PoliViews: A Comprehensive and Modular Approach to the Conceptual Modeling of Genomic Data

Anna Bernasconi[a,*], Alberto García S.[b,*], Stefano Ceri[a] and Oscar Pastor[b]

[a]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy
[b]PROS Research Center, VRAIN Research Institute, Universitat Politècnica de València, Valencia, Spain

ARTICLE INFO

Keywords:
Conceptual Modeling
Data repositories
Data integration
Biological Datasets
Genomics
Scientific Databases

ABSTRACT

The human genome complexity is captured by many signals, representing for instance DNA variations, the expression of gene activity, or DNA's structural rearrangements; a rich set of data types and formats is used to record these signals. Conceptual models can support the description and explanation of the genome's elaborate structure and behavior. Among others, the Conceptual Schema of the Human Genome (CSG) provides a *concept-oriented, top-down* representation of the genome behavior, which is independent of data formats. The Genomic Conceptual Model (GCM) provides instead a *data-oriented, bottom-up* representation, targeting a well-organized, unified description of these formats. In this research, we join the two approaches to achieve PoliViews, a comprehensive model that links (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. Their dynamic connection 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: data records can be semantically described by high-level concepts exploiting their links and, in turn, the continuously evolving abstract model can be extended thanks to the input provided by real datasets. PoliViews enables expressing queries that employ a holistic conceptual perspective on the genome, directly translated onto data-oriented terms and organization. Here, we demonstrate the approach by linking two major genomic data types, namely DNA variation and gene expression. For each type, we consider different eminent data sources; we describe their mapping with the corresponding view in the concepts layer, enabling an *intra-data-type* integration. Then, leveraging on the connections available in the concepts layer, we show how the distinct data types can be interoperated, enabling an *inter-data-type* integration. The PoliViews approach is shown through several examples of biological interest and can be further extended to any kind of genomic information.

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 is underlying our characteristics and behaviors. Many other aspects, such as DNA mutations, the expression of gene activity, DNA's structural rearrangements, and long-distance contacts between DNA regions, are now used to extract complex signals from the DNA, exploiting Next Generation Sequencing [59] technologies; a rich set of data formats is used to represent such signals. The study of genomic information has practical implications in 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 the 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 [46, 15, 3], whereas a holistic view – that incorporates the semantics of different genomic regions – has not been embraced yet. Important conceptual models (CMs) have supported the effort of explaining the elaborate structure of genomic information since 2000 [54, 11]. However, there remains a gap between CMs about genome data (i.e., that represent "genome data as it is" – usually generated in labs without a conceptual characterization) and CMs that are purely about the genome (that model "data as it should be"). In this work, we defend that elements obtained from the first kind of CMs must be connected with their corresponding elements in the

*Co-first corresponding authors
anna.bernasconi@polimi.it (A. Bernasconi); algarsi3@pros.upv.es (A. García S.); stefano.ceri@polimi.it (S. Ceri); opastor@pros.upv.es (O. Pastor)
ORCID(s): 0000-0001-8016-5750 (A. Bernasconi); 0000-0001-5910-4363 (A. García S.); 0000-0003-0671-2415 (S. Ceri); 0000-0003-0671-2415 (O. Pastor)

CMs that represent higher-level conceptual genome knowledge. We characterize the process of connecting concepts with their associated data as "top-down", whereas we use the term "bottom-up" when connecting data to concepts.

A number of works, summarized by the Conceptual Schema of the human Genome (CSG, [53]) produced by the PROS research center, provide a *concept-oriented, top-down* representation of the genome that is independent of the data formats, aiming to give a template of how the genome is supposed to behave, thereby building a general understanding of the language of life [28]. A parallel initiative, represented by the Genomic Conceptual Model (GCM, [9]) produced by the GeCo project [16], provides a *data-oriented, bottom-up* representation, targeting a high-level, abstract description of genomic data formats, focusing on what they capture and how, contributing to favor the joint use of the represented signals [8, 13].

The two existing approaches are different from two perspectives: 1) how they deal with the concepts representing the knowledge of genomics and 2) how they manage their instantiation in the form of data. 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' definitions 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.

The two approaches are not incompatible. On the contrary, they can be interoperated. Genomic information can be interpreted as a dual system that is approached in two opposite directions: 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). In a preliminary conference paper [10], we first presented our intention of connecting these two perspectives. Here, we reinforce our proposal, by describing the comprehensive PoliViews approach to facilitate genome data management with sound CM support.

The manuscript is organized as follows. Section 2 introduces the existing approaches that employ conceptual modeling in genomics, how they individually deal with concepts and data, and motivates the need for the PoliViews effort. Section 3 presents the PoliViews approach, describing its unified conceptual model with a concepts layer, a data layer, and a protocol for establishing links. Section 4 instantiates the approach on two relevant genomic data types, i.e., DNA variation and gene expression. We describe how the relevant parts of the model are built, how data are mapped to concepts, and pose the basis for its joined use. Section 5 provides a wide collection of examples queries that are enabled by PoliViews, working intra-data-type (targeting concepts either of the DNA Variation view or of the gene expression view, across different data sources) and inter-data-type (targeting concepts also across views). Section 6 discusses the benefits of the PoliViews approach, which is applicable to other genomic data types; it will be possible to develop additional views and use them together, towards a more general conceptual modeling-based perspective on the human genome.

2. Background

Conceptual models have been used to describe the elaborate structure and behavior of the genome. The first models representing DNA genomic sequences date back to the late nineties [51, 47], whereas – in the 2000s – Paton et al. proposed a set of data models for describing transcription/translation processes [54], as well as genomic sequences and protein structures [11]. Additional works exploited conceptual models' expressive power for explaining biological entities and their interactions as conceptual data structures [19, 38, 36, 56]. Many of these conceptual approaches have been followed by the proposal of working prototypes of information systems or databases, based on the initial schemas. These include several classic works as the GEDAW UML Conceptual schema [35] (gene-centric data warehouse), the Genome Information Management System [20] (genome-centric data warehouse), the GeneMapper Warehouse [22] (integrating expression data from genomic sources), BioStar [65] (data warehouse capturing biomedical data semantics), BioMart [61] (conceptual modeling-based data warehouse), and GPKB [44] (warehouse for integrating genomic and proteomic information).

This research background has created a set of Conceptual Modeling-based approaches that focus either on the conceptual characterization of the genome structure or on its data-driven application as used in practice. The characterization of the correct connection between the conceptual and the data-driven perspectives emerges as an attractive, relevant problem that we tackle in the paper. To explore these different but complementary dimensions, in this work we have considered two important approaches that have tackled genomics from a conceptual modeling perspective; both of them lead to solid threads of research, as briefly described next.

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 [53] and producing several extensions since then [57, 28]. The schema now results in a rich map of concepts and relationships that support the holistic understanding of different knowledge modules. The most recent version, i.e., the Conceptual Schema of the Genome v3 (CSG) is reported in [29]. The main objective of CSG stands in identifying relevant concepts and their connections, independent of how datasets are really represented in available databases and sources. According to its nature, the CSG model changes constantly, following evolving requirements, whereas the GCM model evolves when new heterogeneous datasets are produced.

The approach devised within the data-driven Genomic Computing (*GeCo*) group, funded by the ERC AdG 693174 (2016-2021), has developed models 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 [8]. Information is divided between: region data (representing actual genomic elements, measured by experiments – using the Genomic Data Model, GDM [46]) and metadata (descriptions of genomic experiments – captured by the Genomic Conceptual Model, GCM [9]), which make data searchable [13]. Finally, the modeled datasets attempt to resolve data-level interoperability, thereby enabling powerful queries using, e.g., the GenoMetric Query Language (GMQL system [45]).

In this work, we propose to join these two independent directions within an innovative approach, named PoliViews, aiming to provide a more complete vision of the steps that go from the collection of genomic data to the understanding of life mechanisms. On the one hand, we employ the CSG as the model that describes concepts (*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 (*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, with possibly multiple SAMPLEREGIONS, i.e., records representing fragments of the genome with specific measured properties. Genomic regions are those typically produced within the scope of large cooperative efforts, open for public use, and made available for secondary research use [7], including for example the Encyclopedia of DNA Elements (ENCODE, [21]), Genomic Data Commons (GDC, [33]), Gene Expression Omnibus (GEO, [5]), Roadmap Epigenomics [39], and the 1000 Genomes Project [1]. Each of SAMPLEREGION instance can be linked to its corresponding concept in the concepts layer; new links are established when specific data types are chosen (in the data layer) triggering the selection of specific views (of the concepts layer).

The benefit is mutual: 1) the GCM is extended by the power of concepts, which enable high-level semantic-aware querying; 2) the CSG is empowered by the links to real-world data, which allow for building computations on experimental instances and obtaining biologically-relevant results.

3. The PoliViews Approach

The general two-layer PoliViews schema contains:

- a *concepts layer* capturing the knowledge available about the human genome mechanisms (inspired by the CSG [29]);
- a *data layer* representing genomic data, with its types and experiments, captured by information structures and formats (inspired by the original GCM [9] for metadata and by the Genomic Data Model [46] for region data).

Data layer. It is centered on the SAMPLE concept, as schematized in Figure 1, and holds two metadata perspectives. The biological perspective contains the REPLICATE to which a sample belongs; this is part of a BIOSAMPLE, extracted from a DONOR. The organizational perspective 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). Many samples are grouped into a DATASET, which is homogeneous in the schema and in the experiment type. Samples contain multiple SAMPLEREGIONS, corresponding to rows in a file, 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. The last two mentioned classes are additions to the PoliViews data layer (with respect to the original GCM): they are necessary to manage the linking between the two layers.

Concepts layer. The concepts layer is based on the last version of the CSG [29], which includes five modules, respectively describing i) the structure of the human genome; ii) protein synthesis; iii) changes in the sequence w.r.t. 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 generated manually by the conceptual modelers of the PROS group. It is also incrementally enriched as new mechanisms are understood 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.

