Using Metadata for Locating Genomic Datasets on a Global Scale

Anna Bernasconi
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Milan, Italy
anna.bernasconi@polimi.it

Abstract

Genomic research benefitted from recent extraordinary improvements in DNA sequencing techniques, leading to the production of enormous amounts of datasets that store information such as nucleotide sequences, gene locations/levels of expression, proteins-DNA interactions. As this has now become a *big data* matter, characterized by an underlying disorganization, there is a strong need for integrative solutions.

In this paper, we devote our efforts to the management of genomic data, to be organized and located using experimental studies descriptions. Such documentation, also referred to as *metadata*, contains fundamental information to understand the content of experimental samples (namely, how the biological material was extracted and processed, in which clinical conditions, with which techniques.) We propose a novel framework to manage metadata of genomic datasets, offering a unified view with respect to a number of heterogeneous data sources (usually big international consortia, but also small research centers) that currently display their metadata in disorganized and very cumbersome formats. The final outcome of this work is a search platform which allows easy location of relevant sources for specific genomic data analysis problems.

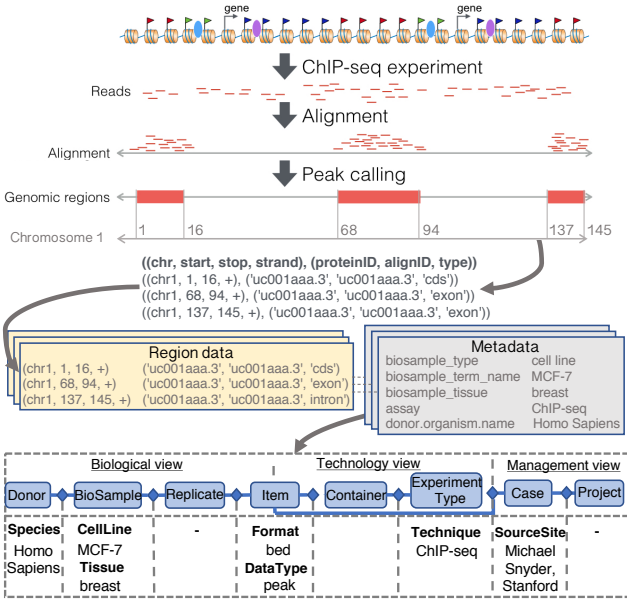

Figure 1: From a DNA fragment, through a sequence of steps, to the memorization of information in region data and its related metadata, made available for exploration in a user interface. Source of upper part of figure: http://www.regulatory-genomics.org/rgt/basic-introduction/

1 Introduction

Genomic research is blooming because of revolutionary technologies to sequence DNA (Next Generation Sequencing), which operate at much faster rates and lower costs than traditional techniques. Such speed-up is achieved by means of massively parallel sequencing, which enables millions of nucleic acids fragments to be handled simultaneously. A single human genome, about 3 billion units of DNA in 23 thousands genes, can now be processed in just a single day and stored in around 200 Gigabytes [CCK+17].

Because of thousands of new experimental datasets becoming available every day, genomics has become a new "big data" generator (see [SLF+15] for com-

parison with other major big data domains). To boost further research, this wealth of data needs to be made available for search and download. Currently, it is distributed across a range of worldwide repositories (nearly 1,000 sequencing centers in 55 countries in universities, hospitals, and other research laboratories), usually coordinated by national research consortia and institutes. Organizations such as the International Cancer Genome Consortium (ICGC, [ZBC+11]), the National Cancer Institute Genomic Data Commons (GDC, [JFGS17]), the National Center for Biotechnology Information (NCBI, [Coo17]), the National Human Genome Research Institute (NHGRI, [Man16]), and the European Bioinformatics Institute (EBI, [LVAL07]) maintain and enrich the repositories of genomic data, that may contain both open and controlled data (i.e., only accessible upon approval from a Data Access Committee). Public data are beneficial to researchers and clinicians who can access and compare them, as well as search for common patterns across a large number of individual.

While data sources have more or less agreed on the definition of protocols and formats for *data* production and transformation, no convergence has been observed for common *metadata* formats. The existing repositories propose their own standards which are only used internally, presume a thorough knowledge of their specific rules, and require tedious manual work to allow for use of data from combined sources.

