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

doi:10.3233/SHTI220189

Automatic Integration of Clinical and Genetic Data Using cBioPortal

Mauricio Brunner[a], Lucas Mullen[b], Federico Jauk[c,e], Javier Oliver[c,h], Federico Cayol[d], José Minata[d], Victor Herrera[f], Walter Pavicic[g], Daniel Luna[a,e], Marcelo Risk[e], Hernan Garcia Rivello[b,e], Sonia Benitez[a,e]

[a] Departamento de Informática en Salud, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
[b] Servicio de Anatomía Patológica, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
[c] Laboratorio de Secuenciación, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
[d] Servicio de Oncología, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
[e] Instituto de Medicina Traslacional e Ingeniería Biomédica, CONICET-IUHI-HIBA, Buenos Aires, Argentina
[f] Departamento de Investigación, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
[g] Instituto Multidisciplinario de Biología Celular (IMBICE), CONICET-CIC-UNLP, Buenos Aires, Argentina
[h] Laboratorio de Biología Molecular del Cáncer. Instituto de Investigaciones Biomédicas de Málaga (IBIMA), Málaga, España

Abstract

Precision medicine seeks to improve the prevention, diagnosis and treatment of patients based on genetic characteristics unique to each person. In oncology, therapeutic decisions have been established based on the genomic characteristics of each patient's tumor.

Data integration is key for the successful implementation of precision medicine since it is necessary for both studying a large volume of data from different sources and working with an interdisciplinary and translational vision.

In this work, a bioinformatic process was successfully implemented that allows the integration of patients' genomic data, from two molecular biology laboratories, with their clinical data provided by their electronic medical records. For this, the REDCap data capture software, the cBioPortal visualization and analysis software, and a computer tool developed to automate the processing and annotation of the information in REDCap were used to be included in cBioPortal, for the "Map of Tumor Genomic Actionability of Argentina" project.

Keywords:

Genomics, Precision Medicine, Medical Informatics.

Introduction

Precision medicine refers to the identification of risk factors and the adaptation of management and treatment options according to the individual characteristics of each patient [1, 2]. Common challenges faced in precision medicine projects include the integration of clinical information with data generated from different sources, such as molecular biology laboratories or DNA sequencing, and the development of interoperable medical records [3]. This integration involves combining data that resides in different sources and providing users with a unified view of it [4]. In Spain, for example, most of the initiatives developed for the implementation of precision medicine arise from projects of sequencing and integration of genomic and clinical data [5].

The Tumor Genomic Actionability Map of Argentina (MAGenTA, by its acronym in Spanish) is a research and development project in precision medicine, and a Technological and Social Development Project (PDTS-CONICET), executed by the Public-Private Associative Consortium (CAPP MAGenTA) constituted between the Hospital Italiano de Buenos Aires (HIBA) and the National Council for Scientific and Technical Research (CONICET), through its Multidisciplinary Institute of Cell Biology (IMBICE) Executing Unit. The main focus of the MAGenTA project is on the clinical actionability of oncological drugs, since in Argentina there are few data on genomic alterations in solid tumors, including studies of single genes or gene panels in small cohorts.

The data generated by the project are publicly accessible and combine genetic and clinical information, thus characterizing the population to allow the health community to know which treatments are currently applicable and, in the future, to know in which clinical trials it is valid to include the patients because they are potentially more sensitive to the drugs under study. Furthermore, they allow planning by the Argentine health system according to the prevalence of molecular alterations in our population [6,7]. In order to carry out MAGenTA project, a second-generation massive sequencing platform (Next Generation Sequencing, or simply NGS) was put into operation at HIBA, working in conjunction with IMBICE, which develops genetics studies for analysis of population ancestry. An NGS approach enables the detection of diagnostic, prognostic, predictive, and follow-up biomarkers to develop cost-effective therapeutic strategies.

