Design and Implementation of a Geis for the Genomic Diagnosis using the SILE Methodology. Case Study: Congenital Cataract

Manuel Navarrete-Hidalgo[1], José Fabián Reyes Román[1,2] and Óscar Pastor López[1]
[1]*PROS Research Center, Universitat Politècnica de València, Valencia, Spain*
[2]*Department of Engineering Sciences, Universidad Central del Este, San Pedro de Macorís, Dominican Republic*

Keywords: Conceptual Modeling, Preventive Diagnosis, GeIS, Precision Medicine, ETL, CMHG.

Abstract: Biomedicine and the preventive diagnosis of diseases open a series of lines of research as diverse as proposed solutions. However, the information that the humans contain within the genome represents a great challenge related to the processing and management of their biological information, whose success will depend directly on the structures that will be generated through the application of conceptual modeling techniques. In this context, this research work presents the development of a prototype of "*Extraction-Transformation-Load*", where biological information can be obtained from multiple scientific repositories that do not have direct interaction. For that reason, the *Conceptual Model of the Human Genome* (CMHG) proposed is used as a holistic representation of the domain with the aim of generating *Genomic Information Systems* (GeIS), which facilitate an efficient management of all the existing knowledge in the genome in order to enhance "*Precision Medicine*" (PM). This work defines a GeIS for the preventive diagnosis of "*congenital cataracts*", whose condition is not related to age and lifestyle, but to the genetic component of each person. In this way, we can provide an early diagnosis and possible means of personalized treatments.

1 INTRODUCTION

Since the beginning of mass sequencing, one of the challenges within the biological context has been to understand the molecular basis of organisms. However, the large amount of data collected has generated a problem with the management of this information. Although a multitude of computational tools and strategies have been developed for the evaluation of this information (Staden, 1979), there is a problem associated with the access and consultation of the data stored in the genomic repositories, because they have offered them in a *heterogeneous*, *redundant* and *dispersed* way, (elements that are part of the so-called "*Genomic Chaos*"). To advance knowledge and its application, most researchers indicate the need of an improvement in the capacity for analysis and management of that information (Solomon, 2014).

Therefore, the extraction of biological information should be automated through the development of software tools for ETL (*Extraction, Transformation and Loading*) (Muñoz et al., 2011), which should be able to consult the different repositories of genomic information and adapt the existing data according to the requirements of a conceptual model, such as the *Conceptual Model of the Human Genome* (CMHG) proposed by the PROS Research Center, whose conceptual representation of the genome opens a standard of *access*, *consultation*, *exploitation* and *generation* of preventive tools for diseases caused by genetics.

These software tools would reduce the interaction time between the human and the genomic repositories to obtain information, because in this phase more resources are invested, and more errors are made (*due to the human factor*). This would allow the direct application of genomic information in public health issues, such as in the preventive detection of cataracts.

Cataract is a disease that affects human and animal vision, mainly caused by the loss of transparency of the lens, which is located just behind the pupil (Boyd, 2016). People with cataracts suffer from blurred vision, inability to appreciate the contour of what they see, loss of color intensity and hypersensitivity to glare, as well as headaches and visual fatigue. This is produced because the lens becomes opaque by the high concentration of proteins in its cells and the development of dense

Navarrete-Hidalgo, M., Reyes Román, J. and Pastor López, Ó.
Design and Implementation of a Geis for the Genomic Diagnosis using the SILE Methodology. Case Study: Congenital Cataract.
DOI: 10.5220/0006705802670274
In *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2018)*, pages 267-274
ISBN: 978-989-758-300-1

bodies (The National Eye Institute, 2017) (*https://nei.nih.gov/*).

One of the ways to improve detection and provide prevention and/or cure mechanisms, as well as to study their direct link with other diseases, would be to analyze genomic repositories for genetic indicators for "*congenital cataract*" whose appearance is not related to the natural aging of the human being but to the genetic component of each person.

The paper is divided into the following sections: Section 2 presents the state of the art (conceptual modeling applied in genomics and cataracts). Section 3 shows a brief description of the CMHG. Section 4 contains the SILE Methodology used to the "*congenital cataract*". Finally, Section 5 presents the conclusions and outlines future work.

