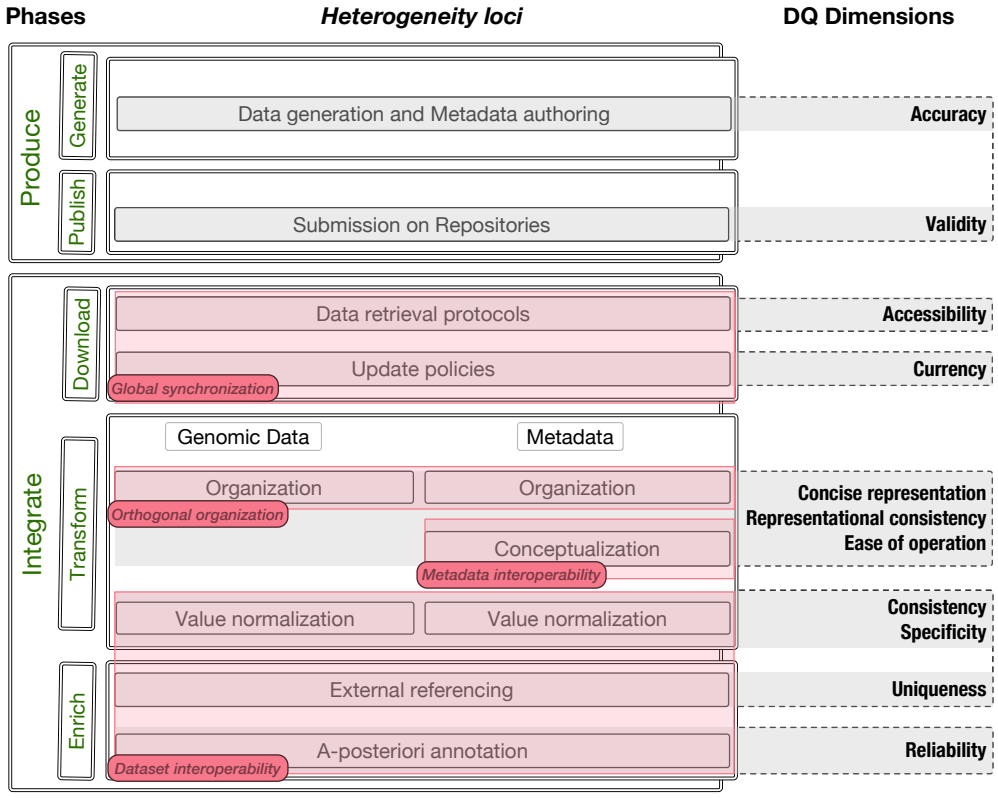
Graphical Abstract

Data quality-aware genomic data integration

Anna Bernasconi



Highlights

Data quality-aware genomic data integration

Anna Bernasconi

- The integration of data and metadata is very relevant in biomedical fields (including genomics), because critical decisions in healthcare depend on it.
- Heterogeneity aspects affect many actors and stages of the genomic data life cycle; data quality dimensions can adequately lead the analysis of problems and related solutions.
- The focus so far has been on quality of genomic signals extracted from raw data, while more efforts are needed on processed data and metadata issues.
- Data quality dimensions should be addressed systematically during data integration of diverse sources, to enable further integrative studies.
- Future data integration approaches will include more and more a data quality-aware *modus operandi* with currency, conciseness, consistency, reliability-driven approaches.

Data quality-aware genomic data integration

Anna Bernasconi[a,<]

[a]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

ARTICLE INFO

Keywords:
data quality
data integration
data curation
genomic datasets
metadata
interoperability

ABSTRACT

Genomic data are growing at unprecedented pace, along with new protocols, update polices, formats and guidelines, terminologies and ontologies, which are made available every day by data providers. In this continuously evolving universe, enforcing quality on data and metadata is increasingly critical. While many aspects of data quality are addressed at each individual source, we focus on the need for a systematic approach when data from several sources are integrated, as such integration is an essential aspect for modern genomic data analysis. Data quality must be assessed from many perspectives, including accessibility, currency, representational consistency, specificity, and reliability.
In this article we review relevant literature and, based on the analysis of many datasets and platforms, we report on methods used for guaranteeing data quality while integrating heterogeneous data sources. We explore several real-world cases that are exemplary of more general underlying data quality problems and we illustrate how they can be resolved with a structured method, sensibly applicable also to other biomedical domains. The overviewed methods are implemented in a large framework for the integration of processed genomic data, which is made available to the research community for supporting tertiary data analysis over Next Generation Sequencing datasets, continuously loaded from many open data sources, bringing considerable added value to biological knowledge discovery.

1. Introduction

Genomics is going to generate the largest "big data" problem for the mankind: between 100 million and 2 billion human genomes are expected to be sequenced by 2025 [114]. High-throughput technologies and, more recently, Next Generation Sequencing [111] have brought increasing amounts of genomic data of multiple types, realizing huge steps towards unravelling human genome mechanisms and applying them for unprecedented personalized medicine outcomes. Prior to data analysis and biological knowledge discovery, data and metadata integration is considered an activity of irrefutable priority, with pressing demands for enhanced methodologies of data extraction, matching, normalization, and enrichment, to allow building multiple perspectives over the genome; these can lead to the identification of meaningful relationships, otherwise not perceivable when using incompatible data representations [107].

Bioinformatics, including genomics in particular, operates traditionally by exploiting the considerable fieldwork on data acquisition, wrangling and analysis from its practitioners. Best practices are accumulated across labs and different projects, shared on forums (e.g., https://www.biostars.org/, http://seqanswers.com/, https://www.researchgate.net/) and collected in the documentation or wiki-guides in code repositories of tools and software. Within these processes, bioinformaticians are mostly concerned with the quality of the experimental data produced by sequencing platforms, for which consolidated pipelines – often composed of many scripts – are available.

In comparison, quality actions that can be performed when aggregating multiple experimental data in systematized ways, have received less attention. However, with the emergence of a culture of data FAIRness [122] and of open and sharable science – promoted by initiatives such as FAIRsharing [109] – caring for data standards in both schemata and values becomes increasingly important, in the same way as implementing integration practices that foster data quality (focusing on accuracy, consistency, currency, and reliability [64]). In its 2012 report on quality of data, IBM found that 1 out of 3 business leaders do not trust the information they use to make decisions (https://www.ibmbigdatahub.com/infographic/four-vs-big-data); this ratio is unacceptable in fields like health-care and precision medicine, that are strongly driven by genomic databases and decision methods.

Recently, we have observed a trend of initiatives that gather tools and data structures to support interoperability among highly heterogeneous systems, to help bioinformaticians perform a set of curation and annotation operations. These include community-driven efforts such as bio.tools [68] (anchored within ELIXIR, https://www.elixir-europe.org/) service providers (EBI [98]), software suites (Bioconductor [67]), or lists (http://msutils.org/). Specific instances include APIs such as BioPython [34], tailored scripts, and field descriptions to be parsed (Bioschemas.org [60]). By using, e.g., the EDAM ontology [69], single initiatives can build bridges among resources, while conforming to well-established operations, types/formats of data and application domains.

In this fashion, most problems are handled within a single database by means of *on-the-fly data integration*, driven by a community-inspired guidance. On the other hand, a more systematic approach of *low-level integration* – based upon experience in building solid data warehouses – has also been adopted, helping to reach stable interoperability among imported sources. In these years we witnessed attempts to this kind of approach at many international centers for genomics (including the Broad Institute – https:

[<]Corresponding author
anna.bernasconi@polimi.it (A. Bernasconi)
ORCID(s): 0000-0001-8016-5750 (A. Bernasconi)

//www.broadinstitute.org/ and Wellcome Sanger Institute https://www.sanger.ac.uk/ that are so far unpublished) as well as in companies (including SciDB, implemented by Paradigm4, https://www.paradigm4.com/try_scidb/). In the context of research applied to real problems of the domain, the data-driven Genomic Computing project (GeCo [27]) has dedicated considerable e orts to integrate sources of data that are open for secondary research use, hence downloadable to a common repository, continuously updated. Such systems provide the advantage of o ering to users practical work environments. Indeed, biologists and clinicians appreciate ready-to-use repositories, while the know-how of bioinformaticians/developers (on scripting and querying technologies) may not be always at hand.

The emergence of the mentioned positive experiences indicates how data quality can be generally addressed, within thousands of cooperative studies that are jeopardized due to the poor quality of genomic data integration. Poor quality arises at very diverse levels: protocols, data units and dictionaries, metadata models and terminologies. We propose a step forward in the holistic understanding of genomic data and metadata integration, describing a number of methods that can be practically employed to resolve heterogeneity. While most of the introduced issues and techniques have commonalities with general data integration problems, we instantiate them in the speci c genomic data context, providing practitioners with easy-to-relate examples to guide their procedures.

