MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)
© 2024 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI231114

Bioinformatics Architecture for Integrating Genomics Data into Electronic Health Records

Mauricio BRUNNER[a,1], Matias BUTTI[bc], Sebastián MENAZZI[b], Hernan CHANFREAU[b], Matias TAJERIAN[a], Alfonso QUIROGA[b], Paula OTERO[a], Daniel LUNA[a], Sonia BENITEZ[a]
[a]*Department of Health Informatics, Hospital Italiano de Buenos Aires*
[b]*GenomIT*
[c]*Universidad Abierta Interamericana, Centro de Altos Estudios en Tecnología Informática, Buenos Aires, Argentina*

ORCiD ID: Mauricio BRUNNER https://orcid.org/0000-0002-3692-9140, Matias BUTTI https://orcid.org/0000-0003-4841-8417, Sebastián MENAZZI https://orcid.org/0000-0002-9860-8457, Hernan, CHANFREAU https://orcid.org/0000-0001-7945-2111, Matias TAJERIAN https://orcid.org/0000-0002-6262-5458, Alfonso QUIROGA https://orcid.org/0000-0002-3248-1713, Paula OTERO https://orcid.org/0000-0002-9812-4591, Daniel LUNA https://orcid.org/0000-0001-5468-9761, Sonia BENITEZ https://orcid.org/0000-0001-6648-1984

Abstract. The adequate management of patients' genomic information is essential for any health institution pursuing the Precision Medicine model. Here we approach a bioinformatic architecture that allows the Institution to store its whole genetic test data in a scalable database, and also the integration of that genetic data with the Electronic Health Record through a Clinical Decision Support System. The system complements patient care by suggesting referral to genetic counseling for patients who are potentially at risk of hereditary breast/ovarian cancer, and allowing for proper follow-up of patients with pathogenic variants in *BRCA1 or BRCA2* genes. The implemented solution uses the FHIR standard and genetic nomenclatures from the Human Genome Variation Society and the HUGO Gene Nomenclature Committee. The architecture is flexible enough to allow any other health institution to integrate -to their information ecosystem- the whole solution or some of the modules according to its degree of digitization progress.

Keywords. Bioinformatics, Health Informatics, Genomics

1. Introduction

Precision Medicine proposes incorporating individual variability in prevention, diagnosis and treatment methods, with the aim of maximizing their effectiveness [1][2]. This variability is largely based on the genetic information of the person. An important milestone in human genetics occurred in 2003, with the completion of the Human Genome Project and the publication of the first complete draft of its sequence. From that moment, the development and translation of different types of genomic tests aimed at

[1]Corresponding author: Mauricio Brunner, mauricio.brunner@hospitalitaliano.org.ar

improving the understanding of the molecular mechanisms underlying most diseases began. Around 2007, the arrival of Next Generation Sequencing (NGS) technologies [3] had a direct impact on the reduction of sequencing costs, which exponentially increased knowledge regarding the genetic causes related to different types of diseases. It wasn't until early 2022 when, applying such technologies, that the full sequence of the human species' genome was finally identified and published [4].

Most genetic tests that are run by NGS result in large volumes of genomic data, and interpretation can sometimes be challenging. The Electronic Health Record (EHR) and Clinical Decision Support Systems (CDSS) allow for the integration of different types of data, both from clinical and laboratory/molecular sources, and this can be particularly useful for supporting professionals in real-time decision making for complex situations, such as genomics. This improves the ability to positively impact patient outcomes [5].

In this work, we present the design and implementation of a bioinformatic architecture aimed to organize the genomic information of the patients of the Hospital Italiano de Buenos Aires, and to make it available in the EHR through the CDSS.

2. Methods

This work is the result of a collaboration between the Health Informatics Department of the Hospital Italiano de Buenos Aires (HIBA) and GenomIT, a Precision Medicine company from Argentina. The interdisciplinary team included medical geneticists, bioinformaticians, medical informatics and software developers and engineers from both institutions.

The scope of this work includes the integration of the data generated by the genetic tests requested by the HIBA Hereditary Cancer Program (ProCanHe) and carried out in the HIBA sequencing laboratory, as well as those performed by other genetic testing providers. ProCanHe is a working group that focuses on the diagnosis, prevention and treatment of different types of hereditary cancer.