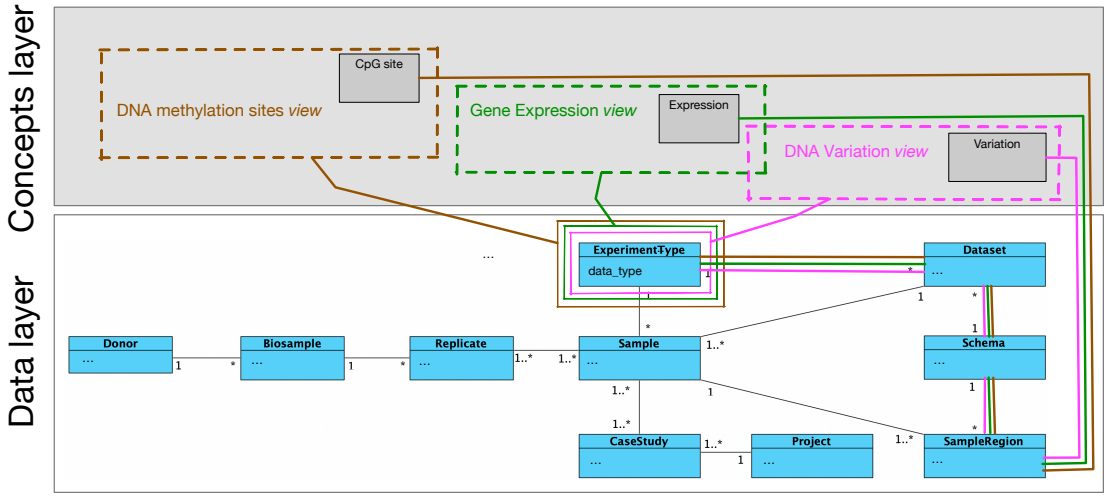

Figure 1: Link the concepts layer and the data layer by means of connections between sample regions and concepts.

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 Figure 1. In their previous description (GCM [9]), data types were forced into containers (i.e., SAMPLES) that flattened their semantics for facilitating integration and processing; 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 or the expression of a gene) 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. Similarly, a SAMPLEREGION measured through a transcriptome profiling experiment, can be represented by its related concept, i.e., the expression levels of the gene BRCA1 detected in a specific tissue in given experimental settings.

Much in the spirit of Ontology-Based Data Access (OBDA [12]) approaches we envision the primary use mechanism of our two-layer schema as an Identification-Selection-GEneration process [30]:

1. Identification of a *genomic data type* (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 explaining the content of that data type.

4. Application of the PoliViews Approach

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 in DNA (frequency of variations among populations, their association with phenotypes, somatic mutations, copy number variation, or structural rearrangements); the behavior of RNA (gene, miRNA, or isoform expression); epigenetic signals (such as DNA methylation, DNA binding, or DNase I Hypersensitive sites); single cell signals and 3D contact matrices—more and more studied in the last years.

In the following, we discuss the case of two of these signals, thus building the DNA Variation view (Section 4.1) and the Gene Expression view (Section 4.2). For each view, many different data representations may be used to indicate the same concepts. Their 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, each of the two next sections first describes the construction of the view in the concepts layer, then shows its mapping to datasets as collected from crucial research projects.

4.1. Modeling DNA Variation

As a first instantiation of our approach, 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 the CSG, as guidance to design the related concepts layer *view*.

The result of the modeling effort is shown in pink in the upper box of Figure 2. The obtained schema is composed of 21 entity classes, six generalizations, one composition, and two aggregations. The central class of this schema is the VARIATION, with a *date*, *name*, *description*, *type* (deletion, insertion or substitution), reference allele, alternative allele, and *ancestral* attribute (representing the allele of the last common ancestor of primates). A variation is located on a named CHROMOSOME. Chromosomes are related to a SPECIES (with a *taxonomy* definition and *scientific*/*common* names). Optionally, VARIATIONS are connected to a specific location, named POSITION on the DNA (with *start*/*end* coordinates on a specific *strand*). For instance, human DNA is made of two strands that are interconnected. In some cases, the position of a variation is located within a confidence interval, recorded in the VARIATIONPOSITION class.

The INDIVIDUAL is another relevant class, representing a specific living being, with several related information (e.g., birth date, blood group, or ethnicity). An individual represents an instance of a given SPECIES; its body has a number of LOCATIONS of interest. A LOCATION is characterized by a *name*, a *description*, and a set of *descriptors* to provide additional information. A specific type of location, particularly useful in genomic sequencing, is the TISSUE (from which a biological sample is derived). VARIATIONS can be identified in a given INDIVIDUAL by means of the READING process, which expresses the particularities of a variation in the individual (such as the *origin* or the *genotype*).

Within a CHROMOSOME, several CHROMOSOMEELEMENTS are hosted (with their *name* and *description*). These include TRANSCRIPTABLEELEMENTS, such as GENES (with a *biotype*, *status*, percentage of G and C nucleotides and their alternative *gene_synonyms*), and REGULATORYELEMENTS that regulate genes, such as ENHANCERS. Elements present possibly multiple ELEMENTPOSITIONS (a specialization of the POSITION class); these are always measured with respect to an ASSEMBLY, i.e., a reference system based on a community-defined sequence (with a *name* and *date*). Each observed chromosome has a CHROMOSOMESEQUENCE, which is also based on the assembly.

The schema allows for three types of VARIATIONS, depending on how their position is considered:

- Variations whose position is known precisely: they are associated with at least one instance of the VARIATIONPOSITION class without confidence interval (i.e., *ci_start* and *ci_end* values are zero).
- Variations whose position is not known precisely: they are associated with at least one instance of the VARIATIONPOSITION class with a specific confidence interval (i.e., *ci_start* and *ci_end* values are not zero).
- Variations whose position is not known: these are not associated with any instance of the VARIATIONPOSITION class.

In the context of a GROUPOFINDIVIDUALS (with *name*, *description*, *geographic_region*, and *size*), a VARIATION presents an ALLELEFREQUENCY, where the *frequency* reports the percentage of the *allele* within the considered pop-

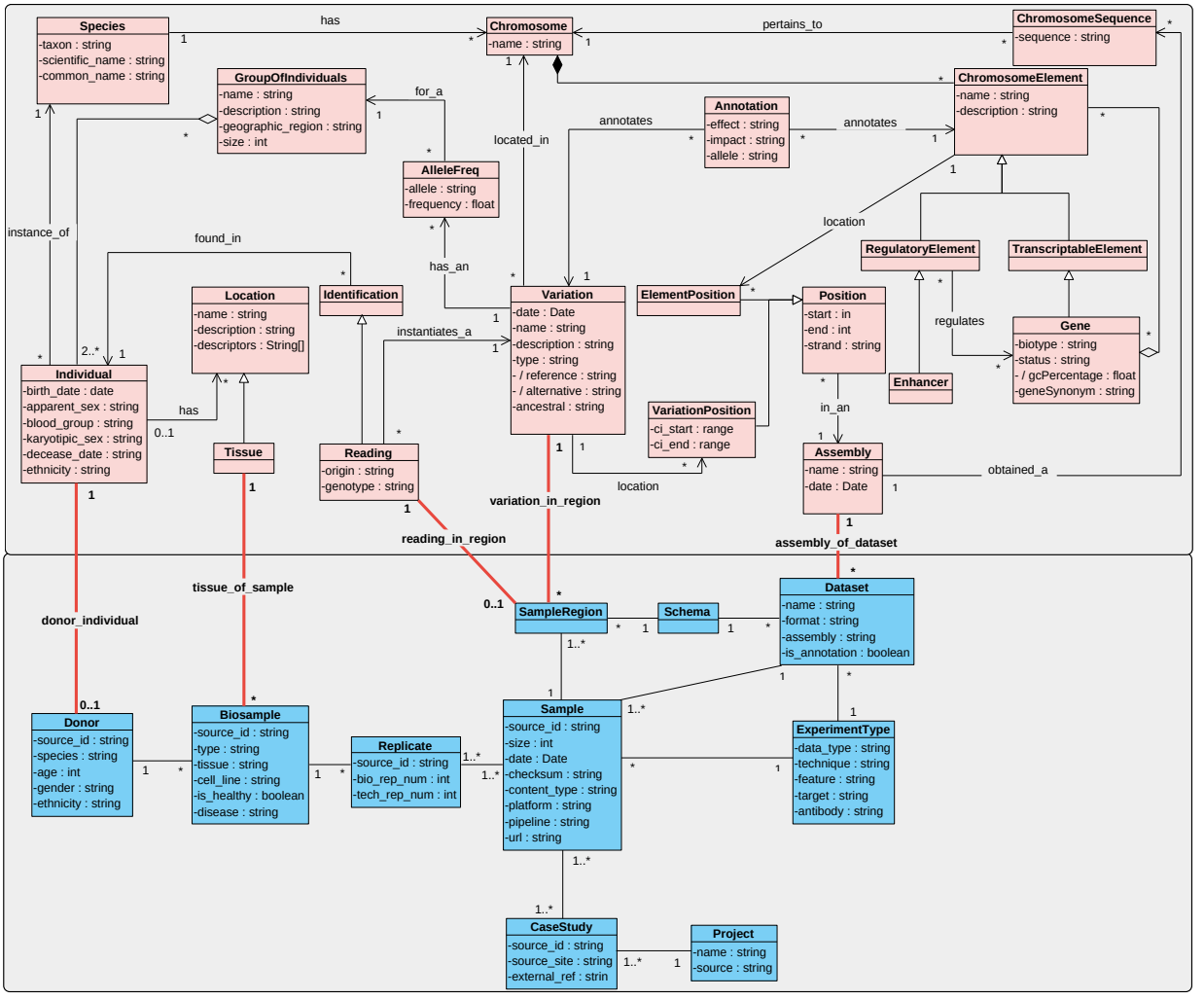

Figure 2: Representation of the DNA variation view (concepts layer, in pink) and the related datasets (data layer, in blue).

ulation. VARIATIONS can alter the functionality of genes; this possibility is represented within ANNOTATION class, with an *impact*, *effect*, and *allele*.

When a full correspondence between the DNA Variation view in the concepts layer and the data layer is established, the complete schema is obtained as in Figure 2, where the top layer (described above) is the concepts layer and the bottom layer is the data layer. Here, the INDIVISUAL (concepts layer) is connected to the DONOR (data layer); the TISSUE (concepts layer) to the BIOSAMPLE (data layer); a READING (concepts layer) corresponds to recording a genomic feature into a SAMPLEREGION (data layer); the VARIATION (concepts layer) is the observed feature, captured in the SAMPLEREGION (data layer). Finally, the ASSEMBLY (concepts layer) becomes a property of DATASETS (data layer).