Data integration under MAGenTA project is laborious, time-consuming, and has a high probability of transcription error because the process is done manually, using different file formats and processed by different bioinformatics tools. Furthermore, this process is repeated many times per year each time new patients are added to the study and their DNA is sequenced. cBioPortal for Cancer Genomics is a web platform for exploring, visualizing, and analyzing multidimensional cancer genomic data [8]. This resource was specifically designed to reduce access barriers to complex data sets, accelerating the translation of genomic data into new biological knowledge, therapies and clinical trials [9].

In this study we describe the computer strategy implemented to automate the integration into cBioPortal of clinical and genomic data from the electronic health record (EHR), the HIBA Sequencing Laboratory (LS-HIBA) and the Population Genetics Laboratory of the IMBICE (LGP-IMBICE) to reduce errors and operator time and to obtain a simplified view of the MAGenTA project dataset.

Methods

This project was carried out at HIBA, a highly complex hospital, with a Health Information system and an EHR certified at level 7 by the Health Information Systems and Management Society (HIMSS). The MAGenTA study included patients with a diagnosis of cancer due to solid tumors who had a tissue sample from a previous surgery or biopsy, sufficient to be able to carry out the study methodology. Patients who refused to participate or give informed consent were excluded.

The Department of Health Informatics (DHI) was in charge of all software implementation at HIBA. In the following paragraphs, the sources of information, the data capture software, the data integration and visualization process are described.

Sources of information

Clinical information

All the clinical information was obtained from the EHR. The clinical information selected to be displayed was: age, sex, history of diabetes, history of alcoholism, history of smoking, type of cancer, tumor stage and primary tumor surgery.

Genetic information

The genomic information was obtained from the HIBA sequencing laboratories (LS-HIBA) and the IMBICE Population Genetics Laboratory (LGP-IMBICE).

The LS-HIBA performed the sample sequencing . For this, the amplicon-based oncology panel "Oncomine Focus Assay" (OFA) [10] was used. For each tumor sample, single nucleotide variants (SNV) and small insertions and deletions (Indels) have been analyzed across 35 genes from DNA samples, and genetic fusions across 23 genes from RNA samples. In total, 52 genes clinically relevant to solid tumors were analyzed.

In the LGP-IMBICE, the population ancestry in the maternal line was analyzed from the mitochondrial DNA of the samples, which makes it possible to determine the haplogroups to define the maternal ancestry of each one of them.

Data storage

REDCap [11,12] is a data capture software that allows us to create dynamic forms to upload information to your database. REDCap version 7.0.11 was used as a storage layer for the study data of the MAGenTA project (Figure 1).

The forms created for this project allowed us to collect clinical information and metadata. Similarly, for genomic information, although the ancestry data were loaded directly into REDCap, it was not the same for the information from NGS sequencing. For this type of data, in REDCap metadata were stored that were then used to obtain genetic information from the VCF (Variant Call Format) variant files and the reports resulting from the analysis, directly from the analysis system data set Ion Reporter [13], version 5.6, used in the HIBA sequencing laboratory (Figure 1).

Data processing

A computer program developed in the DHI in WDL (Workflow Description Language) version 1.0 was used to populate REDCap forms for integration into cBioPortal. This language helps define processing tasks, associate them in workflows (Pipelines) and parallelize their execution, with a readable syntax. The pipeline management system used to run the program was Cromwell [14], a system oriented to scientific workflows.

A computer program developed in the DHI in WDL (Workflow Description Language) version 1.0 was used to populate REDCap forms for integration into cBioPortal. This developed pipeline performed different tasks. The pipeline scripts were developed in the Python programming language, version 3.6, and R, in version 3.4.

To load the "Tumor Type" data in cBioportal, a mapping was performed between the values loaded in REDCap for this variable and the codes for these types of cancer that cBioPortal uses as a short name. To obtain these codes, the application programming interface (API) provided by cBioPortal for programmatic access to data was used. To obtain the records uploaded to REDCap, the API provided by the platform that allows external applications to connect to the system remotely to retrieve or modify data or settings on a scheduled basis was used. To obtain the reports and the VCF variant files that result from the NGS sequencing process, the web services API offered by Ion Reporter was used, which allows automating queries to retrieve information from the system.