Genomic raw files including data generated by an NGS sequencer pipeline in variant call format (VCF) [6] and genetic reports were obtained from Ion Reporter™ (Thermo Fisher Scientific), a system used as an interpretation tool in the sequencing laboratory. The patients evaluated in ProCanHe are not sequenced in the HIBA laboratory, so the genetic variants identified in them were obtained from the secondary databases of the program, in TSV format. Raw VCF files were annotated and filtered. The SnpEff software [7] was used to add information related to the gene, transcript, nomenclature, effect and predicted impact of each genetic variant. VCFTools [8] was used to remove chromosomal regions that were not analyzed in the laboratory by a geneticist. To determine when a new report of a genetic test is completed within the sequencing laboratory, an integration with the Laboratory Information System (LIS) was carried out.

For the definition of genetic rules that activate alerts in the EHR through CDSS, the guidelines of the National Comprehensive Cancer Network (NCCN) [9] were used.

For syntactic interoperability between the developed systems, the HL7 FHIR (Fast Healthcare Interoperability Resources) standard [10] was used.

For semantic interoperability, the nomenclature of the Human Genome Variation Society (HGVS) [11] was used for the description of Genetic variants, and the HUGO Genetic Nomenclature Committee (HGNC) nomenclature [12], for naming approved human genes.

The different elements of software in this architecture were developed using Java, Python, and Bash programming languages. To store the genomic information of the patients, MongoDB v4.2 database technology was used.

3. Results

In 2022, a system was implemented at HIBA that allows for the organization of genetic information from the germline of patients and displays recommendations in the EHR through its CDSS. The architecture of the implemented system is composed of the modules that are described below and that can be seen in Figure 1.

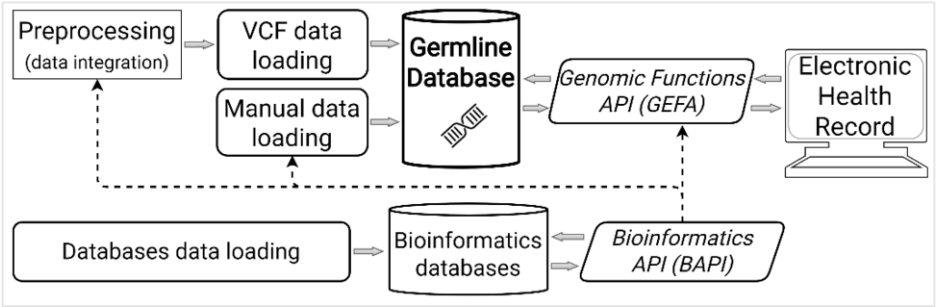

Figure 1. High level architecture.

3.1. Genomic Databases

The germline database is the main component of the architecture. Its implementation allows for storing three types of information in different data collections: tests, genetic variants and sequenced ranges that do not include variants with respect to the reference genome used (wild type regions).

Other sources of information include different databases that contain public genomic data in the bioinformatics database. This database allows us to make inferences, annotations and validations of the patients' results in different parts of the process of uploading data to the germline database. The implemented databases have information for the human reference genomes GRCh37 and GRCh38, and related to variants, genes, transcripts, and genomic nomenclature. Each of these databases is incorporated into MongoDB by a different Extract, Transform and Load (ETL) process defined in a bash script. The weight of all these databases installed locally on the HIBA servers is 131 GB.

3.2. Bioinformatic and clinical REST applications

Two representational state transfer (REST) applications were developed. The Bioinformatics API (BAPI) is the application that contains the services needed to perform validations, inferences and annotations on variants, based on information from bioinformatics databases. On the other hand, the Genetic Functions API (GEFA) contains the high-level bioinformatics functions. Among these functions, we highlight that of making predefined queries to the germline database and processing and integrating the results with information from the bioinformatic databases through BAPI.

It returns the results using the FHIR interoperability standard, which allows us to integrate it with the CDSS. The FHIR MolecularSequence resource was used to represent the genetic variants of the patients.

