A Comprehensive Approach for the Conceptual Modeling of Genomic Data

Anna Bernasconi[1,2,⋆][0000−0001−8016−5750], Alberto García S.[1,⋆][0000−0001−5910−4363], Stefano Ceri[2][0000−0003−0671−2415], and Oscar Pastor[1][0000−0002−1320−8471]

[1] PROS Research Center, VRAIN Research Institute, Universitat Politcnica de Valncia, Spain. abernas@upvnet.upv.es, {algarsi3,opastor}@pros.upv.es
[2] Dept. of Electronics, Information, and Bioengineering, Politecnico di Milano, Italy. {anna.bernasconi,stefano.ceri}@polimi.it

Abstract. The human genome is traditionally represented as a DNA sequence of three billion base pairs. However, its intricacies are captured by many more complex signals, representing DNA variations, the expression of gene activity, or DNA's structural rearrangements; a rich set of data formats is used to represent such signals. Different conceptual models explain such elaborate structure and behavior. Among them, the Conceptual Schema of the Human Genome (CSG) provides a *concept-oriented, top-down* representation of the genome behavior – independent of data formats. The Genomic Conceptual Model (GCM) instead provides a *data-oriented, bottom-up* representation, targeting a well-organized, unified description of these formats. We hereby propose to join these two approaches to achieve a more complete vision, linking (1) a *concepts layer*, describing genome elements and their conceptual connections, with (2) a *data layer*, describing datasets derived from genome sequencing with specific technologies. The link is established when specific genomic data types are chosen in the data layer, thereby triggering the selection of a view in the concepts layer. The benefit is mutual, as data records can be semantically described by high-level concepts and exploit their links. In turn, the continuously evolving abstract model can be extended thanks to the input provided by real datasets. As a result, it will be possible to express queries that employ a holistic conceptual perspective on the genome, directly translated onto data-oriented terms and organization. The approach is here exemplified using the DNA variation data type but is applicable to all genomic information.

Keywords: Conceptual Modeling · Biological Datasets · Genomics

1 Introduction

Representing the human genome DNA as a three billion base pairs' sequence is just a first attempt to capture the complex mechanisms of the life engine that

⋆ A.B and A.G.S. should be regarded as Joint First Authors.

is underlying all our characteristics and behaviors. Many other aspects, such as DNA mutations, the expression of gene activity, DNA's structural rearrangements, long distance contacts between DNA regions, and so on are now used to extract complex signals from the DNA, exploiting Next Generation Sequencing [30]; a rich set of data formats is used to represent such signals. The study of genomic information has practical implications on a number of fields such as cancer genomics, population genomics, and precision medicine. More importantly, being able to interoperate different signals in the context of a same analysis can provide insights and compute properties of the genome that remain otherwise hidden. Genomic data integration has so far been addressed mainly with operational approaches [18,1], whereas a holistic view – that encompasses the meaning of different genomic regions – has not been embraced yet. Conceptual models (CMs) have supported the effort of explaining such elaborate structure and behavior since 2000 [24,6]. However, genome data are frequently generated in practical lab settings without following any sound process of conceptual characterization. This creates a gap between "real" genome data CMs (that represent "genome data as it is") and pure genome CMs (that model "data as it should be"). Components obtained from the first kind of CMs must be connected with their corresponding components in the CMs that represent higher-level conceptual genome knowledge. We refer to the process of connecting concepts with their associated data as a "top-down" process, while we use the term "bottom-up" for connecting data to concepts.

A number of works, summarized by the Conceptual Schema of the Human Genome (CSG, [23]) produced by the PROS research center, provide a *concept-oriented, top-down* representation of the genome that is independent from the data formats, aiming to give a template of how the genome is supposed to behave. This perspective has contributed many valuable results devoted to building a general understanding of the language of life [12]. Another initiative, represented by the Genomic Conceptual Model (GCM, [5]) produced by the GeCo project [10], provides a *data-oriented, bottom-up* representation, targeting a high-level, abstract description of these formats, focusing on what data capture, how they capture it, to favour a joint use of the signals. With this approach, important achievements have been obtained in the area of data integration and search systems for genomics researchers [4,8].

By construction, the CSG model evolves according to upcoming requirements, while the GCM model evolves when new datasets arrive. In this work, we propose to join these two independent directions by explaining how, together, they can provide a more complete vision of the steps involved within the full-stack research that goes from the collection of data to the understanding of life mechanisms. On the one hand, we configure the CSG as the model that describes concepts, now renamed as the *concepts layer*, i.e., the template of the genome, where concepts are genome elements. On the other hand, we employ the GCM as the model that describes data, that is the *data layer*, where classes are real instances of datasets derived from tissues, cell lines, or individual cells that have undergone a sequencing process. The data layer is organized in DATASETS, each

containing multiple SAMPLES. Samples may contain multiple SAMPLEREGIONS that are records representing fragments of the genome with specific measured properties. Each of these records can be linked to the corresponding concept in the concepts layer. New links are established when specific data types or experiments are chosen (in the data layer) triggering the selection of specific views (of the concepts layer). The benefit is two-fold: 1) the GCM is extended by the power of concepts, which enable high-level semantic-aware querying; 2) the CSG is empowered by its links to real-world data, that allow building computations on experimental instances and obtain biologically-relevant results.

