A CENTRALIZED PLATFORM ON HUMAN GENOME FOR SUPPORTING CLINICAL DECISIONS

M.Sc Andrêza Leite de Alencar[1], Dr. Vanilson Burégio[1], M.Sc Jamisson Freitas[2], M.Sc Marcel Caraciolo[2], Dr. Vinícius Cardoso Garcia[3]

[1] Departament of Statistics and Informatics of Federal Rural University of Pernambuco– UFRPE, Recife, Brazil.
[2] Genomika Diagnostics, Recife, Brazil.
[3] Informatics Center of Federal University of Pernambuco – UFPE, Recife, Brazil

Resumo: A Integração de dados é um desafio na área de genética clínica onde analistas precisam manipular múltiplas fontes de dados heterogêneas sobre domínios biológicos e clínicos. Esta pesquisa tem como objetivo prover acesso unificado a essas diversas fontes de dados para dar suporte a variadas decisões clínicas. Neste contexto, um trabalho vem sendo realizado para o projeto e implementação de uma plataforma que permita a integração e o acesso unificado a fontes de dados heterogêneas. Neste artigo detalhamos a arquitetura definida para a plataforma e um cenário de uso *online*, para processamento e anotação de variantes clínicas, que utiliza os principais repositórios de dados públicos de genoma, tais como OMIM, Clinvar, LOVD, ExAC6500, entre outros. Como resultado, esta plataforma provê um serviço de acesso unificado a dados, auxiliando no processo de análise e extração de conhecimento biomédico e suportando o diagnóstico de tumores, cânceres e doenças raras.

Palavras-chave: Heterogeneidade de Dados; Big Data; Genômica.

Abstract*: Data integration is a challenge in the field of clinical genetics, where analysts need to handle multiple heterogeneous data sources on biological and clinical domains. This research aims at providing unified access to these diverse data sources to support various clinical decisions. In this context, a work has been developed for the design and implementation of a platform that allows unified access and integration of heterogeneous data sources. In this paper, we detail the proposed platform architecture and an online usage scenario for processing and annotation of clinical variants, which uses main public genome data repositories such as OMIM, Clinvar, Lovd, ExAC6500, and others. As result, the platform provides a unified data access service to support the process of analysis as well as extraction of biomedical knowledge, helping with the diagnosis of tumors, cancers and rare diseases.*

Keywords*: Data Heterogeneity, Big Data, Genomics*

Introduction

In the last decades the enhancement in computing power has produced impressive data flow expansion which has caused a paradigm shift in the processing of large-scale data. As a result, we have had a boost in volume, velocity and variety of types and sources of data, commonly referred to as Big Data[1]. Scientific research in many fields of knowledge, such as particle acceleration and genome sequencing, produce such massive amounts of data. The access to these various types and sources of data are not frequently standardized, though. That reality leads data scientists to different challenges.

Having an efficient data integration is one of the main challenges faced by professionals of clinical genetics who deals with multiple heterogeneous and distributed catalogs of human genes and genetic disorders. Generally, features of the existing sources of data, used by such professionals, do not establish standards and consequently a diversity of formats and types are available. In applications of variant analysis for molecular diagnostics, for example, a key task consists of matching biological information with clinical data in a way that specialists can determine the potential impact of variants associated to diseases[2,3]. In practice, this matching process requires the use of updated sources of clinical data which typically makes biologists and geneticists spend many labour hours on activities like searching for, parsing, cleaning and integrating data from several databases in complex spreadsheets.

This work presents a platform which offers integrated access to data from a variety of public and private foundations, including OMIM, ClinVar, RefGene, LOVD and ExAC65000. The platform provides a unified set of services, built on top of a consolidated basis of human genome data, that enables data analysis (to perform annotations of variants) as well as biomedical knowledge extraction. The implementation of such platform provide interfaces designed to abstract the complexity of manipulating heterogeneous data sources, in order to simplify the diagnosis of tumors, cancers and rare diseases.

