Briefings in Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Review Paper

Omics Data Management

The road towards data integration in human genomics: players, steps and interactions

Anna Bernasconi [1],*, Arif Canakoglu [1], Marco Masseroli [1] and Stefano Ceri [1]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, 20133, Italy.

*Corresponding author: Tel.: +39-02-2399-3655; Fax: +39-02-2399-3411; E-mail: anna.bernasconi@polimi.it

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Thousands of new experimental datasets are becoming available every day; in many cases they are produced within the scope of large cooperative efforts, involving a variety of laboratories spread all over the world, and typically open for public use. Although the potential collective amount of available information is huge, the effective combination of such public sources is hindered by data heterogeneity, as the datasets exhibit a wide variety of notations and formats, concerning both experimental values and metadata. Thus, data integration is becoming a fundamental activity, to be performed prior to data analysis and biological knowledge discovery, consisting of subsequent steps of data extraction, normalization, matching and enrichment; once applied to heterogeneous data sources, it builds multiple perspectives over the genome, leading to the identification of meaningful relationships that could not be perceived by using incompatible data formats.

In this paper, we first describe a technological pipeline from data production to data integration; we then propose a taxonomy of genomic data players (based on the distinction between *contributors*, *repository hosts*, *consortia*, *integrators*, and *consumers*), and apply the taxonomy to describe about thirty important players in genomic data management. We specifically focus on the integrator players and analyze the issues in solving the genomic data integration challenges, as well as evaluate the computational environments that they provide to follow up data integration by means of visualization and analysis tools.

Keywords: data integration, genomics, metadata, interoperability, genomic databases, bio-ontologies.

1 Introduction

In recent years, benefiting from high-throughput technologies [1], increasing amounts of genomic data of multiple types – deriving from microarray, next-generation sequencing, or single-cell technologies – have become widely available. Gene expression, mutation and variation, transcriptome analysis, chromatin immunoprecipitation sequencing, are only some of the heterogeneous types of data that genomic researchers use and combine in their everyday work.

Genomic data includes *experimental observations*, representing genomic sequences (in raw stages) or regions with their properties (in processed stages) and *metadata*, carrying information about the biological phenomena observed and the performed experiment (i.e., biological material, preparation, donor, etc.), the associated clinical elements, the used technology and assay, and the management aspects, such as case studies and projects/organizations behind data production. As genomic datasets originated from disparate sources are inherently heterogeneous and not interconnected, the use of multiple genomic datasets for analysis and knowledge discovery has raised pressing demands for enhanced data and metadata integration methodologies.

With "genomic data integration" we define the process of combining different types of genomic data and their associated metadata – each of which provides a different and complementary view on the genome – into a single representation, which allows to understand aspects of the genome that cannot otherwise be inferred; metadata are the drivers for the integrated management and linking of data. Genomic data integration must address several technical issues: (i) the need for continually updated data, to guarantee higher quality of results; (ii) the lack of normalization/harmonization between the processing pipelines [2]; (iii) the limited structured metadata information and agreement among models [3]; (iv) the unsystematic use of controlled terminology to allow interoperability [4].

In this review article we focus on human genomic datasets; the term "genomic" is not restricted to DNA data, but used in its typical broader meaning, which includes also transcriptomic, epigenomic, and miRNomic data; however, we do not include other kinds of "omics" data, e.g., proteomics or metabolomics. A rich literature has been produced about genomic data integration (see [5] as specific for genomics, [6] for metabolomics, and [7; 8] for 'omics' in general). However, to the best of our knowledge, a comprehensive review illustrating the actors that are currently playing a role in the integration of genomic data and metadata has not been reported yet. In the following we illustrate how the rest of this review is arranged.

In Section 2, dedicated to the *technological pipeline of genomic data*, from data production to final use of genomic datasets, we describe the steps required for data and metadata production and integration. Each data source and platform may perform only some of these steps, depending on their engagement in the pipeline. We also discuss services and characteristics of the data access options provided to end users.

Section 3 is dedicated to a *taxonomy of the players involved in data production and integration and their interplay*; we describe five main roles of these players and the relationships between them. Specifically, in the landscape of genomic production and integration, we identified the following types: *contributors*, such as laboratories that produce the wet-lab data and associated information; *repository hosts*, organizations handling primary and secondary data archives, such as the well-known Gene Expression Omnibus (GEO, [9]); *consortia*, international organizations who have agreed on broad data collection actions (The Encyclopedia of DNA Elements, ENCODE [10], is a notable example); *integrators*, initiatives whose main objective is combining data collections from other players and provisioning high quality access to integrated resources; and *consumers*, the actual users of the exposed data platforms and pipelines. We also discuss the interactions among different players.

Finally, in Section 4 we describe the *main players* in the three central categories (including 4 repository hosts, 12 consortia, and 13 integrators), specifying which parts of the technological pipeline discussed in Section 2 they address. In particular, we provide a detailed description of the integrative strategy operated by our group within the Genomic Computing project (GeCo, [11; 12; 13]), which has dedicated huge efforts to the whole genomic data integration problem.

2 Technological pipeline of genomic data

Data and their corresponding descriptions, i.e., metadata, are first produced, then integrated. In this section, we give an overview of the relevant technological phases towards final use, distinguishing between data, metadata, and also services and access interfaces built on top of them. Relevant steps are highlighted in the following in bold and comprehensively depicted in Figure 1 (data steps are in grey, metadata ones in purple, service/access ones in green), along with supporting objects (in orange) that guide the definition of each step.

2.1 Production

Every genomic research study starts with nucleic acid **sample collection and preparation**; ensuring high quality samples is important to maximize research efforts and validity of data analysis. This phase deals with privacy issues, for example, related to the use of clinical samples in research; it is impossible to create fully anonymized samples and this leads to issues of identifiable population data.

Methods that determine the nucleotide sequence of DNA and RNA molecules are called sequencing. Next-generation sequencing (NGS) is a high-throughput sequencing technology that enables the reading of billions of nucleotides in parallel. Sequencing (or "**primary analysis**") includes: (i) raw data generation; (ii) analysis of hardware generated raw data; (iii) generation of sequencing reads and their quality score, i.e., billions of short sequencing reads that are stored in text files in FASTQ format.

Typically, production is not driven by any imposed wet-lab standard, unless laboratories are guided by a consortium or other organization (e.g., ChIP-seq can have antibody standards, RNA-seq and DNase-seq can have specific protocols and replicate numbers).

The **metadata generation** is usually performed by the laboratories that generate the raw data; they document it in a rich way, yet approximate in the structure and possibly imprecise in the content. Basic information, about the performed assay, the used sequencing platform, and the analyzed biological material, are collected.

Researchers can then submit data through one of the several data brokers that act as links between production laboratories and ingestion application programming interfaces (APIs) provided by collecting platforms—at times these include web interfaces or web services. Upon submission, the ingestion services sometimes perform basic quality assurance and checking of format consistency, and then deposit the data into their data stores.

2.2 Integration

We describe as part of "data integration" all the steps that follow data production and their preliminary publication. Along the way, a number of issues may be encountered. Thus, we hint at existing methodological solutions chosen by players addressing the mentioned aspects.

During data processing, also referred to as "**secondary analysis**", genomic sequences are reconstructed in a computational way by exploiting overlaps between short sequencing reads. After quality assurance filtering on raw reads, typically the data processing workflow includes alignment of reads to a reference genome, which produces BAM files. The differences between the sequenced genome and the reference one can be identified, for example, by performing variant calling and filtering, which produce VCF files. Other secondary analysis workflows output different file formats[1], e.g., Browser Extensible Data (BED) files from peak calling, or Gene Transfer Format (GTF) files from the identification of differentially expressed genes.

The following steps are at times performed together with the previous ones, other times delegated to other data players that follow up with other data manipulations.

Quality control (QC) is vital for NGS technology experiments. It can be performed during three phases: on the initial extracted nucleic acids (in case they are degraded), after the sequencing library preparation (to verify that the insert size is as expected and that there are no contaminating substances), and after sequencing (most common tools are Sequence Analysis Viewer and FastQC). The more time and effort spent on QC, the better quality results will be. Many players report some kind of QC check in one of these phases; sometimes even just producing quality studies and reporting is considered as QC.

Some players may decide to reprocess part of the data collected elsewhere. Reasons for taking this approach may be several and of different nature, mainly including the need for normalized pipelines, as a means to obtain more homogeneous data ready for analysis. The **normalization of the pipelines** deals with the problem of converting raw data to numerical data such that any expression differences between samples are due solely to biological variation, and not to technical variation introduced experimentally; for example, in microarrays-based experiments technical bias can be introduced during sample preparation, array manufacture and array processing. Selected data types require some processing to achieve compliance to standards (e.g., alignment to a reference sequence, uniform peak calling, thresholding of signal peaks, consistent signal normalization, consistency check between replicates, ...).

