Computer Science and Information Technology 5(3): 105-112, 2017
DOI: 10.13189/csit.2017.050303 http://www.hrpub.org

HPC Services to Characterize Genetic Mutations Through Cloud Computing Federations

Manuel-Alfonso López-Rourich*, **Felipe Lemus-Prieto**, **Javier Corral-García**, **José-Luis González-Sánchez**

Research, Technological Innovation and Supercomputing Center of Extremadura (CénitS), Cáceres, Spain

Abstract One child in every 200 births may be affected by one of the approximately 6,000 monogenic diseases discovered so far. Establishing the pathogenicity of the mutations detected with NGS (Next-Generation Sequencing) techniques, within the sequence of the associated genes, will allow Precision Medicine concept to be developed. However, sometimes the clinical significance of the mutations detected in a genome may be uncertain (VUS, Variant of Uncertain Significance) which prevents the development of health measures devoted to personalize individuals' treatments. A VUS pathogenicity can be inferred thanks to evidences obtained from specific types of NGS studies. Therefore the union of supercomputing through HPC (High-Performance Computing) tools and the cloud computing paradigm (HPCC), within a Data Center federation environment offers a solution to develop and provide services to infer the pathogenicity of a set of VUS detected in a genome, while guaranteeing both the security of the information generated during the whole workflow and its availability.

Keywords VUS, *-omics*, Precision Medicine, HPC, Cloud Computing, Information Security

1 Introduction

The progress of healthcare research has always been one of the main challenges facing humanity. One of the most important discoveries has been the realization that the study of the DNA (Deoxyribonucleic Acid) sequence allows the functioning of fundamental biological processes within organisms to be understood.

DNA sequencing is a set of biochemical techniques whose purpose is to determine both the nucleotides and their order within a DNA molecule. Until 2007 [1], DNA sequencing involved a huge temporal and economic cost. In the context of these problems NGS technology (the second-generation sequencing methods [2]) emerged, and with it a new genetic paradigm which allows whole genome sequencing, or sections, on a broad scale with considerable reductions in the time taken and the cost of processing. However, resequencing the whole exome may be preferred because it allows specialists focusing on the genome section from which proteins which regulate individuals' biological processes are obtained in the transcription process.

However, the move towards the use of multigene panels or exome sequencing in clinical care has caused an increase in variants of uncertain significance (VUS) in genetic counseling within the sequence of specific genes. A key challenge is that the clinical significance of a VUS result for disease risk is in principle unknown.

In this sense, although significant efforts have been made to characterize the clinical significance of a VUS, there is a lack in the adoption of new paradigms which focus in the integration of *-omics*[1] data sets all of which would be useful to improve the pathogenicity prediction of a VUS and to identify the function of the genes involved in the development of hereditary diseases.

Cloud computing allows supercomputing centers not only to deploy HPC services to run scientific applications, such as those which obtain an evidence to infer the clinical significance of a VUS in the least possible time, but also develop elastic and scalable HPC applications as well as foster energy efficiency policies among Data Centers. Therefore the contribution of this paper is the development of solutions which promote the deployment of HPCC services to make the most of HPC federations.

In relation to this work, the COMPUTAEX Foundation[2] has undertaken several projects with the Hospitals Infanta Cristina (Badajoz) and San Pedro de Alcántara (Cáceres), focused on the study of the exome sequence of

[1] -omics aims at identifying the functions of as many genes as possible of a specie by combining the sequencing of different biological variables such as DNA (genomics), RNA (transcriptomics), proteins (proteomics) or even scientific literature (literome)

[2] http://www.computaex.es/

certain patients, processed in the CénitS (Supercomputing, Technological Innovation and Research Center) Data Center by the LUSITANIA supercomputer.