Background

The concept of Big Data is strongly related to the data deluge[4] phenomenon. The data deluge refers to the situation in which the exponential growth in the generation of new data makes its management and analysis increasingly complex. The data deluge phenomenon can be illustrated by the growth of genetic sequences databases. An example of such databases is the GenBank[5], held by the NIH (USA), which consists of an annotated collection of all publicly available DNA sequences. Currently, the Genbank doubles in size every 18 months. This trend has been confirmed in previous years and should be maintained in the years to come.

In fact, genome sequencing is a pioneering application of Big Data. A single human genome consists of about 3 billion base pairs of DNA, and new generation sequencing technologies makes complete sequencing of large-scale genome feasible in terms of cost and time.

The characterization of large-scale human genome involves the generation and interpretation of a huge volume of data in an unprecedented scale, and one of the potential benefits is personalized medicine for supporting clinical decisions. It can impact directly in cancer patients, for instance[6].

There are several projects focused on characterizing large-scale human genome including OMIM, ClinVar, RefGene, LOVD, ExAC65000, 1000 Genomes, among others. Other ongoing projects focus on the characterization of somatic mutations (substitutions, insertions, deletions, etc.) of various cancers. The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are examples of these projects.

A problem for the use of Big Data on health is how to enable identification, access, and citation of (i.e., credit for) biomedical data[7]. Inherent to data discovery is the need for a sustainable and scalable plan to create and maintain a discovery system that allows researchers to readily find and cite biomedical data. Indeed, sustainability and scalability are two intertwined issues that must be addressed so the Big Data to Knowledge (BD2K)[7] can have a lasting effect. An important first step has been to recognize the necessity of assembling and validating ideas drawn from the broader scientific community for developing a Data Discovery Index (DDI)[7].

Although there are many research projects and tools facing those big data challenges on human genome data, an advance is still needed in order to build a system that provides unified access to data from various and heterogeneous sources. It is worth noting that commercially available software for analysis variants are extremely expensive and are not extensible/customizable. Each laboratory has its own internal flow with specific databases, and requiring different update flows.

This paper is organized as follows: Section Methods presents the platform proposal and how it was designed; Section Results presents preliminary results obtained in this research; Section Discussion presents a discussion about practical issues and experiences on this work; Section Conclusion presents conclusion remarks and future work; and Section Acknowledgments.

Methods

This work presents a platform which offers unified access to data from a variety of public and private foundations. Our platform provides specialized services for data access (on a centralized basis of human genome data), data analysis (performing annotation of variants), and biomedical knowledge extraction. The methodology chart presented by Figure 1 shows the development process of this research.

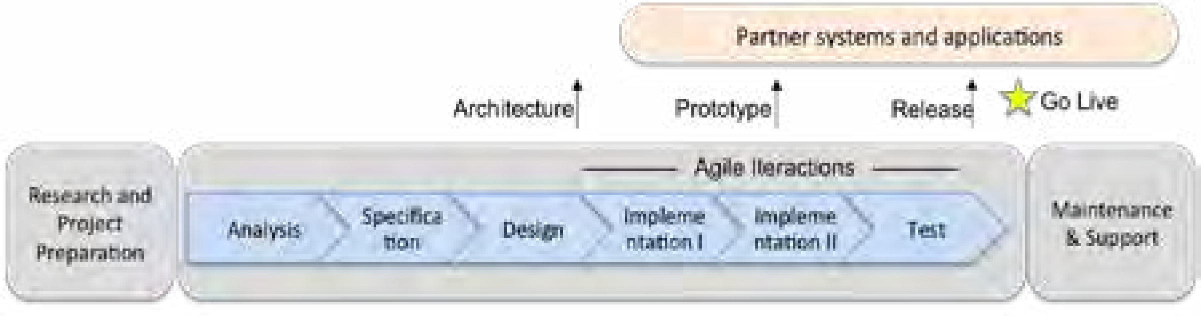

Figure 1. The methodology overview

Next sections present the main components of the platform architecture, an example of use and platform implementation.

a. Architecture overview

The platform architecture follows the lambda architecture framework[8]. Thus, it is a data-processing architecture designed from the ground up for handling massive quantities of data by taking advantage of both batch- and stream-processing approaches. The platform is composed by four layers, namely Batch Layer, Serving Layer, Speed Layer and Meta Layer. These layers work together to provide different views on data that can be consumed by other applications (or "apps" for short). Figure 2 presents an overview of the proposed platform and illustrates the multiple data sources it uses, its different layers, provided data views, interactions between layers and data-consuming apps.

