fact that genome contains information about our ethnic heritage, predisposition to numerous physical and mental health conditions, as well as other phenotypic traits, and ancestors, siblings, and progeny, since genomes of any two closely related individuals are 99.9% identical, in contrast with 99.5%, for two random people.

The current legislation does not offer sufficient technical information about safe and secure ways of storing and processing digitized genomes, therefore, there is need for more restrictive legislation [1] [3].

2 GENOMIC BACKGROUND

Sequencing DNA produces the DNA sequence data, which consists of millions of short reads (raw genomic data). These reads include between 100 and 400 nucleotides (A, C, G, T), depending on the type of sequencer used. They are randomly sampled from a human genome, bioinformatically treated and positioned or aligned to its genetic location in order to produce the Sequence alignment/map file (SAM) and its binary version (BAM). The SAM file of a patient contains hundreds of millions of short reads.
There are privacy-sensitive fields in a short read that should be protected from curious parties. These fields are the position of the short read that indicates the position of the first aligned nucleotide in its content with respect to the reference genome, the cigar string, which shows the variations in the content of the short read with respect to the reference genome, and its content represented by the nucleotides A, T, G and C [2].

3 METHODOLOGY

The purpose of this project is to estimate the storage cost that would involve the incorporation of privacy-preserving schemes into the BAM files.
The first task to do is to convert the BAM files (binary representation of the raw genomic data) to the human friendly format SAM. The next step is to use the SAM APIs to identify the privacy-sensitive fields exposed above.

An example of these fields in a short read could be the following:

Read Bases:
AAGTTAATAGAGAGGTGACTCAGATCCAGAGGTGGAAGAGGAAGGAA GCTTGGAACCCTATAGAGTTGCTGAGGGCCAGGACCAGATCCTGGCCC-TAAAC
Cigar String: 14S69M17S
Reference Position At Read Position: 59999

In order to hide the information contained in theese three fields there are introduced some modifications in the BAM and SAM APIs: the first one is replacing the **Read Bases** with a **random** string with the same length. The corresponding code would be:

samRecord.setReadString(randomString);

This alteration in the short reads doesn't suppose any change in the final size of the new BAM file because the original string and the new one have the same size. The same modification could be applied to the **Cigar String**.

Next, the 32-bit **Reference Position At Read Position** (position field) integer is expanded to a 64-bit **random** integer. For this, the /samtools/BAMRecordCodec.java file needs to be modified replacing

this.binaryCodec.writeInt(alignment.getAlignmentStart() - 1);

by

this.binaryCodec.writeLong(new java.util.Random().nextLong() - 1);

Finally a new field is added to each short read. This field consists of a 64-bit **random** integer that is going to be used in the encryption scheme [4]. To do this we need to add the following line

this.binaryCodec.writeLong(new java.util.Random().nextLong());

to the /samtools/BAMRecordCodec.java file after the line modified above.

To be able to read the new files with this new format we need to replace the following line of the /samtools/BAMRecordCodec.java file

final int coordinate = this.binaryCodec.readInt() + 1;

by

final int coordinate = (int) this.binaryCodec.readLong() + 1;

and add next

final long random = this.binaryCodec.readLong();

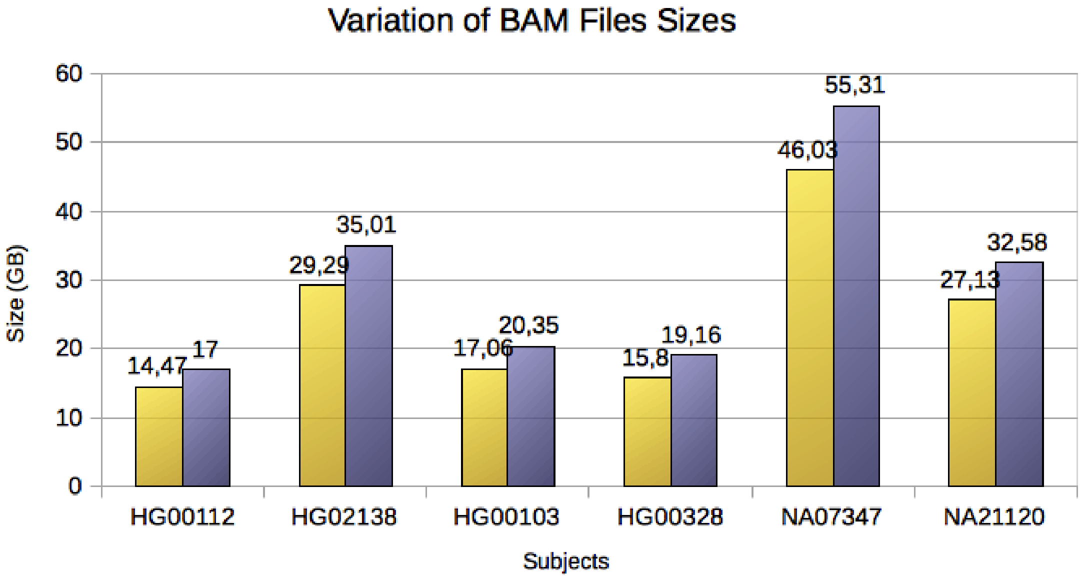
We also need to add the new field to the BAM file like below:

final BAMRecord ret = this.samRecordFactory.createBAMRecord(header, referenceID, coordinate, ***random****, readNameLength, mappingQuality, bin, cigarLen, flags, readLen, mateReferenceID, mateCoordinate, insertSize, restOfRecord);*

The result of these modifications could be a considerably raise of the BAM files size. In the next section we can observe the results of an experiment applied to real BAM files.

4 RESULTS AND CONCLUSIONS

BAM files contain a huge initial amount of data that occupy a considerable part of the memory. As a result of the BAM file modifications introduced above the size of these files increases substantially.
In the next graphic there are some examples with real BAM files, in which we can observe how much the size changes after the modifications.



As we can see, the sizes of these BAM files oscillate between 14GB and 46 GB. The average increase is 4.94GB which suppose an average percentage increase of 19.46%.

Although this quantity is significant in very big files, in this experiment the 46.03GB file grew 9.28GB, with the present technological resources we have, it is still feasible to store this amount of data.

To sumarize, there is need to ensure privacy-preserving of human genomic data for the reasons explained above and there exist methods that allows this in a doable way.

References

[1] Erman Ayday, Emiliano De Cristofaro, Jean-Pierre Hubaux, and Gene Tsudik. The chills and thrills of whole genome sequencing. pages 1–11.

[2] Erman Ayday, Jean Louis Raisaro, Urs Hengartner, Adam Molyneaux, and Jean-Pierre Hubaux. Privacy-preserving processing of raw genomic data. pages 1–4.

[3] Erman Ayday, Jean Louis Raisaro, Jean-Pierre Hubaux, and Jacques Rougemont. Protecting and evaluating genomic privacy in medical tests and personalized medicine. pages 1–6.

[4] The SAM Format Specification Working Group. The sam format specification (v1.4-r985). pages 1–11, 2011.