The contribution of our work is to demonstrate how the union of supercomputing and the cloud computing paradigm, within a Data Center federation environment, can generally be useful to improve the provision of services which are able to meet users' demands. Specifically, we will describe a prototype of a system which aims at helping specialists to infer a VUS pathogenicity thanks to evidences obtained from specific types of NGS studies. We will show how the performance of this prototype can be improved thanks to the utilization of a real Data Centers federation of HPC resources. Besides, we will manifest how this federation can be enriched to be able to meet users' general demands when using supercomputing resources (performance, storage and information security and high availability).

The rest of the paper continues as follows: After introducing our work, Section 2 explains the motivation behind it. Section 3 develops the concept of HPC Federation Data Centers and Section 4 presents the FI4VDI (Federation Infrastructure for Virtual Desktop Infrastructure) project and Section 5 shows a prototype for VUS interpretation. The paper finishes with our conclusions and a description of future areas of research.

2 VUS interpretation

VUS results can subsequently be reclassified as more information becomes available by means of procedures and methods which vary among laboratories. Nevertheless, the key to improving the interpretation of VUS results and therefore to lowering the overall VUS rate is the accumulation and organization of high-quality genetic data and the development of robust methods for data interpretation.

In that sense, given the availability of different results obtained by processing biological variables through distinct genetic sequencing studies (such as re-sequencing or RNA-Seq) a set of research lines can be used to *try* to assess the clinical significance of a variant. For this purpose [3] gathers a set of practice guidelines and recommendations and the following lines of research:

- Looking for a variant in locus specific databases (LSDB), above all if the database allows for the repeated submission of variants.
- Presence or absence of a variant in Single Polymorphisms (SNP) Databases, especially its allele frequency information.
- Co-occurrence (in trans) with known deleterious variants.
- Co-segregation with the disease in the family.
- Occurrence of a new variant concurrent with the incidence of the disease.
- Species conservation.
- *In-silico* prediction of pathogenic effect.
- *In-silico* splice site prediction.
- RNA studies.
- Loss of Heterozygosity (LOH).

From the computational point of view, obtaining this data through different work-flows requires the execution of specific sets of bioinformatics tools within an infrastructure which meets the demands of data security, storage and high-availability. The problem is that sometimes geneticists themselves must assemble the software which is executed in the work-flows steps with the serious risk that the quality of the final data does not satisfy the relevant clinical demands. In the next section we will explore the advantages of deploying cloud computing services over Data Center federations so that users can access ready-to-use workflows which help the clinical significance of a VUS to be inferred.

3 Data Center Federation

A federation of Data Centers is a set of interconnected resources from different Data Centers which may be geographically distributed and combines a centralized government to improve the provision of services.

On the one hand, the provision of cloud computing services through a federation of interconnected and independent sets of cloud resources is an ideal solution because from the point of view of ubiquity, a cloud federation guarantees that cloud providers can both meet the demands of users within a global scale and deploy services which requires a global capability such as gaming, multimedia content delivery distribution (CDN, Content Delivery Network) or eCommerce.

On the other hand, the centralized government of cloud computing resources can benefit the provision of cloud services which are intended not only to execute HPC applications in the least possible time but also to guarantee that users' need such as storage requirements or information protection are met by using the whole set of resources from a Federation.

In this sense, the staff of the supercomputing center CénitS work on the development of solutions with the objective of taking advantage of federations of cloud computing resources to improve the provision of cloud computing services such as the prototype described in Section 5.

Therefore the purpose of the next sections will be to describe those solutions and the general state of art of research related to them.

3.1 Performance

The federated cloud allows desktops to be connected to custom virtual machines which offer optimal performances with high levels of energy efficiency, short deployment times (the entire system can be fully functional in a matter of a few minutes) and ensuring an improvement in its associated productivity.

In this way, thanks to desktop virtualization, almost any system can be encapsulated and delivered to a remote device with minimal hardware features. This has clear advantages over the traditional models, providing much greater flexibility, allowing energy costs to be reduced (when resources are shared and allocated to users according to their needs) and enhancing the integrity of the information.

