Using Conceptual Modeling to Improve Genome Data Management

Óscar Pastor[1[0000-0002-1320-8471]], Ana León Palacio[1[0000-0003-3516-8893]], José Fabián Reyes Román[1[0000-0002-9598-1301]], Alberto García S.[1[0000-0001-5910-4363]1], and Juan Carlos Casamayor[1[0000-0001-5160-9092]]

[1] PROS Research Center, Universitat Politècnica de València,
Camino Vera s/n. 46022, Valencia, Spain
{opastor|aleon|jreyes|algarsi3}@pros.upv.es, jcarlos@dsic.upv.es

Abstract. With advances in genomic sequencing technology, a large amount of data is publicly available for the research community to extract meaningful and reliable associations among risk genes and the mechanisms of disease. However, this exponential growth of data is spread in over thousand heterogeneous repositories, represented in multiple formats and with different levels of quality what hinders the differentiation of clinically valid relationships from those that are less well-sustained and that could lead to wrong diagnosis. This paper presents how conceptual models can play a key role to efficiently manage genomic data. These data must be accessible, informative and reliable enough to extract valuable knowledge in the context of the identification of evidence supporting the relationship between DNA variants and disease. The approach presented in this paper provides a solution that help researchers to organize, store and process information focusing only on the data that is relevant and minimizing the impact that the information overload has in clinical and research contexts. A case-study (*epilepsy*) is also presented, to demonstrate its application in a real context.

Keywords: Genomic Data, Information Systems, Framework, Case Study, CSHG

Author biography

1. Oscar Pastor is Full Professor and Director of the PROS Research Center at the Universitat Politècnica de València (*Spain*). He is currently leading a multidisciplinary project linking Information Systems and Bioinformatics to designing and implementing tools for Conceptual Modeling-based interpretation of the Human Genome information.

2. Ana León Palacio is a Postdoc researcher of the PROS Research Center at the Universitat Politècnica de València (*Spain*). She Works on how to provide a systematic approach to efficiently manage genomic data.

3. Jose Fabián Reyes Román is a Postdoc researcher of the PROS Research Center at the Universitat Politècnica de València (*Spain*). He provides solutions on the genomic domain from a conceptual modeling perspective.

4. Alberto García Simón is a PhD student of the PROS Research Center at the Universitat Politècnica de València (*Spain*). He is currently studying how to improve genome data analysis in the agro-food field.

5. Juan Carlos Casamayor Rodenas is Associate Professor and Researcher of the PROS Research Center at the Universitat Politècnica de València (*Spain*). His areas of research are focused on Databases and Information Systems Design.

Summary key points

- A huge amount of genomic data is publicly available for the research community to extract meaningful and reliable knowledge, while Genomic data generation is growing exponentially.
- Genomic data is heterogeneous, disperse, present a wide range of levels of quality and lacks ontological commitment. These hinders the application of genomic diagnosis into clinical practice.
- The use of conceptual models is essential to deal with complexity in terms of understanding and communication of relevant data. We present a systematic framework -conceptual model-based- to efficiently manage genomic data to improve the efficiency and reliability of genomic diagnosis.
- This framework uses a Conceptual Schema of the Human Genome to give structure and context to the data and it is supported by a data quality methodology specially designed for genomic information management.
- The framework proved to increase the efficiency and reliability to determine variants that are associated with a higher risk of suffering epilepsy.

1 Introduction

During the last decades, technological advances in the field of genomics have allowed techniques like *Next Generation Sequencing* (NGS) [1] to become a routine research tool, improving our understanding on how inherited genetic differences among individuals (DNA variants) are involved in the risk of suffering a disease or present a particular trait. However, the volume of available information that these techniques produce increases in a faster pace than the ability of researchers to connect and analyze it.

The lack of a clear ontological basis to define the key concepts and the continuous evolution of the domain knowledge generate too frequently a different and (*potentially*) ambiguous representation for common concepts [2,3]. Furthermore, all the knowledge is spread in over thousand heterogeneous databases with different sizes, formats and structures. The information may also contain errors caused by the complexity of the biological processes, the noisy nature of experimental data and the diversity of sequencing technologies, that results in a great variability in the quality of the available information [4,5]. These challenges hinder the progress of sound and correct approaches such as *Precision Medicine* (PM) [6], which aim is to provide an accurate prevention, diagnosis and treatment of human diseases considering the genetic variability, the environment and the lifestyle of each person.

