**Data Science for Genomic Data Management: Challenges, Resources, Experiences**

*Stefano Ceri and Pietro Pinoli*

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano

First.last@polimi.it

ORCID: SC - 0000-0003-0671-2415
PP - 0000-0001-9786-2851

Abstract

Purpose: We highlight several challenges which are faced by data scientists who use public datasets for solving biological and clinical problems. In spite of the large efforts in building such public datasets, they are dispersed over many sources and heterogeneous for their formats and sequencing/calling techniques, often meeting highly variable quality standards. Moreover, for most research questions, scientists hardly find datasets with enough samples for building and training machine learning models. Data scarcity depends on the complexity of the genomic domain, with its multi-facets, as well as the lack of organic initiatives to provide standardization across communities.

Method: In this paper we discuss our approach to genomic data management, that can strongly improve the problems of data dispersion and format heterogeneity through high-level abstractions for genomics. We briefly present the computational resources that were recently developed by the GeCo project (ERC Advanced Grant); they include GDM, a Genomic Data Model providing interoperability across data formats; GMQL, a genometric query language for answering data science queries over genomic datasets; and an in-house integrated repository providing attribute-based and keyword-based search over normalized metadata from several open data repositories.

Results and Discussion: We describe these resources at work on a specific research question, and we highlight how we managed to produce a model for addressing such specific research question by overcoming the lack of sufficient samples and labelled datasets.

## 1. CHALLENGES

Data science for genomics suffers from classic problems: data dispersion over many repositories, lack of documentation, presence of heterogeneus and hardly comparable data formats. All these problems are present in most data science scenarios; they require investments in data mapping, normalization and cleaning, which are known as the most laborious data science activities. In addition to these problems, and in spite of the availability of Next Generation Sequencing (a huge technological revolution in DNA processing), we were surprised to realize that data scarcity is an even bigger and unexpected obstacle, preventing data scientists from solving crucial problems, e.g. concerning precision medicine.

Restricting our analysis to genomics of the human population, researchers who work on cancer appreciate the availability of many resources within Genomic Data Commons (GDC) [1] or the International Cancer Genome Consortium (ICGC) [2], which as of today store thousands of experiments, typically reported both with raw data and post-processed format. Specifically, GDC in turns integrates datasets from The Cancer Genome Atlas (TCGA) [3] and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [4] accounting for more than 40 cancer types originating from 30,000 patients. For example, for what it concerns RNA-seq, the NGS experiment to measure gene expression, TCGA publicly provides more

than 10,000 high quality samples, associated with well curated and extremely valuable clinical information of the patients.
However, if one targets a specific cancer type and within it a specific patient population (e.g. wants to compare patients with a given type of cancer at different cancer stages), then numbers immediately turn to few tens or hundreds. If then one needs to relate heterogeneous experiments to the same patient, or to compare normal and cancer cells within the same patient, numbers further decrease. For example, it is often useful to compare the tissues expressing the tumour condition with the normal tissues, so as to understand which aberrations are produced by cancer. Unfortunately, in general normal cells collected for the patients affected by a cancer are much less than cancer cells.