In Section 2 we discuss the state of the art since the earliest works on quality-aware genomic databases management [91, 16]. In Section 3 we focus on processed data (i.e., the signal extracted from raw genomic datasets) and on metadata (i.e., data description), which is the main driver for interoperability and interconnectedness of di erent databases; in this context, we present a taxonomy of data integration procedures that can positively a ect data quality issues. We interpret integration as a set of steps [13], during which practitioners encounter several heterogeneity loci, which are contexts that cause heterogeneity and that may be addressed during the speci c activities of integration.

In Section 4 we describe a collection of problems with related practical examples and solutions proposed as common practices or speci c of our experimented pipelines. In this context, data integration involves: synchronizing the content of a global repository with the data sources, organizing data and corresponding metadata with a unique orthogonal approach, considering interoperability of data descriptions and, more in general, allowing heterogeneous datasets to be used together seamlessly. More pragmatically, we argue that the problem of data quality cannot be addressed as an independent issue. It is entangled with many other aspects regarding data modeling, management, integration and usage. We do not consider quality deriving from original sources as it is not a space where we can intervene *a posteriori*. Instead, we propose a novel angle: addressing data quality dimensions while diverse data sources are being integrated together to enable further applications. In conclusion, in Section 5 we rea rm the need for quality-aware solutions to integration and mention our vision on upcoming approaches dedicated to this matter.

2. Background

From DNA microarrays [70] to Next Generation Sequencing [111], quality in genomics has been usually employed to refer to quality control steps on sequences, typically a pre-processing activity aimed at removing adapter sequences, low-quality reads, uncalled bases, contaminants. Instead, in this review we refer to Data Quality (DQ) in the broader sense de ned by Wang and Strong [119], usually captured by the expression tness for use , i.e., the ability of datasets to meet their users' requirements. DQ is evaluated by means of di erent quality dimensions (i.e., single aspects or components of a data quality concept [115]). State of the art techniques to solve data quality issues in general databases are summarized in [50], under the name of `data cleaning'.

In Figure 1 we appreciate the chronological order of publication of relevant literature. In general, more foundational works of data quality in genomics/biological database have appeared in the early years between 2003 and 2008, building the rst baseline for this subject, while after 2014 we observe more speci c contributions. Müller et al. [91] examine the quality of molecular biological entities databases. Within the production of data, the authors identify intrinsic problems that lead to incorrect data, concluding that traditional data cleaning techniques, used successfully in many other domains, do not t the peculiarities of genomics. While giving a complete review of potentially very dangerous errors in sequence and annotation genomic databases, the discussion leaves aside processed data as well as aspects related to data integration and integrated access to multiple heterogeneous sources.

While in 2005 Martinez and Hammer propose the conceptual integration of data quality measures inside a model of data [80], the research group led by Berti-Équille is more focused on the problems deriving from warehousing genomic data [16, 63, 89]. Their overall experience is summarized in [88], where they claim that metadata describing data preparation and data quality are not exploited enough for ensuring valid results of downstream data analysis. Along the same lines of the pioneer work of Wang and Strong [119], some works propose measures and conceptual frameworks that take into account user-driven quality requirements; see the work by Missier et al. [87], BioGuideSRS [35], BioDQ [81] and the more recent paper by Veiga et al. [118]).

A preliminary work by León et al. [76] classi es the data quality properties that are most relevant for genomics; it was then applied concretely to a Crohn's Disease clinical diagnosis case study [97]. The general framework has been described very recently in Pastor et al. [99]. Rajan et al. [104] have recently proposed to build a knowledge base for assessing quality and characterizing datasets in biomed-

Figure 1: The timeline of publications targeting data quality issues in biological (and more precisely genomic) databases. Red circles represent works describing approaches to resolve duplication; green circles are works on data warehousing or conceptual modeling; blue circles cover expert curation literature; grey circles are for user-driven data quality approaches; black circles are uncategorized works.

ical repositories, thus including also genomics and other translational research data. Other works address data quality on speci c kinds of genomic databases (e.g., by Hedeler and Missier [64] for transcriptomics and proteomics, by Etcheverry et al.[49] for Genome Wide Association Studies, and by Gonçalves and Musen [59] for repositories of biological samples).

As to speci c addressed problems, duplicate detection in biological data was dealt with association rule mining rst by Koh et al. [74], then by Apiletti et al. [2, 3] and by Müller et al. [90], with a focus on contradicting databases. Recent works cover the prevention of redundancy in big data repositories (UniProt KB in [23] and high throughput sequencing in [52]), providing a comparison with other widely used biological large data repositories.

A number of approaches focus on primary data archives quality (i.e., sequence databases such as GenBank [110]) to automatically detect inconsistencies with respect to literature content (see [20, 21]) and to provide benchmarks [31], general categorizations of duplicates [32], de-duplication clustering methods [30], as well as insights on characteristics, impacts and related solutions to the problem of duplication in biological databases [29].

Finally, data integration in a quality-aware perspective includes practices of data curation [108] and of service/process curation [58]. Data curation is explored in [117], where the Eagle-i system is developed to facilitate collaborative curation, and in [102, 101], as a means to deal with con icting and erroneous data in UniProtKB.

Unfortunately, things have not changed much since the rst contributions in this area: data integration is still responsible for solving many data quality problems in this overreaching big data challenge. While the focus until now has been much on quality of original data, not much stress has been dedicated to processed data and to metadata issues, which are critical during the data curation process. We thus consider the challenges reported in the mentioned works and remodel them into data quality-driven methods that are already implemented in a working integration framework.

3. Genomics data quality dimensions

The preliminary generation of genomic data follows guidelines and collections of best practices that are gathered during years of practitioners' experience; they are paired by metadata, describing the produced datasets. These are submitted to repositories or collected by consortia that coordinate big research projects and are appointed with the responsibility of publishing it on their platforms. Unfortunately, the integrated use of data coming from di erent data sources is very challenging, as heterogeneity is met at multiple stages of data extraction (e.g., download protocols, update poli-

cies), integration (e.g., conceptual arrangement, values and terminologies), and interlinking (e.g., references and annotation).

While integrating genomic datasets, either for *ad hoc* use in a research study, or for building long-lasting integrated data warehouses, we deal with various complexities that arise during three phases: i) download and retrieval of data from the (potentially multiple) sources; ii) transformation and manipulation, providing fully or partially structured data in machine-readable formats; iii) enrichment, improving the interoperability of datasets.

With heterogeneity locus we refer to an activity or phase within the genomic data production/integration process that exhibits heterogeneity issues, thus undermining the quality of resulting resources. Dividing production from integration, the taxonomy in Figure 2 keeps track of all the phases in which a genomic data user may need to resolve problems related to non-standardized ways of producing data, making it accessible, organizing it, or enhancing its interoperability. Issues may derive from diverse data and process management habits across different groups that work within the same institution; even more so across different ones. In Figure 2 the heterogeneity loci (listed in the central column) are grouped by production and integration phases (on the left) and are related to data quality dimensions (on the right) that are critical in the represented heterogeneity aspects and are described in the following subsections. In the figure, as in the remainder of the paper, we refer to widely used state-of-the-art definitions of data quality dimensions [119, 105] as well as to more recent ones [5, 9].

From the broader landscape of processed data sources presented in [14], in this review we focus on example applications that refer to a number of open general-purpose resources (Encyclopedia of DNA elements, ENCODE [48]; Roadmap Epigenomics Project [75]; 1000 Genomes Project [33]; Genotype-Tissue Expression Consortium, GTEx [77]; Genome Wide Association Study GWAS Catalog [22]), cancer genomics resources (The Cancer Genome Atlas, TCGA [121]; Genomic Data Commons, GDC [62]; International Cancer Genome Consortium, ICGC [126]), primary archives (Gene Expression Omnibus, GEO [8]), and annotation resources (GENCODE [51] and NCBI RefSeq [95]).

3.1. Accuracy and validity of generated content

Within production, datasets are generated and then published. Generation includes complex practices and challenges, involving quality issues related to accuracy i.e., the degree to which produced experimental data correctly and reliably describe real-world represented events [119]. Such aspects have been thoroughly reviewed in previous works [91, 64, 76]. Much less investigated, instead, are the issues related to metadata authoring (i.e., preliminary compilation of information) [93]. Until very recently, practitioners and investigators from the biomedical community have not recognized metadata creation as a first class activity in their work. As a consequence, accuracy of metadata values is negatively affected and it becomes very hard for many external users to work with it.