Considering only publicly available data, we focused on the need of the genomic research community for a tool which helps to locate and retrieve interesting data to solve biological and clinical questions and also favours data interoperability. We propose a metadata storage system, specific for genomic datasets, with a four-fold contribution: 1. gradual inclusion of all processed datasets from sources considered interesting for *tertiary analysis* (i.e., data analysis in charge of "making sense" of genomic signals [Gab10]); 2. integration of genomic data residing in these heterogeneous data sources to provide a unified view of the comparable concepts; 3. curated representations of metadata, maintained coherent with the current status of the original sources; 4. user-friendly search functionality, based on key-words characterizing the samples, but also on their synonyms and hypernyms, which are retrieved through specialized ontologies.

Fig. 1 illustrates the story behind our effort. The genomic problem can be broken down into a sequence of computational steps. The genome material (e.g., a DNA fragment), by means of an experimental sequencing technique (i.e., ChIP-seq), can be translated first in reads (through *primary analysis* methodologies), then alignments, signals, and finally regions (by means of *secondary analysis* methods).

Within the GeCo Project[1], described in [CBC+17], we use a machine readable representation, which includes "Region data" and the related "Metadata" files. Region data consists of quadruples such as `(chr1,1,16,+)`, which identifies the region contained in chromosome 1 of the human genome, spanning from coordinates 1 to 16 w.r.t. a reference genome, and being located in the positive strand of the double helix structure of DNA. Metadata, instead, contain information about the genomic experiment which generated the data.

In this paper we propose a system which, after submitting metadata through a data integration pipeline, as a final step exposes them by means of a user interface—similar to the one shown at the bottom of the Fig. 1—ready for querying.

With our system, we aim to encourage the use of genomic datasets, allowing easier semantically enriched search and resulting download of processed data. We have previously proposed GMQL [MCP+18], a high-level query language for genomics, and GDM [MKPC16], an integrative model for processed data formats. Using the system described here in combination with the query language and execution engine implemented within the GeCo Project, we aim to help support the specific processes of retrieval, exploration, and analysis of genomic data.

The paper is structured as follows. Section 2 introduces metadata usefulness with a motivating example. Section 3 overviews the overall system which integrates data, driven by the use of the Genomic Conceptual Model. Section 4 explains how we allow novel searches over the database of genomic experiments through a web interface. Section 5 briefly mentions related works in the literature. Finally, Section 6 concludes the paper.

2 Motivating Example

Genomic datasets are typically characterized by explanatory information that can be consulted on the interfaces of data sources; sometimes they are available for download in various semi-structured formats. Generally, aspects described by metadata can be clustered in the following areas: clinical information regarding the physical individual who has donated the biological sample extracted for sequencing; bio specimen information about the tissue (or cell culture) of provenance and the possible pathologies that affect the biological material; the technologies (e.g., platforms), methodologies (i.e., pipelines), and processes used to sequence

[1]Data-Driven Genomic Computing, `http://www.bioinformatics.deib.polimi.it/geco/`

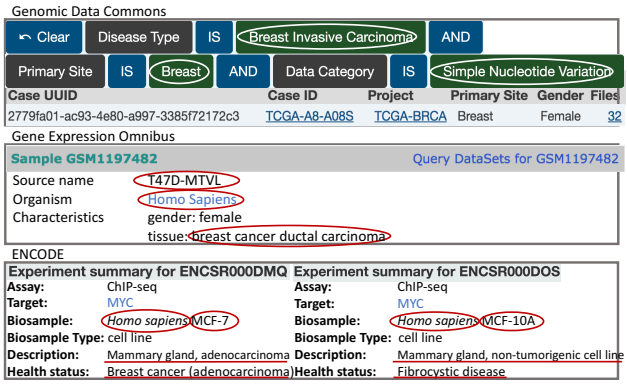

Figure 2: Example of web interfaces of systems: GDC Data Portal [JFGS17] (first rectangle at the top), NCBI Gene Expression Omnibus [BWL+13] (middle), and ENCODE [rE12] (bottom).

the DNA, to align the sequences, and to further produce DNA regions; the formats and data types, which describe the new shape of data, defining what kind of information it delivers; details on the organization aspects that include the program, project, and case study under which the experiment is being conducted. All these aspects are memorized by data sources in various ways. Heterogeneity spans from download protocols and formats to attributes names and values.

