SILE: A Method for the Efficient Management of Smart Genomic Information

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Ana León Palacio

Advisor:
Prof. Dr. Óscar Pastor López

September - 2019

This thesis was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science by the Universitat Politècnica de València, and supported by the Research and Development Aid Program (PAID-01-16) under the FPI grant 2137.

Author:

Ana León Palacio

Advisor:

Prof. Óscar Pastor López

External reviewers:

Prof. Johann Eder, Alpen-Adria-Universität Klagenfurt, Austria
Prof. Jesús Peral, University of Alicante, Spain
Prof. Jolita Ralyté, University of Geneva, Switzerland

Examination committee:

President: Prof. Juan Carlos Trujillo, University of Alicante, Spain
Secretary: Prof. Maribel Santos, Universidade do Minho, Portugal
Speaker: Prof. Jolita Ralyté, University of Geneva, Switzerland

The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.

Marcel Proust

Acknowledgements

This thesis is the result of several years of hard work, during which I received the kind support and help of many people. I would like to extend my sincere thanks to all of them.

Firstly, I would like to express my gratitude to my advisor Prof. Óscar Pastor for his continuous support, his patience, motivation, and immense knowledge. His sense of humor and positivity have taught me to enjoy this adventure as well as to grow personally and professionally. Thank you for accepting to be my supervisor and especially thank you for being my friend.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Juan Carlos Trujillo , Prof. Maribel Santos, and Prof. Jolita Ralyté, as well as the reviewers for agreeing to be a part of the court (Prof. Johann Eder, Prof. Jolita Ralyté, and Prof. Jesús Peral). It has been a great honor to count on your participation in the last stage of the thesis.

Special thanks to José Marín for reviewing the preliminary version of this thesis, for his very valuable advice and his support when the end never seemed to come.

To all my colleagues of *The Hours Cartel*, specially to Ángel, Carlos, Lenin, Nana, Bea, José and Julio. For their friendship, their advice, their complicity and sense of humor, that helped me to disconnect from work and without which these years would not have been the same.

To all the colleagues of the PROS research center, specially to Vero, Alberto, Urko and Ana Cidad, for their support and fruitful discussions.

Last but not the least, I would like to thank my family, specially my parents for the inconditional love and support in the good times and the bad times, throughout the development of this thesis and my life in general. You kept me going and without you I would not have come this far.

Thanks for all your encouragement!

This work has been supported by the Universitat Politècnica de València under the FPI grant 2137.

Abstract

In the last two decades, the data generated by the Next Generation Sequencing Technologies have revolutionized our understanding about the human biology. Furthermore, they have allowed us to develop and improve our knowledge about how changes (variants) in the DNA can be related to the risk of developing certain diseases.

Currently, a large amount of genomic data is publicly available and frequently used by the research community, in order to extract meaningful and reliable associations among risk genes and the mechanisms of disease. However, the management of this exponential growth of data has become a challenge and the researchers are forced to delve into a lake of complex data spread in over thousand heterogeneous repositories, represented in multiple formats and with different levels of quality. Nevertheless, when these data are used to solve a concrete problem only a small part of them is really significant. This is what we call *"smart"* data.

The main goal of this thesis is to provide a systematic approach to efficiently manage smart genomic data, by using conceptual modeling techniques and the principles of data quality assessment. The aim of this approach is to populate an Information System with data that are accessible, informative and actionable enough to extract valuable knowledge.

Resumen

A lo largo de las últimas dos décadas, los datos generados por las tecnologías de secuenciación de nueva generación han revolucionado nuestro entendimiento de la biología humana. Es más, nos han permitido desarrollar y mejorar nuestro conocimiento sobre cómo los cambios (variaciones) en el ADN pueden estar relacionados con el riesgo de sufrir determinadas enfermedades.

Actualmente, hay una gran cantidad de datos genómicos disponibles de forma pública, que son consultados con frecuencia por la comunidad científica para extraer conclusiones significativas sobre las asociaciones entre los genes de riesgo y los mecanismos que producen las enfermedades. Sin embargo, el manejo de esta cantidad de datos que crece de forma exponencial se ha convertido en un reto. Los investigadores se ven obligados a sumergirse en un lago de datos muy complejos que están dispersos en más de mil repositorios heterogéneos, representados en múltiples formatos y con diferentes niveles de calidad. Además, cuando se trata de resolver una tarea en concreto sólo una pequeña parte de la gran cantidad de datos disponibles es realmente significativa. Estos son los que nosotros denominamos datos *"inteligentes"*.

El principal objetivo de esta tesis es proponer un enfoque sistemático para el manejo eficiente de datos genómicos inteligentes mediante el uso de técnicas de modelado conceptual y evaluación de calidad de los datos. Este enfoque está dirigido a poblar un sistema de información con datos que sean lo suficientemente accesibles, informativos y útiles para la extracción de conocimiento de valor.

Resum

Al llarg de les últimes dues dècades, les dades generades per les tecnologies de secuenciació de nova generació han revolucionat el nostre coneixement sobre la biologia humana. És mes, ens han permès desenvolupar i millorar el nostre coneixement sobre com els canvis (variacions) en l'ADN poden estar relacionats amb el risc de patir determinades malalties.

Actualment, hi ha una gran quantitat de dades genòmiques disponibles de forma pública i que són consultats amb freqüència per la comunitat científica per a extraure conclusions significatives sobre les associacions entre gens de risc i els mecanismes que produeixen les malalties. No obstant això, el maneig d'aquesta quantitat de dades que creix de forma exponencial s'ha convertit en un repte i els investigadors es veuen obligats a submergir-se en un llac de dades molt complexes que estan dispersos en mes de mil repositoris heterogenis, representats en múltiples formats i amb diferents nivells de qualitat. A més, quan es tracta de resoldre una tasca en concret només una petita part de la gran quantitat de dades disponibles és realment significativa. Aquests són els que nosaltres anomenem dades *"intel·ligents"*.

El principal objectiu d'aquesta tesi és proposar un enfocament sistemàtic per al maneig eficient de dades genòmiques intel·ligents mitjançant l'ús de tècniques de modelatge conceptual i avaluació de la qualitat de les dades. Aquest enfocament està dirigit a poblar un sistema d'informació amb dades que siguen accessibles, informatius i útils per a l'extracció de coneixement de valor.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Precision Medicine (PM) is an emerging approach for the treatment and prevention of human diseases that takes into account the genetic variability, the environment and the lifestyle for each person [1]. This approach allows doctors and researchers to predict more accurately which treatment and prevention strategies will work for a particular disease and in which groups of people.

Today, when a patient is diagnosed with a disease, he receives the same treatment as others even though it is known that different people may respond differently to the same treatment. Recent advances in science and technology have helped to understand why these differences occur, specially in cancer research. For example, it is known that the tumors have genetic characteristics involved in their growth and spread. These characteristics are different on each tumor and there are drugs that have been proven effective against cancers with specific features and uselessness in others [2]. PM helps to identify which treatments are most likely to respond according to the type of tumor, and prevents the patient from receiving those that are not likely to help. Some of the cancers which genetic characteristics have been studied are melanoma, some leukemias, breast, lung, colon and rectal cancer.

One of the pillars of PM that helps to understand the genetic aspects that make our predisposition to disease and our response to the treatment different from each other is the genetic diagnosis, which consists in the identification of potentially damaging variants in the DNA of a patient. The possibility of doing genetic diagnosis became a reality after the completion of the Human Genome Project (HGP) in April 2003 [3], that was one of the great milestones of research in history. Before the HGP, if researchers wanted to identify genes involved in a specific disease they had to select one or two candidate genes or specific locations in the DNA sequence to see how often the changes occurred in the members of the same family [4]. The processess to carry out this type of research were laborious and didn't reveal much detail. But the HGP allowed,

for the first time, to read nature's complete genetic map of a human being giving sense of the bigger picture. Since the completion of the project, thousands of human genomes have been completely sequenced and most of the research efforts have focused on improving the diagnosis and treatment of diseases, as well as providing new insights in many fields of biology, including human evolution. Furthermore, the agreements made during the project to encourage the free distribution of research data allowed scientists to share their findings with each other, as well as the public. As an example, after an international symposium on 2014, the best genome scientists from more than 20 countries have come together to improve the cooperation and coordination of genomic medicine research around the world [5].

1.1 Motivation

Despite all the efforts of the scientific community and even when new knowledge is accumulated day after day, the human genome is far from being fully understood because the management of the exponential growth of data has become a challenge for researchers. In this context, a first question to be answered is, what is the human genome and which role plays in the development of disease?

The human genome is a code represented with 4 letters, called nucleotides - A (Adenine), G (Guanine), T (Thymine) and C (Cytosine) - that contains the set of instructions required to build a human being. The complete sequence, about 3 billion nucleotides, is encoded as large linear DNA molecules within 23 chromosome pairs stored in the nucleus of each cell [6] (see Figure 1.1).

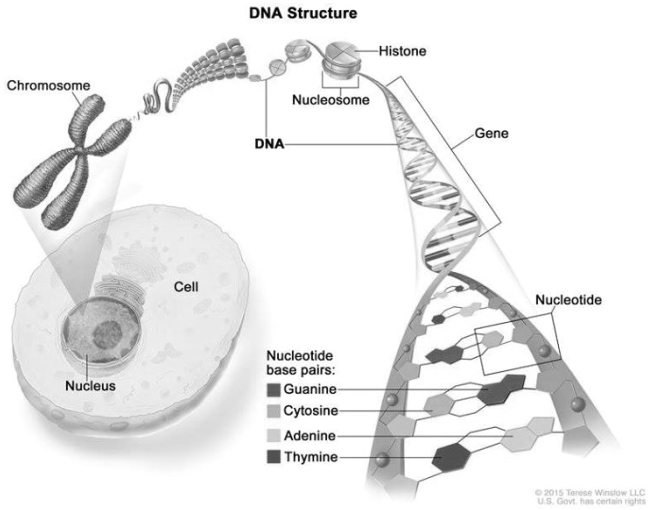

Figure 1.1: DNA, chromosome and cell [7]

The physical characteristics that make us different from each other, the influ-

ence of the environment over our health and even certain personality traits are determined by the configuration of this code. Making an analogy with Software Engineering/Information Systems, any living being is a carbon-based program (instead of a binary silicon based program) that can be executed following a model based on four letters (A, C, G and T) [8]. Nevertheless, as a software program, the human genome may contain defects that can lead to malfunction. For example, a single letter change (called Single Nucleotide Polymorphism or SNP) can be responsible for causing a genetic disease. So, how can scientists identify these small changes in such amount of information?

In order to determine which part of this code is defective, geneticists follow a process consisting in four main stages: Sample extraction, Primary analysis, Secondary analysis and Tertiary Analysis.

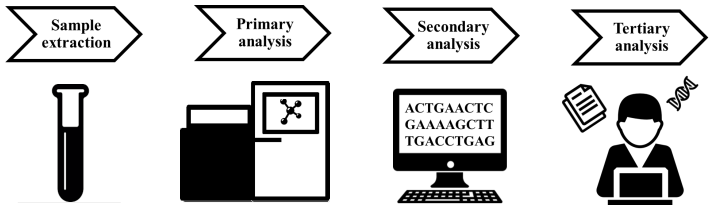

Figure 1.2: Genomic data analysis workflow

As can be seen in Figure 1.2, in the first place a sample is taken from a patient (blood, tissue, saliva, etc.). Then, the biological information is translated into digital data. This is a physical process, called *primary analysis*, that is performed by DNA sequencing machines such as Illumina and Sanger. Next, the original data is compared with a reference genome (also known as a reference assembly) that has been obtained from the DNA sequencing of a set of donors. The aim of this comparison is to determine the differences (variants) present in the DNA sample. This is called *secondary analysis* or *variant calling*. Complex algorithms and bioinformatic pipelines are used during this process. Once the variants are identified, a research must be performed in order to find out if they are liable of causing a genetic disease. Variants occur normally throughout a person's DNA, once in every 300 nucleotides on average, which means that there are roughly 10 million variants in a single human genome [9]. Most of them have no effect on health but others can act as biological markers, helping scientists locate genes that are associated with disease. The identification of relevant markers is a manual process called *variant curation* or *tertiary analysis*, that is mainly performed by searching and reading scientific literature containing relevant population studies. To accomplish this task and extract accurate results it is required to gather and manage as much information as possible.

Technological advances such as Next Generation Sequencing (NGS) and the development of Genome Wide Association Studies (GWAS) have allowed the collection of huge amounts of data in a cheaper and faster way [10]. But the

volume of available information increases in a faster pace than the ability of researchers to connect and analyze it. Additionally, the complexity of human biology makes this a slow task that requires an appropriate technological support for its achievement. It is at this point that Genomic Information Systems (GeIS) gain significance. These systems must support the efficient management of genomic data that ought to be accessible, informative and actionable enough to infer valuable knowledge. This is what we call *smart data*. To this aim, the GeIS must allow the connection of different knowledge fields such as Genomics, Proteomics, Pharmacogenomics, etc. under a structured perspective, and provide appropriate tools to analyze the data and generate new knowledge. Nevertheless, there are a number of problems, inherent to the domain, that make the design and implementation of a GeIS a challenge for experts.

1.2 Problem Statement

One of the problems that affect the biological domain is the lack of a clear ontological basis to define the key concepts of the field. This means that the same concept can be represented in different and sometimes ambiguous ways. For example, for decades in clinical practice and laboratory reporting there were incorrect assumptions about the meaning of *mutation* and *polymorphism*.

A thread published 5 years ago in the ResearchGate portal [11], produced a discusion among more than 300 participants about the difference between a mutation and a polymorphism that is still open. The term *mutation* is frequently and erroneously associated with a pathogenic impact, as well as the term *polymorphism* is erroneously associated with a benign impact [12]. Nevertheless, by definition a mutation is a permanent change in the nucleotide sequence of DNA that occurs in less than 1% of the population [13]. This definition does not imply an impact on gene structure or function. In the same way, a polymorphism is defined as a DNA variant with a frequency above 1%, that may be disease-causing. The distinction between mutation and polymorphism on the basis of their disease-causing capacity can lead to problems of classification. For instance, there is a type of anemia called sickle-cell anemia that is caused by a polymorphism which frequency in the population is >1% [14]. In this case, the disease manifests in people who have two copies of the mutated gene but the appearance of these two copies in the same individual is rare (<1%). However, having only one copy of the mutated gene is frequent (>1%) in other populations where malaria is endemic. These populations have a survival advantage against the malaria and the mutation is inherited through generations. As a consequence, a rare variant that causes disease in one population can persist in another one confering a survival advantage.

Another problem derived from the lack of an ontological commitment is related to the personalized treatments applied in precision medicine. For instance, one of the main reasons to sequence the genome in patients with cancer is the identification of DNA variants in the cancer cells that could be treated with a

personalized treatment focused on destroying these cells and causing less damage to normal cells [13]. But this requires to classify all the DNA variants of the cancer cells in order to differentiate them from the variants that are present in the rest of the cells of the organism. The consequence of a wrong classification would be a toxic effect derived from applying a wrong treatment that would affect both cancer and noncancerous cells. This is what currently happens with chemotherapy, that kill both healthy and cancer cells (see Figure 1.3).

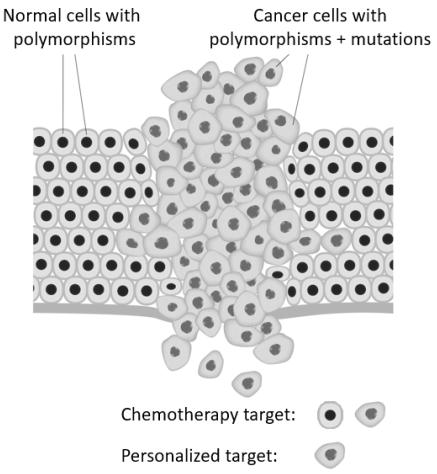

Figure 1.3: Chemotherapy vs personalized treatment.

Besides the ontological problem, there is a second one related to the dispersion of genomic information. All the knowledge is spread in over thousand heterogeneous databases with different sizes, formats and structures. Some of them are dedicated to one organism: Flybase for Drosophila [15], RMD for rice [16], GDB for Human [17], etc. Others provide information about specific parts of the genome: Uniprot for proteins [18], HGMD for genes [19], Reactome for pathways [20], etc. Each of these data repositories represents a different and complementary view of the whole genome. But the understanding of the role that plays each genomic element (genes, proteins, pathways, etc.) requires more information than is provided by one source. For example, the classification of proteins requires the integration of knowledge derived from aminoacid sequences, gene expression data and known protein-protein interactions [21]. Bringing together all these heterogenous and distributed databases can lead to interoperability issues related with semantic heterogeneity, data integrity, data representation and correctness of the interpretation of the data sets obtained from them. Consequently, the harmonization of the different sources requires to convert data into a common format and a common structure, which constitutes a bottleneck in the genomic data analysis workflow.

The third problem to face is that the information may also contain errors

caused by the complexity of biological processes, the noisy nature of experimental data and the diversity of sequencing technologies. This results in 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 [21], experiments performed over a population sample that it is not representative enough can lead to erroneous conclusions, etc. But the quality problems are not only caused by the information itself. If the genomic repositories do not have mechanisms to assure that the submitted information is correct or represented in the correct format it can lead to problems such as redundancy, that decreases the reliability both in the source and the provider. As a conclusion, a huge amount of data is ready to be used but only part of them are relevant enough to extract meaningful conclusions.

All the above mentioned problems constitute what we call the *"genomic data chaos"* associated to the fact of having a huge number of different, complex and diverse data sources where the relevant genomic data is stored in partial views (genes, transcripts, variations, proteins, pathways,etc.), where the holistic perspective is missed out, and where there are problems well-known by the data management community: lack of consistency, different formats for representing similar data, lack of conceptual standards, as well as difficulties with data heterogeneity and data interoperability management. All this leads to data analysis processes that are mainly manual, tedious and repetitive, with no explicit and systematic methods, prone to human errors, and making repetitive navigation through complex hyperlinks unavoidable. This explains why the identification and management of smart genomic data becomes such a complex task.

On this basis, the design of a proper GeIS requires in the first place a sound ontological structure to represent and connect the heterogeneous elements of the domain. Once the structure of the information is clear, the system must support its efficient management. Thus, it is essential to define a systematic process, from the selection of the appropriate data sources and the identification of relevant data, to the final load and exploitation to extract valuable knowledge. But the reliability of the results is highly dependent on the quality of the information managed. This is why the GeIS must count on mechanisms to ensure that the results obtained during the process are of enough quality.

In a previous thesis work [22], the Conceptual Schema of the Human Genome (CSHG) was developed with the aim of providing a clear and precise understanding of the human genome. This goal was achieved by unambiguously describing the relationships between the biological concepts under a holistic perspective. As a continuation of this work and using the CSHG as ontological background, this thesis proposes the definition of a method for the efficient management of smart genomic data, as basis for the development of sound Genomic Information Systems.

1.3 Objectives and Research Questions

In this section we formulate the research questions that are answered in this dissertation in order to achieve four main goals: determine the problems that hinder the management of genomic information (OB1), provide a method for the efficient management of genomic information (OB2), provide a set of quality criteria to ensure that data are reliable and correct (OB3), and validate the contribution of this research work (OB4).

Following the structure provided by Wieringa [23], the research questions are divided into two categories: *Knowledge Questions* (KQ) are asked to gather information about the world, and *Design Problems* (DP) call for the design of an artifact that will improve a problem context and contributes to answer knowledge questions.

Our research starts by obtaining information about the domain of interest in order to determine the problems that hinder the management of genomic information (OB1). This results in a set of knowledge questions:

- Where can the genomic information be found (RQ1)?
- Which problems arise when managing genomic information (RQ2)?

Once the characteristics of the problem domain have been established, our aim is to provide a method that is expected to improve the management of genomic information (OB2). This leads to a set of design problems:

- How can the most suitable genomic data sources be found (RQ3)?
- How can the relevant information be identified (RQ4)?
- How can the information be structured and stored for its further exploitation (RQ5)?

Once the solution to efficiently manage the information has been established, we need to ensure that the data are reliable and correct (OB3). At this point a new knowledge question and a new design problem arise:

- Which are the criteria that genomic information must fulfill to ensure its quality (RQ6)?
- How can the quality of genomic information be measured (RQ7)?

Finally, we need to confirm our contributions and validate the proposed solution (OB4) by answering the following research questions:

- To which extent are the results of our method accurate and valid (RQ8)?
- Do domain experts think that our method is useful to manage genomic information (RQ9)?

In the next section, we present the research methodology followed to achieve the goals and answer the research questions.

1.4 Research Methodology

This research has been developed by using the design science approach of Wieringa, defined as *the design and investigation of artifacts in context* [23].

In this thesis, we define a method (the artifact) to manage genomic information in an efficient way (the context) with the aim of populating Genomic Information Systems with high quality data.

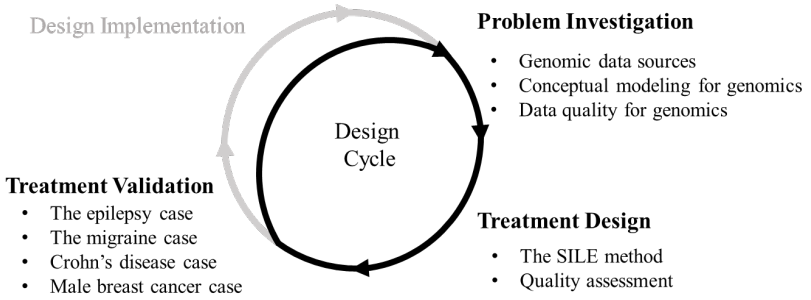

Figure 1.4: Design circle of the research methodology followed in this thesis.

The research has been outlined as an engineering cycle and at the basis of this approach is the design cycle, consisting in three phases:

- **Problem Investigation:** in this phase we describe the problem context, who the stakeholders are and how their goals can be achieved.
- **Treatment Design:** in this second phase, we provide a treatment as the introduction of an artifact in the problem contex.
- **Treatment Validation:** in the third phase, we validate the treatment in order to provide evidence that the artifact helps the stakeholders to achieve their goals.