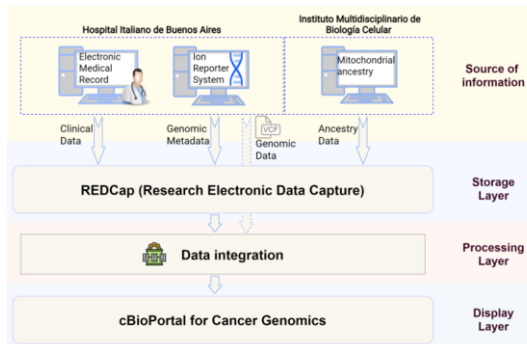

Figure 1 - Information flow. Different levels are observed within the information flow in the process of integrating data from different sources to the cBioPortal viewer.

Information visualization

The data generated by the MAGenTA project were published on the web through the cBioPortal for Cancer Genomics platform. This tool allows fast, intuitive and high-quality access to the molecular profiles and clinical attributes of the MAGenTA project.

To upload study data to cBioPortal it is necessary that they be represented in a set of data files in a specific directory. All the data and metadata files that reference them must be placed in this directory, where each specific type of data must comply with the format required by cBioPortal. This set of files can then be used to import the information into the visualization platform's database and thus make it available as a new study in it.

To help in the import and assembly process of some types of files, cBioPortal provides different tools in the form of scripts, such as validateData, cbioportalImporter and vcf2maf. ValidateData allows us to verify the file formats before

Table 1 – Data upload forms implemented in REDCap

Name	Type	Description
General data	Patient	Allows loading of general patient data such as date of birth and gender.
Comorbidities	Patient	In this form, the history of smoking, diabetes, alcoholism and whether or not the surgery is for a primary tumor are loaded.
Oncology	Patient	Contains options that allow you to load information related to the patient's cancer, such as tumor stage and whether or not it has metastasized, adjuvant chemotherapy, disease relapses, or first-line chemotherapy.
Death record	Patient	Form used to store information about the death of the patient or the date of the last contact with his oncologist.
Ancestry	Patient	It allows the loading of mitochondrial DNA haplogroups.
Pathological anatomy	Sample	From this form, data related to the protocol, the paraffin plug, the date of the protocol and the type of tumor are loaded.
Sequencing laboratory	Sample	In this form, the genomic metadata is loaded, which then allows the genetic variant files to be extracted from the servers of the DNA sequencing equipment.
Oncology Target - Immuno	Sample	Data related to targeted therapies, such as dates, progression, drugs used, and immunotherapy are loaded here. This form can be repeated as many times as necessary if more than one targeted therapy was applied.

importing them and cbioportalImporter allows us to import the set of files with the study information to cBioPortal. Vcf2maf helps us create the mutation data file. This file is the one that contains the information for each sample regarding genetic variants of the SNV and Indels types and must be in MAF (Mutation Annotation Format) mutation annotation format. To use vcf2maf, it is necessary to start from the VCF variant files and have the Variant Effect Predictor (VEP) annotator [15] to obtain extra information on the genetic variants and thus be able to annotate them.

On the other hand, cBioPortal integrates automatically with various external services to provide more information on genetic variants. One such service is OncoKB, which is a precision oncology knowledge base that contains information on the effects and implications of cancer-specific genetic alterations treatment [16]. Since Ion Reporter and OncoKB do not represent fusions in the same way, this database was used to be able to represent Ion Reporter fusions in a format that can be successfully integrated into cBioPortal.

cBioPortal was implemented in version 1.14.1 on the servers of the Hospital Italiano de Buenos Aires.

Results

During 2020, 8 data upload forms were implemented in REDCap, where each of them have one or more responsible for their upload, depending on the area involved. The areas involved were: anatomic pathology, oncology, the HIBA sequencing laboratory and the IMBICE Population Genetics laboratory.

Of the 8 forms created, 5 of them are related to patient information and 3 to information about their samples. The latter are repeated as many times as necessary to increase the biopsy load capacity per patient. The implemented forms are summarized in table 1.