4.1.1. Mapping with Real DNA Variation Datasets

The Cancer Genome Atlas (TCGA,[68]) is a landmark cancer genomics program that sequenced and characterized over 11,000 patients of primary cancer samples, analyzing them with different experiments, including two dedicated to somatic mutations and transcriptome profiling (capturing the expression of genes, as elaborated in the next section). The 1000 Genomes Project (1KGP, [1]) is an international research effort established to create a catalog of common human germline variations, using samples from healthy people. In the GMQL data repository [8], the GeCo group has analyzed all the data fields contained in the datasets' schemas that refer to this data type. Here, we considered 1000

Table 1
Excerpt of the relational schema of the data layer, detailing the 1000 Genomes Project population variation sample regions and the TCGA masked somatic mutations sample regions.

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")

Table 2
Mapping rules for building the relational schema of the DNA Variation *view* in the concepts layer. For ease of reference, lines divide rules related to different classes. Notice that the last five rules do not depend on the choice of the data source (TCGA or 1KGP).

Concept.VARIATION(*gen*(),name,*gen*(),type,reference,alternative,ancestral) ⊇ Data.SAMPLEREGION1KGP(_,_,_,_,_,_,ref.,alt., type,_,name,_,_,_,_,_,_,_,_,_,_,_,anc.,_,_,_,_)
Concept.VARIATION(*gen*(),name,*gen*(),type,reference,*f*(reference,allele1,allele2),null) ⊇ Data.SAMPLEREGIONTCGA(_,_,_,_,_,_,_,type,reference,allele1,allele2,name,_)
Concept.VARIATIONPOSITION(ci_start,ci_end) ⊇ Data.SAMPLEREGION1KGP(_,ci_end,ci_start,_)
Concept.VARIATIONPOSITION([0,0],[0,0]) ⊇ Data.SAMPLEREGIONTCGA(_,_)
Concept.POSITION(start,end,strand) ⊇ Data.SAMPLEREGION1KGP(_,start,end,strand,_)
Concept.POSITION(start,end,strand) ⊇ Data.SAMPLEREGIONTCGA(_,start,end,strand,_,_,_,_,_,_,_,_,_)
Concept.ALLELEFREQUENCY(allele,frequency) ⊇ Data.SAMPLEREGION1KGP(_,_,_,_,_,_,_,allele,_,_,_,_,_,_,_,_,frequency,_,_,_,_,_,_,_,_,_)
Concept.READING("germline",*f*(AL1,AL2)) ⊇ Data.SAMPLEREGION1KGP(_,_,_,_,AL1,AL2,_,"germline")
Concept.READING("somatic",*f*(allele1,allele2)) ⊇ Data.SAMPLEREGIONTCGA(_,_,_,_,_,_,_,_,allele1,_,allele2,_,"somatic")
Concept.CHROMOSOME(name) ⊇ Data.SAMPLEREGION1KGP(name,_)
Concept.CHROMOSOME(name) ⊇ Data.SAMPLEREGIONTCGA(name,_,_,_,_,_,_,_,_,_,_,_,_)
Concept.CHROMOSOMEELEMENT(name,*gen*()) ⊇ Data.SAMPLEREGIONTCGA(_,_,_,_,name,_,_,_,_,_,_,_,_)
Concept.GENE(*f*(geneSynonym),*f*(geneSynonym),*f*(geneSynonym),geneSynonym) ⊇ Data.SAMPLEREGIONTCGA(_,_,_,_,_,geneSynonym,_,_,_,_,_,_,_)
Concept.ANNOTATION(effect,*f*(effect),*f*(allele1,allele2)) ⊇ Data.SAMPLEREGIONTCGA(_,_,_,_,_,_,_,effect,_,_,allele1,allele2,_,_)
Concept.ASSEMBLY(name,*f*(name)) ⊇ Data.DATASET(_,_,_,name,_)
Concept.SPECIES(*f*(scientificName),scientificName,*f*(scientificName)) ⊇ Data.DONOR(_,scientificName,_,_,_)
Concept.INDIVIDUAL(*f*(age),gender,null,null,null,ethnicity) ⊇ Data.DONOR(_,_,age,gender,ethnicity)
Concept.TISSUE(name, *gen*(), *f*(is_healthy,disease)) ⊇ Data.BIOSAMPLE(_,"tissue",name,_,is_healthy,disease)
Concept.GROUPOFINDIVIDUALS(ethnicity,*gen*(),*f*(ethnicity),*gen*()) ⊇ Data.DONOR(_,_,_,_,ethnicity)

Genomes Project datasets [2] and TCGA datasets related to masked somatic mutations [14].

For demonstrating one possible implementation of the PoliViews approach, we employ a relational database representation. 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 and two rows for 1KGP). For sample regions, we employ one table for each different dataset. For simplicity, we refer to SAMPLEREGIONTCGA (with 26 attributes) and SAMPLEREGION1KGP (13 attributes).

Mapping rules are used to describe how datasets information can be mapped into the concepts schema, considering the view that is specific to DNA variation. Table 2 provides the mappings for the TCGA and 1KGP datasets. Each mapping rule is a logic formula (in Datalog-like syntax [17]) 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 Table 1. As an example, the entity VARIATION of the concepts schema is filled using data from the SAMPLEREGION1KGP 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, when the mapping was meaningful, we wrote a different rule for each pair of the output table (in the concepts layer) and input table (1KGP or TCGA in the data layer).

All classes contain an identifier attribute, omitted here for brevity. In some cases, we need to derive new attributes in the concepts layer schema as functions of original attributes. One such example is in the VARIATION table: here, the second attribute *alt* requires combining the values of three attributes in the input table SAMPLEREGIONTCGA. For this, we use the notation *f(...)*. Moreover, dates or descriptions are generated from the system admin (with *gen()*). Here we do not report concepts layer's tables that could not be directly mapped to any attribute of the two data sources considered in this example; this is the case of CHROMOSOMESEQUENCE, for instance, whose attribute *sequence* can be filled by inspecting authoritative sources such as RefSeq [52].

4.2. Modeling Gene Expression

As a second instantiation of the PoliViews approach, we focus on another popular genomic study, i.e., transcription profiling (or expression profiling). This involves the quantification of gene expression of many genes in cells or tissue samples at the transcription level (i.e., in the RNA). This experiment is frequently used in clinical practice to encode the levels of gene expression in patients. Altered levels of gene expression have been associated with a wide range of disorders, such as neurodegenerative diseases [34, 63]. We carefully considered the module of the CSG describing the structure of the human genome; this was employed as guidance to design the concepts layer's *view* on Gene Expression. The result of the modeling effort is shown in green in the upper box of Figure 3.

The obtained schema is composed of 15 classes, four generalizations, and one aggregation. Most of the classes are in common with the previously introduced DNA variation view, therefore are not described in the following. The two central classes are the GENE (already presented) and EXPRESSIONRATE (new); the latter describes the expected expression level of a GENE on a given LOCATION by means of the *default_value* attribute. INDIVIDUALS have their own levels of expression, which can differ from the expected value. These changes are captured by the EXPRESSION class, with a given *value* and *identifier*.

When a full correspondence between the Gene Expression view in the concepts layer and the data layer is established, the complete schema is obtained as in Figure 3. As before, the INDIVIDUAL (concepts layer) corresponds to the DONOR (data layer); the TISSUE (concepts layer) to the BIOSAMPLE (data layer), and the ASSEMBLY (concepts layer) is specified in the DATASET (data layer). Differently, the GENE (concepts layer) now represents SAMPLEREGION (data layer), where its EXPRESSION (concepts layer) is also recorded.

4.2.1. Mapping with Real Gene Expression Datasets

Gene expression datasets are produced and handled by different consortia and provided through different data source platforms. To practically discuss how concepts can be instantiated into data records in real-world scenarios, we consider the use of datasets quantifying gene expression as collected within three eminent research projects:

- TCGA [68], already described above, not only characterizes the DNA variation in cancer patients, but it also presents a large dataset of transcriptome profiling for 33 distinct cancer types–now retrievable through the Genomic Data Commons platform [33]. In the GMQL data repository [8], the GeCo group has analyzed all the data fields contained in the schema of the 'gene expression quantification' dataset [14].

- Gene Expression Omnibus (GEO) [5] is the most general and widely used among repositories. It started in 2002 as a versatile, international public repository for gene expression data [24]; it consequently adopted a

more flexible and open design to allow also non-expression data since 2008. In genomics research, for authors of publications, it is customary (or even required by journals upon submission [48]) to deposit their raw and processed datasets in primary deposition archives. GEO is considered a primary deposition archive [7]; as such, does not impose any fixed schema to the metadata or data formats of the submitted datasets.

- The GTEx Consortium [42] aims at establishing a resource database and associated tissue bank to study the relationship between genetic variation and gene expression and other molecular phenotypes in multiple reference tissues. RNA-Seq data is provided in files divided by tissue and kind of measurement.

Table 3 describes the schemas of the tables designed starting from the presented model. The SAMPLEREGIONTCGAGENEEXPR table (13 attributes) is directly derived from the OpenGDC [14] file schemas of gene expression quantification datasets, translating them into an RDBMS logical schema. The SAMPLEREGIONGTEX (5 attributes) and SAMPLEREGIONGEO (4 attributes) table schemata are extracted using a simple transformation. Indeed, the typical file that can be downloaded from these two sources represents a matrix where rows are genes and columns are patients (or – more in general – biological samples). Each cell stores the expression quantification of that gene in that patient.

RNA-seq data is analyzed using a pipeline that produces reads aligned to the latest version of the reference genome.