Following the design science approach, the technological transfer associated to the real-world design and implementation is out of the scope of this PhD project. Figure 1.4 shows the key points of this thesis according to each phase of the engineering circle.

According to this approach and the objectives described in section 1.3, we start this research work performing an analysis of the state of the art in order to understand the domain and have a deep knowledge of the problem context. This analysis will help us to determine the characteristics of the genomic data sources, identify existing approaches to structure genomic information under a conceptual modeling perspective, and determine the quality of the available information. Next, we verify that there is not a current solution for the problem, and we state the treatment design. It consists in the definition of a method to efficiently manage genomic data, and the criteria required to ensure the quality

of the information. Finally, we validate that the proposed method is useful to solve the problem and achieve the goals of this thesis, using epilepsy and migraine as main case studies, as well as Crohn disease and male breast cancer as complementary case studies.

1.5 Thesis Outline

This thesis is structured in 5 chapters and 5 appendices, according to the three main phases described by the research methodology (Problem Investigation, Treatment Design and Treatment Validation):

- Chapter 2 introduces the problem investigation. In this chapter we analyze the characteristics of genomic data sources, their role as key elements to gather the scientific knowledge and their evolution in the last years. Next, we present a brief history of existing works focused on modeling genomic information and the usefulness of the Conceptual Schema of the Human Genome as basis to develop information systems that are capable of connecting this knowledge under a solid conceptual structure. Finally, we present an analysis of the genomic information available for the scientific community in order to characterize the quality problems that hinder the efficient management of the information. These problems are structured according to six major dimensions: accessibility, completeness, consistency, currency, redundancy and reliability.
- Chapter 3 presents the treatment design. In this chapter we describe the solution proposed to solve the problems described in Chapter 2, consisting in a method to efficiently manage genomic data, from the selection of the adequate data sources to the exploitation of the information to extract valuable knowledge. In addition, the criteria required to ensure the quality of the information managed during the process are defined.
- Chapter 4 focuses on the validation of the treatment in four real cases: epilepsy, migraine, Crohn disease and male breast cancer. The results of the process are reviewed by a group of experts in genetic diagnosis which will allow us to prove the validity of the proposal and the achievement of the objectives.

This thesis ends up with the conclusions and summarizes the main contributions of this work to the scientific and academic community, as well as a discussion of future lines of research (Chapter 5).

Chapter 2

Problem Investigation

The sequencing of the first human genome, took $2.7 billion and almost 15 years to be completed [24]. Since then, the advances in sequencing technologies and the development of new laboratory techniques have allowed the scientific community to sequence DNA in a cheaper and faster way. For example, the cost to generate a whole-exome sequence nowadays is generally below $1,000 and takes a few hours [25]. As a consequence, these practices have become a routine research tool, allowing the development of multitude of research projects around the world. According to [26] about 15,000 sequencing projects have been completed and over 75,000 more are on the way.

In order to gather all the generated knowledge and make it publicly available for the scientific community, a huge amount of genomic repositories have been developed. Nevertheless, what at first was a great advance to carry out new research projects turned into a handicap and has generated a first difficult question to be answered by researchers, where should I go to find the genomic data I need?

2.1 Genomic Data Sources

The majority of the online resources are specialized databases that provide not only information about DNA sequences, but also data on gene expression, macromolecular structures, gene-disease associations and genotype frequencies in diverse populations. This means that the information required to succeed in a complex research is not usually available in only one data source. Frequently, several heterogeneous databases must be queried in order to join all the puzzle pieces together, what makes the finding of the adequate repositories and the connection of the stored information a task that is not trivial at all.

Nowadays, there is no way of knowing with certainty the number of active genomic data sources, which introduces an element of uncertainty in a field

that requires the management of the most accurate and updated data. The problem is due to the fact that the information about these online resources is scattered in several life science journals and around the Web [27]. Actually, only a few web sites provide a guided access in the form of searchable links to most of them. In order to have a clue about the magnitude of the problem we head to catalogs such as the *NAR Molecular Biology Database Collection*[1], the *Online Bioinformatics Resources Collection*[2] or the one provided by the *Human Genome Variation Society*[3].

The *Nucleic Acids Research (NAR)* is an open-access and peer-reviewed scientific journal published by Oxford University Press. The journal publishes each year a special issue dedicated to categorize many of the publicly available online databases related to biology and bioinformatics. According to the issue published on January 2019 [28], the resulting database called the *Online Molecular Biology Database Collection* contains information about 1,656 repositories, divided into 15 main categories and 42 subcategories (see Figure 2.1).

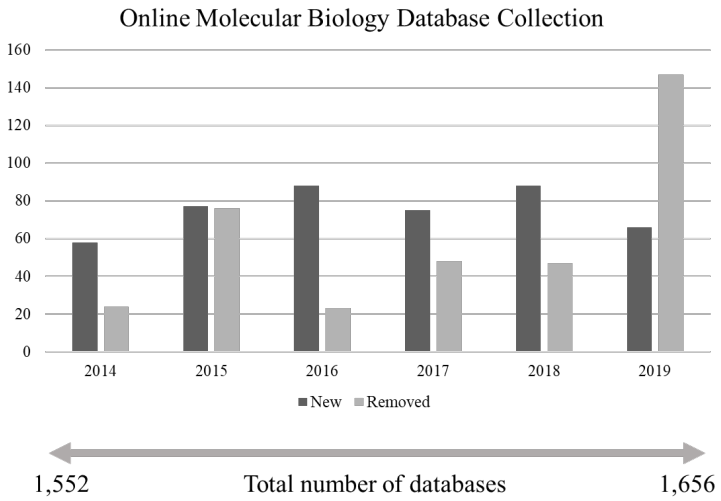

Figure 2.1: Number of databases added and removed from the Online Molecular Biology Database Collection from 2014 to 2019.

The *Online Bioinformatics Resources Collection (OBRC)* at the Health Sciences Library System (HSLS), currently includes 2,458 online bioinformatics databases divided into 13 main categories and 53 subcategories. The categories are created dynamically based on the textual information retrieved.

The *Human Genome Variation Society (HGVS)* promotes the collection, documentation and free distribution of information about genomic variants.

[1] www.oxfordjournals.org/our_journals/nar/database/c/
[2] www.hsls.pitt.edu/obrc/
[3] www.hgvs.org/content/databases-tools

Nowadays, the society maintains a collection of links to 1,750 databases.

Besides the increasing number of databases to be queried, a significant number of them can become obsolete quickly, as can be seen in Figure 2.1. The reason why this happens is due to most of these repositories are developed ad hoc for specific projects and over time they lose the technological maintenance or the information stored is no longer updated.

In spite of the collective efforts made by the scientific community to help researchers to face the growth of genomic repositories, the mentioned catalogs are mere lists of links that differ in the categorization of the content, the quality of the annotations and in their coverage. Therefore, they can be a starting point but not a feasible solution to find and connect the most appropriate sources for each circumstance.

In the next section we describe how conceptual models can provide the ontological support required to achieve this goal.

2.2 Conceptual Modeling 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 under a holistic perspective. However, genomic databases 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 as well as the quality of the conclusions derived from their analysis.

When the research community realized that this issue was becoming an important problem, some solutions were proposed. The first approach was to construct ontologies, with the aim of unifying knowledge and making it interoperable through consistent vocabularies. Examples of such a well-known type of ontologies are *Gene Ontology (GO)* [29], which describes biological processes, cellular components and molecular functions; *Sequence Ontology (SO)* [30], which describes the features and attributes of biological sequences; and *Variation Ontology (VariO)* [31], which defines the effects, mechanisms and sequences of genomic variants. But these ontologies are essentially large terminological resources that describe the terms used in the domain. Consequently, they cannot be considered as "ontologies" in the foundational and more philosophical-oriented sense associated to the term in the Information Systems domain. Neither as a concrete engineering artifact designed for a specific purpose and represented in a specific language as it is the case in the Artificial Intelligence or Semantic Web contexts [32]. As a result, for example it is not obvious how to connect these ontologies and derive a global database schema from them to the benefit of developing information systems for genomic data. For such task, the use of conceptual models proves to be a powerful tool, intended to provide an accurate representation of the relevant concepts of the domain (in ontological

terms, a representation that approximates as well as possible to what should be considered an ideal foundational ontology of the domain).

Conceptual modeling is defined as the *"activity that elicits and describes the general knowledge an Information System needs to know"* [33]. This description is called a conceptual schema and it is widely accepted that its use helps the understanding of complex domains by making a clear definition of the entities involved and the relationships among them.

The idea of applying conceptual modelling to understand the genome has been explored by some authors. It was firstly introduced in 1995 by Chen et al [34] to describe how an extended object data model can be used to capture the properties of scientific experiments. Then, Okayama et al [35] described the conceptual schema of a DNA database and Médigue et al [36] included models for representing genomic sequence data. In 2000, Paton et al [37] presented a set of data models to describe elements involved in transcriptional and translational processes, as well as the variant effects generated by them. Later on, Ram et al [38] applied conceptual modeling principles in the context of 3D protein structure and more recently, Bernasconi et al [39] proposed a conceptual model for describing metadata of experiments. In any case, 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.

The *Conceptual Schema of the Human Genome (CSHG)*, designed by the PROS Research Center at the Universitat Politècnica de València (UPV), has been developed to fill the gap and provide a unified conceptual perspective to the partial views that each database addresses in particular [40]. The 3rd version of the model maintains the essential genome information through 5 main conceptually integrated views:

- The *Structural View*, describes the structure of the genome.
- The *Transcription View*, describes the components and concepts related with the protein synthesis process.
- The *Variation View*, describes the changes regarding the sequence of reference.
- The *Pathway View*, describes information about metabolic pathways.
- The *Bibliography and Data Bank View*, describes where the data comes from.

It is due to its completeness that the CSHG has been selected to establish the ontological background required to bring some structure to the chaos introduced by the growth, size and heterogeneity of genomic databases. But the genomic data chaos problem goes further and once the structure of the information is defined another question arises: is all the available information useful to generate relevant knowledge?

2.3 Data Quality for Genomics

Due to the experimental nature of the genomic domain, the information is susceptible of containing errors or be inaccurate [41]. These errors can be propagated through the data sources, increasing the noise and forcing the researchers to make a great effort to separate the wheat from the chaff. 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. For example, some research has been done in order to determine the quality of a biological data source, such as the one made by the *Human Variome Project* [42]. This proposal encomprises a set of quality criteria divided into 4 main categories: data quality, technical quality, timeliness and accessibility. Nevertheless, it is focused only on databases that store information about DNA variants and it still has no implementation due to the difficulty of gathering all the required metadata, which needs a dedicated team for the manual collection and evaluation of part of the items.

In order to understand the issues that affect the quality of genomic information in general, we performed a study of the most common errors present in different well-known genomic data sources [43].

Table 2.1: Data Quality Dimensions

Dimension	**Definition**
Accessibility	The extent to which data is available or, easily and quickly retrievable.
Completeness	The extent to which data is not missing and all necessary values are represented.
Consistency	The extent to which data is consistent between systems and represented in the same format.
Currency	The extent to which data is sufficiently up-to-date for the task at hand.
Redundancy	The extent to which the database has redundant data or duplicated records.
Reliability	The extent to which data is regarded as true and credible.

This study allowed us to classify the issues found into six major categories called dimensions: Accessibility, Completeness, Consistency, Currency, Redundancy and Reliability. In Table 2.1 the description of each dimension is shown.

Accessibility issues. Even though most of the information stored in the genomic repositories are publicly available, there are some issues that can hin-

der the access, such as restrictions or the lack of mechanisms to automatically query and download the results of a search. Even when these mechanisms are provided, sometimes they only grant access to a view of the data which can lead to completeness issues. For example, the Ensembl database provides an tool to mine the stored information called BioMart [4]. This tool allows the user to query some predefined data sets as well as select the fields to retrieve from the database. But for example, one of the missing fields is the type of DNA variant that is very important to classify them according to the type of change that occurs in the DNA sequence.

Completeness issues. Besides the errors that can occur during the DNA sequencing, the process of marking specific features in a DNA sequence with descriptive information about its structure or its function, called sequence annotation, it is error prone too. These annotations serve as a starting point for assessing the state of the art in a particular field, or as a source for the interpretation of experimental results. But due to the extensive and ever growing amount of available information, manual annotation is a time-consuming task so tools for automated processing and analysis of text are being developed to assist researchers in evaluating the scientific literature [44]. 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) make the automated extraction of relevant biological knowledge a not trivial task. For example, the identification of relationships between proteins or the interaction between drugs and proteins is the basis to understand the metabolic pathways. To do this in practice, the automated curation tools would have to be able to mimic the human ability to infer connections from text. Despite the huge advances in these technologies, the information annotated with these tools can be incomplete, leading to problems such as the presence of missing values which affect the completeness of the databases.

Consistency issues. In order to obtain meaningful results, it is important to incorporate information from different biological repositories into the analysis. If data structures were consistent among systems, the integration from different resources would be easy. But 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 commonly accepted vocabulary, consistency problems are specially 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 recommendations) to 31 (according to the ClinVar database). Even when an ontology is used as basis, some databases map these concepts to their own 'display' terms where common usage differs from the ontology definition, as happens in the Ensembl database.

Currency issues. Another cause of low quality is related to the currency

[4] www.ensembl.org/biomart/martview/

of the information. As the underlying concepts are imperfectly defined, and scientific understanding of them is changing over time, the annotation of most genomes becomes outdated. Nevertheless, as has been shown in Figure 2.1, there are databases that do not have the required technological maintenance or do not review the information stored so they become obsolete quickly. As a consequence, most genome annotations remain static for years or have never been changed since their initial publication [45]. Another problem derived from the lack of updating processess is the presence of old identifiers that can lead to redundancy issues or errors in the analysis of the stored information. For example, the dbSNP database stores information about DNA variants and when there are multiple DNA changes referring to the same location, they are grouped into one cluster and are assigned a reference ID number, called *rs number* or *rs identifier*. If later on two cluster records are found to map to the same location (i.e. are identical), then dbSNP merges those records and consequently, the IDs become obsolete. Thus, it is crucial to take this situation into account when repeating a research after a period of time or when integrating information from different sources. If one of them has old rs numbers, the results of the integration process can be wrong.

Redundancy issues. As the information grows, the above mentioned problems converge into the increase of redundancy in the information collected. The same data can be submitted by different research groups to a database multiple times, or to different databases without cross-reference. An updated version of a record can be entered while the old version still remains. Or there may be records representing the same entity, but with different sequences or different annotations [46]. As 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. The developers of some well-known databases realized that redundancy was becoming a noteworthy problem and started to use de-duplication processes in their datasets. For example, since UniProt release 2015_04, a *Proteome Redundancy Detector* was used in order to discard entries belonging to redundant proteomes of bacterial species in the not reviewed UniProtKB (TrEMBL) set. After applying the method for the first time, 46.9 million entries were removed from the database (see Figure 2.2).

Reliability issues. The problems mentioned above lead to a decrease in 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 are the reason why the use of these repositories is not an extended practice yet, hindering the exploitation of the full potential that these databases can offer.

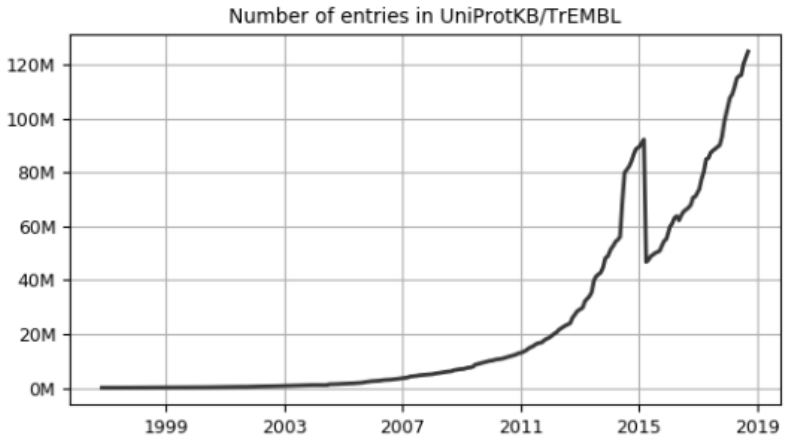

Figure 2.2: Impact of the *Proteome Redundancy Detector* in the growth of the Uniprot database.

2.4 Conclusions

Throughout this chapter, we have referred to the impact that advances in sequencing technologies and laboratory practices have in the growth of genomic data. Thus, the first research question we aimed to answer was where to go in order to find the information required to understand the human genome (RQ1).

As we have seen, thousands of heterogeneous databases are publicly available so geneticists and researchers are forced to delve into a lake of information in order to validate their experiments and extract meaningful conclusions. Furthermore, in order to understand complex systems several databases must be queried. But the finding of the adequate repositories and the connection of the stored information is not a trivial task. Most of the genomic repositories gather information about specific parts of the domain and lack of a conceptual structure to make them interoperable. On the other hand, there is a vast amount of data ready to be explored but only part of them is valuable to serve as a source for interpretation of experimental results. With this in mind, the next research question we aimed to answer was to identify the main problems to face when managing these type of data (RQ2).

After a study performed over some well-known genomic repositories, we have seen that due to the complexity of the biological domain, the specificity of the experiments and the characteristics of the databases, the genomic information is susceptible of containing errors or be inaccurate. There may be records representing the same entity (redundancy), but with different DNA sequences or different annotations (consistency problems). The limitations of the tools used

to mine the biomedical literature lead to the presence of missing values in the databases (low completeness) and the lack of maintenance causes the information to become obsolete (currency problems). All these issues decrease the reliability of the repositories and increase the effort required to find the information useful to solve innovative research hypothesis.

The aim of a Genomic Information System (GeIS) must be to solve, or at least minimize, the impact of the above mentioned problems, decreasing the risk of inconclusive and potentially invalid results. To this end, a systematic approach is required in order to reduce the effort that researchers must invest in so time-consuming tasks such as the selection of the adequate data sources, the identification of high-quality data, and the management of the information with the aim of achieving competitive advantage through its analysis.

Chapter 3

Treatment Design

In the previous chapter, we have looked at the context of this research and the challenges that must be faced when managing genomic information. With the aim of providing a feasible solution, throughout this chapter we describe the main contribution of this thesis: a method to manage genomic information in an efficient way to populate a GeIS with high-quality data. We move towards this goal by solving the following design problems:

- How can the most suitable genomic data sources be found (RQ3)?
- How can the relevant information be identified (RQ4)?
- How can the information be structured and stored for its further exploitation (RQ5)?

The solution to these design problems constitute the core of our approach, the SILE method, that is described in detail throughout section 3.1. Nevertheless, as has been mentioned in chapter 2, once the solution to efficiently manage the information has been established, we need to ensure that the data are reliable and correct. As this requires to face a set of quality challenges, we must tackle a new knowledge question and a new design problem:

- Which are the criteria that genomic information must fulfill to ensure its quality (RQ6)?
- How can the quality of genomic information be measured (RQ7)?

Throughout section 3.2, we answer the above mentioned questions by exploring the concept of *Data Quality*, adapting it to the genomic domain, and describing what is a *Data Quality Methodology* in this context. Next, we propose a quality framework that complements the SILE method to ensure that the data managed are of enough quality to extract meaningful conclusions. The goal to be achieved is to provide correct data that could be used in any GeIS-oriented software platform intended to interpret those data and generate valuable clinical

information in real PM settings (as for instance the VarSearch/GenesLove.Me platform [47] developed in the context of the PROS research center, a natural software platform to check the application of the results of this PhD work). Finally, section 3.3 contains the summary and conclusions of this research work.

3.1 The SILE Method

A GeIS must provide support to four main tasks: i) search for reliable data sources, ii) identification of relevant information, iii) adequate data storage and iv) exploitation of the information and knowledge generation. The mentioned tasks must be performed in a systematic way and due to the characteristics of the domain, it is essential to define a method with a solid ontological background.

To this end, the main contribution of our research work is the SILE method, which acronym refers to the stages that make it up (Search, Identification, Load and Exploitation). A brief description of each stage of the method can be seen in Table 3.1.

Table 3.1: Description of each stage of the SILE method

Stage	**Description**
Search	Search and selection of the adequate data sources to extract information from.
Identification	Identification of the relevant information to satisfy a knowledge requirement.
Load	Storage of the information for its further analysis and exploitation.
Exploitation	Extraction of knowledge from the database by using specific tools to analyze and interpret genomic data.

The core of this approach is the CSHG, which provides the conceptual structure required to connect all the data sources and the stored information under a holistic perspective. In the next sections, we present in detail the main purpose and activities that take place on each stage of the SILE method.

3.1.1 Search

The aim of the Search stage is to identify the genomic repositories that store the information required to succeed in fulfilling the goals of a defined task. As has been seen in Section 2.1, the vast majority of the information generated by biological research centers and biotechnological world-wide consortia

are publicly available to be used by the community: over thousand repositories of open genomic data, that help biologists and clinicians to tackle complex diseases in a multidisciplinary and individualized way. Each repository gathers information from different biological contexts that must be connected in order to infer meaningful conclusions. This has become a challenge for researchers because the existing catalogs differ in the categorization of the data sources and do not provide an intuitive way of representing the connections among them. As a result, the researchers usually search for information in a small set of familiar databases, with the consequent loss of sources that could be essential for their work. To solve this problem, the CSHG has been used to classify the genomic data sources according to their content, as well as a roadmap to identify which repositories are required.

For example, given a certain task such as the identification of DNA variants associated to a disease, the CSHG provides the context of the information that must be searched, and which type of genomic repositories must be queried. In this case, the researcher must focus on querying data sources specialized in the disease of interest (phenotype), the changes in the DNA that are associated to it (the DNA variants and their frequency of appearance in the studied population), their structural context (position in the genome), the functional consequences of each change (affected genes), and the bibliography that supports the findings (see Figure 3.1).