In addition, it is important to highlight that all these features are transparent to the end-user, who will be able to use and manage all the resources using a single frontend, optimizing the efficiency of the processes involved. Another important point resides in the fact that the federation can respond to the growing needs of each environment, on demand, depending on the workloads at a given moment.

Based on HPC and parallel algorithms, some specific tools achieve higher performance compared to traditional DNA and RNA sequencing. They obtain remarkable results when combining the use of virtual machines with multi-tier applications in which presentation, application processing, and data management functions are physically separated. As an example, The OpenNebula [4] OneFlow API is a RESTful (Representational State Transfer) service for the creation, control and monitoring of services composed of interconnected virtual machines with deployment dependencies between them, so that each group of Virtual Machines is deployed and managed as a single entity. In this way, users and administrators can define, execute and manage multi-tiered applications, with the advantages of using auto-scaling policies based on performance metrics and schedule.

3.2 Data storage

From the moment when storage devices emerged, the necessity to store digital data arose. In that sense, many studies have tried to estimate the world's technological capacity to store information in analog and digital devices such as Hilbert et Al. in [5] who argues that in 2007, humankind was able to store 2.9 x 10^{20} optimally compressed bytes, with significant growth rates since the beginning of the study.

However, recent years have witnessed such a huge increase in the amount of data storage needs that Data Centers must work on the development of solutions which meet the demands for high scalability and data access efficiency. Specifically, big data statistics and facts for 2017[3] mark that the digital universe stores 2.7 Zetabytes of data or that Facebook stores, accesses and analyzes more than 30 Petabytes of user generated data.

From the very first file systems which were intended to isolate information chunks into logical files in order to allow users to control how data is stored and retrieved, distributed file systems have made the location of data transparent to users in order to incorporate remote servers and increase storage capacity.

In this sense, DFSs represent the ideal solution for meeting the major challenges in data storage such as scalability or avoiding data transfer bottlenecks mainly because of its capacity to manage a set of storage resources and implement storage policies intended to meet users' demands. An example is HDFS (Hadoop Distributed File System). Basically, HDFS is a distributed, scalable and portable file system with a set of Data Nodes which store redundant chunks of logical files and a Name Node which mainly maps the FS metadata and the pieces of information stored within the Data Nodes. The power of HDFS fault-tolerance policies relies on the use of general purpose servers to deploy Data Node services and in the redundant copies of the file chunks.

Hadoop HDFS is one example among a big number of available options to implement a DFS; one can choose solutions developed for large-scale Linux systems, such as Lustre, or implementations of RAID (Redundant Array of Inexpensive Disks) among independent disks. Another option is the application protocol NFS (Network File System) which allows systems connected to a local network to access files as if it was a local server, regardless of the OS and the network transfer protocol.

With regards to Data Center federations, the concept of Federated Storage, a collection of storage resources (from independent Data Centers which are the nodes of the Federation) governed by a common smart system which rules how data is stored, managed, and migrated throughout the storage network could be developed. The combination of a central management of storage resources and the paradigm of cloud computing will allow organizations to move applications and data globally or create networks which tend to have limitless capacity.

3.3 Information security

Information is one of the most valuable assets today, so it is necessary to protect it appropriately. Information security which regard to ISO/IEC 27000:2016 [6], is understood as the preservation of the following features:

- **Confidentiality:** property that information is not made available or disclosed to unauthorized individuals, entities, or processes.
- **Integrity:** Property of accuracy and completeness.
- **Availability:** property of being accessible and usable upon demand by an authorized entity.

The above model, commonly referred to as CIA (Confidentiality, Integrity and Availability) triad, is designed to guide policies for information security within an organization.

In this sense, a federation of Data Centers can provide many advantages from the point of view of security, especially in the case of both data integrity and availability.

On the one hand, the possibility of storing the same information in different places will contribute to guaranteeing the integrity of such information. On the other hand, the existence of redundant services and data,

[3] http://www.waterfordtechnologies.com/big-data-interesting-facts/

distributed between the different Data Centers within the federation, results in a more fault tolerant system. This system is able to continue providing such services and data even in the case of a disaster (e.g, a Data Center drop), improving the availability of the complete system.