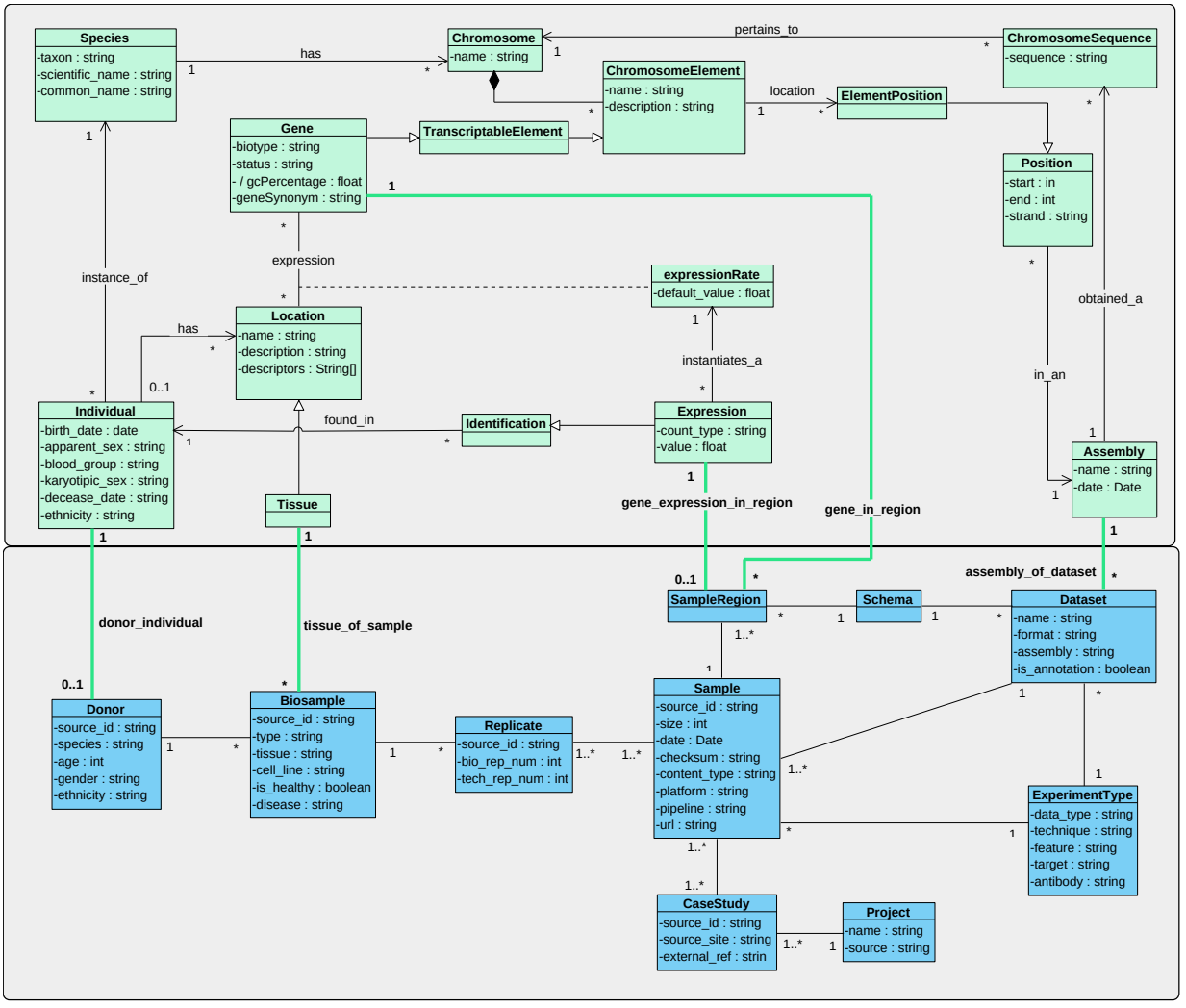

Figure 3: Representation of the Gene Expression view (concepts layer, in green) and the related datasets (data layer, in blue).

Table 3
Excerpt of the relational schema of the data layer, detailing the sample regions for the TCGA gene expression dataset and typical Gene Expression Omnibus and GTEx datasets.

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.SAMPLEREGIONTCGAGENEEXPR(chr,start,stop,strand,gene_id,gene_name,gene_type,unstranded,stranded_first, stranded_second,tpm_unstranded,fpkm_unstranded,fpkm_uq_unstranded)
Data.SAMPLEREGIONGTEX(id,Name,Description,gene_reads,gene_tpm)
Data.SAMPLEREGIONGEO(geneID,rawCounts,genes.tpm)

The produced measurements are named 'raw counts' (i.e., the number of mapped reads summarized and aggregated over each gene). More elaborate estimates of gene expression can be achieved by applying FPKM (fragments per kilobase of exon model per million mapped reads) or TPM (transcripts per million) to the raw counts.

The GEO and GTEx data sources produce one matrix for each kind of employed count. For most experiments, they provide both raw counts and TPMs. In order to standardize the representation of gene expression w.r.t. the previous tables (e.g., prepared for the TCGA data source), we chose to flatten the matrix into several separate records. The simple transformation process is illustrated in Figure 4, showing the two matrices (one for raw counts and one for TPMs) provided for the measurements of thousands of individuals (called 'GTEX-*' in the columns) for the bladder tissue. The output format of a typical sample region contains the gene id, name, and description, and two attributes corresponding to raw counts and TPMs, merged from the two input matrices.

Note that, the INDIVIDUAL class introduced in the Gene Expression view (Figure 3) is fundamental for properly capturing the semantics of this data type. This is a valuable addition that was not included in the models proposed in [10].

Excerpt from the `gene_reads_2017-06-05_v8_bladder.gct` file

id	Name	Description	GTEX-OIZF-1926-SM-7PBZS	GTEX-P44H-2226-SM-E9U4P	GTEX-QEL4-1826-SM-EZ6KU
0	ENSG00000223972.5	DDX11L1	1	0	0
1	ENSG00000227232.5	WASH7P	136	187	116
2	ENSG00000278267.1	MIR6859-1	0	0	0
3	ENSG00000243485.5	MIR1302-2HG	1	0	1
4	ENSG00000237613.2	FAM138A	0	0	2
5	ENSG00000268020.3	OR4G4P	1	1	1

+

Excerpt from the `gene_tpm_2017-06-05_v8_bladder.gct` file

id	Name	Description	GTEX-OIZF-1926-SM-7PBZS	GTEX-P44H-2226-SM-E9U4P	GTEX-QEL4-1826-SM-EZ6KU
0	ENSG00000223972.5	DDX11L1	0.0174	0.0000	0.0000
1	ENSG00000227232.5	WASH7P	3.0830	7.0710	2.6030
2	ENSG00000278267.1	MIR6859-1	0.0000	0.0000	0.0000
3	ENSG00000243485.5	MIR1302-2HG	0.0348	0.0000	0.0344
4	ENSG00000237613.2	FAM138A	0.0000	0.0000	0.0489
5	ENSG00000268020.3	OR4G4P	0.0358	0.0598	0.0355

=

Output format used for the formalization of the `SampleRegionGTEx` table

ID of the sample region	id	Name	Description	gene reads	gene tpm
GTEX-OIZF-1926-SM-7PBZS	0	ENSG00000223972.5	DDX11L1	1	0.0174
GTEX-P44H-2226-SM-E9U4P	0	ENSG00000223972.5	DDX11L1	0	0
GTEX-QEL4-1826-SM-EZ6KU	0	ENSG00000223972.5	DDX11L1	0	0
GTEX-OIZF-1926-SM-7PBZS	1	ENSG00000227232.5	WASH7P	136	3.083
GTEX-P44H-2226-SM-E9U4P	1	ENSG00000227232.5	WASH7P	187	7.071
GTEX-QEL4-1826-SM-EZ6KU	1	ENSG00000227232.5	WASH7P	116	2.603
...	...	...	...	...	...

Figure 4: Graphical representation of the transformation that is applied to original files extracted from the GTEx data source, representing the gene expression quantification (using raw counts in the first excerpt and TPMs in the second excerpt). The result is a set of rows with the schema of the SAMPLEREGIONGTEX table. The transformation is purely syntactical, changing the data structure; values remain intact.

Table 4
Mapping rules for building the relational schema of the Gene Expression *view* in the concepts layer. As above, lines divide rules related to different classes. Notice that the last five rules do not depend on the choice of the data source (TCGA, GTEx, or GEO).

Concept.GENE(gene_type,f(gene_name),f(gene_name),f(gene_id,gene_name)) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR (_,_,_,_,gene_id,gene_name,gene_type,_,_,_,_,_,_)
Concept.GENE(f(Description),f(Description),f(Description),f(Name,Description)) ⊇ Data.SAMPLEREGIONGTEX(_,Name,Description,_,_)
Concept.GENE(f(geneID),f(geneID),f(geneID),geneID) ⊇ Data.SAMPLEREGIONGEO(geneID,_,_)
Concept.CHROMOSOMEELEMENT(gene_name,f(gene_name)) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(_,_,_,_,_,gene_name,_,_,_,_,_,_,_)
Concept.CHROMOSOMEELEMENT(Description,f(Description)) ⊇ Data.SAMPLEREGIONGTEX(_,_,Description,_,_)
Concept.CHROMOSOMEELEMENT(f(geneID),f(geneID)) ⊇ Data.SAMPLEREGIONGEO(geneID,_,_)
Concept.CHROMOSOME(chr) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(chr,_,_,_,_,_,_,_,_,_,_,_,_)
Concept.CHROMOSOME(f(Description)) ⊇ Data.SAMPLEREGIONGTEX(_,_,Description,_,_)
Concept.CHROMOSOME(f(geneID)) ⊇ Data.SAMPLEREGIONGEO(geneID,_,_)
Concept.POSITION(start,end,strand) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(_,start,end,strand,_,_,_,_,_,_,_,_,_)
Concept.POSITION(f(Description),f(Description),f(Description)) ⊇ Data.SAMPLEREGIONGTEX(_,_,Description,_,_)
Concept.POSITION(f(geneID),f(geneID),f(geneID) ⊇ Data.SAMPLEREGIONGEO(geneID,_,_)
Concept.EXPRESSION("unstranded",unstranded) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(_,_,_,_,_,_,_,unstranded,_,_,_,_,_)
Concept.EXPRESSION("stranded_first",stranded_first) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(_,_,_,_,_,_,_,stranded_first,_,_,_,_)
Concept.EXPRESSION("stranded_second",stranded_second) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(_,_,_,_,_,_,_,stranded_second,_,_,_)
Concept.EXPRESSION("tpm_unstranded",tpm_unstranded) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(_,_,_,_,_,_,_,tpm_unstranded,_,_)
Concept.EXPRESSION('fpkm_unstranded",fpkm_unstranded,) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(_,_,_,_,_,_,_,fpkm_unstranded,_)
Concept.EXPRESSION("fpkm_uq_unstranded",fpkm_uq_unstranded) ⊇ Data.SAMPLEREGIONTCGAGENEEXPR(_,_,_,_,_,_,_,fpkm_uq_unstranded)
Concept.EXPRESSION("unstranded",gene_reads) ⊇ Data.SAMPLEREGIONGTEX(_,_,_,gene_reads,_)
Concept.EXPRESSION("tpm_unstranded",value) ⊇ Data.SAMPLEREGIONGTEX(_,_,_,_,gene_tpm)
Concept.EXPRESSION("unstranded",rawCounts) ⊇ Data.SAMPLEREGIONGEO(_,rawCounts,_)
Concept.EXPRESSION("tpm_unstranded",gene.tpm) ⊇ Data.SAMPLEREGIONGEO(_,_,gene.tpm)
Concept.ASSEMBLY(name,f(name)) ⊇ Data.DATASET(_,_,_,name,_)
Concept.SPECIES(f(scientificName),scientificName,f(scientificName)) ⊇ Data.DONOR(_,scientificName,_,_,_)
Concept.INDIVIDUAL(f(age),gender,null,null,null,ethnicity) ⊇ Data.DONOR(_,_,age,gender,ethnicity)
Concept.TISSUE(name, *gen*(), f(is_healthy,disease)) ⊇ Data.BIOSAMPLE(_,"tissue",name,_,is_healthy,disease)

