Received September 18, 2020, accepted October 25, 2020, date of publication October 29, 2020,
date of current version November 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034793

Towards the Understanding of the Human Genome: A Holistic Conceptual Modeling Approach

ALBERTO GARCÍA S., ANA LEÓN PALACIO, JOSE FABIÁN REYES ROMÁN, JUAN CARLOS CASAMAYOR, AND OSCAR PASTOR
PROS Research Center, Universitat Politècnica de València, 46022 Valencia, Spain

Corresponding author: Alberto García S. (algarsi3@pros.upv.es)

This work was supported in part by the Spanish State Research Agency, in part by the Generalitat Valenciana under Grant TIN2016-80811-P and Grant PROMETEO/2018/176, and in part by European Regional Development Fund (ERDF).

ABSTRACT Understanding the human genome is a great scientific challenge, whose achievement requires effective data manipulation mechanisms. The non-stop evolution of both new knowledge and more efficient sequencing technologies generates a kind of genome data chaos. This chaos complicates the use of computational resources that obtain data and align them into specific actions. Conceptual model-based techniques should play a fundamental role in turning data into actionable knowledge. However, current solutions do not give a crucial role in the task of modeling that it should have to obtain a precise understanding of this domain. Hundreds of different data sources exist, but they have heterogeneous, imprecise, and inconsistent data. It is remarkably hard to have a unified data perspective that covers the genomic data from genome to transcriptome and proteome, which could facilitate semantic data integration. This paper focuses on how to design a conceptual model of the human genome that could be used as the key artifact to share, integrate, and understand the various types of datasets used in the genomic domain. We provide a full conceptual picture of relevant data in genomics and how semantic data integration is much more effective by conceptually integrating the diverse types of existing data. We show how such a conceptual model has been built, focusing on the conceptual problems that were solved to adequately model concepts whose knowledge is under constant evolution. We show how the use of the initial versions of the conceptual model in practice has allowed us to identify new features to incorporate in the model, achieving a continuous improvement process. The current version is ready to be used as the key artifact in projects where conceptually combining multiple levels of data helps to provide valuable insights that would be hard to obtain without it.

INDEX TERMS Conceptual modeling, CSHG, evolution, genomics, human genome.

I. INTRODUCTION

Conceptual modeling (CM) is essential for designing and developing correct information systems [2]. We defend the use of CM as a fundamental approach to dive into complex domains to extract knowledge and establish a common ontological framework [14]. In this paper, we report our conceptual experience to improve the representation of the genomic domain. This work provides a sound initial understanding of the domain-relevant information that facilitates communication and helps to achieve more efficient genome data management and improve knowledge generation processes. Nevertheless, more conceptual efforts are strongly required when we look at the main challenge of our work: understanding the human genome.

The genome is an example of how immense and complex a domain can be. We use CM to bring a solution to deal with the human genome from a holistic perspective. The complexity of the genomic domain is mainly due to its lack of foundational ontological knowledge and the big data dimension associated with the management of genomic data. There are plenty of terms and elements that are not clearly defined [31]. Additionally, new elements are discovered every year [37], adding more complexity and interconnection to the domain.

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

Moreover, the functionality that is associated with some specific concepts can change as more knowledge accumulates. For example, junk DNA was considered a useless component for many years, but now its functionality appears to be very relevant for the genome execution model [7]. The lack of knowledge of the scientific community has gotten to the point of not knowing exactly how many genes a human cell contains [13], [23]. The result is an immense, ever-changing domain.

Genome data management is also a complex issue. Its origin dates back to fifteen years ago when techniques like Next Generation Sequencing (NGS) appeared. These new techniques allowed us to rapidly increase data generation [39] due to a reduction in sequencing costs [24] and processing times [18]. However, these advances also caused several problems. The first problem is domain heterogeneity: there are plenty of different standards and formats. The second one is dispersion: there are hundreds of different relevant genomic databases, with many of them created or removed every year [36]. Last, the third problem is the lack of interconnection: all of this data is difficult to integrate or interconnect because of the two previously cited problems. These problems are globally referred to as genomic data chaos, which reinforces the need for the systematic use of CM.

Our recent work focuses on exploring how essential CM is for improving domain understanding and guiding the design and development of Genome Information Systems (GeIS) [34]. The main result is an initial Conceptual Schema of the Human Genome (CSHG). After using it in a set of real-world cases [21], [22] and gathering plenty of feedback, we have identified five problems to be treated. In this paper, we characterize these problems, we describe how we have dealt with each one of them, and we report the results of the subsequent discussions, focusing on the ontological commitments established. The main contribution of the work is to present an extended version of a CSHG that is improved and accurate enough to deal with the data management challenges that are associated with the genomic Medicine of Precision practices.

To describe our conceptual work, the rest of the paper is structured as follows: Section 2 describes the state of the art regarding past efforts oriented to use CM to characterize some parts of the genomic domain more precisely. Section 3 reports our previous work on designing a holistic (not partial) perspective of the problem. We will show how different versions of a subsequent CSHG were required to accommodate the conceptual challenges that appear in a working context that is in continuous evolution. Section 4 discusses the questions that have motivated the last version, where relevant changes were incorporated to generate a stable conceptual schema that is ready to be used in practice. Finally, Section 5 ends with our conclusions and addresses further work.

II. CONCEPTUAL MODELING OF GENOMICS

We are perfectly aware of previous attempts to use CM to better understand and communicate genomic knowledge. While the wide range of information that goes from genotypes to phenotype is very complex and covers a set of different and diverse dimensions (genes, variants, sequences, transcripts, proteins, pathways, clinical phenotypes, etc.), existing attempts to deal with the problem cover only part of that wide range of information.

The most relevant current approaches analyze the problem from two perspectives that, while complementary, are different. On the one hand, some works use a pure CM-perspective, introducing specific conceptual schemas that represent part of the domain of interest. On the other hand, other approaches use the term ‘‘ontology’’ to present domain-dependent descriptions that provide a shared‘‘thesaurus’’ or ‘‘data dictionary’’ to delimit what genomic concepts are to be considered relevant in a specific context. Both perspectives emphasize the importance of making explicit conceptualization a common practice to better understand and communicate genomic concepts. However, they only provide a partial view of the whole genome picture

A very interesting initial proposal on CM-oriented approaches was presented by Paton *et al.* [8], [30] modelling the genome from such a CM perspective. They tried to effectively describe the protein interactions and phenotypic consequences of changes in the genome at three different levels, namely, at the transcriptome, proteome and metabolome levels. Unfortunately, this work was incipient in terms of data complexity and it has been discontinued; nevertheless, their ideas have been a source of inspiration for our work.

A very interesting initial proposal on CM-oriented approaches was presented by Paton *et al.* [8], [30] modeling the genome from such a CM perspective. They tried to describe the protein interactions and phenotypic consequences of changes in the genome at three different levels, namely, at the transcriptome, proteome, and metabolome levels. Unfortunately, this work was incipient in terms of data complexity, and it has been discontinued; nevertheless, their ideas have been a source of inspiration for our work.

Ram [32] used CM to model proteins using an annotation-based approach. Their goal was to search and compare proteins through their 3D structure. The work consisted of defining the semantics of primary, secondary, tertiary, and quaternary structures of proteins. To do this, they described the protein's components, chemical bonding forces, and spatial arrangement along with its associated biological information. The resulting model facilitated the development of user-friendly tools to search and compare proteins by their structure. But again, only a partial dimension of the extensive genome information spectrum was considered: the one related to protein characterization.