- Batch Layer: It stores the master copy of the dataset and pre-computes batch views on that master dataset. The master dataset can be thought of us a very large list of records. The batch layer needs to be able to store an immutable, constantly growing master dataset, and compute arbitrary functions on that dataset. The key word here is "arbitrary." If you're going to pre-compute views on a dataset, you need to be able to do so for any view and any dataset[8].
- Serving Layer: The batch layer builds batch views as the result of its functions. The next step is to load the views somewhere so that they can be queried. Here the serving layer comes in. The serving layer indexes the batch view and loads it up so it can be efficiently queried to get particular values out of the view. The serving layer is a specialized distributed database that loads in batch views, it also makes them able to query, and continuously swaps in new versions of a batch view as they are computed by the batch layer. Since the batch layer usually takes at least a few hours to do an update, the serving layer is updated every few hours[8].

In the batch and serving layers there are no concurrency issues to deal with, and it trivially scales. One missing property is low latency updates. The speed layer provides this property.

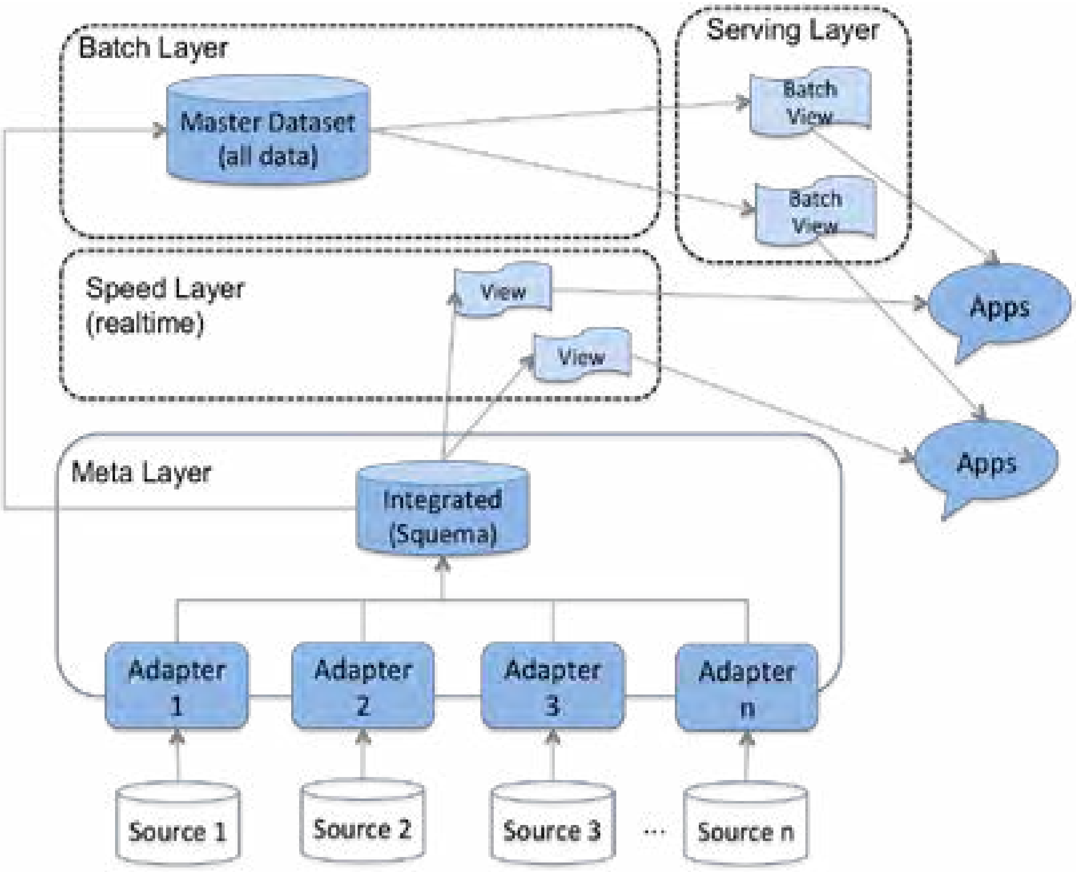

Figure 2. The platform architecture overview

- Speed Layer: The only data not represented in the batch views are data added while the pre-computation is running. At this point, in order to have a fully real time data system, it remains necessary to compensate for those last few hours of data. This is the purpose of the speed layer. Speed layer is similar to the batch layer since it produces views based on the data it receives. There is one big difference though: in order to achieve the fastest latencies possible, the speed layer does not use all new data at once (re-computing them, like the batch layer does). Instead, it updates the real time view as it receives new data. This is called “incremental updates” as opposed to “re-computation updates”. Another big difference is that the speed layer only produces views on recent data, whereas the batch layer produces views on the entire dataset[8].

Another missing property of the platform is the capable of access data from various and heterogeneous sources. The Meta layer fixes this problem.

- Meta Layer: This layer establishes contracts between the platform and the various data sources. It is responsible for acquiring data coming from heterogeneous sources prior to storing and publishing them. It means that the other layers receive the data with the unified representation model arising from the meta layer. The component Adapter converts the scheme of the data from the specifics source models to a common global meta-model (or simply model) adopted by the platform and unifies the internal representation of the data. A translation process is defined according to the structures of the data sources.
- Apps: Real-time and batch data views - provided by the Speed and Serving Layers, respectively - are published as services by the platform as a way to minimize the complexity of applications to use and handle the existing datasets on human genome. Different client and third-party applications can be built on top of our proposed platform. A very important type of client application is data analysis systems for performing annotation of variants. These systems