As publication paves the way to downstream opportunities for integration and analysis, a growing number of scientific journals require, upon submission, that genomic experimental data are contextually submitted to public data repositories [1] (such as GEO, SRA [73] or ArrayExpress [6]). Unfortunately, metadata instances in GEO repository suffer from redundancy, inconsistency, and incompleteness [124], especially due to a lightly regulated submission process. Users are allowed to create arbitrary fields that are not predefined by set dictionaries, many requested information are unstructured, and validity of the fields' values (i.e., the degree of their compliance with syntax format, type, range of the corresponding definitions [5]) is not checked. Information for submitting high-throughput sequencing data is listed at https://www.ncbi.nlm.nih.gov/geo/info/seq.html . A wide literature has been produced to capture structured information from GEO *a posteriori* (e.g., [100, 120]). The scenario of alternative repositories, i.e., NCBI BioSample [7] and EBI BioSamples [44], is witnessed in [59][1].

Once published on public repositories, data become available for a much wider community, they are potentially re-utilized in secondary analysis or integrated in other platforms; disorganization in the conveyance of provenance information and descriptions of generation procedures negatively affects `data lineage' [45].

3.2. Accessibility of open genomic data

Sources display diverse download options including programmatic interfaces (APIs), file transfer protocol (FTP) servers, and simple web interface links (HTTP or HTTPS). According to our analysis of important consortia housing open genomic data: i) ENCODE, GDC, and ICGC provide HTTPS API GET/POST services to retrieve lists of files corresponding to chosen filters and additional services to download the corresponding files one by one; ii) Roadmap Epigenomics, GENCODE, RefSeq, 1000 Genomes, and GWAS Catalog store all files on FTP servers, that can be navigated programmatically; iii) GEO provides a variety of methods (both through its own portal and from alternative interfaces), each concentrated on selected partitions of the entire repository content; iv) GTEx can only be accessed from its HTML website.

Only in some cases metadata information is structured and programmatically available. Sometimes metadata files are associated 1:1 to data files (i.e., each data file has a corresponding metadata file); in these cases they can be downloaded in similar ways as the corresponding data file (e.g., by just adding a parameter in an API call, as in ENCODE, or by calling a similar API endpoint using the same file identifier, as in GDC). In other cases, a single metadata file describes a collection of experiments (e.g., Roadmap Epigenomics) or

[1] This revealing analysis shows many insights, such as: i) in the description field the concept `age' appears in 33 different ways (e.g., AGE, Àge, age (yr-old), Age of patient, age_years); ii) 73% of Boolean metadata values are not actually true or false; iii) 26% of integer metadata values cannot be parsed into integers.

Figure 2: Taxonomy of heterogeneity loci and a ected data quality dimensions during genomic data integration; dimensions are equipped with the references where they are primarily de ned; they are further discussed in dedicated subsections. Pink rectangles are explained in Sections 4.1-4.4.

metadata information need to be retrieved in a number of di erent summary text les, where correspondence between a row and a genomic data le may be obtained using sample IDs (e.g., ICGC or 1000 Genomes). Other times sources have dedicated no e ort in systematizing metadata or bringing metadata to a single place; these can only be gathered from descriptions scattered across Web pages.

Accessibilitymeasures the ability of genomic data consumers to easily and quickly retrieve datasets [119]; it is a critical aspect in this phase, as very speci c modules need to be created for each source, often upon analysis of cumbersome online documentation and understanding of speci c parameters of each portal. Moreover, many well-known open-data databases (such as Cistrome [127], Broad Institute's CCLE [56], and COSMIC [116]) require authentication to access their data; these can only be downloaded and not re-distributed, creating a barrier to integration.

3.3. Currency of retrieved information

Measuring the extent to which data are up to date (the so-calledcurrency[105]) is not trivial, as le version synchronization between integrative solutions and original sources strictly depends on the information about the data update state made available in the speci c scenarios. The analyzed sources provide such kind of information in di erent ways: i) ENCODE, GDC and ICGC store information about last data update and checksums within their metadata; ii) Roadmap Epigenomics is a once-for-all project: it will not be updated (at least in the same distribution), therefore does not give such information; iii) 1000 Genomes organizes copies of its data, sequenced in di erent phases of the project, in different folders of the FTP update information can be inferred from the paths of the les; iv) GENCODE and RefSeq produce di erent versions regularly; they are associated to release dates, available in the folders names used on the FTP server; v) GTEx and GWAS Catalog embed the source data version (and subversion) within the le names (e.g., GTEx_v7_Annotations _SampleAttributesDS.txt or gwas_catalog_v1.0-studies_r2019-12-16).

On the contrary, metadata update information are available only in speci c cases. For example let us consider the case of ENCODE and GDC, which have complex hierarchical metadata structures in JSON. ENCODE centers its model on theExperimententity, includingBiosamples with many Replicates, to whichFiles belong. GDC is centered on the

Patient entity, providing multiple Samples; data are also divided by Project of a certain Tumor Type, for which many Data Types are given. These sources associate an update date to each JSON element representing an entity, such as Experiment, Treatment, Donor. The update date automatically pertains also to the elements contained in the entity (e.g., Experiment.assay, Donor.age, Treatment.pipeline...), allowing a fine-grained definition of last update of each single metadata unit. For all sources where files are downloaded from an FTP server, the upload date of files can be used as reference metadata update date.

3.4. Representational conciseness and consistency

Transformation is necessary to organize genomic data and the related descriptions into formats that allow conciseness and consistency in the representation of information. These dimensions respectively measure the ability to compactly, yet completely, represent data and the ability to present data in a same format, allowing backward compatibility [119]. When targeting further data manipulation and analysis, these requirements consequently translate into ease of operation, i.e. the extent to which data are easily used and customized [119].

Genomic data organization is a hard task because files have many formats with different semantics (e.g., expression matrices, sets of annotations, sets of peaks measured during an experiment or instead corresponding to a specific reference epigenome...). There does not exist any collectively accepted standard for a general yet basic data unit, that is able to concisely represent very heterogeneous input data types (given that rows and columns can express different conceptual entities and with different levels of detail).

Also metadata formats are various: hierarchical ones (such as JSON, XML, or equally expressive) adhere to in-house conceptual models; tab-delimited formats (TSV, CSV or Excel/Google Spreadsheets) present different semantics for rows and columns; completely unstructured metadata formats, collected from Web pages or other documentation provided by sources, need to be understood case by case.

3.5. Value consistency, uniqueness and specificity

Heterogeneity is present not only in representation formats, but also evident in values. Normalization activities may involve adding/standardizing genomic coordinates (e.g., from 0-based coordinates to 1-based or vice-versa) and other positional information, adding associated known genomic regions (e.g., genes, transcripts, miRNA) from standard nomenclatures, or formatting into general/source-specific formats, such as narrowPeak or broadPeak ENCODE's standards. A non-exhaustive list of commonly used genomic formats is found at https://genome.ucsc.edu/FAQ/FAQformat.html.

Also metadata that describe datasets in different sources are often incompatible or incomplete, using various reference ontologies or no terminology at all. The lack of consistency between value domains (i.e., no compliance with semantic rules defined over sets of values [9]) certainly hinders interoperability among sources.

Moreover, as the identity of genomic records is realized using descriptive fields in metadata usually in addition to internal identifiers metadata are in charge of handling uniqueness with respect to instances within a same source, ensuring that no exact duplicates exist for the same experimental data record [105]. Uniqueness is certainly a goal within single sources, while in the genomics domain (and biomedical more in general), it is accepted that entries representing same real-world entities are repeated in different sources, provided that linking references are present and records are aligned (as debated in [113]). This activity, improving lineage and interoperability of the database content, is very critical especially in an application field where resources are typically not well interlinked and information is only present in some databases and with different degrees of value specificity (referred to as level of detail in [105]).

3.6. Reliability of annotations

Annotation, i.e. structural and functional classification of (sub)sequences, is an across-the-board activity of the genomic data life cycle. Typically, annotating means associating genomic regions with labels from Gene Ontology [41] (explaining the related molecular function, biological process, cellular component) or with medical concepts related to the sequence (e.g., from UMLS [17]). The process is described in many works [47, 63, 64], hinting at the related data quality aspects. Annotations are either done by human experts, accurately based on literature evidence and certainly time consuming, or predicted automatically by algorithms that try to infer structural and functional information from similar genes/proteins (worst in terms of accuracy but much less time consuming).

Semantic annotation is instead a typical practice on metadata. As surveyed in Bodenreider [18], ontologies have been widely used in biomedical data management and integration for many years, with the main purpose of improving data interoperability [112]. Many tools are already available to allow semantic annotation with biomedical ontological concepts (see Annotator [71], EBI Zooma (https://www.ebi.ac.uk/spot/zooma/), NIH UMLS MetaMap [4], HeTop [61]). Techniques of text-mining [66] have been put into practice on many sources of biomedical text, including abstracts and experiment description from Gene Expression Omnibus [57, 28, 53], so far one of biggest yet least curated and standardized sources, thus drawing more attention and efforts. The problem of choosing the right ontologies for semantic enrichment is addressed in [96].