Among post-processing activities tailored at enhancing interoperability among different datasets, we mention **data normalization** procedures

[1] https://genome.ucsc.edu/FAQ/FAQformat.html

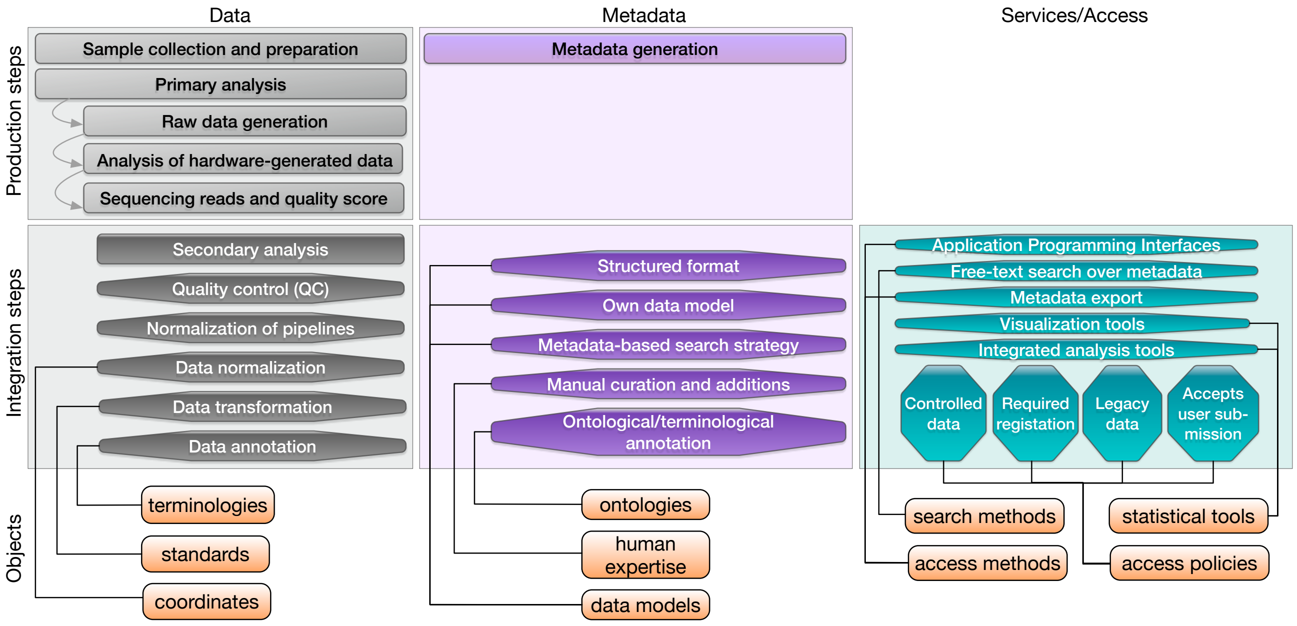

Fig. 1. Diagram of production steps, integration steps, and objects supporting the activities. By reading the diagram from left to right and from top to bottom, we find: the steps involved in producing data (in grey) and metadata (in purple); the steps that are part of integration challenges concerning data (in grey), metadata (in purple), and also information access and provided services (in green). Steps are linked to the supporting objects in orange: these are external data structures or abstract entities needed to define the step execution. Optionality of steps in the diagram is indicated by octagonal shapes, while compulsory steps, which are mainly in production, are rendered as rounded rectangles.

(such as format conversions like normalization of coordinates or re-formatting into narrowPeak or broadPeak standard format in ENCODE), **data transformation** (e.g., matrix-based data formatted as BED data), and **data annotation**. Examples of the latter include: (i) providing positional information (i.e., genomic coordinates) and associated known genomic regions (e.g., genes) in a standardized framework; (ii) allowing joined use of different data types (e.g., gene expression and methylation) based on common gene and sequence identifiers, such as gene IDs from HGNC [14], Entrez Gene [15], or Ensembl [16] terminologies; (iii) merging together in same files multiple expression measures obtained through different calculations, such as FPKM, FPKM-UQ, and counts in gene expression data. As terminologies for structural and functional sequence annotation are various, most integration strategies keep multiple representations together, as a conservative solution for allowing interoperability.

For what concerns metadata, which is the information associated with produced data, it may be organized in a **structured format**. In some cases, integrators apply tailored integration pipelines to extract the needed information to fill their **own** agreed **data model**s, thus performing schema-level integration. Generally, the idea is to redistribute metadata over a few essential entities, e.g., *Project-Sample-File* or *Investigation-Study-Assay* [17], as proposed in the general-purpose ISA-Tab format for structuring metadata [18] and now adopted by the FAIRsharing resource [19]. A number of questions are usually answered during this process: *Is this set of entities minimal? Is it enough to hold all information? Is something lost at this granularity?* Note that, depending on the specific sources, metadata elements have been linked to different entities, and a different base entity has been selected at times. To mention some notable choices:

- ENCODE has centered everything on the *Experiment*, which includes a number of *Biosample*s, from which many *Replicate*s are produced, to which *Item*s belong (sometimes with a many-to-many cardinality as files may be combined from multiple replicas).[2]
- Genomic Data Commons (GDC, [2]) is centered on the *Patient* concept, from which multiple *Samples* are derived. From another perspective, data are divided by *Project*, associated with a *Tumor Type*, for which many *Data Types* are available.
- GEO is organized into *Series* that include *Samples* (whereas these latter ones can be employed in multiple *Series*), sequenced with a *Platform*. A higher-order classification organizes *Series* and *Samples* in *Datasets* and *Profile*s.

Most platforms offer a **metadata-based search strategy**, exploiting the querying possibilities over the new metadata schema. However, sometimes such a search functionality is available even when no new data schema has been applied.

Some players, especially the ones working in connection with a Data Coordination Center, perform **manual curation and additions** to metadata. Within such activities, we mention in particular two: *assigning labels to replicas* and *cleaning metadata names*.

The first activity is used for managing technical and biological replication, which is a common and recommended practice in genomic experiments [20; 21; 22]. In a data model where information is organized based on a hierarchy (e.g., *Experiment/Replicate/File*), it is very likely that metadata will be replicated inside each element. Metadata format such as JSON or XML have a hierarchical structure and can easily represent such data models, by encapsulating each element inside its parent. When a de-normalization of such structures is produced (e.g., to associate with a materialized file also information about the ancestor *Replicate* or *Experiment*), integrators face the problem of assigning labels to metadata in such a way that the one/many-to-many relationship, which is implicit in the JSON or XML syntax, can stay explicit in the data. To overcome this

[2] A complete list of entities is available at `https://www.encodeproject.org/profiles/`.

problem while flattening hierarchical formats, it is customary to assign labels to metadata that belong to ancestors, in such a way that they can be recognized also in the de-normalized version.

The second activity, cleaning metadata names, is needed because, after label assignment, attribute names may become too long. It is highly preferable to make names minimal, so they still express their information without loosing their meaning, even if the semantics of nesting is removed; in this way they are more easily usable within a metadata-based search system and in the connected analysis platforms. As an example, a rather complicated attribute such as `replicates.biosample.donor.organism.scientific_name`, derived from flattening five hierarchical levels in a JSON document, may be simplified into `donor.organism` to facilitate understanding. Redundant information, including duplicated attributes deriving from a comprehensive download approach from the source, may also be removed using similar rule-based mechanisms.

Other widely adopted processes to enhance metadata interoperability include **ontological/terminological annotation** on top of the original or curated metadata. Annotation is a means to achieve metadata normalization, needed to compare metadata terms. Genomics, as many other fields in Bioinformatics, is greatly helped by specialized ontologies, which mediate among terms and enable interoperability. A considerable number of key ontologies are used by many genomic actors: Uber Anatomy Ontology [23] for tissues, Cell Ontology [24] and Experimental Factor Ontology [25] for primary cells and cell lines, Ontology for Biomedical Investigations [26] for assays, Gene Ontology [27] for biological processes, molecular functions and cellular components. All these are employed by ENCODE, that has dedicated great efforts to the systematization of official term names for the description of its data. Experimental Factor Ontology is used by the Genome-Wide Association Studies (GWAS) Catalogue [28] that curates all trait descriptions by mapping them to terms of this ontology. Moreover, GDC enforces a standardization using the National Institutes of Health (NIH[3]) Common Data Elements (CDE[4]) rules. Many attributes present codes referencing terms from the CDE Repository controlled vocabularies. Other relevant resources include: the National Cancer Institute (NCI[5]) Thesaurus [29] for clinical care, translational/basic research, and administrative/public information, and the National Center for Biotechnology Information (NCBI[6]) Taxonomy [30], providing curated nomenclature for all of the organisms in the public sequence databases. Several search services, which integrate a high number of ontologies, are employed in the landscape of genomic data integration. Examples include: BioPortal [31] and Ontology Lookup Service (OLS, [32]), two repositories of biomedical ontologies and terminologies that provide services to annotate search keywords with ontological terms; Ontology Recommender [33], a BioPortal service that annotates free text with a minimal set of ontologies containing terms relevant to the text; Zooma[7], an OLS service providing mappings between textual input and a manually curated repository of text-to-ontology-term mappings. Annotation can also include adding external identifiers pointing to different databases that contain same real-world entities. While redundancy is not accepted within a single source, in the genomics domain it is common across sources, provided that resources are well interlinked and representations are coherent between each other (i.e., metadata values have the same level of detail).

