A Conceptual Approach to Using Relevant Patterns in Genomic Data Analysis

Mireia Costa[2], Alberto García S.[2], Anna Bernasconi[1,*], Stefano Ceri[1] and Oscar Pastor[2]

[2]*PROS Research Center, VRAIN Research Institute – Universitat Politècnica de València, Spain*

[1]*Department of Electronics, Information and Bioengineering – Politecnico di Milano, Italy*

Abstract

Several models have been proposed to represent human genomic information. An interesting approach for supporting genomic applications for health consists of a two-layer representation. In this approach, high-level concepts describing distinct aspects of the human genome at an abstract level are mapped to data representing actual physical measurements. This two-layer method allows users to formulate high-level queries on the concepts and map them onto real datasets. Additionally, the approach is extensible, allowing new conceptual views corresponding to specific genomic features to be mapped to the lower data layer without impacting previous mappings.

We here present how concept-layer and data-layer instances can be composed into patterns corresponding to classic genomic studies: diseases with case-control comparisons, multi-omic representations for the same patients, and comparisons within families for rare genetic diseases. We show that these patterns effectively support genomic data users (i.e., clinicians, geneticists, and bioinformaticians) in genomic analysis practices.

Keywords

Conceptual Modeling, Genomic Datasets, Genomics, Multi-level Querying, Analysis Patterns

1. Introduction

The Human Genome, with its vast complexity, presents challenges in capturing, representing, and utilizing its extensive information; consequently, the landscape of genomic data sources is wide and diverse. Commonly used databases include The Cancer Genome Atlas, a landmark cancer genomics program now embedded within Genomic Data Commons [1]; the 1000 Genomes Project [2], a catalog of common human genetic variation; GTEx [3], a resource database to study the relationship between genetic variation and gene expression in multiple reference tissues; and GEO [4], the most general and widely used among genomic repositories.

In the genomics domain, conceptual models have long been employed to effectively manage and represent extensive data, as well as to accurately depict the structure and functions of

ER2024: Companion Proceedings of the 43rd International Conference on Conceptual Modeling: ER Forum, Special Topics, Posters and Demos, October 28-31, 2024, Pittsburgh, Pennsylvania, USA

*Corresponding author.

micossan@vrain.upv.es (M. Costa); algarsi3@inf.upv.es (A. García S.); anna.bernasconi@polimi.it (A. Bernasconi); stefano.ceri@polimi.it (S. Ceri); opastor@dsic.upv.es (O. Pastor)

0000-0002-8614-0914 (M. Costa); 0000-0001-5910-4363 (A. García S.); 0000-0001-8016-5750 (A. Bernasconi); 0000-0003-0671-2415 (S. Ceri); 0000-0002-1320-8471 (O. Pastor)

the genome. Starting in the late nineties, pioneers such as Okayama et al. [5] ventured into representing DNA genomic sequences in databases. In the 2000s, Paton et al. [6] introduced data models for transcription/translation processes, alongside genomic sequences and protein structures. Subsequent works leveraged conceptual models to articulate biological entities and interactions, leading to databases like GenMapper Warehouse [7] and BioMart [8].

This background research has later motivated conceptual modeling-based approaches focusing on either characterizing the genome's structure conceptually [9] or applying it in data-driven contexts [10, 11]. Bridging these perspectives emerged as a pertinent issue, more recently addressed in [12]. In their proposal, the authors describe a novel conceptual model that merges concepts-based and data-based perspectives for genomic information modeling. Specifically, they link a *concepts-layer* delineating genome elements and their connections to a *data-layer*, detailing real-world datasets from genome sequencing. This dynamic linkage facilitates focused visualization, understanding of commonalities, and complex query expression across genomic data types, expanding the modular view-based approach to genomic data management.

This work focuses on the perspective of data users who frequently need to access and query genomic data resources. Genomic data practitioners typically perform similar types of queries repeatedly. Currently, there are systems that allow for basic data extraction using simple queries (conjunctive/disjunctive) over data. Examples of such systems include [1] for single consortia databases and [13] and [14] for integrated databases. However, while basic queries are supported, more sophisticated approaches tailored for more advanced data analysis purposes are still lacking.

We propose a *pattern-driven approach* to bridge the existing gap, facilitating more complex and specific data analysis tasks. This approach largely leverages the conceptual linking provided by the two-layer conceptual model described in [12], serving as a foundation for generating these query patterns over paramount genomic data sources. The effectiveness of this approach is demonstrated through the instantiation of query patterns that yield significant results in contemporary clinical and genetic research. These patterns include the extraction of datasets for genetic case-control studies [15, 16, 17], integrative multi-omics analyses [18, 19, 20, 21], and family trio analyses [22, 23]. Our proposal aims to offer a flexible and expandable representation of concepts, data, and their typical interconnections, providing simple query templates for guiding concept exploration, inspiring the identification of novel correspondences among data, and enhancing the findability of interoperable data instances.

In the remainder of the manuscript, Section 2 provides notions on the two-layer conceptual model; Section 3 describes our core contribution, i.e., the data analysis patterns and several example instantiations; Section 4 discusses the implications and limitations of the approach; and Section 5 concludes.