If we want to integrate the different perspectives of genomic data, a holistic view must be provided to facilitate the management of all those different genome dimensions, that go from the structural genotype to the functional phenotype. It is at this point that the use of Conceptual Modeling (CM) techniques helps to improve the understanding and the adequate communication of relevant data [7]. Therefore, the development of specific and correct conceptual model-based *Information Systems* (IS) gains significance. These IS allow:

i. The connection of different "*-omic*" fields (*Genomics*, *Proteomics*, *Pharmacogenomics*, etc.) under a common and structured perspective,
ii. The support of efficient management of genomic data that ought to be accessible, informative and actionable enough to infer valuable knowledge, and
iii. The use of appropriate tools to analyze the data and generate new knowledge [8].

On this basis, the design of a proper IS requires in the first place a sound ontological structure to represent and connect the heterogeneous elements of the domain [9,10]. Secondly, the system must support the efficient data management by defining a systematic process, from the selection of the appropriate data sources and the identification of relevant data, to their final load and exploitation to extract valuable knowledge. Finally, considering that the reliability of the results is highly dependent on the quality of the managed information, the IS must count with mechanisms to ensure that the results obtained during the process are of enough quality.

In this paper, we present a framework based on a *Conceptual Schema of the Human Genome* (CSHG) to manage genomic data in an efficient way with the aim of developing and populating an IS with high quality data. The framework is supported by a data quality methodology specially designed for genomic information. The quality of the managed information is ensured by the selection of high-quality repositories, extracting the highest quality information from each one and guaranteeing that the results derived from its analysis can be regarded as reliable and accurate. This approach can be applied to different parts of the biological domain; in this work we describe its application in the context of the identification of relevant genetic and experimental evidence supporting the relationship between DNA variants and disease. Given a particular phenotype, a well-known current problem is the identification of those variants that are relevant for the considered phenotype. This problem is especially important in diagnosis clinical contexts. The framework explained in the paper provides a sound and methodologically well-grounded solution.

The paper is structured as follows. Section 2 presents the relevant research in the domain regarding the use of conceptual models and data quality, and the need of having specific solutions for the genomic domain. Section 3 describes the proposed framework using the concepts and methods that are previously introduced. Along Section 4 the application of the framework is presented using a case study that ends with a discussion about the results that are obtained. Section 5 concludes the paper and proposes future research directions.

2 Conceptual Modeling and Data Quality in Genomics

Understanding the genome is probably the most complex scientific human challenge, and the complexity of the associated data is overwhelming. To face it, two main aspects must be considered: i) the use of conceptual models for the understanding of complex domains and ii) the use of data quality techniques to evaluate the accuracy, reliability and usefulness of the data to obtain meaningful conclusions from their analysis. These aspects provide a solid background to explore effective solutions. Nevertheless, the domain of Genomics has evolved so fast that these techniques must be adapted in order to extract all the potential benefits that the underlying knowledge has for the understanding, prevention and improvement of human health. In the next sections we analyze how both a sound conceptual model and a data quality methodology for genomics help to find out more adequate solutions.

2.1 A Conceptual Model for Genomics

The understanding of complex systems requires the integration of genomic data under well-constructed conceptual structures to describe the relationships between their components. However, the integration of different genomic databases is often challenging because they differ not only in the scope of the information they represent, but also in the way the same information is modeled. This situation hinders the process of retrieval, annotation and integration of heterogeneous datasets and consequently the quality of the conclusions derived from their analysis.

We want to emphasize that most of the existing solutions work at the "*solution space*" instead of working at the "*problem space*". By working at the solution space, we mean that they focus on representing data as they are used in practice, instead of focusing on the problem space, that faces the conceptual representation of the relevant domain concepts. Working at the solution space makes extremely hard to link data whose provenance is diverse, that use different formats and whose semantics is too often too imprecise. In this context, well-known solutions have been proposed by the scientific community. One of these solutions is the construction of ontologies. Examples of such ontologies are:

- *Gene Ontology* [11] describes the knowledge of the biological domain with respect to three aspects, i) molecular-level activities performed by gene products (molecular function), ii) the locations relative to cellular structures in which a gene product performs a function (cellular component) and iii) the larger processes accomplished by multiple molecular activities (biological process).
- *Sequence Ontology* [12] describes the features and attributes of biological sequences that are defined by their disposition to be involved in a biological process. There are also experimental features which are the result of an experiment.
- *Genotype Ontology* [13] represents the levels of genetic variation specified in genotypes.
- *Variation Ontology* [14] describes the effects, consequences and mechanisms of variations on DNA, RNA and/or protein level
- *Phenotype ontology* [15] describes phenotypic abnormalities encountered in human disease by providing an ontology of medically relevant phenotypes, disease-phenotype annotations, and the algorithms that operate on these.