2.3 Services and Access

Organizations operating in the integration field also provide interfaces to access the result of their service. To this end, they must address the issues related to the synchronization of their local database with the original data one. As data size is significant, when updating the interface content, downloading everything from scratch from the original sources should be avoided. Instead, it is necessary to precisely define metrics to compare contents: *What is new, what has been updated, and what is not present anymore?* One possible strategy is defining a partitioning schema. In many cases this is not the simplest possible one (i.e., file by file) since information is typically structured in a complex and hierarchical way. For example, when considering the source ENCODE, metadata can be used to partition the source data repository. API requests can be composed in order to extract always the same partition of data, specifying parameters such as "type = Experiment", "organism = Homo Sapiens" and "file.status = released". Consequently, a list of corresponding files is downloaded; the identifying characteristics of the files (typically including size, last update date, checksum) can be compared with the ones saved in the local database at a previous download session of the same partition. Making a distinction between genomic region data and metadata, the latter are typically smaller in size; in case comparing versions becomes too complicated, metadata may be downloaded each time, as often there are no such things as a data release version or pre-computed checksum values to be checked.

Besides offering updated content, genomic players that host data and make it available through any kind of interface, usually offer also other supporting services. Typically these include: **application programming interfaces** to directly download and extract specific portions of data, or perform rich and structured queries; **free-text search over metadata**, sometimes only on selected kinds of metadata, such as gene names or functional annotations; direct **metadata export**, at times included within the API options, other times as bulk download; **visualization tools** or ready-to-use connections to common visualization browsers (e.g., UCSC or Ensembl genome browser); embedded **integrated analysis tools** to further process and analyze the results retrieved in the interface (diverse use cases include clustering [34], spatial reconstruction [35], visualization [36], and graph-based analysis [37]); possibility to perform **computations on cloud** in a dedicated space, with reserved computational resources.

In addition to openly available datasets, some sources also feature **controlled data**, whose access is only given upon authorization; some just **require a registration step** to access the download functionality. Many players make available **legacy versions of data**, rarely of metadata.

A few players **accept user-data submissions**, either to be included as future content of the platform or to be processed together with publicly available datasets in further computations and analysis.

3 Taxonomy of the players involved in data production and integration and their interplay

The landscape of institutions, private actors and organizations within the scope of genomics is broad and quite blurred. The authors of [9] had previously proposed a tentative classification of sources: *primary resources* publish in-house data; *secondary resources* publish both in-house data and collaborator data; *tertiary resources* accept data to be published from third, unrelated parties. More recently, the Global Alliance for Genomics and Health (GA4GH, [38]), an international, nonprofit alliance formed in 2013, built the Catalogue of Genomic Data Initiatives[8], where they include the following types, not mutually exclusive: *Biobank/Repository*,

[3] https://www.nih.gov/

[4] https://cde.nlm.nih.gov/

[5] https://www.cancer.gov/

[6] https://www.ncbi.nlm.nih.gov/

[7] https://www.ebi.ac.uk/spot/zooma/

[8] https://www.ga4gh.org/community/catalogue/

Consortium/Collaborative Network, *Database*, *GA4GH Driver Project*, *Industry*, *National Initiative*, *Ontology or Nomenclature Tool*, *Research Network/Project*, *Standards*, and *Tool*.

We expand the taxonomy of [9], whereas we compact the one proposed by GA4GH – which in any case only includes initiatives under the alliance's umbrella – by identifying five categories to classify every entity that plays a role in this field, named *genomic data player*. In general terms, data are produced at laboratories (corresponding to the player: *contributor*), deposited at data archives (player: *repository host*), harmonized within programs (player: *consortium*), integrated by systems or platforms that aggregate data from different sources and add value to it (player: *integrator*), and employed by end users, mainly biologists and bioinformaticians (player: *consumer*). These categories are not intended to completely represent the whole possibilities, nor to be exclusive with respect to each other. In the following, we detail each category's characteristics and the interactions among categories, carrying genomic data from production to its integrative use.

Contributor. A contributor generates raw data with any high-throughput platform, using next-generation sequencing or any other technology; it takes care of annotating wet-lab experiment data with a set of descriptive metadata, as well as encrypting and uploading data to archives. A contributor can be a laboratory or hospital, which reports directly to a Principal Investigator holding an independent grant and leading the grant project. In other cases laboratories are part of a bigger program, led by a consortium or national institution. In both cases, it is customary for laboratories to send their data to other players, who carry on the publication and integration process.

Repository host. We call "repository hosts" the organizations standing behind primary data archives (also referred to as "data storage"), recently grown exponentially in size. They host data not only from independent laboratories and companies that gain visibility in this way, but also from consortia that wish to make their data available from such general archives. Moreover, it is customary for authors of biological publications to deposit their raw and processed datasets on these repositories—some journals even require it upon submission [39]. Primary data archives currently face a number of challenges:

- Their primary goal is to pool disparate data into a single location, giving priority to quantity and typically not demanding a structure. However, without any homogenization effort, data is barely useful, impeding analysis and cross-comparison that build an added value with respect to individual experiments [40].
- They archive raw sequencing data, which is usually not immediately usable by the scientific community. The majority of these archives do not provide access to pre-processed published data, leaving this cumbersome task to individual scientists who need to analyze them.
- Usually metadata deriving from contributors' submissions are not sufficient to ensure that each dataset/experiment is reproducible and that the data can be re-analysed. As new technologies, protocols and corresponding annotation vocabularies are constantly emerging, new metadata fields are required and need curation to accurately reflect the data.

Consortium. Consortia provide evolved forms of primary repositories. They usually include many participants and projects, which have to abide to certain policies (see, for example, policies of GDC[9]) and operational conventions for participation (see, for example, experiment guidelines of ENCODE[10]). These policies have to ensure agreement among the parts about sensitive matters such as data access, data submission, and privacy. Guidelines guarantee compatibility among datasets, in order to establish an infrastructure that enables data integration, analysis, and sharing.

Many consortia refer to a Data Coordination Center (DCC) in charge of data and metadata normalization and cleaning, and of all the activities that stand between production and publication. Most well-known DCCs include the ones of ENCODE [41], BLUEPRINT [42], ICGC [43], and 1000 Genomes [44]; Roadmap Epigenomics Consortium [45] has a Data Analysis and Coordination Center (EDAAC[11]) and GTEx has a Laboratory, Data Analysis, and Coordinating Center (LDACC [46]). Along with repository hosts, consortia are required by their own policies to submit their raw sequencing reads and other primary data to controlled access public repositories. The ones that serve this purpose are mainly the European Nucleotide Archive [47], the European Genome-phenome Archive [48], and the Database of Genotypes and Phenotypes [49].

Integrator. An integrator may be a platform, an initiative, or a project whose objective is to overcome the constant need of users to learn how to navigate new query interfaces and to transform data from different sources to be integrated in the analysis. As a secondary purpose, an integrator usually aims at providing visualization and integrative analysis tools for the research community. Integrators do not point to raw data; they instead always reference the sources of their data (either with links to the source portals, or by reporting original identifiers for each data unit).

Consumer. Genomic data and metadata are finally used by biologists, bioinformaticians and data scientists, who download them from sources' platforms and FTP servers to feed a wide variety of tertiary analysis pipelines, including applications in pharmacology, biotechnology, and cancer research.

Interactions among genomic data players are described in Figure 2. Experimental genomic data and metadata are first produced – occasionally also preliminarily processed – by contributors, then published on repositories or directly on consortia's platforms. Within consortia themselves, re-processing may happen, as their pipeline for raw data processing uses community-agreed or consortium's guideline-based algorithms. Intermediate derived results are generated to be later published. In some cases data curated by consortia are re-published also on general archives, such as ENCODE and Roadmap Epigenomics on GEO[12]. Data are finally collected by integrators that expose them on tertiary interfaces, tailored at enhancing interoperability and use. Simpler interfaces are provided to consumers also by many repository hosts and consortia. In rare cases, not depicted in Figure 2 as they are exceptional, integrators may consider important to re-process data of some sources with normalized pipelines to enhance the possibilities of integration. Two long arrows show the taxonomy from different points of view: from a *data perspective*, contributors deal with raw data, repository hosts and consortia with processed data, while integrators make data interoperable and fit for use of consumers; from a *process perspective*, data is produced by contributors, submitted to aggregating platforms, that take care of dissemination to tertiary players, who make it available for its consumption.

