CLEI ELECTRONIC JOURNAL, VOLUME 19, NUMBER 3, PAPER 2, DECEMBER 2016

How to deal with Haplotype data: An Extension to the Conceptual Schema of the Human Genome

José Fabián Reyes Román[1, 2], Óscar Pastor[1], Francisco Valverde[1, 3] and David Roldán M.[1]
[1] PROS Research Center - Universitat Politècnica de València.
Valencia, Spain, 46022
[2] Department of Engineering Sciences - Universidad Central del Este (UCE).
San Pedro de Macorís, Dominican Republic, 21000
[3] Escola Tècnica Superior d'Enginyeria, Departament d'Informàtica,
Universitat de València. Valencia, Spain, 46100
{jreyes | opastor | fvalverde}@pros.upv.es, darolmar@upvnet.upv.es

Abstract

The goal of this work is to describe the advantages of the application of *Conceptual Modeling* (CM) in complex domains, such as genomics. Nowadays, the study and comprehension of the human genome is a major challenge due to its high level of complexity. The constant evolution in the genomic domain contributes to the generation of ever larger amounts of new data, which means that if we do not manage it correctly data quality could be compromised (*i.e., problems related with heterogeneity and inconsistent data*). In this paper, we propose the use of a *Conceptual Schema of the Human Genome* (CSHG), designed to understand and improve our ontological commitment to the domain and also extend (enrich) this schema with the integration of a novel concept: *Haplotypes*. Our focus is on improving the understanding of the relationship between genotype and phenotype, since new findings show that this question is more complex than was originally thought. Here we present the first steps in our data management approach with haplotypes (*variations*, *frequencies* and *populations*) and discuss the database evolution to support this data. Each new version in our *conceptual schema* (CS) introduces changes to the underlying database structure that has essential and practical implications for better understanding and managing the relevant information. A solution based on conceptual models gives a clear definition of the domain with direct implications in the medical field (*Precision Medicine*), in which *Genomic Information Systems* (GeIS) play a very important role.

Keywords: haplotypes, conceptual modeling, GeIS, statistical models, genetic diagnosis.

1 Introduction

As the application of *Next Generation Sequencing* (NGS) technologies contributes to the generation of ever larger amounts of new data, to take full advantage of all this new knowledge we need to build structures to *organize*, *process* and *use* it in order to improve our understanding of the human genome.

Previous studies [1-3] have shown how conceptual models allow us to provide a definition of a domain, so that we can understand the entities involved and their relationship. These studies focused on describing the genome of eukaryotic cells and the interaction between proteins, transcriptome, and other genetic components. Other studies focus specially on proteins by Ram [23]. Pastor *et. al.* describes the Conceptual Schema of the Human Genome (CSHG) [4], [24]. However, this conceptual schema requires to be constantly aligned with the new genomic knowledge and in this paper we extend the aforementioned model to include the specification of haplotypes (which are defined in Section 2). For haplotypes, this model should be extended in two ways: 1) integrating treatment of haplotypes, 2) application of statistical models.

In this context, the goal of the present study, which is based on our previous work [48], is to extend the Conceptual Schema of the Human Genome (CSHG) by including the concepts of haplotypes and statistical models, thus improving the schema's expressiveness. This way, we foresee the creation of a powerful and reliable *Genomic Information System* based on this holistic conceptual schema (CS).

The advances over our previous work [48] are:

- The description of the preliminary steps achieved in the treatment of haplotype data: *e.g.*, the study of repositories, data collection and analysis, data loading and some test queries, and
- The explanation of how the different representations of genomic knowledge affect the structure of the underlying database that is used to manage all the data. This scenario shows how essential it is to have a sound understanding of the relevant information to achieve efficient data management policies.

This paper is divided as follows: Section 2 gives the background to this work and defines the concept of haplotypes. Section 3 explains the research methodology used. Section 4 reviews related work on how haplotype information is currently structured and stored in leading genomic data repositories. Section 5 contains our proposal for the conceptual modeling of haplotypes. Section 6 describes the conceptual alignment that we suggest to formalize haplotype information. Section 7 describes the first steps in haplotype data treatment. Section 8 contains a discussion on database evolution according to the Conceptual Schema of the Human Genome (CSHG). Finally, section 9 presents the lessons learned and outlines future work.

2 Background: Understanding the Haplotype Concept -test case: Alcohol Sensitivity-

We detected the importance of including haplotype treatment in our CSHG on the genetic implications for the pathology of *Alcohol Sensitivity*, in which we did an intensive study of genes and variants that were associated with a predisposition to this disease [5-6].

Alcohol sensitivity occurs when an individual ingests a certain amount of alcohol causing immediate rejection and experiencing discomfort, dizziness and other symptoms. The simple fact of consuming thus causes discomfort and has a future impact on the health of the individual [11], [58]. We have taken this disease as a test case because it occurs in the population as a whole, regardless of social status, age or culture [8].

For the study of these genetic alterations we obtained a lot of information from different repositories of genomic data (such as NCBI: dbSNP [57], PubMed [59], and others). We defined a group of genes [6] closely linked to the disease, including:

Table 1: List of genes and variants associated with alcohol sensitivity

GENE	*OFFICIAL FULL NAME*	*SNP / VARIANTS*
ADHIB	alcohol dehydrogenase 1B (class I), beta polypeptide	rs671; rs1229982; rs1229984; rs1230025
ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	rs671; rs7590720; rs1800497
GABRA2	gamma-aminobutyric acid (GABA) A receptor, alpha 2	rs279836; rs279858; rs279871
PECR	peroxisomal trans-2-enoyl-CoA reductase	rs7590720
PKNOX2	PBX/knotted 1 homeobox 2	rs1426153; rs750338; rs585977; rs10893366
SLC22A18	solute carrier family 22, member 18	rs16928809
DRD2	DRD2 dopamine receptor D2	rs1076560; rs1800497; rs6276

These genes are directly or indirectly associated with alcohol sensitivity. The genes with a direct influence present a high the predisposition to this disease, and the indirect are in a more general perspective (because some are associated with addictive diseases).

After completing the search and identification processes [38], we turned to the medical validation of genes, with the help of our fellow biologists (supporting this validation with papers from high-impact medical journals). We proceed to filter the genes in Table 1, selecting only three relevant genes: (a) ALDH2 (*rs671*); (B) GABRA2 (*rs279836, rs279871*) and (c) ADH1B (*rs1229982, rs1229984*).

We detected a "*haplotype case*" with the GABRA2 gene, in which we found a haplotype composed of three variants (*variations*): rs279871, rs279836 and rs279845 [38]. Initially we worked with individual variations[1], without considering the relationships between them and without considering the rs279845 variant [11-12]).

[1] Variation (*or variants*): naturally occurring genetic differences among organisms in the same species [*Scitable by Nature Edu.*].

The rs279871 form a haplotype with rs279836 and rs279845

rs279871	RefSNP Alleles: A/G	Ancestral Allele: A
rs279836	RefSNP Alleles: A/T	Ancestral Allele: T
rs279845	RefSNP Alleles: A/T	Ancestral Allele: A

From a biological point of view, Haplotypes are a set of SNPs[2] that are inherited and found together in a chromosome and are defined as a group of SNPs of a gene that are very close and tend to be inherited together. This means that a haplotype of the alleles is not separated in the recombination phase and can be transmitted in blocks, allowing combinations of variants to a gene that affects certain phenotypes [7].

There is currently a significant set of genetic diseases in which the influence of haplotypes has been well established, such as breast cancer [9-10] and alcohol sensitivity [11-12], among others [13-14]. If haplotypes are considered in the diagnosis, the outcome can be improved, as the assessed probabilities indicate the level of risk in genomic diagnosis more accurately.

After making this discovery we realized that our conceptual schema should be improved to achieve an integration of information in a more appropriate manner. If we treat this concept effectively, we can improve the results generated in the genetic diagnoses.

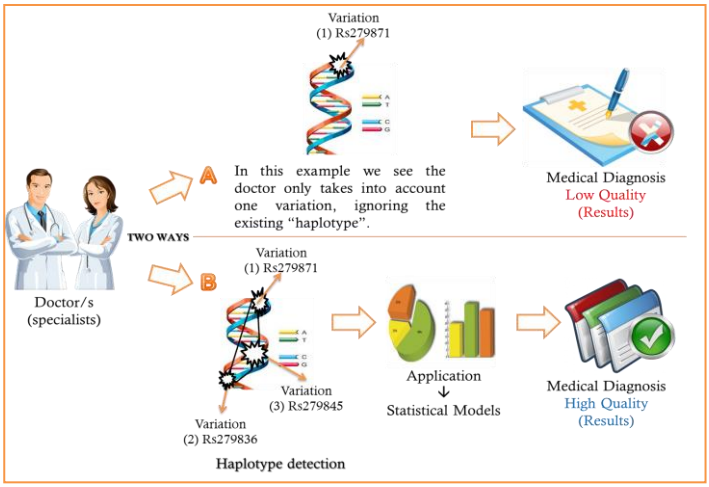

Figure 1: Genetic analysis using "*variants*" versus "*variants + haplotypes*"

Figure 1 shows the differences in genetic analysis focused only on individual variants against studies on variants plus haplotypes in a reference sample. As the figure explains (in Path A) in this situation the doctor looks for a change in the sample associated with a specific disease; but in fact merely checks if the variation (SNP) exists in the analyzed sample. With this information a report is generated on the existence of the variation.

Path B in the figure explains the ideal process we address in this work. The doctor looks for the variants and determines whether there are combinations between all the variants, trying to find haplotypes on different alleles through the analysis of the frequencies of each one. In order to obtain this data, we should apply different statistical models to present a more detailed and comprehensive genetic report. It is widely accepted that Haplotype studies enhance the variation detection rate (with or without combinations) for a specific disease [17]. The reason is that each allele or variation represents a frequency of occurrence in each population, so that with this information we can improve the generation of genetic diagnosis.