As in the case of the DNA Variation view, Datalog-like mapping rules are used to describe how datasets information can be mapped into the concepts schema, considering the view that is specific to Gene Expression. Table 4 provides the mappings for the TCGA Gene Expression, GTEx, and GEO datasets. We wrote a different rule for each pair of the output table (in the concepts layer) and input table (TCGA Gene Expression, GTEx, or GEO in the data layer). Again, all classes contain an identifier attribute, omitted here for brevity. Also here we do not report concepts layer's tables that could not be directly mapped to any attribute of the two data sources considered in this example.

5. Practical Examples

The value of the proposed PoliViews stands in its ability to prospectively enable several data integration processes and guide the query processes aiming to gather and interoperate data that is heterogeneous in formats, provenance, and represented semantics. To demonstrate its usefulness we next provide several examples. First, we focus on *intra-data-type* exploitation of the model, using first its DNA Variation view (see Section 5.1) and then its Gene Expression view (see Section 5.2); these applications allow the representation of datasets sourced from different consortia and platforms. Finally, we propose an *inter-data-type* integration, i.e., a more advanced use of PoliViews, with a series of examples that require the joint use of DNA variation and gene expression information.

5.1. Intra-Data-Type Integration: DNA Variation

The genomics community has produced vast, high-quality, publicly accessible databases of human variants for both the germline and the somatic type. In addition to 1000 Genomes and TCGA, many other data sources exist. Several studies employ these datasets together, mainly based on the location of the point-wise mutations and on the co-occurrence of sets of them. This section reports examples of queries enabled by concept-to-data linking, showing that data improves the representation of genome concepts related to DNA Variation and that, in turn, concepts and their connections improve the knowledge-generation process allowing connections in the otherwise isolated datasets.

Example 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 actual 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 [52], whereas it is located at 156,213,205–156,217,881 according to GENCODE [26]. The concepts layer adequately captures these aspects and it allows for posing generic queries while extracting heterogeneous definitions from the data.

Example 2: Extract mutations whose position is not precisely identified. The concepts layer includes the possibility to represent known imprecise variations: a VARIATION is located in a VARIATIONPOSITION, whose attributes *ci_start* and *ci_end* – respectively representing confidence intervals initial and final position – can augment the typical information *(start, end)* on a specific *strand*, provided in the POSITION class. This kind of variation is commonly found in data sources of variation data, 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).

Example 3: Extract mutations located on enhancers associated with breast cancer. Let us consider the study of a patient genome targeting the presence of mutations on BRCA1, i.e., a specific GENE associated with breast cancer and located at the ELEMENTPOSITION 43,044,295–43,170,245 of the negative strand of CHROMOSOME 17. From the data, mutations located in this range can certainly be retrieved. Note that mutation datasets (such as TCGA's ones) may sometimes report correspondence between variations and their enclosing genes; while this is quite standard information, less studied elements are typically not considered. However, in terms of clinical significance, in addition to genes, it is critical to consider also their regulatory elements. In this case, mutations may be tested also on the ENHANCERS of BRCA1. Several data sources can provide this information. For example, the GH17J043124 ENHANCER is reported by GeneCards [58] on the positive strand with an *ElementPosition* 43,123,800–43,127,201 and by ENCODE [25] with an *ElementPosition* 43,124,247–43,126,961, being currently associated with breast cancer [6]. This connection, however, can be made by employing the concepts layer representation. The schema allows making explicit a relation between positions and elements (including genes and enhancers) that remains instead hidden in the data. Figure 5 shows the described example as a UML instance diagram, where the Single Nucleotide Polymorphism (SNP) from guanine (G) to Adenine (A) is a VARIATION with a VARIATIONPOSITION 43,124,064 that has been observed within the GH17J043124 enhancer when the GRCh38 ASSEMBLY is employed as a reference—in the context of the measurement (READING) performed on the breast TISSUE of an INDIVIDUAL. The mutation is somatic (i.e., an alteration in DNA occurred after conception).

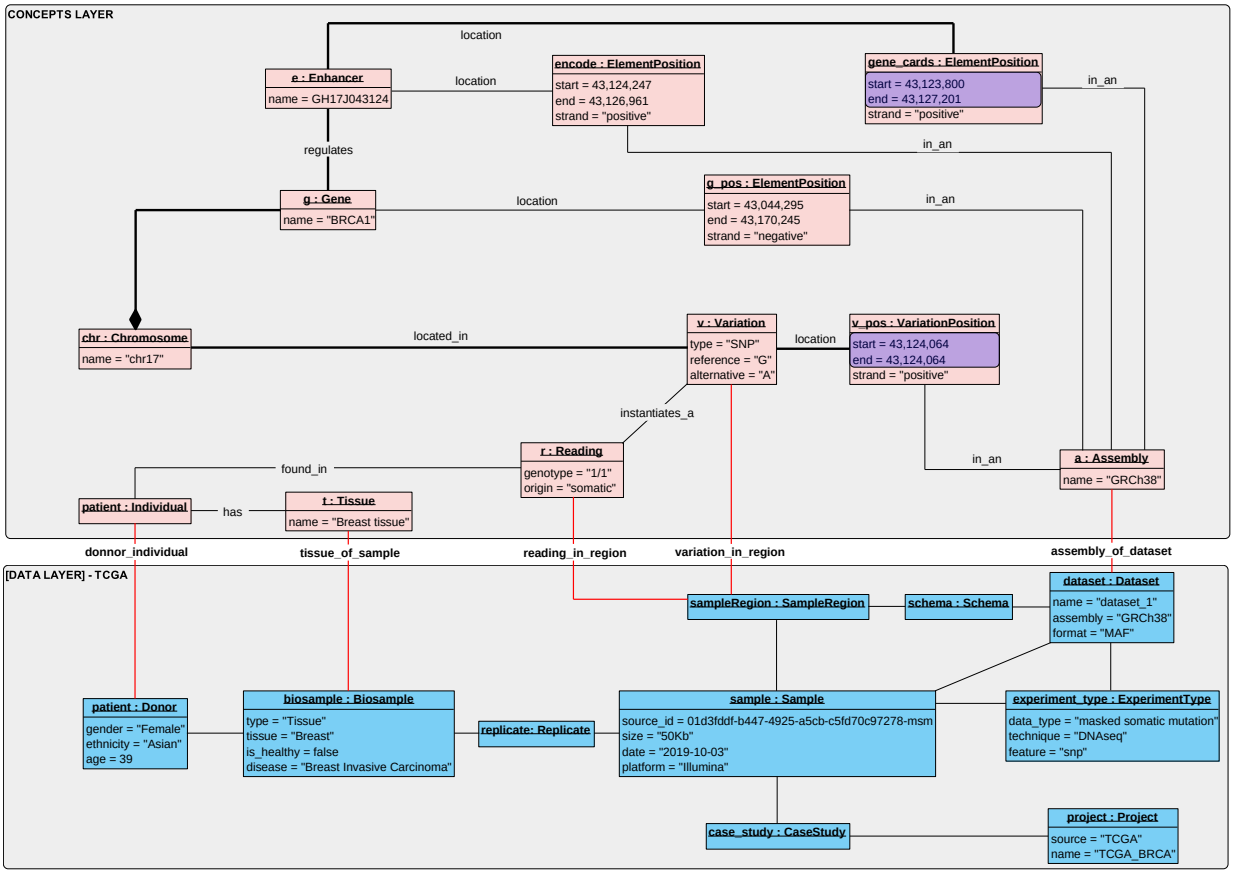

Figure 5: UML instance diagram depicting the scenario described in Example 3.

Example 4: Extract orthologous genes for humans and other species. By exploiting the connection between DONOR (data layer) and INDIVIDUAL of a specific SPECIES (concepts layer) it becomes 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 dogs closely depict the phenotype of human diseases caused by mutations in orthologous genes [32]. By exploiting the findings available for canine genes, candidates for gene-driven therapies may be found, e.g., for Duchenne muscular dystrophy [50].

5.2. Intra-Data-Type Integration: Gene Expression

Measuring gene expression is an important part of genomic studies. The possibility to quantify at which level a particular gene is expressed within a cell or tissue can provide valuable information. Notable uses include the identification of viral infections within a cell (viral protein expression), the characterization of individuals' susceptibility to cancer (oncogene expression), the understanding of bacterial anti-microbial resistance (e.g., beta-lactamase expression, when targeting penicillin), the identification of the molecular signature of a disease [4], or for correlating drug repurposing candidates to a disease [69].

Data that represents the expression of genes is collected widely. GTEx is typically employed as a benchmark of normal data (i.e., extracted from healthy/non-tumor patients), as opposed to diseased (i.e., tumoral in the case of cancer genomics). TCGA is instead typically used for tumor samples. Often, these two sources are used together: there have been efforts to homogenize their data into a common repository (see [66]). GEO, instead, is a very heterogeneous source where researchers deposit datasets supporting their publications, no homogeneous schema is guaranteed by the provided.