As an instantiation of the diagram described in Figure 2, we apply the same taxonomy to a number of relevant genomic players, which will be

[9] `https://gdc.cancer.gov/about-gdc/gdc-policies/`

[10] `https://www.encodeproject.org/about/experiment-guidelines/`

[11] `http://www.roadmapepigenomics.org/overview/edaac`

[12] `https://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html`, `https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/`

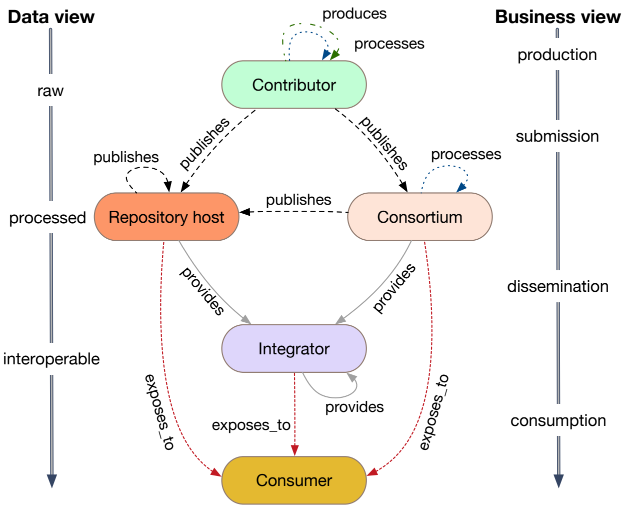

Fig. 2. Diagram of interactions among genomic data players. Nodes are players; arrows, with different colors and textures represent their interactions. Contributors publish either on repositories or consortia platforms; data are then integrated. Consumers retrieve data from repositories, consortia or integrators. The data view represents the perspective of data stage, while the business view shows the process applied to genomic datasets.

described thoroughly in the following. Figure 3 thus shows interactions between example players, starting from the laboratories where data are primarily generated, throughout repositories where data are deposited, consortia where they are curated, and finally integrator interfaces where they are used and explored. The used notation and colors reflect the ones adopted in Figure 2.

We drew the relationships between these players according to their specifications in the documentation and relevant publications, to the best of our knowledge at the time of writing. Some consortia, for instance The Cancer Genome Atlas (TCGA) [50] and GDC, accept submissions both from laboratories gathered under the same organization and from individual submitters that observe the submission guidelines. There are labs that contribute to more projects. Raw experimental data are usually deposited to SRA [51], while GEO (the most used by researchers) and ArrayExpress [52] are employed for publication of data at later stages of processing; complete studies are uploaded to BioStudies [53]. Note that, in the Figure 3 diagram, even primary archives reference to each other.

4 Main genomic data players

We propose a systematic overview of a number of genomic data players, guided by Table 1. The first column of the table contains a list of data sources that contribute to produce, integrate and promote the use of genomic data for research, grouped by the categories of the taxonomy introduced in the previous section. The list is in no way meant to be comprehensive, but should be received as a starting reference. For each mentioned player we show which steps/functionalities are provided. The following columns in Table 1 represent steps described in Section 2 in bold font and depicted in Figure 1.

Inside Table 1 cells, the notation × indicates a step included by the player; an empty cell stands for a step not provided by the player; ~ is used when an answer is only partially positive (e.g., some parts of the step are performed while others not, or the step is only performed under certain conditions); ? is used for an unknown answer, when the documentation and publications describing the player and its services did not allow us to determine an answer. All information contained in the table are filled up to the best of our knowledge, being retrieved from the player's main publications or from the linked online documentation. Our discussed overview is divided in subsections, one for each player type.

4.1 Contributors

As it can be observed from Table 1, contributors, including both independent labs, private submitters, and consortium labs as depicted in Figure 3, are the only ones in charge of sample collection, preparation and primary analysis, followed by generation of metadata; only in some cases they also apply quality control measures before or during secondary analysis activities.

4.2 Repository hosts

Gene Expression Omnibus (GEO, [9]) is the most general and widely used among repositories. It started in 2002 as a versatile, international public repository for gene expression data [54]; it then consequently adopted a more flexible and open design to allow submission, storage and retrieval of a variety of genomic data types, such as from next-generation sequencing or other high-throughput technologies. To include also non-expression data, in 2008 GEO created a new division called "Omix", standing for a mixture of 'omic data' [55].

Data can also be deposited into **Sequence Read Archive** (SRA, [51]) as supporting evidence for a wide range of study types: primarily raw sequence reads and alignments generated by high-throughput nucleic acid sequencers (BAM file format), now expanded to other data including sequence variations (VCF file format) and capillary sequencing reads. As a part of the International Nucleotide Sequence Database Collaboration (INSDC), the SRA is materialized in three instances, one at the European Bioinformatics Institute (EBI[13]), one at the NCBI [56], and one at the DNA Data Bank of Japan (DDBJ, [57]).

ArrayExpress [52] was first established in 2002 only for microarray data. It is now an archive of functional genomics data ranging from gene expression and methylation profiling, to chromatin immunoprecipitation assays. Recently, it also increased the number of stored experiments investigating single cells, rather than bulk samples (i.e., single-cell RNA-seq).

The EBI **BioStudies** [53] database holds high-level metadata descriptions of biological studies, with links to the underlying data databases hosted at the EBI or elsewhere, including general-purpose repositories. Also those that have not been already deposited elsewhere can be hosted at BioStudies.

Discussion. By observing the repository-related rows of Table 1, we conclude that repositories are quite diverse with respect to the data integration steps included in their practice. Generally, they do not perform specific steps on data, however they often require submitters to ensure quality control checks, as GEO and ArrayExpress do, while SRA mentions it as future work. Metadata are not treated uniformly; some organization is enforced, but much information is left also in unstructured format. While services offered by their interfaces are various, they all allow submissions from any user; this is a characterizing feature of the repositories.

Since repositories were growing in diversity, complexity and (unexpressed) interoperability [58], the need for organization and annotation of the available data became primary. To this end, NCBI and EBI have implemented additional, complementary, initiatives on top of repositories. NCBI BioProject and BioSample databases [59] and EBI Biosamples [60] were initiated to help addressing these needs by facilitating the capture and management of structured metadata and data

