Big Data Research 26 (2021) 100253

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Enhancing Precision Medicine: A Big Data-Driven Approach for the Management of Genomic Data

Ana León *, Óscar Pastor

Research Center on Software Production Methods (PROS), Universitat Politècnica de València, Valencia, Spain

ARTICLE INFO

Article history:
Received 16 December 2019
Received in revised form 1 April 2021
Accepted 27 June 2021
Available online 8 August 2021

Keywords:
Big Data
Genomics
Computer science
Theory and methods

ABSTRACT

The management of the exponential growth of data that Next Generation Sequencing techniques produce has become a challenge for researchers that are forced to delve into an ocean of complex data in order to extract new insights to unravel the secrets of human diseases. Initially, this can be faced as a Big Data-related problem, but the genomic data have particular and relevant challenges that make them different from other Big Data working domains. Genomic data are much more heterogeneous; they are spread in hundreds of repositories, represented in multiple formats, and have different levels of quality. In addition, getting meaningful conclusions from genomic data requires considering all of the relevant surrounding knowledge that is under continuous evolution. In this scenario, the precise identification of what makes Genome Data Management so different is essential in order to provide effective Big Data-based solutions. Genomic projects require dealing with the technological problems associated with data management, nomenclature standards, and quality issues that only robust Information Systems that use Big Data techniques can provide. The main contribution of this paper is to present a Big Data-driven approach for managing genomic data, that is adapted to the particularities of the domain and to show its applicability to improve genetic diagnoses, which is the core of the development of accurate Precision Medicine.

1. Introduction

One of the pillars for understanding the genetic aspects that make our predisposition to disease and our response to treatments different from each other is genetic diagnosis. Next-Generation Sequencing (NGS) technologies, such as Whole-Genome Sequencing (WGS) and Whole-Exome Sequencing (WES), have provided researchers with an exponentially growing list of DNA variants that are potentially relevant in calculating the subsequent likelihood of developing specific diseases. These advances and the growing availability of health data have allowed the development of novel approaches such as Precision Medicine, in which health care is individually tailored based on the genetic characteristics, lifestyle, and environment of each patient [1].

While other fields such as Astronomy have faced the challenges of Big Data for decades, it was not until the 1000 Genomes Project was launched in 2008 that Genomics entered the domain, with the total amount of sequence data being produced doubling approximately every seven months [2]. Fig. 1 depicts a chart with the growth of genome sequencing and the growing expectations in the following years. The left axis represents the total number of human genomes sequenced. The right axis represents the worldwide annual sequencing capacity (Tbp: Tera-basepairs, Pbp: Peta-basepairs, Ebp: Exa-basepairs, Zbps: Zetta-basepairs). As can be seen, sequencing capacities are expected to continue growing at a faster pace than the ability of experts to review and analyze the data produced.

What, at the outset, may have seemed a significant step forward for the development of novel approaches such as Precision Medicine has caught researchers and clinicians unaware and forced them to delve into an ocean of complex data in order to extract new insights to unravel the secrets of human disease.

Many efforts in the Big Data community have been oriented to providing efficient solutions to open problems such as NGS read alignment [3–5] and variant calling for detecting rare genetic variants in the DNA sequence with higher confidence [6–8]. Nevertheless, the next step (their classification and interpretation for clinical purposes) remains unsolved and is becoming more and more complex.

Once the genome is sequenced and the variants regarding the sequence of reference are determined, the introduction of this

* Corresponding author.
E-mail addresses: aleon@pros.upv.es (A. León), opastor@pros.upv.es (Ó. Pastor).

https://doi.org/10.1016/j.bdr.2021.100253

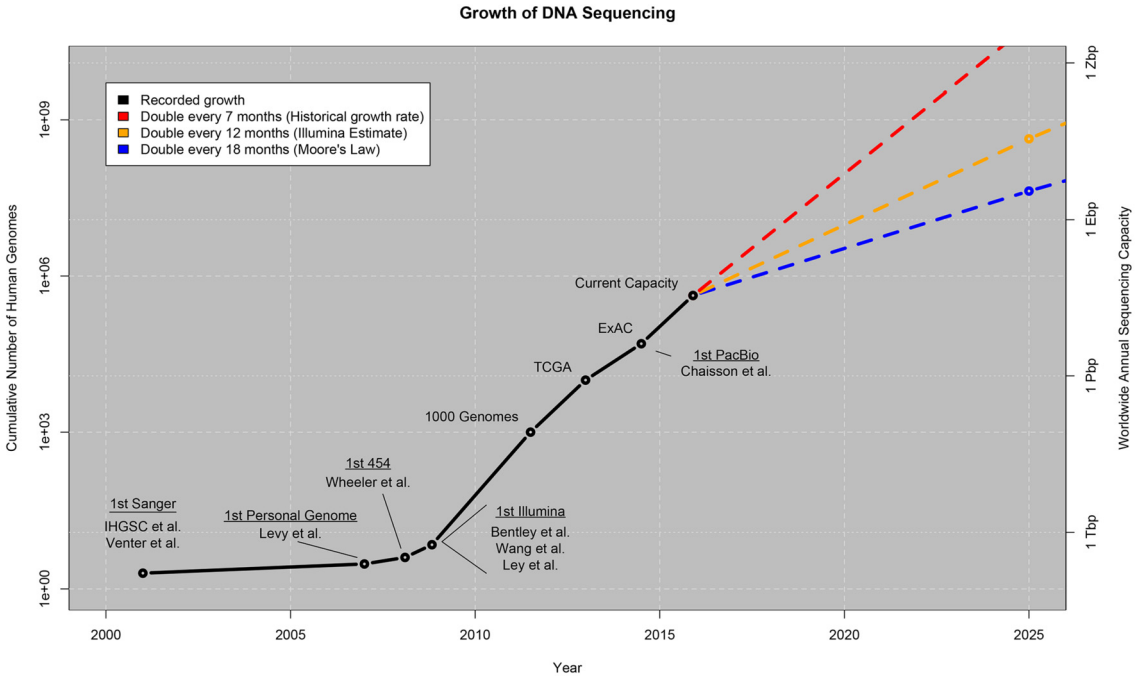

Fig. 1. Growth of DNA sequencing - Image obtained from [2].

knowledge into clinical practice requires answering two important questions: i) Can the impact of these variants lead to a disease phenotype? ii) Does the impact of these variants cause the symptoms observed in the patient?

To answer these questions, all of the additional knowledge about each variant must be considered regarding published studies, population frequencies, gene impact predictions, and segregation data. Surprisingly, the collection and analysis of these data are tasks that are mainly performed manually, which constitutes a bottleneck in the genetic diagnosis workflow and a potential source of errors.

To solve this problem, geneticists, clinicians, and biologists need to change the way they manage the data flow, facing well-known challenges that are inherent to the Big Data life cycle such as collection, integration, cleansing, transformation, storage, processing, analysis, and governance [9]. However, Genomics has specific challenges that must be taken into consideration.

Genomic data are spread in hundreds of repositories. This means that instead of having only one shared ecosystem to process the data, they are distributed among many independent actors with no centralized control or coordination. In addition, genomic data: i) are represented in multiple formats (e.g., FASTQ, VCF, BAM, unstructured text descriptions, and quantitative measurements from laboratories); ii) come from different technologies (e.g., DNA sequencing, protein-DNA binding occupancy, health records, and sensors); iii) cover different domain scopes (e.g., gene regulation, protein interactions, epigenetic interactions, and DNA methylation), and iv) have different levels of quality (e.g., the ChIP-seq data is sparse, noisy and discontinuous). Furthermore, the knowledge of the domain is in continuous evolution because of the ambitious human challenge of understanding the genome.