[2] A single nucleotide polymorphism (SNP), is a variation at a single position in a DNA sequence among individuals. It should be remembered that the DNA sequence is formed from a chain of four nucleotide bases: A, C, G, and T [*Scitable by Nature Edu.*].

3 Research Methodology

This paper deals with and analyzes only a part of a wider PhD research project and is based on the presentation of "*Treatment Design*" [15]. Accordingly, we used the *Engineering Cycle* method proposed by Wieringa [16], which includes the following steps:

a) Implementation evaluation / Problem investigation

In this phase we focus on describing the problem and submit initial proposals for the resolution of the case [15]. Our work with haplotypes and variants explains the necessity of extending our CSHG, because all the possible combinations between variants will help us to improve the impact of the diagnosis of genetic diseases.

b) Treatment design

One of the steps taken to design our solution was to study in depth the state of the art of the subject and then proceed to apply analysis and evaluation techniques to obtain the relevance and the impact factor of haplotypes for genetic diagnosis. Our solution is based on an integrative approach to heterogeneous genomic data sources using *conceptual models* as the main formalism. Our goal is to gather and represent the available knowledge using an unambiguous representation and establish concise relationships among the data. From these models we will develop a GeIS using a Model-driven development (MDD) approach [41].

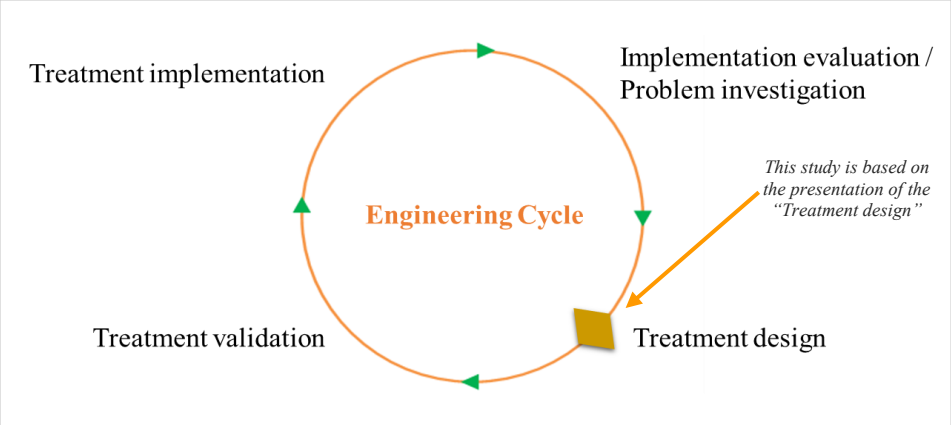

Figure 2: Engineering Cycle by Roel Wieringa

c) Treatment validation

Design validation is a knowledge task [15]. We are going to provide *validation mechanisms* for the outcome from our CSHG (*which integrates the haplotype treatment into the genomic diagnosis*) developed in the previous step.

d) Treatment implementation

Finally, the implementation of our solution will be tested by a selected group of physicians working in the genetic analysis domain and also with a group of users to obtain feedback on the functionality and the value of the information provided.

In the context of this paper we give an overview of the research and development of the second step of the Engineering Cycle (*Treatment Design*), in which we address the following points (*research phases*):

1. Study and analysis of the state of the art (*Section* 4)
2. Study of the domain and analysis of relevant information (*Sections* 2 and 4)
3. Extension of the conceptual schema and the design of Information Systems (*Section* 5)
4. Integration of relevant information into a single genomic database (*Sections* 6 and 7)

4 Related Work

Little work has been done to date on modeling "*haplotypes*". Various studies have tried to achieve a conceptual definition of the entire human genome, but this knowledge quickly becomes obsolete due to the continuous evolution of the domain.

Our research focuses on the integration and use of information on haplotypes in genomic information repositories, such as information systems and databases. For these reasons, we first analyze some of the most important genomic repositories in order to evaluate their schemas, the concepts they use and how haplotypes are stored.

dbSNP

The dbSNP repository is the top source of information on SNPs. This repository facilitates an ER schema which identifies the representation of data on: population, SNP allele frequencies and the summary of populations. dbSNP collects this data in the view "*Frequency calculation submitted by SNP and population*" with construction number 118 dated 11/17/2003. Within this view we find a table called "*b125_SNPMapInfo_35_1*" related to the "*SNP*" table, where there is only one attribute related with haplotypes - "*hap_cnt*"- [45], [56-57].

Table 2: dbSNP attribute identifier

Attribute	***dbSNP_Description***
-hap_cnt	The number of contigs that have the group_term (in *ContigInfo*) with ""*haplotype*"" suffix that the SNP aligns to.

Ensembl

The Ensembl repository mainly provides genomes for vertebrate species. In this repository the schema does not provide any explicit relationship with haplotypes, but it is noteworthy that in the "*features_analyses_core*" view we detected some entities that could be linked to treatment, such as: the table "*Marker_map_location*" and the attribute "*lod_score*"; which is statistical data used in population genetics and in LD calculations [36], [46].

UCSC Genome Bioinformatics

This site contains reference sequences and working draft assemblies for a large collection of genomes. In this case, they present the data as a "*Table schema*" that is somewhat difficult to manage [54]. Using the "*Gene Sorter*" tool [40] we could check the different data provided for a gene, including data associated with haplotypes in the block "*Common Gene haplotype Alleles*", which are generated from the 1000 Genomes project (*http://www.1000genomes.org/*). From the data presented we drafted an outline of the structure to compare with our solution [39], [47].

We also found a set of databases focused on collecting data associated with haplotypes and population frequencies, such as: HapMap [18], ALFRED [19], YHRD [20], D-HaploDB [21], and others [4], [8]. The problem with this type of repository is that it is difficult to manage and access the haplotype information, as this information is widely dispersed, *e.g.*, the data is stored in multiple text files (* .txt, * .csv, etc.). After analyzing different data sources (and schemas) that store haplotype information, we identified three main issues:

1. *Complexity of data management*:

   Data is presented ambiguously, and in many cases is difficult to understand and manipulate, *e.g.*, in the case of dbSNP and Ensembl, the haplotype data is not shown explicitly to final users. In our study we detected that the dbSNP repository uses data from the *HapMap project*, whereas USCS uses data from the *1000 Genomes project*.

   We also found some inconsistencies between these repositories (*i.e.*, contradictory or inconsistent information in the databases, like allele and genotype frequencies). As the genomic environment is continuously evolving, additional knowledge is incorporated. Several sources show data with information on haplotypes, as in the cases mentioned above, but the problem lies primarily in the complexity of management and interpretation of the data (importance, relevance, etc.) [43].

2. *High dispersion and data redundancy:*

This issue is a consequence of different data sources with large amounts of structured and unstructured information in different formats, for example, formats like: *.csv; *.txt; *.xml; *.fasta; and others. This wide range of formats makes it very difficult to process and analyze the data, so it is reasonable to adopt in this domain the benefits of *conceptual schemas*, which allow us to create a structure in which data can be shared effectively and redundancy and other issues can be reduced [52].

A further disadvantage identified in the haplotype data is its wide dispersion and the presence of redundant data [43]. Using a conceptual modeling approach, we will tackle these problems with comprehensive data processing to complement existing genetic diagnostics.

3. *There is no clear formalization of the concept "haplotypes":*

   Currently, the analyzed repositories do not provide a suitable structure (*schema*) to manage haplotypes. In some cases, they do not even represent the same concept, *e.g.* we found a lot differences between dbSNP and Ensembl on how they represent haplotypes at the conceptual level (*i.e.* form of representation, structure and others). We only found a sort of table schema specification in UCSC. dbSNP only shows an attribute associated with the concept of haplotypes in its schema, and for this reason is a very limited definition. Ensembl does not provide a clear definition in its schema. We found the "*lod_score*" attribute that is used in genetics, but not specifically for haplotype treatment. Although UCSC presents data on haplotypes, this repository does not have a conceptual schema.

There are also other alternatives for representing knowledge in general, and in our research we found *ontologies* applied to biological sequences. Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequences (*http://www.sequenceontology.org/*). This ontology defines a *haplotype* as one of a set of coexisting sequence variants of a haplotype block [55] and this approach is interesting for defining types, properties and relationships between entities at a more formal specification level. Our solution seeks to: (i) represent the existing data in this domain, manipulate and manage the information on haplotypes so as to make them easy to use in genomic treatment, (ii) solve existing shortcomings in this domain through the practical application of conceptual schemas, which may be open to extension regardless of the continuing evolution of genomic environment.

5 Conceptual Modeling of Haplotypes

The application of data management techniques in a genomic environment could be thwarted or affected by its special characteristics, such as: high conceptual complexity, large amounts of data and the constant evolution of the community.