More recently, Bernasconi *et al.* [6] have characterized processed genomic data applying CM techniques. They propose a CS to deal with the experimental datasets (genomic data and metadata) used in scientific publications. Their CS describes biological, technological, and management aspects of the experimental datasets. It is an appropriate approach for standardizing the metadata of experimental datasets and improving genomic data integration. However, this solution again focuses on a partial aspect of genomic data, namely, genomic sequence information through experimental datasets used by the scientific community.

Médique *et al.* [25] presented a co-operative computer environment called ''Imagenetrade mark'' developed by applying an object-based model. The tool allows researchers to analyze and annotate genomic sequences. Going further in our argument of partial coverage of genomic information, their paper thoroughly explores well how to model sequences, but only sequences.

These works have provided a valuable contribution to the application of CM in the genomic domain. Even though the reported exercises of using CM to better understand specific parts of the human genome are useful and interesting, they all focus on a particular dimension of the domain. Consequently, there is not a unique, unified ontological commitment, making the integration of information and communication difficult.

Besides these pure CM approaches, the so-called genomic ontologies make up a family of complementary solutions. One significant representative of this approach is the Open Biological and Biomedical Ontology (OBO) Foundry [38]. OBO is an entity whose mission is to provide a set of design ontology principles. Hundreds of so-called ontologies have been defined following their principles, from which tens are already obsolete. The so-called OBO ontologies are loosely hierarchical directed acyclic graphs, e.g., a concept may have more than one parent term. They try to organize domain knowledge into two different dimensions: granularity and relation to time. In our analysis, five selected ontologies that are widely known and that are more related to our work were explored: Human Phenotype Ontology (HPO) [19], Gene Ontology (GO) [3], Sequence Ontology (SO) [16], Protein Ontology (PRO) [27], and Variation Ontology (VO) [40]. The analysis of these five ontologies allows us to show how our ''partial genomic view'' claim is present in them since each ontology focuses on just a given, specific genome dimension, and the whole picture is missing.

- Human Phenotype Ontology (HPO) focuses on phenotype properties. It aims to provide a standardized set of terms to describe phenotypes encountered in humans. It is based on medical literature and contains over thirteen thousand terms. It is used to support differential diagnostics in translational research. There are modifiers to represent the speed of progression, inheritance modes, and frequencies of phenotypes.
- Gene Ontology (GO) focuses on gene characterization. The functionality of genes is studied by providing a standardized vocabulary. More than 40,000 defined terms across 4,500 different species are defined. GO divides its information into three domains: the molecular-level activities performed by gene products, the location inside a cell where a gene product performs a function, and biological processes that are composed of multiple molecular-level activities.
- Sequence Ontology (SO) analyzes genome sequences. It provides a structured and controlled vocabulary to distinguish different sequences of our genome that depend on their positions. For example, a binding site is defined as a region that interacts selectively and non-covalently with other molecules.
- Protein Ontology (PRO) specifies protein-related entities and the relationships between them. PRO uses a system of classification called ''levels of distinction''. There are four levels of distinction: the family-level refers to protein products produced by genes with a common ancestor; the gene-level separates protein product by gene; the sequence-level is used to differentiate protein products that are generated from the same gene but with different alleles in its sequence; the modification-level separates gene protein products that differ due to cleavage or chemical changes to one or more amino acid residues.
- Variation Ontology (VO) provides a standardized description of effects, consequences, and mechanisms of variations. It aims to define unambiguous definitions of variation effects, described at a DNA, RNA, or protein level.

These OBO ontologies help in reducing domain heterogeneity by providing well-defined standards for some specific concepts of the domain. While this is true, there is conceptual vulnerability: they are concept-specific thesauruses of terms or classification systems. These ontologies do not provide a common, clear, and ontological definition of concepts, as [42] explores. Besides, each ontology focuses on a particular part of a specific genomic dimension, and, as a consequence, there is not an explicit link among them. Different OBO ontologies can characterize two related concepts without specifying how they are linked, or one common concept can be represented differently in alternative ontologies. In addition, each ontology focuses on a particular part of a specific genomic dimension, and, as a consequence, there is not an explicit link among them. Different OBO ontologies can characterize two related concepts without specifying how they are linked, or one common concept can be represented differently in alternative ontologies. For instance, phenotypes are characterized in the Unified Phenotype Ontology (upheno).[1] However, the Mammalian Phenotype Ontology (mp)[2] also describes phenotypes, but only the mammalian ones. Going further regarding phenotype characterization, there are plenty of additional ontologies such as the Mouse pathology Ontology (mpath),[3] the Human Phenotype Ontology (hpo),[4] or the Neuro Behavior Ontology (nbo).[5] As can be observed, multiple ontologies try to characterize phenotypes in different ways, but their elements are not interconnected.

[1] http://www.obofoundry.org/ontology/upheno.html
[2] http://www.obofoundry.org/ontology/mfmo.html
[3] http://www.obofoundry.org/ontology/mpath.html
[4] http://www.obofoundry.org/ontology/hp.html
[5] http://www.obofoundry.org/ontology/nbo.html

The lack of a unified, holistic perspective is a common aspect shared by the existing works in the genomic domain. We realized that all of the semantic components of the different conceptual schemas cannot be connected under a common, ontologically well-grounded view. Such a holistic perspective is the core of the contribution of our work. We start with a review of the initial versions of our unified, holistic CSHG in III, and we introduce our final, usable version in Section IV.

III. EVOLUTION OF THE CONCEPTUAL SCHEMA OF THE HUMAN GENOME

The CSHG [33] provides the holistic perspective of the human genome that conforms to the primary goal of this work. It also provides the grounded conceptual background that is needed to characterize the relevant concepts that could make the understanding of the human genome viable. It is a complex task because the human knowledge of genome intrinsics is under constant evolution. Not only that, even the precise characterization of fundamental concepts is under continuous discussion. For instance, conceptually speaking, what exactly is a gene? This concept accepts different definitions, as we will see later. Selecting the correct one determines what conceptual schema is to be designed and, subsequently, the data analysis strategies that are associated with the database schemas generated from the conceptual schema.

This is why the conceptual evolution of our work has advanced, while the ontological commitment associated with the genomic concepts was also changing. Specifically, here we report how the CSHG has been updated twice since its creation (CHSG v1), and we explain why this happened. A description of each version of the CSHG is found below, which provides a clear insight into how important characterizing the human genome conceptually really is:

CSHG Version 1 [29]: The goal of this preliminary version was to model the most basic concepts of the human genome from a unified perspective. Version 1 proposed a gene-centered vision in which genes are the central and most important unit of the human genome. Genes sequences conform to the structural unit of description. Therefore, this version focuses on analyzing individual genes, their mutations, and the consequences of these changes from a global perspective. It is divided into three main views: "Gene-mutation view", "Genome view", and "Transcription view". The Gene-mutation view models how genes are structured and represents allelic variations of genes. The Genome view incorporates individual genome representations. Finally, the Transcription view models the basic components participating in protein synthesis that can be affected by the allelic variations of the genes.

CHSG Version 1.1 [33]: This version included phenotypic information in a new view (Phenotype view) to provide more consistency and completeness to the model. The genotype-phenotype relation reinforces a holistic perspective by explicitly representing how variations are related to phenotypes. On the one hand, the new view models phenotypes, their classification, and their severity; on the other hand, it models how variations are related to phenotypes.

CHSG Version 2 [34]: Version 2 of the CSHG drastically changes how the genome sequence is comprehended and, therefore, represented. Version 1 was manipulating sequences of genes. As we obtained more and more experience in the practical applications of the CSHG, we realized that frequently DNA sequences were coming from other genome structures. This set of potential genome structures was diverse (not only genes but other RNAs, promoters, enhancers, proteins, or transcripts, among others) and open to continuous evolution (when new concepts are proposed, like oncogenes).