To succeed in providing a reliable and accurate genetic diagnosis in this Big Data scenario, some aspects are strictly required: i) retrieval and storage mechanisms to improve data acquisition and ensure scalability; ii) a high level of data quality must be guaranteed to facilitate the knowledge extraction process which is threatened by redundant and uncertain data; and iii) advanced data visualizations that combine appearance and functionality must be provided to extract value from the data.

In order to provide a solution to these problems, the novelty of this paper is to reuse and adapt current Big Data-based techniques to propose a framework that is: i) scalable enough to deal with the increasing amount of data that is being generated; ii) flexible enough to adapt to the dynamicity of the knowledge; and iii) searchable enough to extract valuable insights to support variant classification and interpretation for the genetic diagnosis. The use of a method that is specifically designed for managing genomic data and supported by conceptual modeling conforms the originality of the present work.

To achieve this objective, this work is structured as follows: Section 1 presents the introduction, Section 2 provides more details about the problems that constitute what we call “genomic data chaos”. Section 3 presents the Big Data background that conforms the methodological and technological basis of this work. Section 4 describes the proposed framework, focusing on its architecture and justifying why its components have been selected and how they are combined. Section 5 describes a real example in which the framework has been successfully applied. Finally, Section 6 presents our results, conclusions, and future work.

2. Genomic data chaos

Genetic diagnosis is a complex process that involves the extraction, transcription, and organization of genetic and clinical data from disjointed data sets into an Information System. However, this process is hindered by some issues that are inherent to the domain that must be carefully considered in order to succeed. We focus on the lack of standardized nomenclature, the problem of huge data dispersion, and the strict need for the use of reliable data.

2.1. Lack of a standardized nomenclature

As the genetics field emerged, naming conventions were not defined, which means that there are genes, proteins, organisms, diseases, technologies, and protocols that do not follow any nomenclature standard. This means that the same concept can be represented in different and sometimes ambiguous ways, leading to open debates in the community [10]. Therefore, it is not always

clear how to make correspondences between the different concepts that are represented by each data source, causing redundancies and conflicts.

Great efforts have been made by the community to revert this situation by providing ontologies and nomenclature standards with the aim of unifying knowledge and making it interoperable through consistent vocabularies. Examples of such efforts are the following: HUGO Gene Nomenclature Committee (HGNC) [11], which is responsible for approving unique symbols and names for human genes; HGVS Sequence Variant Nomenclature [12], which proposes a standardized way of naming sequence variants; and the Human Phenotype Ontology (HPO) [13], which proposes a standardized vocabulary of phenotypic abnormalities that are encountered in human disease.

Nevertheless, even when the use of such recommendations is highly encouraged, it is still not a standardized practice, which hinders the process of identifying the genomic elements and also hinders establishing the correct connections among them.

2.2. Huge data dispersion

Besides the ontological problem, there is a second one that is related to the dispersion of genomic information. All of the knowledge is spread over thousands of heterogeneous databases with different sizes, formats, and structures. Some of these sources store information about one organism (e.g., Flybase for Drosophila [14], RAP-DB for rice [15], and GDB for humans [16]). Others provide information about specific parts of the genome (e.g., Uniprot for proteins [17], HGMD for genes [18], and Reactome for pathways [19]). Each of these data repositories represents a different and complementary view of the whole picture. However, the correct understanding of the role that each genomic element plays in the development of a disease requires more information than that provided by only one source.

Bringing together all of these heterogenous and distributed databases can lead to interoperability issues that are related to semantic heterogeneity, data integrity, data representation, and correctness of the interpretation of the data sets obtained from them.

2.3. Lack of reliability

The third challenge to face is related to the lack of reliability because the information may contain errors caused by the complexity of biological processes, the noisy nature of experimental data, and the diversity of sequencing technologies. This leads to a great variability in the quality of the available information. For example, probe design and experimental conditions are known to influence signal intensities and sensitivities for many sequencing technologies [20], experiments performed on a population sample that it is not representative enough can lead to erroneous conclusions [21], and the use of different criteria and methods can lead to conflicts in variant classification [22].

All the above-mentioned problems constitute what we call “genomic data chaos”, which is associated with having a huge number of different, complex, and diverse data sources where the relevant genomic data is stored in partial views, a holistic perspective is missing, and there are problems that are well known to the Big Data community. These include lack of consistency, different formats for representing similar data, lack of conceptual standards, and difficulties with data heterogeneity and data interoperability management.

This chaos leads to data analysis processes that are mainly manual, tedious, and repetitive, and that are no explicit or systematic methods, they are prone to human errors and make repetitive navigation through complex hyperlinks unavoidable.

3. Methodological and technological background

In a previous work [23], we presented the basis for the efficient management of genomic information. The core of this work is the SILE method, which considers the problems of genomic data chaos and proposes a systematic approach that is divided into four main stages:

- Search: In this stage, the most appropriate and relevant sources are selected to extract the required data for the task at hand.
- Identification: During this stage, the relevant and high-quality data from each source is identified.
- Load: In this stage, the data identified as relevant is loaded into the appropriate storage repository taking into account the analytical requirements.
- Exploitation: In this stage, the value from the stored data is extracted in order to fulfill the knowledge requirements.

The method is supported by a sound ontological background and a data quality methodology that ensures the reliability of the data in each stage. It has been validated in different case studies and has proven to be useful in the identification of relevant variants that are associated to different diseases [24,25]. The SILE method provides the conceptual framework, but it requires sound Big Data architectural support in order to be used in a broader and more complex spectrum of diseases. This architecture must be based on existing Big Data solutions that are suitably adapted to the particularities of the domain under investigation (in our case, the genomic domain for diagnostic purposes in the context of Precision Medicine).

A Big Data system that is in charge of supporting the different stages of the SILE method must be based on a suitable infrastructure that fulfills the processing requirements identified by Krishnan in [26]. These that can be summarized as follows:

- Flexible data models to adequately manage complex data.
- Scalable systems to collect and process data either in real-time or in batches.
- Data partitioning to adequately support the volume of data.
- Efficient and fault-tolerant storage mechanisms.
- Replication across multiple nodes.

There are two well-known Big Data architectures that can be used to fulfill these requirements: The Lambda Architecture and the NIST Architecture.

The Lambda Architecture [27] is divided into three main components or layers that satisfy different needs: batch, serving, and speed. The batch layer stores all of the data in an immutable and constantly growing master dataset, the serving layer contains views of precomputed data from the master dataset in an indexed storage, and the speed layer computes the functions that propagate the views between the batch and the serving components.

The NIST Big Data Reference Architecture [28] is an open reference architecture for Big Data that is divided into five main components: System Orchestrator, Data Provider, Big Data Application Provider, Big Data Framework Provider, and Data Consumer (see Fig. 2).