To motivate our effort towards an integrated platform, we introduce an example which simulates the research of data suitable for a genomics project. For illustration purposes, we include just bio specimen information, leaving aside technological and clinical aspects. Consider a comparison study between a human non-healthy breast tissue, suffering from carcinoma, and a healthy sample coming from a similar tissue. A researcher in the field, due to previous experience, knows three portals to locate interesting data for this analysis. The results obtained after some browsing are reported in Fig. 2.

For the diseased data, describing gene expression, the chosen source is GDC Data Portal, an important repository on human cancer mutation data. As it can be seen on the top of Fig. 2, one or more cases (i.e., datasets) can be retrieved by composing a query which allows to locate variation data on "Breast Invasive Carcinoma" from "Breast" tissue.

To compare such data with references, the researcher chooses additional datasets coming from cell lines, i.e., cell cultures which have been permanently established and made immortal. Since cell lines are considered a standard for similar investigations in the past, they are frequently used in place of primary cells to study biological processes. The scientific community tends to accept the derived findings more readily.

A tumor cell line data is found on the GEO web interface (middle rectangle of Fig. 2) where, by browsing thousands of samples, the researcher locates one from "Homo Sapiens" organism, where the analyzed cell type is "T47D-MTVL" and observed disease is "breast cancer ductal carcinoma". On ENCODE, instead, the researcher chooses both a tumor cell line (bottom left of Fig. 2) and a normal cell line (bottom right), to make a control check. "MCF-7" is a cell line started from a diseased tissue afflicted with "Breast cancer (adenocarcinoma)", while "MCF 10A" is its widely considered non-tumorigenic counterpart. Note that considerable external knowledge is necessary in order to find these connections, which cannot be obtained on the mentioned portals. Concerning the disease choice: "breast invasive carcinoma" is the same as "breast carcinoma" (as observed in the annotation from EBI's Expression Atlas [JB15]), which allows to compare GDC's data with the datasets from GEO and ENCODE, since they describe more specific diseases (i.e., "breast cancer (adenocarcinoma)" and "breast cancer ductal carcinoma" are its sub-types, according to the Disease Ontology [KAF+14]). Concerning the cell lines choice: researchers typically query specific databases (such as the cell line browser of the Catalogue Of Somatic Mutations In Cancer[2]) or dedicated forums to discover tumor/normal matched cell line pairs. This information is not encoded in a unique way over sources and is often missing.

3 Integration Procedure

During the design phase we considered four data sources: Genomic Data Commons (GDC, [JFGS17]), containing over 310,000 files, across over 32,000 cases, in 40 projects, covering many aspects of cancer genomics; the Encyclopedia of DNA Elements (ENCODE, [rE12]), with almost 420,000 files, allocated in over 15,000 experiments, part of 6 different projects, related to the functional DNA sequences which intervene at the protein/RNA levels and to the regulatory elements which control gene expression; the Gene Expression Omnibus (GEO, [BWL+13]), an international public repository of high throughput gene expression (and other) data sets submitted by the research community, linked to almost 20,000 published manuscripts; Roadmap Epigenomics Project (REP, [KME+15]) containing 1,936 datasets related to genetic variation in association with human disease based on epigenomics evidence. Other three data sources have been used to validate the approach and we plan to add many others as future work.

The conceived metadata integration process is designed as an incremental procedure. First, we perform