Example 5: Extraction of tumor/healthy gene expression samples for colon adenocarcinoma prognosis. Chen et al. [18] describe a data analysis workflow that exploits gene expression signals to identify a biomarker for colon adenocarcinoma. The data acquisition workflow retrieves TCGA gene expression datasets whose primary site is "colon" and whose analyzed histological type is "colon adenocarcinoma". In parallel, gene expression data extracted from normal tissue is sourced from the GTEx colon sigmoid dataset. The two datasets can be combined together by means of a differential expression analysis conducted with off-the-shelf bioinformatics libraries. Figure 6 shows the GTEx-related instance of the data layer in blue color. It is connected to the green classes, representing instances of the concepts layer classes. The core subschema shows, as an example, the WASH7P GENE, located at ELEMENTPOSITION 14,362–29,570 on the negative strand; the gene has been measured in many different conditions in the data. A colon tissue has been extracted in the context of the GTEx project, using a sample from a healthy DONOR (e.g., a cell line). The EXPRESSION of the gene is calculated as 48 gene reads (unnormalized measure). In a similar way, a TCGA data layer could be instantiated to represent the extraction of a patient affected by a similar cancer type. For each of her/his genes, the count of reads would be retrieved. The information is stored in TSV files that hold the gene expression quantification for many different genes and individuals (each represented as one SAMPLEREGIONGTEX or a SAMPLEREGIONTCGA, stored in two different DATASETS. The PoliViews approach accommodates all the heterogeneous representations allowing for their mapping into homogenous concepts.

In [18], after the gene expression data retrieval is completed, A survival and clinical matrix is also sourced from TCGA (we do not detail the integration process with this kind of data here). The complete workflow has the goal of identifying a hub gene (DAPK3) that is significantly associated with the lymphatic invasion and thus can support the colon adenocarcinoma prognosis.

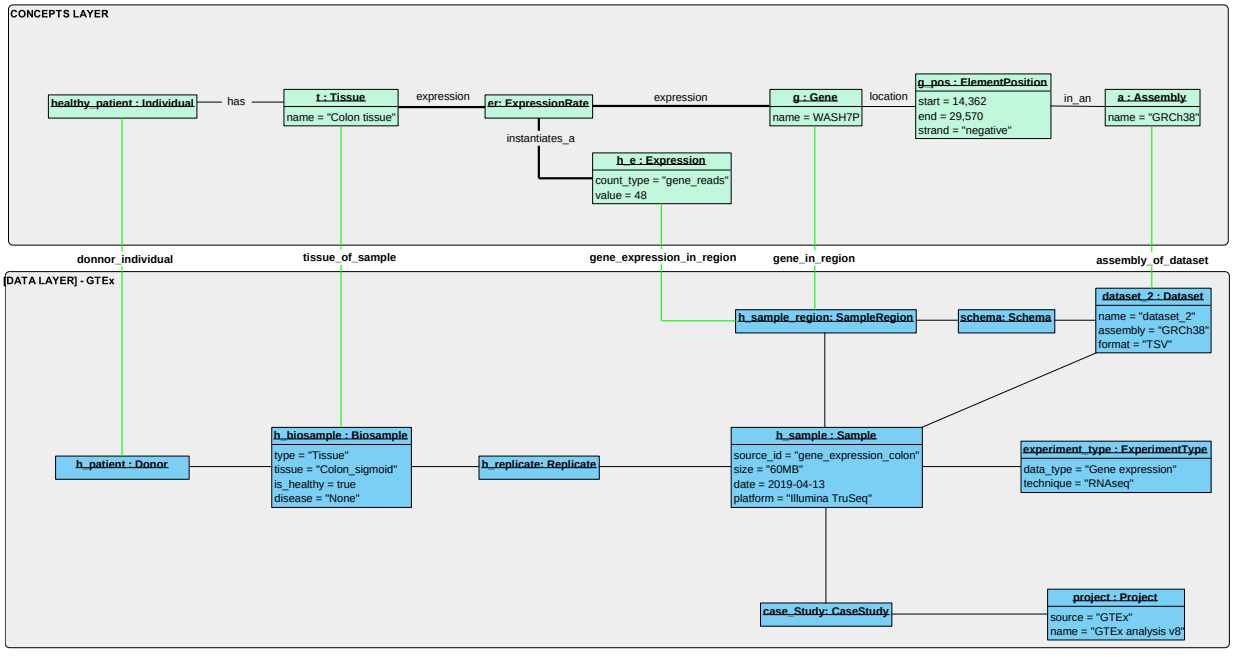

Figure 6: UML instance diagram depicting the scenario described in Example 5.

Example 6: Joint use of tumor/healthy gene expression data for colorectal cancer treatment repurposing. Colorectal cancer is a major cause of cancer deaths worldwide. Many patients are diagnosed at an advanced stage and the 5-year survival rate is only around 30%. Yang et al. [70] employ TCGA and GTEx cohorts contributing to 471 tumor tissues and 349 normal tissues (using the FPKM measure for gene expression). A Weighted Correlation Network Analysis is employed to select the genes that are significantly associated with the analyzed type of tumor. Other statistical analysis steps are applied to build a five-gene prognostic signature (PGM2, PODXL, RHNO1, SCD, and SEPHS1). The data retrieval phase can be mapped on PoliViews in a very similar way as we did with Example 5

in Figure 6.

Note that the signature obtained after the first analysis is further validated using the GSE17536 dataset [27]. The dataset was downloaded from the GEO database, containing 177 samples, to validate the obtained results further. Thanks to the common schemata that these sources share in PoliViews any other GEO sample considered of interest for the study could be added to the pool of collected data. Several studies use similar approaches on a wide plethora of cancer types (e.g., [23] on Renal Cell Carcinoma).

5.3. Inter-Data-Type Integration: Interoperating DNA Variation with Gene Expression data

Figure 7 represents the overall model presented in the manuscript, where the concepts layer includes - for now - the DNA Variation view and the Gene Expression view. Classes related only to DNA Variation are in pink (as in Figure 2): classes related only to gene expression are in green (as in Figure 3), whereas those that are common to both views are colored using a green/pink transition. A GENE is represented through its EXPRESSION that is recorded in the SAMPLEREGION (of a particular data source). Similarly, a VARIATION is captured through a READING operation and recorded into the corresponding SAMPLEREGION. The retrieved sample regions are physical genomic records that can be used for integrated data analysis.

Example 7: Workflow for mapping DNA Variation on highly expressed genes. Settino et al. [60] describe a simple workflow to use the DNA Variation data (and related information on their position on specific elements) together with gene expression data. The first kind of data has been produced using the DMET platform. DMET (not discussed in this manuscript) allows obtaining variation data in a format similar to that of TCGA. In particular, it produces SNPs: each SNP represents a difference in a single DNA building block, called a nucleotide. These are the most common type of genetic variation among people. It is enriched by information on the exon regions (extracted from GENCODE annotation dataset). Only SNPs whose coordinates are included in those of at least one exon region are considered. The second kind of data is extracted from TCGA: Breast Invasive Carcinoma (BRCA) is considered. Only genes with an FPKM count above 3 (i.e., highly expressed) are retained. Finally, the workflow in [60] extracts only SNPs that overlap at least one highly expressed gene in the BRCA dataset.

The same workflow could be applied using directly the TCGA masked somatic mutation dataset and the TGCA gene expression dataset, as done in [45].

Example 8: Interdependence of gene expression and mutations on the same genomes. Several research threads investigated the interdependence of these two signals or exploit them independently to understand pathogenesis mechanisms. The link between gene expression changes and mutation status/effects has been studied previously [64, 40, 49]. On the contrary, others have studied how the functional impacts of somatic mutations in cancer genomes change the expression of genes [31, 37], improving the outcome prediction of certain tumors.

While all the mentioned applications act on tumor-related problems, the same data types have been interoperated also for other research areas. For instance, Weinstein et al. [67] studied adaptive immune response. In this case, it is critical that the link between genetic variability and gene expression at the single-cell level is maintained. This is not trivial to be measured. Weinstein and colleagues proposed a method to simultaneously measure gene expression profiles and genome mutations in single cells.

6. Discussion and Conclusion

The Human Genome is an extremely complex entity, whose enormous information is hard to capture, represent, and operationalize. We have described the concept-driven and data-driven approaches to conceptual modeling for genomics, that guided the development of CSG and GCM. Here, we joined two existing approaches to conceptual modeling of the human genome and proposed PoliViews, a conceptual model that provides both the concept and data viewpoints, 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. Their dynamic connection is established when specific genomic data types are chosen in the data layer, thereby triggering the selection of a view in the concepts layer. Operationally, PoliViews allows us to i) visualize only concepts related to a specific view at a time; ii) appreciate the common concepts and the interconnections between views; iii) when datasets of a specific genomic data type are needed, select them by exploiting the link between concepts and the corresponding SAMPLEREGION; iv) express complex queries that employ a holistic conceptual perspective on the genome, directly translated onto data-oriented terms.

