Extension of the Genomic Conceptual Model to integrate Genome-Wide Association Studies

Federico Comolli

Department of Electronics, Information and Bioengineering
Politecnico di Milano, Milan, Italy

Abstract. The first human genome has been sequenced at the turn of the year 2000. Since then, modern biology has made great progresses, also thanks to the introduction of Next-generation sequencing in the mid-2000s. The growing availability of genomic data led to the birth of tertiary analysis, concerning sense-making and extraction of useful biological information. To deal with data heterogeneity, in the last decade many tools have been introduced to achieve genomic data integration: among them, the Genomic Conceptual Model (GCM) and the META-BASE architecture. The latter one allows to map data from many projects into the GCM through an integration pipeline.
In this work, we proposed an extension of the GCM to integrate two additional sources into the META-BASE architecture, namely: GWAS Catalog (curated by the NHGRI and EBI institutes) and FinnGen (curated by the University of Helsinki). These two sources host Genome-Wide Association Studies (GWAS), useful for explaining the connection between genome variations of single nucleotides and particular traits. They are organized according to different data models but share the same data semantics. As a result of our integration efforts, we enable the interoperable use and querying of GWAS datasets with several other genomic datasets (including TCGA, ENCODE, Roadmap Epigenomics, 1000 Genomes Project, and GENCODE).

Keywords: Data integration · Genomic Datasets · Bioinformatics · Multiomics Studies · GWAS

1 Introduction

Since the mid-2000s, thanks to the introduction of Next-generation sequencing [17], a whole human DNA sequence can be read in a short time and in a cheap way. After being sequenced, the so-called *tertiary analysis* [16] is performed, dealing with sense-making of the huge amount of data produced by the previous analysis. Big amount of data collected by different consortia need to be integrated to allow scientists to extract information useful to understand how life is orchestrated by the DNA and how the sequence of nucleotides affects diseases or phenotype. Data produced in the context of different projects have different formats, resulting into an obstacle for data interoperability required by the tertiary analysis.

A big effort to cope with genomic data heterogeneity has been performed by the GeCo project of Politecnico di Milano developing a conceptual model (the Genomic Conceptual Model [2]), a query language (the GenoMetric Query Language [10]) and a pipeline to integrate genomic data from multiple sources (the META-BASE architecture [1]). One of the purposes of the GeCo project has been to create an integrated genomic repository that collects data from major consortia around the world (e.g., 1000 Genomes [19], Cistrome [24], ENCODE [20], GENCODE [5], Roadmap Epigenomics [8], and TCGA [23]).

In this article, we presented the modelling efforts spent to integrate into the META-BASE architecture a new class of studies called Genome-Wide Association Studies (GWAS). They involve testing genetic variants across the genomes of many individuals to identify genotype-phenotype associations. By comparing groups of people affected by a disease or trait (cases) and without it (controls), the outcomes of these studies comprise the more frequent nucleotides in the cases group against the controls. The difference of GWAS from other studies is the focus of the analysis: single nucleotide polymorphisms (SNPs) for GWAS, whole portions of genome or DNA features for other omics studies. GWAS have revolutionized the field of complex disease genetics over the past decade, providing numerous compelling associations for human complex traits and diseases [18]. Examples of GWAS are [21], which identified 103 SNPs associated to "schizophrenia", and [7], which found 333 SNPs associated to "multiple sclerosis". This work focuses on two GWAS repositories, namely the GWAS Catalog [3] and the FinnGen Project [4].

Integrating multiple omic repositories (i.e., genomic, proteomic and transcriptomic) into the GCM serves to improve the knowledge about the molecular function and disease etiology. Multi-omic studies combine different biological entities to find novel associations between them, paving the road for disease treatments and prevention. The work in this paper focuses on the integration made on metadata describing experiments, rather than on the genomic region attributes, whose transformation is trivial (see [9]).

2 Related Works

Due to the ongoing increase of genomic data, the management techniques for large data can be applied to address the heterogeneity and complexity of the biological field. Many works exploit the conceptual modeling to capture the diverse biological objects and to interpret their relationships (see [22,12,13,15,11,14]). The objective of the cited works is to support biologists to extract insights from raw genomic data. The Genomic Conceptual Model [2], whose extension is introduced in this article, goes further the description and data organization of complex biological integrated repositories; it is an architecture driving the integration of new genomic repositories. The work presented in this article exemplifies how the architecture can be exploited to integrate new datasets, mapping them to a shared conceptual model.