The System Orchestrator ensures that the different applications, data, and infrastructure components all work together. The Data Provider introduces new data or information that come from different sources into the Big Data system for discovery, access, and transformation. The Big Data Application Provider contains the business logic and functionality necessary to transform the data into valuable knowledge through five main activities: collection, preparation, analytics, visualization, and access. The Big Data

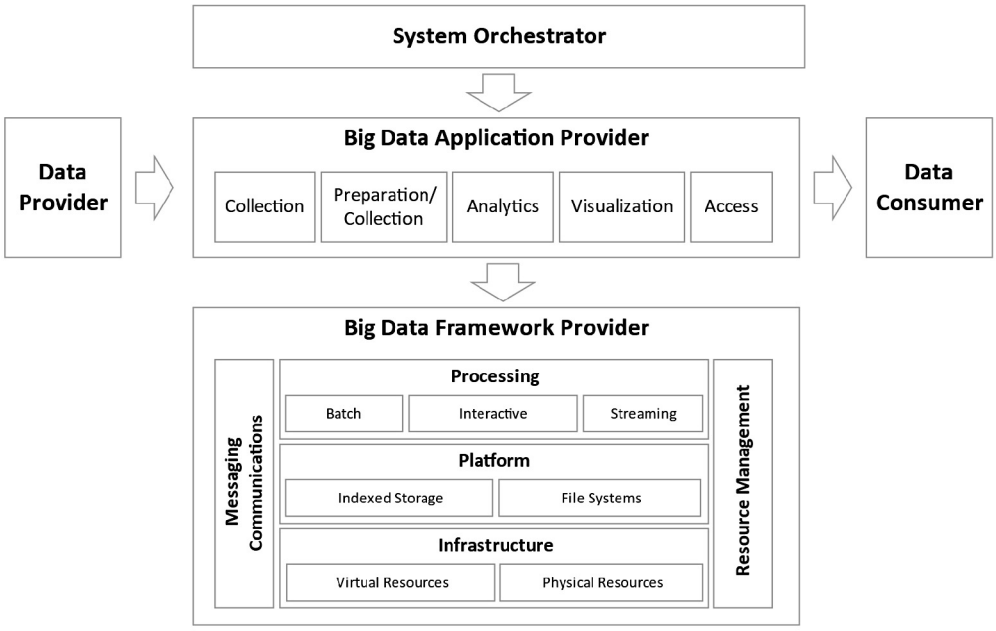

Fig. 2. NIST Big Data Reference Architecture (NBDRA) adapted from [28].

Framework Provider has the resources and services for data storage and processing. Finally, the Data Consumer uses the interfaces or services provided by the Big Data Application Provider to get access to the information of interest.

Besides the NIST and the Lambda architectures, there is a new approach proposed in [29], that extends the NBDRA with new components with the aim of providing a reference architecture to build Big Data Warehousing systems.

4. A Big Data framework for the management of genomic information

Using the four stages of the SILE method as basis along with the recommendations of the above-mentioned works, we present a Big Data framework with the aim of improving the variant classification and interpretation processes in genetic diagnosis. The aim of this framework is to solve, or at least minimize, the impact that the problems derived from genomic data chaos create for researchers and clinicians. To such aim, the presented framework can be used to help data scientists to build information systems that can help clinical experts to analyze and interpret the genetic data, using the results obtained to improve the genetic diagnosis.

One of the characteristics of genomic data is that they are collected in batches and consequently the implications of real-time or streaming data collection do not need to be considered. Nevertheless, these requirements could be added to the framework in a future if needed.

As Fig. 3 shows, the framework simplifies the Lambda Architecture and the NBRDA, and adds a new specific component (Quality Assessment), not explicitly considered in the mentioned architectures, that oversees supporting the core tasks of the Identification stage. The framework is divided into four main components: The Data Provider, the Data Consumer, the Big Data Application Provider, and Big Data Storage (which is a simplification of the Big Data Framework Provider to be adapted to the genome data that is managed).

Following the systematic steps defined by the SILE Method, the different modules of the framework can be implemented. In the following sections, details about how each module supports the tasks of the method are presented.

4.1. The Data Provider and the Data Consumer

The first stage of the SILE method (Search) requires determining the sources that provide the data to the system. These data sources must be determined by the clinical expert and will be part of the Data Provider component. They must provide complete information about the DNA variants, their location in the genome, the frequency of appearance in different populations, and the evidence collected by the scientific community in different published literature. Consequently, they will have different scope, format, and structure. The selection of the sources among all of the currently available ones is a huge problem by itself that is out of the scope of this work.

The Data Consumer can be an end user or another system that uses the results of the Big Data Application Provider to perform different activities such as search, retrieve, download, analysis, reporting, or visualization. As examples of activities that can be performed by the Data Consumer are the processing of a patient sample to find DNA variants that are causative of disease, the generation of a genetic report based on the findings, or just the visualization of the stored data to infer new knowledge. This is closely related with the last stage of the SILE method (Exploitation).

4.2. The Big Data Application Provider

The Big Data Application Provider module is responsible for ensuring the data flowing through the different activities that support the stages proposed by the SILE method.

Once the sources are determined, the Search stage also involves the collection of data from them. This task is supported by the Raw Data Collection module. After the collection of data, the relevant information must be identified. The Identification stage is supported by the Data Preparation and Quality Assessment modules. The Quality Assessment module has been added to the framework due to the importance that this task has for the identification of relevant genomic data. Once the relevant information has been identified, it must be stored into the system. The Load stage is performed on two different types of data storages (Raw Data Storage and Indexed Storage). Finally, the stored information can be analyzed in the Exploitation stage, supported by the Data Exploitation

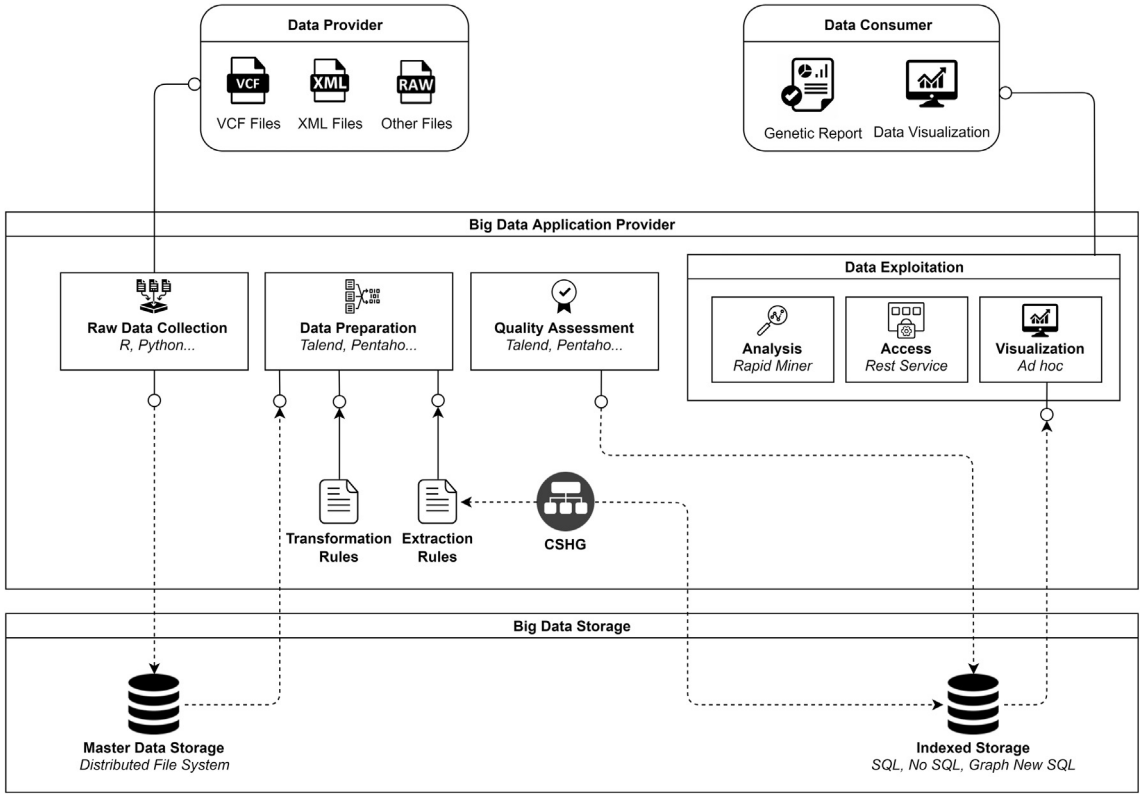

Fig. 3. The Big Data framework that supports the SILE method.

module through three main tasks namely Analytics, Access, and Visualization. In the following subsections we detail each component of the Big Data Application Provider.