2 STATE OF THE ART

Bioinformatics is born from the interaction of *molecular biology* and *computer science*, with the objective of allowing biological data to be processed. This data is characterized by its large size and continuous growth, which is why it is necessary to develop tools to manage the information in an agile and efficient way, as well as new algorithms and statistical solutions oriented to the analysis of the DNA sequences and their variations. In this sense, there are several standards and formats for the representation of nucleotide and protein sequences, programs of comparison of variations and frameworks for the exploration of genetic diseases, as well as databases with all genome sequences, nucleotides, proteins, proteins structures, human genetic diseases, and bibliography available for public study and research.

The main biological databases including sequence data were created in the USA, EU and Japan (see Table 1). Their beginnings date from 1971 and continue to the present day.

On the other hand, all this biological information must be perfectly structured and bounded in an *Information System* (IS) that allows its treatment and exploitation in an efficient way. Understanding the genome is a complex task, and generating a correct conceptual definition (Olivé, 2007) that addresses all current genomic knowledge is essential to understand the -*genomic*- domain. In addition, this conceptual model must be subject to constant evolutions product of new discoveries in the context.

Table 1: Most Popular Databases.

Name	*DB type*	*Source*	*Start*
SNPedia	Polymorphisms	USA	2006
BioCyc	Metabolic pathways	USA	2005
Reactome	Metabolic pathways	EU	2004
Ensembl	Genomics	EU	2000
UCSC	Genomics	USA	2000
dbSNP	Polymorphisms	USA	1998
PubMed	Bibliographical	USA	1996
KEGG	Metabolic pathways	Japan	1995
OMIN	Genetic diseases	USA	1995
EMBL	Nucleotides	EU	1992
DDBJ	Nucleotides	Japan	1986
Uniprot	Proteins	EU	1986
GenBank	Nucleotides	USA	1982
PDB	Proteins	USA	1971

2.1 Conceptual Modeling in Genomics

In the field of genomic bioinformatics, the first works of conceptual modeling were given by Paton (Paton et al., 2000). His essays were supported by previous work on modeling of protein structures (Gray et al., 1990), being a pioneer in the conceptual design of the eukaryotic cell, its genomic organization, transcriptome, proteome and metabolome modeling, among others; using the unified modeling language -*UML*- (*http://www.uml.org/*). In addition, Ram et al. (2004) presents a conceptual modeling approach applied to the protein context, this paper states that the use of conceptual modeling facilitates the representation without semantic loss and comparison and search operations in complex structures, such as in the modeling of the three-dimensional structure of proteins, characterized by the large volume of data.

With these bases, the genome group of the *PROS Research Center*, of the Universitat Politècnica de València, started in 2008 a line of research focused on the modeling of the human genome for analysis and study the most basic expression, genes, and their mutations within a chromosomal segment (Pastor, 2008). However, this model does not consider the existence of processes such as the regulation or coding of a single protein by two different genes, as well as the combined action of multiple genes. As detailed in the work of 2010 (Pastor et al., 2010), this version of the conceptual model of the human genome is called the "*essential*" model and is composed of three views:

- *Genome* View: models human genomes.
- *Gene-Mutation* View: models the entities "*Gene*" and "*Allele*", and the knowledge of their structures.

- *Transcription* View: models transcription processes.

After this, the schema is extended with a fourth view called "*Phenotype View*" that allows for the phenotypic representation, that is, the concrete/visible manifestation of a genotype in a given context. Subsequently, the CMHG migrated from modeling oriented to the study of genes to another centered on the concept of the chromosome (Reyes et al., 2016). This new focusing is called CMHG v2 (Reyes et al., 2017).

2.2 Cataract

Cataract is the most frequent pathology of the lens and is the most common cause of reversible blindness, and produces a series of disabilities as explained in Section 1. According to the location of the opacity, several types of cataract are distinguished: *nuclear*, *cortical* and *subcapsular*. In nuclear cataract, opacity lies in the central part of the lens, while in the cortical and subcapsular the opacity lies in the periphery of the lens and in the central area of the posterior aspect of the lens respectively.