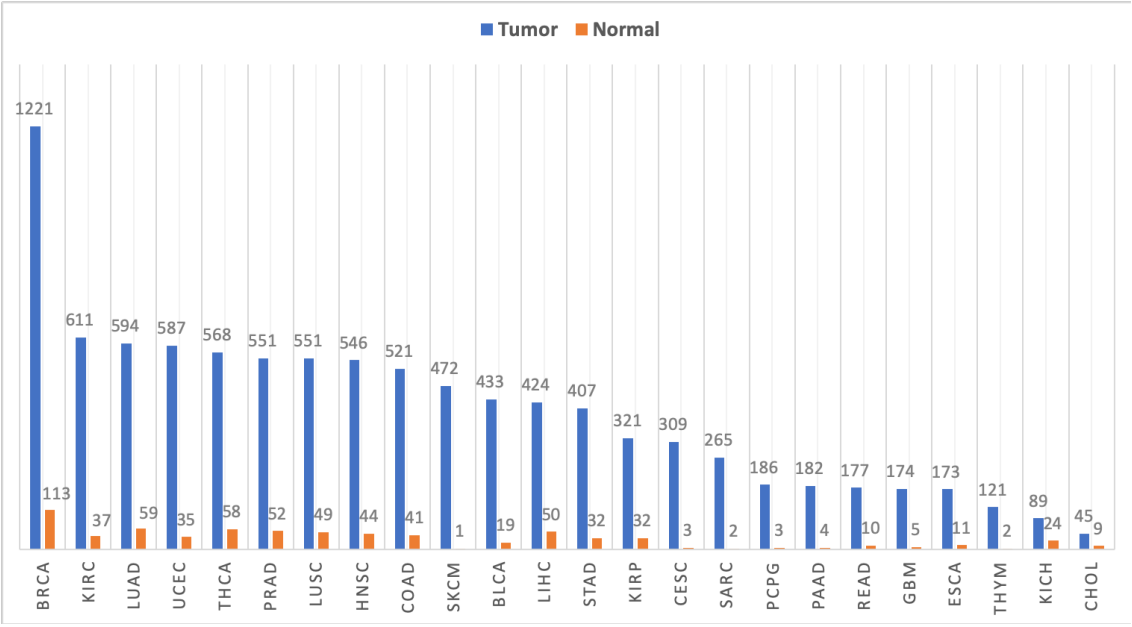


*Figure 1 Number of RNA-seq experiments for 24 out of 33 tumor types. In BLUE are reported the number of tumor samples; in ORANGE the number of match normal samples.*

Figure 1 reports the number of tumour samples (in blue) and the subset of the samples with an associated experiment on normal tissues (in orange) for the 24 tumour types for which at least one normal sample exists. On average, less than the 10% of the tumour samples have an associated normal sample and in most of the cases, tumour-normal associated samples are too few for a sound inference using classical statistical, data mining or machine learning tools.

Another obstacle to data integration comes from the adoption of incompatible types of measures within different data collections. For instance, another source called Genotype-Tissue Expression (GTEx) [5] contains gene expression for different tissues from healthy individuals; thus, in order to mitigate the scarcity of normal tissue data, one would be tempted to integrate the gene expression data for normal cells of cancer patients from TCGA with gene expression of the tissue of origin of the same cancer type. The scientist will find this to be very hard due to the use of different normalization criteria in TCGA and GTEx. Only recently, a package in R/Bioconductor [6] has been released to provide a solution to this problem, that however requires performing case-by-case batch corrections.

Similar difficulties arise when comparing gene expression produced by different technologies (e.g., microarray vs. next generation sequencing) or processed by different laboratories which may adopt different protocols for primary and secondary analysis of data. One interesting example is that of a curated collection of datasets for ovarian cancer, namely the *curatedOvariaData* [7], that merges 30 datasets of gene expression of ovarian cancer patients produced by different technologies and independent laboratories; despite of the valuable effort to integrate metadata of different datasets, these refer to different information items, therefore the resulting merged metadata has many missing values. Suppose a data scientist wants to investigate the difference between two sub-populations of patients (women sensitive and woman resistant

to platinum based therapy), she will only find 2 datasets (GSE30161 and GSE9891) that provide the relevant metadata to discern the two classes of patients and that have been collected by the same microarray platform. Nevertheless, a simple Principal Component Analysis, as reported in Figure 2, will show how the two datasets have very strong batch effects, which compromise their integration.

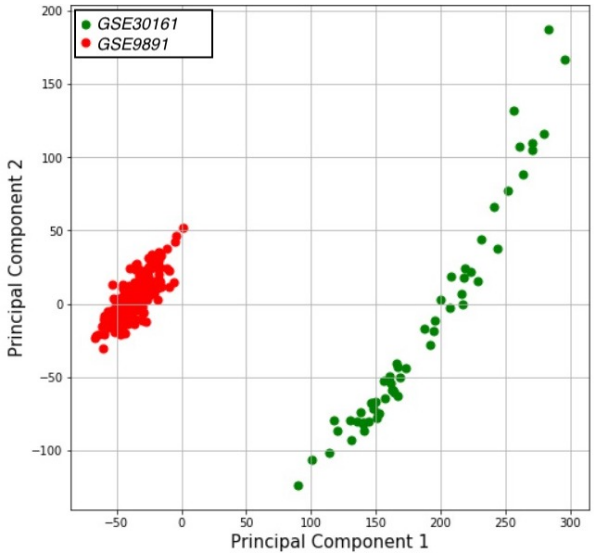


*Figure 2 First two components of the PCA on the two selected datasets form the Curated Ovarian Datasets highlights strong batch effects, that hinders the integration of the two.*

Differences in signal calling pipelines are also affecting the compatibility of signals extracted within dataset collected by the same consortium, to the point that the Cistrome project at Harvard [8] is investing huge amounts of efforts in reprocessing the signal extraction from raw data with higher quality standards of signal calling; these datasets are accessible through data browser, but they are not open for secondary access, therefore their full integration within our integrated repository is not possible.