2. Two-Layer Genomic Representation

Our work takes inspiration from the holistic view presented in [12], which bridges a model of the genomic concepts and a model of the genomic datasets to facilitate genome data management through robust conceptual modeling support. More specifically, we consider a two-layer conceptual model: 1) the "concepts-layer" encapsulates human genome mechanism knowledge; and 2) the “data-layer” portrays genomic data types and experiments through structured information formats. The abstract idea can be appreciated in Figure 1: genomic information can be viewed as a dual system approached in opposite directions: connecting data to pre-existing abstract concepts (top-down) or building concepts based on available data (bottom-up).

A top-down approach initially models biological entities and then verifies data sources; this direction allows us to reveal issues with data structure definition and quality. Conversely, a bottom-up strategy starts from available data and subsequently constructs models to systematize and organize it, aiming to create user-friendly systems for domain experts.

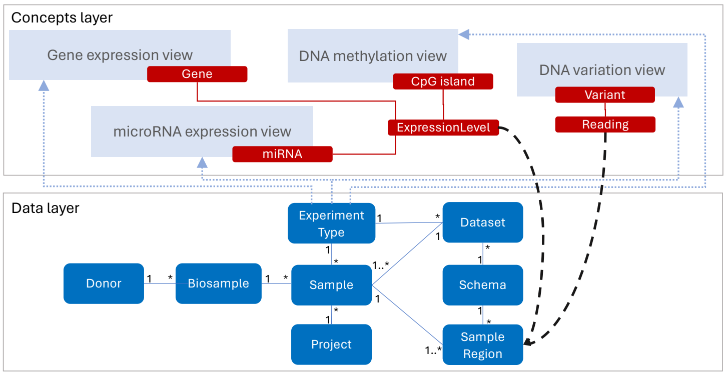

Figure 1: High-level representation of the concepts/data-layers and their connection.

The data-layer (depicted in blue in in Figure 1), centers on the Sample concept, representing a typical genomic data file, which contains a set of SampleRegions, i.e., rows in the file, which represent an interval of the genome on a specific chromosome strand, with start and end coordinates. Multiple samples are collected within Datasets, which are homogeneous in the Schema (i.e., their sample regions have the same columns and semantics) and in the ExperimentType (a description of the experimental assay run to produce the data). The experiment has been performed on biological material, which is described by the BioSample class, which belongs in turn to a Donor (an actual living patient tissue or an immortalized cell line or single cells that have undergone a sequencing process). Samples are grouped within Projects (informing on the management metadata information).

The concepts-layer has different modules (or *views*, depicted as light blue rectangles in Figure 1) describing aspects of the human genome, such as DNA variation, gene or microRNA (miRNA) expression quantification, DNA methylation levels, or any other genomic data type. To each experiment type in the data-layer, we associate a given *genomic data view* (see light blue arrows). Each view includes classes representing concepts that are measurable through genomic sequencing technologies (e.g., the expression levels of genes or the reading of a DNA variation). In Figure 1, these concepts are drawn in red; given concepts could be common to different views (e.g., the “expression level”).

The concepts-layer and the data-layer are linked through relationships between concepts (such as a variation in the DNA) and instances of data-layer classes (i.e., a specific data record). For example, a SampleRegion from a DNA-Seq experiment can be represented by its corresponding concept, a Variant spanning positions 43,044,295 to 43,170,245 on the negative strand of chromosome 17.

New links between the concepts and data-layers can be established when specific data types are selected (in the data-layer), thereby triggering the selection of specific views (of the concepts-layer). Through a classical Ontology-Based Data Access approach [24], it is possible to allow access to datasets of a specific genomic data type by specifying a query on the view of interrelated concepts.

2.1. The Concepts-Layer Model

While the data-layer is static because new genomic data types are simply another instance of the related entities, the concepts-layer is flexible and can grow according to the specific needs of a use case. Figure 2 illustrates a portion of the concepts-layer model containing classes associated with DNA variations, familial relationships, gene expression, miRNA expression, and DNA methylation.