4.2.1. The Raw Data Collection module

Once the data sources that will provide the input data to the system are determined, the next step is to extract the raw data from them. The genetic sources may provide data in different formats (e.g., XML, CSV, VCF, and JSON) and can be collected manually or automatically. The raw data is collected with the same level of detail that they have in the source, so they can serve different analytical purposes. The collection may require the implementation of custom collectors developed using some well-known programming languages such as Java and Python.

The collected data are stored in a raw data storage to be processed in the next stage. This storage must be flexible enough to store data in the raw state and serves not only for data preparation and transformation, but also for other purposes, such as training data science models or text mining tasks.

4.2.2. The Data Preparation module

Once the raw data have been extracted from the data sources, the next stage of the SILE method (Identification) requires the preparation of the data stored in the raw data storage repository for analysis. Depending on the analytical purpose, not all of the data extracted from each source may be required, so determining which information will be analyzed requires specific knowledge about the domain. In this case, the knowledge is provided by the Conceptual Schema of the Human Genome (CSHG) [30], which represents the ontological structure of the core concepts of the genomic domain and the relationships among them. As an example, Fig. 4 shows how the CSHG represents the contextual information about a DNA variant.

The main entities of the conceptual schema are Variation and Gene. The Variation entity represents the changes in the DNA that are the cause of the disease (phenotype) of interest. There are different types of variants, depending on the frequency of appearance in a certain population and the precision of the information associated to them. The Gene entity represents the elements whose alteration derives in a malfunction that leads to the manifestation of the disease. The conceptual schema also represents the information that is associated to the databases from where the information has been extracted to ensure the traceability of the information and to help keep the information updated.

The extraction of the required data must be done following a set of extraction rules defined according to the different analytical tasks required by the clinical expert. Each extraction rule is a logic formula with variables on its left-end side that are computed from the variables on its right-end side. These rules are defined by the data scientist in charge of the system development and maintenance. As an example, Fig. 5 shows how the information about a DNA variant can be extracted from the data coming from different data sources.

According to the example, a variant can be represented by three attributes. The attribute db_variation_id can be extracted from GWAS, ClinVar (requires transformation), Ensembl, and dbSNP databases. The attribute clinically_importance is provided by ClinVar and Ensembl, and the information associated to the attribute other_identifiers is provided by the attribute DOCSUM of dbSNP.

Some data are acquired exactly as they are in the original source, but others need the application of transformations that are specified in a set of transformation rules. The preparation of the data requires all sorts of cleansing, integration, and deduplication tasks that can be performed with tools like Rapid Miner, Talend Open Studio, and Pentaho Data Integration.

4.2.3. Quality Assessment module

One of the key stages of the SILE method is the identification of relevant and high-quality data through the application of a set of

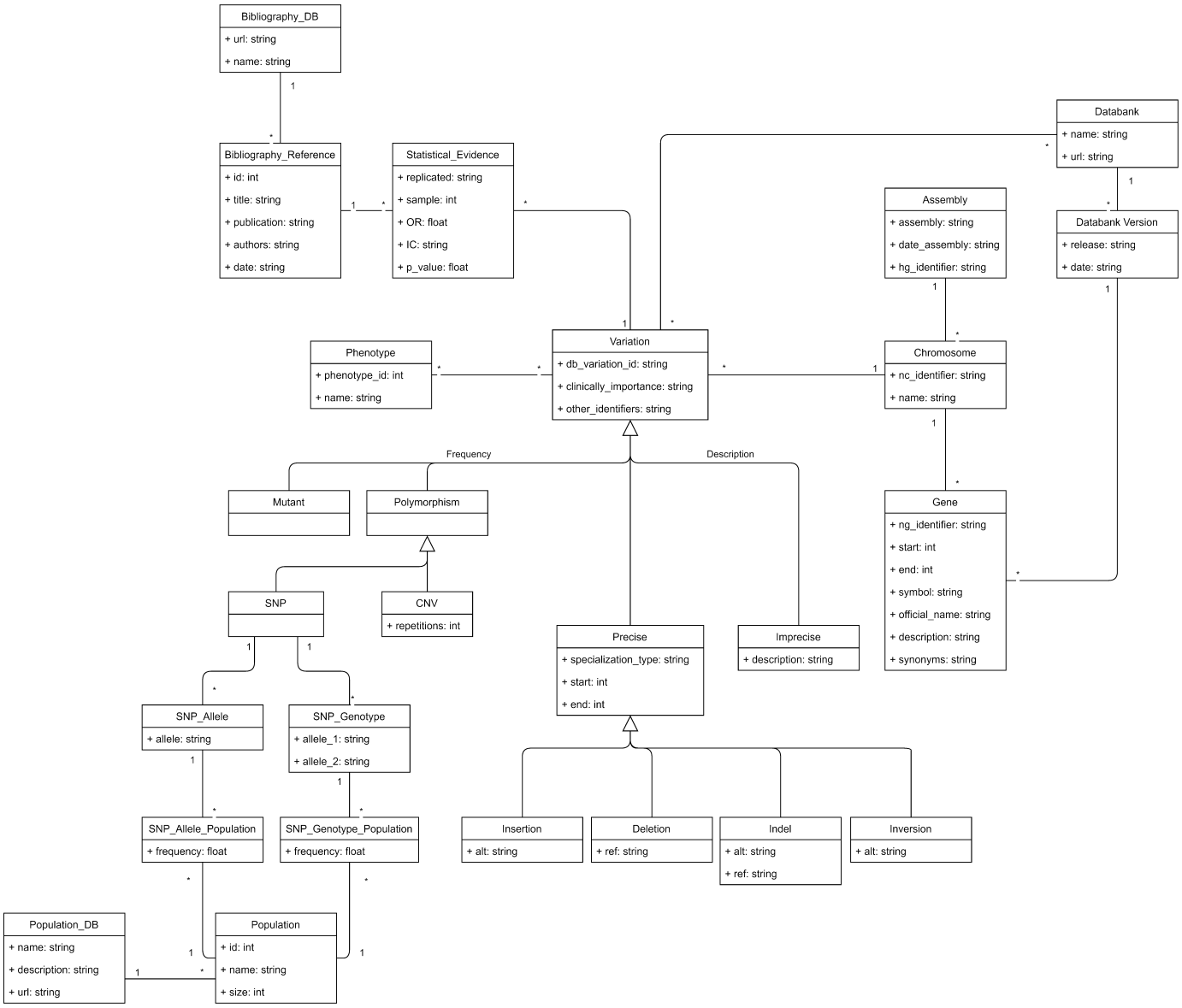

Fig. 4. Simplified view on how the CSHG represents the context of a DNA variant.

Variation(db_variation_id,-,-)	⊇	GWAS(tr(SNP_ID_CURRENT))
Variation(db_variation_id,-,-)	⊇	ClinVar.variation(tr(db_id))
Variation(-, clinically_importance,-)	⊇	ClinVar.clinical_significance(description)
Variation(db_variation_id, clinically_importance,-)	⊇	Ensembl(refsnp_id,clinical_significance)
Variation(db_variation_id, clinically_importance, other_identifiers)	⊇	dbSNP(SNP_ID, CLINICAL_SIGNIFICANCE, DOCSUM)

Fig. 5. Example of the extraction rules required to represent a DNA Variant.

quality metrics. This ensures that the conclusions and the knowledge derived from their analysis are reliable and accurate enough to be used for clinical purposes. These metrics must be defined by the clinical expert and may consider different quality issues in the genomic information such as the presence of conflicts in the literature, the statistical evidence, or the presence of data that is out-of-date [31]. These quality metrics can be implemented with ETL tools like Pentaho or using languages like R, Python, etc.