Furthermore, we detected that there was a lack of consensus in the scientific community when trying to precisely define what a gene is or how it can be characterized. All of that together motivated us to move to what we have called version 2, a new CSHG perspective that is more "chromosome-centric" than "gene-centric". By "chromosome-centric", we mean that any relevant sequence of the human genome is represented as a part of a chromosome: a chromosome element. This gives the expressive facility of modeling any relevant genome component as a part of a chromosome, providing its sequence.

This version is divided into five views: the "structural view", which describes the structure of the human genome; the "transcription view" which models protein synthesis; the "variation view" which characterizes changes in the sequence of the human genome; the "bibliography view" which details information and sources related to elements of the CS; and a new view, called "pathway view", which represents human metabolic pathways, thereby increasing the holistic perspective of the CS.

IV. FROM VERSION 2 TO VERSION 3: CSHG V3

CSHG v2 has been used in a set of real-world use cases for two purposes. First, to validate its correctness and usefulness, Second, to gather and analyze the feedback of genomic domain experts to improve CSHG, ensuring that it is updated by the latest scientific discoveries. As a consequence, a set of problems that should require a more precise conceptual characterization has been identified:

1) We realized that some concepts were too tied to a specific solution (associated with a particular technological implementation). For instance, some attributes of the Gene class refer to specific identifiers of database systems. The "id_hugo" is a unique identifier from the gene class provided by the HGNC database.[6] It is not a universal identifier, and it is not shared among other databases. As a consequence, our CSHG v2 lacks the flexibility to be adapted to different technological platforms when working with new information. In the previous example, if the HGNC database is replaced by a new one, the conceptual schema itself must be

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

updated, together with the implications in its associated database. The question to be answered in this context is the following: *Should specific data source attributes exist in our CHSG, or should universal identifiers be used, independent of any particular data source?*

2) The reference sequence of the human genome is not carved in stone. It is fully dynamic and evolves according to both technological limitations and improvements. These changes are identified by a version number, and multiple versions coexist over time. Having multiple versions of the reference sequence of the human genome implies that the relative position of variations (that are identified and located for a specific version of the human genome) is affected. Without incorporating the notion of "Assembly" in the CS, our CS does not allow multiple versions of the genome to be modeled. A significant question emerges naturally: *Should our CSHG represent multiple genome assemblies, or should an independent instance of the model be generated per assembly?*
3) CSHG v2 represents variations regarding their type and frequency among populations. This representation exhibits limitations, models redundant information, and is over complicated. There are limitations because, in frequency classification, a variation can be either a polymorphism or a mutation and only *Single Nucleotide Polymorphisms* (SNP, a type of polymorphism) have genotype and population information. This means that if a variation is not an SNP, neither genotype nor population information regarding the variation can be described. There is redundancy because variations are represented **at least** twice (i.e., by frequency and by population). Furthermore, nothing prevents a variation from being represented more than twice depending on the frequency among **multiple** populations. For instance, depending on the frequency of appearance, a variation can be defined as a polymorphism in a given population and as a mutation in another one, leading to a variation represented thrice: once per type and twice per frequency. Last, there is complexity because the definition of exclusive disjunction XOR rules is needed to ensure the correctness of the data: a variation that is not precise (type) should never be a polymorphism (frequency); a polymorphism (frequency) should never be an imprecise variation (type); etc. The consequence of this discussion is that *a better classification of variations (i.e., removing limitations and redundancy and reducing complexity) is needed.*
4) CSHG v2 represents the phenotypic effects caused by variations. However, lower-level effects in the organism are not represented (e.g., alterations in the structure of a protein). Adding another way of modeling the effects caused by variations increases the completeness of the model. It also allows domain experts to use our CSHG more precisely (e.g., by identifying variations based on the structural changes they cause at a proteome level). The key question is: *how can the effects caused by variations be enriched in our CSHG?*
5) CSHG v2 misses some concepts and relations that play a relevant role in the transcription process. For instance, the mRNA concept (the intermediate product between the genome and the proteome) is not represented. Additionally, the transcription process is modeled so that transcriptable elements only produce proteins, which is a *correct* but *incomplete* assumption since transcription produces plenty of additional elements, such as non-coding RNA. This problem leads to the following question: *how can the transcription process be enriched so that it is not only correct but also complete?*

In the following Sections, we describe each of the enumerated problems in more detail, explaining how they have been solved and discussing the associated ontological commitments. The development of these five ideas has led us to the evolution of CSHG v3, the main contribution of this paper.

A. A MORE AGNOSTIC APPROACH

We observed that our CSHG v2 lacks *flexibility* because by having concepts with attributes that are associated with a specific solution, two limitations arise. First, it biases domain users to use only those data sources whose identifiers are represented in the CS. However, depending on the working context, some of these attributes may not be used or cannot be obtained. Second, working with new data sources is a problem because it is not possible to model their identifiers without updating the CS itself and its database implementations. As a result, the efficiency of the analysis processes is reduced because domain users focus on how to deal with the data rather than generating knowledge. Therefore, our CSHG needs to avoid references to specific solutions because they are useless outside their particular working context. Three elements represented in multiple data sources are *chromosome elements*, *variations*, and *populations*. They do not always share the same identifier, and they reference specific solutions.

Apart from the limitations explained above, let us use the *variation* concept as an example to demonstrate additional **problems** caused by having concepts specific to particular solutions rather than using a more generic approach. As Figure 1 shows, variations are linked to only one data source (Data bank version) and have five attributes that are identifiers of specific solutions. On the one hand, three of them are used to link variations with chromosome elements: *NC_identifier* links a variation with a chromosome, and both *NG_identifier* and *id_hugo* link the variation to a gene. On the other hand, two of them are used as variation identifiers: *rs_identifier*, which is an identifier used in the Ensembl[7] data source; and *other_identifiers*, which is used

[7] https://www.ensembl.org/

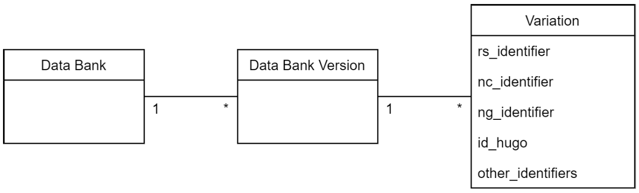

FIGURE 1. Old representation of multiple variation identifiers.

to store HGVS expressions[8](a nomenclature for describing sequence variations) and other identifiers. The **first problem** is which identifier to use if a variation is not in the Ensembl data source. The **second problem** is that there is no easy way to determine how many identifiers a variation can have because some of them can be null, and the *other_identifiers* attribute can store an arbitrary number of identifiers. The **third problem** is that there is a loss of information. Variations are linked to only one data source, but they can store more than one data source identifier in *other_identifiers*, such as Ensembl, ClinVar, gnomAD, PharmGKB, or ClinGen. Therefore, it is not possible to determine to which data source an identifier pertains.

To solve the limitations and problems presented above, an abstraction mechanism to include any data source-specific identifier has been modeled. Therefore, *chromosome elements*, *variations*, and *populations* are not directly linked to a single data version bank anymore. Instead, a new concept that links variations and data bank versions through its specific identifier has been created. Figure 2 shows the new representation of variations as an example. This new approach solves the **problems** regarding variations described above. Variations no longer require attributes that are associated with a specific solution, it is easy to determine how many identifiers a variation has, and there is no loss of information since each identifier is linked to its corresponding data source.

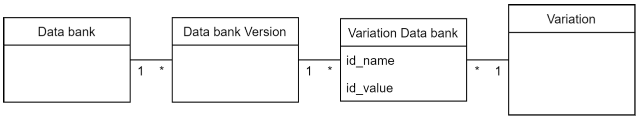

FIGURE 2. New representation of multiple variation identifiers.