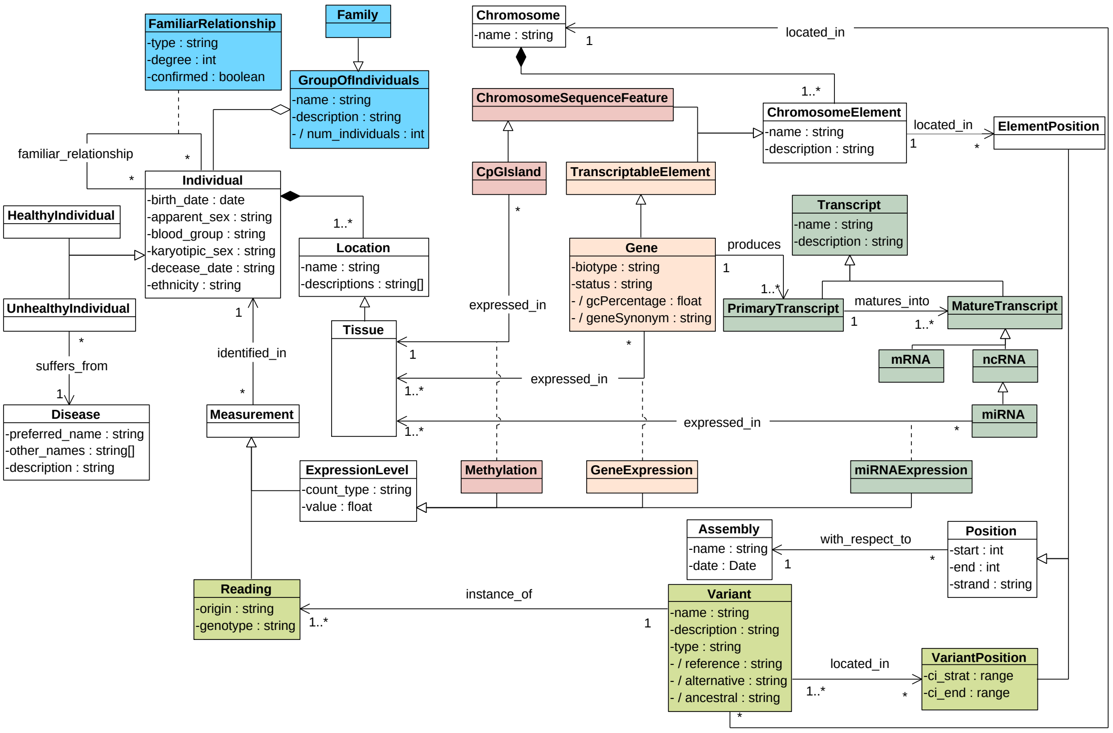

Figure 2: Excerpt of the concepts-layer model. Colors denote different groups of classes: blue (Familiar relationships-associated classes); red (DNA methylation); salmon (Gene Expression); dark green (miRNA expression); olive green (DNA Variation). Classes shared among multiple views are in white.

In this model, the Individual is the primary class, representing a person. Individuals can be classified as a HealthyIndividual or an UnhealthyIndividual based on their diagnosis of a specific Disease. It is possible to establish familial links among individuals (FamiliarRelationship); individuals aggregate in GroupOfIndividuals, such as Family. These aggregations are modeled with the purpose, for instance, of exploring the interaction between

DNA variations and diseases within families; this aspect is crucial for determining patterns of inheritance and the pathogenicity of variants. Individuals are composed of many Locations, such as Tissues.

Different Measurements are performed on individuals. Two types of measurements are captured in the model, as relevant to our portion. The Reading describes the appearance of a DNA variation (or Variant) in an individual. Variants are distinguished by a name and description, a type (substitution, insertion, or deletion), and a set of alleles (i.e., reference, alternative, and ancestral). Each variant can have multiple positions (VariantPosition), each determined according to a specific reference system, also known as Assembly. Variants are crucial in understanding the genetic basis of many diseases.

The second type of measurement is the ExpressionLevel, which is always related to a genetic component. Unlike the previously mentioned type of measurement (i.e., readings), the expression level is specific for a given Tissue, with significant differences among tissues. Three are the different expression levels considered in the excerpt:

- GeneExpression: a biological process that ensures correct *Gene*s (which are TranscriptableElements) are expressed at the right time and in appropriate amounts, enabling cells to perform their functions correctly. Gene expression measurement can help identify differentially expressed genes between normal and cancerous tissues.
- miRNAExpression: a biological process associated with biological components that regulate gene expression. miRNA expression measurements capture the levels of miRNA, a kind of non-coding RNA (ncRNA), corresponding to a MatureTranscript (as opposed to genes, which are a kind of PrimaryTranscript). Measurement of miRNA levels allows for a better understanding of cancer development and progression, providing insight into the regulatory mechanisms underlying cancer.
- DNAMethylation: a biological process altering gene expression, happening in correspondence with CpGIslands, particular featured regions in Chomosomes. Measuring DNA methylation is crucial to understanding how environmental factors affect -for instance- cancer development and progression.

3. Concepts-Data Typical Analysis Patterns

As suggested in [12], the two-layer representation –drawing direct linking between concepts and data in the genomic domain– allows:

(a) real data inspection improving its conceptual representation (e.g., by identifying cases where many different variant positions exist from chromosome elements or variants);

(b) use of abstract knowledge (i.e., concepts-layer) as an extractor of existing datasets (i.e., data-layer), for instance, by leveraging the explicit conceptual relation between positions and elements (including genes and transcripts); and

(c) formulation of inter-data type queries over data (i.e., coming from different Measurement types), by controlling the datasets' concepts that regard different genomic data types.

These aspects could be translated into simple view-driven queries, where concepts are selected in the upper layer and are translated into queries over the data. Here, we propose to make a step forward with respect to (a)–(c): we use conceptual linking as a glue for generating classical *genomic analysis patterns* that are typically used in research practice. In this section, we describe the most relevant ones, selected according to the enduring experience of the authors in the field, developed during several interdisciplinary collaborations with clinicians, biologists, and geneticists. In this work, we focus on:

(1) observational studies in which two existing groups –that differ in outcome (e.g., healthy or non-healthy)– are compared based on a supposed causal attribute (e.g., presence of a DNA mutation);

(2) biological analyses in which the datasets are multiple "omes" (e.g., the genome and the transcriptome) used to study life overlapping multiple layers; and

(3) data analyses that investigate aspects within the genetic hierarchical relationships of families (e.g., causal variations for inherited diseases).

Finally, we show how patterns can be combined in complex patterns, e.g., joining the approach described in (1) and (3); additional patterns can be built along similar lines. Next, we describe patterns one by one by exposing relevant examples in the scientific literature and showing a UML instance diagram [25] depicting the concepts-data-layers linking.