In the following, we present how our background approaches to conceptual modeling in genomics deal with concepts and data (Sec. 2); we describe our vision of a unified conceptual model including a concepts and a data layer, and then illustrate our method for the linking of the two layers (Sec. 3); to exemplify the approach, we focus on the knowledge concerning DNA variation and we show how the two models can be pragmatically connected in this case (Sec. 4). This method is applicable to other genomic data types; in a more general framework, it will be possible to develop additional views and to use them together, towards a more encompassing conceptual perspective on the human genome (Sec. 5).

2 Background

As of today, two main approaches have tackled genomics from a conceptual modeling perspective, as briefly described in the following.

PROS: a top-down approach. The Research Center on Software Production Methods (PROS) at the Universitat Politecnica de Valencia has invested many efforts in studying the genome from a conceptual modeling perspective, introducing the first Conceptual Schema of the Human Genome in 2011 [23] and producing several extensions since then [28,12]. The schema now results into a rich map of concepts and relationships that support the holistic understanding of different knowledge modules. The most recent version, called the Conceptual Schema of the Genome v3 (CSG) is reported in [13]. The employed method is considered top-down, as the main objective stands in identifying relevant concepts and their connections, independently on how datasets are really represented in available databases and sources.

GeCo: a bottom-up approach. The approach devised within the data-driven Genomic Computing (GeCo) group, funded by the ERC AdG 693174 (2016-2021), has instead adopted a bottom-up approach, meaning that models are developed for representing existing data, with the purpose of making data more interoperable and ready for large-scale computations. Open data sources are analyzed and evaluated, understanding their underlying models; selected interesting datasets are imported within an integrative repository [4]. Information is divided between: region data (representing actual genomic elements, measured by experiments – using the Genomic Data Model, GDM [18]) and metadata (descriptions of genomic experiments – captured by the Genomic Conceptual Model, GCM [5]), which make data searchable [8]. Finally, the modeled datasets

attempt to resolve data-level interoperability, thereby enabling powerful queries using, e.g., the GenoMetric Query Language (GMQL system [17]).

Fig. 1. Schematization of the two compared approaches.

Comparing the two approaches. We compare the two existing approaches under two perspectives: 1) how they deal with the concepts representing the knowledge of genomics; 2) how they manage their instantiation in the form of data. Genomic information can be interpreted as a dual system that is approached in two opposite directions, as observed in Fig. 1: on one side, the possibility to connect data to existing concepts that have been modeled in an abstract way (top-down approach), on the other side the possibility to build concepts based on already available data (bottom-up approach). Traditionally, PROS has adopted a top-down perspective, starting from modeling biological entities and only after checking if underlying data sources exist that represent such concepts, possibly unveiling problems in the quality of data structures definition and values. GeCo, instead, has adopted a bottom-up approach, starting from the observation of available data sources and only later building models to systematize, organize and interoperate such existing data, with the purpose of building easy-to-use systems that facilitate domain experts' work. With the intention of connecting these two perspectives, our work contributes a comprehensive approach that integrates them in order to facilitate genome data management by using a sound CM support.

3 Methodological Framework

We describe a general two-layer schema that contains:

- a concepts layer capturing the knowledge available about the human genome mechanisms (inspired by the CSG [13]);
- a data layer representing genomic data, with its types and experiments, captured by information structures and formats (inspired by the original GCM [5] for metadata and the Genomic Data Model [18] for region data).

Making an analogy with the triptych paradigm of Mayr and Thalheim [19], we can interpret our data layer as the one of "languages", enabling the narrative representation of our concepts layer (the "mental reasoning"). Written records (artifact world, our genomic data) stay on one level and – when instantiated – point directly to beliefs and perspectives (mental world, our genome concepts).

The data layer. The data layer (schematized in Fig. 2) is centered on the SAMPLE concept. It holds two metadata perspectives: the biological one contains

the REPLICATE to which a sample belongs, part of a BIOSAMPLE, extracted from a DONOR; the organizational one has the CASESTUDY under which the sample was produced, which is contained in a greater PROJECT. Samples are built when an EXPERIMENTTYPE (e.g., DNA-Seq, RNA-Seq, or ChIP-Seq) is run, expressing information about the sequencing technology and representing a specific *genomic data type* (e.g., DNA variation, gene expression quantification, or binding sites of DNA-associated proteins). With respect to the original GCM, we also have that samples contain multiple SAMPLEREGIONS, typically a file row representing a fragment of the genome on a specific chromosome strand, with start and stop coordinates. All the regions in a sample follow the same SCHEMA. Note that these two classes were added to the data layer (w.r.t. the original GCM) as they are necessary to manage the linking between the two layers. Many samples are grouped into a DATASET, which is homogeneous in the schema and in the experiment type.