There are around 244 biomedical ontologies, with over 5,700,000 terms and 27,000 properties according to the Ontology Lookup Service[1]. These ontologies are essentially large terminological resources that describe the terms used in the domain and the connections between these terms. But the use of all these domain ontologies in practice becomes a big problem. If we are interested in managing the particular data that one ontology describes (what we call a "*vertical*" data query dimension), everything can work reasonably well. But if we are interested in establishing valuable semantic connections among data provided not only by one ontology, but by many of them (what we refer to as an "*horizontal*" data query dimension), we have a problem. Imagine that we want to know the reason why a specific change in the genome (a DNA variant) produces the clinical manifestations of a disease. To solve this task it is required to navigate through the different concepts that connect the chromosomal elements affected by the variant (the transcripts, the genes, the proteins…), the functions that these elements perform (transcription regulation, ion transport, protein degradation…), the biological processes and reactions where these functions are involved (immunological response, transport of elements through the cellular components such as the plasmatic membrane, the normal growth of tissues,..) and the consequences of the malfunction of these processes that can be translated into the observable manifestations of a disease (recurrent infections, malabsorption, growth retardation, etc.). Navigating through all those different concepts to understand the precise semantic connection between genotype and phenotype requires to access many different data sources, each one normally specialized in a partial genomic dimension.

To solve the problem, the holistic perspective that only a global conceptual model can provide is strongly required. This is the essence of the solution presented in this

[1] https://www.ebi.ac.uk/ols/index

paper. A conceptual schema of the human genome is acting as a kind of conceptual database, that provides an unified perspective of all the relevant data that are in practice spread through a set of different data sources (data ontologies, databases, or however their creators call them).

The use of conceptual models proves to be a powerful tool for the understanding and communication of complex domains by making a clear definition of the entities involved and the relationships among them. The idea of applying conceptual modeling to understand the genome has been explored by some authors within which we can highlight:

I. The work presented by Chen *et al.* facilitated an *Object-Protocol Model* (OPM) [16] intending to bring an example of a combination of protocol and object constructions in a framework for the genomic domain, to provide modeling of objects and experiments (*protocol*).

II. The DNA Databank of Japan (DDBJ[2]) used conceptual modeling to design and develop a new version of their Nucleotide Sequence Databank to facilitate a rapid change and growth according to their system requirements [17].

III. The work presented in [18] introduces a cooperative computing environment (called "*Imagenetrade mark*") dedicated to the analysis and annotation of genomic sequences, which has been developed by applying an object-based model.

IV. The work presented by Paton [19] was focused on describing the genome from different perspectives, including the description of the eukaryotic cell genome, the interaction between proteins, the transcriptome and other genetic components, however, their work did not have a fruitful continuation in the domain.

V. In [20], the principles of conceptual modeling were applied in the context of 3D protein structures, which includes the consultation of large amounts of data and a very complex structure.

But these approaches still focus on specific parts of the domain, they are unconnected from each other and do not provide the required global view to understand complex biological systems. New attempts to provide a sound, CM-based solution is being proposed recently. For instance, in the work presented by Bernasconi *et al.,* a conceptual model of genomic metadata is proposed, whose purpose is to consult the underlying data sources to locate relevant experimental data sets [21]. Our work reinforces these approaches by using the Conceptual Schema of the Human Genome (CSHG) [22], developed in a previous work. The CSHG fills the gap providing a unified conceptual perspective to the partial ones that each of the above-mentioned solutions (ontologies and partial conceptual models) provide. Next, we explain in more detail the process that led to the development of a stable version of the CSHG.

Considering that the conceptual background of the genomic domain is under a constant evolution, it is important to highlight that to reach this holistic representation of the human genome, different versions of the model were generated, through a process where the ontological foundations of the relevant concepts were under continuous discussion. This was strongly required to facilitate the understanding and

[2] The DNA Data Bank of Japan

allowed to extend and integrate all the new concepts according to the evolution of the domain and the well-grounded knowledge. Below is a brief detail of each of the generated versions of the CSHG, that have made possible to reach a stable state:

- **CSHG version 1:** It is characterized by being the first attempt to address the holistic description of the genomic domain, and as such it uses a vision of the genome that is focused on its most basic concepts, ignoring some more complex aspects. CSHG version 1 focuses on the analysis of individual genes, their mutations, and their phenotypic aspects. In the modeling stages, four iterations were developed to reach the final version 1 of the CSHG (which can be consulted in [23]). This first version was classified into three main views: a) *Gene-Mutation View*; b) *Genome View*; and c) *Transcription View* [24]
- **CSHG version 1.1:** It is the natural evolution of CSHG version 1. This comprises the inclusion of the phenotypic view in the model. The "*phenotypic*" vision is very important because it provides consistency to the model. The fact of offering a “*genotypic*” vision (genetic information that an organism has), linked to a “*phenotypic*” vision (expression of genotype depending on a certain environment), offers a significant research value and it increases the semantic completeness of the model. As the introduction of the phenotype dimension did not change the essential semantics of the CM, a release change was considered enough.
- **CSHG version 2:** This version changes its central nucleus and goes from representing a "*genecentristic*" vision to a vision centered on the concept of "*chromosome*". This is why a version change was performed. Any relevant part of the genome is easily characterized as a particular chromosome part. This structural change of vision in the model represents the main difference concerning the previous versions of the model (v1 and v1.1). The full development phases that led to this evolution of the conceptual scheme can be consulted in detail in [22]. We decided to organize it in five mains “*views*” (see for instance [22]):
    - *Structural*: it describes the genome structure.
    - *Transcription*: it shows the components and concepts related to protein synthesis.
    - *Variation*: it describes the changes in the sequence of reference.
    - *Pathways*: it describes information about metabolic pathways.
    - *Bibliography and data bank*: it describes the sources of relevant data.