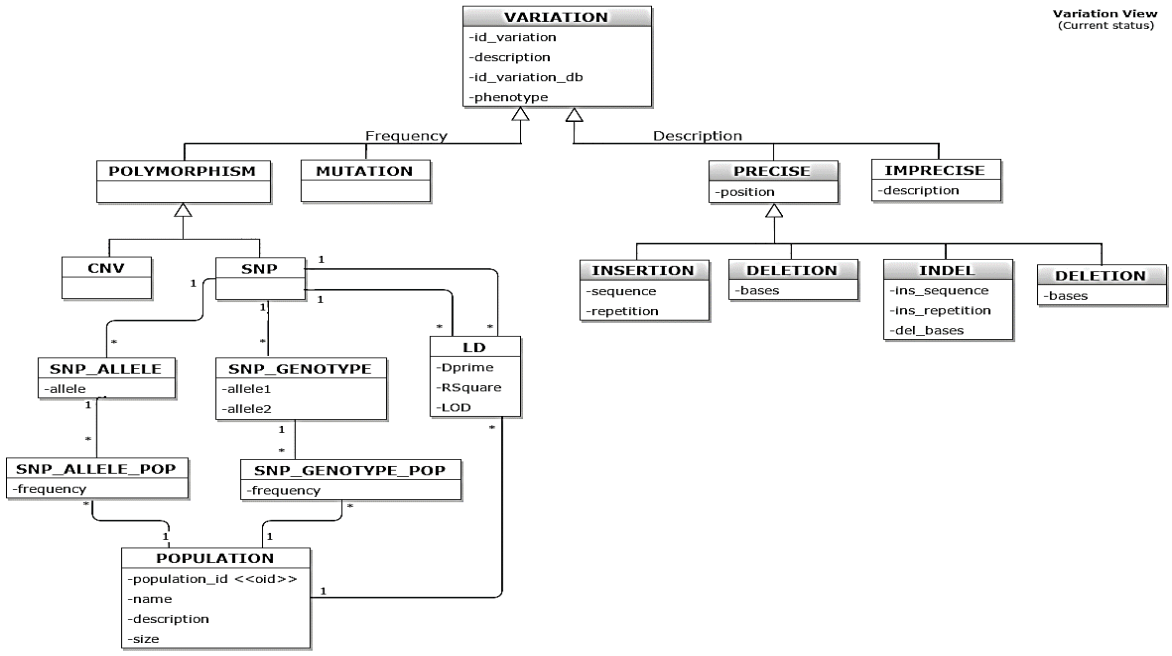

Figure 3: Current status of the "*Variation View*"

Our research group has been able to generate a Conceptual Schema of the Human Genome (CSHG) [24] which has evolved and grown over the years and has advanced the understanding of the human genome.

The latest version of the CSHG (v3) is classified into five views [44], which are:

(1) Structural view: this describes the genome structure (species, chromosomes, etc.).

(2) Transcription view: shows the components and concepts related to protein synthesis. Components involved in going from DNA to the diversity of RNAs.

(3) Variation view: this view models the knowledge related to the differences in the DNA sequence of different individuals.

(4) Pathway view: there are a series of chemical reactions within a cell in the area of biochemical pathways or metabolic pathways intended to enrich the conceptual schema with information on metabolic pathways joining genome components that participate in pathways with phenotype expressions.

(5) Bibliography and data bank view: gives the source of the data and contains a series of bibliographic reference documents for anyone wishing to obtain further information [22], [24].

Our idea of integrating haplotypes strengthens the conceptual schema and also assesses levels of incidence or the risk of "*variations*" in predisposing genetic diseases. Figure 3 shows the current state of the "*variation view*", indicating classes (*grey boxes*) that are currently loaded into our repository (using a data loading process [42], [49]).

In our CSHG (v3) we now use precise variations (*i.e. when the structure and the nucleotides that are involved are clearly defined*). However, when treating haplotypes, other concepts must also be managed, including: frequencies (*allelic and genotypic*) and populations [25], which is difficult to manage in the genomic field. This new scenario means that the data loading process should also consider the concepts represented on the left-hand side (*Frequency generalization*) of Figure 3.

The variations are presented in two groups divided by: *frequency* and *description*. The first is classified according to the variation frequency in the population (occurrence). In this case, there are two types: *mutations* and *polymorphisms*. An alteration in any part of the genetic code is known as a mutation and a single nucleotide change between the genomes of individuals of the same species is known as polymorphism (occurring in at least 1% of the population). The second group conforms to the description given by the variations, classified into two types: *precise* and *imprecise*. The former consists of those that indicate their position in the chromosome and the latter only describe their phenotypic effect but do not give their location within the chromosome.

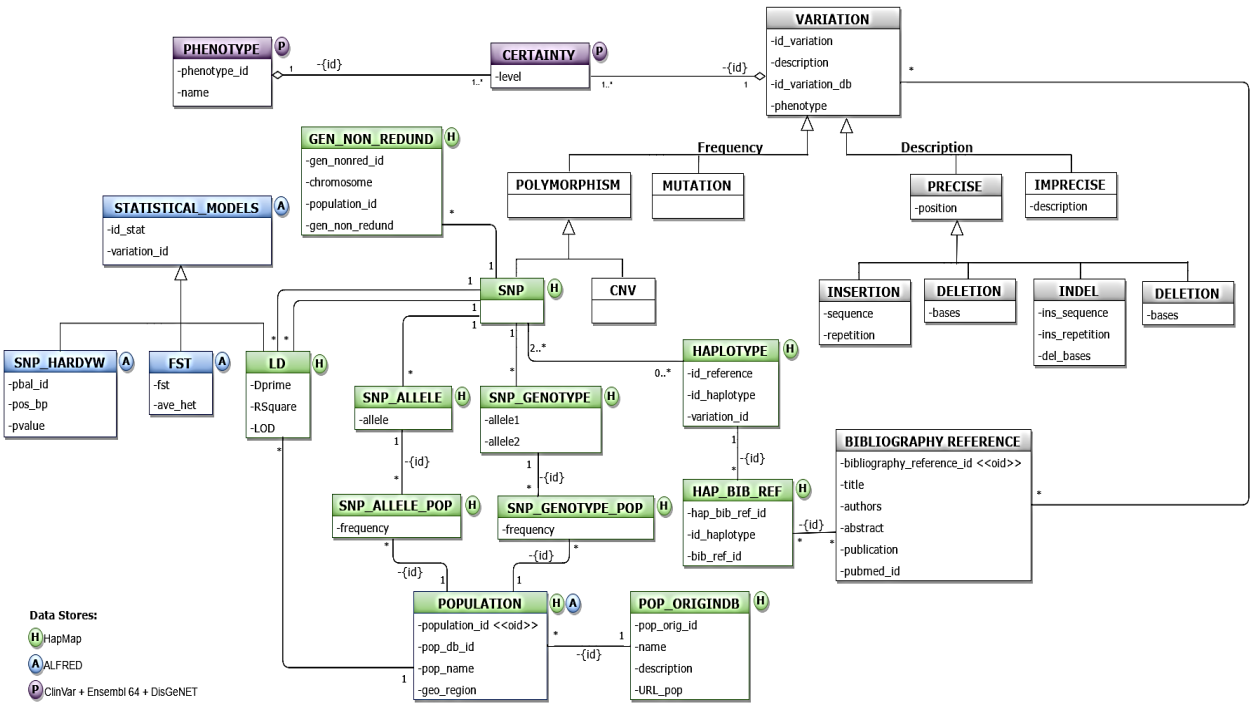

Figure 4: Haplotype integration to the CSHG – Phase II

The "*SNP*" class of our CSHG becomes the root class where new concepts appear. Although the conceptualization of "*SNP_Allele*"; "*SNP_Allele-Pop*"; "*SNP_Genotype*"; "*SNP_Genotype-Pop*"; "*Population*"; "*LD*" classes is represented in our original proposal, the data related to them was neither treated nor loaded into our repository for one of two causes: (a) unavailability of the data (sources, resources, etc.) or (b) we did not consider it to have an appreciable or tangible value.

An SNP is associated with many alleles ("*SNP Allele*") and these sets have a frequency of alleles in a specific population ("*SNP_Allele-Pop*"). Similarly, it also has several types of genotypic data ("*SNP_Genotype*") defined by "*allele1*" and "*allele2*" (the reference allele and the allele changed within the genotype); each allele represents a frequency within a population. The "*Population*" class is used to group all the populations that have been studied for the analysis of variants (variations) in the human genome.

Figure 4 gives our proposal for integrating haplotypes into the CSHG, in which we found a number of added or extended classes. As can be seen, the variation view defined by their "*description*" is preserved, as in the current version. In this figure, variations defined by "*frequency*" have to implement changes that are necessary for working with the haplotype data and statistical models.

The insertions and changes made to the schema are explained as follows: the "*Haplotype*" class, which is associated with the "*SNP*" class, is added, reflecting the relationships present in a Haplotype-SNP by the combination of two or more SNPs. The attributes that define "*Haplotype*" class in the schema are: "*id_reference*" which defines an identifier a link between the different SNPs and haplotype; the "*id_haplotype*" corresponds to the identifier of the haplotype as the "*variation_id*" that identifies the associated variation. As the haplotype should be based on a scientific resource to confirm its medical value, we created the class called "*Hap_Bib_Ref*" to help us the join the "*Haplotype*" and "*Bibliography reference*" classes, to keep our repository linked to various research works on haplotypes.

This class consists of the following attributes: "*population_id*" that serves as an identifier of the population; the identifier of the source "*pop_db_id*"; the name of the population "*pop_name*" and geographical region "*geo_region*". We also associated the new "*Pop_OriginDB*" class, which we use to define the sources that provide the above populations. In this class we define the following attributes: "*pop_orig_id*" which is the identifier of the repository; the "*name*" and "*description*" of the source, and the "*URL_pop*" which has the URL of population data file.

Another novelty in the schema is the addition of the "*Statistical_models*" class, which was defined with the objective of unifying the statistical models that are applied to the data related with variations [26-28], specifically on changes in the value of "*variation_id*", considering that for an SNP or variation this can be zero or many statistical models. Three subclasses of this superclass are derived: "*SNP_HardyW*", "*Fst*" and "*LD*".