Currently, the only effective treatment for cataract is surgical intervention, the most recent and least invasive being *Phacoemulsification* (Emulsifying and aspirating the nucleus of the lens with a high-frequency ultrasonic needle), followed by the *Small Incision Cataract Surgery* (Reinstein et al., 2014). Others like the *Extracapsular Cataract Extraction* and the *Intracapsular Cataract Extraction* (Kirchhof, 2017) are more complex, with slower recovery times, sutures and higher induced injuries. In all of them the lens is removed, and an intraocular lens is placed. At the biological level, the appearance of cataracts is associated with certain common cellular mechanisms (Jobling et al., 2002) (Michael et. al., 2011), which has allowed to know which habits and risk factors associated with cataract development (Rakel, 2014), such as exposure to sunlight and UV, stress, tobacco, excess of alcohol, malnutrition, vitamin deficiencies, obesity, dehydration, chronic diseases, medications, as well as metabolic disorders such as galactosemia and diabetes, plus genetic and hereditary reasons. The latter scenario is centered in the case study, the hereditary *congenital cataract*, which includes four types (Hejtmancik, 2008):

- *Autosomal recessive*: Two copies of the allele are needed to develop the disease.
- *Autosomal dominant*: Located on non-sexual chromosomes, with a single copy of the allele responsible for developing the disease.
- *Sporadic*: It is due to a spontaneous mutation that affected all the cells of the individual including the sexual ones, provoking predisposition to suffer the disease and being heritable, although their parents did not suffer.
- *X-linked:* It is due to an allele predisposed to suffer from the disease located on the sex chromosome X, which causes in women it depends on the dominance or recessiveness of the allele.

3 CONCEPTUAL MODEL OF THE HUMAN GENOME

The *Conceptual Model of the Human Genome* (CMHG) version 2 is composed of six independent but related views (Pastor et al., 2016). These views are *structural*, *transcription, variations*, *phenotype*, *pathways* and *bibliography references*. Once the CMHG has been defined, it is necessary to have an entity that allows the storage and access to this information in a fast and structured way. These operations are guaranteed by a relational database schema called "*Human Genome Database*" (HGDB), which currently models the *structural view*, *variations*, *bibliography references*, and *validations* through tables. For more information, see the full view and description in (Reyes et al., 2016). This model remains in constant study, so it is necessary for the CMGH to continue evolving according to new discoveries and non-contemplated genetic structures, such as haplotypes support (Reyes et al., 2016), as well as the study of quality metrics to ensure the best definition of the domain.

4 SILE METHODOLOGY

SILE (*Search-Identification-Load-Exploitation*) is a methodology whose objective is to perform a load of selective information with "*curated data*", given that the existing information is *dispersed*, *heterogeneous* and often *redundant*. Performing a massive data load would suppose managing and working with invalid or obsolete information. For this, the SILE methodology is composed of four stages. In the first place, a *search* of genes is made in the genomic repositories most used by the scientific community. In the next stage, an accurate *identification* and

validation of genes and variations associated with the genetic disease of study is performed. To do this, there is a group of experts in areas of *molecular biology*, *biomedicine* and/or *biotechnology* coming from different hospital centers with which the PROS Center has collaborations, such as, the *Hospital Universitari i Politècnic La Fe*, the *INCLIVA*, as well as companies specializing in genetic studies, such as *IMEGEN* and *tellmeGen*. Then, the data obtained are *loaded* into the HGDB and finally *exploited* using the genomic framework called "*VarSearch*".

4.1 Search

First, it is necessary to search all genes directly or indirectly associated with congenital cataracts (Table 2). These data were obtained from two reviews (Hejtmancik, 2008), (Shiels & Hejtmancik, 2013) and various research articles (Cobb et al., 2000; Fu & Liang, 2002; Faiyaz-Ul-Haque et al., 2007; Richter et al., 2008; Sagona et at., 2014; Javadiyan et al., 2016).

Table 2: Genes associated with congenital cataract.