However, Data Center federations have some security issues too, especially those related to confidentiality. In this scenario it is difficult to define the roles and responsibilities of each member of the federation over the information stored in the system. To deal with these issues it is necessary to answer the following questions: Who owns the information? Where is the information stored? Who deals with the information?

The previous security issues are even more important when sensitive information (e.g., personal data) is involved. In this situation a set of security measures must be deployed in order to ensure proper legal compliance.

4 FI4VDI project

The main objective of the European FI4VDI[4] project is to develop a Cloud Computing service provision to generate combined infrastructure destined for enterprises of all sizes (small, medium and large), universities and educational or research centers.

As can be seen in Figure 1, the Federation Infrastructure consists of several OpenNebula zones [4] interconnected with a master one. Each slave zone represents a virtualization pool (managed by a specific hypervisor) from a technological center which participates in the FI4VDI project.

In the last phases of the FI4VDI project a pool of prototypes (OpenNebula Virtual Images) related to specific user job have been proposed in order to deploy on-demand services from the OpenNebula IaaS (Infrastructure as a Service) environment.

4.1 Federation members and committed resources

The FI4VDI federation is composed by four research centers from Spain and France. Each center contributes to the federation with a quantity of resources according to its capabilities. Table 1 summarizes initial committed resources by each center. So, at the beginning, the FI4VDI federation had 11 nodes, 28 processors, 232 cores and 1106 GB of main memory. In addition, it had a storage capacity of 33 TB. But this hardware assignation is not static, one of the main advantage of this type of resources federation is that allows to add or release hardware resources in a simple and flexible way, either by a greater contribution of the current members or by the incorporation of new centers.

Following, a brief description of each member of the federation is shown.

[4]http://fi4vdi-sudoe.org/

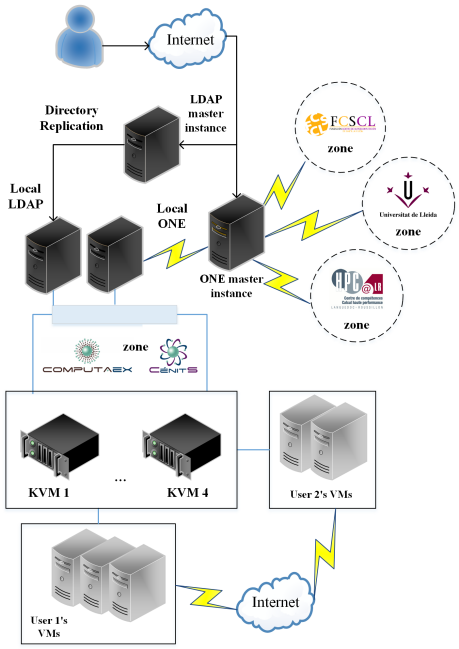

Figure 1. FI4VDI Federation scheme.

4.1.1 CénitS-COMPUTAEX (Spain)

CénitS (Extremadura Supercomputing, Technological Innovation and Research Center) aimes to promote and provide, HPC services and advanced communications to the research communities of Extremadura, or that company or institution that requests it and thus contribute through technological improvement and innovation, improving the competitiveness.

CénitS hosts and manages the supercomputers LUSITANIA and LUSITANIA II. The center is managed by COMPUTAEX (Computing and Advanced Technologies Foundation of Extremadura), a non-profit public foundation dependent on the Regional Government of the Autonomous Community of Extremadura.

4.1.2 FCSCL (Spain)

The Foundation of the Supercomputing Center of Castilla y León (FCSCL) is a public entity created by the Junta de Castilla y León and the University of León, which aims to improve the research tasks of the university, research centers and the companies of Castilla y León.