[2] https://cancer.sanger.ac.uk/cell_lines/

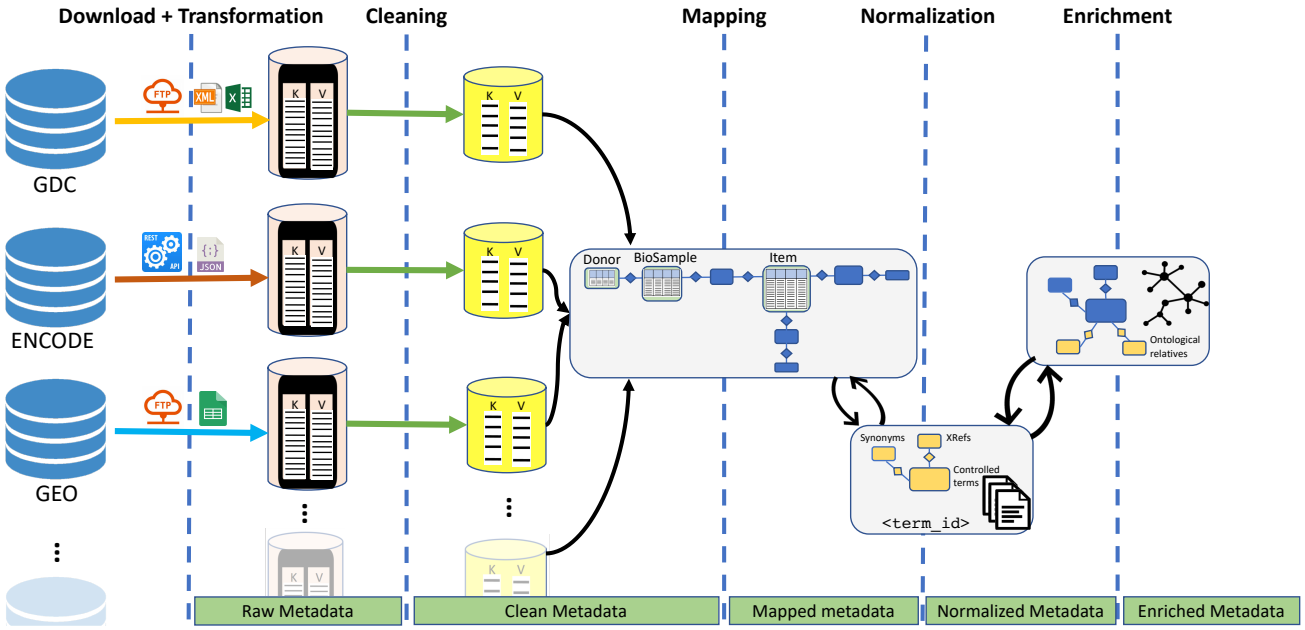

Figure 3: The overall data integration process.

a one-time activity to design the integration rules divided in six steps: download, transformation, cleaning, mapping, normalization, and enrichment (see Fig. 3). Then, we periodically perform data integration sessions where a range of increments with different magnitude are supported: new data sources, portions of data sources, datasets, or basic samples. Rules can be developed in a dynamical fashion and may be modified during design sessions.

The *downloading* phase handles heterogeneity at the distribution and format level. It takes into account various access protocols (such as FTP, HTTP RESTful API, and file bundles) as well as data formats (XML, JSON, CSV, Excel, and Google Sheet) and imports at our repository site the original data and their metadata from the sources.

Through the *transformation*, metadata are translated into a simpler structure of $\langle attribute, value \rangle$ pairs, where the *attribute* describes the kind of represented information and the *value* embodies the actual information (e.g., $\langle$`biosample_type`,"cell line"$\rangle$, $\langle$`target.gene_name`,"RAD21"$\rangle$). This representation is useful for applications which benefit from a semi-structured version of metadata, for example for distributed file systems.

Pairs are consequently *cleaned*, thus producing a collection of clean metadata pairs for each source. As an example, a rather complicated attribute such as `replicates.biosample.donor.organism.scientific_name` is simplified into `donor.organism`, with the aim to facilitate human understanding and also the following integration steps. Similarly, `file.file_type` becomes `file.type`. Redundant information (i.e., duplicated attributes) is removed.

The *mapper* extracts information from some of these pairs and maps it to a relational database.

In [BCCM17] we presented the Genomic Conceptual Model (GCM), shown in Figure 4. This is an entity-relation schema that summarizes the most important information metadata shared between the genomic data sources. GCM's main objective is to recognize a common organization for a limited set of concepts which are supported by most data sources, although with very different names and formats. GCM is centered on the ITEM entity representing an elementary experimental unit and stored as a single file of genomic regions and their attributes. GCM is organized as a star-schema around ITEM. Its three hierarchical dimensions or views, indicated in Figure 4 with big arrows, describe: 1. the *biological* phenomena observed in the experiment: the sequenced replicated sample, the biological material and its preparation, its donor; 2. the *technology* used in the experiment, including parameters used for internal selection and organization of items (i.e., container), and the specific technique; 3. the *management* aspects of the experiment: the case studies and projects/organizations behind its production.