Gene	*Official name*
BCOR	BCL6 corepressor
BFSP2	Beaded filament structural protein 2
CHMP4B	Charged multivesicular body protein 4B
CRYAA	Crystallin alpha A
CRYAB	Crystallin alpha B
CRYGC	Crystallin gamma C
CRYBB1	Crystallin beta B1
CRYBB2	Crystallin beta B2
CRYBB3	Crystallin beta B3
CRYBA4	Crystallin beta A4
CRYGD	Crystallin gamma D
CRYGS	Crystallin gamma S
GCNT2	Glucosaminyl (N-acetyl) transferase 2
GJA3	Gap junction protein alpha 3
GJA8	Gap junction protein alpha 8
HSF4	Heat shock transcription factor 4
LIM2	Lens intrinsic membrane protein 2
MAF	MAF bZIP transcription factor
NHS	NHS actin remodeling regulator
MIP	Major intrinsic protein of lens fiber
PITX3	Paired like homeodomain 3
VSX2	Visual system homeobox 2

In addition, there are studies that highlight the clinical utility of the genomic variations, indicating that the detection rate of mutations in affected patients is 70% using massive sequencing techniques. This indicates that the genomic studies would be useful for early detection of the disease and improving clinical advice, as they provide significant additional diagnostic information, allowing the existence of a personalized medicine at the genomic level (Ding et al., 2017). Jointly, *the Eye Genetics Research Group of Children's Medical Research Institute* (CMRI, 2017) is expanding the genomic knowledge of congenital cataracts, which offers opportunities to analyze this information.

4.2 Identification

Once the cataract related genes are identified, we proceed to the identification of variants, location and cataract types (AD, *autosomal dominant*; AR, *autosomal recessive*; S, *sporadic*; XL, *X-linked*) that have been recently studied, covering a timespan from 2014 to 2016 (Ma. et al., 2016) (Table 3):

Table 3: Variations associated with congenital cataract.

Gene	*Chromosome*	*Type*	*Variation-SNP*
BCOR	Xp11.4	X	rs864309680 rs864309702
CRYAA	21q22.3	AD	rs864309685 rs397515625
CRYAB	11q23.1	AD	rs144451841
CRYGC	2q33.3	AD	rs587778872
CRYBB1	22q12.1	E/AD	rs864309682
CRYBB2	22q11.23	E/AD	rs864309683
GJA3	13q12.11	AD	rs864309687 rs864309691 rs864309694
GJA8	1q21.2	AD E/AD AD/AR E/AD	rs80358205 rs864309677 rs864309684 rs864309688 rs864309703
MAF	16q23.2	AD	rs864309678 rs786205222 rs864309692 rs864309695
MIP	12q13.3	E/AD	rs864309693
NHS	Xp22.13	AD XL/AD	rs864309679 rs111534978

It should be noted that identification allows us to delimit all existing knowledge, and opens an important aspect about what information is obsolete or very relevant.

4.3 Load

The loading step of the data obtained through the application of the SILE methodology in the human genome database is composed of three subprocesses called *extraction*, *transformation* and *loading*. These

subprocesses are part of the ETL concept (*Extraction-Transformation-Load*) that allows and guarantees the extraction of information from different data sources for analysis, conversion and compliance with the constraints of the target system, i.e., the HGDB, like shows the Figure 2.

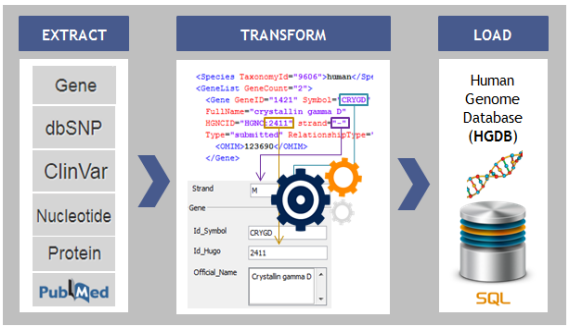

Figure 1: ETL Process.

The table structure of the HGDB currently accommodates the *structural*, *variations*, *bibliography references*, and *validations* view, which can be consulted or reviewed in the following work (Navarrete-Hidalgo, 2017).