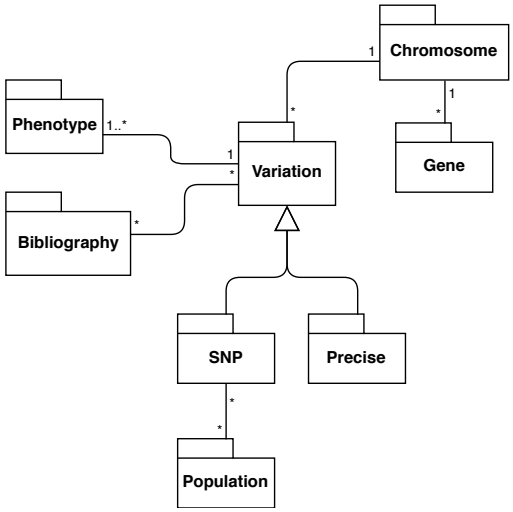

Figure 3.1: Simplified view of the CSHG required to identify variant-disease associations.

Moving through the different views of the CSHG, different tasks can be solved and as long as more details are required, additional parts of the schema can be easily added, such as proteins or pathways.

Once the context of the information to be searched is established, the next step is to identify the specific sources associated to each entity of the CSHG. To ease this task, the repositories collected by the *NAR* catalog, the *OBRC* collection and the *HGVS* society have been studied in depth and mapped to each element of the CSHG [48]. Using a graphical interface developed for this purpose, the researcher can navigate along the conceptual schema and access the databases that provide specific information about the elements of interest in three levels of depth: view, class and attribute (Figure 3.2).

Figure 3.2: Graphical tool for searching genomic databases according to the CSHG [48].

This approach helps the researcher to explore new data sources, identify which ones are the most adequate to extract information from, and to ensure that no relevant repositories have been left aside.

3.1.2 Identification

Once the most adequate repositories to be used as data sources have been selected, the next step is the identification of the relevant information to solve a defined task. To this end, we need to extract only the required information from the selected databases and integrate it under a holistic perspective. The CSHG provides the attributes required to precisely represent each entity, as well as defines how they must be connected to ease their further integration.

Each repository often provides different ways of accessing the stored data. Thus, the first step of the Identification stage is to determine the different access that each data source provides, as well as which one is the most adequate for the task at hand. For example, the ClinVar database provides three ways of accessing its content: i) direct download through the web browser, ii) access via

API and iii) download of the VCF[1] files stored in the FTP site. Each access provides a different level of coverage according to the CSHG, being the VCF files useful for exploratory analysis but not for a deep understanding of the association between the DNA variants and the disease (Figure 3.3).

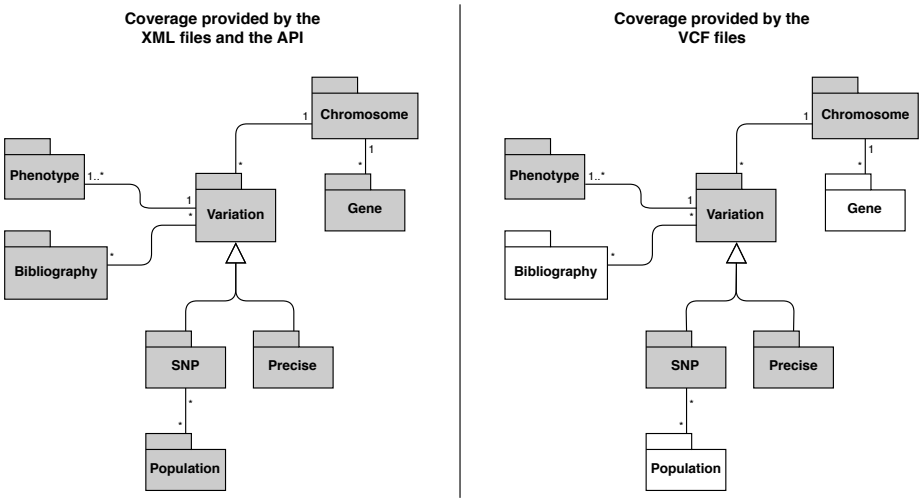

Figure 3.3: Differences in the information coverage according to the type of access to the ClinVar database. Filled boxes represent the entities for which information can be extracted. White boxes represent the entities which information is not provided.

As has been mentioned in chapter 2, each database represents the information with different format and structure, which hinders the integration process. Thus, once the way of extracting the information from each repository is determined, it needs to be harmonized into a common data model so it could be loaded into a database for its further exploitation.

To this aim, it is necessary to establish the conceptual correspondence between the information stored in the databases and the elements represented in the CSHG as a set of mapping rules. Furthermore, it is essential to identify the discrepancies in the definition of the biological concepts and the content of each field (allowed values, variant name format, ontology used, etc.) in order to describe a set of transformation specifications to represent the information under a unified perspective that ease its further integration and management. For example, the Ensembl database considers 8 types of variants while the CSHG considers 7 types. Each variant type in the source of origin must be mapped to the corresponding term in the target schema, including those that do not have

[1]Variant Call Format, a widely accepted file format to manage information about DNA variants, with the main feature of declaring the file structure in the file itself through an initial section of metadata

a direct correspondence: Substitution and Traslocation. Both discrepancies can be solved by heading to the definitions provided by Ensembl[2] and comparing them with those provided by the CSHG [22]:

- Substitution: The Ensembl database defines a substitution as *"a sequence alteration where the length of the deleted sequence is the same as the length of the inserted sequence"*. This type of variants can be considered as a special case of Indels, described by the CSHG as *"an insertion and a deletion that affects two or more nucleotides"*.
- Translocation: The Enmsebl database defines a translocation as *"A region of nucleotide sequence that has translocated to a new position"*. This type of variation can be considered as a special case of Insertion, described by the CSHG as *"an insertion of a nucleotide sequence in a certain position of the DNA"*.

In Table 3.2, the correspondences among the types of variants according to the CSHG and the Ensembl database are shown.

Table 3.2: Correspondence between the Ensembl database and the CSHG variant classification.

	CSHG						
Ensembl	SNP	CNV	Insertion	Deletion	Indel	Inversion	Imprecise
SNP	X						
Insertion			X				
Deletion				X			
Indel					X		
Substitution					X		
CNV		X					
Inversion						X	
Traslocation			X				

Both mapping rules and transformation specifications, require an extensive knowledge of the schemas used by each genomic repository and the target schema, the CSHG. This is a simple case used to illustrate the example; nevertheles, the problem incrases depending on the discrepancies between the databases.

In order to perform the steps described in this stage of the method, the raw information from the genomic repositories are stored into a temporary data

[2] www.ensembl.org/info/website/glossary.html

storage from which they are integrated and transformed following the common structure provided by the CSHG.

3.1.3 Load

The aim of the Load stage is to convert the transformed data into a queryable format that allows their further management and exploitation. To this end, a database that complies with the structure provided by the CSHG is needed, as well as the mechanisms to store the data in the target database.

The success of the load stage depends on what is going to be done with the data once it is loaded into the target database. Some uses could be:

- Perform data analysis.
- Create a tool for search and data exploration.
- Build machine learnign algorithms to infer knowledge.

Another key consideration before performing the load process is to understand the requirements of the work that is going to be performed by the target environment: the volume of data to be managed, their structure, the technology used, and the type of load required.

Additionally, this process requires the raw data to go through a data validation process, including control of duplicates and input data checks for constraints derived from the technology used in the implementation of the target database. For example, for relational databases there must be checks for orphan foreign key values (e.g. values which are present in a foreign key column but not in a primary key column).

3.1.4 Exploitation

The aim of the Exploitation stage is to extract knowledge from the information stored in the database. As has been seen in Section 1.1, one of the tasks that can be performed is to provide support to the *variant curation process* or *tertiary analysis*, key for the enhancement of Precision Medicine (PM). One of the pillars of PM is the genetic diagnosis which consists in the identification of potentially damaging variants in the DNA of a patient. To this end, a tool called VarSearch [47] has been developed by our research group.

The information about the variants present in a biological sample are stored in Variant Call Format (VCF) files, a standard widely accepted by the biological community [49]. The VCF files are processed by VarSearch in order to determine which variants within the file are also among those stored in our database.

As a result, a personalized report is generated, indicating the risk of suffering the disease (see Figure 3.4). Furthermore, VarSearch allows the researcher to go into detail on the characteristics of the variants found, and the evidence that corroborates their relationship with the disease of interest. This tool provides

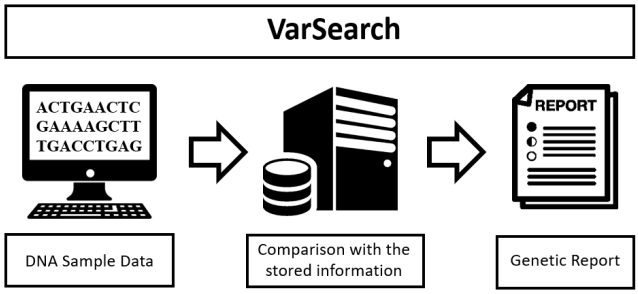

Figure 3.4: VarSearch Framework.

support to the researchers in the genetic data analysis, reducing the time and effort required to determine the variants present in a sample.

The possibility of applying SILE to identify and manage correct data for any phenotype under research conforms an ambitious research challenge, that could lead to a global Genome Database where all the relevant information for any phenotype of interest could be accessible for the adequate stakeholders. In PM terms and moving to the clinical context, this would mean to have a holistic (conceptually-speaking) repository that would include all the information required to perform a correct and valuable genomic diagnosis. Considering that the genome is not only a human feature, but the way in which life manifests in our planet, the level of this scientific challenge is amazing. The work presented in this PhD thesis provides a solution to move in this direction, with a concrete method (SILE) and a sound conceptual modeling basis. Other approaches to extract knowledge from the stored information are oriented to the development of tools to enhance data discovery, enlarge visualization, allow the performance of data analysis operations and contextualize data by augmenting it [50]. To this aim, a specific research on the use of interaction patterns are open and thus out of the scope of this thesis.

Along Section 3.1, we have described a method called SILE to solve three main design problems: How can the most suitable genomic data sources be found (RQ3), how can the relevant information be identified (RQ4) and how can the information be structured and stored for its further exploitation (RQ5).

Due to research in Genomics is under constant evolution, new data sources arise, and because of the experimental nature of the domain, the information is susceptible of containing errors or be inaccurate. Thus, two new design problems arise: Which are the criteria that genomic information must fulfill to ensure its quality (RQ6)? and how can the quality of genomic information be measured (RQ7)? In the next section we present a quality framework that enriches the SILE method and ensures that the information managed is reliable and correct.

3.2 Data Quality Assessment

Since genomic data are prone to errors and quality issues, it is important to assess their quality to achieve advantage through their analysis. Furthermore, decision making based on low genomic data quality may involve serious mistakes with important consequences when applied with clinical purposes. But, before one can address issues involved in analyzing and managing quality in the genomic domain, it is important to well understand what Data Quality (DQ) means.

DQ has been defined by Wang and Strong [19] as *"fitness for use"*, i.e. the ability of a data collection to meet users' requirements. DQ is evaluated by means of different dimensions that can be assessed by using specific metrics in order to get a quantitative measure that represents the quality of the data being managed. But to apply this knowledge properly a sound methodology needs to be defined.

Along Section 3.2.1, we review what a data quality methodology is and how it can be applied to the context of genomics. Then, in Section 3.2.2 and Section 3.2.3 we apply the methododoly to assess the quality of genomic repositories and genomic information.

3.2.1 Data Quality Methodology for Genomics

A Data Quality Methodology (DQM) can be defined as a *"set of guidelines and techniques that, starting from the input information concerning a given reality of interest, defines a rational process for using the information to measure and improve the quality of data of an organization through given phases and decision points"* [51]. To this aim we have defined a DQM specifically for the genomic domain in order to i) ensure Veracity through the selection of high-quality repositories and ii) provide Value by extracting the highest quality data from each one [52]. The proposed DQM is divided into 4 phases: Dimension Description, Metric Description, Requirements Description and DQ Assessment.

Dimension Description. Starting from a detailed description of the knowledge requirements to be solved, the first phase of the methodology consists in describing the interesting dimensions to be measured and their scope. The selection of relevant dimensions in a given scenario is mostly application-dependent. For example, while Currency is a key dimension to determine quality of a genomic database, when measuring the quality of a specific variant for genomic diagnosis we should focus on Reliability.

Metric Description. This phase consists in describing the metrics associated to each dimension. The same dimension can be used to assess quality in more than one scenario and is distinguished by the metrics defined on each case. As an example, Reliability can be measured by two different metrics depending on the context:

- Believability of a database: a metric to measure this dimension is that the database must be supported by well-known institutions and its content reviewed by experts.
- Believability of a variant: a metric to measure this dimension is that there must be at least one publication with credible statistics to support the association between the variant and the studied disease.

Requirements Description. In this phase of the DQM, the minimum levels of quality that must be fulfilled are specifically determined by assigning concrete acceptance criteria to each metric. For example, the number of publications about a gene-disease association must be at least one, the participants in a case-control study must be at least 700, etc.

DQ 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.

The most important tasks that are affected by the quality of the information when performing a tertiary analysis are i) the selection of the genomic data sources and ii) the identification of accurate information. In the next subsection, we explain how the proposed methodology can be connected to the SILE method to support the development of these two tasks.

3.2.2 Data Quality Assessment of Genomic Repositories

Genomic data sources with missing, incomplete or erroneous information hinder the processing and analysis of data. Consequently, experiments based on these sources can yield to incorrect results. Such problems lead to a loss in confidence in the underlying data sources or the provider of the data, and to a rise in effort and frustration for the researcher [41]. Thus, underlying high-quality repositories is of utmost importance.

Taking into account the characteristics of the genomic data sources and the problems described in Section 2.3, the previously defined DQM is useful to determine those repositories with the higher quality according to our requirements.

The first step of the methodology includes the selection and definition of the quality dimensions that are going to be measured. We consider a genomic repository of high-quality if it can be regarded as true and credible, sufficiently up-to-date and also has mechanisms to easily access the information. Thus, the quality dimensions that we are going to measure are: Reliability, Currency and Accessibility.

The next step is to describe the metrics to measure each dimension and the criteria that must be fulfilled. Table 3.3 shows the metrics and the quality criteria for each dimension.

Table 3.3: Metrics and criteria of acceptance for the assessment of genomic repositories

Dimension	**Metric**	**Criteria of acceptance**
Reliability	M1: Curation Process	The information must be manually curated or reviewed by an expert team.
	M2: Submission Process.	There are quality controls to ensure the correctness of the submitted data such as submission forms, automated control for HGVS expressions, etc.
Currency	M3: Database Update	The date of last update is less than 1 year.
		The link to the database is active.
Accessibility	M4: Database Availability	The information stored must be public and freely accessible.
	M5: Information Access	The database provides mechanisms to extract the information such as API, FTP repository, etc.

As metrics for measuring Reliability, we have selected two associated with the curation (addition of metadata coming from the literature) and the submission process. This tasks are key to ensure that the information stored in the database are reliable and correct. Thus, we must check if the repositories provide mechanisms to control both processes, either manually or automatically. In order to measure Currency, we have selected two metrics that check if the link to the database is active and if the information is under review. The update frequency varies depending on the database, from weeks to months. To be sure that no relevant findings are missing due to the advances of the research, we have considered that the database must be updated at least once a year. Finally, being our purpose to assist the management of information by using information systems, we prioritize those repositories that provide free access to the information as well as mechanisms to ease the automation of common tasks such as searching and downloading data.

The check of the above mentioned criteria have been added to the application developed for the Search stage of the SILE method, enriching the decision making process.

3.2.3 Data Quality Assessment of Genomic Information

Even when the genomic data sources with the highest quality have been selected, not all the information they store are reliable enough to populate an information system with the aim of supporting the genetic diagnosis process. The information retrieved from each repository must be checked in order to assess the quality and to avoid or minimize the consequences of using poor quality data. To this aim, we apply the DQM to determine the criteria that the information must

fulfill during the Identification stage of the SILE method.

We consider the genomic information as high-quality if at least there is enough evidence to support the assertions, the information is standardized and there is no conflicts among the different repositories. To this aim, the quality dimensions that are going to be measured are: Consistency and Reliability.

Table 3.4: Metrics and criteria of acceptance for the assessment of genomic information.

Dimension	**Metric**	**Criteria of acceptance**
Consistency	M1: Absence of conflicts	There must not be conflicts among databases in the clinical interpretation of each variant.
		There must not be conflicts among databases related to the structural characteristics of a variant.
Reliability	M2: Assertion Reliability	The relationship between the variant and the disease must be associated to at least one published, peer-reviewed paper with free access.
	M3: Assertion Validity	The reported consequences of a variant must have been independently replicated by at least one group besides the first group reporting the finding.
	M4: Statistical Relevance	The published studies must have at least 700 participants and be replicated.
		For pathogenic variants, the Odds Ratio must be greater than 1.
		For protective variants the Odds Ratio must be less than 1.
		For Genome Wide Association Studies (GWAS), the p-value must be less than $5x10^{-8}$.

The next step is to describe the metrics and the criteria of acceptance as has been done for genomic repositories (see Table 3.4). In order to measure Consistency, we have selected a set of metrics to check the conflicts that could arise when integrating the information from different repositories. The most common problems are i) discrepancies in the interpretation of the clinical significance of a DNA variant due to the use of different assertion protocols, and ii) discrepancies in the structural information such as location in the chromosome due to the use of different reference sequences. Both issues must be carefully considered because they can lead to erroneous results when searching for clinically important variants.

The reliability of the relationship between the DNA variant and the disease is measured according to the associated literature that corroborates the assertion. To this aim, it is important to identify which type of studies can be performed and which ones are useful for the task at hand.

3.2.3.1 Types of Genetic Studies

There are different types of studies that can be performed to identify the genetic causes of a disease. Each type of study has different purposes, as well as advantages and disadvantages that must be carefully considered in order to determine its relevance depending on the task to be performed. The genetic studies can be classified as:

- **Linkage studies:** Genetic linkage is based on the tendency of DNA sequences that are physically close on a chromosome to be inherited together. These studies are a powerful tool to identify the genes of interest for the development of the disease through the study of different biological markers that appear in families with several affected members. It is common to perform these studies over large families, but even when they are useful to identify interesting regions in the human genome, they have a low resolution [53].

- **Association studies:** The aim of these studies is to identify with precision the DNA variants in the genes that are responsible of causing the disease in a population. The association studies require to sequence a concrete part of the DNA in a group of affected and healthy people. Then, the frequencies of appearance of the different variants are compared to identify those that are more frequent in the affected group. The limitation of these studies is that they require to previously know the candidate genes to study [54]. There are two types of association studies:

  - **Genome Wide Association Studies (GWAS):** These are a particular type of association studies where all the variants present in the genome are sequenced, increasing the probability of finding the genes that cause the disease. These studies do not require to previously know the candidate gene because the entire genome is sequenced, but they have a higher probability of having false positives. To minimize this effect, it is required to have a huge number of participants in the study and sctrict statistical thresholds [55].

  - **Cohort studies:** A cohort is a subset of a population that may or may not be exposed to factors that can influence the probability of the occurrence of a particular disease. The cohorts are studied to determine distinguishing characteristics, frequently associated to the influence of the environment, drugs or specific medical treatments [56].

The relevance of each type o study depends on the task to be performed. For example, for exploratory analysis about the influence of genetics in the development of a disease linkage studies are the right ones. But, if we already know that the disease has a genetic basis and we need to specifically identify the variants that causes it, we must focus on association studies. This reduces the amount of information to be managed.

Once the type of studies to consider are established, we must determine if they are statistically relevant to support the assertion about the association between the variant and the disease.

3.2.3.2 Statistical Relevance of Genetic Studies

Among the quality criteria established to measure the statistical evidence of a study, the most widely used are the Odds Ratio (OR) and the sample size. A sample size with sufficient statistical power is critical to the success of genetic association studies to detect causal genes of human complex diseases. A large sample size is more representative of the population, limiting the influence of outliers or extreme observations [57]. A sufficiently large sample size is also necessary to produce results among variables that are significantly different. Following the practices of genetic diagnosis providers such as 23&Me or Promethease, and the recommendations of the stakeholders, we established the minimum sample size for genome association studies in 700 participants.

Odds ratios are used to compare the relative odds of the occurrence of the outcome of interest (e.g. disease or disorder), given exposure to the variable of interest (e.g. health characteristic, aspect of medical history) [58]. The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome:

- OR=1 Exposure does not affect odds of outcome
- OR>1 Exposure associated with higher odds of outcome
- OR<1 Exposure associated with lower odds of outcome

The Interval of Confidence (IC) is used in conjunction with the OR to determine the probability that the value falls within two values. To assure that the OR is reliable, the lower or the upper limit of the IC must be consistent with the calculated odds of the outcome (lower limit >1 for higher risk and upper limit <1 for lower risk). The p-value is critical to control the number of false-positive associations and the threshold of $5x10^{-8}$ has become a standard for genome-wide association studies (GWAS) [59]. For the rest of case studies, the p-value must be <0.0001.

The application of the above mentioned criteria ensures that the information stored in the database during the Load stage of the SILE method is of enough quality to extract reliable and meaningful conclusions in the Exploitation stage.

3.3 Conclusions

Throughout this chapter, we have described a method to manage genomic information in an efficient way, with the aim of populating a GeIS with high-quality

data. To accomplish this task we have tackled 3 design problems:

- How can the most suitable genomic data sources be found (RQ3)?
- How can the relevant information be identified (RQ4)?
- How can the information be structured and stored for its further exploitation (RQ5)?

The SILE method comprises 4 systematic stages (Search, Identification, Load and Exploitation) and uses the CSHG as the conceptual structure required to harmonize the information under a holistic perspective. The aim of the Search stage is to identify the genetic repositories that store the information required to succeed in fulfilling the goals of a defined task. During the Identification stage, the information from the different sources is harmonized into a common data model so it could be loaded into a database for its further exploitation. It requires to describe in detail the different options to access the repositories and the set of transformations to represent the information according to the structure of the CSHG. The aim of the Load stage is to convert the raw data coming from the genomic repositories into a queryable format and ensure the persistency of the information. Finally, the aim of the Exploitation stage is to extract knowledge from the information stored in the database. As an example, a tool called VarSearch has been developed to provide support to the genomic diagnosis by identifying potentially damaging variants in the DNA of a patient, according to the information stored in the database.

The genomic domain is under constant evolution so the repositories and the stored information can become obsolete quickly. Furthermore, genomic information is susceptible of containing errors or be inaccurate. Thus, two new design problems arise:

- Which are the criteria that genomic information must fulfill to ensure its quality (RQ6)?
- How can the quality of genomic information be measured (RQ7)?