3 Background

A shared conceptual schema, the Genomic Conceptual Model (GCM), has been introduced to describe semantically heterogeneous data. Multiple genomic sources have been mapped over the GCM following the META-BASE pipeline, by extracting and transforming the source-specific metadata.

The GCM is an entity-relationship model used to gather metadata of heterogeneous genomic data sources. It is organized as a star-schema centered on the `Item` entity from which depart four sub-schemata (or views), recalling a classic star-schema organization that is typical of data warehouses; they respectively describe biological, technological, management and extraction aspects (more thoroughly described in [2]):

- *Central Entity*: it represents an elementary experimental file of genomic regions and their attributes. Files are typically used by biologists for data extraction, analysis and visualization operations.
- *Biological View*: it consists of the chain of entities `Item`-`Replicate`-`Biosample`-`Donor`, representing the biological elements that contribute to the `Item` production. The `Donor` represents an individual (characterized by Age, Gender and Ethnicity) or strain of a specific organism (Species) from which the biological material was derived or the cell line was established.
- *Technology View*: it describes the technology used to produce the data `Item`. An `Item` is associated by means of a one-to-many relationship with a given `ExperimentType`.
- *Management View*: it consists of the chain of entities `Item`-`Case`-`Project` describing the organizational process for the production of items.
- *Extraction View*: it includes the entity `Dataset`, used to describe common properties of homogeneous items.

4 Data Design

Samples of the typical integrated sources are assigned to single individuals; for each biological sample we can retrieve the information about the donor(s) who provided it. Instead, Genome-Wide Association Studies are based on *cohorts of patients*, so the considered granularity is coarser w.r.t. already integrated datasets. For each GWAS sample we know the cohort size and limited ancestral information, while detailed information about each single component of the cohort is not available. For this reason, in order to include the two sources GWAS Catalog and FinnGen into the META-BASE repository, we have extended the GCM introducing the *GWAS View*, to meet the constraints of the considered class of studies.

Figure 1 illustrates the extended GCM with two added entities, i.e., `Cohort` and `Ancestry`, belonging to the *GWAS View*. In the following, we describe this novel view for including GWAS samples.

Entity Item. It is the central entity of the GCM and it is shared between all its views. A GWAS `Item` contains all the SNPs associated with the phenotype under

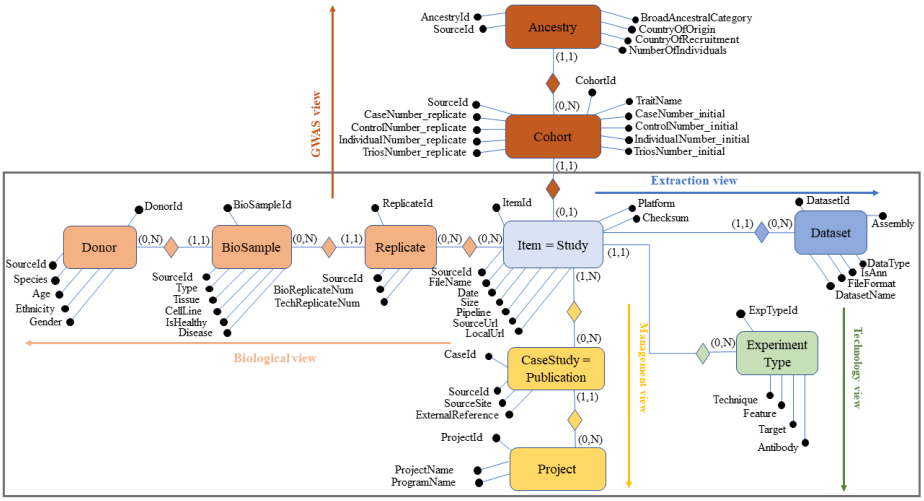


Fig. 1. Extended Genomic Conceptual Model. With respect to its original version (enclosed in the rectangle), it contains a new view (the *GWAS View*). When it gathers GWAS data, the *Biological View* remains empty (optional relationship between `Item` and `Replicate`); on the opposite when receiving different classes of studies, the *GWAS View* stays empty (optional relationship between `Item` and `Cohort`).

consideration. It contains metadata useful to describe how the corresponding region file (list of SNPs) is produced.