Table 1. Resources commited by FI4VDI members

Zone	No. nodes	Features per node			
		CPU			RAM (GB)
		Model	No.	Cores	
CénitS	4	AMD Opteron 6376	2	16	128
FCSCL	3	Intel Xeon x7350	4	4	128
HPC@LR	3	Intel Xeon e5-2650v2	2	8	64
UdL	1	Intel Xeon E5420	2	4	18

4.1.3 HPC@LR (France)

Located in the Languedoc-Roussillon region, the Competence Center for High Performance Computing (HPC@LR) aims to provide a Resource Center devoted to High Performance Computing for entrepreneurs, private and public sector scientists and engineers, as well as teachers.

4.1.4 UdL (Spain)

The main aim of the University of Lleida (UdL) is the education of its students. They provide high standards of teaching with quality services throughout the university community that reach society beyond the campus gates. Students make up most of the UdL community, and the objective is to ensure that they enjoy the teaching and learning processes involved.

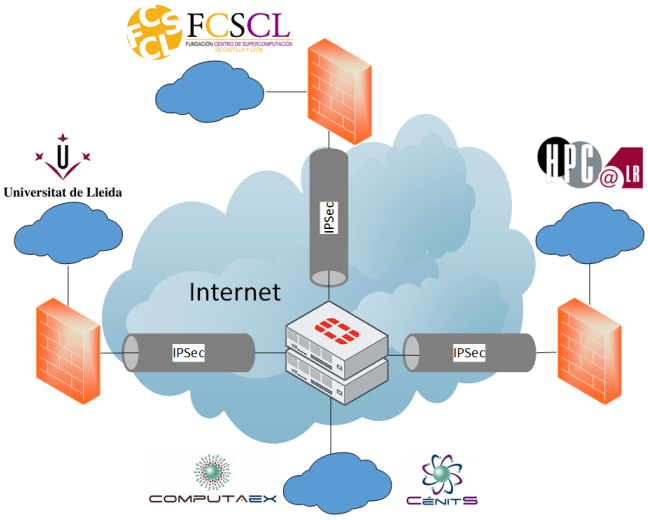

Figure 2. FI4VDI IPSec VPN.

4.2 FI4VDI information security

In order to deploy new services, especially if those services work with sensitive information, it is necessary to undertake appropriate risk evaluation of information systems involved in such services and establish the necessary measures to reduce the associated risk as much as possible. Moreover, it is also necessary to guarantee legal compliance when these services process personal data.

As mentioned above, a HPC resource federation provides security advantages from the point of view of availability and integrity but it has security flaws in the case of confidentiality. Such security flaws are especially worrying when personal data is involved (e.g. patient clinical information). In this context, anonymization is usually the security chosen measure but this technique has been demonstrated to be far from perfect, for example, the combination of publicly available genetic resources and the metadata about DNA donors can reveal the identity of individuals [7], so additional security measures should be deployed.

4.2.1 Main security measures deployed

VPN deployment Most of the security measures that are usually implemented are associated with *data at rest*, but the proposed infrastructure is a federated network composed of four HPC centers which, for example, share the load associated to process data from NGS. So, it was necessary to deploy security measures to protect *data in transit* too. In this sense, an IPSec Virtual Private Network was deployed between the HPC centers for enabling private communication over the Internet.

For a multi site VPN scenario a hub and spoke topology is the most common implementation. A central hub will enable not only connectivity from remote sites to the hub and the hub to the remote sites, but acts as a gateway for remote sites to communicate with each other via the hub. In FI4VDI VPN topology, the Fortigate 1000C cluster of CénitS acts as the hub and the other centers are remote sites as shown in Figure 2.

Chosen topology enables the addition of future HPC partners to the federation easily due to the fact that only a new single VPN needs to be deployed between the VPN concentrator (CénitS) and the network of that new partner.

Authentication and authorization The FI4VDI project federation currently has five OpenNebula zones, one for each supercomputing center (León, Extremadura, Lleida and Montpellier) and one zone defined as master, through which both authentication and authorization data from each zone (users, groups, etc.) are synchronized.