Currently, this stable version (CSHG version 2) of the model is being applied in practice, while continuous evaluation is in progress in order to be ready to generate any updated version that could better suit the needs of the domain. One of the practical cases of application of the model consists of studies on haplotypes and population genetics (which includes the integration of statistical -*biological*- patterns) resulting in the research work reported in [25].

Without renouncing at all to the holistic purpose of the CSHG, in practical terms we need to focus on a particular dimension of interest. Analyzing all the conceptual connections that are represented in the whole model is not possible in one paper. To make the problem treatable, we will focus on this work only on the *Variation view* data, that is especially relevant for facilitating a successful Medicine of Precision practice.

2.2 A Data Quality Methodology for Genomics

Once the conceptual structure is clear, the next step is to define a process to ensure that the information managed has enough quality to be used in the clinical practice. The information to be managed is complex because i) data sources can contain errors, and ii) these errors can be propagated to other data sources that use the original one, increasing the noise and the efforts required for their analysis [26].

The data quality connection is immediate. Even though data quality has been studied for decades, research on the quality of genomic data has just started and there are not sound results yet [27]. Most of the problems to face can be grouped into the assessment of six main categories or dimensions [4]:

- *Accessibility issues*: even though most of the data stored in the genomic repositories are publicly available, there are some issues that can hinder the access, such as the lack of mechanisms to automatically query and download the results of a search.
- *Completeness issues*: due to the extensive and ever-growing amount of available data, the process of manually marking specific features in a DNA sequence with descriptive information about its structure or its function is a time-consuming task. Therefore, some tools for automated processing and analysis of text are being developed to assist researchers in evaluating the scientific literature [28]. Although these tools speed up the annotation process, the heterogeneous nature of written resources and the difficulties of extracting knowledge embedded in free text (inconsistent gene nomenclature, domain-specific languages and restricted access to full text articles,...) mean that the information annotated with these tools can contain missing values, affecting the completeness of the databases.
- *Consistency issues*: genomic databases are very diverse, making extremely laborious to perform even simple queries across databases. As there is no standard format for genome data storage and no universally accepted vocabulary, consistency problems are especially significant when dealing with the terminology used to represent biological concepts. An example of these consistency problems is the classification of the type of DNA variants, which number ranges from 8 types (according to the HGVS[3] recommendations) to 31 (according to the ClinVar[4] database).
- *Currency issues*: as the underlying concepts are imperfectly defined, and scientific understanding of them is changing over time, the annotation of most genomes becomes outdated. There are also databases that do not have the required technological maintenance or do not review the information stored so they become obsolete too quickly. Consequently, most genome annotations remain static for years or have never been changed since their initial publication [29].

[3] https://varnomen.hgvs.org/

[4] https://www.ncbi.nlm.nih.gov/variation/docs/glossary/

- *Redundancy issues*: the lack of a precise ontological commitment tends to increase the redundancy in the collected data. The same entity can be submitted by different research groups to a database with different names, multiple times, or to different databases without a traceable cross-reference [30]. A high level of redundancy leads to an increase in the amount of data to be processed internally by the database and externally by the users. It also hinders the annotation process creating confusion and requiring additional time and effort to resolve missing, duplicate or inconsistent fields.
- *Reliability issues*: all the above-mentioned problems decrease the reliability of the stored information. To minimize these issues, some databases are supported by external experts that manually review the information and correct the errors found. Nevertheless, this is a laborious process that, together with the lack of well-constructed information systems, explains why the use of these repositories is not an extended practice yet, hindering the exploitation of the full potential that these databases can offer.