The conceptual definition of these classes allows us to tackle very important concepts in the world of genomics, known as "*Population Genetics*", which is the study of the forces that alter the genetic makeup of a species. This focus is associated with micro-evolutionary mechanisms like: mutation, natural selection, gene flow and genetic drift [29-30]. The "*SNP_HardyW*" class represents the Hardy-Weinberg model (also known as *panmictic balance*), which is used to calculate genotype frequencies from allele frequencies [31-32], in which data is taken from the sources and applied to the model. This class has the following characteristics: unique identifier of the "*pbal_id*" class; chromosomal position in base pairs "*pos_bp*"; and the P-value "*pvalue*" which indicates the smallest possible level of significance.

$$\textit{Hardy-Weinberg equation}^{3} \quad = \quad \mathbf{p^2 + 2pq + q^2 = 1}$$

Where: p: is the frequency of the "A" allele and q is the frequency of the "a" allele in the population. *In the equation*,

p2: frequency of the homozygous genotype AA,
q2: frequency of the homozygous genotype aa, and
2pq: frequency of the heterozygous genotype Aa.

Another statistical value used in population genetics is the calculation of "*fixation indices*", which allow us to measure the differentiation of the population due to genetic structure, facilitating the comparison of genetic

[3] Hardy-Weinberg equation: this is a mathematical equation that can be used to calculate the genetic variation of a population at equilibrium [*Scitable by Nature Education*].

variability within and between populations. To do this we define the "*Fst*" class, containing the values: "*Fst*" for fixation index and the "*Ave_Het*", which indicates the average heterozygosity. The "*LD*" class defines the "*Linkage disequilibrium*", which occurs when the genotypes in the two loci are not independent of each other.

To calculate the LD, we found three statistical biological parameters, which are: (1) the *sensitivity index* "*DPrime*", which measures the imbalance between the alleles' interaction. (2) The *coefficient of determination* "*RSquare*" which serves to determine the quality of the model, in order to replicate the results, and the proportion of variation in the results presented in the model. (3) *LOD score* "*LOD*": this value refers to the logarithm on the odds of two genes or loci being linked and thus being inherited together more often than usual.

The "*LD*" class is related to the "*SNP*" class, indicating that an SNP can have zero to many LDs. In this extension of the CSHG we integrated all the existing data on phenotypes from different repositories and suggested the "*Phenotype*" class, which is joined to the "*Variation*" class through the intermediate class "*Certainty*", used as the indicator of the incidence level between phenotype-variation (this value is extremely difficult to define, but studies on this topic provide "*estimated values*" within the population).

Finally, the "*Gen_non_redund*" class helps us to offer the results of the cleaned dataset (without redundancy, *i.e.* elimination of inconsistencies and duplicated data), for sets for SNP-genotyping and population. For this class we assign a unique identifier to the data without redundancy "*gen_nonred_id*"; information about chromosome and the studied population, the attributes are "*chromosome*" and "*population_id*" respectively; and total non-redundant data after performing filtering "*gen_non_redund*".

6 Conceptual model validation

To validate our work, we checked whether the extended conceptual model supports the information provided by popular haplotypes repositories: establishing conceptual alignment with data available from popular genomic repositories. Figure 4 shows the elements added to the schema and their sources. The main data source was the HapMap Project in its third phase.

According to our variation view, the HapMap dataset is contained in the following directories: (a) frequencies and (b) ld_data. First, "*frequencies*" is a directory that contains the above SNPs with their frequency in each population. Due to the large amount of information, it is clustered into two groups: the frequencies of SNP variant types, taking into account only one allele of the chromosome ("*Allele_freqs*"), and frequencies of SNP rate variants, taking into account the two alleles ("*Genotype_freqs*").

The table below shows the elements added to the schema, the data repositories and the specific field/table which is aligned with the new element.

Table 3: Schema elements + Data source (origin)

	Schema element (name)	**Origin**	
		Data source	**File / Table**
"*frequencies*" and "*LD_data*"	SNP_Allele	HapMap	*Allele_Freqs_X_Y*
	SNP_Allele-pop	HapMap	*Allele_Freqs_X_Y* → X: *Chromosome*, Y: *Population*
	SNP_Genotype	HapMap	*Genotype_Freqs_X_Y*
	SNP_Genotype-pop	HapMap (*BioQ*)	*Overall Hardy-Weinberg*
	Gen_non_redund	HapMap	*Genotyped non-redundant QC+ SNPs*
	Population	HapMap; ALFRED	*List of populations*
"*frequencies*"	Statistical_models	ALFRED	*Siteswithfstavghet*
	Fst	ALFRED	…
	SNP_HardyW	ALFRED	*Overall Hardy-Weinberg*
	LD	HapMap	*LD_X_Y*
	Phenotype	ClinVar; Ensembl; DisGeNET	*ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/* *Ensembl 64 phenotypes* *all_gene_disease_associations.tsv*

The Linkage disequilibrium data [33] is provided by the "*LD_data*" directory. The large amount of information is divided by chromosomes and populations. Information on population "*Population*" and non-redundant genes "*Gen_non_redund*" is found in specific sections of the main website.

The "*Statistical_models*" class consists of the data extracted from HapMap for "*LD*", and provided by the ALFRED repository for frequencies and statistical processes applied to the biological field. The frequencies dataset from ALFRED is in a file called "*FreqByChrom*" available from their website; these frequencies have been obtained for each chromosome according to the population studied.

This repository also facilitates the extraction of information on the populations they used. For the "*phenotypic*" entity, we find several data sources from which we can extract all the information classified in multiple repositories, such as: *ClinVar*, *Ensembl* and *DisGeNET* (sorted by their scientific relevance) [34-37].

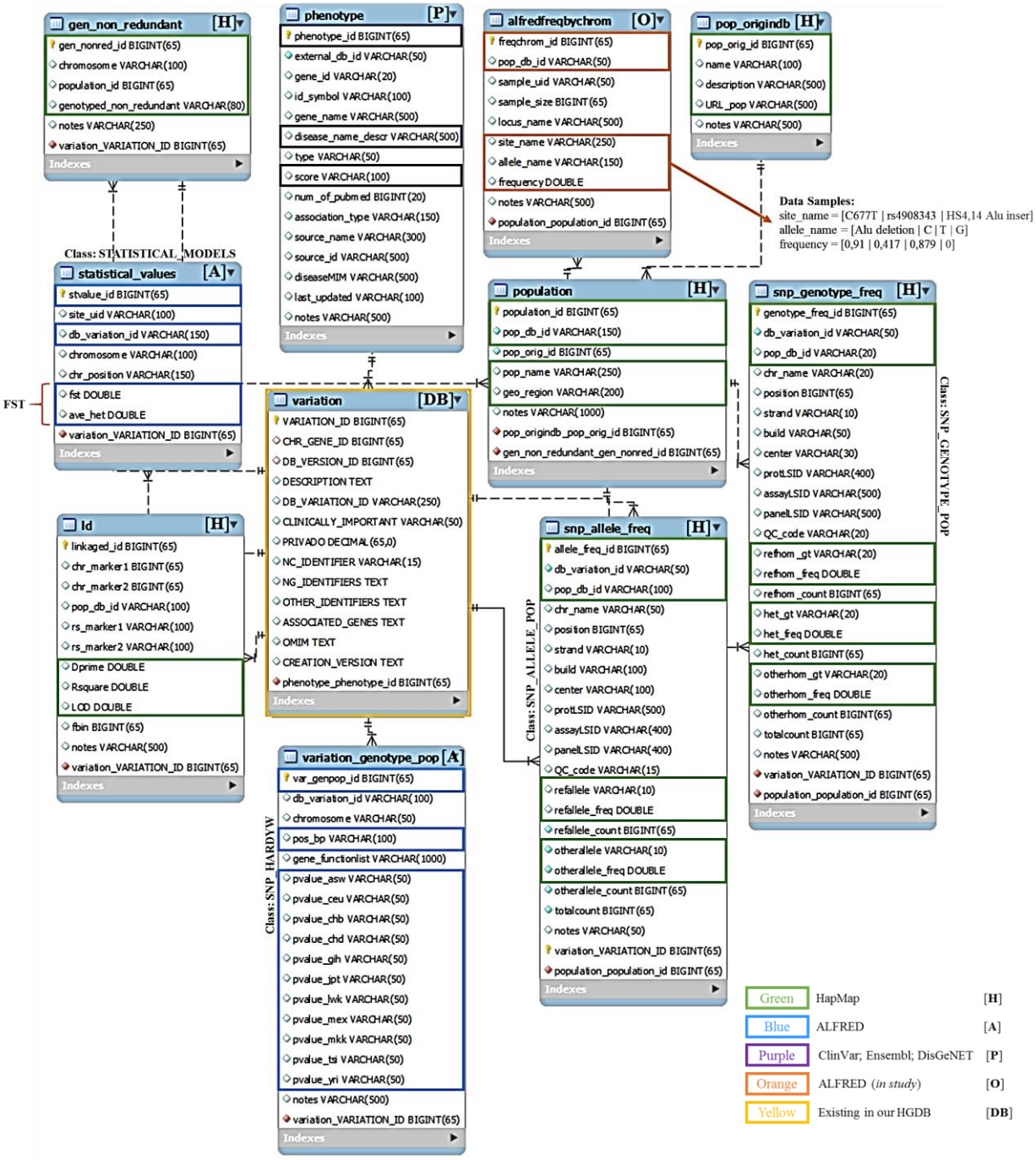

Figure 6: ER model (initial)

7 Development of a haplotype database

This section deals with the treatment of data (haplotypes) from the previously defined *conceptual model* (Figure 4), using the data repositories to define queries relevant to the analysis of the data. As mentioned in the previous section, the data on variations that are part of one or more haplotypes and their respective frequencies were obtained from different repositories. Our data management approach follows five steps:

a) Impact study of existing and available repositories

In this activity we focus on evaluating the usefulness of the data provided and its impact on research and scientific publications, with the aim of defining a framework data source according to the latest advances in the genomic context (see *datasources* in Table 3).

b) Data collection (haplotype frequencies and statistical calculations)

After selecting the data repositories, we proceed to download the files (specified in Table 3). At this stage it is usual to find huge data files, for example:

- Allele frequencies for *chromosome 1* with the studied HapMap populations (11) with an average of 450-1,024MB ±.
- LD data associated with *chromosome 1* for the "*ASW*" population with an average of 930-1,024MB ±.
- In the case of ALFRED repository, we obtained the *chromosomes frequencies* with an average of 500-650MB ±. For files with *phenotypic* data it was much more affordable in size issues.