In order to manage users inside the federation, an authentication and authorization system based on OpenLDAP (Open Source Lightweight Directory Access Protocol) has been deployed. This system is composed of a provider LDAP server and local synchronized copies of it in each of the zones (consumers). To achieve this, the LDAP Sync Replication engine, syncrepl for short, has been used. The syncrepl engine resides at the consumer and it creates and maintains a consumer replica by connecting to the replication provider to perform the initial Directory Information Tree (DIT) content load, followed either by periodic content polling or by timely updates upon content changes.

In this way, when a user tries to access an OpenNebula zone, OpenNebula requests user credentials and privileges from the master LDAP server. If the master is down, credentials would be requested to the local instance instead, thus obtaining a more fault-tolerant authentication system.

Virtual machine backup system Another security advantage of the establishment of a federation of HPC centers is the ability to deploy a virtual machine backup system between the different centers.

Users can restore their virtual machines in case a problem arises (damaged disk, accidental erasure, wrong configuration, etc.), but the copy of the virtual machine can be in the same center or in another center within the federation, with the same virtualization technology (same hypervisor). Thus, obtaining a more robust backup system.

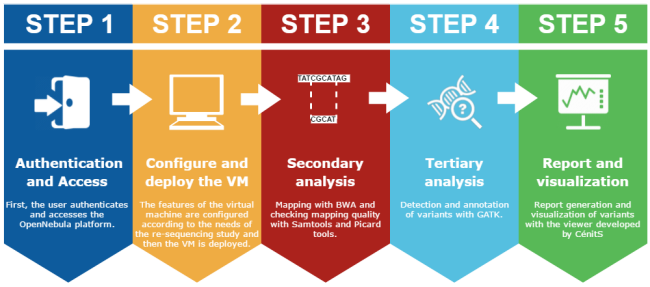

Figure 3. Re-sequencing prototype workflow.

5 Re-sequencing prototype

One of the prototypes is an OpenNebula Virtual Image which contains pre-loaded data and a pipeline (Figure 3) consisting of a set of bioinformatics tools which are connected to obtain high-level genetic information (for instance some of the evidences presented in the Section 2) by processing reads obtained from re-sequencing studies.

5.1 Re-sequencing studies

A re-sequencing experiment consists in sequencing the DNA of an organism to compare its genetic sequence with a consensus reference so that variations between the sequences can be found in order to study hereditary diseases. Its phases are described below:

1. **Primary Phase**: The main objective is to create a collection of clones of DNA sections of interest from which millions of sequences of a certain length, called *reads*, will be generated.

2. **Secondary Phase**: In this phase the reads are mapped (the verb "to align" is also used indistinctly) against a consensus reference sequence. The computational requirements are very high because NGS platforms generate a large volume of reads so both the infrastructure where this phase will be executed and the configuration of the software is critical.

3. **Tertiary Phase**: The purpose of the last phase is to obtain genetic variations such as SNPs (Single-nucleotide polymorphism), a variation that is present to some appreciable degree within a population (e.g. more than 1% [8]) of a single nucleotide that occurs at a specific position in the genome sequence if compared against a consensus reference one and small indels, that is, an insertion or a deletion of a small genetic sequence within an aligned read). This phase ends with the variants annotation[5]. The results from this phase will be stored in a database of annotated variants.

[5]Annotating a genetic variant is a process in which text, from an ontology, or numerical data, from an interval, is associated with it in order to obtain its functionality)

5.2 Prototype

The prototype [9] pipeline consists of a set of bioinformatics tools which perform the three phases of a re-sequencing study. Therefore the whole processing workflow starts with a collection of *reads* and finishes returning a list of variants which will be stored within a relational database.