The concepts layer. The concepts layer is based on the last version of the CSG [13], including five modules, respectively describing i) the structure of the human genome; ii) protein synthesis; iii) changes in the sequence referring to a reference sequence (the "Variation module"); iv) information and sources related to the elements of the conceptual schema; and v) human metabolic pathways. The schema is manually-generated and incrementally enriched as new mechanisms are understood by a team of conceptual modelers or when new research findings are published. Genome knowledge is under continuous progress and understanding the human genome is an open big scientific challenge. For this reason, completeness is obviously not guaranteed and a mechanism to periodically handle needed extensions is employed. We consider this a "work-in-progress" model, where knowledge representation evolves, based on incoming requirements. While building the link with the data layer, it is likely that extensions to the CSG will be required, reinforcing the relevance of accomplishing the essential data-concepts genomic connection that this paper develops.

Data type-driven linking of the two layers. Connections are built between the data and the concepts layers. By selecting specific genomic data types (based on the represented sequencing experiment type) we trigger a mechanism that invokes a specific portion of the concepts schema, as described by Fig. 2. In their previous description within the GCM [5], data types were forced into containers (i.e., SAMPLES) that flattened their semantics for integration and processing benefits; instead, here each data type is "freed" from its container, separately handled, analyzed, and mapped onto its explanation in conceptual terms.

The concepts layer and the data layer are connected by means of relations between concepts (i.e., a variation of DNA) and instances of data layer classes (i.e., the specific data record). For instance, a SAMPLEREGION measured through a DNA-Seq experiment, can be represented by its related concept, i.e., a variation at position 43,044,295–43,170,245 of the negative strand of chromosome 17.

Much in the spirit of Ontology-Based Data Access (OBDA [7]) approaches and in the fashion of ISGE [14], we envision the primary use mechanism of our two-layer schema as follows: 1) **Identification** of a *genomic data type*

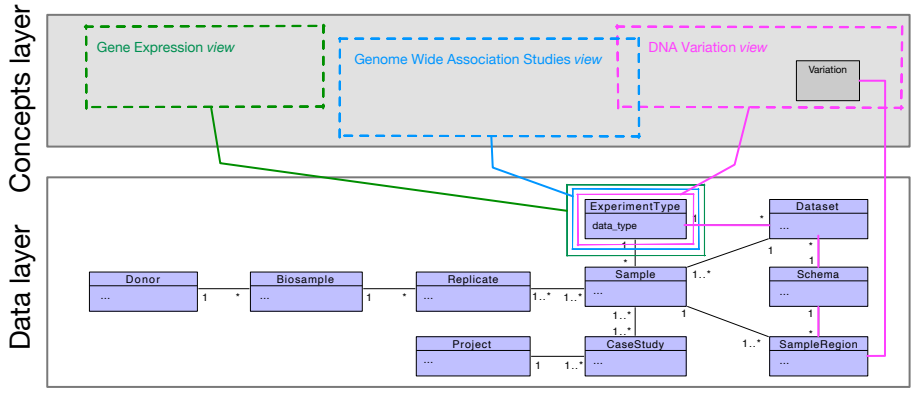

Fig. 2. Link between the concepts layer and the data layer by means of connections between sample regions and concepts.

(ExperimentType in the data layer); 2) **Selection** of the related – possibly multiple – Datasets, which have a corresponding Schema that is followed by the SampleRegions of the dataset (again, in the data layer); 3) **GEneration** of a *view* (in the concepts layer) built around a central concept that represents the SampleRegion of the identified data type. Intuitively, the identification of a genomic data type (within an experiment type) triggers the generation of a specific view of interrelated concepts, comprising only entities and relationships that contribute to explain the content of that data type.

4 Method Application: Modeling DNA Variation

Many datasets are used in the daily practice of geneticists and computational biologists. These represent various types of information captured from the genome and the study of cohorts of patients, including information on the variation of DNA (population variation, its association with phenotype, somatic mutations, copy number variation, or structural rearrangements); the behavior of RNA (gene, miRNA, or isoform expression); or epigenetic signals (such as DNA methylation, DNA binding, or DNase I Hypersensitive sites).

For instantiating our method and describing it in more detail, we focus on one specific type of data, i.e., DNA variation, which includes both population variation and cancer-derived somatic mutations. We carefully considered the DNA variation module of CSG and applied appropriate changes to instantiate the related concepts layer *view*. The color code in Fig. 3 highlights which components have been added (green) or removed (red) with respect to the original model (blue classes) based on [13]; these changes are consolidated in an updated version of the CSG, which is next described so as to explain the evolved concepts layer in full detail.

The obtained schema has 21 entity classes and 2 association classes, with six generalizations and three compositions (one of which is double). The most