After downloading all the files, we generate multiple *gigas* of scattered information with completely different structures to be evaluated. In order to strengthen previous analyses, we decided to complement and compare the information with data generated from the *BioQ* platform (*http://bioq.saclab.net/*) [61], which provides a set of tools to *consult*, *document* and *download* information from relational databases (*genomic*), such as: *1000 Genomes*, *dbSNP*, *Ensembl*, and others (*for specific versions that they manage*).

c) Analysis of the data stored

Due to the heterogeneity of the data, the first thing we did in this phase was to transform all the data from its current format (*see Section 4*) to a common structure. We decided to convert all the files to "**.csv*" format for debugging and analysis, and then took the first genome chromosomes (*i.e.*, 1-3) as a test case for handling and storage issues.

As we can see in Figures 7 and 8, after concluding our analysis of the relevant data for the treatment of haplotypes and statistical factors, we identified the large contribution of the ALFRED (1,063.651 rows) and HapMap (194.417 rows) repositories to our study. We also classified all the knowledge obtained into five categories (*frequencies*, *statistical values*, *phenotypes*, *populations* and *others*). The largest amount of the processed data was on *frequencies* and *statistical values*.

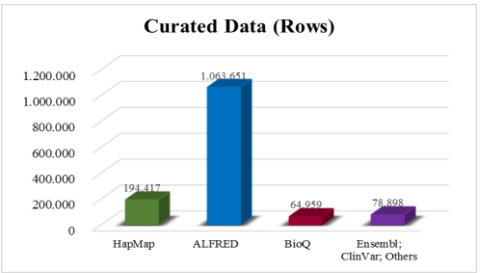

Figure 7: Curated data loaded by data repository

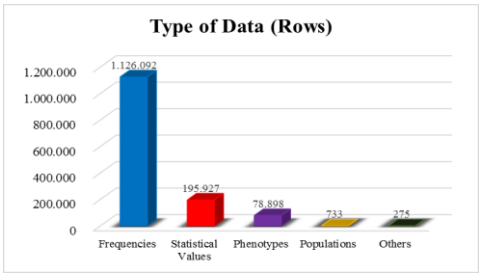

Figure 8: Type of the data stored (total of rows)

d) Massive data load

After completing the analysis and treatment of the data, we did a preliminary *data filtering*, and the next step was to develop an initial database schema for loading all the information. For this, we studied the most appropriate *data structures* to organize, consult and avoid processing problems (*e.g. special signs*) in the data.

The next task was to import the files (**.csv*) using a database management environment (see Figure 9), and then generated a new ER diagram, while maintaining the traceability of the origin of the data used (see Figure 6).

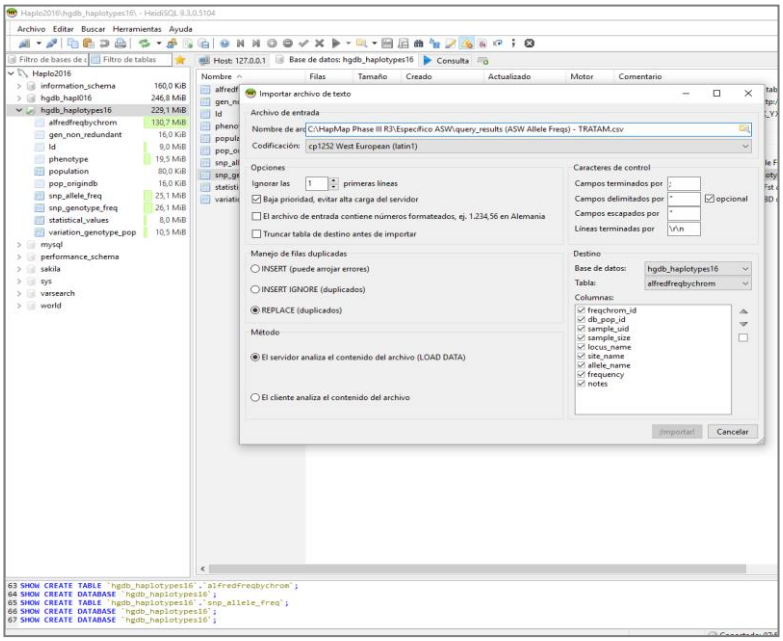

Figure 9: Data import using HeidiSQL

e) Query generation (preliminary)

In this phase we raised a number of *questions* with the aim of finding *answers* through the generation of different *SQL queries* on the previously loaded schema, including the following examples:

```
(1) SELECT COUNT(DISTINCT(id_symbol)) FROM phenotype
```

In this query we included a total of 4,192 genes loaded, which represent an association with one or more diseases or phenotypes.

```
(2) SELECT phenotype_id, id_symbol, disease_name_descr, type, score,
    num_of_pubmed, source_name, diseaseMIM, last_updated, notes FROM phenotype
    WHERE id_symbol = 'GABRA2'
```

phenotype_id	id_symbol	disease_name_descr	type	score	num_of_pubmed	source_name	diseaseMIM	last_updated	notes
72.387	GABRA2	Alcohol dependence	disease		0	NCBI curation	103780	26-oct-11	ClinVar

With this query we obtained the *GABRA2* gene with the identifier 72.387 (*phenotype_id*), which was loaded from the *ClinVar* data source and is associated with "*alcohol dependence*" (see Section 2). Other popular queries were on the subject of: *frequency variations*, *populations* and representation of *statistical data* (biological) obtained for the study.

```
(3) SELECT allele_freq_id, db_variation_id, pop_db_id, chr_name, position,
    strand, refallele, refallele_freq, otherallele, otherallele_freq,
    totalcount FROM snp_allele_freq WHERE position BETWEEN 9200000 AND 9209000
```

allele_freq...	db_variatio...	pop_db_id	chr_na...	position	strand	refallele	refallele_f...	otherallele	otherallele_f...	totalcount
4.831	1009940	ASW	ch1	9.200.729	+	G	0,711	A	0,289	114
4.832	17033526	ASW	ch1	9.202.760	+	A	0,875	G	0,125	112
4.833	7534423	ASW	ch1	9.203.235	+	G	0,877	T	0,123	114
4.834	12402600	ASW	ch1	9.204.303	+	G	0,763	A	0,237	114
4.835	10489436	ASW	ch1	9.208.308	+	A	0,702	G	0,298	114
4.836	12072683	ASW	ch1	9.208.605	+	C	0,491	T	0,509	114

To reduce the amount of data on this query, we decided to fix a search range. Here we show six variations (*db_variation_id*) included on the position of the chromosome *-9.200,000* and *9.209,000-*, which indicates that it belongs to *chromosome 1* (positive strand) for the Asian population (*ASW*). In the specific case of the "*1009940*" variation, this presents allele "*G*" as reference, which has a frequency of *0.711*, and the other allele considered "*A*", which has a frequency of *0.289* in this population (*with a total count of 114 cases*) and so on for each variation.

The statistical data, for example, the study of the different calculations related to the *LD data –biological-*, includes variations and populations. In our next query (4) we give a selection of the most frequently repeated data in the LD table, sorted by the number of repetitions (column "*num*"):

```
(4) SELECT pop_db_id, rs_marker1, rs_marker2, Dprime, Rsquare, LOD, fbin,
    COUNT(*) AS num FROM ld GROUP BY 'rs_marker1' ORDER BY num DESC LIMIT 0 ,
    15
```

pop_db_id	rs_mark...	rs_mark...	Dprime	Rsquare	LOD	fbin	num
CHD	16919558	2804311	0,078	0	0	5	251
CHD	2804311	2641984	1	0,252	7,11	5	250
CHD	2641984	2641983	1	0,261	7,21	5	249
CHD	2641983	7031553	1	1	30,2	5	248
CHD	7031553	9632892	1	0,128	3,4	5	247
CHD	7048037	10975061	1	0,937	26,96	5	247
CHD	2641989	16919558	1	0,005	0,31	5	247
CHD	10975061	7040388	1	0,906	25,24	5	246
CHD	1565793	10815231	1	0,159	4,11	5	246
CHD	9632892	1565793	1	1	34,56	5	246
CHD	9408625	7048037	1	0,968	28,4	5	245
CHD	10815231	10975130	1	0,97	31,24	5	245
CHD	10758683	9408630	1	0,318	8	5	245
CHD	7040388	10758683	1	0,288	6,92	5	245
CHD	2804313	2279619	1	0,332	10,05	5	245

As mentioned in Section 5, "*Linkage disequilibrium*" (LD) allows us to identify when the genotypes at the two loci are not independent of each other, and relies on statistical models used in population genetics, like: *DPrime*, *Rsquare* and *LOD* (described in Section 5).

```
(5) SELECT linkaged_id, pop_db_id, rs_marker1, rs_marker2, Dprime, Rsquare,
    LOD, fbin FROM ld WHERE rs_marker1 = 16919558
```

linkaged_id	pop_db_id	rs_marker1	rs_marker2	Dprime	Rsquare	LOD	fbin
57.936	CHD	16919558	2804311	0,078	0	0	5
57.937	CHD	16919558	2641984	1	0,014	0,38	5
57.938	CHD	16919558	2641983	0,057	0	0	5
57.939	CHD	16919558	7031553	0,068	0	0	5
...	...	...	...	...	...	...	...