After the extraction and transformation process, the data conform a common format and structure (determined by the CSHG), and the quality assessment can be executed to determine if the data comply with the quality thresholds stablished. If this is the case, the data are stored in an indexed storage repository for their further analysis.

4.2.4. Data Exploitation module

The last stage of the SILE method (Exploitation) requires providing access to the data as well as a set of visualization mechanisms and tools that are specifically designed for the user requirements. This task is crucial since it allows access to the data that is stored in both storage systems for different analytical purposes; for example, the modification of the quality thresholds to be adapted to the different disease contexts and the extraction of relevant insights

through the processing of patient samples and the generation of genetic reports.

4.3. The Big Data Storage

This component is a simplification of the Big Data Framework Provider. It considers only the two types of storage required to process batch data (the master dataset and the batch views). The main roles of this module are: i) to provide the infrastructure required to ensure that large and diverse formats of data can be stored and transferred in a cost-efficient, secure and scalable way; ii) to facilitate and organize distributed processing in distributed storage solutions; and iii) to deliver the functionality to query the data and perform runtime operations on the data set.

As already stated, the Master Data Storage must be flexible and scalable enough to store data in the row state, without any constraints on volume and format. Since genome data is provided in batch mode, one solution could be the use of a Distributed File System (DFS). This is an unstructured data storage repository that does not need to have a specific schema or to be modeled in a specific way and allows the data to be explored in the raw state. There are several DFS providers, but one of the most well-known is Hadoop DFS, which is highly fault-tolerant, supports large datasets, and is designed to be deployed on low-cost hardware [32].

The Indexed Storage Repository is in charge of storing the computed views that come from the Master Data Storage and must comply with a specific structure. In this case, it is the one provided by the ontological background specified in the CSHG. Several options are available for this purpose, from SQL to NoSQL/New SQL systems or graph databases. Unlike the Master Data Storage, the Indexed Storage is updated as new data, which are useful from an analytical point of view, enter the system. It must also consider scalability, but the volume would be less that the one managed by the first storage system, and, therefore, other characteristics such as query performance, data consistency, and security are preferable.

Most of the modules above described, can be automated. Nevertheless, due to the current complexity of the domain, some tasks must be performed manually. For example, when a new data source is added to the system, the extraction and transformation rules must be defined by the data scientist in collaboration with the clinical expert that has deeper knowledge of the domain, to correctly harmonize the data into a common structure. Once this is done, the extraction, transformation, quality assessment, and the rest of the tasks can be automated.

5. Evaluation in a real example: epilepsy

In order to prove that the framework proposed is useful to solve the problems mentioned in the introduction, we have applied it to a real example in collaboration with a group of experts in genetic diagnosis. The aim of the example is to provide an Information System that helps the experts to find the relevant DNA variants that are associated to a higher risk of having epilepsy.

Epilepsy is a spectrum condition that has a wide range of seizure types, which what makes the gathering and analysis of the genetic information a challenge [33]. Furthermore, epilepsy means the same thing as “seizure disorders”, and the word “epilepsy” does not indicate anything about the cause of the person’s seizures or their severity. Many people with epilepsy have more than one type of seizure and may have other symptoms of neurological problems as well, which can be defined as an epilepsy syndrome. In addition, most individuals with genetically determined epilepsy are thought to have a polygenic basis in which multiple genes of low-to-moderate risk interact (sometimes with an environmental contribution) to produce the epileptic disease [34]. Thus, to provide an accurate genetic diagnosis, it is crucial to manage as much information as possible, which is a complex and time-consuming task for researchers. To such aim we have followed the different stages of the SILE method in order to build an information system to support the process.

To accomplish the objective of this example, different data sources were selected to provide a global view of the genetic context of the disease:

- DNA variants: The data about all of the DNA variants that the scientific community has studied in connection with epilepsy have been extracted from three different databases, namely ClinVar [35], dbSNP [36], and Ensembl [37].
- Genes: The data about the genes whose alteration could lead to epilepsy have been extracted from Entrez Gene [38] and HGNC [39].
- Genomic context: The data about the location of the variants in the genome have been extracted from NCBI Assembly [40].
- Population studies: The data about the frequency of appearance and the populations where the variants were studied have been extracted from 1000 Genomes [41].
- Published literature: The data about the different studies performed by the scientific community have been extracted from PubMed [42] and GWAS Catalog [43].

Each of these sources has particularities when accessing and downloading the required data, which hinder the integration process. For example, to access ClinVar data the database provides the following options: i) performing a manual search using the website and downloading the data in CSV format, ii) downloading the entire database using the FTP server, iii) accessing a partial view of the data in VCF format, and iv) using the REST API (known as e-utils) [44]. Each of these options allows access to different content, with the most restrictive one being the manual search. In addition, the processing of the downloaded data is also different because the XML or JSON structure of the result provided by e-utils is totally different from the XML structure of the FTP files. In addition, VCF has a specific format that is different from the ones mentioned above. This can be extended to the rest of the sources, which gives an idea about the complexity of the task to be performed.

In our example, a specific connector to query and extract the information from each data source has been implemented using R and the different APIs that each source provides. For the data coming from NCBI (ClinVar, dbSNP, Gene and Assembly) we used the Rentrez library, that provides a useful wrapper for the functions provided by Entrez and is optimized to work in R. For Ensembl we used the biomaRt package, and for the rest of sources we implemented the corresponding REST services to extract the raw data. These connectors have been included in a R package that experts can reuse to extract data from these sources about other interesting diseases.

After the data sources were determined, we defined the set of mapping and transformation rules. These rules were also implemented in R, extending the functionality of the package with the possibility of integrating the data coming from the selected sources into a homogenized dataset.

As Fig. 6 shows, data about 11,506 DNA variants, 1,509 genes, and 844 studies on four populations have been extracted from the sources and stored in the Master Data Storage. The current evaluation of this amount of data, which is mainly performed manually, is a task that can take weeks or even months to complete. The bottleneck that this situation produces when trying to provide a more accurate diagnosis for each patient is understandable, considering the increasing interest in characterizing the genetic causes of each disease and the huge number of existing diseases.

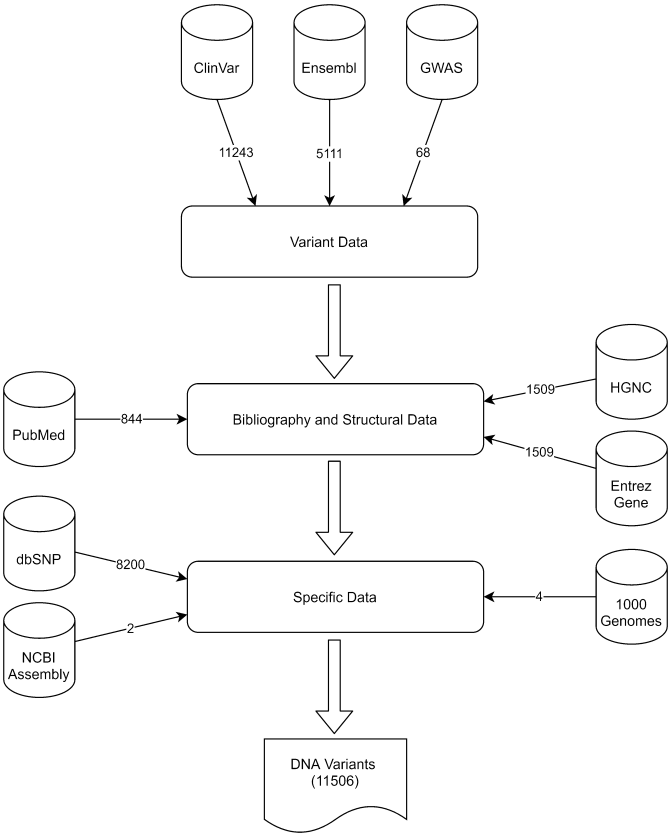

Fig. 6. Data extraction associated with epilepsy.