This approach removes any possible bias by not explicitly representing any attribute associated with specific solutions, and it gives domain users the needed flexibility to model any new data source without having to update the CS. In conclusion, our CSHG is as generic and technological independent as possible: **Universal identifiers that are independent of any specific data source are used as primary identifiers, while an abstraction mechanism is provided to include as many data source-specific identifiers as possible**.

B. HOW TO MODEL MULTIPLE ASSEMBLIES

Since the first version of our CSHG, a fundamental question has insistently prowled around in our mind: *How should the sequence of the human genome be modeled?* CSHG v2 models the human genome assuming that only one sequence of reference is considered, i.e., each chromosome has a unique sequence. Therefore, chromosome elements and variations are located in one and only one position in the reference sequence. However, a relevant limitation was identified by domain experts when working with the CS: *in the real world the reference sequence of the human genome is not unique, and multiple versions coexist. Moreover, domain experts need to work with more than one of them at the same time*. For instance, the rs11571636 variation [1] is located at different positions depending on the version of the human genome sequence: chromosome 13, position 32.905.026 in the GRCh37[9] assembly, and chromosome 13, position 32.330.889 in the GRCh38[10] assembly.

The coexistence of multiple versions of the reference sequence of the human genome is caused by the existing limitations regarding sequencing technologies (assembly software) in the genomic domain. Currently, it is not possible to read the entire sequence of the genome at one time. Instead, it is split into many smaller sequences that are read multiple times to form overlapping sequences. The sequences with the highest quality are joined to form *contigs*, and the contigs are joined to form *scaffolds* (non-contiguous contigs that are separated by gaps of known length but unknown sequence) [41]. These scaffolds compose the *assembly*, which contains the *chromosome sequences*, like the GRCh37 or the GRCh38 [17].

To solve this problem, we have analyzed three different approaches below. Each of them contains a figure that illustrates the modifications performed on the CS and an example of how the resulting CS is instantiated. For simplicity, the provided examples only represent one chromosome (chromosome 13) and two assemblies (GRCh37 and GRCh38).

The first approach generates a new instance of the CS for each assembly (see Fig. 3). Although this approach allowed us to have multiple assemblies, it was discarded for three reasons. First, the concept of the *assembly* was not represented in the CS, and it should be represented since it is an important and widely used concept in the working domain. Second, domain users need to work with multiple assemblies *at the same time* because variations of interest can be identified in one or several assemblies. This fact should be properly represented in the CS. Third, this approach introduces unnecessary redundancy since, for any new instance of the CS, the rest of the information is duplicated.

The second approach includes the concept of the *assembly* into the CS and instantiates a set of chromosomes per assembly (see Fig. 4). This approach solves the problems

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

[9] Reference sequence of the human genome, build 37 https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/

[10] Reference sequence of the human genome, build 38 https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/

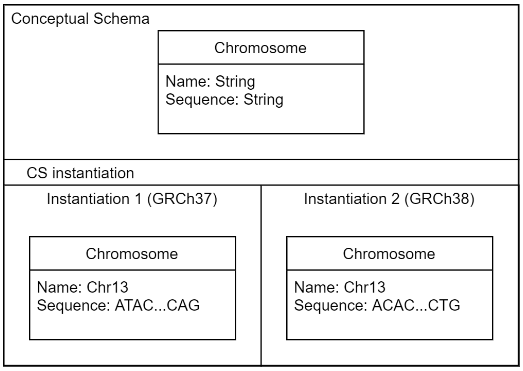

FIGURE 3. The first approach: a CS instantiation per assembly.

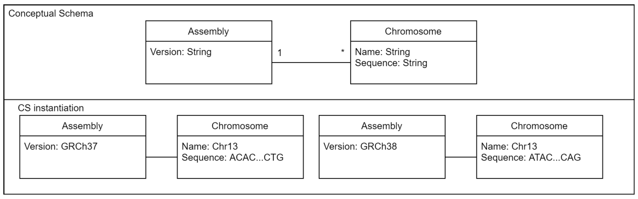

FIGURE 4. The second approach: assembly as a concept itself.

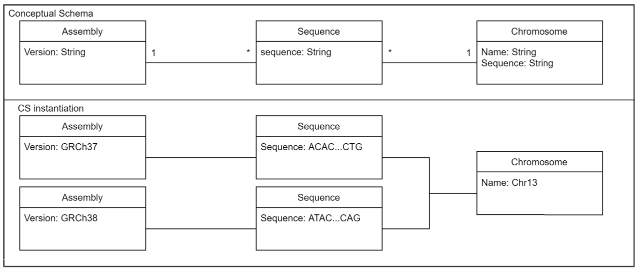

FIGURE 5. The third approach: assembly and sequence as concepts themselves.

of the first approach since the concept of *assembly* is represented in the CS, multiple assemblies can be represented and worked out at the same time, and unchanged parts of the CS are not CS are not duplicated. Although this allowed us to have multiple chromosome sequences per genome, this approach is not conceptually precise. The reason is that the chromosome concept is a *unique* biological entity and should not be instantiated multiple times. Indeed, what changes and should be instantiated multiple times is its *sequence*, which is based on the assembly it is linked to.

The third approach consists of extracting the *sequence* from the chromosomes as a new concept (see Fig. 5). This approach solves the problems of the second approach while keeping its advantages since chromosomes are instantiated only once. With this approach, a chromosome, *as a unique entity*, has several sequences depending on the existing assemblies. We concluded that this is the most suitable and conceptually accurate approach regarding assemblies, chromosomes, and their sequences.

Having chromosomes with multiple sequences of reference means that the ''chromosome element'' and ''precise variation'' classes are no longer located in one and only one position. This change causes a loss of contextual information because it is not possible to know the locations of these classes among the different assemblies. To deal with this, the locations of these classes have been extracted into a new class. Consequently, it is properly identified in the different assemblies (i.e., they are no longer located once, but rather once per assembly). In conclusion, **the CSHG is now capable of representing multiple genome assemblies with their corresponding chromosome sequences**. With these changes, the model is more accurate and flexible and can facilitate domain experts' work. Moreover, sequencing technologies in the genomic domain are improving rapidly, and now the CSHG will be able to deal with a growing number of assemblies.

C. A NEW WAY OF REPRESENTING VARIATIONS

The concept of variation is classified based on two criteria: its frequency and its type (description). Concerning its frequency, a variation is a polymorphism if it appears with a higher than one percent frequency in a given population; otherwise, it is a mutant variation. A Polymorphism is a *Single Nucleotide Polymorphism* (SNP) if the variation only changes one nucleotide; otherwise, it is a *Copy Number Variation* (CNP). Concerning its type, a variation is a *precise variation* when its location and change are known; otherwise, it is *imprecise*. A precise variation can either be an insertion, a deletion, an indel (insertion and deletion), or an inversion (if a region of the genome sequence is inverted). Three issues have been identified as a consequence of this classification.

The first issue is that this representation has **limitations** regarding four *concepts* tied to SNP variations:

1) Population: a set of individuals that share a common characteristic whose genome has been sequenced to find variations.
2) Haplotype: a group of SNPs that tend to occur and be inherited together and can be linked to a specific disease.
3) Allele frequency: the frequencies of appearance of the alleles of a variation in a specific population. The alleles of a variation are the list of possible nucleotides of a variation. It includes both the allele of reference (unique) and the potential alternative alleles. For instance, the list of alleles of a variation that changes an A for a T is A (reference allele) and T (alternative allele).
4) Genotype frequency: the frequencies of appearance of the genotypes of a variation in a specific population. The genotypes of a variation are the combination of two (as we have two copies of each chromosome) of the possible alleles of a variation. The genotypes of the previous example are AA (reference homozygote), AT (heterozygote), and TT (alternative homozygote).