[13] https://www.ebi.ac.uk/

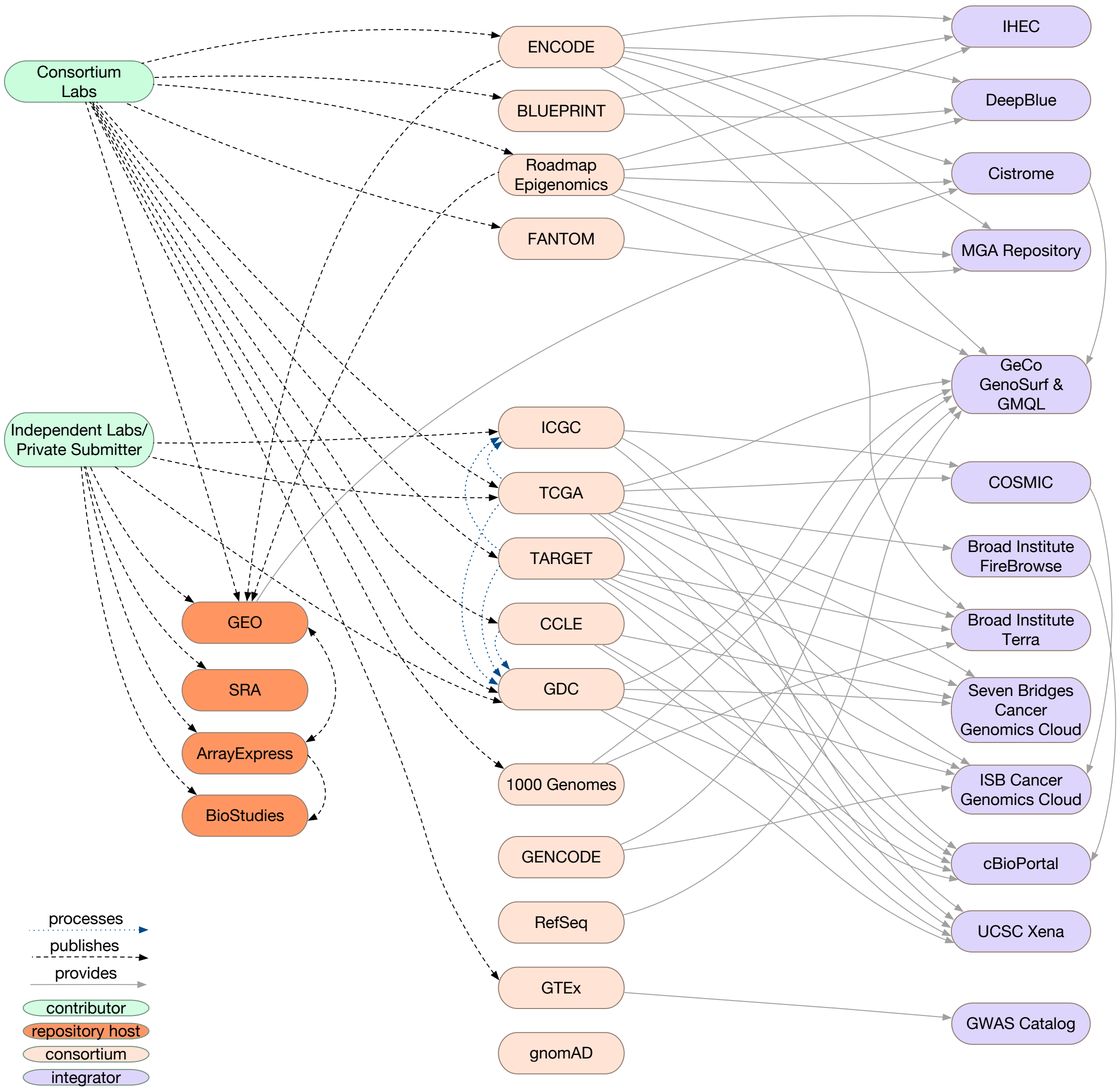

Fig. 3. Diagram of example players and their most important interactions. Note that, with respect to rows referring to consortia in Table 1, two more nodes are shown here: TCGA and TARGET, as they contribute with their data to many other players. They are not discussed separately in Table 1 as, currently, their data is only made available through other platforms (the most important are GDC and ICGC); the old TCGA portal was dismissed and TARGET does not have its own one.

for diverse biological research projects and samples represented in their archival databases.

4.3 Consortia

Consortia are usually focused on particular aspects of what we generically call "genomics". The following four work all on matters related to epigenomics.

The Encyclopedia of DNA Elements (**ENCODE**) Consortium [10] is an ongoing international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). Primary goal of the project is to characterize functional features in DNA and RNA expression in a wide number of cell lines. The project's integrative effort is presented in [61]; ENCODE DCC also published interesting results regarding its achievements [41; 62; 63; 64], respectively reporting on ontologies used for annotation, metadata organization, storage system, and duplication prevention.

BLUEPRINT [65] is an EU-funded consortium under the umbrella of the International Human Epigenome Consortium (IHEC). It was set up to develop new high-throughput technologies to perform epigenome mapping, and to analyze diverse epigenomic maps comprehensively, making them available to the scientific community as an integrated resource. Besides being available through IHEC resources, the BLUEPRINT built its own Data Analysis Portal [42], as the first platform based on EPICO, an open access reference set of libraries to be used to develop data portals for comparative epigenomics.

The **Roadmap Epigenomics Consortium** [45] was born in 2015 from the NIH with the aims of: (i) understanding the biological functions of epigenetic marks and evaluate how epigenomes change; (ii) designing and improving technologies, i.e., standardized platforms, procedures, and reagents, that allow researchers to perform epigenomic analysis and to study epigenetic marks efficiently; (iii) creating a public resource of disease-relevant human epigenomic data to accelerate the application of epigenomics approaches.

FANTOM [66] is an international research consortium created to perform functional annotations of the mammalian genomes, including *Homo Sapiens*. The object of the project has recently moved from understanding the transcripts to understanding the whole transcriptional regulatory network.

The following three consortia are instead working on problems related to cancer genomics.

Genomic Data Commons (GDC, [2]) is an information system for storing, analyzing and sharing genomic and clinical data from cancer patients. It aims to give democratic access to such data, improve sharing and promote approaches of precision medicine that can diagnose and treat cancer. Ultimately, the goal is to become the one-stop cancer genomics knowledge base; however, consolidation and harmonization of genomic and clinical data are ongoing and they will require a long process. GDC was created mainly to help individual investigators and small programs to meet NIH and the NCI genomic data sharing requirements, and thus to store their data in a permanent home. In addition, GDC now includes data from big cancer programs, such as The Cancer Genome Atlas (TCGA, [50]) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET), also shown as consortia nodes in Figure 3. While GDC is technically a cancer knowledge network, we classify it as a *consortium* as it has a very broad mission: it accepts user submissions, it performs quality control, it provides storage and it also redistributes the data. What particularly distinguishes it from a simple *repository host* or an *integrator* is the great effort dedicated to harmonizing data (standardizing metadata, re-aligning data and re-generating tertiary analysis data using new pipelines [67]) deriving from incoming submissions and from the included cancer programs. TCGA, instead, is a terminated program; it no longer accepts samples for characterization. It used to expose the data by means of its own portal, while now it relies on the GDC infrastructure. As its concluding project, in 2018 the TCGA program produced the Pan-Cancer Atlas [68], a collection of analysis performed cross-cancer type. In addition to including many single-cancer-type projects, the last datasets that GDC platform makes available are the ones produced within the context of the Pan-Cancer project.

The **International Cancer Genome Consortium** (ICGC, [43]) was established in 2011 to launch and coordinate a large number of research projects de-centralized in many countries of the world, sharing the common goal of explaining the genomic changes present in many forms of cancer. Its Data Portal hosts data from other large-scale projects focused on cancer research, such as TCGA and TARGET, as shown by the two dotted incoming arrows in Figure 3.

The **Cancer Cell Line Encyclopedia** (CCLE, [69]), for almost 1,500 human cancer cell lines, collects gene expression, chromosomal copy number and massively parallel sequencing data.

The last five consortia we mention are instead focused on various matters, such as variation across populations, transcriptomics, exome sequencing, and annotation.

Launched to become one of the largest distributed data collection and analysis projects in genomics, the goal of the **1000 Genomes Project** [44] was to find most genetic variants with frequencies of at least 1% in the studied populations. In 2015 the International Genome Sample Resource (IGSR, [70]) was established to expand and improve the legacy inherited from the 1000 Genomes Project.

The **Genotype-Tissue Expression Consortium** (GTEx, [46]), supported by the NIH Common Fund, aims at establishing a resource database and associated tissue bank to study the relationship between genetic variation and gene expression and other molecular phenotypes in multiple reference tissues. The results of this transcriptomics-focused project help the interpretation of findings from genome-wide association studies (GWASs) by providing data and resources on expression quantitative trait loci in many tissues and diseases.

The **GENCODE** project [71] produces high-quality reference gene annotations and experimental validation for human and mouse genomes. It aims at building an encyclopedia of genes and gene variants, by identifying all gene features in the human and mouse genomes, using a combination of computational analysis and manual annotation.

The NCBI **RefSeq** project [72] provides a comprehensive manually annotated set of reference sequences of genomic DNA, transcripts, and proteins—including, for example, genes, exons, promoters, enhancers, etc.. Exploiting the data from the INSDC, it provides a stable reference for genome annotation, analysis of mutations and studies on gene expression.

The Exome Aggregation Consortium (ExAC, [73]) unites a group of investigators who are aggregating and harmonizing exome sequencing data from other large-scale projects. According to the most updated news (end of 2016, [74]), the ExAC provided sequences from almost 61,000 individuals belonging to studies about different diseases and populations. Recently, the ExAC browser has been dismissed in favour of the Broad Institute **Genome Aggregation Database** (gnomAD, [75]), which more than doubles the previous sample size.

Discussion. From Table 1 we observe that consortia are generally concerned with coordination of data transformation, from secondary analysis activities, to quality control filtering and normalization/annotation of data. Almost all the analyzed consortia definitely include a pipeline normalization step in their activities, as this is the characterizing step of this type of player (only BLUEPRINT and GENCODE did not mention information about uniform workflows in their documentation). Instead, the approach towards metadata curation and tertiary analysis tools is diversified and does not show a unique trend. As to the "metadata-based search strategy" column in Table 1, we specify that some consortia just provide a very limited functionality of this kind (e.g., 1000 Genomes can only filter by population, technique and data collection, and Roadmap Epigenomics by tissue and data type only). While ENCODE and GDC present sophisticated search interfaces, other ones are quite basic. As to providing APIs and visualization tools, GENCODE and 1000 Genomes were assigned the ~ symbol since they do not offer such services natively, but exploit the ones of Ensembl.

4.4 Integrators

As consortia, also integrators tend to clusterize based on the sub-branch of genomics they cover. This happens mostly because rules and common-practices are better shared within a same branch, following similar purposes, while cross-branch projects are more rare. We first mention four integration organizations that collect data from epigenomics-focused consortia.