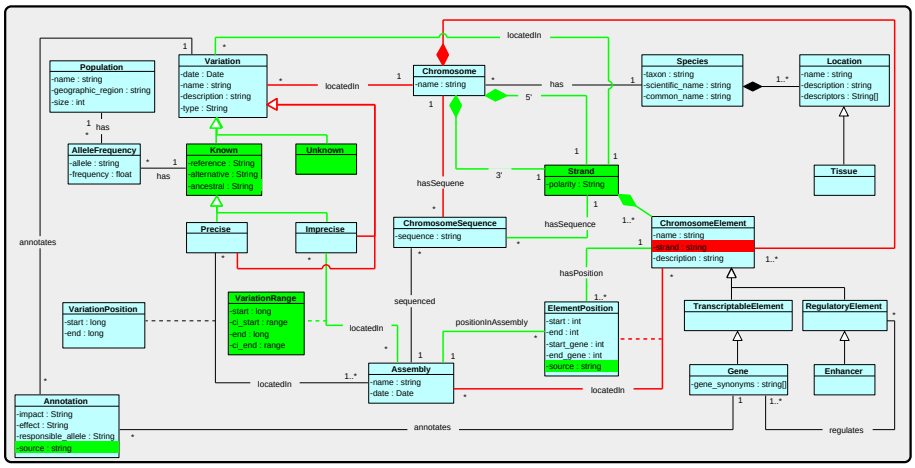

Fig. 3. Conceptual view dedicated to the DNA variation data type. Blue classes are derived from the CSG [28]; green classes, attributes, and relationships have been added here; red attributes and relationships are removed for the purpose of this effort.

important class is the VARIATION one, with a *date*, *name*, *description*, and *type* (deletion, insertion or substitution); it is located on a specific STRAND (with positive or negative *polarity*), which contributes to compose a CHROMOSOME (with a *name*). Chromosomes are related to a SPECIES (with a taxonomy definition and scientific/common name), made of LOCATIONS (with *name*, *description*, and *descriptors*), such as TISSUES. On a strand, several CHROMOSOMEELEMENTS can be hosted (with their *name* and *description*). These include TRANSCRIPTABLEELEMENTS, such as GENES (with their alternative *gene_synonyms*) and REGULATORYELEMENTS that regulate genes, such as ENHANCERS. Elements present possibly multiple ELEMENTPOSITIONS (*start* and *end* positions on the chromosome, the genes on which they insist, and the information *source* from which the position has been obtained); these are measured with respect to an ASSEMBLY, i.e., a reference system based on a community-defined sequence (with a *name* and *date*). Each strand of the observed chromosome has a CHROMOSOMESEQUENCE, which is also determined based on the assembly.

Variations may be specialized according to how their position is considered. If the position is not determined, we call the variation UNKNOWN; else it is KNOWN. Known variations have alleles called *reference* (the base reported by the reference sequence in that position), *alternative* (the mutated base), and *ancestral* alleles. If the exact position is available, we call the variation PRECISE; if the position is reported within a range, we call it IMPRECISE. Precise variations record the VARIATIONPOSITION – with *start* and *end* coordinates – as an association class. Imprecise variations are also related to an assembly, but their association is characterized by a VARIATIONRANGE class that sets *start* and *end* positions within intervals of confidence (called *ci_start* and *ci_end*).

In the context of a POPULATION (with *name*, *geographic_region*, and *size*), a known variation has an ALLELEFREQUENCY, with a *frequency* indication reporting the percentage of presence of the *allele* within the considered population. Variations can alter the functionality of genes; we represent this with the ANNOTATION class, with an *impact*, *effect*, *responsible_allele*, and information *source*. In Fig. 3, we applied notable additions (green elements) to the original CSG:

- A STRAND class was added such that a chromosome is made of two strands and a VARIATION is exhibited only on one of them (i.e., variations can be read from 5' to 3' or from 3' to 5').
- The *ancestral* attribute was added in the KNOWN class to represent the allele of the last common ancestor of primates.