The GCM, used as a global model, drives a schema integration process. We specify *ad hoc* mappings between the entities of the global schema and cleaned attributes from the sources. Consider the entity DONOR from the global schema, which contains information on the individual from which the biological sample is extracted. As a basic example, to fill one of its attributes, e.g., *Ethnicity*, we derive the value related to the at-

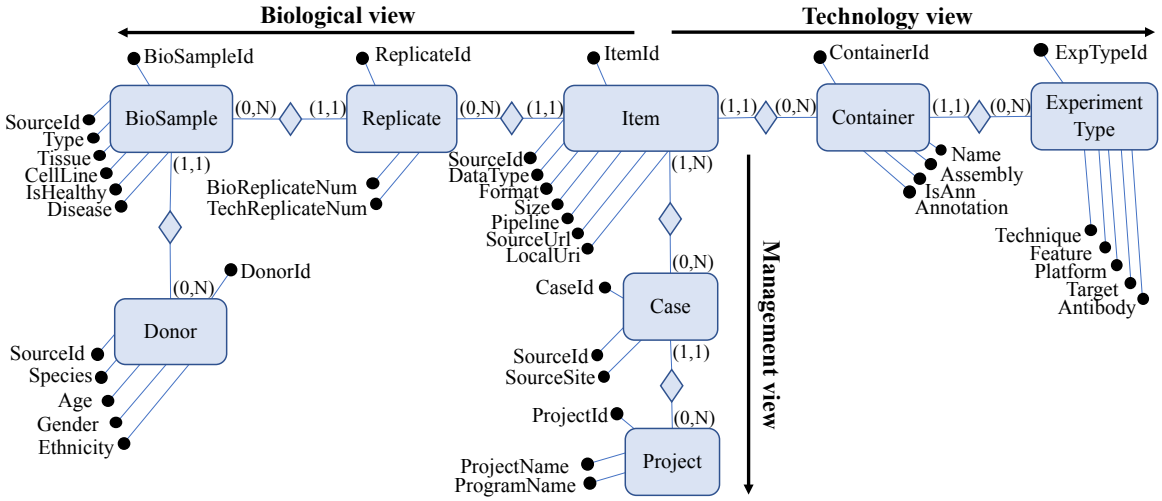

Figure 4: Genomic Conceptual Model presented in [BCCM17].

tribute `donor.ethnicity` from the data source ENCODE, while we extract and concatenate `demographic.ethnicity` and `demographic.race` from GDC; from GEO this is usually a missing information, while rarely we can find `characteristics.donor_ethnicity`. The mapper also handles duplicates: if two samples, respectively called 'ENCBS236ISK' (from ENCODE) and 'GSM2192006' (from GEO) contain information about an external reference, i.e., univocal indication of being the same real-world entity, they are registered in our repository with a unique identifier.

The values mapped into the global schema are then *normalized*. Normalization acts to ensure metadata consistency at the semantic level. This phase involves linking values to controlled vocabularies or biomedical ontologies, typically manually curated by expert curators. Fig. 5 shows a biological sample entity from the global schema. From the original source only the information in the blue solid boxes is retrieved. This is then completed through normalization, which adds the information in the red dashed boxes. For example, the disease information "Breast cancer (adenocarcinoma)" is equipped with a synonym "Mammary adenocarcinoma" and DOID:3458, the corresponding concept identifier in the Disease Ontology [KAF+14].

Finally, values are *enriched* by means of external ontologies. During this phase, values that have super-concepts or sub-concepts in the biomedical ontologies are enriched with all concepts in a *is a* relationship within three steps in the ontology graph (see information in the green dotted boxes in Fig. 5). For example, the value "breast", corresponding to the attribute *Tissue*, is enriched by both its super-concept "Female reproductive gland" and its sub-concept "Mammary duct", among others. Details of normalization and enrichment pipelines are available in [BCCC18].

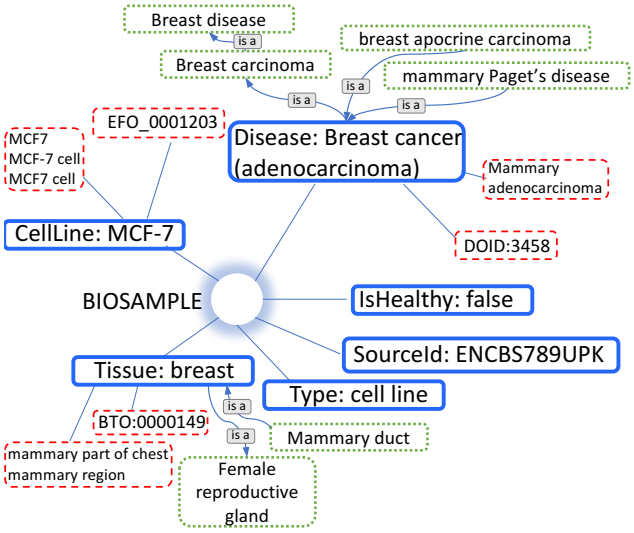

Figure 5: Example of normalization and enrichment of a BIOSAMPLE tuple from the ENCODE data source.