These *concepts* are linked to SNP variations in version 2 of the CS because, historically, genomic population studies have only been applied to them. However, this has changed due to advances in sequencing technologies. Domain experts reported that they need to work with population information regarding variations that are not SNPs, which is not possible with the CSHG v2. Therefore, any precise variation should be able to store information about these four *concepts*.

The second issue is that the CS models the concept of variation in a **redundant** way. A variation is modeled at least twice: once per frequency and once per variation. For instance, the rs11571636 variation is a precise variation (type), more specifically, an indel. It also is a mutant variation (frequency) since it does not appear in any population with a frequency higher than one percent. This issue worsens when new information regarding the variation is obtained. For instance, the rs11571636 variation will be modeled thrice if a genomic population study shows that it is an SNP in a specific population: once per type (indel) and twice per frequency (Mutant and SNP). Domain experts found it to be counter-intuitive to have the same variation instantiated thrice. They also indicated that this approach was overcomplicated, which leads us to the third issue, discussed below.

The third issue refers to the **complexity** of the CS. The definition of exclusive disjunction XOR rules are needed to ensure the correctness of the data: a polymorphism (frequency) can only be precise (type); an imprecise (type) can only be a mutant (frequency); an SNP (frequency) can only be an indel (type) whose change is one nucleotide; a CNP (frequency) can only be an insertion (type). These rules confuse domain experts and add another layer of complexity in the management of the data.

The efficiency of the analysis processes is reduced as a consequence of these issues. Apart from that, an additional conflict arises since the terminology used in the CS is controversial. The terms mutation and polymorphism lead to confusion as incorrect assumptions are made: the term mutation is assumed to have pathogenic effects, while polymorphism is assumed to have benign effects. Therefore, the term variation is preferred over mutation and polymorphism [35].

A different approach has been used to overcome these problems. In the CSHG v3 population, allele frequency, genotype frequency, and haplotype *concepts* are linked to the precise variation class instead of the SNP class. This change solves the *first issue* because now any precise variation can store information about population and frequencies.

After this change, should variations still be classified based on the frequency criterion? There are two reasons not to keep this classification in the CS. First, since, with the new associations, any precise variation can have information regarding population and frequencies, this classification loses its original reason. Second, the terms used in the frequency classification, namely polymorphism, and mutant, are controversial. Therefore, specialization by frequency is not required anymore, and variations are only classified based on their type in the CSHG v3. This change solves the *second issue* since it removes the redundancy from the CS.

The *third issue* has been solved because, with the two previous changes, there is no need to define exclusive disjunction XOR rules to ensure the correctness of the data in the CSHG v3.

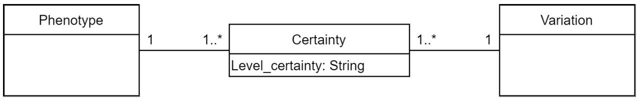

FIGURE 6. Representations of the effects caused by variations in the CSHG v2.

As a result of these changes, **the classification of variations has been improved**, and its completeness is reinforced: **the former limitations of the CSHG v2 have been overcome** because any precise variation in v3 can store information regarding alleles, genotypes, haplotypes, and populations. **The representation of variations is no longer redundant** because variations are only classified based on the type criterion. By solving the two previous limitations, **the complexity of the CS has been reduced naturally**. The CS gained simplicity and expressiveness through an exercise of concept reevaluation and simplification.

D. MODELING THE EFFECTS CAUSED BY VARIATIONS

Our CSHG v2 represents the effects caused by variations at a phenotypic level that focuses on the *consequences* of the structural changes. But it does not represent the structural changes themselves. It is a high-level perspective based on the premise that phenotypes emerge as the result of genotype-environment interactions [5]. Variations are linked to phenotypes with a given certainty (see Fig. 6). This certainty is provided by domain experts (submitters), and it indicates how strong the evidence supporting a variation-phenotype link is. There are six levels of certainty [11]:

1) Level one: a variation-phenotype link without any evidence.
2) Level two: a variation-phenotype link with evidence provided by a single submitter.
3) Level three: a variation-phenotype link with evidence provided by multiple submitters whose interpretations conflict.
4) Level four: a variation-phenotype link with evidence provided by multiple submitters without interpretation conflicts.
5) Level five: a variation-phenotype link validated by a panel of experts.
6) Level six: a variation-phenotype link with evidence obtained from following a set of strictly-defined practice guidelines.

Three improvements have been identified regarding how the effects caused by variations are modeled in the CSHG v2. First, as explained above, the level of certainty depends on the evidence provided by *submitters*, but they are not modeled in the CS. On the one hand, It is relevant to know **who** submits the evidence supporting variation-phenotype links. On the other hand, it is necessary to know **when** the evidence was provided. Therefore, the CS must include submitters.

Second, the CSHG v2 does not model the variation's pathogenicity regarding the phenotypes they are linked to. For instance, Sickle cell disease (SCD) is caused by blood cell deformation. What is the pathogenicity of a variation that is linked to CSD? Is it protective (reducing the chances of suffering CSD) or pathogenic (causing CSD)? (i.e., what is its *effect*?) Domain experts require this information to conduct their analysis processes. Therefore, its addition is needed.

Third, the actual perspective is unable to describe the effects caused by variations at a structural level. For instance, the CSHG v2 identifies variations linked to CSD, but it does not identify the specific structural change that deforms the blood cells. Therefore, a complementary perspective that models those changes is required.

The first improvement is performed by adding the submitters to the CS (see Fig. 7). Now, any variation-phenotype link is provided by a submitter on a specific date. As a result, when a variation has multiple interpretations that conflict, the domain expert can identify **who** provided them and **when**.

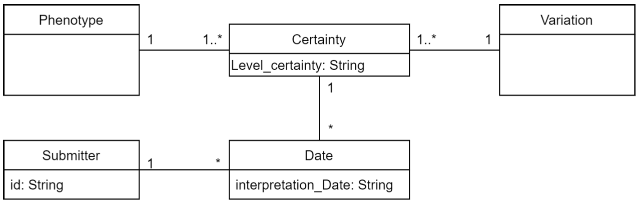

FIGURE 7. Inclusion of the submitters.

The second improvement is done by adding the pathogenicity of the variations regarding the phenotype they are linked to (see Fig. 8). A new attribute called "clinical significance" has been added to the Certainty class. It has been defined following the ClinVar recommendations [10]. They include, on the one hand, the ACMG/AMP recommended terms [35] and, on the other hand, additional terms to provide more precise links between variations and phenotypes. The most commonly used clinical significance terms are:

1) Benign: a variation that is not responsible for causing a particular phenotype.
2) Pathogenic: a variation that is responsible for causing a particular phenotype.
3) Protective: a variation that decreases the factor of a phenotype.

The third improvement is done by including a new, low-level representation of the structural changes caused by variations (see Fig. 9). It has been modeled following the VCF standard [4] and its annotation field [9]. A new class, called "annotation", identifies the "impact" and the "effect" of the variation alleles in chromosome elements and transcripts. It links the alleles of variations rather than the variations because each allele can have a different effect. Based on its putative impact, the annotation's impact classifies the effect of a variation into the following:

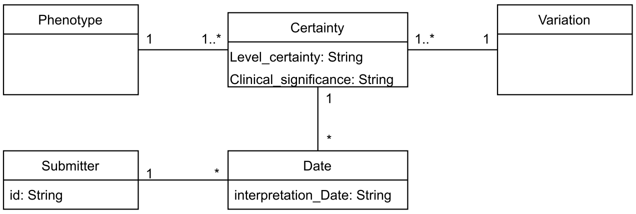

FIGURE 8. Inclusion of the clinical significance.