However, guidelines to achieve more standard annotation outcomes are still lacking. Reliability [119] of results (i.e., the extent to which annotations can be confidently used to connect and compare datasets) remains a critical aspect of annotation, being dependent on both the algorithm and the acceptance of the ontology in the biomedical community (which itself results from many factors, sometimes hard to measure).

4. Quality-aware methods for data integration

During the research activity documented in [14] we analyzed about 30 data repository hosts, consortia databases, platforms, and interfaces that integrate heterogeneous datasets. We performed various genomic data excavation sessions with the perspective goal of understanding the most important open data sources to be included in a rich processed data repository. Within this process, we experienced several cases of heterogeneity located in the specific loci depicted in Figure 2 (see pink rectangles of different sizes, marked with labels that characterize Sections 4.1-4.4), necessarily resulting into data quality problems. In the following discussion, we focus on the loci related to integration phases. For each, first we provide paradigmatic real-world instances. Then, we formalize the problem into overarching questions, specifying the data quality dimensions that are addressed at this stage (as listed in the previous section). Finally, we outline methods that are employed to resolve the issue, from literature and from integration efforts realized in the GeCo project.

4.1. Global repository synchronization with data sources

In the following we provide two example problems regarding data synchronization on the widely employed TCGA and ENCODE sources. Two additional examples, based on ICGC and 1000 Genomes, are available in the Section 1 of Supplementary material.

Example 1. Until 2016, TCGA data was available through a data portal that provided metadata only in XML format, using *biospecimen supplement* and *clinical supplement* that described respectively the biological samples analyzed in the experiments and the patient history, clinical information, and treatments. TCGA has undergone a transition towards the new GDC portal, where the data has been, by now, almost completely transferred. However, there are significant inconsistencies related to metadata. All supplements have been maintained and are still downloadable, but they nowhere fit in the new described data model, available at https://gdc.cancer.gov/developers/gdc-data-model-0. Instead, an entirely new collection of metadata, available through programmatic interface, has been defined, divided in four main endpoint groups (Project, Case, File, Annotation). The documentation of available fields is at https://docs.gdc.cancer.gov/API/Users_Guide/Appendix_A_Available_Fields/. GDC migration is still ongoing; nevertheless documentation is not consistently updated and it is common to find fields that are already visible in the interface facets (and APIs) but that indeed have null values for all instances in the database. Moreover, not all datasets that were available in the previous portal are now available in the new portal. For these reasons, synchronizing the content of an integrated repository with the one of GDC becomes very critical.

Example 2. ENCODE source elements in JSON schemata used for searching metadata through Elasticsearch (https://www.elastic.co/) are changed very often, as documented in about 90 Changelogs, one for each JSON entity corresponding to a profile. A complete list of ENCODE's data model entities (i.e., profiles) is at https://www.encodeproject.org/profiles/. However, metadata instances change also their values. For example, to keep track of the change of about 10 attribute-value pairs in the experiment ENCSR635OSC (https://www.encodeproject.org/experiments/ENCSR635OSC/), only a simple comment in the metadata was added (i.e., `"submitter_comment": "IMPORTANT! Biorep licate 2 was previously annotated as liver from a 4 year old female. It has now been corrected to be liver from a 32 year old adult male."`).

Problem formulation. How can changes on genomic data sources be taken into account to be reflected on integrated repositories, guaranteeing currency? How can it be done systematically, overcoming accessibility issues?

Method 1 Source partitioning. When targeting integrated systems up to date, the main difficulty is to identify data partitioning schemes specific for each source (as discussed in [13]); a partition can be repeatedly accessed and source files that are modified within the partition (or added to it) can be recognized, avoiding selectively the download of the source files that are not changed. Suppose we are interested in downloading a certain updated ENCODE portion (e.g., transcriptomics experiments on human tissue, aligned to reference genome hg19). We produce an API request to the endpoint https://www.encodeproject.org/matrix/, specifying the parameters `type = Experiment`, `replicates.library.biosample.donor.organism.scientific_name = Homo+sapiens`, `status = released`, `assembly = hg19`, and `assay_slims = Transcription`. In 1000 Genomes, as there are no API available, we instead navigate the FTP server directly and check the most updated release available on ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release; we consequently enter the relative folder (e.g., 20190312_biallelic_SNV_and_INDEL) and download all chromosomes files.

Method 2 Event-based update. We periodically check source websites and FTP servers for new data. We use a relational database (called `Importer_db` in the following) to manage the synchronization process between the data sources and our local repository. The `Sources` table has many `Datasets`, each of which corresponds to files (i.e., the genomic region data files). Each run of the download process checks unique properties of data files such as `URL`, `OriginLastUpdate`, `OriginSize` and `Hash`, used to compare the local copy of the file with the original one on the data source: new files are stored and processed; missing files (i.e., deprecated on the source) are copied to an archive; matching files that have identifying values different from the corresponding local values stored in the `Importer_db` are re-downloaded.

For different sources ad hoc software modules can be developed to periodically check for changes in the Changelog, schema definitions, documentation, in search for motivation to update our local copy of the source data. As an ex-

ample, GEO o ers to registered users a mechanism to be noti ed when new data is available, relevant for a search saved previously (https://www.ncbi.nlm.nih.gov/geo/info/faq.html#notifications). TCGA2BED [46] was realized to handle data acquisition and transformation for TCGA source and OpenGDC [26] provides an updated framework to acquire synchronized data also from GDC portal.

4.2. Orthogonal data and metadata organization

Examples. While there is common agreement on the terminology used to de ne genomic data types (e.g., mutation, copy number variation, chromatin accessibility), data types are typically not rendered using the same machine-readable formats (e.g., there exist both VCF-like and ICGC-like mutation formats (https://docs.icgc.org/submission/guide/icgc-simple-somatic-mutation-format/), gene expression data may be presented as sample/gene matrices or just as lists of genes with expression values per aliquot). Sometimes formats are de ned at experiment time to suit particular needs of the data; they are documented in plain text attachments. An example of format definition of ENCODE tsv les representing gene expression matrix is https://www.encodeproject.org/documents/c2bbcf04-9b9d-41aa-883f-bbba9bc45e68/. In this kind of documents, some speci cations may further confuse data organization, as matrix cell values are allowed to contain *ad hoc* formatting semantics (e.g., from an ENCODE format de nition document: The value in the cell contains two strings, one for TPM values and another for FPKM values, separated by underscore; each string contains values for each replicate, separated by colon.).

Many formats also fail at keeping representation levels orthogonal; for example, properties that represent values aggregated over a multitude of regions are sometimes displayed as part of single regions, repeated in each of them. In 1000 Genomes variation data, each line expresses one mutation and contains, as a property, the measure of allele frequencies across entire geographic populations (i.e., thousands of samples).

Additionally, data from speci c projects are simultaneously provided by di erent portals, that however re-shape it in several ways: ENCODE portal includes Roadmap Epigenomics data, re-processed using distinct pipelines and with completely di erent data schemata and metadata; TCGA data appears in both GDC and ICGC with very dissimilar representation both for data (one textual le for each aliquot from a patient, as opposed to one big spreadsheet containing independent lines, each connected to a patient) and for metadata.

Problem formulation. There is no agreement towards a basic genomic data unit for tertiary analysis. A common choice is to prepare one le for each experimental session; lines are genomic regions associated to some properties. Other times data units are huge matrices of patients or samples crossed with genes, miRNA, or other encoded sequences. Each source and each data type, thus, needs its own basic unit . Can genomic data be expressed using a unique model that is general enough to represent all analyzed formats (*concise and consistent representation*), and that also allows *ease of operation*?

Method Genomic Data Model and sample identity. The need for de ning a genomic basic data unit is emerging: a single piece of information that contains genomic regions with their properties and is identi able with an entity that is interesting for downstream analysis (e.g., a patient, a biological sample, a reference epigenome...). Any set of downloaded les with their input format should be convertible through a transformation relation into a set of genomic basic data units. We de ne as transformation relation cardinality the pair $X : Y$, where X is the cardinality of the set of les from the input source and Y is the cardinality of the output set of basic units into which the input is transformed for downstream use in an integrative system; X, Y is a fraction in lowest terms.