4 User-friendly Search Platform

The web platform ensures easy and fast location of datasets from the considered set of repositories. We provide the URL endpoint for download from our system, when the dataset is available (as it was retrieved from the original system or transformed into processed data to make it suitable for tertiary analysis). Otherwise, we provide the original source URL for download. The laborious integration process is designed to make data querying easier. An example instance of a user query on our interface can be appreciated in the lower part of Fig. 1. This query for genomic experiments data works regardless of how requested values are expressed. For example, due to the mapping efforts made during the integration process, by using the DONOR column *Species*, the user can also reach data that was documented through alike concepts, such as *organism*,

rather than abbreviations, such as *Sp.*, or words in other languages, such as the italian equivalent *specie*. Moreover, due to the normalization and enrichment efforts made during the integration process, a search for samples with donors of "Homo Sapiens" species will result in a selection of samples which were marked with this annotation or, alternatively, with synonyms (e.g., "man", "Human"), abbreviations (e.g., "H. sapiens"), misspellings (e.g., "Homo sapeins"), or even sub-concepts (e.g., "Homo sapiens neanderthalensis"). Similarly, concept-based search holds also for other attributes.

To support these functionalities, the system rewrites user queries to instrument wider searches, which also cover synonyms, hyponyms, and other kinds of similarities.

5 Related works

Many works in literature use conceptual models in the genomics—and more in general biomedical—field. However, they employ conceptual models' expressive power to explain biological entities and their interactions [WZR05, RPCV16]. Instead, we propose to use a conceptual model as the driving principle to achieve data integration.

In the state of the art there have been multiple attempts to offer integrated access to heterogeneous sources. Some of these are: BioKleisli [DOTW97] (to provide read access to complex structured data), BioMart [SHD+15] (for biomedical databases), NIF [GBM+08] (in the neuroscience field), and DATS [SGBRS+17] (for scientific datasets in general).

Also some of the genomics consortia mentioned earlier have provided methodologies to organize metadata (see the BioProject database [BCG+12], Encode Data Coordination Center [HSC+16], and Genomic Data Commons [JFGS17]). However, these are not frameworks which are general enough to make possible including all genomic data sources, regardless of how far apart the sub-areas on which their data focus.

Also DeepBlue [ALBL16], an interesting starting point in terms of easy-to-use interfaces, only handles epigenomic data (i.e., study of epigenetic modifications on the cell), a small area compared to the whole genomics.

DNADigest [KWR+16] is an effort that investigates the problem of locating genomic data to download for research purposes. Their work differs from ours since, even allowing a dynamical and collaborative curation of metadata, they only provide means to locate raw data. Instead, we provide processed data ready to be used for tertiary analysis.

6 Conclusions

Following the need to make genomic datasets and their information collectively searchable, we are proposing a framework to manage, integrate and enrich semantically the experimental data documentation. We are soon delivering an online platform for genomic data querying driven by metadata, which will be appreciated by the genomic research community. This will be an important resource for: 1. conducting research activities by using directly our processed data, available from the data repository we are currently developing as a major research project (further details are omitted for anonymity reasons); 2. locating data through the URL endpoints of the original data sources.

Acknowledgements

This research is funded by the ERC Advanced Grant 693174 GeCo (Data-Driven Genomic Computing), 2016-2021.

References

[ALBL16] Felipe Albrecht, Markus List, Christoph Bock, and Thomas Lengauer. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome. *Nucleic acids research*, 44(W1):W581–W586, 2016.

[BCCC18] Anna Bernasconi, Arif Canakoglu, Andrea Colombo, and Stefano Ceri. Ontology-driven metadata enrichment for genomic datasets. In *International Conference on Semantic Web Applications and Tools for Life Sciences*, volume 2275. CEUR-WS, 2018.

[BCCM17] Anna Bernasconi, Stefano Ceri, Alessandro Campi, and Marco Masseroli. Conceptual modeling for genomics: Building an integrated repository of open data. In *International Conference on Conceptual Modeling*, pages 325–339. Springer, 2017.