For the process of integrating the data and metadata stored in REDCap into cBioportal, a computer tool or pipeline for processing and annotation of clinical and genomic data was developed. Each of the processes that occur within the pipeline are processed with a different script. They can be seen in Figure 2 and are described below.

1) Getting the raw data from REDCap

The pipeline begins with the download of clinical data and genomic metadata uploaded to the RedCap platform, where members of the research team upload the data from the MAGenTA project. This information is downloaded in CSV (Comma-separated values) format using the API provided by REDCap to interact with the system. Each record in the downloaded file corresponds to an anonymized patient.

2) Filtering records

The next step in the pipeline, takes the previously downloaded file and performs a filter whose purpose is to discard incomplete records. This filtering differentiates between the data that are indispensable and those that are not. Among the essential information, it is required that the variables have their value assigned: date of birth, date of completion of the hospital protocol, type of tumor, and analysis code. The analysis code is the code supplied by the genetic sequencing team, from which the results of the sequencing were interpreted and allows us to obtain both the report generated with the genetic variants of interest and the VCF files with the results of said sequencing from the Ion Reporter system.

During this process, a file where the filtering events are recorded (Log File) is created.

3) Annotation

This step processes the filtered file from the previous step. The cancer types loaded in the REDCap forms are mapped with the abbreviated defined cancer type that is used in cBioPortal to associate a sample with a predefined type of cancer. Another process that occurs here is adding information to the file to enrich its content, and making modifications to its format, such as translations from English to Spanish. In this step, the patient's age is also calculated based on their date of birth and the date of the protocol, and the information of the samples is anonymized. Once this processing is finished, a new file is obtained with the enriched information, and a second Log file in which the observations made during the process are recorded, such as a wrongly loaded date or an incoherent age of the patient. The file with information will be used by subsequent steps to obtain data necessary for their respective processing.

4) First report: records discarded from REDCap

From the two Log files generated in the two previous steps, a script generates a PDF report (Portable Document Format). The

report classifies the information into Errors, Information and Warnings and concisely shows the observations made by the filtering and annotation processes. Logs classified as Errors will not be uploaded to cBioPortal.

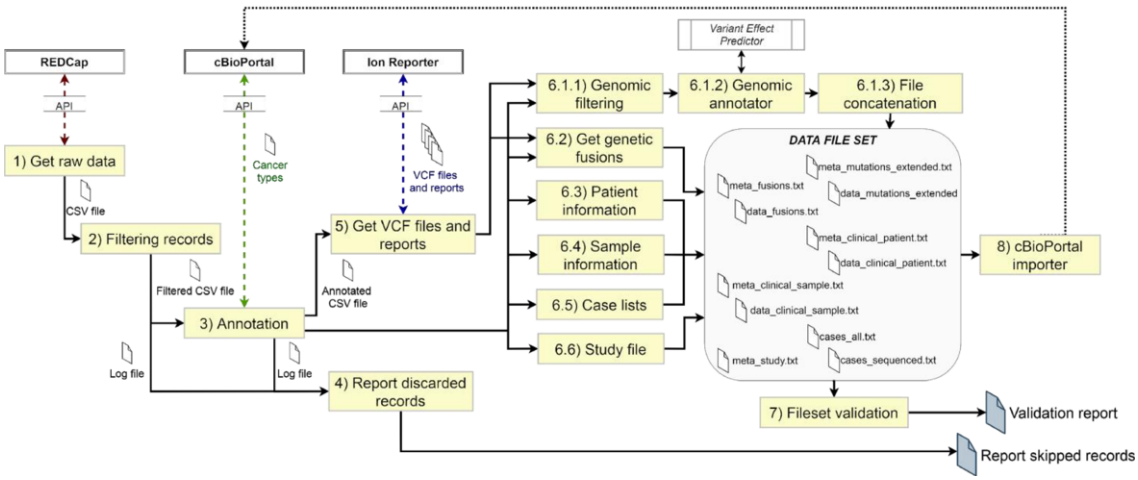

Figure 2 - clinical and genomic data processing and annotation pipeline developed for the integration of data and metadata stored in REDCap and Ion Reporter to cBioportal.