3.1. Case-Control Studies

Case-control studies constitute a commonplace method in clinical research aimed at comparing diverse genomic datasets from ill individuals (*cases*) and unaffected individuals (*controls*) to delineate genetic elements that contribute to increased susceptibility or severity to disease. Publicly accessible data repositories such as The Cancer Genome Atlas (TCGA, [1]) for cases and The GTEx Consortium [3] for controls are fundamental for increasing sample sizes and identify cases and controls in scenarios where they were previously unavailable, thereby enhancing the efficiency and robustness of genomic studies. Two types of case-control analyses are typically produced:

1. At the *population-level*, examining cases and controls using data that is specific to a particular tissue, with the purpose of investigating how diseases or phenotypic traits affect that tissue.
2. At the *patient-level*, analyzing both healthy and diseased tissue samples extracted from the same patient to determine the impact of cancer processes on a given tissue.

Population-level case-control. The advantages of population-level case-control analyses focusing on the same tissue type have been extensively described in the literature. In the concrete domain of cancer genomics, such studies typically involve comparing healthy tissues (derived from patients without cancer) of the same tissue type as those giving rise to cancer (obtained from patients with a specific cancer subtype).

For instance, Aran *et al.* [15] combined data from the TCGA and GTEx projects to analyze gene expression disparities between healthy and across eight tissues and corresponding tumor types. This approach facilitated the comprehension of tumor development and the discovery of novel biomarkers, critical for effective prevention and therapeutic stratagem selection.

In Figure 3 we present a simplified version of the scenario outlined in [15], featuring two distinct patients: one afflicted (extracted from TCGA, with ID = "TCGA-A2-A04N") and one healthy (extracted from GTEx, with ID = "074b0792-df3c-4b59-9f50-793bc14bcb81") individual.

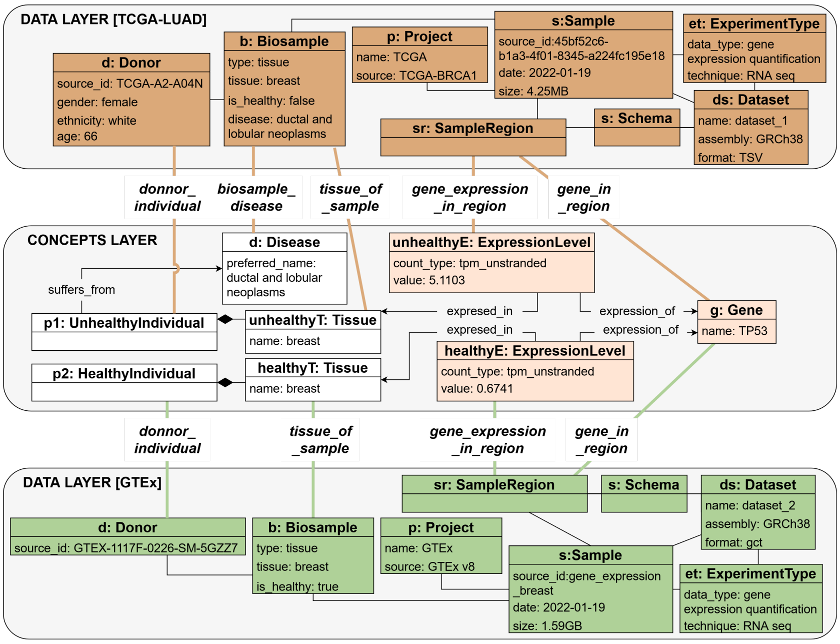

Figure 3: Selection of population-level case-control instances. Samples are extracted from TCGA (top rectangle, showing one example) selecting non-healthy donors, while one sample is extracted from GTEx representing the tissue-specific (i.e., breast tissue) gene expression of typical healthy patients.

Note that the afflicted patient has a non-healthy biosample (`is_healthy = false`) associated with the *Ductal and Lobular Neoplasm* disease.

The process of selecting cases and controls with specific conditions and from the same tissue type within such datasets is nontrivial and necessitates sophisticated instance modeling. For example, identifying patients who are "male, white, and 79 years old" within TCGA [1] is not feasible. Consequently, pairing cases and controls entail not only ontological mediation (via the concepts-layer) but also an understanding of the data sources. Our proposed approach streamlines this process by enabling a consistent representation of biological concepts and a technologically-independent data representation. The expression levels are observed on a specific gene (TP53), which is fixed at the conceptual level and therefore searched in the SampleRegions of GTEx and TCGA Samples to extract appropriate values.

Patient level case-control. Numerous studies have highlighted the clinical advantages of performing case-control analyses at the patient level. For individual patients, the analysis involves comparing samples from adjacent tumor tissue (considered as *control*) and tumor tissue (*case*). Collectively, these two types of samples are referred to as *paired samples*.