The **International Human Epigenome Consortium** (IHEC, [76]) coordinates large-scale international efforts towards the production of reference epigenome maps. For a wide range of tissues and cell types, the regulome, methylome, and transcriptome are characterized. As a second phase, the consortium is expanding its focus from data generation to the application of integrative analyses and interpretation on these datasets, with the goal of providing a standardized framework for clinical translation of epigenetic knowledge. We classified the IHEC as an *integrator* rather than a *consortium* as the normalization work is mainly carried on by the members institutions (or consortia themselves) that are part of it (for instance ENCODE, BLUEPRINT, Roadmap Epigenomics). The main outcome of the IHEC is instead its Data Portal, which can be used to view, search, download, and analyse the data already released by the different associated projects.

DeepBlue [80] is a data server that was developed to mitigate the lack of mechanisms for searching, filtering and processing epigenomic data,

Table 1. Overview of the steps towards data integration included by genomic data players. Rows represent players (with reference of main publication, when available) and are grouped by player type. Columns represent the steps for genomic data integration described in Section 2 and are grouped according to their progression in a typical pipeline. Used notation: × indicates that a certain step is included/performed by the player; ~ indicates an uncertain answer (i.e., in some cases the service/step is provided just in few studies or for some data types); ? indicates that the player's documentation and publications did not allow determining an answer; empty cell indicates that the service/step is not provided.

		Production			**Integration**																				
					Data					Metadata					Services						Access			
		sample collection and preparation	primary analysis	metadata generation	secondary analysis	quality control	pipeline normalization	data normalization/transformation	data annotation	structured format	own data model	metadata-based search strategy	manual curation/additions	ontology/terminology annotation	application program interfaces	meatadata-free-text search	metadata export	visualization tools	integrated analysis tools	computations on cloud	include controlled data	required registration	legacy versions	accept data submissions
Contributor	Laboratories	×	×	×	~	~																		
Repository	ArrayExpress [52]					×			×	×			×		×	×	?	×	×		×		×	×
	BioStudies [53]											×		×		×	?				×	~	~	×
	Gene Expression Omnibus (GEO, [9])					×				~	~	~	×				×	×	×				×	×
	Sequence Read Archive (SRA, [51])					~				×	×	×			×		×						?	×
Consortium	1000 Genomes Project Consortium [44]				×	×	×	×	×	~		~			~		×	~					×	
	BLUEPRINT [65; 42]				?	?	?	×	?	×	×	×		×	×		×	×	×				?	
	Cancer Cell Line Encyclopedia (CCLE, [69])				×	×	×	×	×			×				×	×	×	×			×	×	
	ENCODE [10] Consortium				×	×	×	×	×	×	×	×	×	×	×	×	×	?	?		×		×	
	FANTOM [66]				×	×	×	?	×									×	×				×	
	GENCODE [71]					×	?	×	×						~			~					×	
	Genome Aggregation Database (gnomAD, [75])				×	×	×											×					×	
	Genomic Data Commons (GDC, [2])				×	×	×	×		×	×	×	×	×	×		×	×			×		×	×
	Genotype-Tissue Expression Consortium (GTEx, [46])				×	×	×	×	×						×			×			×		×	
	International Cancer Genome Consortium (ICGC, [43])					×	×	×	×	×	×	×		×	×		×		×		×	~	×	×
	RefSeq [72]					×	×	×	×				×										~	
	Roadmap Epigenomics Consortium [45]				×	×	×	?	×	×		~					×	×						
Integrator	Broad Institute Firehose/FireBrowse				?	?	?	?	?			×			×	~		×	×				×	×
	Broad Institute Terra									×		×		~		~	×	×	×	×	×	×		×
	cBioPortal [77]				×		×	×	×			×			×	~	×	×	×				×	~
	Cistrome [78]				×	×	×		×	×	×	×	×	×		×	×	×	×			×		
	COSMIC [79]								×	×	×	×	×	×		×	×	×	×			×	×	
	DeepBlue [80]								×	×	×	×	×	×	×	×	×		×					×
	GeCo GenoSurf & GMQL [12]							×	×	×	×	×	×	×	×	×	×	×	×	×			×	×
	GWAS Catalog [28]					×		×	×	×	×	×	×	×	×	×	×	×						×
	International Human Epigenome Consortium (IHEC, [76])					×	×	×		×	×	×	×	×	×		×	×	×		×	~	×	×
	ISB Cancer Genomics Cloud [81]									×		×			×			×	×	×	×	×		×
	MGA Repository [82]				×		×	×	×	×			×		×	×		×	×					
	Seven Bridges Cancer Genomics Cloud [83]									×		×	?	?	×	~	×	×	×	×	×	×	×	×
	UCSC Xena [84]									×	?	×	?		×	×		×	×	×			×	×

within the scope of the IHEC. DeepBlue made a precise work of data integration by homogenizing many epigenomic sources, including data from ENCODE, BLUEPRINT, Roadmap Epigenomics among others. It uses a clear distinction between region data and metadata, manages both experiment and annotation related datasets, defines a set of mandatory metadata attributes – while storing additional ones as key-value pairs – and uses metadata to locate region data.

In the **Cistrome** Data Portal [85] users can find data relevant to transcription factor and chromatin regulator binding sites, histone modifications and chromatin accessibility. Such data is useful to a number of studies, including differentiation, oncogenesis and cellular response to environmental changes. As to the last available publication [78], its database contains about 100,000 samples, both for human and mouse organisms. It includes data of ChIP-seq and chromatin accessibility from ENCODE, Roadmap Epigenomics and GEO, which has been carefully curated and homogeneously re-processed with a new streamlined analysis pipeline, detailed at `http://cistrome.org/db/#/about/`. Comparison between Cistrome enriched region signal peaks and the ones in ENCODE, which they are derived from, showed that they are significantly different.

The **MGA repository** [82] is a database of both NGS-derived and other genome annotation data, which are completely standardized and equipped with metadata. It does not store raw sequence files, but instead lists of base positions in the genome corresponding to reads from experiments, e.g., ChIP-seq. Ten model organisms are represented.

The following seven integrators are mainly working in the field of cancer genomics. Notice that, in Figure 3, the cluster formed by consortia and integrators working in the cancer domain is the most connected—integrators retrieve datasets from the most important consortia portals.

The Catalogue Of Somatic Mutations In Cancer (**COSMIC**, [79]) catalogue is the most comprehensive global resource for information on somatic mutations in human cancer. It contains 6 million coding mutations across 1.4 million tumour samples, which have been (primarily) manually curated from over 26,000 publications.

Broad Institute maintains both **Firehose/FireBrowse**[14] and **Terra**[15] platforms as aggregators of genomic data. The first one mainly imports TCGA data and offers a number of visualization options over it. The second one is a new large-scale project that also includes cloud computational environments.

Along with the Broad Institute, the **Seven Bridges Cancer Genomics Cloud** [83] and the **Institute for Systems Biology (ISB) Cancer Genomics Cloud** [81] are the other two systems funded by the NCI to store massive public datasets (first of all TCGA ones) and together provide secure scalable computational resources for analysis.

The **cBio Cancer Genomics Portal** (cBioPortal, [77]) was designed to address the data integration problems that are specific of large-scale cancer genomics projects, such as TCGA—including the Pan-Cancer Atlas datasets, TARGET, and ICGC. In addition, it also makes the raw data generated by large projects more easily and directly available to cancer researchers.

UCSC Xena [84] is a high-performance visualization and analysis tool that handles both large public repositories (e.g., CCLE, GDC Pan-Cancer, TARGET and TCGA) and private datasets. Its characterizing aspects are strong performances and a privacy-aware architecture, working across multiple hubs simultaneously. Target users are cancer researchers with and without computational expertise.

The NHGRI-EBI **GWAS Catalog** [28] is a collection of all published genome-wide association studies that enable investigations to identify causal variants, understand disease mechanisms, and establish targets for novel therapies. It adds manually curated metadata for publication, study design, sample and trait information. Many information from GTEx are also integrated.

As a last player, we present the ERC-funded **data-driven Genomic Computing** project (GeCo, [11]), which aims at providing a new focus on data extraction, querying, and analysis by raising the level of abstraction of models, languages, and tools for genomic tertiary analysis. The main research product the project has developed is a cloud-based data engine for genomic region-based data and metadata supporting a new query language for genomics, called GenoMetric Query Language (GMQL, [86]), based on the data model described in [87]. The associated GMQL query system [12] uses Apache Spark[16] on arbitrary servers and clouds.

Within the project, we analyzed thoroughly the issues related to the integration pipeline described in Section 2, and we proposed a unique approach implemented in a software architecture provided at `https://github.com/DEIB-GECO/Metadata-Manager/`; it keeps our (meta)data repository updated periodically and, as indicated in Table 1, includes all the integration steps for data and metadata that follow secondary analysis. Specifically, we already integrate data from ENCODE, TCGA, GDC, Roadmap Epigenomics, 1000 Genomes, GENCODE, and Refseq (plus metadata from Cistrome), and we are currently considering ICGC and Chip-seq data from GEO. Among the analysed integrators, the GeCo approach is the only one that joins together a broad range of genomic data, which spans from epigenomics to all data types typical of cancer genomics (e.g., mutation, variation, expression, etc.), until annotations.