The origins of the information, used for this work were the biological repositories of the *National Center for Biotechnology Information* (https://www.ncbi.nlm.nih.gov/). These are: *Gene,* offers all the information associated with the genes. *dbSNP*, compiles simple nucleotide polymorphisms among several species. *ClinVar*, shows information on variations, types, nucleotide changes, etc. *Nucleotide*, links genome sequences and data to genes and their transcripts. *Pubmed,* database of research articles (Gibney et. al., 2011). Although each source contains very specific information and encourages a trivial data location, the characteristics of the views require the need to consult several repositories and be able to relate them in an agile and concurrent way.

The programming language used for application development and ETL subprocesses has been *Java* in version 8.144 (Oracle Corporation, 2017). The choice of this language is due to its extensive documentation, libraries, as well as its robustness, security and portability, as it is independent of the architecture and platform. On the other hand, the output format of the repositories is the meta-tagged language (*XML, https://www.w3.org/XML*), Java has an XML processing API called SAX (Oracle Corporation, 2017), free access and event-based. That is, SAX (*Simple API for XML*) parses the document sequentially and as it locates a start or end tag launches an event to read the attributes of the same, or perform another action. SAX is characterized by having reduced memory consumption and is oriented to reading large XML files.

4.3.1 ETL (Extraction)

The extraction is the first stage of the ETL process and its main function is to extract the data from the different sources of information, which may be local repositories or located on the Internet. In addition to accessing them, the extraction stage must analyze them, check their structure, and in case of not complying with the necessary requirements, reject them, since during this phase the data are converted to a previous format that serves as intermediate for stage of transformation. In this sense, the origin of the information required comes from the biological and scientific repositories of NCBI: *dbSNP*, *ClinVar*, *Gene*, *Nucleotide*, *Protein* and *Pubmed*.

ClinVar is consulted to obtain the detail of the variation, that is to say, the change of nucleotide, its chromosome position, as well as the type of variation carried out. To obtain gene information, such as his official name, abbreviation, description, synonyms, the *Gene* repository is queried. Thirdly, to obtain the information of the proteins resulting from the variation, the *Protein* repository is consulted, from which it's official name and its amino acid sequence are obtained. On the other hand, to obtain the chromosome sequences and the transcribed elements of the variation, *Nucleotide* is consulted. Finally, *PubMed* is consulted, from which all the bibliographical sources and research articles are obtained, being the views documented by the repositories. However, there is a problem of access and consultation to these repositories that the present work intends to face. First, the relationship and consultation between them is not direct.

In addition, access to each one is done individually, sequentially and with complex parameterized queries from which the information is extracted in XML format. The XML information of each repository has a particular structure, repetitive information in different elements and tags that need to be analyzed to discard those which are less descriptive. All the extracted information has to be processed later and transformed to obtain the identifiers and arguments of consult, which causes that the stage of extraction of some repositories does not begin until the transformation finished, and both the identifiers and the rest of attributes are obtained. The query consists of a URL composed of four elements that allow specific arguments (Table 4).

Table 4: Arguments allowed in the query URL.

Argument	*Parameter*	*Description*
db	gene snp pubmed nuccore protein	Selection of the database to consult.
id	Identifier	Identifier of the item to consult.
retmode	xml txt json	Output format of the data. By default, is XML.

In the case of *ClinVar*, it is important to note that the construction of the URL varies slightly. While access to *dbSNP* is performed with the unique identifier of the variation, the step to *ClinVar* may result in one or more queries with one or more identifiers depending on whether the variation makes one or more nucleotide changes.

To know these identifiers, an intermediate query must be made in which the body of the previous URL is maintained, but with new arguments as: *dbfrom*: Source database (SNP). *db*: Target database (*ClinVar*). *id*: dbSNP variant identifier. This query generates an XML with ClinVar variations inside the tags <*Link*> <*Id*> and </*Id*> </*Link*>. Also, it shows the query parameters used, the source database (*dbSNP*), the identifier and the query database (*ClinVar*). Once the *ClinVar* identifiers are obtained the next step is to formulate the URL with the parameters required for the database, whose structure is: *Body URL,* gives access to fetch (NCBI Explorer tool). *db*: Selection of the database to be consulted (*ClinVar*). *rettype*: Record type returned (*Variation*). *id*: ClinVar variant identifier.