Entity Cohort. Each `Item` of the GCM has its corresponding `Cohort` which includes the information about the groups of people from which the biological sample is collected. An `Item` is obtained by comparing the DNA sequences of the cases (people affected by the phenotype) against the controls (people not showing that phenotype). Moreover, a sample can have one initial stage and one or more replicate stages. Some GWA studies, besides cases and controls, can be based upon groups of individuals or trios. The entity `Cohort` holds the cardinalities of the cases, controls, individuals or trios that provide the corresponding `Item`, both of the initial stage or replicate stage(s). Among all its metadata, the attribute "TraitName" is most relevant as GWA studies are driven by a phenotype (or trait, endpoint, disease).

Entity Ancestry. A `Cohort` can be partitioned into many `Ancestries`, each one containing given ancestral information about the current partition. The country of origin, the ancestral category or the country from which the participants are selected are possible available information about a partition.

Figure 2 shows how the source-specific metadata of GWAS Catalog and FinnGen are mapped into the attributes of the GCM: in the left column we list the attributes of the GCM, in the central column we report the attributes of

GCM	GWAS Catalog	FinnGen
	Ancestry	
broad_ancestral_category	broad_ancestral_category_X	--
country_of_origin	country_of_origin_X	--
country_of_recruitment	country_of_recruitment_X	"Finland"
number_of_individuals	number_of_individuals_X	n_cases + n_controls
ancestry_source_id	study_accession	phenocode
	Cohort	
trait_name	mapped_trait	name
case_number_initial	initial_sample_description [MANUAL]	n_cases
control_number_initial	initial_sample_description [MANUAL]	n_controls
individual_number_initial	initial_sample_description [MANUAL]	--
triosNumber_initial	initial_sample_description [MANUAL]	--
case_number_replicate	replication_sample_description [MANUAL]	--
control_number_replicate	replication_sample_description [MANUAL]	--
individual_number_replicate	replication_sample_description [MANUAL]	--
trios_number_replicate	replication_sample_description [MANUAL]	--
cohort_source_id	study_accession	phenocode
	Dataset	
name	dataset_name	dataset_name
data_type	"gwas"	"gwas"
format	"gdm"	"gdm"
assembly	"GRCh38"	"GRCh38"
is_annotation	"false"	"false"
	Item	
item_source_id	study_accession	phenocode
size	manually_curated__origin_file_size	manually_curated__local_file_size
date	manually_curated__origin_last_modified_date	manually_curated__download_date
checksum	manually_curated__origin_md5	manually_curated__local_md5
platform	platform__snps_passing_qc_	--
file_name	study_accession + "gdm"	phenocode + "gdm"
	Experiment_type	
technique	genotyping_technology	"FinnGen_technique"
	Case_study	
case_source_id	pubmedid	phenocode
source_site	study	"https://www.finngen.fi/en"
path_https	link	externalRef
	Project	
program_name	"Gwas Catalog"	"FinnGen"
project_name	"Gwas Catalog"	"FinnGen"

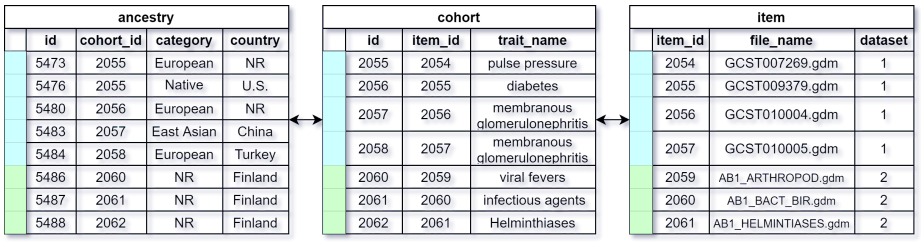
Fig. 2. Attribute mapping from source-specific metadata to Genomic Conceptual Model. In this figure are not reported the items of the *Biological View* since it has no corresponding GWAS metadata.

GWAS Catalog, while in the right column we provide the attributes of FinnGen. The metadata enclosed by quotes are meant to be treated as values to fill the corresponding GCM attributes. Let us consider the metadata *assembly*; for both the two sources the value "GRCh38" is manually inserted since it is not contained in any of the raw attributes.

Some GCM attributes are obtained by the concatenation of two raw metadata (e.g., *file_name* = *study_accession* + "gdm") or by their sum, if they are numeric (e.g., *number_of_individuals* = *n_cases* + *n_controls*).