include the addition of various metadata to the detected variants, as the chromosomal position, functional effect of the protein and other data, allowing filtering and prioritizing variants in an automated way.

b. Use scenario

The platform's client systems for analysis can be applied in the scenario described as follows.

Imagine that the mastology division of a local hospital sends samples to a laboratory for research and diagnosis of genes related to breast cancer - BRCA1 and BRC2. After sequencing, the DNA samples would be added to the platform run queue for analysis of mutations in these two genes in a volume of tens of hundreds per month. The variants analysis system identifies the end of the sequencing reaction, and uploads the files of the patients to the server in order to start the analysis pipeline.

With the presence of parallel processing and high performance algorithms, the platform can perform (in hours) the analysis of these samples. After identifying DNA variants and querying clinical databases, the platform makes the ranking and classification of variants according to the possible pathogenicity, and data become available through a friendly web interface, where the analyst can check, filter and select the relevant genetic alterations. Finally, the specialist completes analysis and makes a review of the digital report, which is automatically filled by the platform. The new procedure increases productivity by having specialist's focus on the process analysis, rather than on the tool. This advantage justifies the implementation of such a heavy tool.

The platform also can support the specialist in order to filter out specific variants that shoud not be prioritized and help him at one of the main issues in the variant's report conclusion, which is the classification of the variants. Generally, the biologist or the geneticist must have to apply several filters among multiple parameters (columns or annotations in this context) to prioritize the most important variants that could explain the clinical hypothesis of the patient. The main issue is that the current tools such as spreadsheets and business intelligence databases might lead to long and complex queries when the analysis becomes more specific with several filters and options. The platform's client using big data concepts helps not only the final user on querying multiple variants but as also the bioinformaticians responsible to gather all data from several data sources and link them in a clear and schematic way without complaining about data schema.

c. Implementation outline

Computing arbitrary functions and analysis on an arbitrary dataset in real time is a daunting task. No single tool provides a complete solution. Instead, one has to use a variety of tools and techniques to build a complete Big Data system. For this purpose, we use the Lambda Architecture[8], which is designed to solve the problem of computing arbitrary functions on arbitrary data in real-time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. In addition to those layers, we define the Meta Layer, (which we have already presented) to insert new functions and address heterogeneity problems.