We have addressed these problems by defining a DQM, specially designed for genomic information. The aim of the methodology is to ensure Veracity through the selection of high-quality repositories, and provide Value by extracting the highest quality data from each one. The DQM has been connected with the Search and Identification stages of the SILE method. This ensures that the information stored in the database during the Load stage and the results derived from its analysis during the Exploitation stage can be regarded as reliable and valuable.

Chapter 4

Treatment Validation

In chapter 3, a method to manage genomic data in an efficient way was introduced and in this chapter we discuss its validity in terms of the contribution to fulfill the stakeholders goals. To this aim, we work toward answering the following research questions:

- To which extent are the results of our method accurate and valid? (RQ8).
- Do domain experts think that our method is useful to manage genomic information? (RQ9).

In order to validate the accuracy of the method and its usefulness, we have been working with a company specialized in generating personal genome reports. In this real, industrial working context we head to two experts in genetic diagnosis (a geneticist and a clinician) that were responsible of the corresponding teams in charge of finding and selecting relevant DNA variants, and we performed four different case studies: two initial cases and two complementary cases. The aim of the first initial case study was to compare the results obtained by using the method with the results obtained by the group of experts in the search of the genetic causes of an already studied disease (migraine), in order to validate the accuracy of the results. The aim of the second initial case study was to use the method to search for the genetic causes of a disease (epilepsy), currently under study and particularly difficult because of the complexity of the disease context, so we can measure the usefulness of the method to guide the research process. After performing these two initial case studies, two more were faced (Crohn's disease and male breast cancer) in order to reinforce the idea of generalization of the SILE method to any kind of phenotype. Our final goal was to provide a solution that could be used to systematize the management of relevant variants for a full-coverage genomic diagnosis.

In order to test the method for these four case studies, we developed the GeIS that would support the different stages of the process. The details about

the development of the GeIS are described in section 4.1. Along sections 4.2, 4.3 and 4.4, we describe the execution of the method for the case studies and detail the results. Finally, we draw conclusions in section 4.5.

4.1 Building a Genomic Information System

The task to be performed for the validation is the identification of the DNA variants related with the risk of suffering four different diseases: migraine, epilepsy, Crohn and male breast cancer.

The first step to build the GeIS that supports the process was to define the ontological background required to represent the information under a holistic perspective. Using the CSHG as basis, we have determined the view that would provide support to the process (see Figure 4.1).

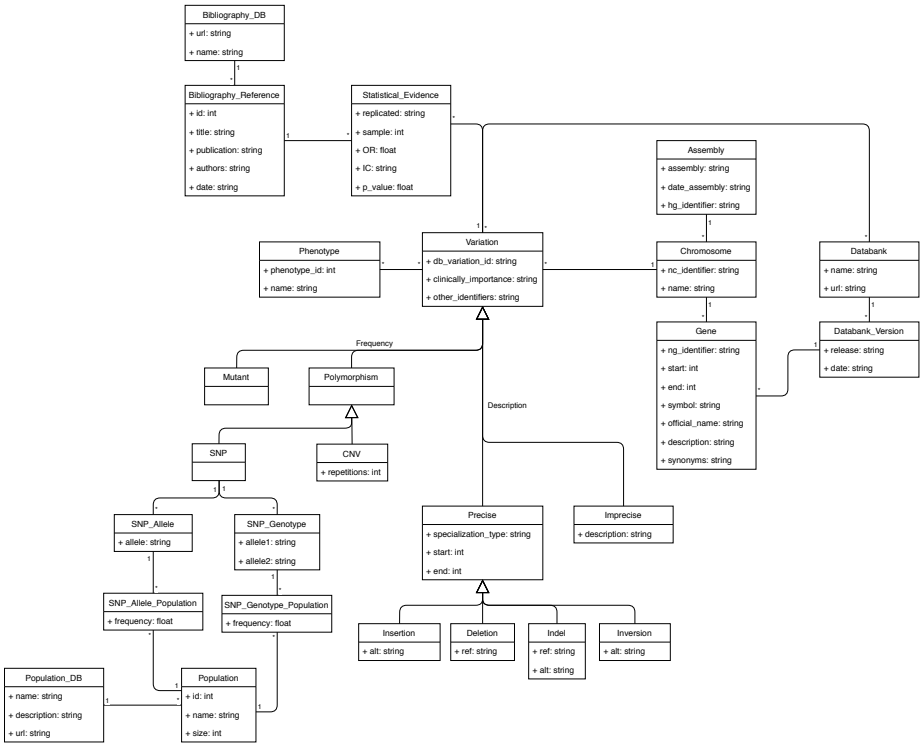

Figure 4.1: Conceptual schema used to design the GeIS.

The main entities of the conceptual schema are *Variation* and *Gene*. The *Variation*[1] entity represents the changes in the DNA that are considered to be

[1]The terms Variation and Variant are frequently used as synonyms by the genomic repositories. Along this thesis we have used the term Variant due to its increasing use. Nevertheless,

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 notion of SNP is associated to its frequency of appearance in a certain population and the amount of information about the combination of alleles that the individual carries, known as genotype. According to the frequency of appearance the variants can be: *Mutations*, if the frequency is lower than 1%, and *Single Nucleotide Polymorphisms (SNP)* or *Copy Number Variations (CNV)* if the frequency is greater or equal to 1%. A SNP is a change that affects a short region of the genome, usually a single letter change. A CNV is a phenomenon in which a section of the genome is repeated [60]. The number of repeats varies between individuals and affects a considerable number of nucleotides. According to the precision of the information there are *Precise* variants, which change and position in the DNA sequence are known, and *Imprecise* variants, which position and change are not known. Precise variants can be *Insertions*, *Deletions*, *Indels* and *Inversions*. The association of a variant with the development of a disease must be supported by the literature and the statistical evidence.

The *Gene* entity represents the elements which alteration derives in a malfunction that leads to the development of the disease. In addition, the GeIS must store the information associated to the databases where the information has been extracted from to ensure the traceability of the information and help to keep the information updated. A complete description of each attribute of the conceptual schema can be seen in Annex A.

Once the conceptual schema was determined, we started the development of the GeIS to support the four stages of the SILE method: Search, Identification, Load and Exploitation.

4.1.1 Searching for Genomic Data Sources

The first stage of the SILE method requires the support to the search of the most adequate data sources to solve the task at hand. To this end, we used the tool described in Section 3.1.1 in order to extract an initial set of useful data sources. Doing this, the noise produced by the great amount of available repositories is reduced, and this step helped us to focus on selecting those that are reliable and useful enough.

As genomic repositories cover a wide range of species and experimental data, we performed a search using the conceptual schema as a guide for the search. We focused on databases that store genomic information about humans, germline variants (those that are inherited from parent to child) and databases that do not result of gathering the information from other sources (e.g. GenesCards comprises information from other sources such as HGNC and Entrez Gene so we prefer to head to the original sources). At the end of the search, we identified

The CSHG was built considering the term Variation. As there is not consensus or formal recommendations about it, we will use both terms as synonyms from now on.

9 repositories that can be used to extract the information: PubMed, NCBI Assembly, GWAS Catalog, ClinVar, Ensembl, dbSNP, HGNC, Entrez Gene and 1000 Genomes. A brief description of their main characteristics is introduced below.

PubMed[2] is a resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM). The data source comprises over 29 million citations and abstracts for biomedical literature, life science journals, as well as behavioral sciences, chemical sciences, and bioengineering [61].

NCBI Assembly[3] is a database that provides information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data [62].

GWAS Catalog[4] is a publicly available, manually curated resource of all published genome wide association studies, collaboratively produced and developed by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI). They curate eligible studies within 1-2 months of publication, dependent on the availability of literature, and the data is released on a weekly cycle.

ClinVar[5] is a public archive of relationships among human variants and phenotypes, with supporting evidence [63]. The database has been developed by the NCBI.

The Ensembl Variation[6] database stores information about variants and, where available, associated disease and phenotype information. The database has been developed as a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute and is updated approximately five times each year with new genome assemblies and additional data as it becomes available [64].

dbSNP[7] is a repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms. The database has been developed in collaboration with the NHGRI and the NCBI [65].

The HUGO Gene Nomenclature Committee[8] database is a curated online repository that stores information about approved gene nomenclature, gene groups and associated resources including links to genomic, proteomic and phenotypic information.

Entrez Gene[9] is a NCBI's database for gene-specific information. The database focuses on the genomes that have been completely sequenced, that

[2] www.ncbi.nlm.nih.gov/pubmed/
[3] www.ncbi.nlm.nih.gov/assembly/
[4] www.ebi.ac.uk/gwas/home
[5] www.ncbi.nlm.nih.gov/clinvar/
[6] www.ensembl.org/index.html
[7] https://www.ncbi.nlm.nih.gov/projects/SNP/
[8] www.genenames.org/
[9] www.ncbi.nlm.nih.gov/gene/

have an active research community or that are scheduled for intense sequence analysis [66].

The 1000 Genomes Project[10] is the largest public catalogue of human variant and genotype data. The goal of the 1000 Genomes Project is to provide a resource of almost all variants, including polymorphisms and structural variants, as well as their haplotype contexts [67].

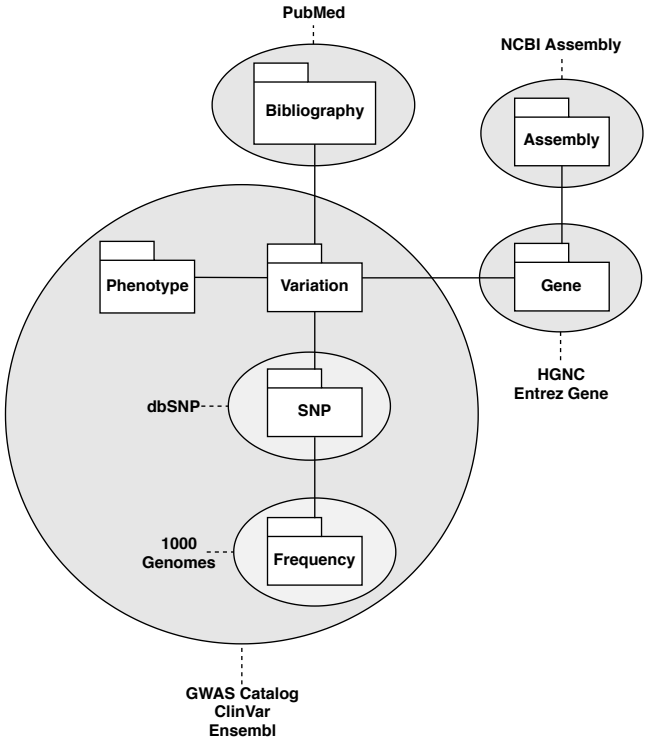

Figure 4.2: Data sources coverage according to the CSHG.

As can be seen in Figure 4.2, the initial set of genomic repositories covers different parts of the conceptual schema. The structural part is covered by HGNC, Entrez Gene and Assembly, that provide information about the location of the variants in the DNA sequence. The bibliography that supports the assertion among the variants and the studied disease is provided by PubMed. The specific information about the variants is covered by the GWAS Catalog, ClinVar and Ensembl. More details are provided by dbSNP, which focuses on a specific type of variants (polymorphisms), and 1000 Genomes that provides information about their frequency of appearance in diverse populations.

In addition, the databases conform the quality criteria established in section

[10] www.internationalgenome.org/

3.2.2: the stored information are manually curated by experts, the submission process has enough quality controls to ensure the correctness, the databases are active and updated as well as they provide mechanisms to extract the information.

Besides the databases mentioned above, we found one repository specific for genomic variants associated with the epilepsy disease: The Lafora Database. Nevertheless, this database is not currently maintained so the information is not reviewed or updated, which does not conform with our quality criteria. Therefore, the Lafora Database has not been selected as a valid source.

Considering the dynamism of the selection process that we are describing, the use of this approach and the search tool developed to assist the process, the information system can be easily extended by adding new sources at any moment, as long as they meet the quality requirements.

4.1.2 Identifying Relevant Information

The second stage of the SILE method consists in the identification of the relevant information from each data source. To this aim, the GeIS must support the extraction of the information from the repositories, the integration following the structure provided by the conceptual schema, and its analysis according to the quality criteria established in Section 3.2.3. The core of this part of the GeIS follows an Extraction-Transformation-Load architecture as can be seen in Figure 4.3.

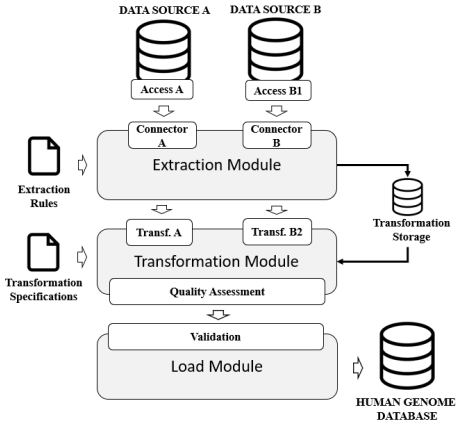

Figure 4.3: Architecture to support the SILE method.

4.1.2.1 The Extraction Module

The first step to extract the information from each data source is to determine the mechanisms that they provide to access the stored data, and select the most adequate according to our needs. In our case, we have selected as first choice

the mechanisms that allow a high level of automation, such as REST services. In Table 4.1, the names of the REST services selected to access each data source are described.

Table 4.1: Types of access used to extract the information from the databases.

Data Source	**Access Name**	**Access Type**
PubMed	E-Utilities	RESTful
NCBI Assembly	E-Utilities	RESTful
GWAS Catalog	GWAS API	RESTful
ClinVar	E-Utilities	RESTful
Ensembl	Biomart	RESTful
dbSNP	E-Utilities	RESTful
HGNC	genenames.org	RESTful
Entrez Gene	E-Utilities	RESTful
1000 Genomes	Biomart	RESTful

The REST service used to access the databases provided by the NCBI (PubMed, Assembly, ClinVar, dbSNP and Genes) is called E-utilities. It is a set of server-side programs that provide a stable interface to search for and retrieve data from 38 databases, covering a variety of biomedical data [68]. Another REST service used is Biomart, a data-mining tool that provides access to the databases that the Ensembl project comprises (Ensembl Genes, Ensembl Variation, Ensembl Regulation and Vega) [69]. In our case, Biomart is used to extract information from Ensembl Variation, that includes information from 1000 Genomes too. HGNC provides a way of searching and fetching data from the database within a script/program called genenames.org. Using this REST web-service we can get the information from the database either in XML and JSON format [70]. Finally, the information from the GWAS Catalog is extracted through the REST service in Hypertex Application Language (HAL) format.

Once the mechanisms to access the repositories are established, the next step is to implement the connectors that will extract the data. To this end we need to define a set of mapping rules. For illustration purposes, in Table 4.2 we provide some of the mappings related with the *Variation* entity.

Each mapping rule is a logic formula with variables in its left end side that are computed from the variables in its right end side. As an example, the entity *Variation* of the CSHG is filled with data from the GWAS Catalog database, together with data from ClinVar, Ensembl and dbSNP. According to the CSHG, a variation is represented by three attributes (db_variation_id, clin-

Table 4.2: Examples of mapping rules.

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)

ically_importance and other_identifiers). The attribute *db_variation_id* can be extracted from GWAS, ClinVar, Ensembl and dbSNP. Both GWAS and ClinVar require the transformation of the raw data to meet the required format. 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. All the required mapping rules to build the GeIS are described in Annex B.

The connection modules to the selected databases and the information extraction procedures have been implemented with R (using the Rentrez and BiomaRt libraries) as well as Python scripts. The reason why these programming languages have been selected is that they are widely used by the Bioinformatics Community and many libraries and documentation can be found on the web. To help the user to query and extract the information from each data source a graphical interface has been built (see Figure 4.4).

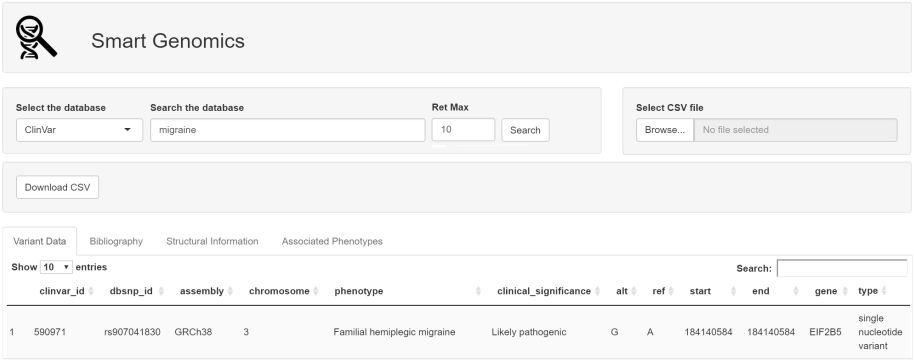

Figure 4.4: Graphical User Interface for data extraction.

The user can selet the database to be queried in the drop down menu on the left, and specify a key term for the search (in this case a phenotype). He can also specified the number of records to be retrieved from the repository. The results of the search are grouped in 4 different tabs:

- Variant Data: Contains information about the DNA variants retrieved.
- Bibliography: Contains informacion about the associated bibliography.
- Structural informacion: Contains information about the location of the

variants in the genome.

- Associated phenotypes: Contains information about all the phenotypes associated to the variants.

The information retrieved can be downloaded into a CSV file specified by the user. In addition, the user can upload and view any CSV file that has been previously created using the *"Select CSV file"* dialog on the right side.

The extracted data are stored into a temporary relational database from which it is transformed and checked.

4.1.2.2 The Transformation Module

The Transformation module is responsible of transforming and integrating the data into a common structure. Some of the attributes are acquired exactly as they are in the original source, others need the application of transformations (denoted as *tr* in the mapping specifications), and others are computed as a combination of multiple source fields (denoted as *comb* in the mapping specifications). These transformations are specified in a set of transformation rules. As an example, the format of the attribute Variation.*db_variation_id* comprises the *rs* prefix followed by an integer number. As the variant identifier coming from the GWAS Catalog is represented only by the integer number, it requires the addition of the *rs* prefix. The set of all the transformation rules required to transform the raw data into the correct format is summarized in Table 4.3.

Table 4.3: Transformation Rules

Database	**Name**	**Action**
PubMed	tr(PubDate)	Extract Year
GWAS	tr(REPLICATION SAMPLE SIZE)	NA = NO: YES
GWAS	tr(STRONGEST SNP-RISK ALLELE)	Extract risk allele
GWAS	tr(SNP_ID_CURRENT)	Add prefix: 'rs'
ClinVar	tr(db_id)	Add prefix: 'rs'
ClinVar	tr(variant_type)	See Annex C
1000 Genomes	tr(population)	Extract population name
dbSNP	tr(SNP_ID)	Add prefix: 'rs'
dbSNP	tr(DOCSUM)	Extract NG Identifier
dbSNP	tr(DOCSUM)	Extract NC Identifier
dbSNP	tr(CHRPOS)	Extract Position

As we already discussed in Section 3.1.2, it is also required to define a shared

set of homogenized values for those attributes that use different ontologies to represent the stored values. In this case, the type of the DNA variants which correspondences with the conceptual schema can be seen in Annex C.

The integration and transformation processes require following a number of consecutive steps:

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

The transformation rules for each data source and the integration process have been implemented using Pentaho's Data Integration[11] (Kettle), because it is a flexible and complete ETL tool, freely accessible. In Figure 4.5, a simplified view of the Kettle transformation and integration entities can be seen.

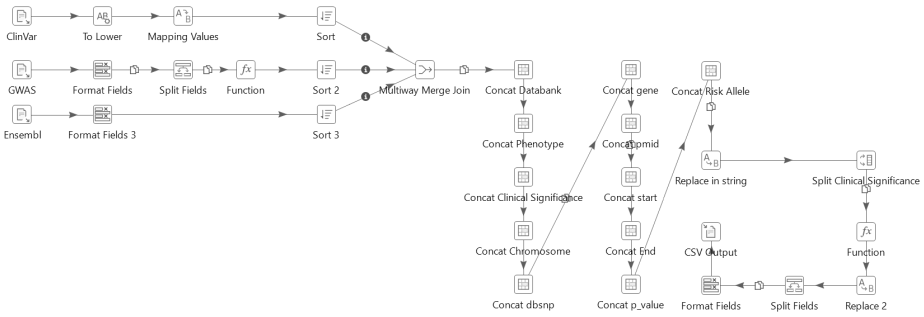

Figure 4.5: Simplified view of the transformation and integration processes performed with Kettle.

The first part of the pipeline takes the information from the different repositories (ClinVar, Ensembl and GWAS) and prepares the data for the integration, harmonizing and formating the different columns of each data source. Then, the transformed data is sorted by variant identifier and combined into a common dataset. During the final stage of the pipeline, the values of the common columns are merged and the results are stored in a CSV file.

Once the transformation and the integration processes are finished, the data are ready to perform the last step in the identification of relevant information:

[11] https://community.hitachivantara.com/docs/DOC-1009855-data-integration-kettle

the quality assessment. As has been explained in Section 3.2, not all the information coming from the repositories are reliable enough to perform a genetic diagnosis so it is important to measure its quality in order to ensure that only the information with the highest quality is considered. To this aim we have implemented the quality criteria explained in section 3.2.3 as a new set of Kettle transformations.

In order to optimize the quality assessment, the different criteria are applied following the workflow described in Figure 4.6.