As a paradigm that more generally includes the interval-based genomic data representations (see BEDTools [103] and BEDOPS [94] for example), an interesting candidate for expressing such basic unit is represented by the sample of the Genomic Data Model (GDM, [84]). A sample can express heterogeneous DNA features, such as variations (e.g., a mutation in a given DNA position), peaks of binding or expression (i.e., genomic regions with higher read density), or structural properties of the DNA (e.g., break points, where the DNA is damaged, or junctions, where the DNA creates loops). GDM is based on the notion of *dataset*, i.e. a collection of samples. A sample, in turn, consists of two parts: the *region data*, describing the characteristics and DNA location of genomic features, and the *metadata*, describing general properties of the sample, in the form of key-value pairs; in GDM format there is one metadata le for each region data le.

Some sources provide a data le for each experimental event, for example ENCODE. In this case, the transformation has a 1:1 cardinality, i.e., to each ENCODE produced le, it corresponds one GDM sample. Other sources include more complex formats, such as MAF, VCF, and gene expression matrices. In these cases, the transformation phase takes care of compiling one single data le for each patient or univocally identi ed sample in the origin data. The transformation cardinality is thus 1:N, N being the number of patients or biological samples.

Metadata also feature diverse formats in the analyzed sources: i) hierarchical formats (JSON, XML, or equally expressive) require applying a attening procedure to create key-value pairs the key results from the concatenation of all JSON/XML elements from the root to the element corresponding to a value; ii) tab-delimited formats (TSV, CSV or Excel/Google Spreadsheets) strictly depend on the semantics of rows and columns (e.g., 1 row = 1 epigenome, 1 row = 1 biological sample) they often require pivoting tab-delimited columns into rows (which corresponds to creating key-value pairs); iii) two-columns tab-delimited formats (such as GEO's SOFT les) are translated into GDM straightforwardly; iv) completely unstructured metadata for-

Table 1
Census of 13 important data sources reporting for each: the processed data types that can be downloaded (along with metadata), their physical formats, and the semantic cardinality of the transformation relation with respect to the GDM output format [84].

Source	Data type	Data format, cardinality[1]	Metadata format, cardinality[1]
ENCODE	peaks	BED, 1:1	JSON, 1(experiment):#samples
	transcription	TSV, 1:1	JSON, 1(experiment):#samples
	transcription	GTF, 1:1	JSON, 1(experiment):#samples
GDC	mutations	MAF, 4:1	JSON, 4:1
	gene expression	TXT, 3:1	JSON, 3:1
	methylation, cnv, quanti cations	TXT, 1:1	JSON, 1:1
ICGC	mutation, methylation, miRNA/gene expression	TSV, 1:#donors	TSV, 1:#donors
Roadmap Epigenomics	peaks	BED, 1:1	Spreadsheet, 1:(#samples/#epigenomes)[2]
	transcription	TSV, 2:#epigenomes	Spreadsheet, 1:(#samples/#epigenomes)[2]
GENCODE	annotations	GTF, 1:#annotation_types	region data le + webpage, X:1[3]
RefSeq	annotations	GFF, 1:#annotation_types	region data le + webpage, X:1[3]
1000 Genomes	variation	VCF, 23:#individuals	TSV, 4:#individuals
GEO	expression	BED, 1:1	HTML/SOFT, 1:# les_from_sample
GTEx	expression	GCT, 1:#donors	TXT, 1:#donors
GWAS Catalog	associations	TSV, 1:1	-
CISTROME	peaks	BED, 1:1	TSV, 1:1
CCLE	various	GCT/TXT, 1:#cell_lines	TXT, 1:#cell_lines
COSMIC	various	TSV, 1:#individuals	TSV, X:1[3]

[1] Expressed as X : Y; this ratio represents the number X of data (resp. metadata) units used in the origin source to compose Y data (resp. metadata) le(s) in GDM format. [2] Each reference epigenome is used for many data types, thus many GDM samples. The same epigenome-related metadata is replicated into many samples. [3] In these cases it is di cult to build a numerical relation many meta are retrieved from the data les themselves, in addition to manually curated information.

mats, collected from Web pages or other documentation provided by sources, need case-speci c manual processing.

Table 1 shows transformation relation cardinalities regarding both data and metadata input formats, targeting the GDM output format. We analyzed di erent data types in a number of important data sources, that possibly include les with di erent formats. Note that, while for descriptive purposes we indicate physical formats (e.g., TSV, TXT, JSON), the indication of cardinalities also embeds a semantic information: how many data units are represented in one le. Following the mapping from input sources into GDM format we can solve systematically the heterogeneity of data formats and prepare the GDM datasets as sets of GDM samples that are uniform in their schema. The Supplementary material (Section 2) provides additional details on this method.

4.3. Metadata interoperability

Examples. Metadata heterogeneity can also be analyzed from other perspectives. From a schema point of view (i.e., how each piece of information is identi ed and interrelated with others), when searching for disease-related attributes, we nd diverse possibilities: Disease type in GDC, Characteristics tissue in GEO, Health status in ENCODE. From the values point of view, when searching for breast cancer-related information, we nd multiple expressions, pointing to comparable samples, e.g., Breast Invasive Carcinoma (GDC), breast cancer ductal carcinoma (GEO), Breast cancer (adenocarcinoma) (ENCODE).

Roadmap Epigenomics expresses ages of samples using a unique column AGE (Post Birth in YEARS/Fetal in GESTATIONAL WEEKS/CELL LINE CL) together for three di erent kind of classes and, consequently, with three different measure units. Example values for single instances are Fetus (GW unknown) , CL , Unknown, Unknown, 45Y , or 49Y, 59Y, 41Y, 25Y, 81Y , where 5 values are put together to express that the related epigenome is derived from 5 individuals. Also the information about donors and its interrelation with other attributes is confusing. The column Single Donor (SD) / Composite (C) discriminates between epigenomes deriving from one or more donors. Yet, at the column DONOR / SAMPLE ALIAS (containing identi ers), it happens that an epigenome labeled as deriving from a single donor, contains instead multiple IDs. No further explanation is available to clarify the semantics; other dependent columns such as sex and ethnicity become also unclear. Paradoxically, one donor, identi ed by the string H-22772, turns out to be present in two di erent epigenomes, one derived from lung and one from heart tissue. This problem is easily propagated to other sources, as Roadmap Epigenomics experiments are replicated in ENCODE repository. Here, di erent experiments present the same external tags (e.g., roadmap-epigenomics:UW H22772).

Problem formulation. How can metadata be employed to provide a global and integrated view of data sources (`representational consistency')? Does such representation

Figure 3: GeCo integration process, from the download of a source partition (based on the prior de nition of a GDM dataset), to its transformation and all following phases. Cleaning, mapping, enriching and checking are only performed on the metadata, while data are left unchanged. Metadata are attened from the GCM relational implementation back to the le-based GeCo repository. Data and corresponding metadata of each dataset are loaded into a cloud-based data engine for queries on genomic region and metadata [83].

help users in querying data straightforwardly (ease of operation)?

Method Genomic Conceptual Model for metadata normalization. In literature there are works that use conceptual modeling to better explain relations between biological entities [63, 106, 97]. However, conceptual modeling can serve brilliantly also the purpose of organizing metadata from heterogeneous sources into one global view. The Genomic Conceptual Model (GCM, [15]) is an Entity-Relationship model used to describe metadata of genomic data sources. The main objective of GCM is to recognize a common set of concepts (about 40) that are semantically supported by most genomic data sources, although with very di erent syntax and forms. GCM is a star-schema inspired to classic data marts [19] centered around the REGION entity, representing a genomic basic data unit, such as the GDM elementary sample. The four dimensions of the star describe the biology of the experiment, the used technology, its management aspects, and the extraction parameters for internal organization of items. A complete integration framework (described in [13]) can be employed to download, transform, clean and integrate metadata at the schema level, importing them into the relational database gem_db that implements the GCM physically. Data constraints checks (name existence and value dependencies in [15]) are performed based on a set of manually introduced rules, but also on automatically generated ones, inspired by the works on data cleaning using association rules mining [3] and much in the fashion of [82], who uses rules as a means to generate recommendations for suitable metadata additions to datasets. The conceptual representation of GCM widely helped domain users in nding data more easily from a unique query interface, without having to deal with heterogeneous access points, metadata formats and models, as demonstrated in [10].

4.4. Large-scale dataset interoperability

Example on data. Within ICGC gene annotation is not consistent among di erent datatypes (e.g., sequence-based gene expression datasets use Ensembl Gene IDs [125], like ENCODE gene quanti cation data and TCGA gene expression quanti cation, while array-based gene expression datasets use the gene name convention of HGNC [123]). Within annotation databases themselves, data may be incomplete. For example in GENCODE's comprehensive gene annotation les (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/) not all exons and transcripts regions have a corresponding gene region that includes them. While searching for correct coordinates of a gene, users may alternatively calculate the start as the one of its left-most transcript/exon and the stop as the one of its right-most transcript/exon, but this procedure could be not always accurate. Such shortcomings are consequently propagated in all processed data sources where the reference gene annotation is used to codify signals data (e.g., ENCODE). Furthermore, secondary sources use di erent releases versions to annotate di erent les (to date, GENCODE has 34 releases, out of which only 6 are still maintained for the new GRCh38 assembly). This makes it hard to consistently compare les from a same source that have been annotated using di erent reference sets.