Data scarcity in the context of a specific research question results in strong limitations in the adoption of classical data science tools, such as Machine Learning (ML), given the unbalance between the huge number of features (e.g., ~60,000 transcripts, millions of DNA mutations) and the very limited amount of samples. Such difficulties explode when trying to use more complex Deep Learning (DL) models, which require to estimate an enormous number of parameters. The biggest data collections (for example ARCHS4 [9] that contains more than 130,000 RNA-seq experiments) are strong candidates for adopting ML-based analysis, but the feature space is so big and the number of classes so high that using DL methods hardly outperforms traditional ML methods.

In most cases, computational experts supporting the biologists and clinicians usually ends up designing specialized statistical tests, able to cope with limited amount of data, rather than using out-of-the-box ML or DL tools.

## 2. GECO RESOURCES

So far, the bioinformatics research community has been mostly challenged by *primary analysis* (production of sequences in the form of short DNA segments, or ''reads'') and *secondary analysis* (alignment of reads to a reference genome and search for specific genomic features on the reads, such as variants/mutations and peaks of expression). The most important emerging problem is the so-called *tertiary analysis*, concerned with sense making, e.g., discovering how heterogeneous regions interact with each other, by integrating heterogeneous DNA features, such as variants or mutations in a DNA position, or signals and peaks of

expression, or structural properties of the DNA, e.g., break points (where the DNA is damaged) or junctions (where DNA creates loops). According to many biologists, answers to crucial genomic questions are hidden within genomic data already available in public repositories, but suitable tools for processing them are lacking. Our research in genomic data management is developed within a EU-funded project[1] whose main goal is to provide better abstractions, languages, systems, and tools for fostering genomic data integration during tertiary data analysis. A bird eyes' view of the project is shown in Figure 3. The biologists and clinicians may interact with two main systems, delivered by our project.

The former is a **repository of public datasets**, where we integrate data from ENCODE [10], GDC [2], Roadmap Epigenomics [11] and Cistrome [8] (we plan to add more sources while the project continues). The important aspect of this repository is the format unification achieved across the different sources. Data files available at the sources are transformed to a same representation, called the **Genomic Data Model**, GDM [12], which essentially forces every data type used by the data files to become a mapping from regions to a data type-specific feature vector. Format transformations come as the result of significant efforts: for instance, the transformation of TCGA-supported data types to GDM is a long process, with several syntactic and semantic transformations (see TCGA2BED [13].) Metadata, i.e. descriptive information associated to each data file, are also transformed to a unique conceptual representation (called the **Genomic Conceptual Model**, GCM [14]), which includes 8 entity types and about 40 attributes.

The latter is a cloud-based data manager for region-based data, supporting a new query language for genomics, called **GenoMetric Query Language**, GMQL [15]. The language derives from classical abstractions of relational databases and is the composition of orthogonal operations, which apply to either one or two datasets. Unary operations include Selection, Projection, Merge, Group, Extract, Sort, Cover; binary operations include, Join, Map, Union, Difference. While some of the operation are rather straightforward extensions of relational abstractions to the GDM model, other operations are domain-specific; among them, Cover (extracting regions based on their accumulation along the genome), Map (extracting aggregate properties of regions of one dataset that overlap with regions of another one, used as reference) and Join (extracting regions from two datasets that satisfy distance-based predicates). Another peculiarity of the language is the support of both region and metadata processing; the latter allows the progressive construction of the description of query results from the description of the operands.

The associated **GMQL query system** [15] has a modular architecture including an intermediate representation supporting operations over regions and metadata which are executed by the Apache Spark engine, a data frameworks on the cloud that proved to be extremely efficient in supporting massive genomic queries [16], with a high-level technology-independent repository abstraction, supporting different repository types (e.g., local file system, Hadoop File System, or others), several system interfaces, including an intuitive public Web-based interface[2], as well as two programmatic interfaces: a pyGMQL library for Python[3] and a RGMQL package[4] for the R/Bioconductor environment.