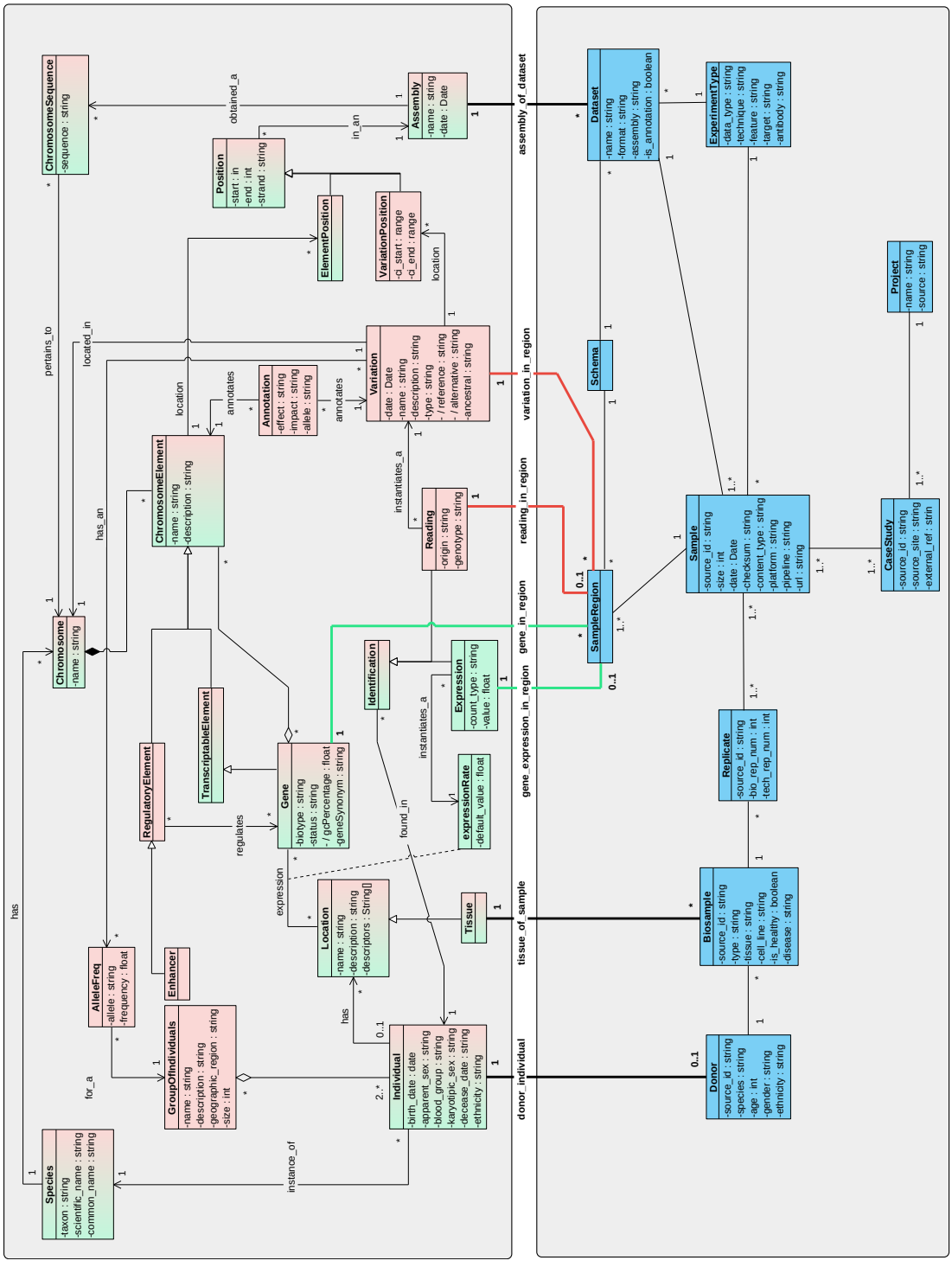

Figure 7: PoliViews schema representing the DNA Variation and Gene Expression views in the concepts layer, connected to their respective data representations in the bottom layer. The classes that only pertain to the DNA Variation view are depicted in pink; those that only pertain to the gene expression view are in green; those that pertain to both views are in green and pink.

The approach is here exemplified using the DNA variation and gene expression data types, showing that the new conceptual model can support interesting queries and applications, acting on a single dataset, on different integrated datasets with the same view, or on different integrated datasets across views. New views can be created that correspond to all genomic information and data types.

PoliViews extends with several novelties previous work published in the ER 2023 conference [10] where the first idea of a modular view-based approach to genomic data management was presented. We applied several changes to the conceptual layer, such as the introduction of the concept of "individual" (on which, e.g., the variation or the gene expression is measured) and the generalization of the variation's position handling strategy. We added the conceptual and data view about the quantification of expressions of genes and introduced the idea of intra-data-type (e.g., using together different data sources from one view) versus inter-data-type integration (i.e., using together data sources coming from different views). Finally, we provided several new examples, especially regarding the Gene Expression view and intra-data-type use cases. As immediate extensions, we plan on adding a concepts layer extension that allows calculating expression rate levels based on the population considered in the data. We will next add a new view dedicated to the methylation values of CpG sites (i.e., regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases on the positive strand) mapping to DNA Methylation signals (used in [41, 45]).

Here, we discussed the conceptual challenge of generating views for all the most relevant genomic data types, while carefully designing their links. We showed the variation-related and the gene expression-related information, but we will next take data types one by one and generate extensions of the concepts layer view by view. The scope of the presented effort is limited to data design and explanation, but it will be further exploited to achieve effective data querying, pending an integration and implementation effort. In this direction, we envision a holistic system that, based on accurate view-specific contents, is able to provide a synergic perspective on the genome. The system will enable the combined use of multiple views, with selective mechanisms that activate one area or the other. To achieve this, the PROS group will enrich the CSG entities by inspecting new datasets and the GeCo will understand how the datasets are connected conceptually. Significant future joint activities are envisioned that will integrate other well-known open genomic data sources [7] and, possibly, also population-specific or nation-scale sequencing initiatives [62], whose datasets are not openly made available yet.

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 [55] or identify susceptibility to tumorigenesis by exploiting genome-wide association studies [43]. 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.

Authorship contribution statement. **Anna Bernasconi**: Conceptualization, Source investigation, Data mapping, Examples, Writing - original draft; **Alberto García S.**: Conceptualization, Models design, Data mapping, Examples, Writing - review & editing; **Stefano Ceri**: Conceptualization, Writing - review & editing; **Oscar Pastor**: Conceptualization, Writing - review & editing.

Declaration of Competing Interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. This work was supported by the Valencian Innovation Agency through the OGMIOS project (INNEST/2021/57), the Generalitat Valenciana through the CoMoDiD project (CIPROM/2021/023), and the Spanish State Research Agency through the DELFOS (PDC2021-121243-I00) and SREC (PID2021-123824OB-I00) projects, MICIN/AEI/10.13039/501100011033 and co-financed with ERDF and the European Union NextGenerationEU/PRTR. S.C. is supported by the PNRR-PE-AI FAIR project funded by the NextGenerationEU program.

References

[1] 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68.

[2] Alfonsi, T., Bernasconi, A., Canakoglu, A., Masseroli, M., 2022. Genomic data integration and user-defined sample-set extraction for population variant analysis. BMC bioinformatics 23, 401.

[3] Augustyn, D.R., Wyciślik, Ł., Mrozek, D., 2021. Perspectives of using Cloud computing in integrative analysis of multi-omics data. Briefings in functional genomics 20, 198–206.
[4] Bai, J.P., Alekseyenko, A.V., Statnikov, A., Wang, I.M., Wong, P.H., 2013. Strategic applications of gene expression: from drug discovery/development to bedside. The AAPS journal 15, 427–437.
[5] Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al., 2012. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Research 41, D991–D995.
[6] Bass, J.I.F., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., Yi, S., Hill, D.E., Vidal, M., Walhout, A.J., 2015. Human gene-centered transcription factor networks for enhancers and disease variants. Cell 161, 661–673.
[7] Bernasconi, A., Canakoglu, A., Masseroli, M., Ceri, S., 2021. The road towards data integration in human genomics: players, steps and interactions. Briefings in Bioinformatics 22, 30–44. doi:10.1093/bib/bbaa080.
[8] Bernasconi, A., Canakoglu, A., Masseroli, M., Ceri, S., 2022a. META-BASE: A Novel Architecture for Large-Scale Genomic Metadata Integration. IEEE/ACM Transactions on Computational Biology and Bioinformatics 19, 543–557.
[9] Bernasconi, A., Ceri, S., Campi, A., Masseroli, M., 2017. Conceptual modeling for genomics: building an integrated repository of open data, in: International Conference on Conceptual Modeling, Springer. pp. 325–339.
[10] Bernasconi, A., García S, A., Ceri, S., Pastor, O., 2022b. A comprehensive approach for the conceptual modeling of genomic data, in: Conceptual Modeling: 41st International Conference, ER 2022, Hyderabad, India, October 17–20, 2022, Proceedings, Springer. pp. 194–208.
[11] Bornberg-Bauer, E., Paton, N.W., 2002. Conceptual data modelling for bioinformatics. Briefings in Bioinformatics 3, 166–180.
[12] Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rosati, R., 2007. Ontology-based database access., in: SEBD, pp. 324–331.
[13] Canakoglu, A., Bernasconi, A., Colombo, A., Masseroli, M., Ceri, S., 2019. GenoSurf: metadata driven semantic search system for integrated genomic datasets. Database 2019.
[14] Cappelli, E., Cumbo, F., Bernasconi, A., Canakoglu, A., Ceri, S., Masseroli, M., Weitschek, E., 2020. OpenGDC: Unifying, Modeling, Integrating Cancer Genomic Data and Clinical Metadata. Applied Sciences 10, 6367.
[15] Cappelli, E., Felici, G., Weitschek, E., 2018. Combining dna methylation and rna sequencing data of cancer for supervised knowledge extraction. BioData mining 11, 1–23.
[16] Ceri, S., Bernasconi, A., Canakoglu, A., Gulino, A., Kaitoua, A., Masseroli, M., Nanni, L., Pinoli, P., 2017. 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, Springer. pp. 46–57.
[17] Ceri, S., Gottlob, G., Tanca, L., et al., 1989. What you always wanted to know about Datalog (and never dared to ask). IEEE Transactions on Knowledge and Data Engineering 1, 146–166.
[18] Chen, H.M., MacDonald, J.A., 2021. Network analysis identifies dapk3 as a potential biomarker for lymphatic invasion and colon adenocarcinoma prognosis. IScience 24, 102831.
[19] Chen, J.Y., Carlis, J.V., 2003. Genomic data modeling. Information Systems 28, 287–310.
[20] Cornell, M., Paton, N.W., Hedeler, C., Kirby, P., Delneri, D., Hayes, A., Oliver, S.G., 2003. GIMS: an integrated data storage and analysis environment for genomic and functional data. Yeast 20, 1291–1306.
[21] Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al., 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic acids research 46, D794–D801.
[22] Do, H.H., Rahm, E., 2004. Flexible integration of molecular-biological annotation data: The GenMapper approach, in: International Conference on Extending Database Technology, Springer. pp. 811–822.
[23] Durślewicz, J., Klimaszewska-Wiśniewska, A., Antosik, P., Grzanka, D., 2023. Low expression of matr3 is associated with poor survival in clear cell renal cell carcinoma. Biomedicines 11, 326.
[24] Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30, 207–210.
[25] ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.
[26] Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al., 2021. GENCODE 2021. Nucleic acids research 49, D916–D923.
[27] Freeman, T.J., Smith, J.J., Chen, X., Washington, M.K., Roland, J.T., Means, A.L., Eschrich, S.A., Yeatman, T.J., Deane, N.G., Beauchamp, R.D., 2012. Smad4-mediated signaling inhibits intestinal neoplasia by inhibiting expression of β-catenin. Gastroenterology 142, 562–571.
[28] García, A., Palacio, A.L., Román, J.F.R., Casamayor, J.C., Pastor, O., 2020. Towards the understanding of the human genome: a holistic conceptual modeling approach. IEEE Access 8, 197111–197123.
[29] García, A., Palacio, A.L., Román, J.F.R., Casamayor, J.C., Pastor, O., 2021. A conceptual model-based approach to improve the representation and management of omics data in precision medicine. IEEE Access 9, 154071–154085.
[30] García S, A., Casamayor, J.C., Pastor, O., 2021. ISGE: A Conceptual Model-Based Method to Correctly Manage Genome Data, in: International Conference on Advanced Information Systems Engineering, Springer. pp. 47–54.
[31] Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M.G.D., Jädersten, M., Dolatshad, H., Verma, A., Cross, N.C., Vyas, P., et al., 2015. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nature communications 6, 5901.
[32] Gopinath, C., Nathar, T.J., Ghosh, A., Hickstein, D.D., Remington Nelson, E.J., 2015. Contemporary animal models for human gene therapy applications. Current gene therapy 15, 531–540.
[33] Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., Staudt, L.M., 2016. Toward a shared vision for cancer genomic data. New England Journal of Medicine 375, 1109–1112.
[34] Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R.K., Buckberry, S., Vargas-Landin, D.B., Poppe, D., Pflueger, J., Lister, R., Rackham, O.J.L., Petretto, E., Polo, J.M., . A single-cell atlas of entorhinal cortex from individuals with alzheimer's