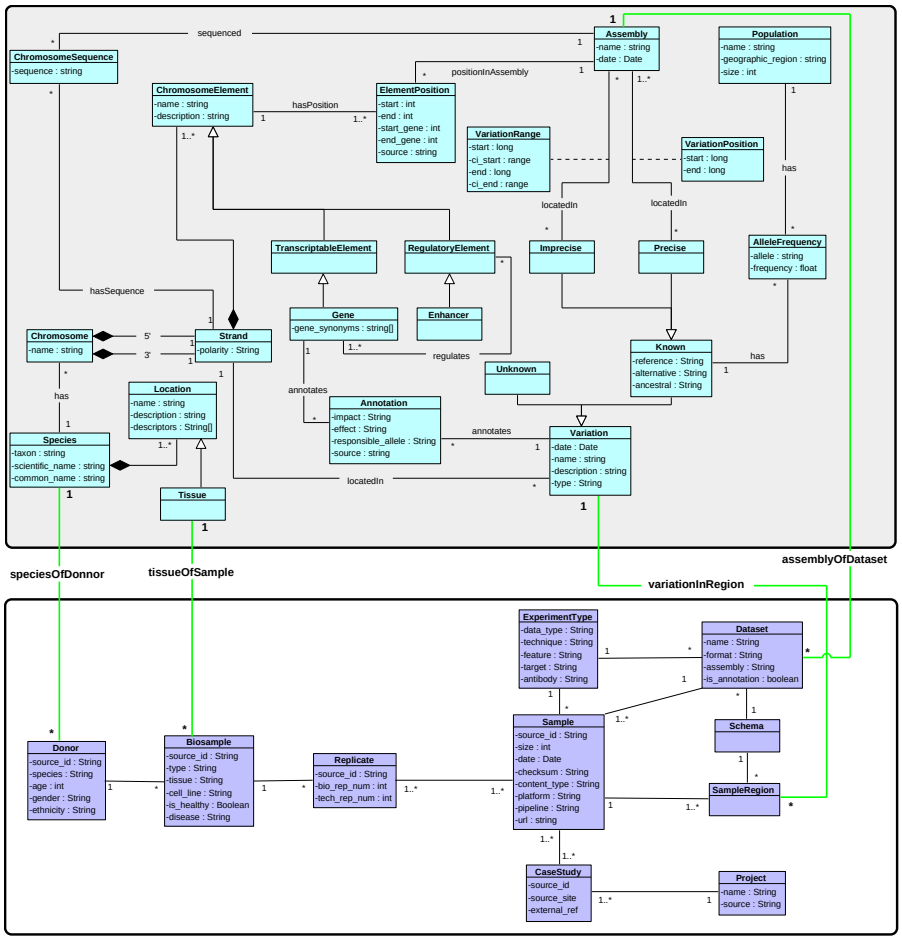

Fig. 4. Representation of the DNA variation information comprising the concepts view and related datasets.

- The concept previously represented by the IMPRECISE class (i.e., variation for which coordinates were unknown) was updated to include variations with uncertain positions (within confidence intervals, represented by ranges), whereas the new UNKNOWN class was added to capture the original concept.
- The KNOWN variation class was added to generalize properties of both PRECISE and IMPRECISE variations.
- The *source* attribute was added in the ANNOTATION class to identify the origin of such assertion (e.g., a research group or automated annotation software).
- The original association class ELEMENTPOSITION was transformed into a regular class, to overcome the limitation of only allowing one-to-one correspondences between the two linked classes. This enables, for example, the characterization of a same CHROMOSOMEELEMENT in terms of coordinates (in the same assembly) provided by different authoritative *sources* (e.g., RefSeq or GENCODE).

When a full correspondence between the concepts layer and the data layer is established, the complete schema is obtained as in Fig. 4. Here, connections are made between DONOR (data layer) and SPECIES (concepts layer); BIOSAMPLE (data) and TISSUE (concepts): SAMPLEREGION (data) and VARIATION (concepts); and DATASET (data) and ASSEMBLY (concepts).

4.1 Mapping with Real Datasets

Many different data representations may be used to indicate same concepts. Semantic integration can be achieved by using the conceptual layer as a pivot of data representations. To practically discuss how concepts can be instantiated into data records in real world scenarios, we consider the use of datasets representing human variation as collected within two important research projects. The Cancer Genome Atlas (TCGA,[31]) is a landmark cancer genomics program that sequenced and characterized over 11,000 patients of primary cancer samples, analyzing them with different experiments, including one dedicated to somatic mutations. The 1000 Genomes Project (1KGP, [26]) is an international research effort established to create a catalogue of common human germline variation, using samples from healthy people. In the GMQL data repository [4,17] (`http://gmql.eu/gmql-rest/`), we analyzed all the data fields contained in the datasets' schemas that refer to these data types. Specifically, we considered 1000 Genomes Project datasets (for both the hg19 and GRCh38 assemblies) and TCGA datasets related to masked somatic mutations (for both the hg19 and GRCh38 assemblies [9]).

To demonstrate a possible implementation of the proposed approach, we employ a relational database representation. The top part of Table 1 describes the schemas of the tables designed starting from the presented model. Note that most tables are directly derived from a translation from the class diagram into an RDBMS logical schema. The central SAMPLE class (a file in the repository) has one-to-many SAMPLEREGIONS, which correspond to a specific SCHEMA (an auxiliary table with a row for each dataset, in the example two rows for TCGA

Data.Donor(source_id,species,age,gender,ethnicity)
Data.BioSample(source_id,type,tissue,cell_line,is_healthy,disease)
...
Data.Sample(source_id,size,date,checksum,content_type,platform,pipeline,url)
Data.SampleRegion1KGP(chr,start,stop,strand,AL1,AL2,ref,alt,mut_type,length,id,quality,filter,DP,AF,AC,
AFR_AF,AMR_AF,EUR_AF,EAS_AF,SAS_AF,AA,IMPRECISE,CIEND,CIPOS,"germline")
Data.SampleRegionTCGA(chrom,start,end,strand,gene_symbol,entrez_gene_id,variant_classification,
variant_type, reference_allele, tumor_seq_allele1, tumor_seq_allele2, dbsnp_rs,"somatic")