In [16], Kim *et al.* analyzed paired samples from patients with *Colon Adenocarcinoma*, showing that this type of analysis significantly impacts the prediction of cancer recurrence. Oh and Lee [17], instead, examined the differences in gene expression between paired samples in *Lung Adenocarcinoma* and *Breast Invasive Carcinoma*, among others. Using machine learning models, they concluded that such analyses can aid in predicting the prognosis of certain cancers, thus facilitating appropriate clinical treatments. Both studies employed the TCGA public resource

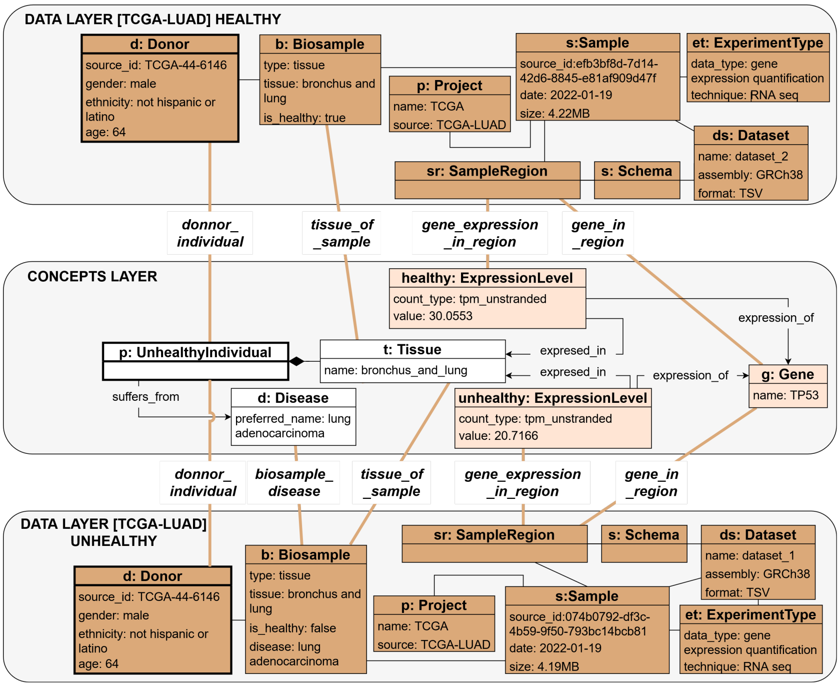

Figure 4: Selection of patient-level case-control instances. Note that the model allows for the selection of data derived from different samples (i.e., one healthy and one unhealthy) from the same patient. This extraction can be repeated for many patients to build a dataset of *paired samples*.

to obtain patient data. In Figure 4 we show the case of a patient possibly included in this patient-level case-control analysis [17]. This patient, diagnosed with Lung Adenocarcinoma (id = "TCGA-44-6146"), holds paired samples available in TCGA.

TCGA is a repository that provides information on files resulting from specific genomic analyses of healthy and tumor tissues from cancer patients. The analytical nature of TCGA makes the search for paired samples challenging, requiring advanced data processing and search functionality to identify analyses from the same patient (e.g., files associated with the same Donor). Currently, this is not allowed even by the updated TCGA major entry point [1]. Note that, in the concepts-data framework, the data-layer enables easy identification of the patient (or Donor) from whom each sample originates, while the concepts-layer facilitates the identification of both samples as belonging to the same individual.

3.2. Integrative Multi-Omics Studies

Multi-omics approaches are innovative frameworks that integrate multiple omics datasets to enhance understanding of genetic disease [26]. Particular attention is given to multi-omics studies to study cancer's molecular and clinical features. Here, areas of research include segmentation into subtypes, improvement of survival predictions and therapeutic outcomes, and uncovering key pathophysiological processes across different molecular layers. Again, we refer to the TCGA data source, while other cancer genomics could also be considered (see the

ICGC [27]). For multi-omics analysis, two types of inquiries typically hold significant interest:

1. At the *population-level*, analyzing a specific disease or phenotypic trait, by using genomic samples that refer to different genomic data types (i.e., features in the genome).
2. At the *patient-level*, analyzing a specific disease or phenotypic trait, considering specific patients whose samples have been analyzed according to multiple genomic data tests (i.e., for whom multiple data types are available).

Population-level multi-omics. Associating variation signatures or gene expression/methylation/miRNA profiles with diagnostic/prognostic values is of high importance in cancer research. Pinoli *et al.* [18] examine the rich presence of variants, abnormal methylation levels, as well as copy number alteration events, in the proximity of specific topological structures for 26 cancer types. Mehrgou and Teimourian [19] utilize gene expression, methylation, and miRNA datasets from both TCGA and GEO sources to derive insights on *Colorectal cancer*, with applications in diagnosis, prognosis, and targeted therapy.

Patient-level multi-omics. Focusing on specific tissues, we aim to find patients whose biological samples have been analyzed using different genomic experiments (i.e., for whom multiple data types are available). Grouping data by the same patient enables building richer disease models. We call this pattern *one-to-one linking* and to the multiple samples derived from the same patient as *linked multi-omics samples* (connecting mutations, expression, and epigenomic signals such as methylation levels).