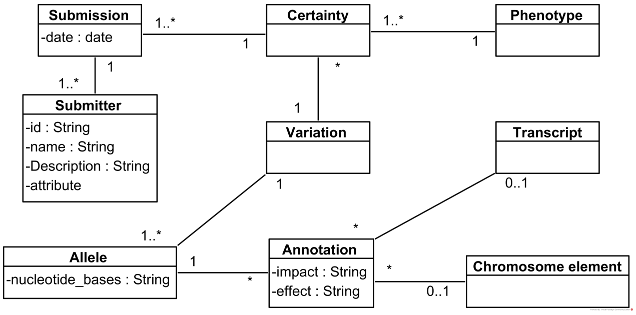

FIGURE 9. Inclusion of the structural changes.

1) HIGH: A disruptive change is triggered. A disruptive change causes a chromosome element truncation or a loss of transcript functionality. For instance, a stop codon prevents a protein to from being completed.
2) MODERATE: A non-disruptive change is triggered. For instance, an enzyme's sequence is modified, so its reactions are carried out less effectively.
3) LOW: A harmless change is triggered. For instance, a variation changes a CTT codon to a CTC one: both codons are translated to the same amino acid (Leucine).
4) MODIFIER: This is a special case where the variation is located in a non-coding region.

On the other hand, the annotation's effect is much more specific because it describes the exact effect of a variation, e.g., the duplication of a gene or the appearance of a termination codon in the sequence of a transcript.

The CSHG v3 **improves the description of the effects caused by variations**. First, it is enriched by including the variation's pathogenicity and the submitters in the CS. Second, a new, low-level approach that focuses on structural changes and its implications at a genome and proteome level is added. As a consequence, the holistic perspective of the CSHG is increased, and the efficiency of domain expert analysis processes is boosted.

E. RETHINKING THE GENE EXPRESSION PROCESS

The protein-coding process is modeled at three levels in the CSHG v2, namely, DNA, RNA, and amino acid (see Fig. 10). The first level contains chromosome elements: on the one hand, we have the protein-coding genes and their exons (gene parts that are transcribed) and introns (gene parts that are not transcribed). On the other hand, we have regulatory elements. The second level contains transcripts, which are the result of gene transcription. The third level includes proteins, which are the result of transcript translation. In this context, we emphasize the following three issues:

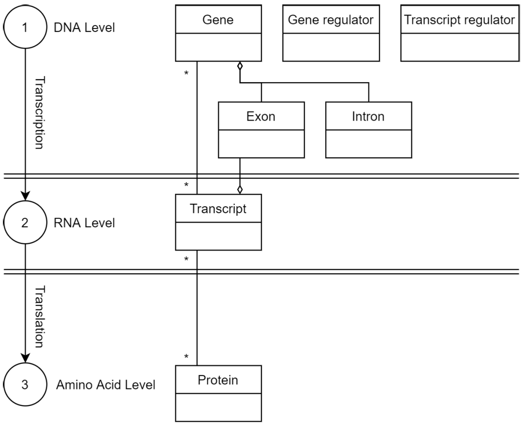

FIGURE 10. Representations of the transcription process in the CSHG v2.

1) Inspired by the *central dogma of molecular biology* [12], the CS assumes that genes only code for proteins. However, this assumption is not complete since **they also produce additional products called non-coding RNA (ncRNA)** that do not code for proteins [15]. Unlike protein-coding RNAs, ncRNAs are linked to diverse regulatory functions and are specific to different organs or tissues. The fact is that ncRNAs are increasingly gaining attention in the field of precision medicine [28]. Consequently, the modeled transcription process needs to be updated to consider those additional elements produced by genes rather than only proteins.
2) Domain experts observed that **the concept of messenger RNA (mRNA) is not present in the CSHG v2.** It is *transcribed* from genes and *translated* into proteins. This concept is represented as "transcript" in the CSHG v2. However, after a series of discussions, it has been determined that the mRNA is a *type* of transcript and that there are additional types of transcripts, like the ncRNA of the previous issue, which are not translated into proteins but are semantically relevant. Therefore, the mRNA concept should be represented in our CSHG.
3) There is a misunderstanding regarding regulatory elements in the transcription process. There are two types, the ones that exist at a DNA level and the ones that exist at the RNA level. **Both of them are modeled as chromosome elements at the DNA level**, which is not correct because RNA regulatory elements do not exist at the DNA level; they only exist at the RNA level. Besides, the CS does not model *what* is regulated by them. For instance, **what** gene is regulated by a given enhancer? Therefore, their conceptualization must be improved.

A series of changes in the CS have solved the first two issues. First, the concepts of mRNA and ncRNA have been included as types of transcripts (see Fig. 11). An mRNA is defined as a transcript that codes for proteins, while ncRNA is defined as a transcript that does not. Examples of ncRNA include transference RNA (tRNA), ribosomal RNA (rRNA), or long ncRNA (lncRNA). As a consequence, genes no longer code only for proteins but also for ncRNAs, and now proteins are translated from mRNAs rather than from the generic concept of "transcript".

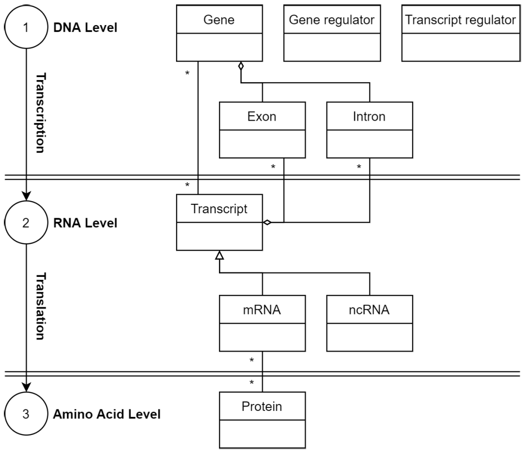

FIGURE 11. Addition of mRNA and ncRNA to the transcription process.

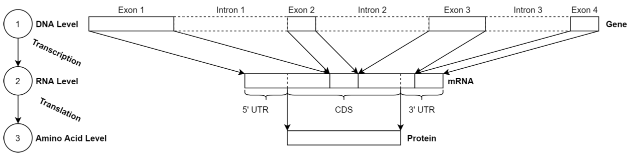

FIGURE 12. The protein-coding process of an mRNA.

Second, the parts of the mRNA sequence have been modeled. It is important to characterize its structure. It is composed of three elements: one coding sequence (CDS) and two untranslated regions (UTR). The *CDS* is the sequence of the mRNA that is translated into the protein. The UTRs, which are a *3' UTR* region before the CDS sequence and a *5' UTR* region after it, are not translated.

Third, transcripts are no longer only composed of exons for two reasons: i) even if it is true that mRNA is usually composed of gene exons, there is a mechanism, called intron retention (IR) [26], which is responsible for not removing introns in the transcription process; and ii) ncRNA transcripts are obtained from introns. Therefore, transcripts are now composed of exons and introns with a minimum cardinality of zero, which allows us to represent the following:

1) A transcript composed of exons (Fig. 12): It is the basic protein-coding process that has two steps. In the first step, a set of exons is transcribed into a mRNA transcript and the introns are discarded. In the second step, the CDS sequence is translated into the protein and the UTR regions are discarded. The protein can then perform its activities.
2) A transcript composed of exons and introns (Fig. 13): The process is the same as the one explained above but with the additional consideration that an intron can also be transcribed due to the IR effect. Consequently, the mRNA sequence is changed and the translated protein is different.