Further extensions of our GMQL system will support data sharing and federated processing of data located in distinct GMQL instances running on different systems/clouds, by automatically splitting and transferring the processing where the data to be evaluated are located. This will avoid as much as possible expensive downloads and transfers of big data quantities between repositories, also allowing taking advantage of processing resources available in different organizations.

[1] Advanced ERC Grant N. 693174, "GeCo", data-driven Genomic Computing, 2016-2021.
[2] http://www.gmql.ue/
[3] https://pygmql.readthedocs.io/
[4] https://bioconductor.org/packages/release/bioc/html/RGMQL.html

Figure 3 shows the interactions supported by the two systems; we expect computational biologists and bioinformaticians to interact with it from their desktop interfaces. While they normally access the individual data sources and then are left with the burden of data integration, in our environment they can connect to a Web-based search interface, called **GenoSurf**, that operates upon the integrated metadata extracted from all the sources and converted into GCM; GenoSurf performs attribute-based and keyword-bases searches and selects the datasets of interest. Such datasets are integrated thanks to complex data integration pipelines, that periodically test the data sources and load new data files within existing datasets; they reside within the big data repository, where each use can also load private datasets. Once the relevant datasets are selected, they can be inspected by using the GMQL language, supported by a variety of interfaces: a Web interface, two language embeddings for Python and R, two workflow-based embedding for Galaxy and FireCloud. Query execution is in any case supported by Spark, supported by major cloud vendors (e.g. Google, Amazon) and by the servers running Hadoop and HDFS.

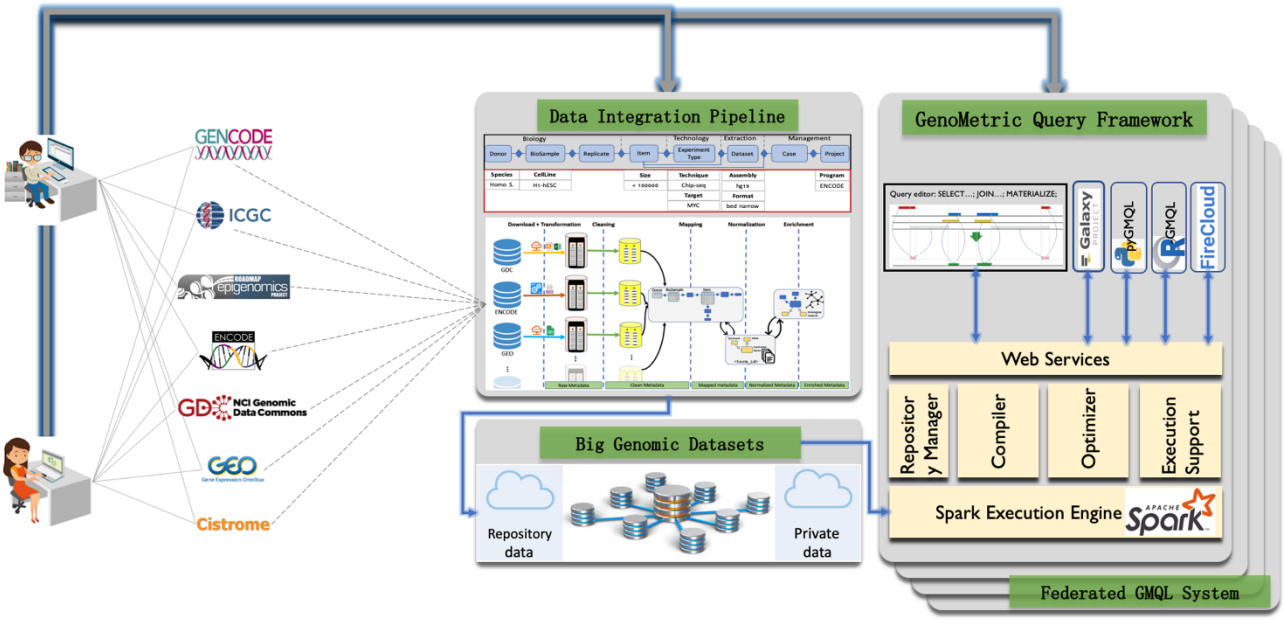


*Figure 3 Bird eyes' view on the GeCo project, highlighting both the resources which are made available by the project and their internal architecture; the repository hosts an attribute-based interface for supporting end-user interactions and a data preparation pipeline which runs periodically, whenever sources add new datasets or files; the query framework includes a variety of user interfaces all served by a Web Service layer, which wraps classical data management components (repository manager, compiler, optimizer, execution support) all acting upon a cloud-based architecture which is based upon the Spark Execution Engine.*