There is a class of systems called "batch processing systems" that are built to do exactly what the batch layer requires. They are very good at storing immutable, constantly growing datasets, and they expose computational primitives to allow you to compute arbitrary functions on those datasets. Hadoop is the canonical example of a batch processing system, and we will consider to use it in the next versions of the platform. In this first version we use the Data Base Management System (DBMS) PostgreSQL (Specifically the JSON Field) to implement the properties of the batch layer.

A serving layer database only requires batch updates and random reads. It does not need to support random writes. This is a very important point, as random writes cause most of the complexity in databases. By not supporting random writes, serving layer databases can be very simple. This simplicity makes them robust, predictable, easy to configure and to operate. ElephantDB is an appropriate database to implement the serving layer requirements.

The speed layer requires databases that support random reads and random writes. Because these databases support random writes, they are orders of magnitude more complex than the databases used in the serving layer, both in terms of implementation and operation.

The meta layer uses translation algorithms, implemented in Python language, which converts data from the source models to the common model adopted inside the platform. The schema presented by Figure 3 shows the model components structure and relationships. The constructors established for the initial model have been defined to allow storage of information such as repositories (Depository) and their versions (Version). It also allows traceability and versioning of data. Other constructors are Dataset (which can be compared to a table or relationship in a ER); records (Register, which can be compared to a tuple in a model ER) and metadata that describe the data (DataField which can be compared to a column of a relation in an ER model). It also maintains a log (Logger) for saving a historic of operations performed on the data.