Concept.Variation(*gen*(),name,*gen*(),type)
⊇ Data.SampleRegion1KGP(_,_,_,_,_,_,_,_,type,_,name,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_)
Concept.Variation(date,name,description,type)
⊇ Data.SampleRegionTCGA(_,_,_,_,_,_,_,type,_,_,_,name,_)
Concept.Known(reference,alternative,ancestral)
⊇ Data.SampleRegion1KGP(_,_,_,_,_,_,reference,alternative,_,_,_,_,_,_,_,_,_,_,_,_,_,_,ancestral,_,_,_,_)
Concept.Known(reference,*f*(reference,allele1,allele2),`null`)
⊇ Data.SampleRegionTCGA(_,_,_,_,_,_,_,reference,allele1,allele2,_,_)
Concept.Imprecise() ⊇ Data.SampleRegion1KGP(_,`true`,_,_,_)
Concept.Precise() ⊇ Data.SampleRegion1KGP(_,`false`,_,_,_)
Concept.VariationRange(start,ci_start,end,ci_end)
⊇ Data.SampleRegion1KGP(_,start,end,_,ci_end,ci_start,_)
Concept.VariationPosition(start,end)
⊇ Data.SampleRegion1KGP(_,start,end,_)
Concept.VariationPosition(start,end)
⊇ Data.SampleRegionTCGA(_,start,end,_,_,_,_,_,_,_,_,_,_)
Concept.Species(*f*(scientificName),scientificName,*f*(scientificName))
⊇ Data.Donor(_,scientificName,_,_,_)
Concept.Location(name, *gen*(), *f*(is_healthy,disease))
⊇ Data.Biosample(_,"tissue",name,_,is_healthy,disease)
Concept.Chromosome(name) ⊇ Data.SampleRegion1KGP(name,_)
Concept.Chromosome(name) ⊇ Data.SampleRegionTCGA(name,_,_,_,_,_,_,_,_,_,_,_,_,_)
Concept.Strand(polarity) ⊇ Data.SampleRegion1KGP(_,_,_,strand,_)
Concept.Strand(polarity) ⊇ Data.SampleRegionTCGA(_,_,_,strand,_,_,_,_,_,_,_)
Concept.ChromosomeElement(name,*gen*())
⊇ Data.SampleRegionTCGA(_,_,_,_,name,_,_,_,_,_,_,_,_,_)
Concept.Gene(geneSynonym)⊇ Data.SampleRegionTCGA(_,_,_,_,_,geneSynonym,_,_,_,_,_,_,_)
Concept.Assembly(name,*f*(name))
⊇ Data.Dataset(_,_,_,name,_)
Concept.AlleleFrequency(allele,frequency)
⊇ Data.SampleRegion1KGP(_,_,_,_,_,_,_,alt,_,_,_,_,_,_,AF,_,_,_,_,_,_,_,_,_,_,_,_)
Concept.Annotation(effect,*f*(effect),*f*(ref,allele1,allele2))
⊇ Data.SampleRegionTCGA(_,_,_,_,_,_,_,effect,_,ref,allele1,allele2,_,_)
Concept.Population-AlleleFrequency("African","Africa",1418,allele,frequency)
⊇ Data.SampleRegion1KGP(_,_,_,_,_,_,_,allele,_,_,_,_,_,_,_,_,frequency,_,_,_,_,_,_,_,_,_)
...

Table 1. Top part: relational schema of the data layer, with the 1000 Genomes Project population variation dataset and the TCGA masked somatic mutations dataset. Bottom part: examples of mapping rules for building the relational schema of the concepts layer; we assume Population-AlleleFrequency to be a single table obtained as the join of tables derived from the Population and AlleleFrequency classes.

and two rows for 1KPG). For sample regions we employ one table for each different dataset. For simplicity, in this example we refer to SampleRegionTCGA and SampleRegion1KGP (only considering their GRCh38 versions).

Mapping rules are used to describe how datasets information can be mapped into the concepts schema, considering the view that is specific for DNA variation. The bottom part of Table 1 provides the mappings for the TCGA and 1KGP datasets. Each mapping rule is a logic formula (in Datalog-like syntax [11]) with variables in its left end side (LHS) that are computed from the variables in its right end side (RHS). The order of the variables follows the one indicated in the upper part of the table (e.g., the SampleRegion1KGP table has 26 fields and the SampleRegionTCGA table has 13 fields). As an example, the entity Variation of the concepts schema is filled using data from the SampleRe-

GION1KGP table, using the attributes in its 9th and 11th position (originally called *mut_type* and *id*) that map to the *type* and *name* attributes of the output VARIATION table. Similarly, the same VARIATION entity is filled using also data from the SAMPLEREGIONTCGA table, using the attributes in its 8th and 12th position (originally called *variant_type* and *dbsnp_rs*) that map to the *type* and *name* attributes of the output VARIATION table. Note that we wrote a different rule for each pair of output table (in the concepts layer) and input table (1KGP or TCGA in the data layer), when the mapping is meaningful.

In some cases, we need to derive new fields in the concepts layer schema as functions of original fields. One such example is in the KNOWN table: here, the second field *alternative* requires combining the values of three fields in the input table SAMPLEREGIONTCGA. For this, we use the notation *f(...)*. Moreover, names or descriptions are generated from the system admin (with *gen()*). A particular case is the one of POPULATION and ALLELEFREQUENCY tables: here the computation of the attributes of the second table (*allele* and *frequency*) depends on the values of the first. The values coming from the input table (e.g., AFR_AF from the SAMPLEREGION1KGP schema) denote the allele frequency only for a specific population. We thus represent this case using, as output table, the joined table that contains together the information of the population matching with its allele/frequency information. Here we did not report concepts layer's tables that could not be directly mapped to any field of the two data sources considered in this example; this is the case of CHROMOSOME, for instance, whose attribute *sequence* can be filled by inspecting authoritative sources such as RefSeq [22].