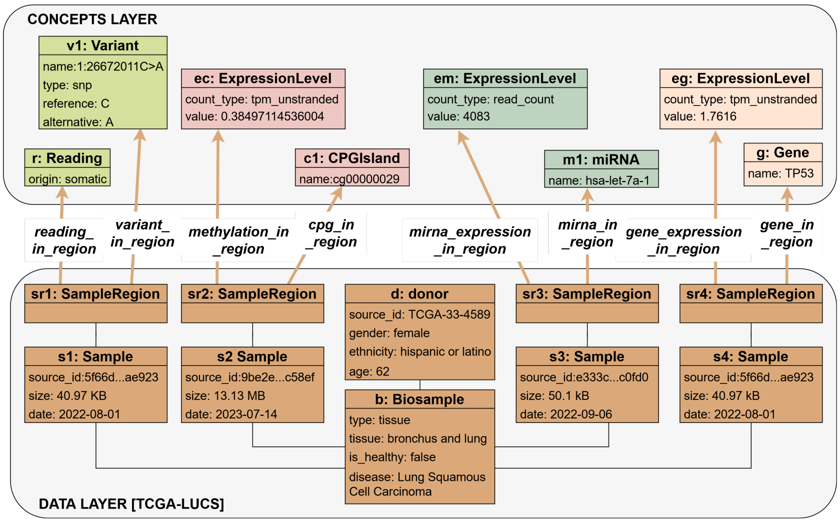

Figure 5: Selection of patient-level multi-omics instances. Several samples can be extracted from TCGA to represent different genomic data types (i.e., DNA variation, DNA methylation, miRNA expression, gene expression). It can be repeated for many patients to build a dataset of *linked multi-omics samples*.

Figure 5 illustrates a patient (ID = "TCGA-33-4589") with lung adenocarcinoma for whom data on variants, methylation levels, miRNA, and gene expression are available in TCGA. The comprehensive data of this patient facilitates multiple analyses with significant clinical applications. For instance, in [20], miRNA and gene expression data were analyzed in patients with lung adenocarcinoma to classify patients based on survival. This has crucial implications for cancer prognosis, enabling the identification of patients who may require more intensive monitoring due to a poor prognosis. Similar studies have been conducted for survival prediction in breast cancer, utilizing gene and miRNA expression, DNA methylation, and CNV data [21].

3.3. Family Trio Analyses

Rare disorders are conditions with a low frequency in the population and often have a genetic component. Despite the significance of genetics in these disorders, most patients remain undiagnosed after standard genetic testing [8]. Family trio analysis involves comparing the genetic information of the patient with that of their parents. Consequently, it is possible to identify *de novo variations*, i.e., DNA variations unique to the patient and not inherited from either parent. This kind of analysis has been shown to positively impact rare disease contexts, by improving diagnostic and serving as a powerful tool in identifying disorder-causing variations [23].

Figure 6 illustrates information about a family trio reported in [22]. In this study, the authors examine family trios in the context of *Amyotrophic Lateral Sclerosis* (ALS) to identify risk factor variants associated with this devastating disease. They identified several *de novo* variations, such as the v1 instance of the Variant class in Figure 6. This variant is considered *de novo* because it was identified only in the affected individual (son instance of the UnhealthyIndividual class) and not in either parent (father and mother instances in the concepts-layer). The identified variants helped the authors improve the understanding of the genetic role in ALS.

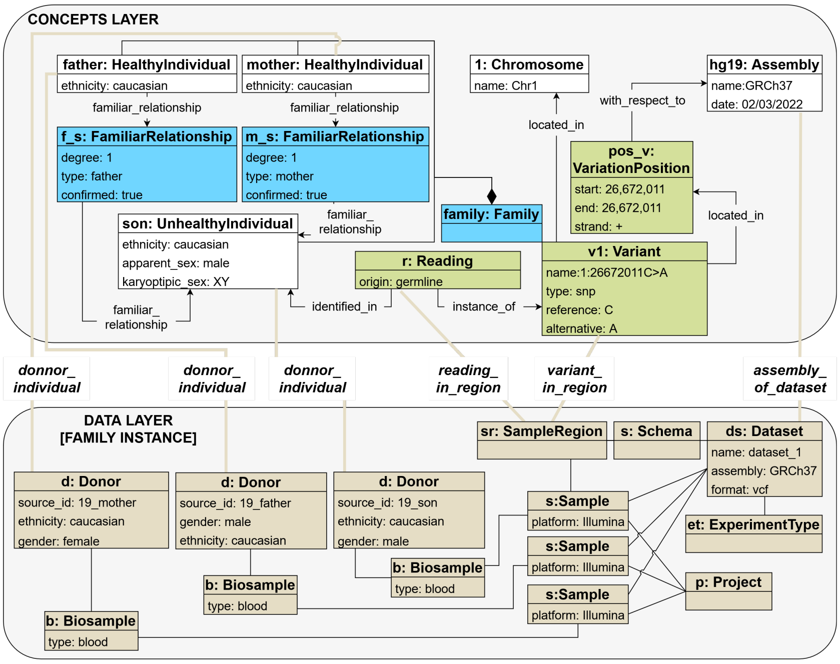

Figure 6: Selection of samples of patients related in a family-trio pattern, using the information presented in the concepts-layer. This process can be repeated for many families to build a comprehensive dataset of *family-trio* samples.

In this particular pattern, the data-layer represents fundamental information about the experiment, such as the sequencing technology, which cannot be represented in the concepts-layer. On the other hand, the concepts-layer allows us to infer that the variant shown in Figure 6 is de novo, as it captures the familial relationships between individuals. The ontological connection between both models provides a holistic representation of all the relevant information needed for family trio analysis.

An important genomic data source, the 1000 Genomes Project, collected a huge dataset intending to identify all the genetic variants with frequencies of at least 1% in several worldwide populations. The last release of the project covered 26 populations and observed single nucleotide variants (SNVs) and insertions/deletions (indels) from different 602 parent/child trios produced within the project [28]. The pattern described in Figure 6 can be reproduced on 1000 Genomes data to perform a database-wide analysis on *family-trio* samples.