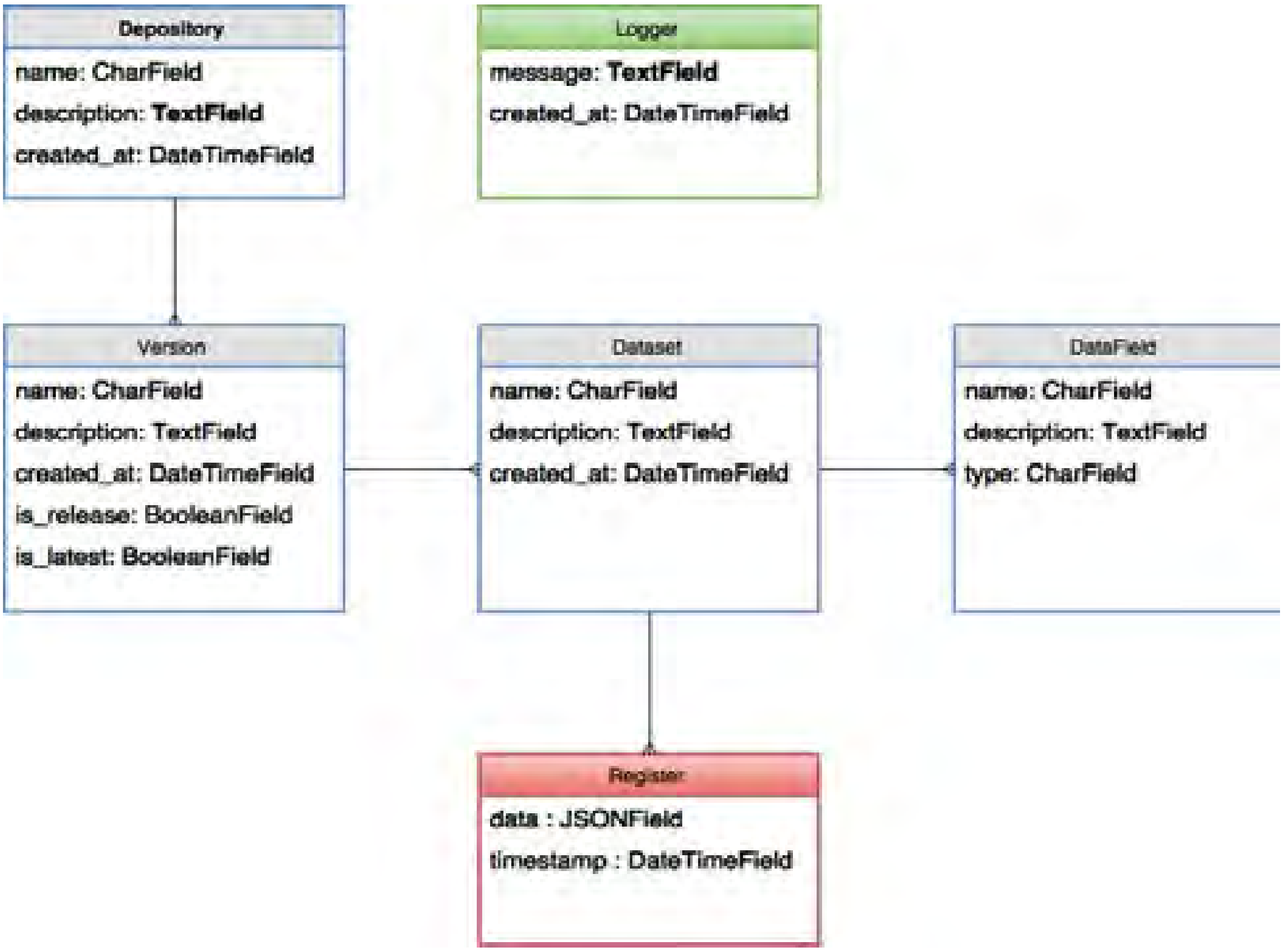

Figure 3. Meta-model

The platform client system for analysis was implemented in Python language and screams of the annotation system is presented by Figure 4 and 5.