This query states, for example, that in "CHD" individuals (*Beijing, China*) presenting variations "*16919558*" and "*2804311*" there is a: DPrime = *0.078*, Rsquare and LOD = *0*. In this case we can say that these variations applied to the statistical model are not highly dependent on one another, which indicates a probability that there are other variations with a higher complicity.

Thus, by launching query (5) we could see the calculation of LD for the variation (*rs_marker1*) "*16919558*" and how this compares with a total of *251* variations (*rs_marker2*) for the "*CHD*" population.

In the treatment of this data we developed each step with different control parameters in order to generate a reliable and solid outcome. It is noteworthy that despite the great *heterogeneity* and *dispersion of the data* among the different repositories analyzed (*genomic data*), we can reduce these problems (*raw data*) by collecting and applying conceptual modeling and data management techniques and thus generate and manage repositories with *curated data*.

In our research work we seek to exploit this data (*haplotypes*) in a new way, taking advantage of the current knowledge on genetic variations and themes related to "*population genetics*", which positively collaborates in the detection of genetic diseases with personalized attention (*personalized medicine*).

8 Database evolution according to the conceptual model

In this section we focus on how the way in which the different representations of genomic knowledge are presented affects the structure of the databases used to manage all the data. This is an important part of this work as data is available in different omics data sources and the selected database structure determines how the data is to be managed. The role in this context is to keep an overall conceptual perspective of this data management problem, regardless of the database structure, to obtain a precise conceptual workbench to integrate data correctly.

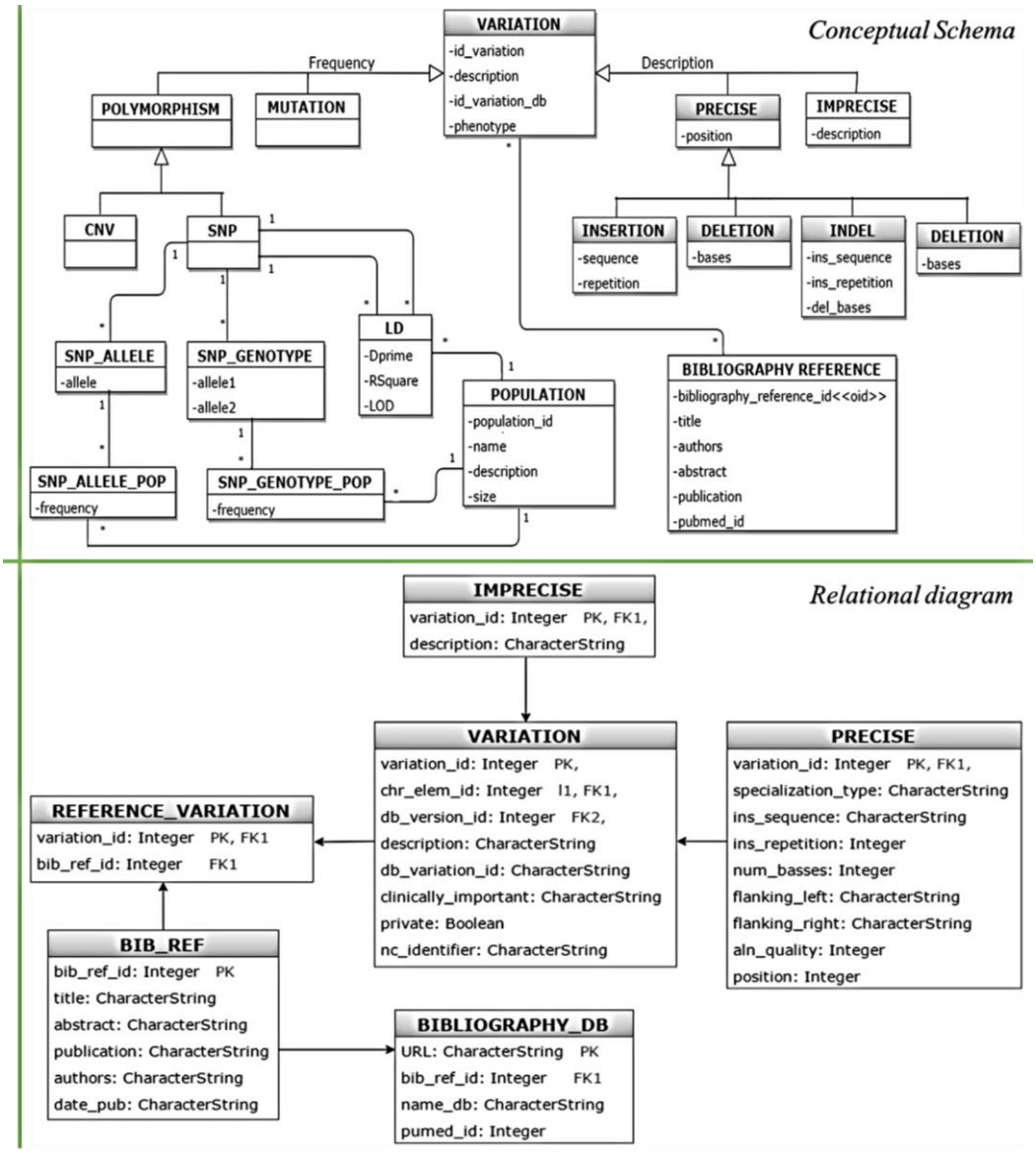

Figure 9: Previous version (*current*)

Figures 9 and 10 show a fragment of a relational database diagram generated from the existing version (*previous*) and our extended version of the CSHG representing the concepts used in this view.

Our previous version only represents the data loaded into our genomic repository, or variations with a "*description*". A relational diagram was generated with the "*precise*" and "*imprecise*" definitions (with their respective references), omitting the part related to frequencies and populations (not stored).

As mentioned in Sections 2 and 4, haplotypes should be integrated into the conceptual schema because they greatly improve the expressiveness and detail of genetic diagnoses.

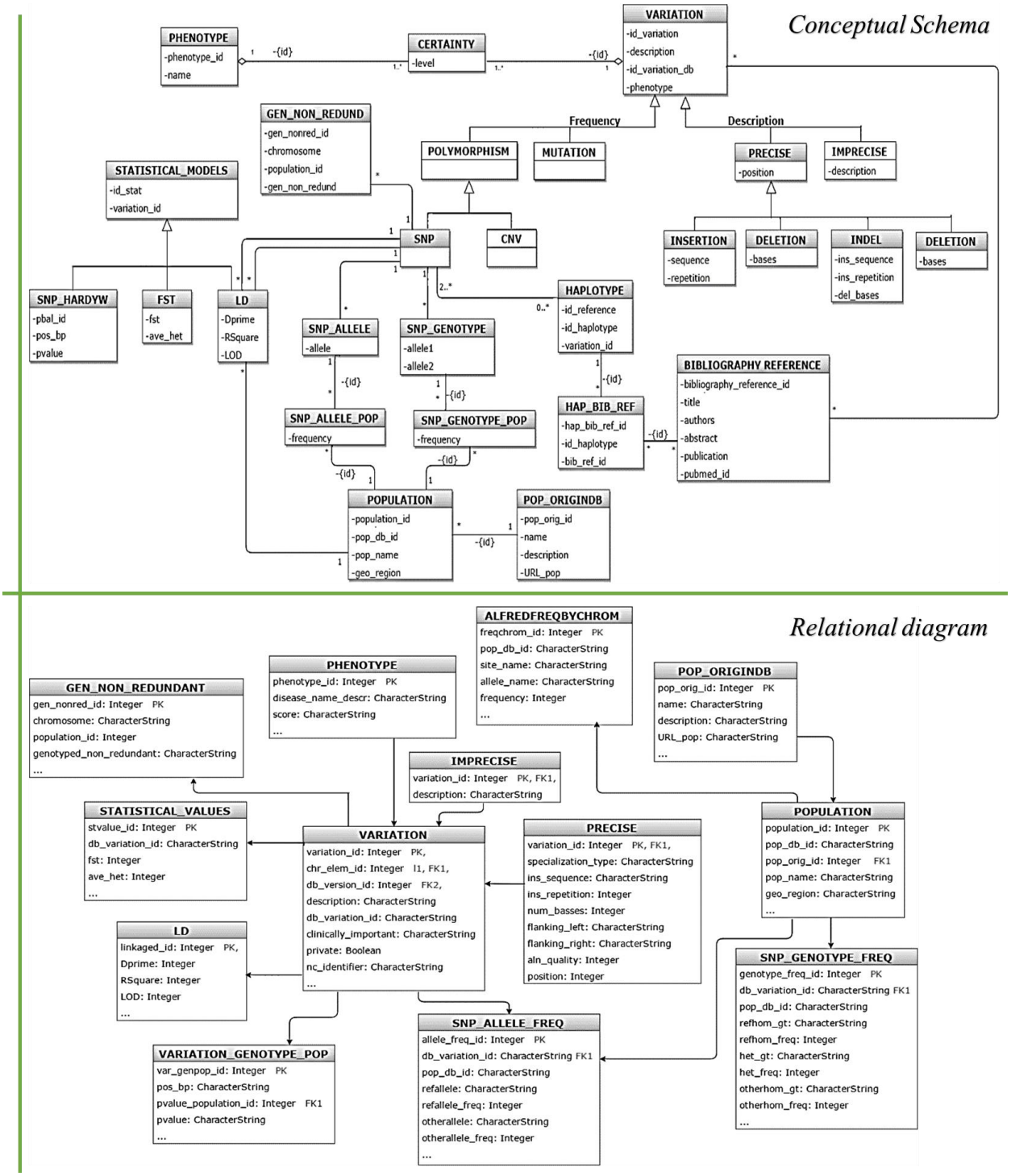

Figure 10: New version (*extension*)

Figure 10 shows our extended conceptual schema with the integration of haplotypes. In this case we include all the elements of the variation by "*description*" and "*frequency*", generating a more complete relational diagram. This new representation seeks to reduce dispersion issues and prevent redundancy.