4.2 Examples of Applications

This section reports examples of queries that are enabled by concept-to-data linking, showing that: a) data improves the representation of genome concepts *within a specific view* (bottom-up); b) concepts and their connections improve the knowledge generation process allowing connections *across views* generated by different data types (top-down). Examples 1 and 2 demonstrate case (a) while examples 3 and 4 show case (b).

Ex 1. Extract positions of chromosome elements provided by different sources. Intuitively, one would expect that a specific gene was located in a uniquely defined range on a chromosome. However, its positions are identified by means of complex measurements which depend on the used technology or employed bioinformatics algorithm/parameters. Indeed, when such a query is posed to real data sources, we find multiple distinct positions. For instance, in the hg19 assembly, the PAQR6 gene is located in chromosome 1 at 156,213,111–156,217,908 according to RefSeq, whereas it is located at 156,213,205–156,217,881 according to GENCODE. The concepts layer adequately captures these aspects and it allows to pose a generic query while extracting heterogeneous definitions from the data.

Ex 2. Extract mutations whose position is not precisely identified. The concepts layer includes the possibility to represent known imprecise variations, which are commonly found in variation data sources such as the 1000 Genomes Project. For

instance, a 297 bases-long variation could be located between position 14,477,084 (with a range of uncertainty that spans from 22 bases before, up to 18 bases after) and position 14,477,381 (with uncertainty between 12 and 32 bases).

Ex 3. Extract mutations located on enhancers associated to breast cancer. Let us consider the study of a patient genome targeting presence of mutations on BRCA1, i.e., a specific gene that is associated to breast cancer, located at position 43,044,295–43,170,245 of the negative strand of chromosome 17. From data, it can be observed that no relevant mutations are present in this range. However, in terms of clinical significance, in addition to genes, it is critical to consider also their regulatory elements. In this case, mutations should be tested also on the enhancers of BRCA1. Several data sources can provide this information. For example, the GH17J043124 enhancer is reported by GeneCards [29] at positions 43,123,800–43,127,201 and by ENCODE [27] at positions 43,124,247–43,126,961, being currently associated to breast cancer [2]. Note that mutation datasets (such as TCGA's ones) may sometimes report correspondence between variations and their enclosing genes; while this is a quite standard information, less studied elements, such as enhancers, are not typically considered. This connection, however, can be made by employing the concepts layer representation. The schema allows to make explicit a relation between positions and elements (including genes and enhancers) that remains instead hidden in the data.

Ex 4. Extract orthologous genes for humans and other species. By exploiting the connection between Donor (data layer) and Species (concepts layer) it would be possible to select genes of *Homo Sapiens* and genes of, e.g., canine models, which are orthologous (i.e., genes in different species that evolved from a common ancestral gene by speciation). Notably, over 58% of genetic diseases seen in the dogs closely depict the phenotype of human diseases caused by mutations in orthologous genes [15]. By exploiting the findings available for canine genes, candidates for gene-driven therapies may be found, e.g., for Duchenne muscular dystrophy [21].

5 Discussion and Conclusion

In this work we have described the concept-driven and data-driven approaches to conceptual modeling for genomics, that have guided the development of CSG and GCM. We then described a method for linking these models so as to generate an encompassing conceptual model that provides both the concept and data viewpoints. We applied our approach to the DNA variation case, showing that the new conceptual model can support interesting queries and applications, both acting on a single dataset and on several integrated datasets.

This work inspires future developments within the two projects and significant future joint activities that will integrate several available open data sources [3]. For what concerns the CSG model, the most substantial issue that will be addressed as future work is the inclusion of the notion of "individual". Indeed, DNA variation data, as well as many other genomic signals – here not

discussed – do express information of this kind. Examples taken from the analyzed domain include 1) the person's genotype, which comprises the *allele1* and *allele2* attributes, concerning on which of the two chromosome copies the associated variant is located; 2) the *origin* (i.e., nature) of the variation, which could be somatic (occurring from damage to DNA in an individual cell during a person's life, not passed from parent to child) or germline (occurring in a sperm/egg cell, copied into every cell in the body, possibly passing from generation to generation). This may, for instance, enable studies on overlaps between variations that are recorded both as somatic and as germline in public databases [20]. The missing notion of "individual" is being investigated within the CSG working group and the upcoming results will be reported on this effort as well. For what concerns the GCM model, work has been so far driven by the requirement of creating a large repository (hosting, at the time of the writing, about 550 thousand files within a large database of 9 terabyte [17]). As a consequence of this initial choice, today GCM misses opportunities for conceptual data linking, that will drive its future extensions.

Regarding the joint effort described here, the most important challenge stands in generating views for all most relevant genomic data types, while carefully designing their links. In this paper, we show the variation-related information, but we will next take data types one by one and generate extensions of the concepts layer view by view. In this direction, we envision a holistic system that, based on the accurate view-specific contents, is able to provide a synergical perspective on the genome. The system will enable the combined use of multiple views, with selective mechanisms that activate one area or the other.