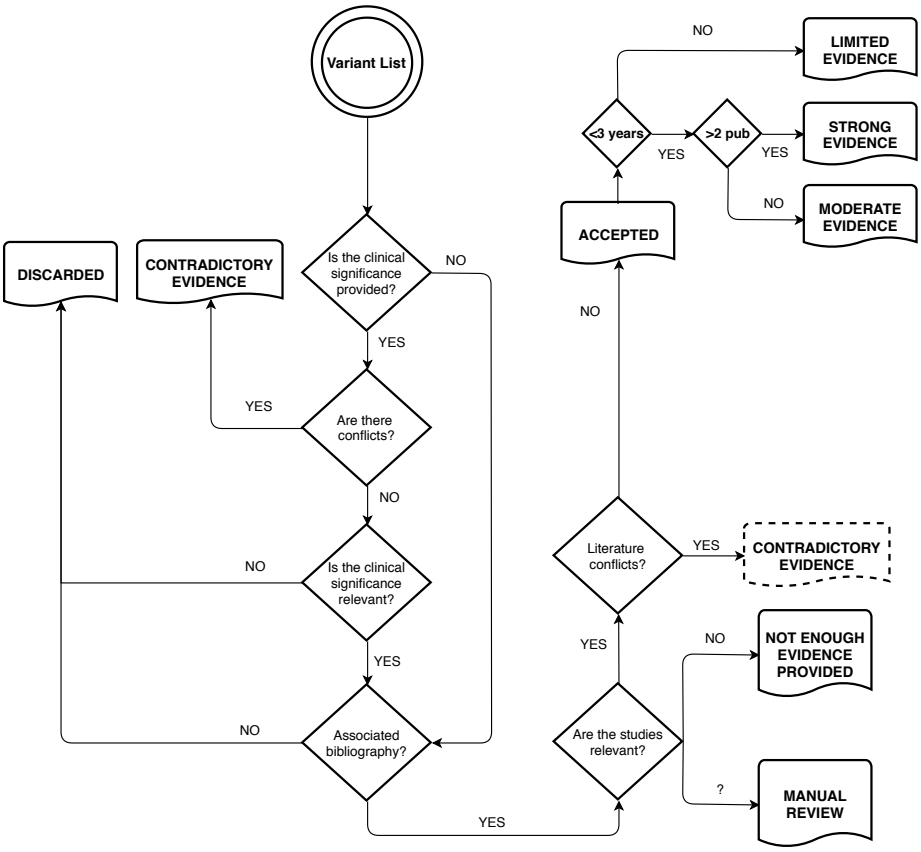

Figure 4.6: Data Quality Workflow.

One of the quality criteria specified in the workflow requires to determine the relevancy of the associated bibliography. As has been seen in Section 3.2.3.1, the first step to determine if a study is relevant is to identify its type. This is an information that the databases do not provide so, a relevant question arises: how can be the bibliography classified according to its type?

There are two ways to identify the articles according to its type: i) with the

keywords defined by the authors and ii) using the MeSH terms associated to each study. MeSH (Medical Subject Headings) is a thesaurus of controlled vocabulary provided by the National Library of Medicine [71]. The terms provided by this thesaurus are used to describe the content of each article and for indexing them in the PubMed database. The MeSH terms required to classify the bibliography associated to the DNA variants are described in Table 4.4.

Table 4.4: MeSH terms used to classify the bibliography by type.

Study type	**MeSH terms (Identifier)**
Linkage Study	Genetic Linkage (D008040)
	Linkage Disequilibrium (D015810)
Association Studies	Case-Control Studies (D016022)
	Genetic Association Studies (D056726)
GWAS Studies	Genome-Wide Association Study (D055106)
Cohort Studies	Cohort Studies (D015331)

Despite the wide use of the MeSH terms, not all the articles are fully described and sometimes it is required to check the keywords provided by the author or the abstract to complete the information. Figure 4.7 shows the workflow followed to classify the bibliography according to the different study types. One or more labels will be assigned to each study according to the conditions established for each type:

- In the case of the GWAS studies, it will first be determined if they come from a GWAS source such as GWAS Catalog. If they do not, the MeSH term *Genome-Wide Association Study* will be searched. In case the term does not appear, the title, abstract and associated keywords will be reviewed in search of one of the following terms: *genome-wide, genome wide, whole-genome, whole genome* and *genomewide*. If none of them indicates that it is a GWAS study, the category OTHER will be considered.

- For association studies, the MeSH terms *Case-Control Studies* and *Genetic Association Studies* will be searched. If they do not appear, the term *case-control* will be searched in the abstract. If none of them indicates that the study is of type ASSOCIATION, the category OTHER will be considered.

- For cohort studies, the MeSH term *Cohort Studies* will be searched. If it does not appear, the category OTHER will be considered.

- For linkage studies, the MeSH terms *Genetic Linkage* and *Linkage Disequilibrium* will be searched. It they do not appear, the category OTHER will be considered.

- If the article under review cannot be classified as any of the above mentioned types, it will be labelled as OTHER.

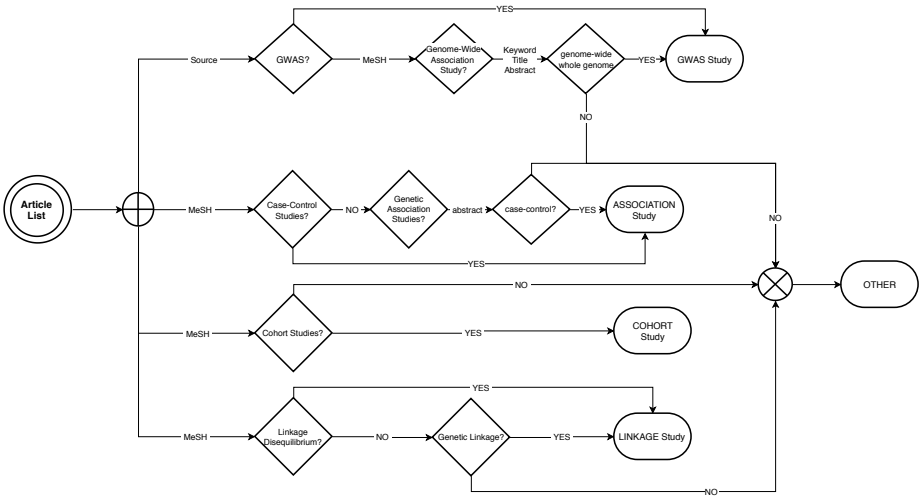

Figure 4.7: Workflow to classify bibliography according to the type of study.

Only those articles classified as Cohort Studies, Association Studies or GWAS Studies are going to be considered as relevant to provide a genetic diagnosis. Once these articles are identified, their statistical relevance can be measured as well as the appearance of contradictory evidence.

Along the quality assessment, the DNA variants are classified in four main groups:

- **Variants Discarded:** This group includes the variants which clinical significance is not relevant for the task at hand (e.g. benign, likely benign, uncertain, etc.). It also includes those variants without bibliography to support the association with the disease of interest.
- **Variants with contradictory evidence:** This group includes the variants which associated studies are relevant but contain contradictory evidence (e.g. one study considers the variant as pathogenic while other considers the variant as benign), as well as those which clinical significance has been proved to be contradictory in the ClinVar database.
- **Variants without enough evidence provided:** This group includes the variants which associated studies are not replicated or not statistically significant (i.e. not enough number of participants and OR, IC and p_value out of the minimum accepted requirements).
- **Variants Accepted**: This group includes the variants that have fulfilled all the quality criteria and thus can be considered useful to provide a genetic diagnosis.

Finally, the accepted variants are classified according to the level of the evidence provided. To this end, we check the number of associated studies and the date on which the last study has been performed. If the variant has more than two studies and less than three years have elapsed since the publication of the last one, the evidence can be considered as *Strong*. Three years is considered by experts time enough for new studies that contradict the evidence to appear. If there is only one study but less than three years have elapsed since its publication, the evidence can be considered as *Moderate*. In any other case, the evidence is considered as *Limited*. These criteria are based on the recommendations provided by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology [72].

At the end of the Identification stage, the information extracted from the genomic repositories has been harmonized into a common data model and checked in order to determine its validity. Therefore,it is prepared to be stored adequately to assure its persistency.

4.1.3 Information Storage

With the aim of preparing the information for its further exploitation, the GeIS must support its storage in a target database that conforms the structure provided by the conceptual schema. To this aim, we have developed the *Human Genome Database (HGDB)* which structure can be seen in Figure 4.8.

The HGDB is a relational database. We used this technology because it is well-known and widely accepted, 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. 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. We are aware that new technologies such as NoSQL databases have arisen and consequently, a new research line is under development within our research group with the aim of exploring these alternative database representations and compare how efficient and effective they are. Even though in this PhD work we report a relational-based implementation, the conceptual model-based strategy that we use warranties a semantic independence between the conceptual level and the logical level. This implies that changing the database model would preserve the conceptual schema, and only the data transformations should be adapted to any new, selected database model.

The storage of the information into the HGDB is performed by the Load Module and requires to carry out the needed checks derived from the technology used to build the database. Due to the HGDB is a relational database we must avoid the unsuccessful load due to constraint violations or wrong data types. To this aim, the checks to be performed are:

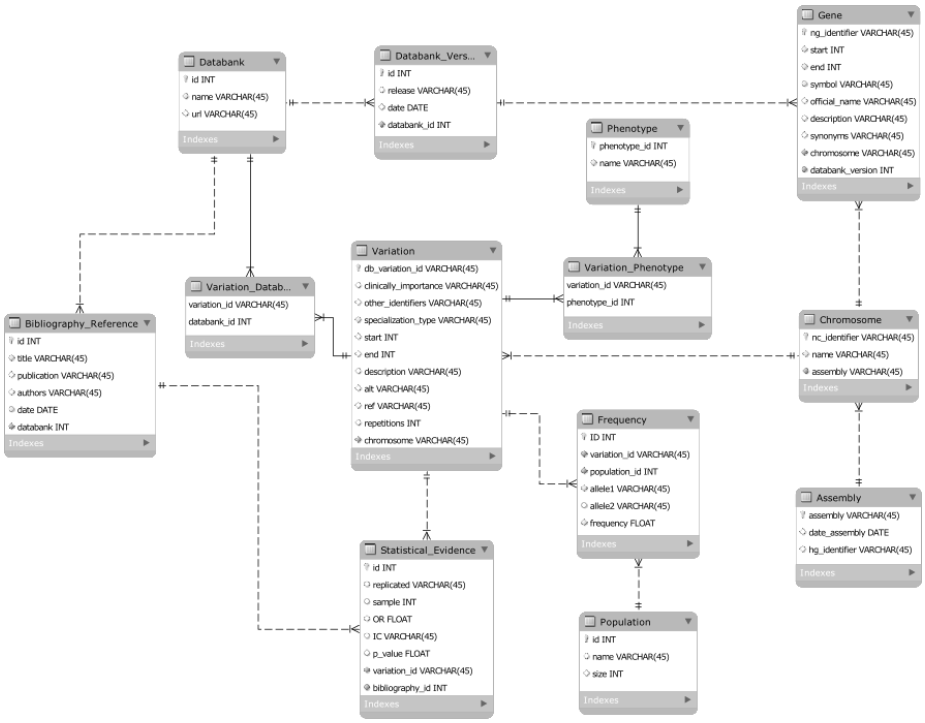

Figure 4.8: The Human Genome Database.

- **Attribute Value Constraints:** Implies checking column data types and not null constraints.

- **Referential Integrity Constraints:** Implies checking foreign key constraints.

- **Key Constraints:** Implies checking unique key constraints. Primary key constraints can be implemented as a combination of a unique key rule and not null rules of each column that is part of the primary key.

A new set of Kettle transformations were implemented in order to support the check of the above mentioned constraints, as well as the final load of the information in the HGDB. Once the load process is finished, the information is ready for the Exploitation phase what can be performed by tools such as VarSearch (described in Section 3.1.4).

The storage of the information is the last stage in the building of the GeIS that provides support to the tasks to be performed for the validation of the SILE method. In the next sections we expose the results of applying the SILE method in the different case studies with the support of the GeIS.

4.2 The Migraine Case

The aim of this case study is to determine the accuracy of the results provided by the SILE method by comparing them with those manually obtained by the stakeholders in a previous research. The task to be performed is the identification of the DNA variants associated to the risk of suffering migraine, in order to perform a genetic diagnosis from a patient sample.

As the most prevalent and disabling neurological disorder, migraine affects the lives of millions of people worldwide, and for many there are still no effective treatments [73]. Migraine attacks cause severe throbbing pain or a pulsing sensation, usually on just one side of the head. It is often accompanied by nausea, vomiting, and extreme sensitivity to light and sound. They can cause significant pain for hours to days and can be so severe that the pain is disabling. The genetic causes associated to migraine are being studied by looking at the DNA of large families where migraine is passed down at every generation [74]. In recent years, it has become very clear that migraine is mainly a disorder of the nervous system and significant progress has been made in the understanding of the causes as well as new ideas for treatments have arisen. This is an important step in working out why some people are predisposed to suffering this condition.

The accuracy of the SILE method in this case study is measured by determining the presence of false positives (variants incorrectly considered as relevant) and false negatives (variants incorrectly considered as irrelevant) in the results that pass all the quality criteria that have been established. False positive results have a negative impact due to they increase the amount of information to be reviewed by the experts and consequently it leads to a loss of confidence in its usefulness. Furthermore, they could have a direct impact in the health of the patients whom could undergo unnecessary treatment or take unnecessary drugs. On the other hand, having false negatives could imply the loss of information that could delay the diagnosis and treatment of the disease.

To test the accuracy of the SILE method, the first step is to extract, transform and integrate the information coming from the genomic repositories that have been selected as relevant. The number of variants extracted vary from one genomic repository to another (174 from Ensembl, 351 from ClinVar and 203 from GWAS Catalog). This is due to the scope of the repositories. While ClinVar stores information about different types of variants (SNPs, insertions, deletions, etc.), Ensembl stores information about SNPs and GWAS focuses on GWAS studies.

The information about the DNA variants must be integrated with the associated bibliography (219 articles from PubMed) and the structural information about the genes affected by the DNA changes (385 genes from HGNC and Entrez Gene), as can be seen in Figure 4.9.

Most of the variants are SNPs so extra information must be added from the

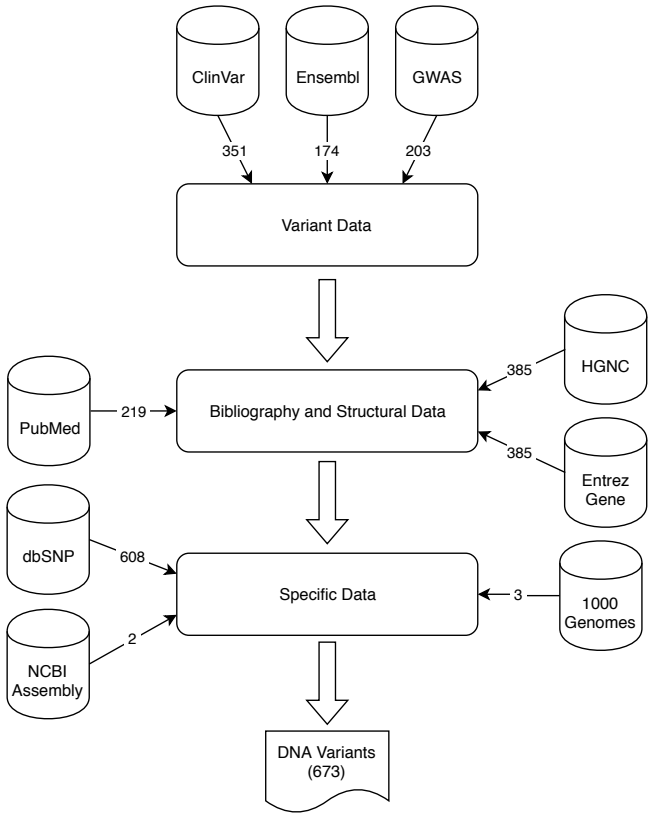

Figure 4.9: Extraction and integration of the DNA variants associated with migraine.

dbSNP repository. The reported studies associated to the variants have been performed over three main populations (European, Han Chinese and Dutch) and the location of each variant are expressed according to two different reference sequences or assemblies: GRCh38 and GRCh37. All the extracted information went through the transformation and integration process described in Section 4.1.2 and after removing duplicates, 673 unique DNA variants were collected.

If we look at the common information among the data sources, only 55 variants are present both in ClinVar and Ensembl (see Figure 4.10). Furthermore, there are not variants present in the three repositories, what justifies the importance of considering as much sources as possible when performing the research to ensure the completeness of the information. The omission of any of the selected sources could lead in the loss of hundreds of potentially valid variants as well as the loss of information crucial for the research process.

Once the information has been extracted and integrated, it is ready to per-

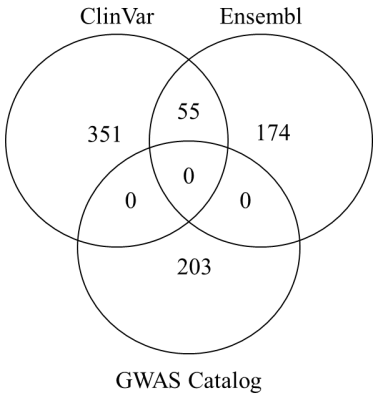

Figure 4.10: Common results among databases in the migraine case study.

form the quality assessment in order to determine which variants are relevant enough to be used in a genetic diagnosis. In Figure 4.11, the evolution of the results along the quality process are shown. Using as starting data set the 673 variants determined in the previous phase, the different metrics are systematically applied in order to classify the results according to their level of relevance:

- **Variants discarded:** about the 3% of the variants are discarded because the clinical significance is not relevant for the task at hand (i.e. benign, likely benign, uncertain and likely pathogenic). Attending to the associated evidence, about 54% of the variants have been discarded due to there is not bibliography to support their connection with the disease.
- **Variants with contradictory evidence:** about 5% of the variants are discarded due to conflicts in the interpretation of the clinical significance (e.g. one submitter considers the variant as benign and another one considers the variant as pathogenic).
- **Variants without enough evidence provided:** the 30% of the variants are discarded because the associated studies are not statistically significant. Nevertheless, not all the databases provide this information so the GeIS returns an additional result: a list with 49 variants that require a manual review of the associated bibliography.
- **Variants accepted:** only 4 variants pass all the established criteria and thus are considered as relevant.

The number of variants to be manually reviewed (491) is considerable lower than the number of variants present in the initial dataset (673), as well as the effort that the revision implies. In Table 4.5, a summary of the variants selected as relevant by the GeIS is shown.

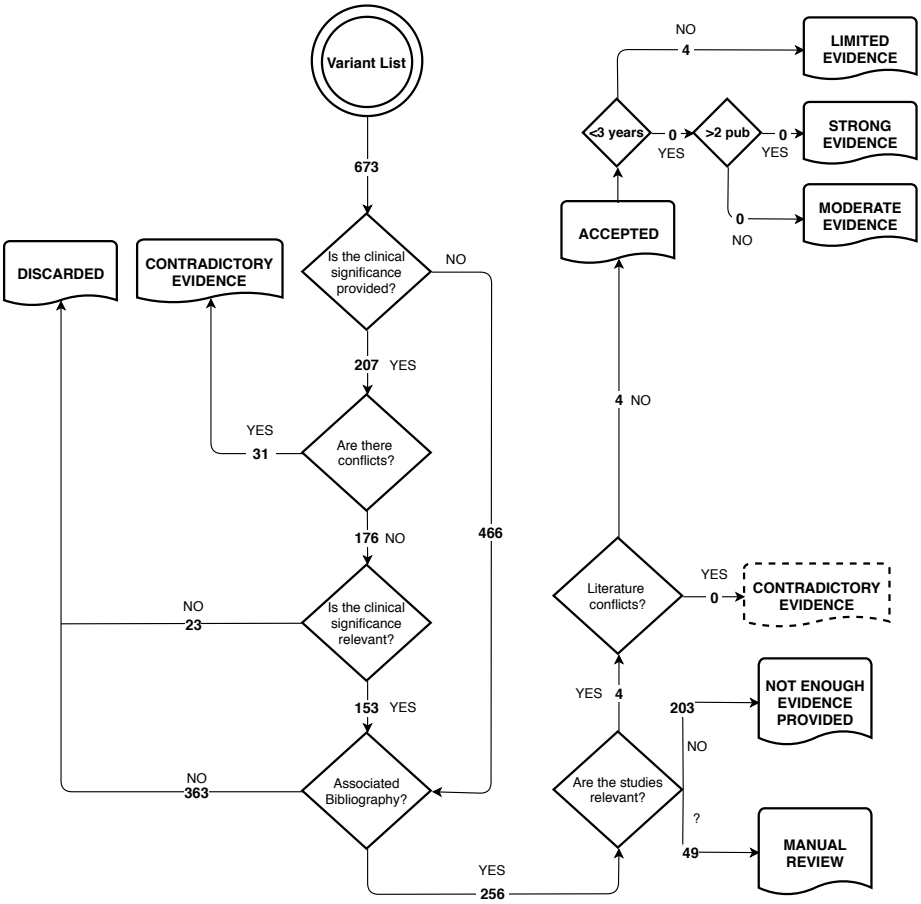

Figure 4.11: Quality assessment results for migraine case study.

Due to only one article is associated to each variant and more than 3 years have lasted since they have been published, the evidence can be considered as *Limited*. This means that although the increasing number of studies, there are not sound results yet and the conclusions derived from the genetic diagnosis must be carefully considered.

4.3 The Epilepsy Case

The aim of the second case study is to determine if the SILE method is useful to manage the genomic information in a particularly difficult disease context, the epilepsy. The task to be performed is the same as for the first case study: to determine the DNA variants associated to the risk of suffering the disease. Nevertheless, epilepsy is a spectrum condition with a wide range of seizure types

Table 4.5: Variants associated with the risk of suffering migraine.

ID	Gene	Chromosome	Clinical Significance	Bib. ID/Year
rs10166942	TRPM8	2	Risk Factor	21666692 (2011)
Clinvar:12388	TNF	6	Risk Factor	14718719 (2004)
rs1835740	LOC 101927066	8	Risk Factor	20802479 (2010)
rs2651899	PRDM16	1	Risk Factor	21666692 (2011)

what makes the gathering and analysis of the genetic information a challenge [75]. 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 [76]. Thus, in order to provide an accurate genetic diagnosis it is crucial to manage as much information as possible, which is a time consuming task for researchers. To this aim, we validate the usefulness of the SILE method to collect and manage the genomic information associated to epilepsy in terms of time saving and support to the research process.

Table 4.6: Number of variants associated with the risk of suffering epilepsy (grouped by keyword).