## 3. Experience

Recently, we approached several biological questions by using GeCo models and tools, including research on gene expression (precision medicine for the early identification of patient subclasses), transcription factors (mutual interactions and interplay with protein-protein networks) and mutations (synthetic lethality, pancancer analysis of mutations affecting topological domain boundaries). We next describe one of these studies, concerning the following research question: "Given a cell line and a pair of transcription factors, do they cooperate in the regulation of gene expression?" – this experience shows both the challenges in dealing with data scarcity and the use of GeCo resources.

Transcription factors (TFs) are proteins that enter the cell nucleus and bind to the DNA and contribute to many different molecular functions, such as organization of the chromatin, creation of DNA 3D structure of the genome and regulation of gene expression by enhancing or repressing their transcription. In our model, two TFs cooperate to gene regulation if they form a protein complex that binds the DNA in genes' promoters, short DNA sequences just upstream the start of a gene. Therefore, we would expect to observe those two TFs binding the genome close to each other, in particular in correspondence of promotorial regions.

Besides wet-laboratory experiments to identify protein complexes, some analytical strategies have been proposed in the past. In particular, it is known that every TF can bind to the DNA only in specific positions where the sequence has a specific pattern of nucleotides, called motif, that make such binding possible; analytical studies were used to identify on the reference sequence of the genome those locations which are dense of positions where two TF could be close, regardless of the specific DNA activity. Such approach, however, does not reflect the actual bindings that a TF may have in specific cells or biological conditions; these factors are known to influence TFs cooperation. in our study, we instead considered active bindings, i.e. positions where it is possible to experimentally determine that the protein effectively binds to the DNA.

The first step, of course, consisted in the identification of valuable sources of data; we found the NarrowPeak collection of ChIP-seq experiments provided by the ENCODE consortium to be appropriate for this research. ChIP-seq is NGS experimental technique executed on a biological sample (typically, an immortalized cell line) targeting a given protein; the outcome of the experiment consists in a list of positions within the genome of the cell line where the target protein is actively binding the DNA. Despite the fact that ENCODE is the largest collection of high quality curated ChIP-seq experiments, we could identify only three human cell lines (namely K562, GM12878 and HepG2) for which a sufficiently high number of experiments had been collected. This was the first limitation derived by data scarcity that we faced.

The data extraction and cleaning phase involved GMQL, and was aimed at selecting only reliable ChIP-seq regions (filtering them on the associated signal quality score) and merging replicas of the same experiment. Then, as a first attempt, we tried to associate to each pair of TFs a feature vector, so as to apply supervised learning. Such feature vectors comprised statistics on the frequency of binding of the two TFs as well as cross statistics such as the average, the median and the standard deviation of the reciprocal distance at which the two TFs binds the DNA. Unfortunately, this classic data science approach did not work, because in literature only a small fraction of the TF pairs is already known and therefore, we had very few labelled data (order of tens, compared to thousands TF-TF pairs).

Using unsupervised learning tools we got a similar outcome. Whatever set of features we tested, it was not possible to strictly divide the pairs of TFs in two classes (i.e., the ones forming a complex and the ones not), indeed biological processes are usually complex phenomena, with a continuous set of possible outcomes. For example, the binding of two TFs may be influenced by the presence of a third TF competing for either one of the two; therefore. the two still interact but the effect of such interaction is much less observable.

Finally, we developed a novel approach which can be considered as prototypical of a bottom-up data science approach for biological research. We computed the minimal distances at which the two TFs are found on the DNA; then we performed a bootstrap to estimate the null distribution (which represent the mean value of our observation), and we developed a test allowing to predict the two classes of cooperating vs non-cooperating TF pairs; furthermore, and more important, we could associate a confidence value to the prediction. After considering several features as candidates, we finally discovered that the short and long tails of the distributions best served the cause of dividing the two classes; the two tails correspond to the frequency at which pairs of binding sites where observed at increasing distance (up to 2,250, that in our model corresponds to the length of promoters). Comparing such frequencies with the corresponding frequencies in the null distribution led to satisfactory results and to a good predictor [17]. In Figure 4, we show the two distributions for prototypical cases of known pairs of TF which respectively cooperate (Myc, MAX) and do not cooperate (CTCF and Myc).