3.4. Complex Patterns

Above, we have demonstrated traditional data analysis patterns in genomics. However, more complex analysis patterns have gained attention.

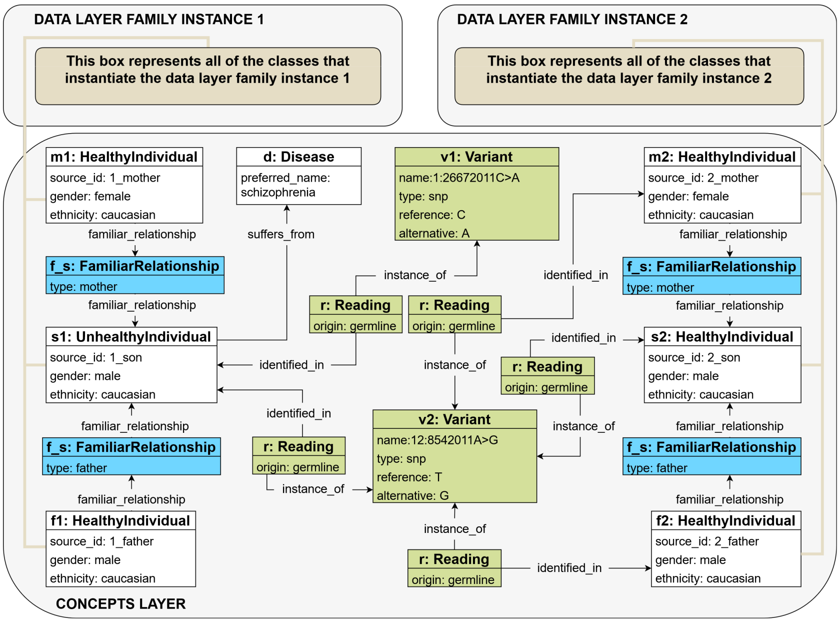

Figure 7: Selection of two family-trio patterns to build a complex family-trio case-control study. Each family instance is condensed into a box in the data-layer. Analysis of Family 1's data-layer reveals that the affected son carries two candidate variants potentially linked to the observed disease. In contrast, Family 2's data-layer depicts a healthy family member carrying one of the candidate variants. Utilizing Family 1 as the *case* and Family 2 as the *control*, the complex pattern represented in the concepts-layer facilitates the identification and exclusion of one of the two candidate variants (i.e., v2).

One example of a complex pattern involves family-trio case-control analysis, which corresponds to performing case-control analyses (see Section 3.1) on family trios (see Section

3.3). Specifically, it compares the genetic information of families with an affected individual to families with no affected individuals. In [29], this strategy was employed to identify genes and mutation types that are highly associated with Schizophrenia. Figure 7 illustrates that scenario, by using only two families (for simplification purposes). Here, it can be observed that the v2 VARIANT appears in the offspring of both families; this insight can be used to rule out the association of this variation with schizophrenia (as s1 is an UNHEALTHYINDIVIDUAL, whereas s2 is a HEALTHYINDIVIDUAL). Conversely, the v1 VARIANT appears only in the family member s1, who is affected by schizophrenia; however, it does not appear in any of her/his parents, which offers strong evidence of the potential relationship between variant v1 and schizophrenia.

Another example of a complex pattern is multi-omics case-control analyses. Here, experts compare different data types from patients with a certain characteristic (cases) to those without it (controls) to determine if there is a clinically relevant relationship between any omic feature and the characteristic under study. For instance, in [30], the authors used multi-omic data to predict the risk of developing asthma, and in [31], they employed this type of analysis to predict the development of preeclampsia.

4. Discussion

The two-layer framework described in Section 2 allows us to incorporate new concepts and relationships according to a data-agnostic approach. Indeed, as acquired knowledge in genomics is constantly evolving, new concepts will be added and changes will recur in the concepts-layer. The model will not remain the same as the one presented in Figure 2, which -in turn-extends previous work [12]. Conversely, the data-layer is typically not impacted by genomic concepts' changes as long as all genomic data can be represented as SAMPLES containing *SampleRegions*. Even when experts' understanding of genomic-related knowledge mutates, possibly impacting the interpretation of data analysis results, data keeps the same model; this favors the maintainability of potential data mappings, processing pipelines, and bio-tools that leverage this representation.

Here, we show that a strong connection between data and concepts compensates for the limitations of approaches that consider the layers separately, allowing a holistic representation of the genomic domain. The identification of interesting patterns of analysis and the consequent reasoning can only be explained by using an interactive two-layer representation. Our rationale is to use the conceptual linking as a glue for generating classical patterns of case-controls, multi-omics, or family-trios by having the conceptuals-layer model in the middle and instantiating the data-layer model as many times as needed. At the patient-level, case-controls typically have two-instance-replication (see Figure 4). Instead, multi-omics have many-instance-replication; we showed four, in the example of Figure 5, replicating the data model for four *genomic data types* thereby creating one copy for each sample of a same patient. In this way, we let classical genomic data analysis patterns emerge, where we "pivot" upon ontological knowledge (concepts-layer) as the mediator across several instantiations of the data-layer, playing clearly identified roles. We demostrated the capability to perform queries with high complexity, which facilitates the extraction of relevant data from highly-heterogeneus disorganized repositories and the advancement of data exploitation in the domain.