Example on metadata. Metadata are a ected by the even more complicated issue of ontology misalignment. Ontology CL [86] and EFO [79] reference same concepts: the specific instances in the two ontologies have di erences in the values and schema. NCIT [42] and UBERON [92], both including parts of the human body, also show inconsistencies: while hypothalamus is considered a synonym of BRAIN in NCIT, it is a sub-concept of brain in UBERON (ve levels more speci c, traversing both relationships of sub-

sumption (is_a) and containment (part_of).

Using ontologies as a base for further semantic annotation, many algorithms still produce a relevant number of inaccurate annotations (see [54, 28]), which result in harder work for the downstream integration process.

Problem formulation. How can datasets understand each other? Can we normalize data with respect to commonly adopted terminologies (consistency, `specificity' and `uniqueness') and confidently exploit the currently available external resources (reliability')?

Method 1 Data enrichment. A fruitful approach with annotation is the inclusive one: integrators may add as many information as possible, considering the most accepted resources in the field. For structural and functional annotation of genomic regions and sequences, including adding for example gene/transcript/exon identifiers or biological process related to a protein, multiple reference databases may be queried (RefSeq, GENCODE, Ensembl, Entrez [78], HGNC), as documented in TCGA2BED [46] and OpenGDC [26], or performed during the integration process of several datasets from Roadmap Epigenomics and transcriptomics data from ENCODE. Large-scale data integration in genomics can be achieved using cross-references (see [55]); its success strictly depends on a correct use of persistent identifiers [85]. See the Supplementary material Section 3 for more details on this method.

Method 2 Metadata enrichment. The process of annotating existing structured metadata with ontological terms, their definitions, synonyms, ancestors, and descendants can be done in an iterative way, automated with respect to the querying of online annotation systems and semantic match computation, but also assisted by an expert manually checking the obtained links [12]. This enhancement of metadata (using specialized biomedical ontologies) can be seen as the construction of a knowledge graph of the content of the repository [11]; it is useful to instrument the search of datasets described by such metadata in a semantically enriched fashion (see GenoSurf interface [25]). See the Supplementary material Section 4 for more details on this method.

5. Discussion and outlook

The integration of data and metadata is of growing relevance in biomedical fields (including genomics), because critical decisions in the domains of health-care such as precision medicine depend on it. As individualized predictions become more difficult, they require approaches that combine multiple sources and multiple data types (from genomics, transcriptomics, epigenomics, etc.), possibly completed with clinical data. Heterogeneity aspects affect many actors and stages of the data life cycle. In such situations, data quality dimensions can adequately lead the analysis of problems and related solutions. We have reviewed works that have contributed to data quality-driven approaches in genomics; even with community-driven approaches that propose on-the-fly data integration, the focus has so far been on quality of origin data sources and not so much on the overall process that channels data together for subsequent use. Thus, we have introduced a novel perspective: we have shown a taxonomy of integration phases that directly impact quality of genomic databases and interfaces during data integration; we have detailed the issues related to such phases, providing examples, questions to be addressed, and methods that we experimented during the creation of a repository of high quality, which inspired the discussions of this review paper.

The repository, currently with more than 250k processed items, results from the GeCo project effort. Figure 3 shows the sequential software modules (https://github.com/DEIB-GECO/Metadata-Manager) to integrate genomic sources, by solving all the analyzed heterogeneity aspects. Phases are recorded in the importer_db: a given dataset is downloaded and periodically synchronized with the origin source, transformed into the GDM format (achieving orthogonal data/metadata organization); metadata are cleaned, simplifying redundant attribute names, mapped into the gmql_db (unique conceptual representation, towards interoperability of metadata), semantically enriched and checked with respected to constraints. The relational representation is attuned to load datasets into a file-based engine for further biological querying [83].

While the described approaches have been successfully implemented in practical contexts [13], future challenges include applying the proposed solutions to complex contexts such as the one of clinical data and translational medicine, that ultimately will need to be also iterated with genomic data. We are aware of important work that is being conducted in parallel on health data [40, 72, 24, 65, 38], also employing the data warehouse paradigm as a guarantor of up-to-date de-duplicated data within a public network of research centers [37], usually oriented to support analytics [39]. Several works already address data quality for precision medicine [36, 43, 97], revieweing the use of genomic data in the medical context, whereas my review is focused primrily on issues of quality in genomic data integration (data comparability, metadata definitions, data standards, ...) encompassing all possible uses of genomic data.

In this review, we have shown how resolving quality while building a data repository can effectively create usable integrated environments for researchers. Since many of the described approaches may be useful for other researchers even in dynamic data integration assets these will be provided through convenient external programmatic access. Starting from this baseline, we envision a data integration process that includes seamlessly evaluation of quality parameters, towards data and information that are more directly employable in genomic analysis and biological discovery. Predictably, future data integration approaches will include more and more a data quality-aware *modus operandi* with the following characteristics: i) currency-driven synchronization of sources, ii) concise/orthogonal/common data representations, iii) light and interoperable data descriptions, iv) reliability-tailored dataset linkage. All in all, this review

highlights trends in genomic data and information integration, which will ideally guide and improve future e orts and activities.

6. Funding

This work was supported by the European Research Council Executive Agency under the EU Framework Programme Horizon 2020, ERC Advanced Grant number 693174 GeCo (data-driven Genomic Computing).

7. Acknowledgements

The author would like to thank Professor Stefano Ceri and Professor Cinzia Cappiello for fruitful discussions and for providing precious suggestions and inspiration during the preparation of the manuscript.

8. Con icts of interest

The author declares no con ict of interest.

References

[1] , 2002. Microarray standards at last. Nature 419, 323. URL: https://doi.org/10.1038/419323a .

[2] Apiletti, D., Bruno, G., Ficarra, E., Baralis, E., 2006. Data cleaning and semantic improvement in biological databases. Journal of Integrative Bioinformatics 3, 219 229.

[3] Apiletti, D., Bruno, G., Ficarra, E., Baralis, E., 2009. Extraction of constraints from biological data, in: Biomedical Data and Applications. Springer, pp. 169 186.

[4] Aronson, A.R., Lang, F.M., 2010. An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17, 229 236.

[5] Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G., Schwarzenbach, J., 2013. The six primary dimensions for data quality assessment. DAMA UK Working Group , 432 435URL: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf online (Jan. 12th, 2021, date last accessed).

[6] Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N.A., Petryszak, R., Papatheodorou, I., et al., 2018. ArrayExpress update from bulk to single-cell expression data. Nucleic acids research 47, D711 D715.

[7] Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T., et al., 2011. Bioproject and biosample databases at ncbi: facilitating capture and organization of metadata. Nucleic acids research 40, D57 D63.

[8] Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al., 2012. Ncbi geo: archive for functional genomics data sets update. Nucleic acids research 41, D991 D995.

[9] Batini, C., Scannapieco, M., 2016. Data and Information Quality: Dimensions, Principles and Techniques. 1st ed., Springer Publishing Company, Incorporated.

[10] Bernasconi, A., Canakoglu, A., Ceri, S., 2019a. Exploiting conceptual modeling for searching genomic metadata: A quantitative and qualitative empirical study, in: Guizzardi, G., Gailly, F., Suzana Pitangueira Maciel, R. (Eds.), Advances in Conceptual Modeling, Springer International Publishing, Cham. pp. 83 94.

[11] Bernasconi, A., Canakoglu, A., Ceri, S., 2019b. From a conceptual model to a knowledge graph for genomic datasets, in: Laender, A.H.F., Pernici, B., Lim, E.P., de Oliveira, J.P.M. (Eds.), Conceptual Modeling, Springer International Publishing, Cham. pp. 352 360.

[12] Bernasconi, A., Canakoglu, A., Colombo, A., Ceri, S., 2018. Ontology-driven metadata enrichment for genomic datasets, in: Baker, C.J.O., Waagmeester, A., Splendiani, A., Beyan, O.D., Marshall, M.S. (Eds.), International Conference on Semantic Web Applications and Tools for Life Sciences.

[13] Bernasconi, A., Canakoglu, A., Masseroli, M., Ceri, S., 2020a. META-BASE: a novel architecture for large-scale genomic metadata integration. IEEE/ACM Transactions on Computational Biology and Bioinformatics https://doi.org/10.1109/TCBB.2020.2998954 .