	Epilepsy	Seizures	Lafora	Unique Variants
ClinVar	6,932	6,238	78	**11,243**
Ensembl	4,632	1,252	4	**5,111**
GWAS Catalog	61	7	0	**68**
Unique Variants	**7,224**	**6,200**	**76**	**11,506**

The first step is to extract, transform and collect all the information associated to the keywords "epilepsy", "seizures" and "Lafora" (a type of progessive myoclonic epilepsy). As can be seen in Table 4.6, the number of DNA variants collected (11,506) are considerable higher than those obtained for the first case study (673), which gives an idea of the complexity of the disease context.

For each row and column, the table shows the number of unique results, due to the same variant can be stored in more than one database or belong to more than one type of epilepsy. As a result, the total number of DNA variants to be analyzed is 11,506. The information about the DNA variants must be integrated

with the associated bibliography (844 articles from PubMed) and the structural information (1,509 affected genes), as can be seen in Fig 4.12.

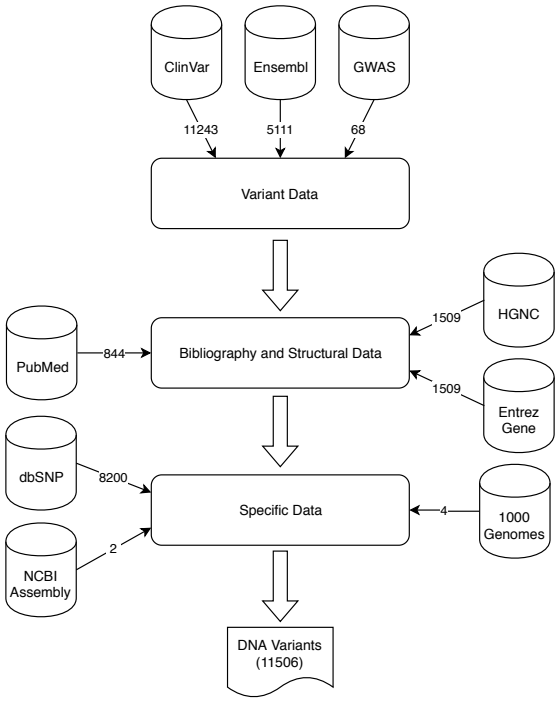

Figure 4.12: Extraction and integration of the DNA variants associated with epilepsy.

From the total amount of variants, 8,200 are SNPs so extra information must be added from the dbSNP repository. The reported studies associated to the variants have been performed over 4 main populations (European, Han Chinese, African American and Korean). As in the migraine case study, the location of each variant is expressed according to two different reference sequences or assemblies: GRCh38 and GRCh37.

Once the integration process is finished, the information is ready to perform the quality assessment which process can be seen in Figure 4.13.

At the end of the process, the variants are classified according to their relevance for the task at hand:

- **Variants discarded:** about the 11% of the variants are discarded due to the clinical significance is not relevant (benign, uncertain, etc.). It is very significant that almost 64% of the variants are discarded due to they do not have associated bibliography.

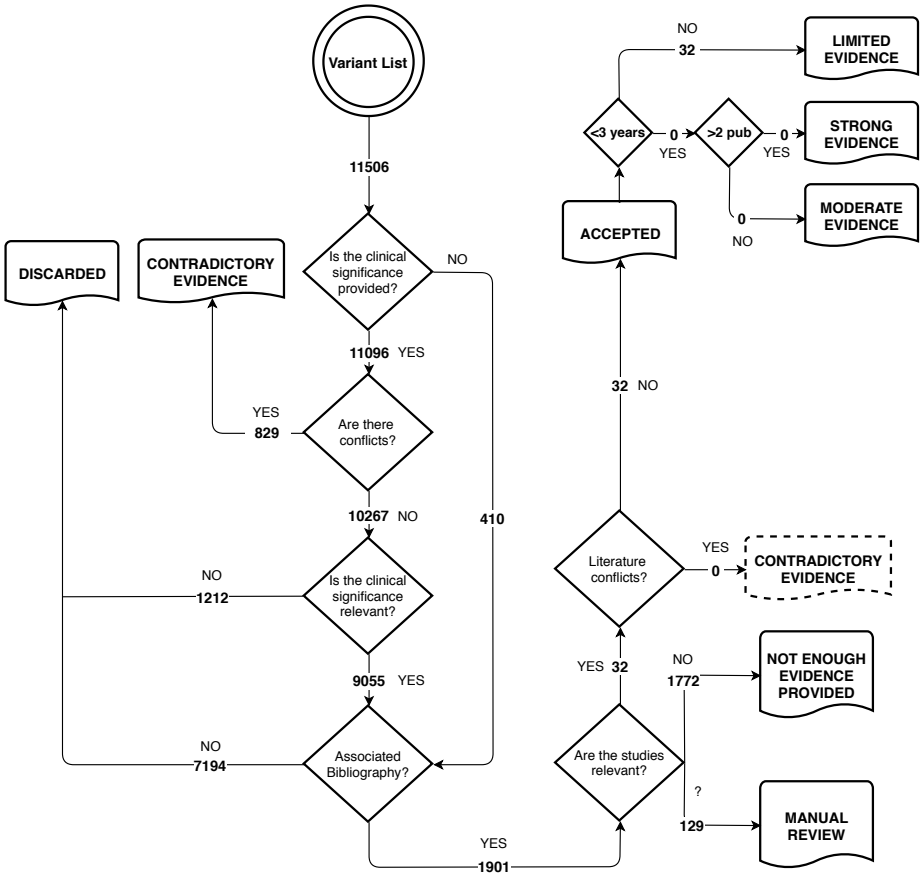

Figure 4.13: Quality assessment results for the epilepsy case study.

- **Variants with contradictory evidence:** over 7% of the variants are discarded due to conflicts in the interpretation of the clinical significance (e.g. one submitter considers the variant as benign and another one considers the variant as pathogenic).
- **Variants without enough evidence provided:** about 16% of the variants variants are discarded due to the associated studies are not statistically significant. In addition, 129 variants require a manual review due to the information needed to perform the analysis is not provided by the repositories.
- **Variants accepted:** finally, 32 variants have pass all the quality criteria and thus they are considered as relevant. In Annex D, a summary of the selected variants is shown.

As happened in the first case study, the SILE method results in a substantial

time saving considering the initial amount of variants extracted from the sources. Due to only one article is associated to each variant and more than 3 years have lasted since they have been published, the evidence is considered as *Limited*. The results are consistent with the existent knowledge about the disease. The complexity of the biological mechanisms that causes the epilepsy as well as the difficulty in its classification hinder the process of identification of relevant information if it has to be performed manually. Nevertheless, with the SILE method we were able to differentiate those variants that can be considered as relevant among the huge amount of available information.

Comparing the results obtained in both case studies, we can see the importance of providing methods to measure the quality of the information in fields were the knowledge is evolving at a fast pace. As can be seen in Table 4.7, less than 1% of the initial variants for both case studies can be considered as relevant.

Table 4.7: Comparison of results between both case studies.

	Migraine	**Epilepsy**
Discarded	57%	75%
Contradictory evidence	5%	7%
Not enough evidence	30%	16%
Manual review	7%	1%
Accepted	1%	<1%

By identifying the information that it is not relevant (at this moment), we can decrease the amount of time required to analyze the evidence that is valuable and we can help the stakeholders to focus only on those that can provide new meaningful insights about the area of study. In the next section we are going to discuss the results obtained in terms of validity and usefulness for the stakeholders.

4.4 Results

In order to answer the research questions mentioned at the beginning of this chapter, the results of both case studies have been presented to two experts in genetic diagnosis (a geneticist and a clinician), that lead the team on applied genomics in a company dedicated to the genomic diagnosis in a context of industrial Medicine of Precision.

4.4.1 Migraine Case Study Results

The aim of the first case study was to answer the RQ9: To which extent are the results of our method accurate and valid?

The accuracy of the SILE method was measured by determining the presence of false positives (variants incorrectly considered as relevant) and false negatives (variants incorrectly considered as irrelevant). After the analysis of the results and the comparison with those obtained by the stakeholders, the results were:

- Number of false positives: 1
- Number of false negatives: 0
- Number of new findings: 1
- Number of results in line with current research: 2 (100%)

According to these results, all the relevant variants have been identified which is one of the main concerns when dealing with information with the aim of providing a genetic diagnosis. Furthermore, one relevant variant that the stakeholders did not considered has been found. As a consequence, it is going to be included in their catalog of variants to be checked for this disease.

According to the variant classified as false positive (rs1835740), a new discussion has been opened. The cause why the stakeholders discarded the variant as relevant was because there is one study that has not found association between the variant and the disease in the Spanish population [77]. As we are not focusing on any specific population, we cannot discard the variant, that has been proved to be relevant in other populations such as Finland, The Netherlands, Germany and Denmark [78]. Due to the discrepancies among populations must be carefully studied to provide new insights about the disease, the identification of these cases is going to be considered in a future version of the information system. Overall, the results obtained by applying the SILE method have been considered as accurate by the stakeholders.

4.4.2 Epilepsy Case Study Results

The aim of the second case study was to answer the RQ10: Do domain experts think that our method is useful to manage genomic information?

The method is considered useful if it can help the researchers in the identification of potentially relevant biomarkers for the diagnosis of a complex disease such as epilepsy, whose analysis is still under open investigation and it is not yet included in their catalog of selected phenotypes. From the initial dataset of 11,243 variants, only 32 have been considered as relevant. After the analysis of the results, the conclusions are:

- Number of results in line with the current research: 15 (100 %)
- Number of new findings: 6
- Number of potentially significant biomarkers: 11

According to these results, the stakeholders have considered the method as useful due to 15 variants are in line with the current results obtained by them. Furthermore, 6 variants that they have not found are going to be added to the catalog of variants to be checked for the disease. Due to the complexity of the information to be managed, the stakeholders had to reduce the research and focuse only on a few types of epilepsy. With the SILE method we were able to handle the data associated to a higher amount of epilepsy types and as a result, 11 variants were considered as potentially significant for a further increase in the coverage of the genetic service provided by the stakeholders.

4.4.3 Complementary Case Studies

The good results obtained in the validation has pushed us to start a generic validation to analyze in depth the particularities of the method when applied to different phenotypes. Our final goal is to make SILE become a fundamental method to manage genome data providing the right, valuable contents to the most advanced Genome Information Systems. To this aim, we report here the results of the last two applications of the method that have been done at the request of the stakeholders with the aim of corroborate its value in different scenarios. The selected phenotypes were Crohn disease and male breast cancer.

4.4.3.1 Crohn Disease Case Study

Crohn disease, also known as inflamatory bowel disease type 1, is a chronic disorder that involves an abnormal funtion of the immune system which causes inflammation in the digestive system [79]. The causes of Crohn results from a combination of genetic, environmental and lifestyle factors. Even when the inheritance pattern of this disease is unclear, it is known that about 15% of affected people have a first-degree relative with the disorder [80]. The understanding of Crohn's disease is in an early stage and the identification of relavant DNA variants is useful to provide insights into disease biology and potential therapeutic targets.

As can be seen in Figure 4.14, the number of variants initially extracted from the genomic repositories is 739. There is a huge amount of bibliography (2,976 articles) about studies performed over 120 genes and 6 populations (European, Japanese, Jewish, Korean, Polish and Southern European).

Despite the evidence provided, how many variants can be considered as relevant for the genetic diagnosis? To answer this question, we performed the data quality assessment which results can be seen in Figure 4.15.

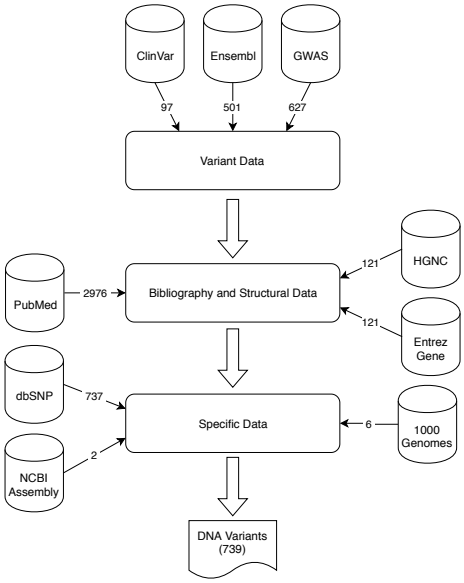

Figure 4.14: Extraction and integration of the DNA variants associated with Crohn disease.

From the initial amount of 739 DNA variants and 2976 articles, only 29 variants and 25 articles can be considered as relevant to perform a genetic diagnosis. Over 37% variants were discarded because they do not have a relevant clinical significance or associated bibliography. Over 57% of the variants were discarded due to the studies performed were not statistically relevant enough. Only 5 variants required manual review because the information needed to perform the analysis was not provided. After the manual review, none of them was considered as relevant. One of the variants included in the final results (rs2241880) has a moderate level of evidence because there is one study performed less than three years ago (2016). The rest of variants have a limited level of evidence. A summary of the information associated to the variants found can be seen in Annex E.

4.4.3.2 Male Breast Cancer Case Study

Less than 1% of all breast cancers occur in men, what makes it a rare disease. As a consequence, few cases are available to study so the population samples used in the studies are very small. Nevertheless, when a number of these small studies are grouped together, new insights can be derived from them [81]. The understanding of the genetic causes of male breast cancer is important because it can drastically impact the medical management for patients and their relatives.

As can be seen in Figure 4.16, the initial amount of DNA variants that can

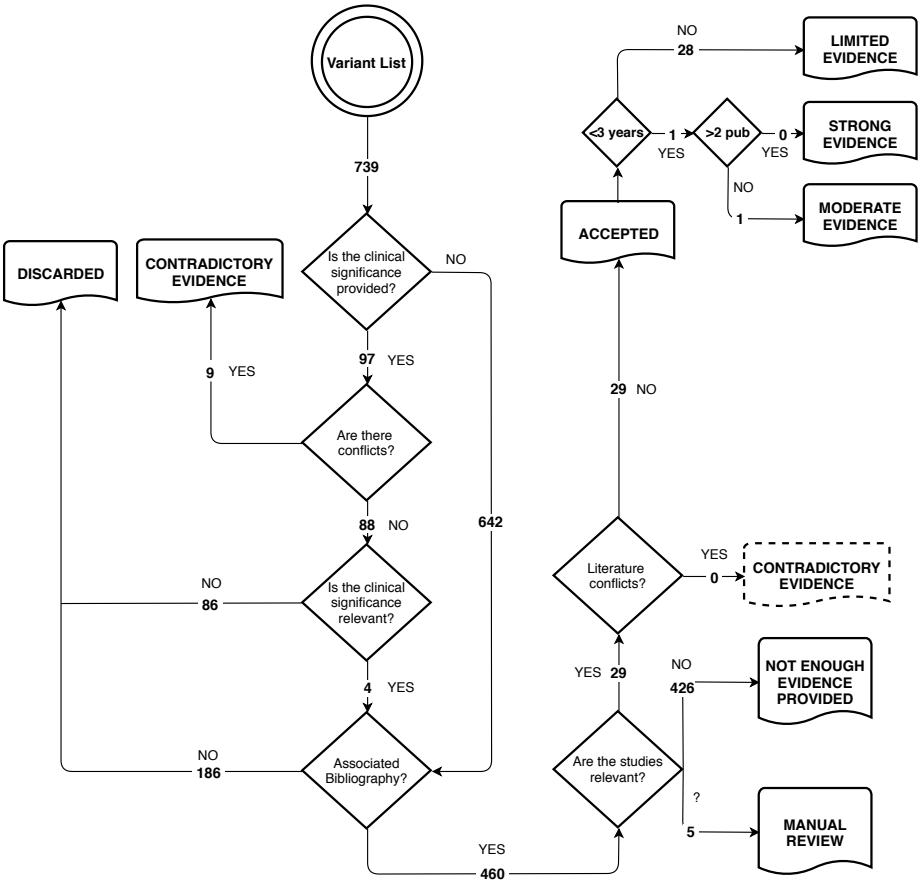

Figure 4.15: Quality assessment results for the Crohn case study.

be found in the genomic repositories is 117, with 1,266 articles associated and 21 genes under study. Nevertheless, only one population (European) has been screened, in contrast with the other case studies, where the populations under study were diverse.

Once the initial list of variants has been established, we performed the quality assessment to determine how many of them are relevant for genetic diagnosis. Due to the number of men that suffer this cancer is low, none of the studies associated to the variants pass the quality filters stablished in 700 cases and 700 controls. As a consequence none of the variants can be considered as relevant according to our initial criteria.

After presenting the results to the stakeholders, they changed the quality criteria to accept those studies performed over 500 individuals (cases and controls). One of the strengths of the SILE method is its versatility, that allows

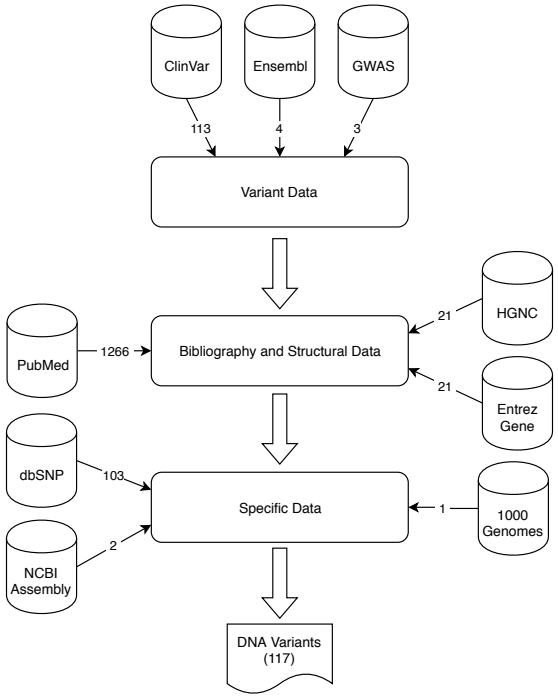

Figure 4.16: Extraction and integration of the DNA variants associated with male breast cancer.

the adjustment of the quality criteria depending on the context of the disease under study without incrementing the cost of its performance. The results of this new implementation are shown in Figure 4.17.

From the initial amount of 117 DNA variant, only 8 fulfill the new requirements. Most of the variants have been discarded due to the clinical significance was not clear (likely benign, likely pathogenic or uncertain) or not relevant enough.

From the variants selected as relevant, 6 of them have a moderate level of evidence because there are two studies performed less than 3 years ago (2016 and 2017). The other variants have a limited level of evidence. A summary of the information associated to the variants can be seen in Table 4.8.

The results from both complementary case studies were satisfactorily validated by the stakeholders, what corroborates the usefulness of the method in different scenarios.

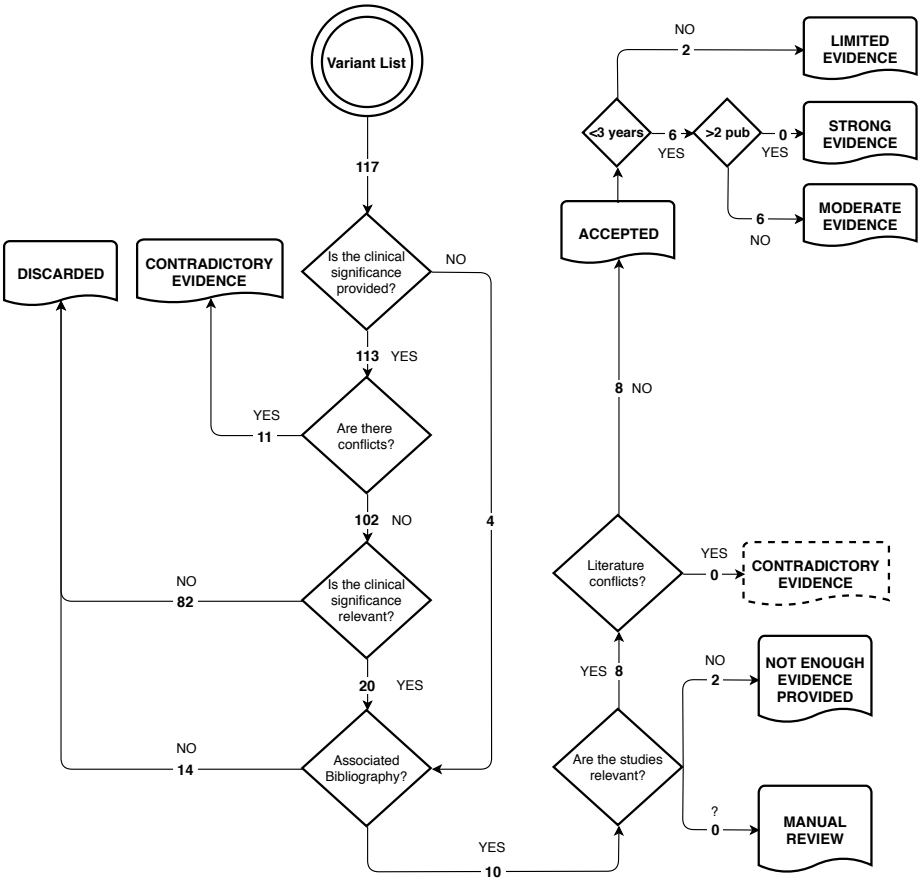

Figure 4.17: Quality assessment results for the male breast cancer case study.

4.5 Conclusions

Throughout this chapter, we have validated the solution proposed in terms of the adequacy of the contribution to fulfill the stakeholders goals. In order to validate the accuracy of the method and its usefulness, we proposed two case studies and developed the GeIS that supports the different stages of the process: from the selection of the adequate data sources to the final exploitation of the information.

As can be seen in both case studies, the management of genomic information 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 proper information system support. The SILE method has been considered accurate and useful for the stakeholders, as much from the precision of the results as from the support to the management of the information to achieve new research

Table 4.8: Variants associated with the risk of suffering male breast cancer.

ID	Gene	Chromosome	Clinical Significance	Bib. ID/Year
rs1314913	RAD51B	14	Increased Risk	26248686 (2015) 23001122 (2012)
rs3803662	TOX3	16	Increased Risk	21949660 (2011) 23001122 (2012)
rs397507751	BRCA2	13	Pathogenic	28008555 (2017)
rs80359276	BRCA2	13	Pathogenic	28008555 (2017)
rs80359492	BRCA2	13	Pathogenic	28008555 (2017)
rs80359598	BRCA2	13	Pathogenic	26360800 (2016)
rs80359677	BRCA2	13	Pathogenic	28008555 (2017)
rs80359763	BRCA2	13	Pathogenic	28008555 (2017)

goals. As a result, two new implementations of the method were performed at the request of the stakeholders: one for the study of the genetic causes of Crohn disease and another one for male breast cancer. The results of these implementations confirmed the usefulness of the SILE method, and open the opportunity to propose SILE as a fundamental method to provide valuable contents for those GeIS that are intended to become the cornerstone of the modern Medicine of Precision.