One of the most interesting tools of the prototype is the one which supports the diagnosis of hereditary diseases which has to be performed by geneticists in a genetic counseling consultation (that is interpreting the clinical significance of the variant in order to infer its relation with the development of the disease under investigation). In fact, the prototype does not automatically interpret the clinical significance of a VUS but provides tools so that specialists can manually interpret the evidences that are automatically gathered.

What is important to understand is the fact that depending on the sequencing study and the biological variable which the geneticist wants to analyze, the sequencing platform will generate data from which users will be able to obtain other evidence which is not available in the re-sequencing prototype. However, the whole workflow would change and other kinds of tools and pre-loaded data will be needed.

For this reason, given that OpenNebula allows system administrators to create Virtual Images with a customized S.O containing programs required by those VM users such as this prototype programs (which allows the deployment of VMs which are fully prepared to obtain different clinical evidences), the platform developed is ideal.

5.3 Prototype workflow

The workflow has been developed in relation to the phases of a re-sequencing study which were explained in the subsection 5.1). The steps of the workflow which can be shown in the Figure 3 are the following ones:

5.3.1 Authentication and Access

The first step consist of accessing to the OpenNebula platform. Once the user has been authenticated, he or she will be able to navigate through the different zones of the federation and check both the already VMs which may have previously been deployed by the user and virtual resources from the federation such as the available virtual images or the cloud resources.

5.3.2 Configure and deploy the VM

Depending on a user' needs (specifically the amount of *reads* to be processed which depends on the percentage of the genome that was sequenced, among other issues), he or she will be able to customize the resources of the new VM. Given users' requests and the available resources among the federation Data Centers, the platform will launch the VM in the most suitable federation zone. After that, the VM is deployed and the user will have a workflow to process their genetic sequences through the prototype pipeline itself.

5.3.3 Secondary analysis

The next step of the general workflow consists of aligning sequences themselves against an indexed consensus reference sequence of the human genome provided by the UCSC (University of California San Diego)[6]. In this case, the tools which the prototype uses to perform the phase are firstly BWA-SW [10]. However, the generated SAM (Sequence Aligment/Map) file must be processed. Specifically, we included Samtools[7] and PicardTools[8] to both sort and index the alignment results, to mark the duplicated alignments or to recalibrate the alignment quality scores among other issues. The final result of this phase is a BAM (Binary Aligment/Map) whose content is ready to call variants in the next phase.

5.3.4 Tertiary analysis

The purpose of this phase is to detect SNP and small indels and to annotate them. Concerning variant calling, we used GATK [11] (Genome Analysis Toolkit), a very popular suite which includes lots of useful tools. In addition, its users community is enormously useful to refine the analysis performed with the tools.

One important issue about variants calling is the fact that we have to avoid false positives and the annotation must describe their functionality in case it is possible and the variant is not a VUS. Concerning the first issue, we use a GATK tool called VQSR (Variant Quality Recalibration Score) which calibrates the accuracy of the variants with the help of databases of variants which were detected in the IBS (Iberian population in Spain) from public projects such as 1000Genomes. On the other hand, annotations are assigned by using both VEP (Variant Effect Predictor) [12] and the Ensembl v75 Perl API [13].

5.3.5 Report and visualization

The prototype incorporates a viewer developed by the staff of the CénitS center. This tool presents a main interface where users can choose the sample of reads which has been processed in order to get the whole list of detected variants. In addition, since the size of the variants list may be too big, the tool incorporates filters by the whole set of fields of each variants such as the type (SNVs, Single Nucleotide Variation or *indels*) or the chromosome whose sequence has been altered.

Concerning the interpretation of the clinical significance of the detected variants, the viewer incorporates functionality to check some of the evidences (whether it is a VUS or not) which has been shown in previous sections: among others the conservation of the affected protein or *in-silico* predictions of the variants pathogenic effect.

Since it may not be possible to infer the clinical significance of a variant, the contribution of our viewer to interpret the significance of a VUS is to present some evidences so that a specialist can interpret it. However, he or she will have to access to other kind of evidences by means of the performing of clinical testing or the processing of other kinds of -omics data.