The progressive numbers nearby the attributes describing the ancestries of GWAS Catalog refer to the multiple ancestries linked to a single cohort. Let us consider the instance of the GCM proposed in Figure 3; the item with accession "GCST007269" is linked to the cohort with id "2055", which is linked to two different ancestries (respectively ids "5473" and "5476"). The attributes of these

two ancestries, before being mapped, are referred with two different progressive numbers.



ancestry

id	cohort_id	category	country
5473	2055	European	NR
5476	2055	Native	U.S.
5480	2056	European	NR
5483	2057	East Asian	China
5484	2058	European	Turkey
5486	2060	NR	Finland
5487	2061	NR	Finland
5488	2062	NR	Finland

cohort

id	item_id	trait_name
2055	2054	pulse pressure
2056	2055	diabetes
2057	2056	membranous glomerulonephritis
2058	2057	membranous glomerulonephritis
2060	2059	viral fevers
2061	2060	infectious agents
2062	2061	Helminthiases

item

item_id	file_name	dataset
2054	GCST007269.gdm	1
2055	GCST009379.gdm	1
2056	GCST010004.gdm	1
2057	GCST010005.gdm	1
2059	AB1_ARTHROPOD.gdm	2
2060	AB1_BACT_BIR.gdm	2
2061	AB1_HELMINTIASES.gdm	2

Fig. 3. An instance of the Genomic Conceptual Model containing four items from GWAS Catalog (light blue) and three items from FinnGen (green). In this figure are reported only the entities corresponding to the *GWAS View*. Moreover, are reported only the relevant attributes to show the proper cardinalities between the entities, the full list of the attributes is reported in Figure 2.

The attributes marked with the label "MANUAL" are not reported as they are but need a syntactic transformation. The metadata `initial_sample_description` and `replication_sample_description` are written in plain text: their values are parsed to fill the GCM attributes of the cohort. Here we report the example item extracted from study accession "GCST005538": `initial_sample_description` = 1,726 European ancestry cases, 5,482 European ancestry controls; `replication_sample_description` = 1,912 European ancestry cases, 5,938 European ancestry controls, 781 African American cases, 876 African American controls. As a result of our transformation, the attributes of the corresponding cohort become: CaseNumber_initial = 1726, ControlNumber_initial = 5,482, CaseNumber_replicate = 1,912 + 781, ControlNumber_replicate = 5,938 + 876.

5 The case of "traitName"

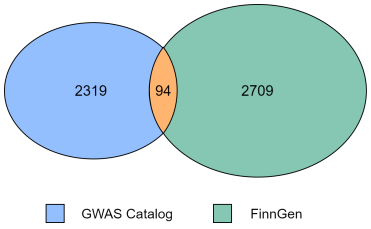
GWAS studies follow the phenotype-first approach: the participants of these studies are classified according to their clinical manifestations. The feature of GWAS studies is to search for SNPs given a phenotype. This is the reason why it is interesting to understand the set of phenotypes present in both the sources considered in this work.

All traits in GWAS Catalog are mapped over the EFO ontology [6]. Traits in the GWAS Catalog are highly diverse and include diseases (e.g., Type II diabetes), disease markers (e.g., measurements of blood glucose concentration), and non-clinical phenotypes (e.g., hair color). The Experimental Factor Ontology was chosen as the ontology to represent GWAS Catalog traits as it is highly adaptable and extensible. It is freely available in OWL format from the EFO

website and can be browsed in the Ontology Lookup Service. At the moment of writing (July 2021), the GWAS Catalog contains 2,413 different traits from the EFO ontology. Each study is characterized by one or more traits contained into the source-specific attribute "mapped_trait", comma-separated.

FinnGen phenotypes are instead harmonized over the International Classification of Diseases (ICD) revisions 8, 9 and 10, cancer-specific ICD-O-3, (NOMESCO) procedure codes, Finnish-specific Social Insurance Institute (KELA) drug reimbursement codes and ATC-codes [4]. The latest release at the moment of writing (July 2021) is the fifth one. In its manifest all the files available are listed, each with its corresponding phenotype. The fifth release contains 2803 different phenotypes.

Applying exact string matching between the list of phenotypes of the two sources, 94 traits are found to be shared. More sophisticated semantic matches are subject to future extension of this work.



GWAS Catalog	
1	schizoaffective disorder-bipolar type
2	schizophrenia
3	treatment refractory schizophrenia

FinnGen	
1	schizophrenia, schizotypal and delusional disorders
2	schizoaffective disorder
3	schizotypal disorder
4	schizoid personality disorder
5	schizophrenia
6	schizophrenia or delusion
7	schizophrenia or delusion (more controls excluded)

Fig. 4. Left: Intersection of the sets of phenotypes of the two sources GWAS Catalog and FinnGen. The intersection is obtained through exact matching of the two sets and it represents a small portion of both of them. Right: Traits related to "schizophrenia". The blue table reports the phenotypes of GWAS Catalog; the green table is dedicated to FinnGen. Only one trait is shared, all the others require domain experts to be correctly mapped.

