MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220255

Building a Genome Archiving and Communication System Integrated into a Health Information Systems

Mauricio Brunner[a], Federico Jauk[b], Nereo Candenas[a], Alfredo Cancio[a], Daniel Luna[a,c], Sonia Benitez[a,c]

[a] *Departamento de Informática en Salud, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina*
[b] *Laboratorio de Secuenciación, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina*
[c] *Instituto de Medicina Traslacional e Ingenieria Biomedica (IMTIB), Buenos Aires, Argentina*

Abstract

The use of next-generation sequencing technologies in clinical practice has increased the volume of information that must be stored, processed, and interpreted.

In this work, a description of the implementation of a genomic communication and archiving system (GACS) is presented. This GACS will allow us to store, share and search the genomic files and genetic variants obtained as a result of genetic laboratory tests.

Keywords:

Genomics, Precision Medicine, Data Integration.

Introduction

Personalized precision medicine seeks to use genomic approaches to improve treatments, prevent disease, and promote people's health. This implies the integration of different types of data with electronic medical records, which in turn opens up new lines of research and opportunities to improve health management [1]. The development of 'next generation' sequencing (NGS) technology [2] and its rapid price drop in recent years allowed healthcare institutions to incorporate sequencing equipment into their laboratories and use genomic information to make clinical decisions. At the same time, results processing and interpretation required highly complex computing platforms. Files with genetic data are used and generated both in the NGS sequencing process and in its bioinformatic processing. These files must be managed efficiently for reasons of order, control, quality and information security. The size of these files increases exponentially as the scope of genome sequencing increases, thus there is a trend towards storing complete sequencing data separate from the Electronic Health Record (EHR) [3] [4]. In this study we show how at the Hospital Italiano de Buenos Aires (HIBA), a highly complex hospital accredited with level 7 by the Healthcare Information and Management Systems Society, the genomic data of the patients were integrated into a genomic information archiving and communication system (GACS).

Methods

In the GACS we store files and genetic variants in a similar way to how an Picture Archiving and Communication System (PACS) does with the images generated by different medical devices (Figure 1).

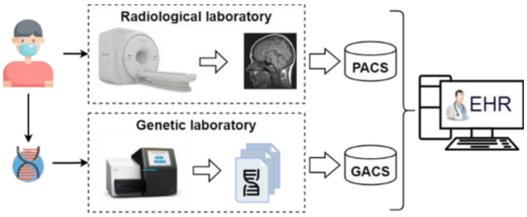

Figure 1– Analogy between GACS and PACS

The information flow of the implemented GACS can be seen in figure 2 and is summarized in the following paragraphs.

Architecture

The GACS is made up of two non-relational MongoDB [5] databases and a file system. MongoDB databases store the variants resulting from genetic studies, in such a way that they can be quickly consulted by a genetic information communication service. The first database stores the information from germline genetic tests, while the second stores somatic line tests. The file system is used in the GACS for the storage of raw files generated in the sequencing process and intermediate files generated in the bioinformatic processing. These files include:

- BAM files (Binary Alignment Map): These files contain the information of each sequence obtained by the sequencer about where and how they are mapped to a human reference genome.
- Raw uBAM files (Unmapped BAM): These files are a variant of the BAM file format in which the data read does not contain any mapping information.
- Variant Call Format (VCF) files: These are plain text files with a specific format used in bioinformatics to store sequence variations.

The Dell EMC Isilon platform [6] has been used on the HIBA servers for both, the databases and the file system. This network-attached storage platform allows us to scale-out for the storage, backup and archiving of data.

A relational data model stores the metadata that relates the patient to their laboratory samples. In this model, these samples are also associated with their different genetic studies carried out, and with the path of the genomic files within the GACS file system. For this model, the Oracle Database Manager [7] was used.