The result of this query contains all the variation data in XML format, which must be processed in the next step in order to adapt the information to the requirements of the destination information system, i.e., the Human Genome Database (HGDB).

4.3.2 ETL (Transform)

Once the information is extracted from the different genomic repositories, the transformation stage is responsible for the processing and validation of the data, in order for the information to comply with the requirements and rules of the target repository system. For this, changes of format and codification of values are performed, as well as obtaining of data from other repositories, generation of new identifiers, combination or division of attributes.

The following describes some of the transformations carried out to comply with the requirements of the CMGH and allow the successful loading of the data in the HGDB:

a) The CMHG contemplates four types of possible variations: *insertion*, *deletion*, *indel* and *inversion*. In contrast, *ClinVar* represents ten.
b) The identifier that NCBI assigns to the transcripts and genes is obtained by processing the "*Others_identifiers*" attribute of the "*Variation*" class, eliminating the change produced by the variation in each context.
c) *ClinVar* represents the strand by means of the "-" and "+" signs, whereas in the CMHG the characters "M" are defined from minus to "-" and "P" from plus to "+".

4.3.3 ETL (Load)

The data loading is the last stage of the ETL process (and the SILE "*Load*" stage). It should be noted that the loading stage introduces all extracted and transformed data into the destination repository (HGDB). The information to enter is unique and does not allow the existence of duplicates, which is why before carrying out any load the genome version (i.e., *GRCh build 38*) must be checked and new studies regarding the variation must be checked. In case the information stored in the HGDB is the same as that extracted from the different repositories, the load will not be carried out.

The prototype developed and one of the main contributions of this research work begins with the establishment of the connection to the database. For this, the prototype has a section of management of connections with which direct communication with the server is verified. Once established and validated, the application reports on the state of the connection and allows the information load extraction process (from the genomic repositories: *ClinVar, dbSNP, Gene, Nucleotide and Pudmeb*).

If the information extracted from the repositories is useful and if it is desired to be entered into the database for subsequent *exploitation*, prior to insertion, the application performs a duplication check, and later, regardless of whether it performs the validation, introduces all the biological information obtained in the HGDB.

4.4 Exploitation

The last stage of the SILE methodology is the exploitation of the data cured through the *VarSearch* genomic framework, whose function is to obtain knowledge by processing the information contained in patient samples (provided by VCF or Sanger files,

which represent and store variations), and that available one in the database of the human genome.

The web-based "*VarSearch*" genomic framework starts with the selection and processing of the VCF study file, which will contain the information of the variations organized into blocks, representing the position in which the variation occurs, the chromosome, the unique identifier of the variation, as well as the reference alleles and other biological data. Once the analysis of the VCF file is finished, the application shows two types of results: a) *variations found*, which indicate the relationship with the cataract; and b) *variations not found*, that is, those found in the study VCF file (*sample*), but whose information is not supported / stored by the *Genomic Information System* (GeIS), and therefore, are not related to cataracts. The detail of the exploitation process can be consulted in (Navarrete-Hidalgo, 2017).

However, the GeIS and the knowledge obtained does not have a direct application in the population, since the genomic framework acts as an interface between the HGDB and domain experts (i.e., *geneticists*, *clinical laboratories*, *biologists*, among others). Its application to end-users is intended to be facilitated through what are known as "*direct genetic tests to the consumer*", and for them a parallel project called "*GenesLove.Me*" has been developed (Reyes et al., 2017). The purpose of this web service is to offer various tests for genetically based conditions, such as: *androgenic alopecia*, *lactose intolerance*, *alcohol sensitivity* or *dupuytren*'s disease (see in detail in *http://geneslove.me*).