This representation means that variations can be treated more specifically, covering all the relevant factors in current genomic knowledge. When the diagram identifies the relationships between variations as part of a haplotype in the genome, the information is enhanced (*with a higher level of detail*) and points to more accurate results. The

fact of incorporating population values into variations plays a key role because of its great impact on the results for end-users (*diagnostic* / *personalized medicine*).

This way of representing the data on genome variations defines a new conceptual and logical schema that is very helpful for managing information (*loading and handling*) and provides a substantial improvement in performance.

Our aim in this section is to show that an accurate representation of the genome data requires precise knowledge of the selected conceptual schema intended to guide the potential different possibilities in which the relevant data can be stored and managed. Although different representations can be made of the same data sources, only a well-defined conceptual schema can provide the unified conceptual perspective needed for managing data correctly and understanding the way in which the data is stored.

9 Lessons Learned and Future Work

The application of conceptual models (CM) to the bioinformatics field is essential for developing *Genomic Information Systems* (GeIS). Because of recent advances in genomic data repositories, the extension of the previously proposed CSHG is a must. We believe that our conceptual alignment is correct because it extends genetic knowledge in the CSHG through the integration of new concepts related with haplotypes, thus achieving a valid conceptual representation for the different data sources analyzed.

The main contributions of this work are focused on two aspects:

(1) The *extension of the conceptual schema (CS)*, which has great advantages over other representations for its ease of use, understanding and adaptation of genomic data, which are in constant evolution. With our schema we address various gaps identified in the current management of genomic data.

(2) The *generation of a new logical schema* that adds value in order to analyze, use and exploit the knowledge of haplotypes generated over the years.

One lesson that has been learned is that the application of conceptual modeling [60] techniques contributes to the specification of a sound and reliable schema, to address the various problems detected in genomic data management (see Section 2). The incorporation of conceptual models (CM) into the genomic domain has facilitated data processing, resulting in more powerful "*Information Systems*" that seek to solve cases related to the large set of heterogeneous data sources (responsible for the high dispersion of data) and incompatible data structures (schemas), among others (see Section 4).

As can be seen in Table 3, when assessing the status of the data associated with haplotypes and frequencies (allelic and genotypic), we find that these repositories provide the information in a wide range of formats, as for example: text (txt) and comma-separated value (csv) files. Firstly, we decided to trace all existing knowledge on this topic; after finishing our conceptual alignment (see Section 6), as an initial step we generated a physical schema to upload and retrieve all the information more efficiently. For this we used the support of database tools such as *HeidiSQL* [50] and *MySQL Workbench* [51]. All the data was joined with these tools and then an initial schema was generated (Figure 6) that would display a novel perspective of the data. Figure 6 shows a database schema generated after extracting the different repositories (HapMap, ALFRED, etc.), in which each table gives the source of the data (with the database symbol, *e.g.* [H]: HapMap).

Finally, after completing this second step of the research methodology, we found that the genomic domain requires the use of approaches that contribute to the organization and management of the "*current genomic chaos*". With the support of conceptual modeling techniques, the problem is easier to understand and can be solved systematically.

In this work, we developed a conceptual model for the treatment of haplotypes and validated it by determining whether current genomic knowledge (*in the different repositories*) could be supported by our model. After completing the validation process, we introduced a database into the proposed conceptual model and explained the evolution of the database according to the evolution of the conceptual model.

An additional contribution of this work is the novel conceptual model that combines the knowledge of haplotypes with statistical factors in order to avoid the problems of heterogeneity and the dispersion of data.

Future research work will be aimed at the development of Steps 3 and 4 of the Engineering Cycle (as mentioned in Section 3). The main goal is to check whether current sources could be managed correctly and effectively. Additionally, we will validate the treatment of haplotypes and statistical factors by including this

repository in the development of a software tool for genetic diagnosis. Finally, we want to extend the model with studies on the treatment of *haplogroups*, which include subjects with a similar genetic profile who share a common ancestor.

Acknowledgements

The authors thanks to the members of the PROS Center Genome group for fruitful discussions. This work has been supported by the Ministry of Higher Education, Science and Technology (*MESCyT*) of the Dominican Republic, and It also has the support of Generalitat Valenciana through project IDEO (PROMETEOII/2014/039).

References

[1] N. W. Paton, S. A. Khan, A. Hayes, F. Moussouni, A. Brass, K. Eillbeck and S. G. Oliver, "Conceptual Modelling of Genomic Information", *BIOINFORMATICS*, vol. 16, no. 6, pp. 548-557, 2000. DOI: 10.1093/bioinformatics/16.6.548

[2] O. Pastor, "Enforcing Conceptual Modeling to Improve the Understanding of Human Genome", *Asia-Pacific Conference on Conceptual Modelling (APCCM2010)*, 2010.

[3] E. Bornberg-Bauer and N. W. Paton, "Conceptual data modelling for bioinformatics", *Henry Stewart Publications 1467-5463, Briefings in Bioinformatics*, vol 3, no 2, pp. 166–180, 2002. DOI: 10.1093/bib/3.2.166

[4] A. Martin and M. Celma, "Integrating Human Genome Variation Data: An Information System Approach", *22nd International Workshop on Database and Expert Systems Applications (DEXA)*, IEEE, pp. 65-69, 2011. DOI: 10.1109/DEXA.2011.45

[5] A. M. Sánchez, "Sensibilidad al Alcohol y la Predisposición a beber", *Ciencia al Día Internacional*, vol. 3, no. 2, ISSN 0717-3849, pp. 1-13, 2000. [Online]. Available: http://www.ciencia.cl/CienciaAlDia/volumen3/numero2/articulos/articulo3.html

[6] D. M. Dick and T. Foroud, "Candidate Genes for Alcohol Dependence: A Review of Genetic Evidence from Human Studies", *Alcoholism: Clinical & Experimental Research*, vol. 27, no. 5, pp. 868-879, 2003. DOI: 10.1097/01.ALC.0000065436.24221.63

[7] P. L. Fernández, J. M. Ladero Quesada, J. C. Leza Cerro, I. Lizasoain H., "Drogodependencias: Farmacología - Patología – Psicología – Legislación", Ed. Médica Panamericana, Edición 3, ISBN: 8498354684, 9788498354683, 725 páginas, 2009.

[8] J. F. Reyes Román, "Integración de Haplotipos al Modelo Conceptual del Genoma Humano utilizando la Metodología SILE", *Tesis de Máster en Ing. Software, Métodos Formales & Sistemas de Información*, Universitat Politècnica de València (UPV), Valencia, España, 2013. [Online]. Available: http://hdl.handle.net/10251/43776

[9] B. Campos, O. Díez, C. Álvarez, L. Palma, M. Domènech, J. Balmaña and M. Baiget, "Análisis del haplotipo en portadores de la mutación 6857delAA en el gen BRCA2 en 4 familias con cáncer de mama u ovario hereditario", *Medicina Clínica*, vol. 123, no. 14, pp. 543-545, 2004. DOI: 10.1157/13067548

[10] A. S. Álvarez, J. T. Márquez, F. R. Vargas and M. R. Romero, "Asociación del cáncer de mama con los polimorfismos T-66G y G-156GG del gen SPP1 y las concentraciones séricas de osteopontina", *Ginecol Obstet Mex*, ISSN-0300-9041, vol. 80, no. 1, pp. 22-29, 2012.

[11] TSRI – The Scripps Research Institute, "*The Effects of Alcohol on the Brain*", 2016. [Online]. Available: http://www.scripps.edu/newsandviews/e_20020225/koob2.html

[12] H. J. Edenberg, D. M. Dick, X. Xuei, H. Tian, L. Almasy, L. O. Bauer and J. Kwon, "Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations", *The American Journal of Human Genetics*, vol. 74, no. 4, pp. 705-714, 2004. DOI: 10.1086/383283

[13] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel and S. N. Liu-Cordero, "The Structure of Haplotype Blocks in the Human Genome", *Science*, vol. 296, no. 5576, pp. 2225-2229, 2002. DOI: 10.1126/science.1069424

[14] D. Fonseca, C. Silva, H. Mateus and C. M. Restrepo, "Deletions Identification in Female Carriers of Duchenne's Muscular Distrophy", *Acta Medica Colombiana*, ISSN 0120-2448, vol. 33, no. 2, pp. 63-67, 2008.

[15] R. Wieringa, “Introduction to design science methodology”, *Slides based on a book on Design Science methodology*, REFSQ 2013, Doctoral Symposium, 2013. [Online]. Available: https://refsq.org/wp-content/uploads/2013/05/Wieringa-2013-REFSQ-DS-Introduction-to-design-science-methodology-slides.pdf

[16] R. J. Wieringa, “Design science methodology for information systems and software engineering”, *Springer*, 2014. DOI: 10.1007/978-3-662-43839-8

[17] P. Delves, S. Martin, D. Burton and I. Roitt, “Inmunología: Fundamentos”, *Ed. Médica Panamericana*, ISBN 9500608995, 9789500608992, EAN: 9786077743934, Edition 12ª, pp. 1-548, 2014.