5) Getting VCF files and reports

In this step, information from the Ion Reporter sequencing laboratory reporting system is downloaded into a predetermined data directory. The information obtained is the report published in the final analysis of the sequencing process and the complete VCF genetic variant file.

Then, taking as input the enriched file generated in the annotation step and the reports and VCF files obtained in the previous step, the 6 processes that assemble the data set are triggered in parallel with the formats that cBioPortal needs to be able to load information to the portal.

6.1) Generation of genomic information files: SNVs and INDELs

The first of these processes generates the files that contain the genetic information corresponding to SNV and Indels. To produce these files 3 different scripts have been used:

6.1.1) Genomic filtering

In this step, unreported gene variants and fusions are filtered from the VCF file of each sample.

6.1.2) Genomic annotation

A script processes each filtered VCF file using the vcf2maf tool. This tool annotates VCF files using the VEP annotator and formats them to the mutation annotation format (MAF).

6.1.3) File concatenation

The third script concatenates all the MAF files into a single MAF file that contains the SNVs and Indels of all the samples. In this same step, the metadata file that cBioPortal needs to reference the MAF file is also generated.

6.2) Generation of genomic information files: Fusions

In this process, data on gene fusions are obtained. We obtain this information from previously downloaded reports. Fusions found in the reports with their corresponding identifier in the database OncoKB to be displayed properly on cBioportal are mapped. In this step, an only fusion file is created for all the samples and the metadata file that references it with the proper format to be able to upload the fusions to cBioPortal.

6.3) Generation of patients' clinical information files

Here the data and metadata files are generated with the clinical information of the patients. The clinical data file is a two-dimensional matrix with the clinical attributes corresponding to the comorbidities and general data of the patients.

6.4) Generation of samples' clinical information files

In this process a script generates the data and metadata files with the clinical information of the analyzed samples, including the type of cancer and the patient code to which they belong to be able to make the mapping between the patient identifiers and samples.

6.5) Generation of case lists files

This script generates files where list all the samples that may be selectable to make different types of queries in cbioPortal. For MAGenTA project this list contains all analyzed samples.

6.6) Generation of MAGenTA study file

In this step, a single file containing the metadata of the MAGenTA study is generated, including an identifier, a name, a description, and the types of cancer that we can find within the project. At this point in the pipeline, all the files that make up the complete data set necessary to be loaded into cBioPortal with the study data for the MAGenTA project have already been generated.

7) Second report: data set validation

The last step in this process uses the validateData tool. This tool is provided by cBioPortal and facilitates the loading of new studies in its database by allowing the automatic validation of the format of each of the files that make up the study data set.

Once the dataset has been validated, the next step is its upload to cBioPortal. For this purpose, we use the cbioportalImporter tool, provided by cBioPortal too. This tool was excluded from the pipeline since it is necessary to perform a manual interpretation of the validation to corroborate the integrity of the format of the files in the data set.

The tool developed allowed us to update the information on the portal in a process of approximately 30 minutes. This process obtains the REDCap data, integrates it with the genomic data

stored in the Ion Torrent system, processes it, annotations, and generates a set of files that is used to validate and later upload to cBioPortal. When this procedure had been performed manually, for 93 samples, it had required approximately 24 hours of hand make work.

In this work the data corresponding to 186 samples evaluated in the MAGenTA project have been integrated so far. Loading data cBioPortal yielded the first preliminary results of the Project Magenta, among which are: The age of the patients included in the study ranged between 0 and 90 years, with a female-male ratio of 1.5: 1. 33 patients with tumor stage I, 37 with stage II, 37 with stage III, 50 with stage IV, and 23 patients with no data. Between 0 and 4 clinically relevant genomic alterations were detected per patient with the genetic panel studied, including SNV, Indels and gene fusions, with an average of 0.91 alterations per patient. At the DNA level, 154 variants were detected in 126 patients, along 28 genes. The genes with the highest prevalence of variants were KRAS (30 cases), BRAF (25 cases), PIK3CA (18 cases). 10 of the 35 genes included in the panel (DNA) did not show clinically relevant variants. At the RNA level, 11 gene fusions were detected in 11 patients. The genes with the presence of fusions were ALK (5 cases), ROS1 (2 cases), BRAF, MET, NTRK3 and RET (1 case each). 17 of the 23 genes included in the panel (RNA) did not show fusions. More than one variant was detected at the DNA and / or RNA level in 34 patients.