To address these problems and assure the veracity and value of the information, a Data Quality Methodology (DQM) must be defined. A DQM consists in a set of guidelines and techniques to define specific metrics in order to get a quantitative measure that represents the quality of the data. The methodology is divided into four main phases:

- *Dimension description:* the interesting dimensions to be measured are described together with their scope. For example, to determine the quality of a genomic repository, one of the dimensions to be measured can be Currency.
- *Metric description*: it describes the metrics associated to each dimension. For example, the currency of a database can be measured by its last update
- *Requirements description:* its objective is to define the minimum levels of quality that must be fulfilled by assigning concrete acceptance criteria to each metric. For example, the requirement for the last update metric is less than one year, which means that a database must have been updated less than one year ago.
- *Data Quality Assessment*: Once the dimensions, metrics and minimum requirements are established, a sound data quality assessment can be made by comparing the collected information and the minimum acceptance criteria that have been defined in the previous phase. For example, if the selected database has been updated less than one year ago, it can be considered as current.

These phases are mostly application-dependent so they can be adapted to any possible given scenario. Using the artifacts resulting from each phase, the quality assessment can be performed allowing the selection of the data that accomplish the levels of quality established by the user.

3 A Framework for the Efficient Management of Genomic Information

The framework presented in this paper uses the CSHG and the DQM described in the previous section and provides support to the four main tasks required to develop and populate an IS with relevant data:

- *The selection of the most adequate and reliable data repositories.*
- *The identification and management of the relevant information.*
- *The transformation of the subsequent data into a queryable format using the adequate technology, making them persistent.*
- *The analysis and extraction of the underlying knowledge.*

The schema of the framework (see Fig. 1) is divided into three sections that represent the main concepts that have been explained along this work. At the top of the figure, the different stages that conform the data quality methodology are represented (*Dimension Description*, *Metric Description*, *Requirements Description* and *Data Quality Assessment*). At the bottom of the figure, the ontological support is defined as i) a Conceptualization Process that must be carried out to precisely define the concepts of the domain, and ii) a formalization of this conceptualization that results in the conceptual model. Finally, at the center of the figure, the four main sequential tasks that must be performed to populate an IS with relevant data can be found (*Repository Selection*, *Information Identification*, *Information Persistence* and *Knowledge Extraction*). The data quality methodology and the ontological support interact with the three first tasks providing a set of input elements:

- The CSHG is the core of the ontological support and provides a unified structure to the information that comes from the different repositories. A set of mapping and transformation rules help to consistently represent the external data according to the structure defined by the CSHG. It also provides the association and constraint rules required to define the structure of the target database that will store the information. This process requires a conceptualization of the relevant information and its precise representation using the appropriate languages such as the diagrams provided by the *Unified Modeling Language* (UML, *https://www.uml.org/*).
- The DQM supports the process of data repository selection and relevant data identification by providing the different sets of quality criteria to be applied on each stage.

Finally, once the information is appropriately stored, a set of tools can be used to analyze and visualize the data in order to extract valuable knowledge.

In the next section, we are going to present the application of the framework in the context of a case study: the systematic identification of relevant genetic and experimental evidence supporting the relationship between DNA variants and epilepsy.

4 Presentation of the Framework through a Case Study

In order to explain the application of the framework, this section is structured as follows: first we explain how to prepare the conceptual and technological support to build the IS that will support the process (Section 4.1 to 4.4) and then we describe its execution to extract the required results (Section 4.5). These results have been positively validated by a group of experts in genetic diagnosis.

4.1 Defining the Ontological Structure

As it has been explained in previous sections, the key to build a sound IS is to ground it on a solid ontological commitment. In this example, we are focusing on solving a concrete task -the identification of relevant variants that are clinically related to the phenotype under study (*epilepsy*)-. This task requires to manage data about a specific part of the genomic domain. Using the holistic perspective provided by the CSHG as a basis, we can determine the parts of the model that provide the required structure to the data that is going to be managed. The resulting conceptual schema can be seen in Fig. 2.

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 *Phenotype* (disease) of interest. There are different types of variants, depending on i) the frequency of appearance in a certain *Population* (*Mutant, Polymorphism, CNV* and *SNP*) and ii) the precision of the information associated to them. If the location of the variant is known it is classified as *Insertion, Deletion, Indel* or *Inversion*. If the location of the variant is unknown, it is classified as *Imprecise*.

The *Gene* entity represents the elements whose alteration derives in a malfunction that leads to the development of the disease. As the traceability of the data must be ensured, the schema also considers the information associated to the databases where the genomic data have been extracted from, represented by the classes *Bibliography_DB*, *Databank* and *Population_DB*. This helps to keep the information continuously updated.

4.2 Selecting Relevant Sources

Once the conceptual schema is specified, the next step is the selection of the most adequate data sources to populate it. This process is based on the application of a set of data quality criteria in order to reduce the noise produced by the great number of publicly accessible repositories (over 1,500 according to a report of 2019 Molecular Biology Database Collection in the journal Nucleic Acids Research [31]). The selection of the adequate repositories is based on a set of quality criteria that considers the assessment of 3 dimensions and 5 metrics with their corresponding criteria of acceptance (see Table 1).