[18] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang and P. K. H. Tam, "The international HapMap project.", *Nature*, vol. 426, no.6968, pp. 789-796, 2003. DOI: 10.1038/nature02168

[19] K. H. Cheung, M. V. Osier, J. R. Kidd, A. J. Pakstis, P. L. Miller and K. K. Kidd, "ALFRED: an allele frequency database for diverse populations and DNA polymorphisms", *Nucleic Acids Research*, vol. 28, no.1, pp. 361-363, 2000. DOI: 10.1093/nar/28.1.361

[20] S. Willuweit, L. Roewer and International Forensic Y Chromosome User Group, "Y chromosome haplotype reference database (YHRD): update", *Forensic Science International: Genetics*, vol. 1, no. 2, pp. 83-87, 2007. http://dx.doi.org/10.1016/j.fsigen.2007.01.017

[21] K. Higasa, K. Miyatake, Y. Kukita, T. Tahira and K. Hayashi, "D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples", *Nucleic Acids Research*, 35. suppl 1: D685-D689, 2007. DOI: 10.1093/nar/gkl848

[22] A. M. Mayordomo, “Human genome conceptual modeling: An ontological framework for the design and implementation of genomic Information Systems”, *Sixth International Conference on Research Challenges in Information Science (RCIS)*, IEEE, pp. 1-6, 2012. DOI: 10.1109/RCIS.2012.6240461

[23] S. Ram and W. Wei, "Modeling the semantics of 3D protein structures”, *In Conceptual Modeling–ER 2004*, Springer Berlin Heidelberg, pp. 696-708, 2004. DOI: 10.1007/978-3-540-30464-7_52

[24] O. Pastor, A. M. Levin, J. C. Casamayor, M. Celma, L. E. Eraso, M. J. Villanueva and M. Pérez-Alonso, "Enforcing conceptual modeling to improve the understanding of human genome", *Fourth International Conference on Research Challenges in Information Science (RCIS)*, IEEE, pp. 85-92, 2010. DOI: 10.1109/RCIS.2010.5507367

[25] The International HapMap Consortium*, “A haplotype map of the human genome”, *Nature*, vol. 437, no. 7063, pp. 1299-1320, 2005. DOI:10.1038/nature04226

[26] M. Stephens, N. J. Smith and P. Donnelly, “A New Statistical Method for Haplotype Reconstruction from Population Data”, *The American Journal of Human Genetics*, vol. 68, Issue 4, pp. 978–989, 2001. DOI: 10.1086/319501

[27] D. Fallin and N. J. Schork, “Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data”, *The American Journal of Human Genetics*, vol. 67, Issue 4, pp. 947–959, 2000. DOI: 10.1086/303069

[28] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson and G. A. Poland, “Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous”, *The American Journal of Human Genetics*, vol. 70, Issue 2, pp. 425–434, 2002. DOI: 10.1086/338688

[29] UVIGEN, “Genética de Poblaciones”, *Universidad de la República Uruguay*, 2003. [Online]. Available: http://uvigen.fcien.edu.uy/utem/Popgen/popintro.html

[30] A. Barbadilla, "Genética de Poblaciones", *Departamento de Genética y Microbiología*, Universidad Autónoma de Barcelona, 2009. [Online]. Available: http://bioinformatica.uab.es/divulgacio/genpob.html

[31] “Modelo de Hardy-Weinberg”, *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, 2015.

[32] J. C. Barrett, B. Fry, J. D. M. J. Maller and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps", *Bioinformatics*, vol. 21, no. 2, pp. 263-265, 2005. DOI: 10.1093/bioinformatics/bth457

[33] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter and E. S. Lander, "Linkage disequilibrium in the human genome", *Nature*, vol. 411, no. 6834, pp. 199-204, 2001. DOI:10.1038/35075590

[34] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church and D. R. Maglott, "ClinVar: public archive of relationships among sequence variation and human phenotype", *Nucleic Acids Research*, vol.

42, D1: D980-D985, 2014. DOI: 10.1093/nar/gkt1113

[35] J. J. Johnston and L. G. Biesecker, "Databases of genomic variation and phenotypes: existing resources and future needs", *Human molecular genetics*, ddt384, 2013. DOI: 10.1093/hmg/ddt384

[36] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark and R. Durbin, "The Ensembl genome database project", *Nucleic Acids Research*, vol. 30, no. 1, pp. 38-41, 2002. DOI: 10.1093/nar/30.1.38

[37] J. Piñero, N. Queralt-Rosinach, Á. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron and L. I. Furlong, "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes", *Database 2015*, vol. 2015, pp. bav028, 2015. DOI: 10.1093/database/bav028

[38] J. F. Reyes Román and O. Pastor, "Use of GeIS for Early Diagnosis of Alcohol Sensitivity", *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016)*, ISBN 978-989-758-170-0, vol. 3, BIOINFORMATICS, pp. 284-289, 2016. DOI: 10.5220/0005822902840289

[39] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans and R. A. Harte, "The UCSC genome browser database: 2015 update", *Nucleic Acids Research*, vol. 42, D1, pp. D764-D770, 2015. DOI: 10.1093/nar/gku1177

[40] W. J. Kent, F. Hsu, D. Karolchik, R. M. Kuhn, H. Clawson, H. Trumbower and D. Haussler, "Exploring relationships and mining data with the UCSC Gene Sorter", *Genome res.*, vol. 15, no. 5, pp. 737-741, 2005. DOI: 10.1101/gr.3694705

[41] O. Pastor, S. España, J. I. Panach and N. Aquino, "Model-driven development", *Informatik-Spektrum*, vol. 31, no. 5, pp. 394-407, 2008. DOI: 10.1007/s00287-008-0275-8

[42] L. Munoz, J. N. Mazon and J. Trujillo, "ETL process modeling conceptual for data warehouses: a systematic mapping study", *Latin America Transactions*, IEEE, vol. 9, no. 3, pp. 358-363, 2011. DOI: 10.1109/TLA.2011.5893784

[43] A. Martinez F., "Wiki-Genome: A model-driven genome data management environment", *Sixth International Conference on Research Challenges in Information Science (RCIS)*, IEEE, pp. 1-6, 2012. DOI: 10.1109/RCIS.2012.6240460

[44] O. Pastor, J. C. Casamayor, M. Celma, L. Mota, M. Á. Pastor and A. M. Levin, "Conceptual modeling of human genome: Integration challenges", *In Conceptual Modelling and Its Theoretical Foundations*, Springer Berlin Heidelberg, pp. 231-250, 2012. DOI: 10.1007/978-3-642-28279-9_17

[45] NCBI. *dbSNP ER Schema*, 2015. ftp://ftp.ncbi.nih.gov/snp/database/b124/mssql/schema/erd_dbSNP.pdf

[46] Ensembl. *Ensembl: Features_Analyses_Core Schema*, 2016. http://www.ensembl.org/info/docs/api/core/features_analyses_core.pdf

[47] UCSC. *UCSC Genome Bioinformatics: Haplotypes representation*, 2015. https://genome.ucsc.edu/goldenPath/help/haplotypes.html

[48] J. F. Reyes Román, O. Pastor, F. Valverde and D. Roldán M., "Including haplotypes treatment in a Genomic Information Systems Management", *XIX Ibero-American Conference on Software Engineering (CIbSE 2016)*, pp. 11-24, 2016.

[49] ETL-Tools.Info, "*Proceso ETL*", 2016. http://etl-tools.info/es/bi/proceso_etl.htm

[50] HeidiSQL, "*Basic help on using HeidiSQL*", 2016. http://www.heidisql.com/help.php

[51] O. Heurtel, "PHP y MySQL: domine el desarrollo de un sitio web dinámico e interactivo", *Ediciones ENI*, 2014.

[52] C. J. Date, "Introducción a los sistemas de bases de datos", *Pearson Educación*, pp. 15-19, 2001.

[53] J. P. Febles R. and A. González P., "Aplicación de la minería de datos en la bioinformática", *Acimed*, ISSN 1024-9435, vol. 10, no. 2, pp. 69-76, 2002. [Online]. Available: http://eprints.rclis.org/5150/1/aplicacions.pdf

[54] UCSC. *UCSC Genome Bioinformatics: Schema for Haplotypes - GRCh38 Haplotype to Reference Sequence Mapping Correspondence, 2016.* http://ucscbrowser.genap.ca/cgi-bin/hgTables?db=hg38&hgta_group=map&hgta_track=altLocations&hgta_table=altLocations&hgta_doSchema=describe+table+schema

[55] The Sequence Ontology. *Haplotype (Current_SVN)*, 2016. http://sequenceontology.org/browser/current_svn/term/SO:0001024

[56] L. Phan, "dbSNP and dbVar: NCBI Databases of Simple and Structural Variations", *In Plant and Animal Genome XXIII Conference*, Plant and Animal Genome, 2015. [Online]. Available: https://pag.confex.com/pag/xxiii/webprogram/Paper16208.html

[57] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin, "dbSNP: the NCBI database of genetic variation", *Nucleic Acids Research*, vol. 29, no. 1, pp. 308-311, 2001. DOI: 10.1093/nar/29.1.308

[58] E. García G., G. Lima M., L. Aldana V., P. Casanova C., and V. Feliciano Á., "Alcoholismo y sociedad, tendencias actuales", *Al de Medicina Militar*, vol. 33, no. 3, 2014. Available: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0138-65572004000300007&lng=es&tlng=es.

[59] PubMed web, "Using PubMed", 2016. http://www.ncbi.nlm.nih.gov/pubmed

[60] A. Olivé, "Conceptual modeling of information systems", *Springer Science & Business Media*, ISBN 978-3-540-39390-0, 2007. DOI: 10.1007/978-3-540-39390-0

[61] S. F. Saccone, J. Quan and P. L. Jones, "BioQ: tracing experimental origins in public genomic databases using a novel data provenance model", *Bioinformatics*, vol. 28, no. 8, pp. 1189-1191, 2012. DOI: 10.1093/bioinformatics/bts117