A graphical representation of the intersection of the sets of phenotypes (with exact matching) is provided in Figure 4. In the same figure, we report an example of the matching between the phenotypes of the two repositories. In both tables we report all the phenotypes resulting by searching for the word "schizophrenia". Only one common phenotype is spotted using exact matching; more correspondences may be found with semantic match (note that mapping phenotypes requires further effort and experts validation).

6 Datasets interoperability

Using the GenoMetric Query Language (GMQL) [10] many genomic datasets belonging to the GeCo repository can be jointly queried based on the values of some corresponding attributes. The GMQL operators exploit the transformed source-specific attributes and not the original ones, to ease the matching between

attributes. The most significant attributes of GWA studies are the names of the phenotypes (attribute *traitName* of the cohort) and the ancestral information of the cohort upon which the studies are based on (attributes *countryOfOrigin* and *countryOfRecruitment* of the ancestry).

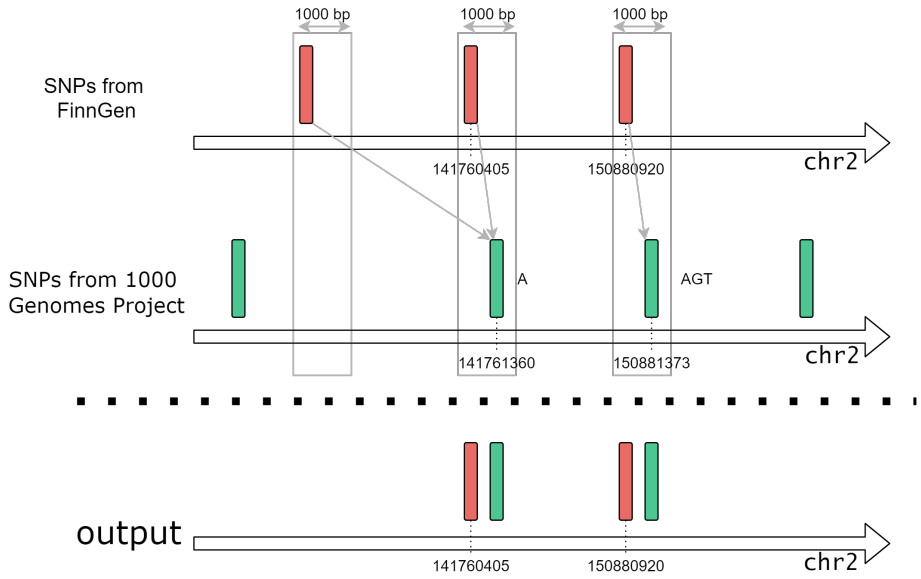


Fig. 5. GMQL query exploiting the integrated datasets FinnGen and 1000 Genomes Project. The query merges the two datasets to extract the couples made of SNPs that are closer than 1000 base pairs. It has been enabled by the META-BASE pipeline and the shared GCM.

Figure 5 graphically represents a query that jointly exploits the FinnGen dataset and the data from 1000 Genomes Project. The genomes sequenced in the 1000 Genomes Projects are not selected with regard to phenotype, so to provide a resource of variants that supports a deeper understanding of newly discovered loci influencing human disease. The projects include SNPs with allele frequencies as low as 1% across the genome and 0.1-0.5% in gene regions, as well as structural variants. It includes genomes from 26 different populations, including the Finnish one.

In the considered query, the SNPs related to the Finnish population are selected from 1000 Genomes, thereby enabling comparisons with the SNPs from FinnGen dataset. From this latter dataset we select only the SNPs which have been found as related to schizophrenia. The goal of the query is to analyze which SNPs of the two datasets are particularly close, specifically within an interval of 1000 base pairs. In the example, only two pairs of SNPs are eventually extracted in the output, as they meet the set distal constraint. Note that the querying

of the two initial datasets was only enabled thanks to the integration efforts of metadata and region attributes introduced in this article and building up on [2,1].

7 Conclusions

GWAS studies are important because they allow to find numerous compelling associations for human complex traits and diseases. Once such genetic markers are identified, they can be used to understand how genes contribute to the diseases and to develop better prevention and treatment strategies. These associations have led to insights into the architecture of disease susceptibility (through the identification of novel disease-causing genes and mechanisms) and to advances in clinical care (for example, the identification of new drug targets and disease biomarkers) and personalized medicine (for example, risk prediction and optimization of therapies based on genotype).