disease reveals cell-type-specific gene expression regulation 22, 2087–2097. doi:10.1038/s41593-019-0539-4. number: 12 Publisher: Nature Publishing Group.

[35] Guerin, É., Marquet, G., Burgun, A., Loréal, O., Berti-Équille, L., Leser, U., Moussouni, F., 2005. Integrating and warehousing liver gene expression data and related biomedical resources in GEDAW, in: International Workshop on Data Integration in the Life Sciences, Springer. pp. 158–174.

[36] Ji, F., Elmasri, R., Zhang, Y., Ritesh, B., Raja, Z., 2005. Incorporating concepts for bioinformatics data modeling into EER models, in: The 3rd ACS/IEEE International Conference on Computer Systems and Applications, IEEE. pp. 189–192.

[37] Jia, P., Zhao, Z., 2017. Impacts of somatic mutations on gene expression: an association perspective. Briefings in bioinformatics 18, 413–425.

[38] Keet, C.M., 2003. Biological data and conceptual modelling methods. Journal of Conceptual Modeling 29, 1–14.

[39] Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al., 2015. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330.

[40] Li, X., Lalić, J., Baeza-Centurion, P., Dhar, R., Lehner, B., 2019. Changes in gene expression predictably shift and switch genetic interactions. Nature communications 10, 3886.

[41] Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., Ma, J., He, B., Yang, J., 2021. Evaluating dna methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. Frontiers in Cell and Developmental Biology 9, 619330.

[42] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al., 2013. The genotype-tissue expression (gtex) project. Nature genetics 45, 580–585.

[43] Mamidi, T.K.K., Wu, J., Hicks, C., 2019. Integrating germline and somatic variation information using genomic data for the discovery of biomarkers in prostate cancer. BMC cancer 19, 1–12.

[44] Masseroli, M., Canakoglu, A., Ceri, S., 2016. Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction. IEEE/ACM Transactions on Computational Biology and Bioinformatics 13, 209–219.

[45] Masseroli, M., Canakoglu, A., Pinoli, P., Kaitoua, A., Gulino, A., Horlova, O., Nanni, L., Bernasconi, A., Perna, S., Stamoulakatou, E., Ceri, S., 2018. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. Bioinformatics 35, 729–736.

[46] Masseroli, M., Kaitoua, A., Pinoli, P., Ceri, S., 2016. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. Methods 111, 3–11.

[47] Médigue, C., Rechenmann, F., Danchin, A., Viari, A., 1999. Imagene: an integrated computer environment for sequence annotation and analysis. Bioinformatics (Oxford, England) 15, 2–15.

[48] n.a., 2002. Microarray standards at last. Nature 419.

[49] Nagy, Á., Győrffy, B., 2021. mutarget: a platform linking gene expression changes and mutation status in solid tumors. International journal of cancer 148, 502–511.

[50] Nghiem, P.P., Kornegay, J.N., 2019. Gene therapies in canine models for duchenne muscular dystrophy. Human Genetics 138, 483–489.

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

[52] O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al., 2016. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic acids research 44, D733–D745.

[53] Pastor, O., Levin, A.M., Celma, M., Casamayor, J.C., Virrueta, A., Eraso, L.E., 2011. Model-based engineering applied to the interpretation of the human genome, in: The Evolution of Conceptual Modeling. Springer, pp. 306–330.

[54] Paton, N.W., Khan, S.A., Hayes, A., Moussouni, F., Brass, A., Eilbeck, K., Goble, C.A., Hubbard, S.J., Oliver, S.G., 2000. Conceptual modelling of genomic information. Bioinformatics 16, 548–557.

[55] Przytycki, P.F., Singh, M., 2017. Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes. Genome Medicine 9, 79.

[56] Rechenmann, F., 2012. Data modeling: the key to biological data integration. EMBnet. journal 18, 59–60.

[57] Román, J.F.R., Pastor, Ó., Casamayor, J.C., Valverde, F., 2016. Applying conceptual modeling to better understand the human genome, in: International Conference on Conceptual Modeling, Springer. pp. 404–412.

[58] Safran, M., Rosen, N., Twik, M., BarShir, R., Stein, T.I., Dahary, D., Fishilevich, S., Lancet, D., 2021. The genecards suite, in: Practical Guide to Life Science Databases. Springer, pp. 27–56.

[59] Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. Nature methods 5, 16–18.

[60] Settino, M., Bernasconi, A., Ceddia, G., Agapito, G., Masseroli, M., Cannataro, M., 2019. Using gmql-web for querying, downloading and integrating public with private genomic datasets, in: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 688–693.

[61] Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A.J., Bucci, G., Buetti, I., Burge, S., Cabau, C., Carlson, J.W., Chelala, C., Chrysostomou, C., Cittaro, D., Collin, O., Cordova, R., Cutts, R.J., Dassi, E., Genova, A.D., Djari, A., Esposito, A., Estrella, H., Eyras, E., Fernandez-Banet, J., Forbes, S., Free, R.C., Fujisawa, T., Gadaleta, E., Garcia-Manteiga, J.M., Goodstein, D., Gray, K., Guerra-Assunção, J.A., Haggarty, B., Han, D.J., Han, B.W., Harris, T., Harshbarger, J., Hastings, R.K., Hayes, R.D., Hoede, C., Hu, S., Hu, Z.L., Hutchins, L., Kan, Z., Kawaji, H., Keliet, A., Kerhornou, A., Kim, S., Kinsella, R., Klopp, C., Kong, L., Lawson, D., Lazarevic, D., Lee, J.H., Letellier, T., Li, C.Y., Lio, P., Liu, C.J., Luo, J., Maass, A., Mariette, J., Maurel, T., Merella, S., Mohamed, A.M., Moreews, F., Nabihoudine, I., Ndegwa, N., Noirot, C., Perez-Llamas, C., Primig, M., Quattrone, A., Quesneville, H., Rambaldi, D., Reecy, J., Riba, M., Rosanoff, S., Saddiq, A.A., Salas, E., Sallou, O., Shepherd, R., Simon, R., Sperling, L., Spooner, W., Staines, D.M., Steinbach, D., Stone, K., Stupka, E., Teague, J.W., Dayem Ullah, A.Z., Wang, J., Ware, D., Wong-Erasmus, M., Youens-Clark, K., Zadissa, A., Zhang, S.J., Kasprzyk, A., 2015. The BioMart community portal: an innovative

alternative to large, centralized data repositories. Nucleic Acids Research 43, W589–W598.

[62] Stark, Z., Dolman, L., Manolio, T.A., Ozenberger, B., Hill, S.L., Caulfied, M.J., Levy, Y., Glazer, D., Wilson, J., Lawler, M., et al., 2019. Integrating genomics into healthcare: a global responsibility. The American Journal of Human Genetics 104, 13–20.

[63] Su, L., Chen, S., Zheng, C., Wei, H., Song, X., . Meta-analysis of gene expression and identification of biological regulatory mechanisms in alzheimer's disease 13.

[64] Vural, S., Simon, R., Krushkal, J., 2018. Correlation of gene expression and associated mutation profiles of apobec3a, apobec3b, rev1, ung, and fhit with chemosensitivity of cancer cell lines to drug treatment. Human Genomics 12, 20.

[65] Wang, L., Zhang, A., Ramanathan, M., 2005. BioStar models of clinical and genomic data for biomedical data warehouse design. International Journal of Bioinformatics Research and Applications 1, 63–80.

[66] Wang, Q., Armenia, J., Zhang, C., Penson, A.V., Reznik, E., Zhang, L., Minet, T., Ochoa, A., Gross, B.E., Iacobuzio-Donahue, C.A., et al., 2018. Unifying cancer and normal rna sequencing data from different sources. Scientific data 5, 1–8.

[67] Weinstein, J.A., Zeng, X., Chien, Y.H., Quake, S.R., 2013a. Correlation of gene expression and genome mutation in single b-cells. PLoS One 8, e67624.

[68] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., et al., 2013b. The cancer genome atlas pan-cancer analysis project. Nature genetics 45, 1113–1120.

[69] Wu, P., Feng, Q., Kerchberger, V.E., Nelson, S.D., Chen, Q., Li, B., Edwards, T.L., Cox, N.J., Phillips, E.J., Stein, C.M., et al., 2022. Integrating gene expression and clinical data to identify drug repurposing candidates for hyperlipidemia and hypertension. Nature Communications 13, 46.

[70] Yang, F., Cai, S., Ling, L., Zhang, H., Tao, L., Wang, Q., 2021. Identification of a five-gene prognostic model and its potential drug repurposing in colorectal cancer based on tcga, gtex and geo databases. Frontiers in Genetics 11, 622659.