For what concerns data downloading from sources, we follow a partition-driven approach to sync our local instances with the origin ones. We do not re-process data, but perform many transformation and normalization tasks, as proven by our work on TCGA data [88], where we reported the development of an automatic pipeline to transform into BED format the data available at the original TCGA portal (`https://tcga-data.nci.nih.gov/` – now deprecated), based on the hg19 reference assembly. The TCGA data has now been migrated to the GDC portal, which provides data for the GRCh38 assembly; we transformed into BED format also this updated version of the TCGA data (see `http://www.bioinformatics.deib.polimi.it/openGDC/`). Metadata are transformed by keeping information about replicates and cleaned to only maintain relevant information. Then, they are imported into a unique conceptual representation, the Genomic Conceptual Model (GCM, [89]), including 40 attributes.

We performed an assessment of different ontology search services and selected the best ones to annotate ten GCM metadata attributes with ontological terms, their definitions, synonyms, ancestors, and descendants, in order to instrument a semantically enriched search of datasets linked to such metadata [90]. On the basis of our experience, too many levels of ontological enrichment would bring to unnecessary sophistication, which would not be appreciated by users. Thus, we included ontology terms' super- and sub-concepts up to a small number of levels of depth, typically between three and five, as a reasonable trade-off that also guarantees acceptable query performances. We expose a fast metadata-based search engine over the GMQL repository of genomic datasets at the GenoSurf[17] web interface [13]; the interface is designed to allow user-friendly surfing upon integrated data. It has been evaluated and improved thanks to user feedback, by using a survey submitted to a large audience of bioinformaticians and genomics practitioners [91], who are the main target of our systems. So far GeCo's achievements have employed principally a model/system-driven approach, leading to significant limitations of usability and intuitiveness of the interfaces. We have become aware of the need for treating applications as first-class citizens to feed and consolidate the system; thus, major ongoing efforts are directed to produce a workflow-driven approach that makes data search and analysis processes more attractive for domain experts—with strong domain expertise, but small computer science and programming knowledge.

Discussion. Integrators, as reported in Table 1, are in general concentrated on metadata and services; however, some of them do re-process also data (e.g., Cistrome), and many of them transform and augment it in various ways. Table 1 also proves that genomic data integration means going through the pipeline described in Section 2.

5 Conclusion

In this paper we reviewed the steps and characteristics of production, integration, and accessibility of genomic data and related metadata. We presented a broad set of actors involved in the data life cycle, dividing them in taxonomical categories and inspecting their relationships. In the collection we provided, international initiatives are either focused on given diseases (e.g., cancer for TCGA), or on specific technologies (e.g., epigenetics for Roadmap Epigenomics). Meanwhile, we are assisting world-wide to the emergence of a new generation of large-scale genomic *national* initiatives [92]: some employ population-based sequencing (see *All of US* [93] from NIH in the United States—aiming

[14] Firehose: `https://gdac.broadinstitute.org/`; FireBrowse: `http://firebrowse.org/`

[15] `https://terra.bio/`

[16] `https://spark.apache.org/`

[17] `http://www.gmql.eu/genosurf/`

at sequencing 1 million American volunteers' genomes, *ChinaâĿ™s Precision Medicine Initiative* [94], *GenomeDenmark* [95], *Estonian Genome Project* [96], *Qatar Genome Programme* [97]); others are testing large numbers of cancer or rare disease patients (for example *100,000 Genomes Project* [98]—a UK Government project that is sequencing whole genomes from UK National Health Service patients, *Saudi Human Genome Program* [99], *Turkish Genome Project* [100]). Still other nations are focused on developing infrastructure to later achieve similar results (for instance *FinnGen* [101] and *GenomeCanada* [102]). As their data access models are unarguably not open for research (with rare exceptions), we did not discuss them in this review; no integrators include them yet either. All these projects share issues of data governance and privacy protection [103], but they certainly also represent a wealth of information, which will be considered within the scope of future data integration efforts, giving a new and substantial boost to the potential of genomic data analysis. In the specific, GeCo will continue its mission towards data integration to support clinical and biological research, using powerful data extraction and analysis models and implementations, towards user-friendly platforms.

Funding

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

References

[1]Schuster SC. Next-generation sequencing transforms today's biology. Nature methods. 2007;5(1):16.

[2]Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. New England Journal of Medicine. 2016;375(12):1109–1112.

[3]Posch L, Panahiazar M, Dumontier M, et al. Predicting structured metadata from unstructured metadata. Database. 2016;2016:baw080.

[4]Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Scientific data. 2019;6:190021.

[5]Hamid JS, Hu P, Roslin NM, et al. Data integration in genetics and genomics: methods and challenges. Human genomics and proteomics: HGP. 2009;2009:869093.

[6]Cambiaghi A, Ferrario M, Masseroli M. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Briefings in bioinformatics. 2017;18(3):498–510.

[7]Gomez-Cabrero D, Abugessaisa I, Maier D, et al. Data integration in the era of omics: current and future challenges. BMC Systems Biology. 2014;8(Suppl 2):I1.

[8]Manzoni C, Kia DA, Vandrovcova J, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings in bioinformatics. 2016;19(2):286–302.

[9]Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2012;41(D1):D991–D995.

[10]Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic acids research. 2017;46(D1):D794–D801.

[11]Ceri S, Bernasconi A, Canakoglu A, 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. Springer; 2017. p. 46–57.

[12]Masseroli M, Canakoglu A, Pinoli P, et al. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. Bioinformatics. 2018;35(5):729–736.

[13]Canakoglu A, Bernasconi A, Colombo A, et al. GenoSurf: metadata driven semantic search system for integrated genomic datasets. Database: The Journal of Biological Databases and Curation. 2019;2019.

[14]Yates B, Braschi B, Gray KA, et al. Genenames. org: the HGNC and VGNC resources in 2017. Nucleic Acids Research. 2017;45(Database issue):D619.

[15]Maglott D, Ostell J, Pruitt KD, et al. Entrez Gene: gene-centered information at NCBI. Nucleic acids research. 2010;39(suppl_1):D52–D57.

[16]Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. Nucleic acids research. 2018;46(D1):D754–D761.

[17]Sansone SA, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. Nature genetics. 2012;44(2):121.

[18]Sansone SA, Rocca-Serra P, Brandizi M, et al. The first RSBI (ISA-TAB) workshop:"can a simple format work for complex studies?". OMICS A Journal of Integrative Biology. 2008;12(2):143–149.

[19]Sansone SA, McQuilton P, Rocca-Serra P, et al. FAIRsharing as a community approach to standards, repositories and policies. Nature biotechnology. 2019;37(4):358–367.

[20]Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research. 2012;22(9):1813–1831.

[21]Yang Y, Fear J, Hu J, et al. Leveraging biological replicates to improve analysis in ChIP-seq experiments. Computational and structural biotechnology journal. 2014;9(13):e201401002.

[22]Schurch NJ, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? Rna. 2016;22(6):839–851.

[23]Mungall CJ, Torniai C, Gkoutos GV, et al. Uberon, an integrative multi-species anatomy ontology. Genome biology. 2012;13(1):R5.

[24]Meehan TF, Masci AM, Abdulla A, et al. Logical development of the cell ontology. BMC bioinformatics. 2011;12(1):6.

[25]Malone J, Holloway E, Adamusiak T, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010;26(8):1112–1118.

[26]Bandrowski A, Brinkman R, Brochhausen M, et al. The ontology for biomedical investigations. PloS one. 2016;11(4).

[27]Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic acids research. 2019;47(D1):D330–D338.

[28]Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research. 2018;47(D1):D1005–D1012.

[29]de Coronado S, Wright LW, Fragoso G, et al. The NCI Thesaurus quality assurance life cycle. Journal of biomedical informatics. 2009;42(3):530–539.

[30]Federhen S. The NCBI taxonomy database. Nucleic acids research. 2012;40(D1):D136–D143.

[31]Whetzel PL, Noy NF, Shah NH, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research. 2011;39(suppl_2):W541–W545.

[32]Jupp S, Burdett T, Leroy C, et al. A new Ontology Lookup Service at EMBL-EBI. In: International Conference on Semantic Web Applications and Tools for Life Sciences; 2015. p. 118–119.

[33]Martínez-Romero M, Jonquet C, O'Connor MJ, et al. NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. Journal of biomedical semantics. 2017;8(1):21.

[34]Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–2912.

[35]Satija R, Farrell JA, Gennert D, et al. Spatial reconstruction of single-cell gene expression data. Nature biotechnology. 2015;33(5):495.