After establishing the extraction and transformation rules, the clinical experts defined the quality criteria and thresholds that the variants should fulfill to be considered as useful in a genetic diagnosis. Figs. 7 and 8 show the quality workflows that have been implemented using Pentaho Data Integration (PDI) to perform the quality assessment once the integration and cleansing is done. We used PDI due to the clinical experts considered this tool more intuitive than the implementation with other environments such as Python and R.

Starting from a dataset of variants, the different criteria that have been previously identified classify the variants according to their relevance for clinical purposes into four main categories: Discarded, Contradictory Evidence, Not Enough Evidence Provided, and Accepted. Accepted variants are also classified according to the strength of the associated evidence.

Starting from a list of literature identifiers, the different articles are classified according to the type of study performed. The classification is based on the analysis of the title, the abstract, the keywords, and the MeSH terms provided by PubMed. Depending on the type, a set of quality metrics are applied to assess the population and the statistical evidence regarding the number of participants.

The DNA variants that comply with the quality criteria are stored in the Indexed Storage, which in this case has been implemented as a relational database. This technology has been selected because it is well-known and widely accepted, it has a solid technological background, and it provides an intuitive organization based on the table structure that is close to the way the concepts are represented in the CSHG. In addition, data integrity is an essential feature of the relational databases; as well as they provide strong data typing, validity checks, and referential integrity that ensure the accuracy and consistency of the data. The Indexed

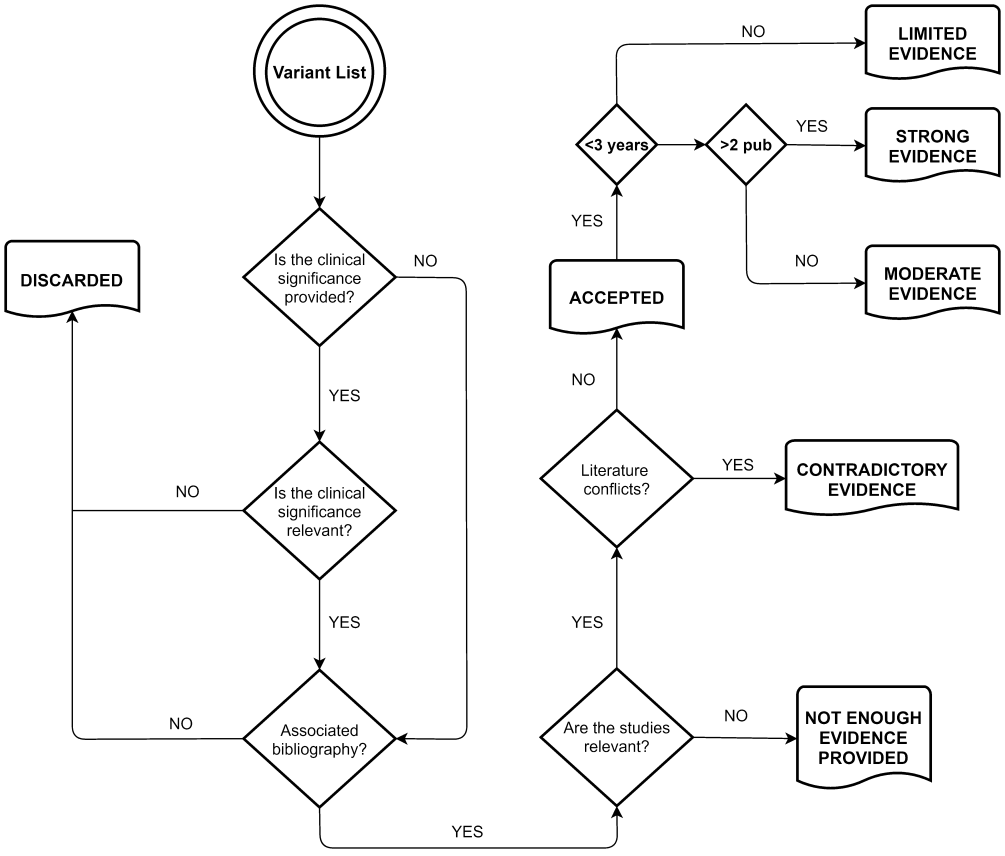

Fig. 7. Data Quality workflow to assess the relevancy of a DNA variant.

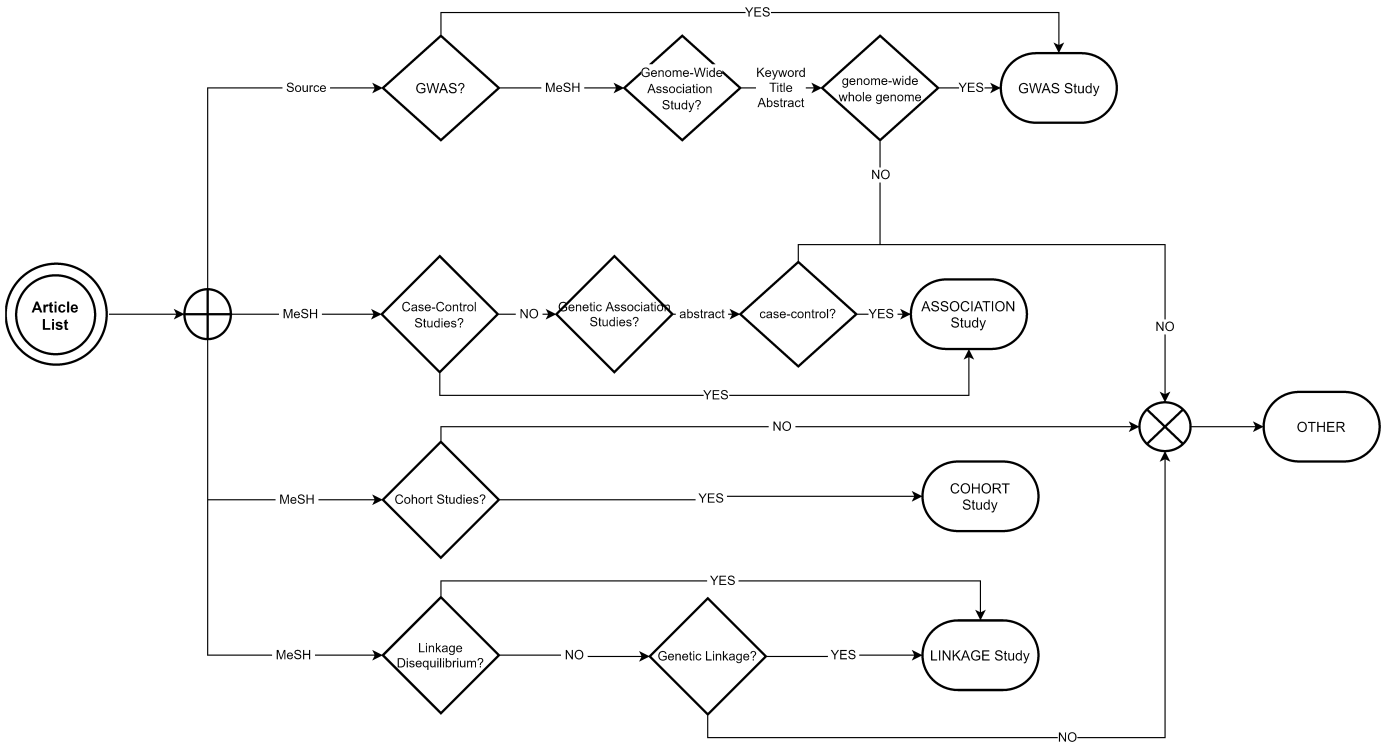

Fig. 8. Data Quality workflow to assess the statistical relevance of a genetic study.