In addition, the transformation and integration processes require following a sequence of steps:

- *Data Processing*: it involves cleaning the extracted data as well as converting all the values to the required data types.
- *Data Transformation*: this step refers to the implementation of the transformation rules required to represent the information coming from the different repositories into a common structure.
- *Data Integration*: it consolidates the data under a single unified view.
- *Data Deduplication*: this step involves removing duplicate copies of repeating data.

After this process, the data is prepared to be filtered with the aim of reducing the noise and the negative consequences of analyzing low quality data.

4.3 Identifying the Relevant Information

As not all the information coming from the repositories is reliable enough to perform a genetic diagnosis, it is important to measure its quality in order to ensure that only the most reliable one is considered. To this aim we have defined and implemented a set of quality criteria in agreement with the group of experts in genetic diagnosis that will validate the results of the process. These criteria are summarized in Table 2.

We are aware of the complexity of the domain. Due to this complexity, the definition of the quality criteria can be adapted to any context of application by changing the default values or defining new criteria. At the end of the process, only the information that conforms the established criteria can be prepared to be stored adequately ensuring its persistency.

4.4 Information Persistence and Knowledge Extraction

With the aim of preparing the information for its further exploitation, it must be stored in a target database that conforms the structure provided by the CSHG. The selection of the adequate technology depends on the volume of the information to be managed as well as the further data analysis requirements. In addition, the process of providing persistency to the information requires to carry out the needed checks to provide strong data typing as well as referential integrity to ensure the accuracy and consistency of the data. In our case, we built a relational database because it is a well-known and widely accepted technology, with a solid technological background, and it provides an intuitive organization based on the table structure that is familiar to most users and close to the way the concepts are represented in the CSHG. It also adapts well to the type of queries that the problem requires, facilitating an effective data exploitation. These characteristics simplify the development and use of the database. In addition, data integrity is an essential feature of the relational databases. They provide strong data typing and validity checks as well as referential integrity, which ensure the accuracy and consistency of the data. Nevertheless, we are aware that other technologies such as NoSQL databases could be useful too, but on this case a SQL database fully complies with our requirements.

The final aim of this approach is to extract knowledge from the information stored in the database through the use of analytical and graphic tools specifically designed for this domain. In our case, we use the information to provide support to the identification of potentially damaging variants in the DNA of a patient, represented in a VCF file (the most accepted file format to represent and managing variants). This is a very complex and time-consuming task that is mainly performed manually. A tool called "*GenesLove.Me*" [32] has been developed to provide support to this task. Using the high-quality information stored in our IS, the tool generates an automatic report with the DNA variants present in a patient's sample, what is a valuable help for the geneticists/clinicians that only have to validate the results.

As a proof of concept, we explain in the next subsection the results obtained after processing a specific dataset with variants associated with epilepsy.

4.5 **Results**

We used the IS developed considering the previously explained framework to determine the relevant variants that are associated with a higher risk of suffering epilepsy. The selected sources to extract the data (according to the stablished quality criteria) have been *PubMed*[5], *NCBI Assembly*[6], *GWAS Catalog*[7], *ClinVar*[8], *Ensembl*[9], *dbSNP*[10], *HGNC*[11], *Entrez Gen*[12] and *1000 Genomes*[13]. This selection provides a reliable coverage of the most relevant genome data sources that is to be considered in the analyzed working domain. Using the automated connectors implemented to access data from each source, an initial extraction allowed us to integrate data from 11,506 variants, 1,509 genes and 844 articles [33, pp. 57-58].

Without the support of an IS, this information must be integrated and analyzed manually, which is human-error prone and it implies a waste of time and human resources. After applying the criteria defined to identify relevant variants according to the requirements of the experts, only 32 variants were considered as relevant to perform a genetic diagnosis (see Table 3, that summarizes the most relevant results). Almost 64% variants were discarded because they do not have a relevant clinical significance or associated bibliography to verify the assertion, and about 16% of the variants were discarded because the studies performed were not statistically relevant enough according to our criteria (presented in Table 2). This gives an idea of the importance of managing high-quality, accurate data in genomic contexts.

The results were validated by a group of experts as clearly relevant, reducing the effort required to query, integrate and analyze the information coming from the different sources.