6 Conclusions and Future work

The approach we consider in the present article allows potential users (doctors, geneticists, researchers, etc.) to characterize genetic mutations without strong background knowledge in bioinformatics. The use of a virtualization platform such as OpenNebula provides two significant advantages: the ability to deploy different virtual machines depending on the particular needs of each user (DNA, RNA or proteins studies) and a greater energy efficiency (machines can be turned on or off depending on the demand). In addition, the establishment of a Data Center federation enhances system features such as performance, scalability, flexibility or security.

Information security is always important, but especially when working with sensitive data (eg genetic information). Our approach includes the necessary security measures to ensure and adequately protect these data both at rest and in transit.

The future work that better encapsulates this work is our proposal of a free service to infer the diagnosis of Primary Immunodeficiency Diseases (PID). This is a very promising line of work whose main challenge is that the information needed to characterize PIDs is not stored centrally. To solve this problem we propose the development of techniques which foster the generation of new value for standardized digital -omics data through their dissemination and exchange between stakeholders in healthcare. To enhance the service, a federated system of European HPC resources is proposed, so that computing, storage, security and availability demands from NGS services can be satisfied.

Other lines of future work are those which are intended to develop work-flows to obtain other lines of evidence to infer the clinical significance of a VUS to enrich the number of Virtual Images which are available.

[6] http://hgdownload.cse.ucsc.edu/downloads.html
[7] http://samtools.sourceforge.net/
[8] https://broadinstitute.github.io/picard/

Acknowledgements

This research has been partially funded by the project SOE4/P3/E804, FI4VDI-SUDOE, "Federation Infrastructure for Virtual Desktop Infrastructure" and by the European Union through the European Regional Development Fund (ERDF) Programme: Extremadura Operational Programme 2014-2020, Thematic objective 1: Research and Innovation, and Thematic Objective 2: Information and communication technologies.

REFERENCES

[1] K. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata.

[2] A. Grada, K. Weinbrecht. Next-Generation Sequencing: Methodology and Application, Journal of Investigative Dermatology: Vol. 133, e11, 2013.

[3] Y. Wallis, S. Payne et al. Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics. Clinical Molecular Genetics Society and Association of Clinical Cytogenetics and the Dutch Society of Clinical Genetic Laboratory Specialists, 2013.

[4] R. Moreno-Vozmediano, R. Montero. I.M. Llorente. IaaS Cloud Architecture: From Virtualized Datacenters to Federated Cloud Infrastructures, IEEE Computer, Vol.45, 65-72, 2012.

[5] M. Hilbert and P. Lopez. The World's Technological Capacity to Store, Communicate, and Compute Information, Science, vol.332, 60-65, 2011.

[6] ISO/IEC. ISO/IEC 27000:2016 Information technology. Security techniques. Information security management systems. Overview and vocabulary. Geneva, Switzerland: ISO/IEC, 2016.

[7] John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, Vol. 339, No. 6117 (18 January 2013), p. 262.

[8] Single-nucleotide polymorphism/SNP — Learn Science at Scitable. Online available from www.nature.com.

[9] M.A. Rourich, J.L. González-Sánchez et Al. Automatic Genetic Sequences Processing Pipeline in the Cloud for the Study of Hereditary Diseases, 8th Iberian Grid Infrastructure Conference, 129-142, 2014.

[10] Li and Durbin. Fast and accurate long-read alignment with Burrows-Wheeler Transform, Bioinformatics: Epub. [PMID: 20080505], 2010.

[11] G. A. Auwera, M. O. Carneiro, C. Hartl et Al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline, Current Protocols in Bioinformatics, 43:11.10.1-11.10.33, 2013.

[12] McLaren, Pritchard, Rios et Al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor, Bioinformatics 26(16): 2069-70, 2010.

[13] Flicek, Amode, Barrell et Al. Ensembl 2014, Nucleic Acids Research: 42, D749–D755, 2014.