Storage is called the Human Genome Database and it has been implemented using MySQL.

Once the data is stored in the Human Genome Database, the user can access a dataset of high-quality and reliable DNA variants as well as all of the contextual information about their position in the genome, the population where they were studied, and all of the relevant information required to perform an accurate genetic diagnosis. The data can be accessed through a web application that structures the content according to a set of visualization patterns specific for the genomic domain [45,46]. The rest of the data that do not have sufficient quality remain in the Master Data Storage and are available to be queried if the quality criteria are changed by the user or they are required to perform different types of analysis. As more data are collected from the sources, they are evaluated, and if they comply with the quality criteria, the Human Genome Database is updated.

To enhance the genetic diagnosis, another tool is under development which allows clinically relevant DNA variants to be found in a patient's sample. In this case, the user provides a VCF file with the information about the genome of a patient. The system returns a genetic report with the DNA variants that have been found to be associated with a higher risk of having epilepsy based on the data stored in the Human Genome Database. This tool has been developed as a prototype called GenesLove.Me.

6. Results

From the 11,506 DNA variants stored in the Master Data Storage, only 32 were stored in the Human Genome Database and thus are considered to be relevant to perform genetic diagnosis. These are the variants that will be checked and reported as highly reliable when a patient sample is analyzed.

The rest of the variants were discarded because contradictory evidence was found, not enough evidence to verify the relationship with the disease was provided, or because the results were not statistically significant. A further study of the discarded variants is required from the scientific community since the genetic causes of epilepsy are not yet fully understood; therefore, they should not be used if an accurate genetic diagnosis must be performed. Nevertheless, since knowledge is evolving so fast, as more evidence is collected, it will be added to the Information System and will contribute to re-evaluating the stored data and improving the process in a more efficient way thanks to its scalability.

Another advantage of using this Big Data framework is the possibility of analyzing different types of epilepsy and seizures. While the experts were focusing only on a part of the domain due to the complexity and amount of data to review, with this system, an analysis of the whole spectrum was performed allowing the possibility of finding new biomarkers with evidence that had initially been discarded or missed.

To validate the results, we consulted with companies that face the stated problem when interpreting genomic data. These companies provide genetic diagnosis services performed by geneticists and clinicians. As mentioned in the introduction, surprisingly, the diagnosis process is performed mainly manually or, in the best case, supported by tools such as Excel, which do not provide a suitable integration and analysis environment for this purpose. This constitutes an important loss of human resources for the companies because the study of complex diseases such as epilepsy can take weeks.

We have made an initial validation of the framework with some of the phenotypes that they manage (in addition to epilepsy), and they have confirmed the advantages that our approach introduces in their daily work. Once the selection of the data sources and the implementation of the system were finished, the time required to determine the relevant variants for a disease were decreased, improving their capacity to explore new knowledge and consider the exploration of additional data sources.

7. Conclusions and future work

The huge amount of data that technological advances in genomics produce has opened the door to new paradigms, such as Precision Medicine, for the prevention, diagnosis, and treatment

of human diseases. Nevertheless, these advances have occurred in such a short period of time that researchers and clinicians feel overwhelmed and without the appropriate tools to manage and extract valuable insights.

Big Data methods and techniques are expected to facilitate the right context to obtain the needed solutions. However, Big Data solutions need to be adapted to the particularities and complexity of the genomic domain. The application of a Big Data-driven approach that takes into consideration the special characteristics of genomic data constitutes a step forward in solving the bottleneck that tasks such as variant classification and variant interpretation produce.

The Big Data framework proposed in this work has been designed to provide an effective and efficient solution and complies with two important requirements of this domain: scalability to be able to gather the increasing amount of data, and flexibility to adapt to the fast-changing evolution of knowledge. It has proven to be useful in improving the genetic diagnosis of a particularly difficult disease (epilepsy) and is currently under evaluation in other disease contexts such as research in pediatric oncology and sudden cardiac death. In the use case presented it has been shown how the framework can be useful to solve the heterogeneity problems associated to the study of epilepsy. Nevertheless, for other diseases such as cancer, the main problem is the volume of the data to be managed (hundreds of thousands of variants). For such cases, the framework is prepared to provide the required support to analyze the data.

This proposal is in constant evolution, trying to solve emerging problems such as the specification of extraction and transformation rules that, in the real example, have been performed manually. This requires deep knowledge of each data source. New methods to automate the generation of the mapping rules required to perform the integration of the different datasets are also being evaluated. This will facilitate the addition of new data sources, which is a key task in such a complex domain. In addition, a new module to perform text mining is under development in order to deal with one of the most challenging tasks of genomic research, the extraction of data from the published literature.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Spanish State Research Agency (grant number TIN2016-80811-P) and the Generalitat Valenciana (grant number PROMETEO/2018/176), and co-financed with ERDF.