3.3. Manual and VCF genetic data ingestors

The manual ingestor is a web application that allows for loading test results and variants to the germline database. It uses BAPIs to perform real-time inference and validation, to reduce loading time and the risk of making typing mistakes when registering variants. The VCF ingestor is used to upload tests together with the genetic variants to the germline database once the report has been uploaded in the LIS. This ingestor runs automatically every day and incorporates all the genetic variants obtained in the test (previously reported or not), from the annotated VCF genome file, into the database. It also allows for storing the sequenced ranges where no genetic variants have been detected (essential for determining the absence of relevant variants).

3.4. Preingestor

We have named preingestor a preprocessing service responsible for obtaining the different inputs from the VCF ingestor. It integrates with the LIS database to determine which genetic tests were reported in the last 24 hours. Once the tests to be processed for uploading to the germline database have been determined, the preingestor integrates with the Ion Reporter™ sequencing laboratory interpretation system to get the VCF file and the genetic variants. It then runs the annotation process for the VCF file and finally runs the ingestor. The complete data flow of the preingestor consists of an ETL process where each step of the workflow is carried out by a different script. This process is managed by the Cromwell workflow management system [13], implemented in the HIBA servers.

3.5. Genetic rules in the CDSS

With regard to genetic rules, two types of rules were specified to be shown in the EHR: a patient follow-up rule, which, for example, recommends the attending physician to refer the patient to dermatology for increased risk of suffering from melanoma due to having a pathogenic or probably pathogenic variant in the *BRCA1* or *BRCA2* genes, and rules for referral to genetic counseling for patients at increased risk of having a pathogenic or likely pathogenic variant in the same genes. This risk may be due to the diagnosis of: breast cancer in women under 46 years of age, breast cancer in men, ovarian cancer, or pancreatic cancer, among other rules.

4. Discussion

The architecture described in this work was successfully implemented in the Italian Hospital of Buenos Aires. The main components of this architecture use interoperability standards such as FHIR for communication with the EHR or CDSS and the VCF, HGNC and HGVS standards for the genetic variants stored in the database. We consider that they could be implemented in any health institution that has implemented an information system that uses health information standards. However, we acknowledge an effort

would have to be made for the integration with the specific systems of each institution (LIS and interpretation systems of each sequencing laboratory). This solution includes the complete storage of the results of a genetic test carried out with NGS sequencers, since it allows for storing both the genetic variants and the sequenced chromosomal ranges where no variants were found. For this reason, we consider that it is a great contribution to introduce Precision Medicine in a health institution through a CDSS. On the other hand, NGS technologies generate raw data that occupies hundreds of gigabytes of storage. Although these raw files must also be kept for backup, security and possible future reanalysis purposes, their incorporation into the architecture is still challenging.

5. Conclusions

The implementation of an architecture such as the one presented in this work makes it possible to incorporate germline genomics into EHR, allowing for health professionals who are not specialists in genetics to incorporate genomic data into clinical decisions and therefore improve patient care.

References

[1] Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, NIH. 2015. The Precision Medicine Initiative Cohort Program: Building a Research Foundation for 21 st Century Medicine. [cited 7 Nov 2022]. Available: https://cutt.ly/PMICP
[2] Katsnelson A. Momentum grows to make “personalized” medicine more “precise.” Nat Med. 2013;19: 249–249.
[3] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17: 333–351.
[4] Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376: 44–53.
[5] Sitapati A, Kim H, Berkovich B, Marmor R, Singh S, El-Kareh R, et al. Integrated precision medicine: the role of electronic health records in delivering personalized treatment. Wiley Interdiscip Rev Syst Biol Med. 2017;9. doi:10.1002/wsbm.1378
[6] Schickhardt C, Fleischer H, Winkler EC. Do patients and research subjects have a right to receive their genomic raw data? An ethical and legal analysis. BMC Med Ethics. 2020;21: 7.
[7] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6: 80–92.
[8] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158.
[9] Carlson RW. Breast Cancer: NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®). 2012.
[10] HL7 FHIR Release 4.3. [cited 4 Feb 2023]. Available: https://www.hl7.org/fhir/?
[11] Den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016;37: 564–569.
[12] Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res. 2021;49: D939–D946.
[13] Red Team at Broad Institute. Home - Cromwell. [cited 4 Feb 2023]. Available: https://cutt.ly/CROM