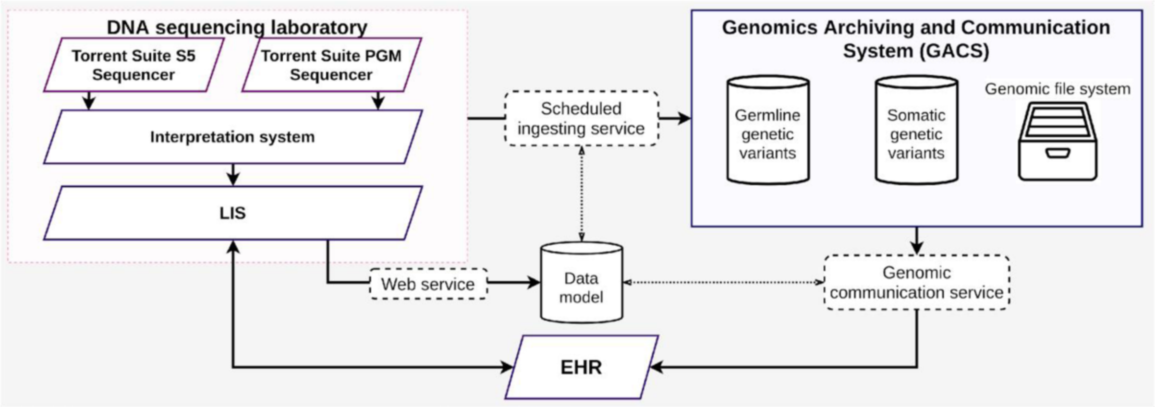
DNA sequencing laboratory
Torrent Suite S5 Sequencer
Torrent Suite PGM Sequencer
Interpretation system
LIS
Scheduled ingesting service
Genomics Archiving and Communication System (GACS)
Germline genetic variants
Somatic genetic variants
Genomic file system
Web service
Data model
Genomic communication service
EHR

Figure 2– GACS architecture

Integration process

A web service developed in the Java 11 language and the Spring Boot framework [8] is used by the laboratory information system (LIS) to write in the relational model tables each time a new genetic study report is made. The main data stored in the model are: sample and patient identifiers, genetic test code, test date, reporting system, sequencer equipment identifier, and name, type and directory of the genomic files generated in the sequencing process.

A daily computer process searches the tables for new records. Each time a new sample is found, a workflow is triggered that is responsible for searching for the reported variants and the genomic files generated in the sequencing equipment to upload them to the GACS. Finally, the metadata in the model is completed with the directories where they were saved.

Results

During the months of January to March of the year 2021, two NGS Ion Torrent sequencers and the Laboratory Information System were integrated into the GACS. The information was produced by the genetic tests carried out in the HIBA sequencing laboratory. These tests correspond to the following gene panels: BRCA1 / BRCA2 germline genetic test, Cystic Fibrosis Test, Hereditary Hemorrhagic Telangiectasia Test and Oncomine Focus Assay Solid Tumor Somatic Test.

Considering genomic files and databases, 2.7 TB of data have been integrated into the GACS so far. 352 genomic files have also been associated with 63 patients. These files include BAM, VCF, and uBAM files. Among this volume of information are results of tests for the years 2016, 2017, 2018, 2019, 2020 and 2021.

Conclusions

Implementing a GACS can provide efficient and centralized storage of all genomic information of patients to ensure their integrity and their subsequent use. While it can be technically challenging and requires a lot of storage, it can aid in diagnostics, information security, research purposes, and as supporting documentation for patient studies.

References

[1] Abul-Husn NS, Kenny EE. Personalized Medicine and the Power of Electronic Health Records. Cell. 2019;177: 58–69.

[2] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26: 1135–1145.

[3] Alterovitz GSync for Genes. ONC November 2017 report. [cited 17 Mar 2021]. Available at: https://www.healthit.gov/sites/default/files/sync_for_genes_report_november_2017.pdf.

[4] Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems. JAMA. 2013;309: 1237–1238.

[5] MongoDB Documentation. [cited 17 Sep 2021]. Available at: https://docs.mongodb.com/

[6] Almacenamiento Nas de Dell EMC isilon | Dell Technologies. [cited Sep 17 2021]. Available at: https://www.delltechnologies.com/es-ar/storage/isilon/isilon-a200-archive-nas-storage.htm

[7] Oracle Database management. [cited 17 Sep 2021]. Available at: https://www.oracle.com/database/technologies/manageability.html

[8] Spring Boot [Internet]. Spring. [cited 2021Sep17]. Available at: https://spring.io/projects/spring-boot

Address for correspondence

To contact the authors, write by email to mauricio.brunner@hospitalitaliano.org.ar.