[BCG+12] Tanya Barrett, Karen Clark, Robert Gevorgyan, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*, 40(D1):57–63, 2012.

[BWL+13] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, et al. NCBI GEO: archive for functional genomics data

sets - update. *Nucleic acids research*, 41(Database issue):D991–D995, 2013.

[CBC+17] Stefano Ceri, Anna Bernasconi, Arif Canakoglu, et al. Overview of geco: A project for exploring and integrating signals from the genome. In *International Conference on Data Analytics and Management in Data Intensive Domains*, pages 46–57. Springer, 2017.

[CCK+17] Stefano Ceri, Arif Canakoglu, Abdulrahman Kaitoua, et al. Data-driven genomic computing: Making sense of signals from the genome. 2017.

[Coo17] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 45(Database issue):D12, 2017.

[DOTW97] Susan B Davidson, Christian Overton, Val Tannen, and Limsoon Wong. Biokleisli: A digital library for biomedical researchers. *International Journal on Digital Libraries*, 1(1):36–53, 1997.

[Gab10] R. Gabe. A hitchhiker's guide to next generation sequencing. http://blog.goldenhelix.com/grudy/a-hitchhikers-guide-to-next-generationsequencing-part-2/, 2010. Accessed: 2018-05-17.

[GBM+08] Amarnath Gupta, William Bug, Luis Marenco, et al. Federated access to heterogeneous information resources in the neuroscience information framework (nif). *Neuroinformatics*, 6(3):205–217, 2008.

[HSC+16] Eurie L Hong, Cricket A Sloan, Esther T Chan, et al. Principles of metadata organization at the encode data coordination center. *Database*, 2016, 2016.

[JB15] Simon Jupp and Tony Burdett. A new ontology lookup service at embl-ebi. In J. Malone et al., editors, *Proceedings of SWAT4LS International Conference 2015*, 2015.

[JFGS17] Mark A Jensen, Vincent Ferretti, Robert L Grossman, and Louis M Staudt. The nci genomic data commons as an engine for precision medicine. *Blood*, 130(4):453–459, 2017.

[KAF+14] Warren A Kibbe, Cesar Arze, Victor Felix, et al. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*, 43(D1):D1071–D1078, 2014.

[KME+15] Anshul Kundaje, Wouter Meuleman, Jason Ernst, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[KWR+16] Nadezda V Kovalevskaya, Charlotte Whicher, Timothy D Richardson, et al. Dnadigest and repositive: connecting the world of genomic data. *PLoS biology*, 14(3):e1002418, 2016.

[LVAL07] Alberto Labarga, Franck Valentin, Mikael Anderson, and Rodrigo Lopez. Web services at the european bioinformatics institute. *Nucleic acids research*, 35(suppl_2):W6–W11, 2007.

[Man16] Teri A Manolio. Implementing genomics and pharmacogenomics in the clinic: The national human genome research institute's genomic medicine portfolio. *Atherosclerosis*, 253:225–236, 2016.

[MCP+18] Marco Masseroli, Arif Canakoglu, Pietro Pinoli, Abdulrahman Kaitoua, et al. Processing of big heterogeneous genomic datasets for tertiary analysis of next generation sequencing data. *Bioinformatics*, page bty688, 2018.

[MKPC16] Marco Masseroli, Abdulrahman Kaitoua, Pietro Pinoli, and Stefano Ceri. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods*, 111:3–11, 2016.

[rE12] Consortium ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[RPCV16] José F Reyes Román, Óscar Pastor, Juan Carlos Casamayor, and Francisco Valverde. Applying conceptual modeling to better understand the human genome. In *International Conference on Conceptual Modeling*, pages 404–412. Springer, 2016.

[SGBRS+17] Susanna-Assunta Sansone, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, et al. Dats, the data tag suite to enable discoverability of datasets. *Scientific data*, 4:170059, 2017.

[SHD+15] Damian Smedley, Syed Haider, Steffen Durinck, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):589–598, 2015.

[SLF+15] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, et al. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[WZR05] Liangjiang Wang, Aidong Zhang, and Murali Ramanathan. BioStar models of clinical and genomic data for biomedical data warehouse design. *Int. J. Bioinformatics Res. Appl.*, 1(1):63–80, April 2005.

[ZBC+11] Junjun Zhang, Joachim Baran, Anthony Cros, et al. International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, 2011, 2011.