[5] https://www.ncbi.nlm.nih.gov/pubmed/

[6] https://www.ncbi.nlm.nih.gov/assembly/

[7] https://www.ebi.ac.uk/gwas/

[8] https://www.ncbi.nlm.nih.gov/clinvar/

[9] https://www.ensembl.org/index.html

[10] https://www.ncbi.nlm.nih.gov/snp/

[11] https://www.genenames.org/

[12] https://www.ncbi.nlm.nih.gov/gene/

[13] https://www.internationalgenome.org/

4.6 Final Remarks

With the ever-increasing volume of information generated for curing or treating diseases and cancers, conceptual model-based technologies, tools, and techniques should play a critical role in turning data into actionable knowledge to meet unstated and unmet medical needs. This is the main objective of the work presented in this paper.

The case study that has been introduced can be generalized to any phenotype for which precise genome data have been registered. This facilitates the jump from prototyping to real application of the framework, defining what a sound and validated workflow framework means and determining how to balance agility needs in the identification of relevant variants while compliance and consistency of requirements are preserved. What is more important, this approach provides a sound methodological background to deal with potential inconsistencies and uncertainties in definitions and meta-data across the multiple datasets that form the basis of the complex, big genomic data world.

5 Conclusion and Future Work

The management of genome data is a complex and time-consuming task that requires a great effort from the researchers if they do not have the support of a systematic process and a well-grounded IS. The lack of a consistent process and the usage of non-standardized data result in sub-optimal identification of relevant (clinically speaking) variants and longer periods of time to obtain an accurate information. In this work, the non-standardized problem is prevented by using sound conceptual modeling background (introduced in Section 2.1), and a data quality methodology (presented in Section 2.2) to provide the required consistent process support.

Through the conceptual framework, we have stated the importance of using conceptual models to reduce the bottleneck that researchers must face when managing inconsistent, heterogeneous, dispersed and unaffordable huge amounts of biological data in continuous growth. The Conceptual Schema of the Human Genome provides a solid ontological structure to the data coming from diverse and heterogeneous repositories.

The data quality methodology ensures that the managed information has enough quality to support the genetic diagnosis in a clinical context. The application of the framework to determine the variants associated with the risk of suffering epilepsy has been proved to be useful in reducing the effort and time required to perform the entire process. The criteria applied in the case study were adapted to the needs of the experts that collaborated in the validation. Nevertheless, we are aware about the complexity of the domain and that the characteristics of each disease differ in which criteria must be considered and how they should be applied. To address these challenges, we are working on a continuous improvement of the CSHG to consider all the new knowledge that the community is generating day after day, such the role of haplotypes and pathways in the development of complex disease.

Additionally, we are working on providing an implementation of specific guidelines for the classification of DNA variants such as the ones provided by the American College of Medical Genetics (ACMG) and the Association for Molecular Pathology (AMP) [34], and the evaluations and adaptations performed by research groups like

ClinGen [35]. At the same time, we are preparing new case studies to corroborate that the results can be considered as accurate in as many scenarios as possible.

As a conclusion, the use of conceptual models and data quality techniques is the basis to build *dynamic*, *scalable* and *efficient* IS that could interoperate with other systems. It allows to efficiently manage all the data required to understand the complex mechanisms that conform the genomic domain, ensuring a global data structure (key for interoperability) and a high-quality data management environment (key for generating accurate and reliable knowledge).

Acknowledgements

The authors would like to thank the members of the PROS Research Center Genome group for fruitful discussions regarding the application of Conceptual Modeling in the medical field. This work has been developed with the financial support of the Spanish State Research Agency and the Generalitat Valenciana under the projects TIN2016-80811-P and PROMETEO/2018/176 and co-financed with ERDF.

References

1. McCombie, W. R., McPherson, J. D., & Mardis, E. R. (2019). Next-generation sequencing technologies. Cold Spring Harbor perspectives in medicine, 9(11), a036798.
2. Condit C, Achter PJ, Lauer I, *et al*. The changing meanings of "mutation:" A contextualized study of public discourse. *Human Mutation*, 2002;19(1):69-75.
3. Karki R, Pandya D, Elston RC, *et. al.* (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Medical Genomics*;8(1):37.
4. León Palacio A, Reyes Román JF, Burriel V, *et. al.* (2016). Data Quality Problems When Integrating Genomic Information. *in: ER 2016 Workshops. LNCS, Springer International Publishing*;173-182.
5. Hamid JS, *et al.* (2009). Data Integration in Genetics and Genomics: Methods and Challenges. *Human Genomics and Proteomics*.
6. Baudhuin, L. M., Biesecker, L. G., Burke, W., Green, E. D., & Green, R. C. (2020). Predictive and precision medicine with genomic data. Clinical Chemistry, 66(1), 33-41.
7. Olivé, A. (2007). Conceptual Modeling of Information Systems. Springer, Heidelberg.
8. León A, Pastor O, *et. al*. (2018). From Big Data to Smart Data: A Genomic Information Systems Perspective. *In 12th International Conference on Research Challenges in Information Science (RCIS)*;1-11.
9. Guizzardi, G., Herre, H., & Wagner, G. (2002, October). On the general ontological foundations of conceptual modeling. In International Conference on Conceptual Modeling, pp. 65-78. Springer, Berlin, Heidelberg.
10. Amaral, G., & Guizzardi, G. (2019, September). On the Application of Ontological Patterns for Conceptual Modeling in Multidimensional Models. In European Conference on Advances in Databases and Information Systems, pp. 215-231. Springer, Cham.
11. Ashburner, *et al*. (2000). Gene Ontology: tool for the uni_cation of biology. *Nature Genetics*,25(1):25-29.
12. Eilbeck K, *et al*. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology*,6(5):R44.
13. Geno-Ontology, *https://github.com/monarch-initiative/GENO-ontology/* [*Last access*: February 17, 2020]