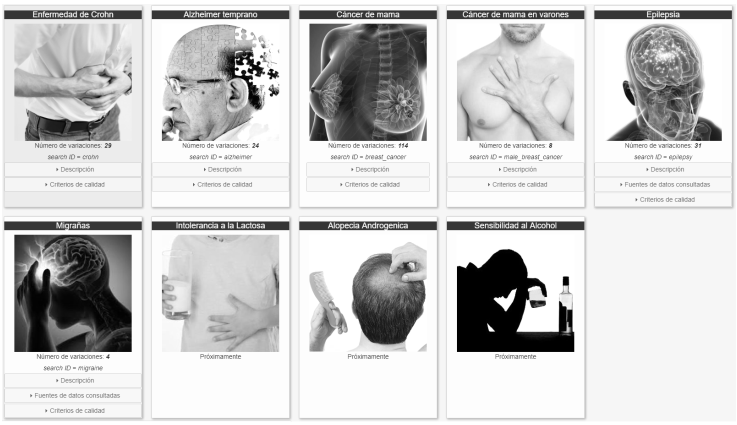

Figure 4.18: Available phenotypes through the GenesLove.Me platform.

After the validation process, a fruitful communication path has been established and new opportunities to check the validity of the method in other projects have arisen. One of the projects is a collaboration with the child oncology group of the Hospital La Fe (Valencia), focused on adding the genetic risk of Neuroblastoma, a type of childhood cancer, to the clinical trials performed in the patients in order to determine the effectiveness of the treatments.

A second project derived from the results obtained in this thesis is the development of an online platform that serves as basis for the genetic testing of the studied phenotypes. The platform is called GenesLove.Me [12] and the users can access the information about the DNA variants as well as the quality criteria used to determine their relevance. Figure 4.18 shows the prototype interface where the user can access the available phenotypes.

This prototype serves as basis for a bigger goal of the research center, the development of an industrial platform for the management of genomic data that takes advantage of the high quality results obtained by the SILE method.

[12] https://varsearch2.dsic.upv.es:4342

Chapter 5

Conclusions and Future Work

We started this thesis by formulating the research questions associated to four main goals: determine the problems that hinder the management of genomic information, provide a method for the efficient management of genomic information, provide a set of quality criteria to ensure that data are reliable a correct, and validate the contribution of this research work.

After introducing the characteristics of genomic data sources and the problems to face when trying to connect and integrate the knowledge they collect, we proposed a method (SILE) to manage the genomic information in an efficient way. The method is supported by a conceptual model (the CSHG) that provides the solid conceptual structure required to connect the genomic information. In addition, the SILE method is also supported by a data quality methodology that guarantees that the information managed is of enough quality. This ensures that the conclusions derived from its analysis are reliable and accurate.

Finally, we tested the validity of the method by building a Genomic Information System (GeIS) that supports its application in four case studies, and verified the results with the stakeholders.

In this last chapter of the thesis, we use the results of this work to answer the research questions we formulated in the first chapter (section 5.1), present its impact in terms of publications, academic works and academic projects (section 5.2), and suggest directions for future work (section 5.3).

5.1 Answers to Research Questions

Throughout this research work, we have been working toward answers for the research questions associated to the goals defined in the first chapter.

5.1.1 Results of Objective 1

Along chapter 2, we answer the set of knowledge questions associated to the first objective of this thesis: determine the problems that hinder the management of genomic information.

RQ1. *Where can the genomic information be found?*

The advances in sequencing technologies, such as Next Generation Sequencing (NGS) have allowed the development of multitude of research projects around the world. Most of the knowledge generated by these projects are publicly available for the scientific community in over thousand online genomic repositories. These resources are specialized databases that provide not only information about DNA sequences, but also data on gene expression, macromolecular structures, gene-disease associations and genotype frequencies in diverse populations. Nowadays, there is no way of knowing with certainty the number of active genomic data sources due to the information about them is scattered in various life science journals and around the Web. Furthermore, a significant number of them become obsolete quickly due to the loss of technological maintenance or the lack of updates. We answered this question in section 2.1.

RQ2. *Which problems arise when managing genomic information?*

In order to understand complex biological systems and validate their experiments, geneticists and researchers are forced to delve into a lake of information as well as connect as much databases as possible. However, the genomic information differs not only in its scope but also in the way the same concepts are modeled. Along section 2.2, we describe the current efforts of the community to provide ontologies and conceptual models to represent the key concepts of the domain. In addition, due to the complexity of the biological processes, the experimental nature of the research and the different techniques and protocols used by the laboratories, there is a high variability in the quality of the results, increasing the noise and the effort required to extract meaningful conclusions. Following this premise, we performed a study of the most common errors present in the genomic data sources (section 2.3), which helped us to define the problems that researchers must face when managing genomic information. These problems can be summarized as: the selection of the adequate data sources, the identification of high-quality data, and the storage and analysis of the information with the aim of achieving competitive advantage.

5.1.2 Results of Objective 2

Along the first part of chapter 3 (section 3.1), we answered the set of research questions associated with the second objective of this thesis: to provide a method for the efficient management of genomic information.

The SILE method comprises 4 systematic stages (Search, Identification, Load and Exploitation) and uses the Conceptual Schema of the Human Genome

(CSHG) as the conceptual structure required to harmonize the information under a holistic perspective.

RQ3. *How can the most suitable genomic data sources be found?*

The aim of the first stage of the method is to identify the genetic repositories that store the information required to succeed in fulfilling the goals of a defined task. By using a graphical interface developed for this purpose and the CSHG, the researcher can explore new data sources, identify which ones are the most adequate to extract information from, and ensure that not relevant repositories have been left aside. This question has been answered in section 3.1.1.

RQ4. *How can the relevant information be identified?*

The aim of the second stage of the SILE method is to identify the required information from each repository and harmonize it into a common data model so it could be loaded into a database for its further exploitation. This requires to have a deep knowledge of the underlying structure of each repository as well as to define a set of extraction and transformation rules that are fully described in section 3.1.2.

RQ5. *How can the information be structured and stored for its further exploitation?*

The data coming from the genomic repositories must be transformed into a queryable format to ensure its persistency and allow the extraction of knowledge. Both tasks are performed in the third and fourth stages of the SILE method. Along sections 3.1.3 and 3.1.4, we answer this research question and present a tool called VarSearch that has been developed to this aim.

5.1.3 Results of Objective 3

The last part of chapter 3 (section 3.2) is dedicated to answer the set of research questions associated to the third objective of this thesis: to provide a set of quality criteria to ensure that data are reliable and correct.

RQ6. *Which are the criteria that genomic information must fulfill to ensure its quality?*

Using the results of the study described in section 2.3, we determined a set of quality dimensions and quality metrics that genomic information must fulfill to ensure that the information managed is reliable and valuable.

RQ7. *How can the quality of genomic information be measured?*

In order to sistematically apply the dimensions and metrics, we defined a Data Quality Methodology specially designed for genomic information. The methodology has been applied in the Search and Identification stages of the SILE method as explained in sections 3.2.2 and 3.2.3.

Once the method and the quality criteria have been established, we performed the validation by answering a new set of research questions.

RQ8. *To which extent are the results of our method accurate and valid?*

The accuracy of the SILE method was measured by determining the presence of false positives and false negatives in the results obtained after its application in a case study: the identification of the DNA variants associated to the risk of suffering migraine, in order to perform a genetic diagnosis from a patient sample. After the analysis of the results and the comparison with those obtained by the stakeholders, we concluded that the results of the method are accurate for the task at hand. The answer to this research question is explained in section 4.2.

RQ9. *Do domain experts think that our method is useful to manage genomic information?*

To answer this research question we performed a second case study which aim is the identification of potentially relevant biomarkers for the diagnosis of a complex disease: epilepsy. After the analysis of the results, the stakeholders have considered the method as useful so it was applied to two additional case studies to complement their work: Crohn disease and male breast cancer. The application of these case studies is detailed in sections 4.3 and 4.4.

5.2 Thesis Impact

This research work has been validated through the publication of the results in different international forums. In addition, some academic works have been developed as complementary work as well as the participation in research projects. In this section the mentioned contributions are summarized.

5.2.1 Publications

The results that have been obtained during the development of this research have been published and presented in forums of high impact in the field of Conceptual Modeling and Information Systems; all of them with international projection:

A. León, O. Pastor, *Infoxication in the Genomic Data Era and Implications in the Development of Information Systems*, Accepted in the IEEE Thirteen International Conference on Research Challenges in Information Science (RCIS), 2019.

A. León, O. Pastor, *Smart Data for Genomic Information Systems: the SILE Method*, in: Complex Systems Informatics and Modeling Quarterly Journal (CSIMQ), Article 97, Issue 17, pp. 1-–23, 2018.

A. León, A. García, J.C. Casamayor, J. Reyes, *Genomic Data Management in Big Data Environments: The Colorectal Cancer Case*, in: Advances in Con-

ceptual Modeling. ER 2018. Lecture Notes in Computer Science, vol 11158, pp. 319–329, 2018.

A. León, I. Pascual, O. Pastor, *Genomic Information Systems applied to Precision Medicine: Genomic Data Management for Alzheimer's Disease Treatment*, in: Designing Digitalization (ISD2018 Proceedings), 2018.

A. León, O. Pastor, J.C. Casamayor, *A Method to Identify Relevant Genome Data: Conceptual Modeling for the Medicine of Precision*, in: Proceedings ER, pp. 597-–609, 2018.

A. León, O. Pastor, *From Big Data to Smart Data: A Genomic Systems Perspective*, in: Proceedings RCIS, pp. 1-–11, 2018.

A. León, O. Pastor, *Towards an Effective Medicine of Precision by using Conceptual Modelling of the Genome*, in Proceedings ICSE-SEHS, pp. 14-17, 2018.

A. León, J. Reyes, V. Burriel, F. Valverde, *Data Quality Problems When Integrating Genomic Information*, in: ER 2016 Workshops. LNCS, Springer International Publishing, pp. 173-–182, 2016.

Additional contributions:

In addition to the main publications, this research served as contribution for different works that have been developed in parallel such as:

A. Escalera, A. León, O. Pastor, *Sistemas de Información para una Medicina de Precisión: Identificación de Variaciones Genómicas para el Diagnóstico del Cáncer de Mama*, Accepted in the XXII Ibero-American Conference on Software Engineering (CibSE), 2019.

J.F. Reyes Román, A. León Palacio, O. Pastor López, *Software Engineering and Genomics: The Two Sides of the Same Coin?*, in: Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering, SCITEPRESS - Science and Technology Publications, pp. 301—307, 2017.

V. Burriel, J.F. Reyes Román, A. Heredia Casanoves, C.E. Iñiguez-Jarrín, A. León Palacio, *GeIS based on Conceptual Models for the Risk Assessment of Neuroblastoma*, in: 2017 11th International Conference on Research Challenges in Information Science (RCIS), IEEE, pp. 451—452, 2017.

O.P. López, A.L. Palacio, J.F.R. Román, J.C. Casamayor, *Modeling Life: A Conceptual Schema-centric Approach to Understand the Genome*, Conceptual Modeling Perspectives, Springer International Publishing, pp. 25—40, 2017.

5.2.2 Academic Works

The research performed throughout this thesis work has allowed the lead and development of several academic works that have contributed to the achievement of the objectives, as well as the obtaining of the corresponding bachelor 's and master's degrees by the students involved in them:

- **Biotecnología para la medicina de precisión: Identificación de variaciones genómicas para el diagnóstico de cáncer de mama.** Final Degree Project in Biomedical Engineering (2018). Alba Escalera Balsera. Advisor: Óscar Pastor López. Co-advisor: Ana León Palacio.
- **Un proceso para la identificación sistemática de variaciones genómicas: Aplicaciones a la medicina de precisión.** Master's Thesis in Software Engineering, Formal Methods and Information Systems (2018). Simranpreet Kahur. Advisor: Óscar Pastor López. Co-advisor: Ana León Palacio.
- **Exploración de bases de datos genómicas dirigida por modelos conceptuales.** Master's Thesis in Software Engineering, Formal Methods and Information Systems (2018). Vanessa Solís Cabrera. Advisor: Óscar Pastor López. Co-advisor: Ana León Palacio.

In addition to the above mentioned works, three more master's theses are under development, two in the field of Biomedichal Engineering and one in the field of Software Engineering.

5.2.3 Teaching Experience

It is also worth mentioning that another important result of this thesis has been the generation of material that has been used for the development of the teaching content of the following subjects:

- **Contribution of the Biomedical Engineer**: Bachelor's Degree in Biomedichal Engineering.
- **Design and Management of Genomic Information Systems:** Bachelor's Degree in Computer Science Engineering.
- **Bioinformatics:** Bachelor's Degree in Biomedichal Engineering.
- **Information Systems Applied to Bioinformatics: Management of Genomics Data:** Master's Degree in Software Engineering, Formal Methods and Information Systems.

5.2.4 Research Projects

In addition to the above mentioned contributions, I have collaborated in the following research projects:

- **Un Método de producción de software dirigido por modelos para el desarrollo de aplicaciones Big Data (DataME)**. Spanish Ministry of Science and Innovation. From December 2016 to December 2020. Ref. TIN2016-80811-P.
- **Innovative services for Digital Enterprises with ORCA (IDEO)**. Generalitat Valenciana. From January 2014 to December 2017. Ref. PROMETEO/2014/039.

5.3 Future Work

Based on the work proposed and the validation performed, in this section we formulate directions for future work.

Along this research work, we applied the SILE method in a group of case studies where its reliability and usefulness have been proved. However, the complexity of the genetic causes of the diseases requires its application in as many cases as possible to identify those situations where the method could be improved to make it more useful. For instance, some diseases are caused by the combination of DNA variants in more than one gene so it is required the study of what is called *haplotypes*: a group of alleles in an organism that are inherited together from a single parent. These alleles by itself are not relevant to develop the disease, but when a combination of them are present the probabilities are substantially increased which could affect the identification process and the established quality criteria.

The SILE method has been validated by applying it to answer a specific research question of interest for the stakeholders: the identification of DNA variants associated to the risk of suffering a disease. Nevertheless, the method can be applied to answer other research questions to cover different parts of the biological domain such as proteomics and pharmacogenomics. To this end, we must review the underlying conceptual schema to ensure that the required elements are correctly defined and to formulate new case studies to validate its reliability and usefulness.

Another important aspect of the proposed solution is to prove its validity to update the information identified in a first application of the method. As has been mentioned in the introduction, this domain is evolving at a fast pace and it is very common that in a short period of time new findings that contradict the current knowledge appear. Thus, it is important to verify that the information remains valid. To this aim, it is required to determine the issues that could arise during the update process and how the different stages of the method must be applied.

The tools developed for the search of genomic databases and the information system that assists the SILE method must be improved to enhance the user experience. For example, due to the limitations of the metadata provided by the

genomic repositories, part of the results provided by the GeIS must be manually verified in order to assure its validity. The automation of this process, using text mining to extract the required information from the literature, would improve its efficiency and would bring significant time savings for researchers.

The technology used to store the information is a relational database, nevertheless it would be interesting the study of the different database technologies and their performance according to the task to achieve.

As a final conclusion, we found that the proposed method are a step in the right direction, but we are aware that more research must be done and more cases should be analyzed to verify the working of the method in a greater number of situations. We continue in this line of work with the development of new research projects and academic works, hoping to continue to show the positive results we are achieving so far.

Bibliography

[1] Reza Mirnezami, Jeremy Nicholson, and Ara Darzi. Preparing for Precision Medicine. *New England Journal of Medicine*, 366(6):489–491, 2012.

[2] National Cancer Institute. Precision Medicine in Cancer Treatment, Accessed March 2019. https://www.cancer.gov/about-cancer/treatment/types/precision-medicine.

[3] Francis S. Collins, Michael Morgan, and Aristides Patrinos. The human genome project: Lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

[4] Christian Gilissen, Alexander Hoischen, Han G Brunner, and Joris A Veltman. Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, 20(5):490–497, may 2012.

[5] National Human Genome Research Institue (NIH). International collaboration aims to speed development of genomic medicine, Accessed March 2019. https://www.genome.gov/27561782/2015-news-feature-international-collaboration-aims-to-speed-development-of-genomic-medicine/.

[6] Hussein Abdel-Haleem. The Origins of Genome Architecture. *Journal of Heredity*, 98(6):633–634, 08 2007.

[7] National Human Genome Research Institue (NIH). What is Cancer?, Accessed March 2019. https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[8] Oscar Pastor. Conceptual Modeling of Life: Beyond the Homo Sapiens. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 18–31. Springer, Cham, 2016.

[9] Kara Rogers. International HapMap Project, Accessed May 2019. Available at: www.britannica.com/event/International-HapMap-Project.

[10] Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), Accessed January 2019. Available at: www.genome.gov/sequencingcostsdata.

[11] Mohamed Abu-farha. What is the difference between polymorphism and a mutation?, Accessed February 2019. Available at: www.researchgate.net/post/What_is_the_difference_between-_polymorphism_and_a_mutation.

[12] Celeste M. Condit, Paul J. Achter, Ilon Lauer, and Enid Sefcovic. The changing meanings of "mutation:" A contextualized study of public discourse. *Human Mutation*, 19(1):69–75, jan 2002.

[13] Roshan Karki, Deep Pandya, Robert C. Elston, and Cristiano Ferlini. Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Medical Genomics*, 8(1):37, dec 2015.

[14] Frédéric Piel, Anand Patil, Rosalind Howes, Oscar Nyangiri, Peter Gething, Thomas Williams, David Weatherall, and Simon Hay. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature communications*, 1:104, nov 2010.

[15] Rachel Drysdale. FlyBase. In *Methods Mol Biol*, pages 45–59. 2008.

[16] National Center of Plant Gene Research. Rice Mutant Database, Accessed March 2019. http://rmd.ncpgr.cn/.

[17] S. I. Letovsky, R. W. Cottingham, C. J. Porter, and P. W. Li. GDB: the Human Genome Database. *Nucleic acids research*, 26(1):94–9, jan 1998.

[18] Alex Bateman et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, jan 2017.

[19] Peter D Stenson et al. Human Gene Mutation Database (HGMD ®): 2003 update. *Human Mutation*, 21(6):577–581, jun 2003.

[20] David Croft et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, jan 2014.

[21] Jemila S. Hamid, Pingzhao Hu, Nicole M. Roslin, Vicki Ling, Celia M. T. Greenwood, and Joseph Beyene. Data Integration in Genetics and Genomics: Methods and Challenges. *Human Genomics and Proteomics*, 2009.

[22] José Fabián Reyes Román. *Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano*. PhD thesis, Universitat Politècnica de València, 2018.

[23] Roel Wieringa. *Design Science Methodology for Information Systems and Software Engineering*. Springer-Verlag, Berlin, Heidelberg, 2014.

[24] National Human Genome Research Institute (NIH). The Human Genome Project Completion: Frequently Asked Questions, Accessed January 2019. Available at: www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/.

[25] National Human Genome Research Institute. The Cost of Sequencing a Human Genome, Accessed May 2016. Available at: www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/.

[26] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Olena Verezemska, Michelle Isbandi, Alex D. Thomas, Rida Ali, Kaushal Sharma, Nikos C. Kyrpides, and T. B. K. Reddy. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Research*, 45(D1):D446–D456, 2017.

[27] Y.-B. Chen, Ansuman Chattopadhyay, Phillip Bergen, Cynthia Gadd, and Nancy Tannery. The Online Bioinformatics Resources Collection at the University of Pittsburgh Health Sciences Library System–a one-stop gateway to online bioinformatics databases and software tools. *Nucleic Acids Research*, 35(Database):D780–D785, jan 2007.

[28] Daniel J. Rigden and Xosé M. Fernández. The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection. *Nucleic Acids Research*, 47(D1):D1–D7, jan 2019.

[29] Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, may 2000.

[30] Karen Eilbeck, Suzanna Lewis, Christopher Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):R44, 2005.

[31] Mauno Vihinen. Variation Ontology for annotation of variation effects and mechanisms. *Genome Research*, 24(2):356–364, feb 2014.

[32] Giancarlo Guizzardi. Summary for Policymakers. In *Climate Change 2013 - The Physical Science Basis*, pages 1–30. Cambridge University Press, 2007.

[33] Antoni Olivé. *Conceptual Modeling of Information Systems*. Springer Berlin Heidelberg, 2007.

[34] I.-M.A. Chen and V.M. Markowitz. Modeling scientific experiments with an object data model. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 391–400. IEEE Comput. Soc. Press, 1995.

[35] T. Okayama, T. Tamura, T. Gojobori, Y. Tateno, K. Ikeo, S. Miyazaki, K. Fukami-Kobayashi, and H. Sugawara. Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library. *Bioinformatics*, 14(6):472–478, jan 1998.

[36] Claudine Médigue, François Rechenmann, Antoine Danchin, and Alain Viari. Imagene: An integrated computer environment for sequence annotation and analysis. *Bioinformatics*, 15(1):2–15, 1999.

[37] N. W. Paton, S. Khan, A. Hayes, F. Moussouni, A. Brass, K. Eilbeck, C. a Goble, S. J. Hubbard, and S. G. Oliver. Conceptual modelling of genomic information. *Bioinformatics*, 16(6):548–557, jun 2000.

[38] Sudha Ram and Wei Wei. Modeling the Semantics of 3D Protein Structures. In *Genome*, pages 696–708. Springer, Berlin, Heidelberg, 2004.

[39] Anna Bernasconi, Stefano Ceri, Alessandro Campi, and Marco Masseroli. 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)*, volume 10650 LNCS, pages 325–339, 2017.

[40] José F. Reyes Román, Óscar Pastor, Juan Carlos Casamayor, and Francisco Valverde. 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)*, pages 404–412. Springer International Publishing, 2016.