The cBioPortal visualization and analysis platform implemented with the information generated in the MAGenTA study can be accessed from https://magenta.hospitalitaliano.org.ar.

Discussion

Combining the use of REDCap with the cBioPortal platform and a bioinformatic pipeline for data integration between both systems, the system was created that allows the registration, storage, exploration and visualization of data from different sources that are generated in the medicine project of precision in oncology MAGenTA.

cBioPortal is an easy-to-use tool allowing the visualization and analysis of the data to obtain preliminary results from the MAGenTA project, but the assembly of the data set to be uploaded to the platform can be complicated if you do not have technical knowledge and can take several days. The use of a bioinformatics pipeline decreased the hands-on time for integrating the information stored in REDCap and in the DNA sequencers to the cBioPortal platform. In addition, it made the report generation possible that summarize the content of the REDCap records and alert to possible incorrectly loaded or incomplete data, and that the process of updating data in cBioPortal does not require advanced technical knowledge to be carried out.

Developing this pipeline is useful for research projects in precision medicine such as MAGenTA, since the number of cancer patients included in it is prospective, that is, it increases month by month for an indeterminate time and periodic updates are necessary in the analysis platform cBioPortal.

Conclusions

The combination of REDCap and cBioPortal tools, plus a computer pipeline to integrate them, is a very good option for precision medicine projects in oncology in which the information must be updated with a specific frequency and over time, as it is collected.

Acknowledgements

We want to thank Heidy Díaz de Arce, Leandro Ortega, Marcela Martinez, Miguel Fantin, Andrea Mayordomo and the MAGenTA group for their contribution to this project.

References

[1] Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med. 2010;363: 301–304.
[2] Bousquet J, Anto JM, Sterk PJ, Adcock IM, Chung KF, Roca J, et al. Systems medicine and integrated care to combat chronic noncommunicable diseases. Genome Med. 2011;3: 43.
[3] Propuesta de Recomendaciones para una Estrategia Estatal de Medicina Personalizada de Precisión. [cited 20 Nov 2020]. Available: https://bit.ly/32vvXqG
[4] Lenzerini M. Data integration: a theoretical perspective. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02. New York, New York, USA: ACM Press; 2002. p. 233.
[5] Medicina Personalizada de Precisión en España: Mapa de Comunidades. [cited 20 Nov 2020]. Available: https://bit.ly/3gr1pyj
[6] Medicina de precisión: se presentó plataforma de secuenciación masiva del genoma. In: Argentina.gob.ar [Internet]. 25 Oct 2019 [cited 19 Jun 2020]. Available: https://bit.ly/3emeNkZ
[7] MinCyT. MIA Medicina de precisión. In: Mercado de Innovación Argentina [Internet]. [cited 19 Jun 2020]. Available: https://mia.gob.ar/proyectos/seguimientos/45
[8] Fang P, Yan Z, Labrousse P, Liu W, Biroschak J, Wright J, et al. Abstract 1397: Oncomine focus assay: Simultaneous detection of clinically relevant hotspot mutations, CNVs, and gene fusions in 52 oncogenes relevant to solid tumors. Clinical Research (Excluding Clinical Trials). 2016. doi:10.1158/1538-7445.am2016-1397
[9] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42: 377–381.
[10] Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform. 2019;95: 103208.
[11] Ion ReporterTM Server System. [cited 6 Aug 2020]. Available: https://bit.ly/3v9QVrx
[12] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2: 401–404.
[13] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6: l1.
[14] Red Team at Broad Institute. Home - Cromwell. [cited 28 Jul 2020]. Available: https://cromwell.readthedocs.io/
[15] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17: 122.
[16] Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;2017. doi:10.1200/PO.17.00011

Address for correspondence

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