5 CONCLUSIONS AND FUTURE WORK

The design and implementation of a GeIS for genomic diagnosis has been carried out using the SILE methodology. In this work, a large part of the processes has been automated, and in a greater proportion the loading stage. This prototype ETL has worked well for the case study, *congenital cataract*. It has been shown that the developed ETL application works as a tool for obtaining biological information from multiple repositories from a single input parameter, which reduces human interaction with data sources from hours to seconds. To obtain the biological information that documents the variations different scientific repositories of the NCBI have been studied, such as *Gene, ClinVar, dbSNP, Nucleotide, Protein and Pubmed*. In this sense, there would be a line of improvement regarding the exploration of new data sources and their standardization.

The most important future work would be to provide the application with mechanisms for automatic updating of stored information. This would allow a direct comparison between the existing information and that obtained from the repositories, which are continuously updated with new biological data, as well as the selective loading of information depending on its version, veracity and biomedical utility. On the other hand, the application allows for a possible conversion of desktop tool to web tool, thus allowing its use from place, device and platform.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the PROS Research Centre Genome group for the fruitful discussions regarding the application of CM in the medicine field. In addition, we would like to thank Fernando Cervera, Rubén Casatejada and Mariano Collantes as experts in Biology for their contribution to this research. This work has been supported by the Generalitat Valenciana through project IDEO (PROMETEOII/2014/039) and the MICINN through project DataME (ref: TIN2016-80811-P).

REFERENCES

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research, 6(7)*, 2601–2610.

Muñoz, L., Mazón, J. N., Trujillo, J., Munoz, L., & Mazon, J.-N. (2011). ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study. *IEEE Latin America Transactions*, 9(3), 360–365.

Solomon, B. D. (2014). Obstacles and opportunities for the future of genomic medicine. *Molecular Genetics & Genomic Medicine, 2(3)*, 205–209.

Boyd, Kierstan. (2016) What Are Cataracts? American Academy of Ophthalmology. https://www.aao.org/eye-health/diseases/what-are-cataracts. [On line; accessed 5-September-2017].

The National Eye Institute (NEI). Facts About Cataract. https://nei.nih.gov/health/cataract/cataract_facts. [On line; accessed 15-May-2017].

Olivé, A. (2007). Conceptual modeling of information systems. *Springer-Verlag. Berlin Heidelberg*.

Paton, N. W., Khan, S. A., Hayes, A., Moussouni, F., Brass, A., Eilbeck, K., Goble, C. A., Hubbard, S. J. &

Oliver, S. G. (2000) "Conceptual modelling of genomic information," *Bioinformatics, vol. 16*, p. 548.

Gray, P. M., Paton, N. W., Kemp, G. J., & Fothergill, J. E. (1990). An object-oriented database for protein structure analysis. *Protein Eng,* 3(4), 235–243.

Ram, S. & Wei, W. (2004). Modeling the semantics of 3D protein structures. ER2004, 696-708.

Pastor, O. (2008). Conceptual modeling meets the human genome. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 5231 LNCS*, pp. 1–11).

Pastor, O., Levin, A., Celma-Giménez, M., Casamayor, J. C., Eraso, L., Villanueva, M. J. & Perez-Alonso, M. (2010). Enforcing Conceptual Modeling to Improve the Understanding of Human Genome. *4th. RCIS Proceedings 2010*, 85-92.

Reyes Román J. F., Pastor Ó., Casamayor J.C., Valverde F. (2016). Applying Conceptual Modeling to Better Understand the Human Genome. In: Comyn-Wattiau I., Tanaka K., Song IY., Yamamoto S., Saeki M. (eds) *Conceptual Modeling. ER 2016. LNCS, vol 9974, 404-412. Springer, Cham*.

Reyes Román J. F., León Palacio A. & Pastor López Ó. (2017). 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*. ISBN 978-989-758-250-9, 301-307.

Reinstein, D. Z., Archer, T. J. & Gobbe, M. (2014). Small incision lenticule extraction (SMILE) history, fundamentals of a new refractive surgery technique and clinical outcomes. *Eye and Vision*, 1, 3.

Kirchhof, B. Graefes Arch Clin Exp Ophthalmol (2017) 255: 1685.