The integration of GWAS with other classes of genomic data is fundamental to reach interoperability and answering complex biological questions. The exact interpretation of the SNPs found in GWAS is not trivial for at least two reasons. First, the outputs of GWAS are often large clusters of SNPs in linkage disequilibrium, making it difficult to distinguish causal SNPs from neutral variants in linkage. Second, even assuming the causal variants can be identified, interpretation is limited by incomplete knowledge of non-coding regulatory elements, their mechanisms of action and the cellular states and processes in which they function. For the aforementioned reasons, it is important to further investigate GWAS data by merging different genomic datasets and by performing multi-omic analyses.

We started from the GCM, which proposed an integrative schema solution for several genomic repositories. We extended it by adding three entities that are relevant for GWAS and we implemented the data import and transformation pipeline to store the datasets of two new data sources (i.e., GWAS Catalog and FinnGen) within the META-BASE repository. We detailed the mapping effort performed for this purpose and finally demonstrated the usefulness of this work by using a domain-specific language that interrogates FinnGen together with 1000 Genomes datasets for extracting SNPs relevant to the schizophrenia trait.

Acknowledgement. This research is funded by the ERC Advanced Grant 693174 GeCo (data-driven Genomic Computing).

References

1. Bernasconi, A., et al.: META-BASE: a novel architecture for large-scale genomic metadata integration. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020)
2. Bernasconi, A., et al.: Conceptual modeling for genomics: building an integrated repository of open data. In: Int. Conf. on Conceptual Modeling. pp. 325–339. Springer (2017)

3. Buniello, A., et al.: The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research **47**(D1), D1005–D1012 (2019)
4. FinnGen Project: 5th release. https://finngen.gitbook.io/documentation/ (2020)
5. Frankish, A., et al.: GENCODE reference annotation for the human and mouse genomes. Nucleic acids research **47**(D1), D766–D773 (2019)
6. GWAS Catalog team: GWAS catalog website. https://www.ebi.ac.uk/gwas/
7. International Multiple Sclerosis Genetics Consortium: Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science **365**(6460) (2019)
8. Kundaje, A., et al.: Integrative analysis of 111 reference human epigenomes. Nature **518**(7539), 317–330 (2015)
9. Masseroli, M., et al.: Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. Methods **111**, 3–11 (2016)
10. Masseroli, M., et al.: GenoMetric Query Language: a novel approach to large-scale genomic data management. Bioinformatics **31**(12), 1881–1888 (2015)
11. Palacio, A.L., et al.: A method to identify relevant genome data: conceptual modeling for the medicine of precision. In: Int. Conf. on Conceptual Modeling. pp. 597–609. Springer (2018)
12. Pastor, O.: Understanding the human genome: a conceptual modeling-based approach. In: Int. Conf. on Database and Expert Systems Applications. pp. 467–469. Springer (2010)
13. Pastor, O., et al.: Enforcing conceptual modeling to improve the understanding of human genome. In: 2010 Fourth Int. Conf. on Research Challenges in Information Science. pp. 85–92. IEEE (2010)
14. Rambold, G., et al.: Meta-omics data and collection objects (MOD-CO): a conceptual schema and data model for processing sample data in meta-omics research. Database **2019**
15. Román, J.F.R., et al.: Applying conceptual modeling to better understand the human genome. In: Int. Conf. on Conceptual Modeling. pp. 404–412. Springer (2016)
16. Rudy, G., Helix, G.: A hitchhiker's guide to next-generation sequencing. http://www.goldenhelix.com/pdfs/whitepapers/Hitchhikers-Guide-to-NGS.pdf (2010)
17. Schuster, S.C.: Next-generation sequencing transforms today's biology. Nature methods **5**(1), 16–18 (2008)
18. Tam, V., et al.: Benefits and limitations of genome-wide association studies. Nature Reviews Genetics **20**(467-484) (2019)
19. The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. Nature **467**(7319), 1061–1073 (2010)
20. The ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)
21. Wang, K., et al.: A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophrenia Research **124**(1), 192–199 (2010)
22. Wang, L., et al.: Biostar models of clinical and genomic data for biomedical data warehouse design. Int. j. of bioinform. research and applications **1(1)**, 63–80 (2005)
23. Weinstein, J.N., et al.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)
24. Zheng, R., et al.: Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Research **47**(D1), D729–D735 (2018)