5. Conclusion

In this paper, we explain five basic patterns and then hint at how they can be composed to form more complex patterns. This framework will allow easy extension to novel query patterns that will go along with the rapidly-evolving state of genomic knowledge In current practice, domain experts typically navigate genomic data source interfaces and download data without a clear formalization of the underlying concepts and their semantic relationships. In this work, we describe a conceptual modeling-based framework that enables a unified querying strategy. Building on this, we envision a next-generation genomic data query builder that, starting from high-level concepts, allows users to execute abstract queries. This approach will relieve practitioners from the complexities of data formats and heterogeneity, enabling them to seamlessly formulate data extractions that align more closely with the classical problem formulations they are familiar with. The patterns presented in this work demonstrate initial prototypes of modular queries that can be implemented in such a system. Prospectively, this data query builder will be the main component of a visual model-driven query system for practitioners, where the conceptual model in the concepts layer is used to identify data instances in the underlying data layer.

Acknowledgements. This work is supported in part by CIPROM/2021/023), PDC2021-121243-I00), PID2021-123824OB-I00, MICIN/AEI/10.13039/501100011033, and ACIF/2021/117 grants.

References

[1] R. L. Grossman, et al., Toward a shared vision for cancer genomic data, New Engl J Med 375 (2016) 1109–1112.

[2] 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (2015) 68.

[3] J. Lonsdale, et al., The genotype-tissue expression (gtex) project, Nat Genet 45 (2013) 580–585.

[4] T. Barrett, et al., NCBI GEO: archive for functional genomics data sets–update, Nucleic Acids Res 41 (2012) D991–D995.

[5] T. Okayama, et al., Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library, Bioinformatics 14 (1998) 472–478.

[6] N. W. Paton, et al., Conceptual modelling of genomic information, Bioinformatics 16 (2000) 548–557.

[7] H.-H. Do, E. Rahm, Flexible integration of molecular-biological annotation data: The GenMapper approach, in: International Conference on Extending Database Technology, Springer, 2004, pp. 811–822.

[8] D. Smedley, et al., The BioMart community portal: an innovative alternative to large, centralized data repositories, Nucleic Acids Res 43 (2015) W589–W598.

[9] A. García, et al., Towards the understanding of the human genome: a holistic conceptual modeling approach, IEEE Access 8 (2020) 197111–197123.

[10] A. Bernasconi, et al., Conceptual modeling for genomics: building an integrated repository of open data, in: Int. Conference on Conceptual Modeling, Springer, 2017, pp. 325–339.

[11] S. Gundersen, et al., Recommendations for the fairification of genomic track metadata, F1000Research 10 (2021).

[12] A. Bernasconi, et al., PoliViews: A comprehensive and modular approach to the conceptual modeling of genomic data, Data Knowl Eng 147 (2023) 102201.

[13] F. Albrecht, et al., DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets, Nucleic Acids Res 44 (2016) W581–W586.

[14] A. Canakoglu, et al., GenoSurf: metadata driven semantic search system for integrated genomic datasets, Database-Oxford 2019 (2019).

[15] D. Aran, et al., Comprehensive analysis of normal adjacent to tumor transcriptomes, Nat Commun 8 (2017).

[16] J. Kim, et al., Transcriptomes of the tumor-adjacent normal tissues are more informative than tumors in predicting recurrence in colorectal cancer patients, J Transl Med 21 (2023) 209.

[17] E. Oh, H. Lee, Transcriptomic data in tumor-adjacent normal tissues harbor prognostic information on multiple cancer types, Cancer Med 12 (2023) 11960–11970.

[18] P. Pinoli, et al., Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries, PloS one 15 (2020) e0227180.

[19] A. Mehrgou, S. Teimourian, Update of gene expression/methylation and mirna profiling in colorectal cancer; application in diagnosis, prognosis, and targeted therapy, Plos one 17 (2022) e0265527.

[20] K. Asada, et al., Uncovering prognosis-related genes and pathways by multi-omics analysis in lung cancer, Biomolecules 10 (2020) 524.

[21] L. Tong, et al., Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis, BMC Med Inform Decis 20 (2020) 225.

[22] A. Chesi, et al., Exome sequencing to identify de novo mutations in sporadic als trios, Nat Neurosci 16 (2013).

[23] M. Mousa, et al., Whole-exome sequencing in family trios reveals de novo mutations associated with type 1 diabetes mellitus, Biology 12 (2023) 413.

[24] D. Calvanese, et al., Ontology-based database access, in: SEBD, 2007, pp. 324–331.

[25] G. Booch, et al., The Unified Modeling Language User Guide, Addison-Wesley, Reading, MA, 1999.

[26] S. Graw, et al., Multi-omics data integration considerations and study design for biological systems and disease, Mol Omics 17 (2020).

[27] J. Zhang, et al., The international cancer genome consortium data portal, Nat Biotechnol 37 (2019) 367–369.

[28] M. Byrska-Bishop, et al., High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios, Cell 185 (2022) 3426–3440.

[29] B. Xu, et al., De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia, Nat Genet 44 (2012) 1365–1369.

[30] X.-W. Wang, et al., Benchmarking omics-based prediction of asthma development in children, Respiratory Research 24 (2023) 63.

[31] A. Rahnavard, et al., Molecular epidemiology of pregnancy using omics data: advances, success stories, and challenges, Journal of Translational Medicine 22 (2024).