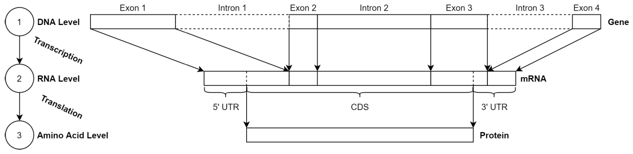

FIGURE 13. **The protein-coding process of an mRNA with IR.**

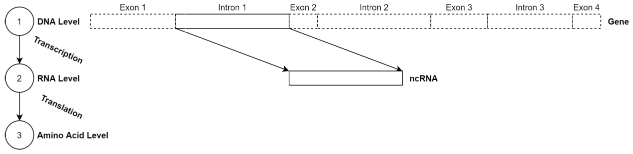

FIGURE 14. **The ncRNA-coding process.**

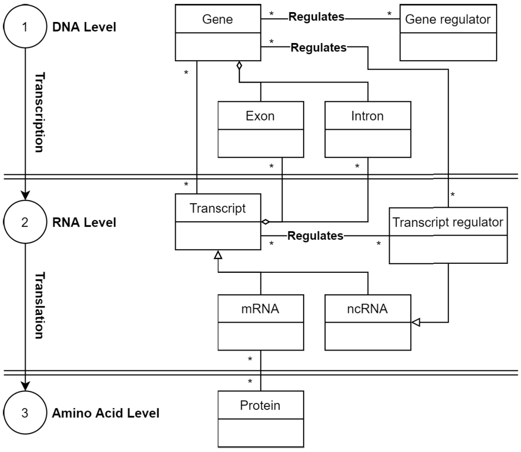

FIGURE 15. **Reevaluation of the regulatory elements of the transcription process.**

3) A transcript composed of introns (Fig. 14): It is the way ncRNAs are generated. An intron of the gene is transcribed into a RNA sequence that will not be translated. After the transcription, the ncRNA can perform its activities.

To solve the third issue, a complete reevaluation of regulatory elements has been performed in the CSHG v3 (see Fig. 15). First, regulatory elements are now divided into DNA and RNA regulatory elements. **This classification does not indicate what type of elements they regulate but rather at which level they exist**. Therefore, it is important to note that an RNA-level regulatory element can regulate DNA-level elements. For instance, an ncRNA can prevent the initiation of the transcription process of a gene.

DNA-level regulatory elements are "passive" because they do not actively regulate gene expression. Instead, they are specific regions where "active" regulatory elements bind. There are three types: promoters, enhancers, and silencers [20]. Promoters are the gene transcription starting point. Enhancers can be bound by activators to boost gene transcription. Silencers can be bound by repressors to inhibit gene transcription.

RNA-level regulatory elements are "active" because they bind to the "passive" ones and actively regulate gene expression. On the one hand, they act at the RNA level. For instance, micro RNAs (miRNAs) inhibit gene expression by increasing mRNA degradation speed. On the other hand, they also act at the DNA level. For instance, lncRNAs bind to silencers to prevent gene expression.

As a result, the transcription process has been enriched and is more complete and versatile. Can a protein-coding gene be modeled? Can an ncRNA coding gene be modeled? Can a gene that codes for both protein and ncRNA be modeled? The answer to these questions is a resounding yes. In addition to that, regulatory elements are classified into DNA-level and RNA-level elements, and it is possible to know what elements they regulate. The CSHG v3 offers a far more precise representation of reality.

V. CONCLUSION AND FUTURE WORK

A correct interpretation of genome data greatly requires getting a shared understanding of the domain. Such a complex problem cannot be faced without the use of CM techniques. In this paper, we have used a CS to demonstrate the benefits that can be obtained: First, it improves the communication with domain stakeholders, i.e., doctors, geneticists, or biologists. Second, it eases knowledge transference by having a shared ontological commitment to discuss. Third, it provides a solid background for developing better software solutions. As a consequence, more efficient exploitation of the information can be achieved.

To deal with an ever-changing domain that requires continuous updating, the CSHG needs to be continuously adapted. In this work, we present the experience that we have accumulated during the elaboration of the different updates performed. The initial version focused on creating a semantic and content description of the most relevant concepts of the domain based on a gene-centered vision. Version 2 changed to a chromosome-centered one to simplify the CS and provide a more flexible approach. With the new Version 3, we have increased its flexibility by expanding the interactions among the different parts of the CS and including new, sound domain information that was missing.

It is important to remark that one of the main benefits of the conceptualization work presented in this paper is the possibility of making relevant data and relationships "visible" that were "invisible" in the previous version of the CS, improving the data analytic tasks that can be performed using the CS as the basic knowledge artifact.

We also want to emphasize that these changes respond to real domain-user needs that were requested. The changes in Section IV-A ease the selection, addition, and subtraction of data sources. The changes in Section IV-B allow working with variations from multiple assemblies at the same time. The changes in Section IV-C permit supports with population frequencies in every type of variation. The changes in Section IV-D increase the degree of knowledge of variation-caused effects. The changes in Section IV-E open a wide range of new analyses regarding ncRNA and regulatory elements.

Future work is oriented towards enriching the model semantics and introducing new relevant concepts such as the role of introns and ncRNAs. These genome components are gaining relevance in the context of precision medicine, and its introduction would reinforce the type of data analytic tasks that can be designed with the support of the CSHG. Also, a process that is very relevant in the biotechnology domain called protein splicing (which removes protein segments) is going to be evaluated as a potential addition to the CS.

ACKNOWLEDGMENT

The authors would like to thank the members of the PROS Research Center Genome Group for fruitful discussions regarding the application of Conceptual Modeling in the medical field.

REFERENCES

[1] *rs11571636 RefSNP Report—dbSNP—NCBI*. Accessed: Feb. 13, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/snp/rs11571636

[2] D. Aguilera, C. Gómez, and A. Olivé, "Enforcement of conceptual schema quality issues in current integrated development environments," in *Advanced Information Systems Engineering* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7908. Berlin, Germany: Springer, 2013, pp. 626–640.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, pp. 25–29, May 2000.

[4] (2018). *The Variant Call Format Specification v4.3*. [Online]. Available: https://samtools.github.io/hts-specs/VCFv4.3.pdf

[5] T. M. Baye, T. Abebe, and R. A. Wilke, "Genotype–environment interactions and their translational implications," *Personalized Med.*, vol. 8, no. 1, pp. 59–70, 2011.

[6] A. Bernasconi, S. Ceri, A. Campi, and M. Masseroli, "Conceptual modeling for genomics: Building an integrated repository of open data," in *Conceptual Modeling* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10650. Cham, Switzerland: Springer, 2017, pp. 325–339

[7] C. Biémont and C. Vieira, "Junk DNA as an evolutionary force," *Nature*, vol. 443, no. 7111, pp. 521–524, Oct. 2006.

[8] E. Bornberg-Bauer, "Conceptual data modelling for bioinformatics," *Briefings Bioinf.*, vol. 3, no. 2, pp. 166–180, Jan. 2002.

[9] P. Cingolani, F. Cunningham, W. McLaren, K. Wang. (Jan. 2018). *Variant Annotations in VCF Format*. [Online]. Available: http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf

[10] (2017). *ClinVar: Representation of Clinical Significance in ClinVar and Other Variation Resources at NCBI*. [Online]. Available: https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/

[11] (2019). *ClinVar: Review Status in ClinVar*. [Online]. Available: https://www.ncbi.nlm.nih.gov/clinvar/docs/review_status/

[12] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970

[13] H. R. Crollius, O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau, C. Fischer, C. Fizames, P. Wincker, P. Brottier, F. Quétier, W. Saurin, and J. Weissenbach, "Estimate of human gene number provided by genome-wide analysis using tetraodon nigroviridis DNA sequence," *Nature Genet.*, vol. 25, no. 2, pp. 235–238, Jun. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10835645

[14] L. M. L. Delcambre, S. W. Liddle, O. Pastor, and V. C. Storey, "A reference framework for conceptual modeling," in *Proc. Int. Conf. Conceptual Modeling*, Springer, 2018, pp. 27–42.