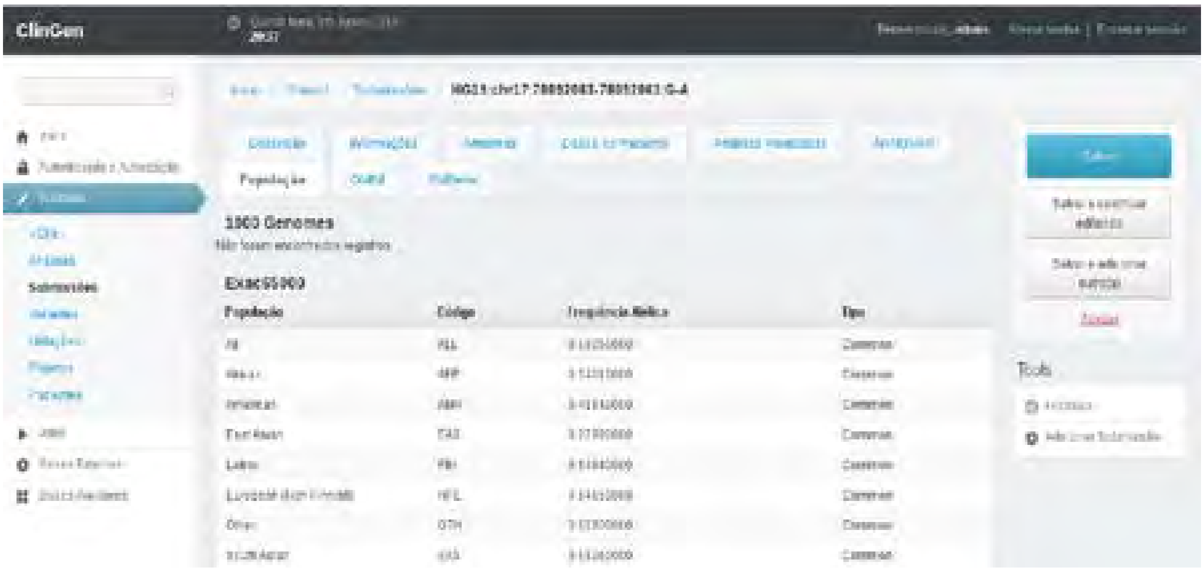

Figure 4. Annotation System

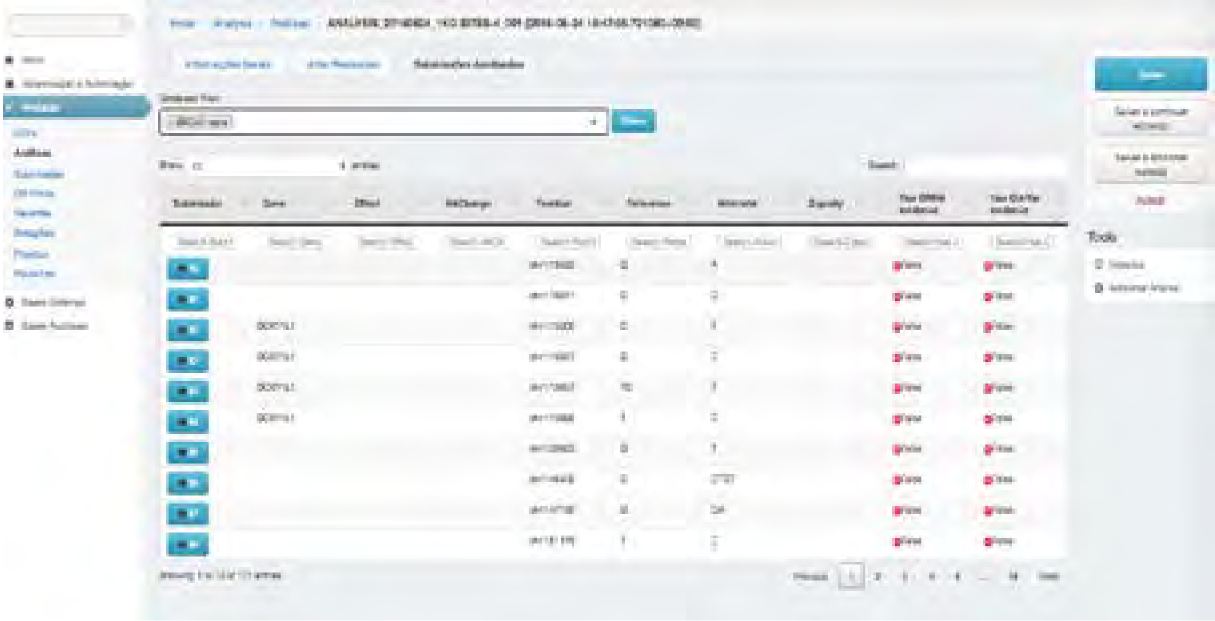

Figure 5. Filtering and Analysis System

Results

The platform presented in this work provides unified data access service (on a consolidated human genome database) and tools for data analysis and biomedical knowledge extraction in order to simplify the work of scientists in the diagnosis of rare diseases, tumors and cancers.

Before the use of the platform, data analysis was a long and expensive process performed with spreadsheets. Geneticists used to use filters to select and combine variants for reports. Nowadays, the length of time required for the analysis process was reduced as well as experts effort.

Using the platform doctors and analysts (geneticists, biologists and biomedical) have the opportunity to build their bioinformatics experiments and to perform analysis with various filters and functionalities in an intuitive and practical way.

The main impacts of the platform are highlighted below.