Users will then be allowed to ask questions that, for example, connect datasets on variation at the DNA level to variation at the amino acid level (i.e., proteins). More complex queries could compare somatic and germline variations (by means of "differential mutation analysis") to identify genes that are likely involved in a given disease [25] or identify susceptibility to tumorigenesis by exploiting genome-wide association studies [16]. More broadly, queries could span from mutations to their interaction with phenotype evidence, using their position within annotated genome elements, possibly also connecting it to interactions with the epigenome or the tridimensional organization of the genomic chain. All of these queries would benefit from the approach described in this work, facilitating in a natural way the interoperability between different data types connecting their corresponding views.

Acknowledgement. This research is funded by the ERC Advanced Grant 693174 GeCo (Data-Driven Genomic Computing), INNEST/2021/57, and MICIN/AEI/ 10.13039/501100011033.

References

1. Augustyn, D.R., et al.: Perspectives of using Cloud computing in integrative analysis of multi-omics data. Briefings in functional genomics **20**(4), 198–206 (2021)

2. Bass, J.I.F., et al.: Human gene-centered transcription factor networks for enhancers and disease variants. Cell **161**(3), 661–673 (2015)
3. Bernasconi, A., et al.: The road towards data integration in human genomics: players, steps and interactions. Briefings in Bioinformatics **22**(1), 30–44 (2021). https://doi.org/10.1093/bib/bbaa080
4. Bernasconi, A., et al.: META-BASE: A Novel Architecture for Large-Scale Genomic Metadata Integration. IEEE/ACM Transactions on Computational Biology and Bioinformatics **19**(1), 543–557 (2022)
5. Bernasconi, A., et al.: Conceptual modeling for genomics: building an integrated repository of open data. In: International Conference on Conceptual Modeling. pp. 325–339. Springer (2017)
6. Bornberg-Bauer, E., et al.: Conceptual data modelling for bioinformatics. Briefings in Bioinformatics **3**(2), 166–180 (2002)
7. Calvanese, D., et al.: Ontology-based database access. In: SEBD. pp. 324–331 (2007)
8. Canakoglu, A., et al.: GenoSurf: metadata driven semantic search system for integrated genomic datasets. Database **2019** (2019)
9. Cappelli, E., et al.: OpenGDC: Unifying, Modeling, Integrating Cancer Genomic Data and Clinical Metadata. Applied Sciences **10**(18), 6367 (2020)
10. Ceri, S., et al.: Overview of GeCo: A project for exploring and integrating signals from the genome. In: International Conference on Data Analytics and Management in Data Intensive Domains. pp. 46–57. Springer (2017)
11. Ceri, S., et al.: What you always wanted to know about Datalog (and never dared to ask). IEEE Transactions on Knowledge and Data Engineering **1**(1), 146–166 (1989)
12. García, A., et al.: Towards the understanding of the human genome: a holistic conceptual modeling approach. IEEE Access **8**, 197111–197123 (2020)
13. García, A., et al.: A conceptual model-based approach to improve the representation and management of omics data in precision medicine. IEEE Access **9**, 154071–154085 (2021)
14. García S, A., et al.: ISGE: A Conceptual Model-Based Method to Correctly Manage Genome Data. In: International Conference on Advanced Information Systems Engineering. pp. 47–54. Springer (2021)
15. Gopinath, C., et al.: Contemporary animal models for human gene therapy applications. Current gene therapy **15**(6), 531–540 (2015)
16. Mamidi, T.K.K., et al.: Integrating germline and somatic variation information using genomic data for the discovery of biomarkers in prostate cancer. BMC cancer **19**(1), 1–12 (2019)
17. Masseroli, M., et al.: Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. Bioinformatics **35**(5), 729–736 (08 2018)
18. Masseroli, M., et al.: Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. Methods **111**, 3–11 (2016)
19. Mayr, H.C., et al.: The triptych of conceptual modeling. Software and Systems Modeling **20**(1), 7–24 (2021)
20. Meyerson, W., et al.: Origins and characterization of variants shared between databases of somatic and germline human mutations. BMC bioinformatics **21**(1), 1–22 (2020)
21. Nghiem, P.P., et al.: Gene therapies in canine models for duchenne muscular dystrophy. Human Genetics **138**(5), 483–489 (2019)

22. O'Leary, N.A., et al.: Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic acids research **44**(D1), D733–D745 (2016)
23. Pastor, O., et al.: Model-based engineering applied to the interpretation of the human genome. In: The Evolution of Conceptual Modeling, pp. 306–330. Springer (2011)
24. Paton, N.W., et al.: Conceptual modelling of genomic information. Bioinformatics **16**(6), 548–557 (2000)
25. Przytycki, P.F., et al.: Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes. Genome Medicine **9**(1), 79 (2017)
26. 1000 Genomes Project Consortium: A global reference for human genetic variation. Nature **526**(7571), 68 (2015)
27. ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)
28. Román, J.F.R., et al.: Applying conceptual modeling to better understand the human genome. In: International Conference on Conceptual Modeling. pp. 404–412. Springer (2016)
29. Safran, M., et al.: The genecards suite. In: Practical Guide to Life Science Databases, pp. 27–56. Springer (2021)
30. Schuster, S.C.: Next-generation sequencing transforms today's biology. Nature methods **5**(1), 16–18 (2008)
31. Weinstein, J.N., et al.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)