[14] Bernasconi, A., Canakoglu, A., Masseroli, M., Ceri, S., 2020b. The road towards data integration in human genomics: players, steps and interactions. Brie ngs in Bioinformatics https://doi.org/10.1093/bib/bbaa080.

[15] Bernasconi, A., Ceri, S., Campi, A., Masseroli, M., 2017. Conceptual modeling for genomics: Building an integrated repository of open data, in: Mayr, H.C., Guizzardi, G., Ma, H., Pastor, O. (Eds.), Conceptual Modeling, Springer International Publishing, Cham. pp. 325 339.

[16] Berti-Équille, L., Moussouni, F., 2005. Quality-aware integration and warehousing of genomic data, in: Proceedings of the 2005 International Conference on Information Quality (MIT ICIQ Conference), Sponsored by Lockheed Martin, MIT, Cambridge, MA, USA, November 10-12, 2006.

[17] Bodenreider, O., 2004. The uni ed medical language system (umls): integrating biomedical terminology. Nucleic acids research 32, D267 D270.

[18] Bodenreider, O., 2008. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of Medical Informatics , 67.

[19] Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., Paraboschi, S., 2001. Designing data marts for data warehouses. ACM Transactions on Software Engineering and Methodology 10, 452 483.

[20] Bouadjenek, M.R., Verspoor, K., Zobel, J., 2017a. Automated detection of records in biological sequence databases that are inconsistent with the literature. Journal of biomedical informatics 71, 229 240.

[21] Bouadjenek, M.R., Verspoor, K., Zobel, J., 2017b. Literature consistency of bioinformatics sequence databases is e ective for assessing record quality. Database: The Journal of Biological Databases and Curation 2017.

[22] Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al., 2018. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research 47, D1005 D1012.

[23] Bursteinas, B., Britto, R., Bely, B., Auchincloss, A., Rivoire, C., Redaschi, N., O'Donovan, C., Martin, M.J., 2016. Minimizing proteome redundancy in the uniprot knowledgebase. Database: The Journal of Biological Databases and Curation 2016.

[24] Callahan, T.J., Bauck, A.E., Bertoch, D., Brown, J., Khare, R., Ryan, P.B., Staab, J., Zozus, M.N., Kahn, M.G., 2017. A comparison of data quality assessment checks in six data sharing networks. eGEMs (Generating Evidence & Methods to improve patient outcomes) 5.

[25] Canakoglu, A., Bernasconi, A., Colombo, A., Masseroli, M., Ceri, S., 2019. Genosurf: metadata driven semantic search system for integrated genomic datasets. Database: The Journal of Biological Databases and Curation 2019.

[26] Cappelli, E., Cumbo, F., Bernasconi, A., Canakoglu, A., Ceri, S., Masseroli, M., Weitschek, E., 2020. OpenGDC: Unifying, Modeling, Integrating Cancer Genomic Data and Clinical Metadata. Applied Sciences 10, 6367.

[27] Ceri, S., Bernasconi, A., Canakoglu, A., Gulino, A., Kaitoua, A., Masseroli, M., Nanni, L., Pinoli, P., 2017. Overview of geco: A project for exploring and integrating signals from the genome, in: International Conference on Data Analytics and Management in Data Intensive Domains, Springer. pp. 46 57.

[28] Chen, G., Ramírez, J.C., Deng, N., Qiu, X., Wu, C., Zheng, W.J.,

Wu, H., 2019a. Restructured geo: restructuring gene expression omnibus metadata for genome dynamics analysis. Database: The Journal of Biological Databases and Curation 2019.

[29] Chen, Q., Britto, R., Erill, I., Je ery, C.J., Liberzon, A., Magrane, M., Onami, J.i., Robinson-Rechavi, M., Sponarova, J., Zobel, J., et al., 2019b. Quality matters: Biocuration experts on the impact of duplication and other data quality issues in biological databases. bioRxiv , 788034.

[30] Chen, Q., Wan, Y., Zhang, X., Lei, Y., Zobel, J., Verspoor, K., 2018. Comparative analysis of sequence clustering methods for deduplication of biological databases. Journal of Data and Information Quality (JDIQ) 9, 1 27.

[31] Chen, Q., Zobel, J., Verspoor, K., 2017a. Benchmarks for measurement of duplicate detection methods in nucleotide databases. Database: The Journal of Biological Databases and Curation 2017.

[32] Chen, Q., Zobel, J., Verspoor, K., 2017b. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. Database: The Journal of Biological Databases and Curation 2017.

[33] Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M., et al., 2012. The 1000 genomes project: data management and community access. Nature methods 9, 459.

[34] Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kau , F., Wilczynski, B., et al., 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422 1423.

[35] Cohen-Boulakia, S., Biton, O., Davidson, S., Froidevaux, C., 2007. Bioguidesrs: querying multiple sources with a user-centric perspective. Bioinformatics 23, 1301 1303.

[36] Hulsen, T., Jamuar, S.S., Moody, A.R., Karnes, J.H., Varga, O., Hedensted, S., Sprea co, R., Ha er, D.A., McKinney, E.F., 2019. From big data to precision medicine. Frontiers in Medicine 6, 34.

[37] Ross, T.R., Ng, D., Brown, J.S., Pardee, R., Hornbrook, M.C., Hart, G., Steiner, J.F., 2014. The hmo research network virtual data warehouse: a public data model to support collaboration. eGEMs (Generating Evidence & Methods to improve patient outcomes) 2.

[38] Savitz, S.T., Savitz, L.A., Fleming, N.S., Shah, N.D., Go, A.S., 2020. How much can we trust electronic health record data?, in: Healthcare, Elsevier. p. 100444.

[39] Spengler, H., Gatz, I., Kohlmayer, F., Kuhn, K.A., Prasser, F., 2020. Improving data quality in medical research: A monitoring architecture for clinical and translational data warehouses, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE. pp. 415 420.

[40] Weiskopf, N.G., Weng, C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association 20, 144 151.

[41] Consortium, G.O., 2019. The gene ontology resource: 20 years and still going strong. Nucleic acids research 47, D330 D338.

[42] de Coronado, S., Wright, L.W., Fragoso, G., Haber, M.W., Hahn-Dantona, E.A., Hartel, F.W., Quan, S.L., Safran, T., Thomas, N., Whiteman, L., 2009. The nci thesaurus quality assurance life cycle. Journal of biomedical informatics 42, 530 539.

[43] Cruz Correia, R., Ferreira, D., Bacelar, G., Marques, P., Maranhão, P., 2018. Personalised medicine challenges: quality of data. International Journal of Data Science and Analytics 6, 251 259.

[44] Courtot, M., Cherubin, L., Faulconbridge, A., Vaughan, D., Green, M., Richardson, D., Harrison, P., Whetzel, P.L., Parkinson, H., Burdett, T., 2018. Biosamples database: an updated sample metadata hub. Nucleic acids research 47, D1172 D1178.

[45] Cui, Y., Widom, J., Wiener, J.L., 2000. Tracing the lineage of view data in a warehousing environment. ACM Transactions on Database Systems (TODS) 25, 179 227.

[46] Cumbo, F., Fiscon, G., Ceri, S., Masseroli, M., Weitschek, E., 2017. TCGA2BED: extracting, extending, integrating, and querying the cancer genome atlas. BMC Bioinformatics 18, 6.

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

[48] ENCODE, C., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57 74.

[49] Etcheverry, L., Marotta, A., Ruggia, R., 2010. Data quality metrics for genome wide association studies, in: 2010 Workshops on Database and Expert Systems Applications, IEEE. pp. 105 109.

[50] Fan, W., 2015. Data quality: From theory to practice. Acm Sigmod Record 44, 7 18.

[51] Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al., 2018. GENCODE reference annotation for the human and mouse genomes. Nucleic acids research 47, D766 D773.

[52] Gabdank, I., Chan, E.T., Davidson, J.M., Hilton, J.A., Davis, C.A., Baymuradov, U.K., Narayanan, A., Onate, K.C., Graham, K., Miyasato, S.R., et al., 2018. Prevention of data duplication for high throughput sequencing repositories. Database 2018, bay008.

[53] Galeota, E., Kishore, K., Pelizzola, M., 2020. Ontology-driven integrative analysis of omics data through onassis. Scienti c Reports 10, 1 9.

[54] Galeota, E., Pelizzola, M., 2017. Ontology-based annotations and semantic relations in large-scale (epi) genomics data. Brie ngs in bioinformatics 18, 403 412.

[55] Gasteiger, E., Jung, E., Bairoch, A.M., 2001. Swiss-prot: connecting biomolecular knowledge via a protein database. Current issues in molecular biology 3, 47 55.

[56] Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al., 2019. Next-generation characterization of the cancer cell line encyclopedia. Nature 569, 503.