- Updated databases: The platform updates the unified database whenever a new database appears or it is updated. By maintaining an updated database, it is possible to have results delivered to the patient with greater details. It also allows re-examining previous clinical cases which were not solved during the first analysis. This automated reanalysis can result in reclassification of the variant, and in a new notification for the experts and patients.
- Productivity: the computing power of the platform have a direct impact in the amount of processed and visualized data. Analysts, who are the main stakeholders, achieve higher productivity in the analysis, save time to study new exams or to optimize existing procedures. Moreover, remote access to the data analysis adds flexibility to the process. Altogether, it reduces cost of annotation process and increases the number of diagnoses and customers served (e.g., laboratory patients, research institutes and hospitals).
- Decision support: Since the platform can handle annotation with key scientific references of main global genomic medicine databases (e.g., Pubmed), medical doctors may have more detailed and updated information to better interpret reports. This directly impacts on the work of medical doctors, geneticists, biomedics and the population

Discussion

The main challenges of dealing with big data on health and genomic applications are related to: privacy and data security, data streaming and storage, lack of training in data science, discoverability of and access to biomedical data[9]. Having an unified access platform can enable scientists to overcome the problems discussed below:

- Overlapped and decentralized data sources: Online catalogs of clinical data are available on the Web, but besides not being unified they are extremely overlapped. OMIN and Clinvar are examples of online repositories with overlapped data. Consumers of such data (e.g., genetics clinics) have difficulties in creating new consistent applications on top of them. It happens because they need an extra effort to analyze, understand and deal with overlapped information extracted from these various and distributed sources.
- Lack of standards: in addition to the overlapped and decentralized data sources, there is a lack of standards for data publishing. Each publisher chooses which datasets to provide and how to publish them. Often, there is also no agreement between publishers, laboratories and clinics. As a result, existing services that consume clinical data and the data formats and types themselves may vary significantly.
- Usability: although it is necessary to build internationalized systems with friendly user interfaces for genome annotation and data management, most existing tools do not have support for other languages (Portuguese, for instance). For this reason, users have still to work with text files and complex spreadsheets.
- Cost: licenses of existing tools are expensive. It increases the total cost of the systems which use them. Further, the lack of appropriate systems with friendly user interfaces also increase the labor cost, once it makes biologists and geneticists spend many labour time in their activities.

Conclusion

By providing unified access to a set of genetics databases, the platform hereby presented contributes to genome data analysis and biomedical knowledge extraction. It simplifies the work of scientists and supports medical/doctor's decisions in the diagnosis of tumors, cancers and rare diseases. The proposed platform also has direct impact on the cost of the annotation process, which otherwise is a long and expensive process, manually performed with excel spreadsheets. It is worth noting that several other areas and data types require unified access to heterogeneous data sources and therefore can benefit from this kind of platform implementation. Future work includes the delivery of other versions of the platform with the implementation of new features.

Acknowledgments

We would like to acknowledge the support of FACEPE, Genomika, UFPE and UFRPE.

References

[1] Sagiroglu, S.; Sinanc, D. Big data: a review. International Conference on Collaboration Technologies and Systems (CTS), pp. 42-47, 2013

[2] Anguita, A., et al. (2010) A review of methods and tools for database integration in biomedicine. Curr. Bioinform., 5, 253–269

[3] Peterson, Thomas A., Emily Doughty, and Maricel G. Kann. "Towards precision medicine: advances in computational approaches for the analysis of human variants." Journal of molecular biology 425.21 (2013): 4047-4063.

[4] Hey AJ, Trefethen AE. The data deluge: an e science perspective. In: Grid computing: making the global infrastructure a reality. 2003. P.809-24

[5]Genbank Statistics. [cited 2016 May 17]. Available from: http://www.ncbi.nlm.nih.gov/genbank/statisticshttp://www.ncbi.nlm.nih.gov/genbank/statistics

[6] Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med. 2011;17(3):297-303.

[7] Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of health's big data to knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21(6):957-8.

[8] Marz, N. Big data : principles and best practices of scalable realtime data systems. [S.l.]: O'Reilly Media, 2013.

[9] Silva, F. A. B. Big Data e Nuvens Computacionais: Aplicações em Saúde Pública e Genômica. Journal of health Informatics. v. 8, n. 2 (2016).

Contato

Profa. Andrêza Leite de Alencar
DEINFO/UFRPE
Rua Dom Manoel de Medeiros, s/n. Campus
Dois Irmãos CEP: 52171-900 - Recife/PE
- Brazil
Phone: +55-81-33206491
e-mail: ala4@cin.ufpe.br