Jobling, A. I. & Augusteyn, R. C. (2002). What causes steroid cataracts? A review of steroid□induced posterior subcapsular cataracts. *Clinical and experimental optometry*, 85(2), 61-75.

Michael, R. & Bron, A. J. (2011). The ageing lens and cataract: a model of normal and pathological ageing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1568), 1278-1292.

Rakel, D. Medicina integrativa (2ª edición).

Hejtmancik, J. F. (2008, April). Congenital cataracts and their molecular genetics. *Seminars in Cell and Developmental Biology*.

Pastor, O., Reyes, J.F., Valverde, F. (2016). Conceptual Schema of the Human Genome (CSHG). *Tech. Rep*.

Reyes Román, J.F., Pastor López, O., Roldán Martínez, D. & Valverde Giromé, F. (2016). How to deal with Haplotype data: An Extension to the Conceptual Schema of the Human Genome. *CLEI Electronic Journal.* 19(3):1-21.

Shiels, A., & Hejtmancik, J. F. (2013, August). Genetics of human cataract. *Clinical Genetics*.

Cobb, B. A., & Petrash, J. M. (2000). Structural and functional changes in the ??A-crystallin R116C mutant in hereditary cataracts. *Biochemistry*, 39(51), 15791–15798.

Fu, L. & Liang, J. J. N. (2002). Detection of protein-protein interactions among lens crystallins in a mammalian two-hybrid system assay. *Journal of Biological Chemistry*, 277(6), 4255–4260.

Faiyaz-Ul-Haque, M., Zaidi, S. H. E., Al-Mureikhi, M. S., Peltekova, I., Tsui, L. C. & Teebi, A. S. (2007, August). Mutations in the CHX10 gene in non-syndromic microphthalmia/anophthalmia patients from Qatar [5]. *Clinical Genetics*.

Richter, L., Flodman, P., Barria von-Bischhoffshausen, F., Burch, D., Brown, S., Nguyen, L., Turner, J., Spence, M. A. & Bateman, J. B. (2008), Clinical variability of autosomal dominant cataract, microcornea and corneal opacity and novel mutation in the alpha A crystallin gene (CRYAA). *Am. J. Med. Genet*., 146A: 833–842.

Sagona, A. P., Nezis, I. P. & Stenmark, H. Association of CHMP4B and Autophagy with Micronuclei: Implications for Cataract Formation. *BioMed Research International.* 2014.

Javadiyan, S., Craig, J. E., Souzeau, E., Sharma, S., Lower, K. M., Pater, J. & Burdon, K. P. (2016). Recurrent mutation in the crystallin alpha A gene associated with inherited paediatric cataract. *BMC Research Notes*, 9, 83.

Ding, X.-X., Zhu, Q.-G., Zhang, S.-M., Guan, L., Li, T., Zhang, L. & Zhang, H.-Q. (2017). Precision medicine for hepatocellular carcinoma: driver mutations and targeted therapy. *Oncotarget*, 8(33), 55715–55730.

Eye Genetics Research Group (CRMI). http://www.cmri.org.au/Research/Research-Units/Eye-Genetics. [On line; accessed 5-May-2017].

Ma, A. S., Grigg, J. R., Ho, G., Prokudin, I., Farnsworth, E., Holman, K., …& Jamieson, R. V. (2016). Sporadic and Familial Congenital Cataracts: Mutational Spectrum and New Diagnoses Using Next-Generation Sequencing. *Human Mutation*, 37(4), 371–384.

Navarrete-Hidalgo, M. (2017). Diseño e Implementación de un Sistema de Información Genómico para el Diagnóstico de la Catarata Congénita utilizando la Metodología SILE.

Gibney, G., & Baxevanis, A. D. (2011). Searching NCBI databases using Entrez. *Current protocols in human genetics*, 6-10.

GeneLoves.Me. http://geneslove.me/. [On line; accessed 15-May-2017].

Reyes R., J. F., Iñiguez, C. & Pastor, O. (2017). GenesLove.Me: A Model-based Web-application for Direct-to-consumer Genetic Tests. *Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering*. ISBN: 978-989-758-250-9, DOI: 10.5220/0006340201330143, 133-143.