References

[1] A. Alzu'bi, L. Zhou, V.J.M. Watzlaf, Genetic variations and precision medicine, Perspect. Health Inf. Manag. (2019) 1–14.
[2] Z.D. Stephens, S.Y. Lee, F. Faghri, R.H. Campbell, C. Zhai, M.J. Efron, et al., Big Data: astronomical or genomical?, PLoS Biol. 13 (7) (2015) e1002195, https://doi.org/10.1371/journal.pbio.1002195. Available from: https://www.researchgate.net/publication/279863341_Big_Data_Astronomical_or_Genomical. (Accessed 17 March 2021).
[3] H. Yeo, C.H. Crawford, Big Data: cloud computing in genomics applications, in: Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015, 2015.
[4] A. Tarasov, A.J. Vilella, E. Cuppen, I.J. Nijman, P. Prins, Sambamba: fast processing of NGS alignment formats, Bioinformatics (2015), https://doi.org/10.1093/bioinformatics/btv098.
[5] V. Simonyan, R. Mazumder, High-performance integrated virtual environment (hive) tools and applications for big data analysis, Genes (Basel) (2014), https://doi.org/10.3390/genes5040957.
[6] R. Bao, L. Huang, J. Andrade, W. Tan, W.A. Kibbe, H. Jiang, G. Feng, Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing, Cancer Inform. (2014), https://doi.org/10.4137/CIN.S13779.
[7] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. (2010), https://doi.org/10.1101/gr.107524.110.
[8] U.S. Evani, D. Challis, J. Yu, A.R. Jackson, S. Paithankar, M.N. Bainbridge, A. Jakkamsetti, P. Pham, C. Coarfa, A. Milosavljevic, F. Yu, Atlas2 cloud: a framework for personal genome analysis in the cloud, BMC Genomics (2012), https://doi.org/10.1186/1471-2164-13-s6-s19.
[9] C. Costa, M.Y. Santos, Big Data: state-of-the-art concepts, techniques, technologies, modeling approaches and research challenges, IAENG Int. J. Comput. Sci. (2017).
[10] ResearchGate, What is the difference between polymorphism and a mutation?, https://www.researchgate.net/post/What_is_the_difference_between_polymorphism_and_a_mutation, 2019. (Accessed 15 December 2019).
[11] Hugo Gene Nomenclature Committee (HGNC), https://www.genenames.org/, 2019. (Accessed 15 December 2019).
[12] Human Genome Variation Society (HGVS), Sequence variant nomenclature, http://varnomen.hgvs.org/, 2019. (Accessed 15 December 2019).
[13] The Human Phenotype Ontology (HPO), https://hpo.jax.org/app/, 2019. (Accessed 15 December 2019).
[14] FlyBase, A database for drosophila genes and genomes, https://flybase.org/, 2019. (Accessed 15 December 2019).
[15] The Rice Annotation Project Database (rap-db), https://rapdb.dna.affrc.go.jp/, 2019. (Accessed 15 December 2019).
[16] S.I. Letovsky, R.W. Cottingham, C.J. Porter, P.W.D. Li, GDB: the human genome database, Nucleic Acids Res. (1998), https://doi.org/10.1093/nar/26.1.94.
[17] UniProt, https://www.uniprot.org/, 2019. (Accessed 15 December 2019).
[18] The Human Gene Mutation Database (HGMD), http://www.hgmd.cf.ac.uk/ac/index.php, 2019. (Accessed 15 December 2019).
[19] Reactome, https://reactome.org/, 2019. (Accessed 15 December 2019).
[20] J.S. Hamid, P. Hu, N.M. Roslin, V. Ling, C.M.T. Greenwood, J. Beyene, Data integration in genetics and genomics: methods and challenges, Hum. Genomics Proteomics (2009), https://doi.org/10.4061/2009/869093.
[21] N. Shah, Y.C.C. Hou, H.C. Yu, R. Sainger, C.T. Caskey, J.C. Venter, A. Telenti, Identification of misclassified ClinVar variants via disease population prevalence, Am. J. Hum. Genet. (2018), https://doi.org/10.1016/j.ajhg.2018.02.019.
[22] S. Yang, S.E. Lincoln, Y. Kobayashi, K. Nykamp, R.L. Nussbaum, S. Topper, Sources of discordance among germ-line variant classifications in ClinVar, Genet. Med. (2017), https://doi.org/10.1038/gim.2017.60.
[23] A. León Palacio, Ó. Pastor López, Smart data for genomic information systems: the SILE method, Complex Syst. Inform. Model. Q. (2018), https://doi.org/10.7250/csimq.2018-17.01.
[24] A.L. Palacio, I.P. Fernández, O.P. López, Genomic information systems applied to precision medicine: genomic data management for Alzheimer's disease treatment, in: C.S.B. Andersson, B. Johansson, S. Carlsson, C. Barry, M. Lang, H. Linger (Eds.), Int. Conf. Inf. Syst. Dev., Lund, Sweden, 2018, https://aisel.aisnet.org/isd2014/proceedings2018/eHealth/6.
[25] A. León Palacio, A. García Giménez, J.C. Casamayor Ródenas, J.F. Reyes Román, Genomic data management in Big Data environments: the colorectal cancer case, in: C. Woo, J. Lu, Z. Li, T. Ling, G. Li, M. Lee (Eds.), Adv. Concept. Model. ER 2018, in: Lect. Notes Comput. Sci., Springer, Cham, Xi'an, 2018, pp. 319–329.
[26] K. Krishnan, Data Warehousing in the Age of Big Data, 2013.
[27] N. Marz, J. Warren, Big Data: Principles and Best Practices of Scalable Realtime Data Systems, Manning Publications Co., 2015.
[28] NBD-PWG, NIST Big Data Interoperability Framework: Volume 6, Reference Architecture Interfaces, National Institute of Standards and Technology, Technical Report NIST SP 1500-6, June 2018.
[29] M.Y. Santos, C. Costa, Big Data: Concepts, Warehousing, and Analytics, FCA, Lisboa, 2019.
[30] J.F. Reyes Román, Ó. Pastor, J.C. Casamayor, F. Valverde, Applying Conceptual Modeling to Better Understand the Human Genome, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2016.
[31] A.L. Palacio, Ó.P. López, J.C.C. Ródenas, A method to identify relevant genome data: conceptual modeling for the medicine of precision, in: J. Trujillo, K.C. Davis, X. Du, Z. Li, T.W. Ling, G. Li, M.-L. Lee (Eds.), Concept. Model. - 37th Int. Conf. {ER} 2018, Xi'an, China, Oct. 22-25, 2018, Proc., Springer, Xi'an, 2018, pp. 597–609.
[32] Haddop, HDFS architecture guide, https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, 2019. (Accessed 15 December 2019).
[33] C.T. Myers, H.C. Mefford, Advancing epilepsy genetics in the genomic era, Gen. Med. (2015), https://doi.org/10.1186/s13073-015-0214-7.
[34] I.E. Scheffer, Epilepsy genetics revolutionizes clinical practice, Neuropediatrics (2014), https://doi.org/10.1055/s-0034-1371508.
[35] M.J. Landrum, et al., ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 46 (D1) (2018 Jan 4) D1062–D1067, https://doi.org/10.1093/nar/gkx1153.

[36] S.T. Sherry, M. Ward, K. Sirotkin, dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation, Genome Res. 9 (1999) 677–679.
[37] A.D. Yates, et al., Ensembl 2020, Nucleic Acids Res. 48 (D1) (08 January 2020) D682–D688, https://doi.org/10.1093/nar/gkz966.
[38] Gene [Internet], Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, 2004. Available from: https://www.ncbi.nlm.nih.gov/gene/.
[39] B. Braschi, P. Denny, K. Gray, T. Jones, R. Seal, S. Tweedie, B. Yates, E. Bruford, Genenames.org: the HGNC and VGNC resources in 2019, Nucleic Acids Res. 47 (D1) (2019 Jan 8) D786–D792.
[40] Assembly [Internet], Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, 2004. Available from: https://www.ncbi.nlm.nih.gov/assembly/.
[41] S. Fairley, E. Lowy-Gallego, E. Perry, P. Flicek, The international genome sample resource (IGSR) collection of open human genomic variation resources, Nucleic Acids Res. 48 (D1) (08 January 2020) D941–D947, https://doi.org/10.1093/nar/gkz836.
[42] PubMed [Internet], Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, 2004. Available from: https://www.ncbi.nlm.nih.gov/pubmed.
[43] A. Buniello, et al., The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, Nucleic Acids Res. 47 (Database issue) (2019) D1005–D1012.
[44] E. Sayers, E-utilities Quick Start. 2008 Dec 12 [Updated 2018 Oct 24]. In: Entrez Programming Utilities Help [Internet], Bethesda (MD): National Center for Biotechnology Information (US), 2010. Available from: https://www.ncbi.nlm.nih.gov/books/NBK25500/.
[45] C. Iñiguez-Jarrín, A. García, O.P. López, GenDomus: interactive and collaboration mechanisms for diagnosing genetic diseases, in: International Conference on Evaluation of Novel Approaches to Software Engineering, vol. 2, SCITEPRESS, 2017 April, pp. 91–102.
[46] C. Iñiguez-Jarrín, et al., Guidelines for designing user interfaces to analyze genetic data. Case of study: GenDomus, in: International Conference on Evaluation of Novel Approaches to Software Engineering, Springer, Cham, 2017.