14. Vihinen M. (2014). Variation Ontology for annotation of variation effects and mechanisms. *Genome Research*,24(2):356-364.
15. Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, et al. (2018). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research. doi: 10.1093/nar/gky1105
16. Chen I.-M.A, Markowitz V.M. (1995). Modeling scientific experiments with an object data model. *In Proceedings of the Eleventh International Conference on Data Engineering,* 39-400.
17. Okayama T, *et al.* (1998). Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library. *Bioinformatics*, 14(6):472-478.
18. Medigue C, Rechenmann F, Danchin A, *et. al*. (1999). Imagene: An integrated computer environment for sequence annotation and analysis. *Bioinformatics*, 15(1):2-15.
19. Paton NW, Khan SA, Hayes A, *et al*. (2000). Conceptual modelling of genomic information. *Bioinformatics*, 16(6):548-557.
20. Ram S, Wei W. (2004). Modeling the Semantics of 3D Protein Structures. *In Genome*, 696-708.
21. Bernasconi A, Ceri S, Campi A, *et. al*. (2017). Conceptual modeling for genomics: Building an integrated repository of open data. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (10650):325-339.
22. Reyes Román JF, Pastor O, Casamayor JC, *et. al*. (2016). Applying Conceptual Modeling to Better Understand the Human Genome. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 404-412.
23. Pastor O, Levin AM, Celma M, *et. al*. (2011). Model-Based Engineering Applied to the Interpretation of the Human Genome. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (6520):306–330.
24. Reyes Román, J. F. (2018). Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano. *PhD Thesis, Universitat Politècnica de València*, doi: https://riunet.upv.es/handle/10251/99565.
25. Reyes Román JF, Pastor O, Valverde F, *et. al*. (2016). How to deal with Haplotype data: An Extension to the Conceptual Schema of the Human Genome. *CLEI Electronic Journal*, 19(3):58-106.
26. Muller H, Naumann F. (2003). Data quality in genome databases. *In Eighth International Conference on Information Quality (ICIQ 2003)*, 269-284.
27. Vihinen M. *et al*. (2016). Human Variome Project Quality Assessment Criteria for Variation Databases. *Human Mutation*, 37(6):549-558.
28. Wilco W.M. Fleuren and Wynand Alkema. (2015). Application of text mining in the biomedical domain. Methods, 74:97-106.
29. Steven L. Salzberg. (2007). Genome re-annotation: A wiki solution? Genome Biology.
30. Chen, Q., Zobel, J., and Verspoor, K. (2017). Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. Database, 2017:baw163.
31. Rigden DJ, Fernández XM. (2019). The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection. *Nucleic Acids Research*, 47(D1): D1-D7.
32. Reyes Román, J. F., García, A., Rueda, U., & Pastor, Ó. (2019). GenesLove. Me 2.0: Improving the Prioritization of Genetic Variations. In International Conference on Evaluation of Novel Approaches to Software Engineering (pp. 314-333). Springer, Cham.
33. León Palacio, A. (2019). SILE: A Method for the Efficient Management of Smart Genomic Information. PhD Thesis, Universitat Politècnica de València; doi: https://doi.org/10.4995/Thesis/10251/131698.

34. Richards S, Aziz N, Bale S, et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, 17:405–424.

35. Kelly MA, et al. (2018). Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. *Genetics In Medicine*, (20):351.

Figure Legends

Page 10 - Fig. 1. Framework for the management of genomic information.

Page 12 - Fig. 2. Conceptual schema representing the information used in the case study

Table Legends

Page 13 - Table 1. This table summarizes the dimensions, metrics and criteria of acceptance used to select relevant genomic repositories.

Page 14 - Table 2. This table summarizes the dimensions, metrics and criteria of acceptance used to select relevant genomic data from a repository.

Page 16 - Table 3. This table represents a summary of the results obtained after the classification of the variants according to their relevance for the task at hand [33].