[41] Heiko Müller and Felix Naumann. Data quality in genome databases. In *Eighth International Conference on Information Quality (ICIQ 2003)*, pages 269–284, 2003.

[42] Mauno Vihinen, John M. Hancock, Donna R. Maglott, Melissa J. Landrum, Gerard C. P. Schaafsma, and Peter Taschner. Human Variome Project Quality Assessment Criteria for Variation Databases. *Human Mutation*, 37(6):549–558, 2016.

[43] Ana León, José Reyes, Verónica Burriel, and Francisco Valverde. Data Quality Problems When Integrating Genomic Information. In *Advances in Conceptual Modeling*, pages 173–182. Springer International Publishing, 2016.

[44] Wilco W.M. Fleuren and Wynand Alkema. Application of text mining in the biomedical domain. *Methods*, 74:97–106, mar 2015.

[45] Steven L. Salzberg. Genome re-annotation: A wiki solution? *Genome Biology*, 2007.

[46] Qingyu Chen, Justin Zobel, and Karin Verspoor. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database*, 2017:baw163, 2017.

[47] José Fabián Reyes Román, David Roldán Martínez, Alberto García Simón, Urko Rueda, and Óscar Pastor. VarSearch: Annotating Variations using an e-Genomics Framework. In *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering*, pages 328–334. SCITEPRESS - Science and Technology Publications, 2018.

[48] Vanessa Solís Cabrera. Exploración de Bases de Datos Genómicas Dirigida por Modelos Conceptuales (Unpublished masther's thesis), 2018. Universitat Politècnica de València, Valencia, Spain.

[49] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 2011.

[50] Carlos Iñiguez-Jarrín, José Ignacio Panach, and Oscar Pastor López. Defining interaction design patterns to extract knowledge from big data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[51] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 2009.

[52] Ana Leon Palacio and Oscar Pastor Lopez. From Big Data to Smart Data: A Genomic Information Systems Perspective. In *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, pages 1–11. IEEE, 2018.

[53] M. Dawn Teare and Jennifer H. Barrett. Genetic linkage studies. In *An introduction to genetic epidemiology*, pages 39–60. Bristol University Press, 2017.

[54] Cathryn M. Lewis and Jo Knight. Introduction to Genetic Association Studies. *Cold Spring Harbor Protocols*, 2012(3), mar 2012.

[55] William S. Bush and Jason H. Moore. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12), dec 2012.

[56] Adnan Custovic and Angela Simpson. What are we learning from genetic cohort studies? *Paediatric Respiratory Reviews*, 7:S90–S92, jan 2006.

[57] Eun Pyo Hong and Ji Wan Park. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics*, 2012.

[58] Magdalena Szumilas. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 2010.

[59] Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J. Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, 2008.

[60] Steven A McCarroll and David M Altshuler. Copy-number variation and association studies of human disease. *Nature Genetics*, 39:S37, jun 2007.

[61] Bethesda (MD): National Center for Biotechnology Information (US). PubMed Help, Accessed January 2019. Available from: www.ncbi.nlm.nih.gov/books/NBK3827/.

[62] P.A. Kitts, Deanna M. Church, F. Thibaud-Nissen, J. Choi, V. Hem, V. Sapojnikov, R.G. Smith, T. Tatusova, C. Xiang, A. Zherikov, M. DiCuccio, T.D. Murphy, K.D. Pruitt, and A. Kimchi. Assembly: a resource for assembled genomes at ncbi. *Nucleic Acids Research*, 44(D1):D73–D80, 2016.

[63] Landrum et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018.

[64] Zerbino et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018.

[65] Kitts et al. The Database of Short Genetic Variation (dbSNP), Accessed January 2019. Available from: https://www.ncbi.nlm.nih.gov/books/NBK174586/.

[66] Murphy et al. Gene Help: Integrated Access to Genes of Genomes in the Reference Sequence Collection, Accessed January 2019. Available from: https://www.ncbi.nlm.nih.gov/books/NBK3841/.

[67] Ewan Birney and Nicole Soranzo. The end of the start for population sequencing. *Nature*, 526:52, sep 2015.

[68] National Center for Biotechnology Information. Entrez Programming Utilities Help [Internet], Accessed April 2019. https://www.ncbi.nlm.nih.gov/books/NBK25497/.

[69] Kinsella Rhoda et al. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*, 2011.

[70] HUGO Gene Nomenclature Committee. REST web-service help, Accessed April 2019. https://www.genenames.org/help/rest/.

[71] U.S. National Library of Medicine. Medical Subject Headings - Preface, Accessed March 2019. Available from: www.nlm.nih.gov/mesh/intro_preface.html#pref_rem.

[72] Richards et al. 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. *Genetics in Medicine*, 17(5):405–423, may 2015.

[73] Padhraig Gormley, Bendik S. Winsvold, Dale R. Nyholt, Mikko Kallela, Daniel I. Chasman, and Aarno Palotie. Migraine genetics: from genome-wide association studies to translational insights. *Genome Medicine*, 2016.

[74] Arn MJM van den Maagdenberg. Migraine genetics: New opportunities, new challenges. *Cephalalgia*, 36(7):601–603, 2016.

[75] Candace T. Myers and Heather C. Mefford. Advancing epilepsy genetics in the genomic era, 2015.

[76] Ingrid E. Scheffer. Epilepsy genetics revolutionizes clinical practice, 2014.

[77] Celia Sintas, Oriel Carreno, Jessica Fernandez-Morales, Pilar Cacheiro, Maria-Jesus Sobrido, Bernat Narberhaus, Patricia Pozo-Rosich, Alfons Macaya, and Bru Cormand. A replication study of a GWAS finding in migraine does not identify association in a Spanish case-control sample. *Cephalalgia : an international journal of headache*, 32(14):1076–1080, oct 2012.

[78] Anttila Verneri et al. Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. *Nature genetics*, 42(10):869–873, oct 2010.

[79] National Cancer Institute. Crohn disease, Accessed March 2019. https://ghr.nlm.nih.gov/condition/crohn-disease.

[80] Jimmy Z. Liu and Carl A. Anderson. Genetic studies of Crohn's disease: Past, present and future. *Best Practice & Research Clinical Gastroenterology*, 28(3):373–386, jun 2014.

[81] Breastcancer.org. Male Breast Cancer, Accessed March 2019. https://www.breastcancer.org/symptoms/types/male_bc.

A Conceptual Schema Description

Table A1: Conceptual Schema Description

Class	**Attribute**	**Description**
Assembly	assembly	Name of the reference sequence
	date_assembly	Date of the reference sequence
	gh_identifier	Identifier provided by the UCSC Genomic Institute
Bibliography_DB	url	URL of the bibliography data source
	name	Name of the bibliography data source
Bibliography_Reference	id	Identifier of the publication
	title	Title of the publication
	publication	Formatted citation
	authors	List of authors
	date	Publication date
Chromosome	nc_identifier	Identifier of the chromosomic sequence of reference
	name	Name of the chromosome
CNV	repetitions	Number of times a sequence of nucleotides has been repeated
Databank	name	Name of the data source
	url	URL of the data source
Databank_Version	release	Identifier of the release
	date	Release date
Deletion	ref	Sequence of nucleotides that has been deleted
Gene	ng_identifier	Identifier of the genomic sequence of reference
	start	Start of the gene in the DNA sequence
	end	End of the gene in the DNA sequence
	symbol	Symbol of the gene
	official_name	Official name of the gene
	description	Description of the gene
	synonyms	Any other names used to represent the gene
Imprecise	description	Description of the variation

Table A2: Conceptual Schema Description (Cont.)

Class	**Attribute**	**Description**
Indel	ref	Sequence of nucleotides that has been deleted
	alt	Sequence of nucleotides that has been inserted
Insertion	alt	Sequence of nucleotides that has been inserted
Inversion	alt	Sequence of nucleotides that has been inverted
Phenotype	phenotype_id	Identifier of the phenotype
	name	Name of the phenotype
Population	id	Identifier of the population
	name	Name of the population
	size	How many people conforms the studied population
Population_DB	name	Name of the database
	description	Description of the database
	url	URL of the database
Precise	start	Start of the variation in the DNA sequence
	end	End of the variation in the DNA sequence
SNP_Allele	allele	Sequence of nucleotides that conforms the allele
SNP_Allele_Population	frequency	Appearance frequency of the allele in the population
SNP_Genotype	allele1	Sequence of nucleotides in one strand of the chromosome
	allele2	Sequence of nucleotides in the other strand of the chromosome
SNP_Genotype_Population	frequency	Frequency of the genotype in the studied population
Statistical_Evidence	replicated	Wheter the study has been replicated
	sample	How many people participates in the study
	OR	Odds Ratio
	IC	Interval of Confidence
	p_value	p-value
Variation	db_variation_id	Identifier of the variation provided by the data source
	clinically_importance	Effect of the variation
	other_identifiers	HGVS expressions associated to the variation

B Mapping Rules

Table B1: Mapping Rules

Assembly(assembly,-,-)	⊇	ClinVar.assembly_set(assembly_name,-,-)
Assembly(assembly, date_assembly, hg_identifier)	⊇	Assembly(assemblyname, seqreleasedate, ucscname)
Bibliography_DB(ulr, name)	⊇	PubMed("https://www.ncbi.nlm.nih.gov/pubmed/", "PubMed")
Bibliography_Reference (id, title, publication, authors, date)	⊇	PubMed.DocSum(Id, Title, comb(Title, AuthorList, tr(PubDate), Volume, Issue, Pages, FullJournalName), comb(AuthorList), tr(DatePub))
Bibliography_Reference (id,-,-,-,-)	⊇	GWAS(PUBMEDID,-,-,-,-)
Bibliography_Reference(id,-,-,-,-)	⊇	ClinVar.IdList(Id,-,-,-,-)
Bibliography_Reference(id,-,-,-,-)	⊇	Ensembl(pmid,-,-,-,-)
Chromosome(-, name)	⊇	GWAS(-, CHR_ID)
Chromosome(-, name)	⊇	ClinVar.assembly_set(-, chr)
Chromosome(-, name)	⊇	Ensembl(-, chr_name)
Chromosome(nc_identifier, name)	⊇	dbSNP.DocSum(tr(DOCSUM), CHR)
Chromosome(nc_identifier, name)	⊇	EG.GenomicInfoType(ChrAccVer, ChrLoc)
Databank(name, url)	⊇	GWAS("GWAS Catalog", "https://www.ebi.ac.uk/gwas/home")
Databank(name, url)	⊇	ClinVar("ClinVar", "https://www.ncbi.nlm.nih.gov/clinvar/")
Databank(name, url)	⊇	Ensembl("Ensembl", "https://www.ensembl.org/index.html")
Databank(name, url)	⊇	dbSNP("dbSNP", "https://www.ncbi.nlm.nih.gov/projects/SNP/")
Databank(name, url)	⊇	HGNC("HGNC", "https://www.genenames.org/")
Databank(name, url)	⊇	Gene("Entrez Gene", "https://www.ncbi.nlm.nih.gov/gene/")
Databank(name, url)	⊇	Assembly("NCBI Assembly", "https://www.ncbi.nlm.nih.gov/assembly/")
Databank_Version(release, date)	⊇	GWAS("v1.0.2", "2018-10-29")
Databank_Version(release, date)	⊇	ClinVar(DbBuild, LastUpdate)
Databank_Version(release, date)	⊇	Ensembl("94", "October 2018")
Databank_Version(-, date)	⊇	LF(-, "March 13, 2018")
Databank_Version(release, date)	⊇	dbSNP(DbBuild, LastUpdate)
Databank_Version(-, date)	⊇	HGNC(-, lastModified)

Table B2: Mapping Rules (Cont. I)

Databank_Version(release, date)	⊇	Gene(DbBuild, LastUpdate)
Databank_Version(release, date)	⊇	Assembly(DbBuild, LastUpdate)
Deletion(ref)	⊇	ClinVar.assembly_set(ref)
Gene(-,-,-, symbol,-,-,-)	⊇	GWAS(-,-,-, MAPPED_GENE,-,-,-)
Gene(-,-,-, symbol,-,-,-)	⊇	ClinVar.gene(-,-,-, symbol,-,-,-)
Gene(-,-,-, symbol,-,-,-)	⊇	Ensembl(-,-,-, associated_gene,-,-,-)
Gene(ng_identifier,-,-, symbol,-,-,-)	⊇	dbSNP.DocSum(tr(DOCSUM),-,-, GENE,-,-,-)
Gene(-,-,-, symbol, official_name,-, synonyms)	⊇	HGNC.doc(-,-,-, symbol, name,-, com(alias_symbol))
Gene(-,-,-, symbol, official_name, description, synonyms)	⊇	EG.DocumentSummary(-,-,-,NomenclatureSymbol, NomenclatureName, Summary, OtherAliases)
Gene(-, start, end,-,-,-,-)	⊇	EG.GenomicInfoType(-, ChrStart, ChrStop,-,-,-,-)
Indel(alt)	⊇	GWAS(tr(STRONGEST SNP-RISK ALLELE))
Indel(ref, alt)	⊇	ClinVar.assembly_set(ref, alt)
Indel(-,alt)	⊇	Ensembl(-, associated_variant_risk_allele)
Insertion(alt)	⊇	GWAS(tr(STRONGEST SNP-RISK ALLELE))
Insertion(alt)	⊇	ClinVar.assembly_set(alt)
Insertion(alt)	⊇	Ensembl(associated_variant_risk_allele)
Inversion(alt)	⊇	GWAS(tr(STRONGEST SNP-RISK ALLELE))
Inversion(alt)	⊇	ClinVar.assembly_set(alt)
Inversion(alt)	⊇	Ensembl(associated_variant_risk_allele)
Phenotype(-, name)	⊇	ClinVar.trait(-, trait_name)
Phenotype(-, name)	⊇	GWAS(-, MAPPED_TRAIT)
Phenotype(-, name)	⊇	Ensembl(-, phenotype_description)
Population(id, name, size)	⊇	1G.populations(tr(population))
Population(-, name, size)	⊇	dbSNP.DocSum(-, GLOBAL_POPULATION, GLOBAL_SAMPLESIZE)
Population_DB(name,-, url)	⊇	1000Genomes(“1000 Genomes”,-, “https://www.ensembl.org/index.html”)

Table B3: Mapping Rules (Cont. II)

Precise(-, start,-)	⊇	GWAS(-, CHR_POS, -)
Precise(start, end)	⊇	ClinVar.assembly_set(start, stop)
Precise(specialization_type,-,-)	⊇	ClinVar.variation(tr(variant_type,-,-))
Precise(-, start, end)	⊇	Ensembl(-, chrom_start, chrom_end)
Precise(specialization_type, start,-)	⊇	dbSNP.DocSum(SNP_CLASS, tr(CHRPOS),-)
SNP_Allele(allele)	⊇	Ensembl(minor_allele)
SNP_Allele(allele)	⊇	1G.populations(allele)
SNP_Allele_Population(frequency)	⊇	Ensembl(minor_allele_frequency)
SNP_Allele_Population(frequency)	⊇	1G.populations(frequency)
SNP_Allele_Population(frequency)	⊇	dbSNP.DocSum(GLOBAL_MAF)
SNP_Genotype(allele1, allele2)	⊇	1G.population_genotypes(genotye)
SNP_Genotype_Population(frequency)	⊇	1G.population_genotypes(frequency)
Statistical_Evidence(replicated, sample, OR, CI, p_value)	⊇	GWAS(tr(REPLICATION SAMPLE SIZE), INITIAL SAMPLE SIZE, OR or BETA, 95% CI, P-VALUE)
Statistical_Evidence(-,-,-,-, p_value)	⊇	Ensembl(-,-,-,-, p_value)
Variation(db_variation_id,-,-,-)	⊇	GWAS(comb(tr(SNP_ID_CURRENT), STUDY ACCESSION,-,-)
Variation(db_variation_id,-,-)	⊇	ClinVar.variation(comb(measure_id,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.DocSum(tr(SNP_ID), CLINICAL_SIGNIFICANCE, DOCSUM)

C Variant Type Mapping

Table C1: Variant Type Mapping

Source Databases	**Conceptual Schema**
Alu Deletion	Deletion
Alu Insertion	Insertion
Complex Chromosomal Rearrangement	Precise
Complex Substitution	Substitution
Copy Number Gain	CNV
Copy Number Loss	Deletion
Copy Number Variation	CNV
Deletion	Deletion
Deletion/Insertion	Indel
Duplication	CNV
HERV Deletion	Deletion
Indel	Indel
Insertion	Insertion
Interchromosomal Translocation	Precise
Intrachromosomal Translocation	Precise
Intron Splice	Variation
Inversion	Inversion
LINE1 Deletion	Deletion
LINE1 Insertion	Insertion
Microsatellite	CNV
Mobile Element Deletion	Deletion

Table C2: Variant Type Mapping (Cont.)

Source Databases	**Conceptual Schema**
Mobile Element Insertion	Deletion
Monomeric Repeat	CNV
Multiple Nucleotide Polymorphism	SNP
Multiple Nucleotide Variation	Indel
No Alteration	Imprecise
Novel Sequence Insertion	Insertion
Sequence Alteration	Variation
Single Nucleotide Variant	SNP
Substitution	Indel
Substitution/Insertion	Indel
SVA Deletion	Deletion
SVA Insertion	Insertion
Tandem Duplication	CNV
Translocation	Precise

D Variants associated with Epilepsy

Table D1: Variants associated with the risk of suffering Epilepsy.

ID	Gene	Chromosome	Clinical Significance	Bib. ID/Year
rs12744221	RNF115	1	Association	23962720 (2014)
rs72698613		4	Association	23962720 (2014)
rs61670327		5	Association	23962720 (2014)
rs492146	GSTA4	6	Association	23962720 (2014)
rs72700966	PTPRD	9	Association	23962720 (2014)
rs143536437	ARHGAP11B	15	Association	23962720 (2014)
rs11861787		16	Association	23962720 (2014)
rs771390	C1orf94	1	Association	22949513 (2012)
rs13026414		2	Association	22949513 (2012)
rs11890028	SCN1A	2	Association	22949513 (2012)
rs72823592	NFE2L1	17	Association	22949513 (2012)
rs12720541	PLA2G4A	1	Association	22949513 (2012)
rs2717068		2	Association	22949513 (2012)
rs10496964	ZEB2	2	Association	22949513 (2012)
rs10030601		4	Association	22949513 (2012)
rs12059546	CHRM3	1	Association	22949513 (2012)
rs39861	MAST4	5	Association	22949513 (2012)
rs6732655	SCN1A	2	Association	25087078 (2014)
rs28498976	PCDH7	4	Association	25087078 (2014)
rs111577701	GOLIM4	3	Association	25087078 (2014)
rs535066	GABRA2	4	Association	25087078 (2014)
rs12987787	SCN1A	2	Association	25087078 (2014)
rs2947349	VRK2, FANCL	2	Association	25087078 (2014)
rs1939012	MMP8	11	Association	25087078 (2014)
rs1044352	PCDH7	4	Association	25087078 (2014)
rs55670112		5	Association	25087078 (2014)
rs7587026	SCN1A	2	Association	25087078 (2014)
rs346291		6	Association	20522523 (2010)
rs2601828	ADCY9	16	Association	20522523 (2010)
rs1490157	ZNF385D	3	Association	20522523 (2010)
rs16944	IL1B	2	Association	22160471 (2012)

E Variants associated with Crohn's Disease

Table D2: Variants associated with the risk of suffering Crohn's disease.

ID	Gene	Chromosome	Clinical Significance	Bib. ID/Year
rs1000113	IRGM	5	Association	17554300 (2007) 25191865 (2014) 19098858 (2009) 23365659 (2013)
rs1004819	IL23R	1	Association	26678098 (2015) 20066736 (2010) 19334001 (2009) 17786191 (2007) 17068223 (2006)
rs10210302	ATG16L1	2	Association	17554300 (2007)
rs10889677	IL23R	1	Association	26678098 (2015) 17786191 (2007)
rs11209026	IL23R	1	Protective	26678098 (2015) 17554300 (2007) 17068223 (2006) 17447842 (2007)
rs1128535	TRAIP	3	Association	18200509 (2008)
rs11805303	IL23R	1	Association	21253534 (2010)
rs12037606			Association	17554300 (2007)
rs13361189	IRGM	5	Association	26066377 (2015) 25526194 (2014) 25009628 (2014) 22573572 (2013) 18580884 (2008) 17554300 (2007) 17554261 (2007)
rs17221417	NOD2	16	Association	17554300 (2007)
rs17234657	intergenic	5	Association	19174780 (2009)
rs1793004	NELL1	11	Association	17684544 (2007)
rs1800629	TNF	6	Drug Response	12190096 (2002)

Table D3: Variants associated with the risk of suffering Crohn's disease. (Cont.)

ID	Gene	Chromosome	Clinical Significance	Bibliography ID
rs1992660	PTGER4	5	Association	17684544 (2007)
rs1992662	PTGER4	5	Association	17684544 (2007)
rs2076756	NOD2	16	Association	21209938 (2010) 17684544 (2007)
rs2201841	IL23R	1	Association	20066736 (2010) 17068223 (2006)
rs2241880	ATG16LI	2	Association	25738374 (2015) 25731871 (2016) 22573572 (2013) 21172187 (2007) 20066736 (2010) 19590455 (2009) 19575361 (2009) 19491842 (2009) 19337756 (2009) 17894849 (2007) 17484864 (2007) 17200669 (2007)
rs2542151	PTPN2	18	Association	24127071 (2014) 22457781 (2012) 20403149 (2010) 17554300 (2007)
rs4958847	IRGM	5	Association	24232856 (2013) 19491842 (2009) 18580884 (2008)
rs6596075	IBD5	5	Association	17554300 (2007)
rs6601764		10	Association	17554300 (2007)
rs6908425	CDKAL1	6	Association	17554300 (2007) 18587394 (2008)
rs7076156	ZNF365	10	Risk Factor	21257989 (2011)
rs7753394		6	Association	17554300 (2007)
rs7807268		7	Association	17554300 (2007)
rs8111071		19	Association	17554300 (2007)
rs9469220		6	Association	17554300 (2007)
rs10761659	IL23R	10	Association	17554300 (2007)