[15] S. R. Eddy, "Non-coding RNA genes and the modern RNA world," *Nature Rev. Genet.*, vol. 2, pp. 919–929, Dec. 2001.

[16] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, "The sequence ontology: A tool for the unification of genome annotations," *Genome Biol.*, vol. 6, no. 5, p. R44, 2005.

[17] (2019). *Genome Reference Consortium: GRCh38.p13-Genome-Assembly-NCBI*. [Online]. Available: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39

[18] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: Ten years of next-generation sequencing technologies," *Nature Rev. Genet.*, vol. 17, no. 6, pp. 333–351, Jun. 2016.

[19] S. Köhler, L. Carmody, N. Vasilevsky, J. O. B. Jacobsen, D. Danis, J. P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry, and D. Osumi-Sutherland, "Expansion of the human phenotype ontology (HPO) knowledge base and resources," *Nucleic Acids Res.*, vol. 47, no. 1, pp. D1018–D1027, 2019.

[20] D. S. Latchman, "DNA sequences, transcription factors and chromatin structure," in *Eukaryotic Transcription Factors*. Amsterdam, The Netherlands: Elsevier, 2004, pp. 1–22.

[21] A. L. Palacio, I. P. Fernández, and O. Pastor, "Genomic information systems applied to precision medicine: Genomic data management for alzheimer's disease treatment," in *Proc. Int. Conf. Inf. Syst. Develop. (ISD)*, C. S. B. Andersson, B. Johansson, S. Carlsson, C. Barry, M. Lang, and H. Linger, Eds. Lund, Sweden: Lund Univ., 2018. [Online]. Available: https://aisel.aisnet.org/isd2014/proceedings2018/eHealth/6

[22] A. L. Palacio and Ó. P. López, "Towards an effective medicine of precision by using conceptual modelling of the genome," in *Proc. Int. Workshop Softw. Eng. Healthcare Syst. (SEHS)*, Gothenburg, Sweden, 2018, pp. 14–17.

[23] F. Liang, I. Holt, G. Pertea, S. Karamycheva, S. L. Salzberg, and J. Quackenbush, "Gene index analysis of the human genome estimates approximately 120,000 genes," *Nature Genet.*, vol. 25, no. 2, pp. 239–240, 2000.

[24] E. R. Mardis, "A decade's perspective on DNA sequencing technology," *Nature*, vol. 470, no. 7333, pp. 198–203, 2011.

[25] C. Medigue, F. Rechenmann, A. Danchin, and A. Viari, "Imagene: An integrated computer environment for sequence annotation and analysis," *Bioinformatics*, vol. 15, no. 1, pp. 2–15, Jan. 1999.

[26] I. P. Michael, L. Kurlender, N. Memari, G. M. Yousef, D. Du, L. Grass, C. Stephan, K. Jung, and E. P. Diamandis, "Intron retention: A common splicing event within the human Kallikrein gene family," *Clin. Chem.*, vol. 51, no. 3, pp. 506–515, Mar. 2005.

[27] D. A. Natale, C. N. Arighi, J. A. Blake, J. Bona, C. Chen, S. C. Chen, K. R. Christie, J. Cowart, P. D'Eustachio, A. D. Diehl, and H. J. Drabkin, "Protein ontology (PRO): Enhancing and scaling up the representation of protein entities," *Nucleic Acids Res.*, vol. 45, no. 1, pp. D339–D346, Jan. 2017.

[28] Q. Nguyen and P. Carninci, "Expression specificity of disease-associated lncRNAs: Toward personalized medicine," in *Long Non-Coding RNAs in Human Disease* (Current Topics in Microbiology and Immunology), vol. 394. Cham, Switzerland: Springer, 2016, pp. 237–258.

[29] O. Pastor, A. M. Levin, M. Celma, J. C. Casamayor, A. Virrueta, and L. E. Eraso, "Model-based engineering applied to the interpretation of the human genome," in *The Evolution of Conceptual Modeling* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6520. Berlin, Germany: Springer, 2011, pp. 306–330.

[30] N. W. Paton, S. A. Khan, A. Hayes, F. Moussouni, A. Brass, K. Eilbeck, C. A. Goble, S. J. Hubbard, S. G. Oliver, "Conceptual modelling of genomic information," *Bioinformatics*, vol. 16, no. 6, pp. 548–557, Jun. 2000.

[31] H. Pearson, "What is a gene?" *Nature*, vol. 441, pp. 398–401, May 2006.

[32] S. Ram, and W. Wei, "Modeling the semantics of 3D protein structures," in *Conceptual Modeling—ER 2004*, (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3288. Berlin, Germany: Springer, 2004, pp. 696–708.

[33] J. F. R. Román and Ó. Pastor, "Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano," Ph.D. dissertation, Univ. Politècnica de València, Valencia, Spain, Feb. 2018. [Online]. Available: https://riunet.upv.es/handle/10251/99565

[34] J. F. R. Román, Ó. Pastor, J. C. Casamayor, and F. Valverde, "Applying conceptual modeling to better understand the human genome," in *Conceptual Modeling*, (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9974. New York, NY, USA: Springer-Verlag, 2016, pp. 404–412.

[35] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, and K. Voelkerding, "Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology," *Genet. Med.*, vol. 17, no. 5, pp. 405–423, May 2015.

[36] D. J. Rigden and X. M. Fernández, "The 27th annual nucleic acids Research database issue and molecular biology database collection," *Nucleic Acids Res.*, vol. 48, no. 1, pp. D1–D8, 2020.

[37] A. Smirnov, C. Schneider, J. Hör, and J. Vogel, "Discovery of new RNA classes and global RNA-binding proteins," *Current Opinion Microbiol.*, vol. 39, pp. 152–160, Oct. 2017.

[38] B. Smith, M. Ashburner, C. Rosse, J. Bard, K. Eilbeck, W. Bug, W. Ceusters, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, L. J. Goldberg, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnol.*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007.

[39] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: Astronomical or genomical?" *PLoS Biol.*, vol. 13, no. 7, 2015, Art. no. e1002.

[40] M. Vihinen, "Variation ontology for annotation of variation effects and mechanisms," *Genome Res.*, vol. 24, no. 2, pp. 356–364, Feb. 2014.

[41] R. H. Waterston, E. S. Lander, and J. E. Sulston, "On the sequencing of the human genome," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 6, pp. 3712–3716, 2002.

[42] R. S. Yon, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nature Rev. Genet.*, vol. 9, pp. 509–515, May 2008.

Alberto García S. is currently pursuing the Ph.D. degree with the PROS Research Center, Universitat Politècnica de València, Spain. He is also studying how to improve genome data analysis.

ANA LEÓN PALACIO is currently a Postdoctoral Researcher of the PROS Research Center, Universitat Politècnica de València, Spain. She works on how to provide a systematic approach to efficiently manage genomic data.

JOSE FABIÁN REYES ROMÁN is currently a Postdoctoral Researcher of the PROS Research Center, Universitat Politècnica de València, Spain. He develop solutions on the genomic domain from a conceptual modeling perspective.

JUAN CARLOS CASAMAYOR is currently an Associate Professor and a Researcher of the PROS Research Center, Universitat Politècnica de València, Spain. His research interest includes databases and information systems design.

OSCAR PASTOR is currently a Full Professor and the Director of the PROS Research Center, Universitat Politècnica de València, Spain. He is also leading a multidisciplinary project linking information systems and bioinformatics to designing and implementing tools for conceptual modeling-based interpretation of the human genome information.