[57] Giles, C.B., Brown, C.A., Ripperger, M., Dennis, Z., Roopnarinesingh, X., Porter, H., Perz, A., Wren, J.D., 2017. Ale: automated label extraction from geo metadata. BMC Bioinformatics 18, 509.

[58] Goble, C., Stevens, R., Hull, D., Wolstencroft, K., Lopez, R., 2008. Data curation+ process curation= data integration+ science. Briefings in bioinformatics 9, 506 517.

[59] Gonçalves, R.S., Musen, M.A., 2019. The variable quality of metadata about biological samples used in biomedical experiments. Scienti c data 6, 190021.

[60] Gray, A.J., Goble, C.A., Jimenez, R., 2017. Bioschemas: From potato salad to protein annotation., in: International Semantic Web Conference (Posters, Demos & Industry Tracks).

[61] Grosjean, J., Merabti, T., Dahamna, B., Kergourlay, I., Thirion, B., Soualmia, L.F., Darmoni, S.J., et al., 2011. Health multiterminology portal: a semantic added-value for patient safety. Stud Health Technol Inform 166, 129 138.

[62] Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., Staudt, L.M., 2016. Toward a shared vision for cancer genomic data. New England Journal of Medicine 375, 1109 1112.

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

[64] Hedeler, C., Missier, P., 2008. Information quality management challenges for high-throughput data. Biological Database Modeling , 81.

[65] Henley-Smith, S., Boyle, D., Gray, K., 2019. Improving a secondary use health data warehouse: Proposing a multi-level data quality framework. eGEMs (Generating Evidence & Methods to improve patient outcomes) 7.

[66] Huang, C.C., Lu, Z., 2016. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brie ngs in bioinformatics 17, 132 144.

[67] Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Car-

valho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al., 2015. Orchestrating high-throughput genomic analysis with bioconductor. Nature methods 12, 115 121.

[68] Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Menager, H., Kala², M., Schwämmle, V., Grüning, B., Beard, N., Lopez, R., et al., 2019. The bio.tools registry of software tools and data resources for the life sciences. Genome biology 20, 1 4.

[69] Ison, J., Kala², M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., Rice, P., 2013. Edam: an ontology of bioinformatics operations, types of data and identi ers, topics and formats. Bioinformatics 29, 1325 1332.

[70] Ji, H., Davis, R.W., 2006. Data quality in genomics and microarrays. Nature biotechnology 24, 1112 1113.

[71] Jonquet, C., Shah, N.H., Musen, M.A., 2009. The open biomedical annotator. Summit on translational bioinformatics 2009, 56.

[72] Kahn, M.G., Callahan, T.J., Barnard, J., Bauck, A.E., Brown, J., Davidson, B.N., Estiri, H., Goerg, C., Holve, E., Johnson, S.G., et al., 2016. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMs (Generating Evidence & Methods to improve patient outcomes) 4.

[73] Kodama, Y., Shumway, M., Leinonen, R., 2011. The sequence read archive: explosive growth of sequencing data. Nucleic acids research 40, D54 D56.

[74] Koh, J.L., Lee, M.L., Khan, A.M., Tan, P.T., Brusic, V., 2004. Duplicate detection in biological data using association rule mining, in: Proceedings of the second European Workshop on Data Mining and text Mining in Bioinformatics, Citeseer. pp. 35 41.

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

[76] León, A., Reyes, J., Burriel, V., Valverde, F., 2016. Data quality problems when integrating genomic information, in: International Conference on Conceptual Modeling, Springer. pp. 173 182.

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

[78] Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T., 2010. Entrez gene: gene-centered information at ncbi. Nucleic acids research 39, D52 D57.

[79] Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., Parkinson, H., 2010. Modeling sample variables with an experimental factor ontology. Bioinformatics 26, 1112 1118.

[80] Martinez, A., Hammer, J., 2005. Making quality count in biological data sources, in: Proceedings of the 2nd international workshop on Information quality in information systems, pp. 16 27.

[81] Martinez, A., Hammer, J., Ranka, S., 2008. Biodq: data quality estimation and management for genomics databases, in: International Symposium on Bioinformatics Research and Applications, Springer. pp. 469 480.

[82] Martínez-Romero, M., O'Connor, M.J., Egyedi, A.L., Willrett, D., Hardi, J., Graybeal, J., Musen, M.A., 2019. Using association rule mining and ontologies to generate metadata recommendations from multiple biomedical databases. Database: The Journal of Biological Databases and Curation 2019.

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

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

[85] McMurry, J.A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., Courtot, M., Deck, J., Dumontier, M., Fellows, D.K., et al., 2017. Identi ers for the 21st century: How to design, provision, and reuse persistent identi ers to maximize utility and impact of life science data. PLoS biology 15.

[86] Meehan, T.F., Masci, A.M., Abdulla, A., Cowell, L.G., Blake, J.A., Mungall, C.J., Diehl, A.D., 2011. Logical development of the cell ontology. BMC Bioinformatics 12, 6.

[87] Missier, P., Embury, S., Greenwood, M., Preece, A., Jin, B., 2006. Quality views: capturing and exploiting the user perspective on data quality, in: Proceedings of the International Conference on Very Large Data Bases, pp. 977 988.

[88] Moussouni, F., Berti-Équille, L., le Développement, F., 2013. Cleaning, integrating, and warehousing genomic data from biomedical resources. Elloumi M., Zomaya AY, editors.(Hoboken, NJ: John Wiley and Sons, Inc.) , 35 58.

[89] Moussouni, F., Berti-Equille, L., Rozé, G., Loréal, O., Guérin, E., 2007. Qdex: a database pro ler for generic bio-data exploration and quality aware integration, in: International Conference on Web Information Systems Engineering, Springer. pp. 5 16.

[90] Müller, H., Freytag, J.C., Leser, U., 2012. Improving data quality by source analysis. Journal of Data and Information Quality (JDIQ) 2, 1 38.

[91] Müller, H., Naumann, F., 2003. Data quality in genome databases, in: Eighth International Conference on Information Quality (ICIQ 2003), November 7-9. 2003, pp. 269 284.

[92] Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., Haendel, M.A., 2012. Uberon, an integrative multi-species anatomy ontology. Genome biology 13, R5.

[93] Musen, M.A., Sansone, S.A., Cheung, K.H., Kleinstein, S.H., Crafts, M., Schürer, S.C., Graybeal, J., 2018. Cedar: Semantic web technology to support open science, in: Companion Proceedings of the The Web Conference 2018, International World Wide Web Conferences Steering Committee. pp. 427 428.

[94] Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al., 2012. Bedops: high-performance genomic feature operations. Bioinformatics 28, 1919 1920.

[95] O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al., 2015. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research 44, D733 D745.

[96] Oliveira, D., Butt, A.S., Haller, A., Rebholz-Schuhmann, D., Sahay, R., 2019. Where to search top-k biomedical ontologies? Brie ngs in bioinformatics 20, 1477 1491.

[97] Palacio, A.L., López, Ó.P., Ródenas, J.C.C., 2018. A method to identify relevant genome data: conceptual modeling for the medicine of precision, in: International Conference on Conceptual Modeling, Springer. pp. 597 609.

[98] Park, Y.M., Squizzato, S., Buso, N., Gur, T., Lopez, R., 2017. The ebi search engine: Ebi search as a service making biological data accessible for all. Nucleic acids research 45, W545 W549.

[99] Pastor, O., León, A.P., Reyes, J.F.R., García, A.S., Casamayor, J.C.R., 2020. Using conceptual modeling to improve genome data management. Brie ngs in Bioinformatics https://doi.org/10.1093/bib/bbaa100.

[100] Posch, L., Panahiazar, M., Dumontier, M., Gevaert, O., 2016. Predicting structured metadata from unstructured metadata. Database: The Journal of Biological Databases and Curation 2016.

[101] Poux, S., Arighi, C.N., Magrane, M., Bateman, A., Wei, C.H., Lu, Z., Boutet, E., Bye-A-Jee, H., Famiglietti, M.L., Roechert, B., et al., 2017. On expert curation and scalability: Uniprotkb/swiss-prot as a case study. Bioinformatics 33, 3454 3460.

[102] Poux, S., Magrane, M., Arighi, C.N., Bridge, A., O'Donovan, C., Laiho, K., Consortium, U., et al., 2014. Expert curation in uniprotkb: a case study on dealing with con icting and erroneous data. Database: The Journal of Biological Databases and Curation 2014.

[103] Quinlan, A.R., Hall, I.M., 2010. Bedtools: a exible suite of utilities for comparing genomic features. Bioinformatics 26, 841 842.

[104] Rajan, N.S., Gouripeddi, R., Mo, P., Madsen, R.K., Facelli, J.C., 2019. Towards a content agnostic computable knowledge repository