[36]Loraine AE, Blakley IC, Jagadeesan S, et al. Analysis and visualization of RNA-Seq expression data using RStudio, Bioconductor, and Integrated Genome Browser. In: Plant Functional Genomics. Springer; 2015. p. 481–501.

[37]Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature reviews genetics. 2011;12(1):56.

[38]Terry SF. The global alliance for genomics & health. Genetic testing and molecular biomarkers. 2014;18(6):375–376.

[39]Microarray standards at last. Nature. 2002;419(323).

[40]Barrett T, Suzek TO, Troup DB, et al. NCBI GEO: mining millions of expression profiles—database and tools. Nucleic acids research. 2005;33(suppl_1):D562–D566.

[41]Hong EL, Sloan CA, Chan ET, et al. Principles of metadata organization at the ENCODE data coordination center. Database. 2016;2016:baw001.

[42]Fernández JM, de la Torre V, Richardson D, et al. The BLUEPRINT data analysis portal. Cell systems. 2016;3(5):491–495.

[43]Zhang J, Bajari R, Andric D, et al. The international cancer genome consortium data portal. Nature biotechnology. 2019;37(4):367–369.

[44]1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68.

[45]Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317.

[46]Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. Nature genetics. 2013;45(6):580.

[47]Harrison PW, Alako B, Amid C, et al. The European nucleotide archive in 2018. Nucleic acids research. 2019;47(D1):D84–D88.

[48]Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nature genetics. 2015;47(7):692–695.

[49]Tryka KA, Hao L, Sturcke A, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic acids research. 2014;42(D1):D975–D979.

[50]Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113.

[51]Kodama Y, Shumway M, Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. Nucleic acids research. 2011;40(D1):D54–D56.

[52]Athar A, Füllgrabe A, George N, et al. ArrayExpress update–from bulk to single-cell expression data. Nucleic acids research. 2018;47(D1):D711–D715.

[53]Sarkans U, Gostev M, Athar A, et al. The BioStudies database—one stop shop for all data supporting a life sciences study. Nucleic acids research. 2017;46(D1):D1266–D1270.

[54]Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30(1):207–210.

[55]Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic acids research. 2008;37(suppl_1):D885–D890.

[56]Sayers EW, Agarwala R, Bolton EE, et al. Database resources of the national center for biotechnology information. Nucleic acids research. 2019;47(Database issue):D23.

[57]Kodama Y, Mashima J, Kosuge T, et al. DNA data bank of japan: 30th anniversary. Nucleic acids research. 2017;46(D1):D30–D35.

[58]Rigden DJ, Fernández XM. The 2019 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Research. 2019;47(D1):D1–D7.

[59]Barrett T, Clark K, Gevorgyan R, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic acids research. 2011;40(D1):D57–D63.

[60]Courtot M, Cherubin L, Faulconbridge A, et al. BioSamples database: an updated sample metadata hub. Nucleic acids research. 2018;47(D1):D1172–D1178.

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

[62]Malladi VS, Erickson DT, Podduturi NR, et al. Ontology application and use at the ENCODE DCC. Database. 2015;2015:bav010.

[63]Hitz BC, Rowe LD, Podduturi NR, et al. SnoVault and encodeD: A novel object-based storage system and applications to ENCODE metadata. PloS one. 2017;12(4):e0175310.

[64]Gabdank I, Chan ET, Davidson JM, et al. Prevention of data duplication for high throughput sequencing repositories. Database. 2018;2018:bay008.

[65]Adams D, Altucci L, Antonarakis SE, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nature biotechnology. 2012;30(3):224.

[66]Lizio M, Harshbarger J, Shimoji H, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome biology. 2015;16(1):22.

[67]Gao GF, Parker JS, Reynolds SM, et al. Before and after: Comparison of legacy and harmonized TCGA genomic data commons' data. Cell systems. 2019;9(1):24–34.

[68]Hoadley KA, Yau C, Hinoue T, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173(2):291–304.

[69]Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503.

[70]Clarke L, Fairley S, Zheng-Bradley X, et al. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic acids research. 2016;45(D1):D854–D859.

[71]Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic acids research. 2018;47(D1):D766–D773.

[72]O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2015;44(D1):D733–D745.

[73]Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285.

[74]Karczewski KJ, Weisburd B, Thomas B, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic acids research. 2017;45(D1):D840–D845.

[75]Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv. 2019;p. 531210.

[76]Bujold D, de Lima Morais DA, Gauthier C, et al. The international human epigenome consortium data portal. Cell systems. 2016;3(5):496–499.

[77]Cerami E, Gao J, Dogrusoz U, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer discovery. 2012;2(5):401.

[78]Zheng R, Wan C, Mei S, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic acids research. 2018;47(D1):D729–D735.

[79]Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic acids research. 2018;47(D1):D941–D947.

[80]Albrecht F, List M, Bock C, et al. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. Nucleic acids research. 2016;44(W1):W581–W586.

[81]Reynolds SM, Miller M, Lee P, et al. The ISB Cancer Genomics Cloud: a flexible cloud-based platform for cancer genomics research. Cancer research. 2017;77(21):e7–e10.

[82]Dréos R, Ambrosini G, Groux R, et al. MGA repository: a curated data resource for ChIP-seq and other genome annotated data. Nucleic acids research. 2017;46(D1):D175–D180.

[83]Lau JW, Lehnert E, Sethi A, et al. The Cancer Genomics Cloud: collaborative, reproducible, and democratized—a new paradigm in large-scale computational research. Cancer research. 2017;77(21):e3–e6.

[84]Goldman M, Craft B, Brooks A, et al. The UCSC Xena Platform for cancer genomics data visualization and interpretation. BioRxiv. 2018;p. 326470.

[85]Mei S, Qin Q, Wu Q, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic acids research. 2016;45(D1):D658–D662.

[86]Masseroli M, Pinoli P, Venco F, et al. GenoMetric Query Language: a novel approach to large-scale genomic data management. Bioinformatics. 2015;31(12):1881–1888.

[87]Masseroli M, Kaitoua A, Pinoli P, et al. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. Methods. 2016;111:3–11.

[88]Cumbo F, Fiscon G, Ceri S, et al. TCGA2BED: extracting, extending, integrating, and querying The Cancer Genome Atlas. BMC bioinformatics. 2017;18(1):6.

[89]Bernasconi A, Ceri S, Campi A, et al. Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data. In: Mayr HC, Guizzardi G, Ma H, et al., editors. Conceptual Modeling. Cham: Springer International Publishing; 2017. p. 325–339.

[90]Bernasconi A, Canakoglu A, Colombo A, et al. Ontology-Driven Metadata Enrichment for Genomic Datasets. In: Baker CJO, Waagmeester A, Splendiani A, et al., editors. International Conference on Semantic Web Applications and Tools for Life Sciences. vol. 2275 of CEUR Workshop Proceedings; 2018. .

[91]Bernasconi A, Canakoglu A, Ceri S. Exploiting Conceptual Modeling for Searching Genomic Metadata: A Quantitative and Qualitative Empirical Study. In: Guizzardi G, Gailly F, Suzana Pitangueira Maciel R, editors. Advances in Conceptual Modeling. Cham: Springer International Publishing; 2019. p. 83–94.

[92]Stark Z, Dolman L, Manolio TA, et al. Integrating genomics into healthcare: a global responsibility. The American Journal of Human Genetics. 2019;104(1):13–20.

[93]Collins FS, Varmus H. A new initiative on precision medicine. New England journal of medicine. 2015;372(9):793–795.

[94]Cyranoski D. China embraces precision medicine on a massive scale. Nature News. 2016;529(7584):9.

[95]Genome Denmark;. Available from: http://www.genomedenmark.dk/english/. (20 November 2019, date last accessed).

[96]Leitsalu L, Metspalu A. From Biobanking to Precision Medicine: The Estonian Experience. In: Genomic and precision medicine. Elsevier; 2017. p. 119–129.

[97]Qatar Genome;. Available from: https://qatargenome.org.qa/. (20 November 2019, date last accessed).

[98]Caulfield M, Davies J, Dennys M, et al. The National Genomics Research and Healthcare Knowledgebase. figshare; 2017. Available from: https://figshare.com/articles/GenomicEnglandProtocol_pdf/4530893/5.

[99]Abu-Elmagd M, Assidi M, Schulten HJ, et al. Individualized medicine enabled by genomics in Saudi Arabia. BMC medical genomics. 2015;8(1):S3.

[100]Alkan C, Kavak P, Somel M, et al. Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. BMC genomics. 2014;15(1):963.

[101]FinnGen Research Project;. Available from: https://www.finngen.fi/. (20 November 2019, date last accessed).

[102]Genome Canada;. Available from: https://www.genomecanada.ca/. (20 November 2019, date last accessed).

[103]Dankar FK, Ptitsyn A, Dankar SK. The development of large-scale de-identified biomedical databases in the age of genomics—principles and challenges. Human genomics. 2018;12(1):19.