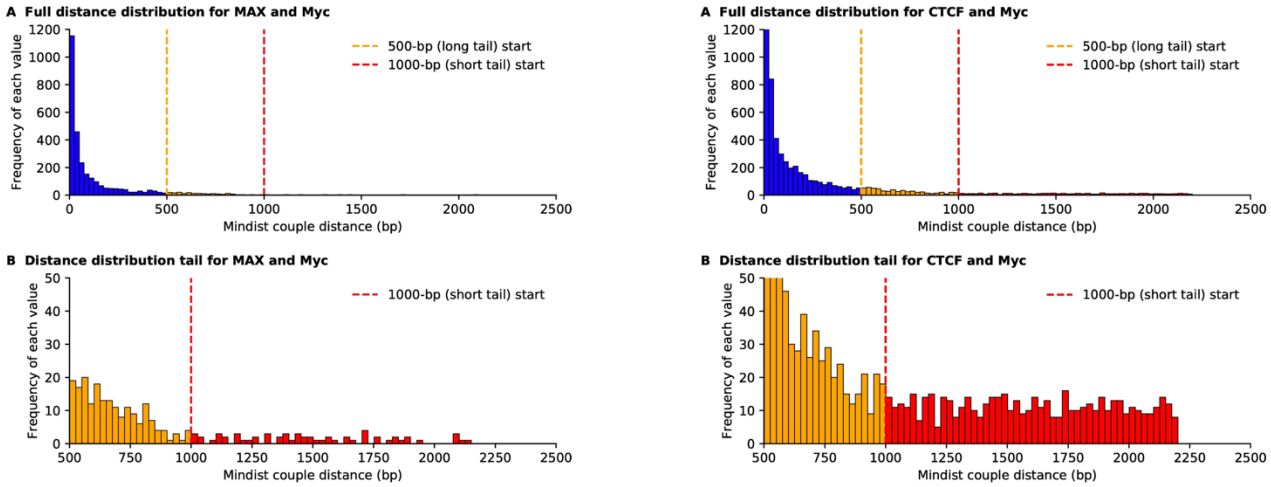


*Figure 4 The short and long tails of the distribution of minimal distances between pairs of active TF can be used to separate cooperating pairs from non-cooperating pairs. Panel on the left shows a pair of TFs (MAX and Myc) that are known to cooperate, while panel on the right show a pair of TFs that do no cooperate (Myc and CTCF).*

## 4. Discussion

Compared to classic data science, in genomics we often do not have enough labelled data; even when observed features of biological phenomena are available, they are not strictly black or white. Moreover, even for "simple" observations (e.g., a gene is over/down expressed) there could be many confounding effects, mostly unknown or even unmeasurable. Building a complete model that, for example, describes all the details of how a given TF contributes to gene regulation is typically impossible. Many top-quality papers, even in good journals, describe partial models and then provide an experimental validation which is based on few examples confirming that the theoretical approach is plausible, but cannot really provide a firm, statistical validation.

Thus, building a model that, for example, describes how a certain factor, say a TF, participates in the regulation of the gene is practically impossible. In these cases, we resort to statistical methods for associating scores (e.g., p-values, enrichments, folds), and these are of paramount importance to understand and analyse data. Typically, we build a null distribution by bootstrapping; for example, if we were asked to assess the relationship between the given TF and the expression of the given gene, we would start by building a null distribution from all the available TFs and genes; such null would depict the expected contribution of TFs to genes' expression and then it would be possible to understand if the role of our targets TF/gene pair is significant or not.

Of course, these difficulties in applying classic data science to genomics may significantly reduce in the near future. According to [18], by 2025 genomics will become the biggest source of big data, as the total size of genomic raw data will exceed by 20 to 40 times the size of astronomical data (the current leading scientific discipline for mass data production) as well as the total of videos loaded on YouTube channels. Projects such as the 100K genomes[5] of the ongoing effort for sequencing 500K Finnish citizens[6] are remarkable examples of the ongoing trend. While several independent bodies are producing very large datasets, a parallel effort should be spent on data standardization and normalization, so as to empower the community of biological and clinical researchers with datasets which are not just very large, but also comparable and of high quality.

[5] doi:10.6084/m9.figshare.4530893.v2
[6] https://www.finngen.fi/en

References


[1] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe e L. M. Staudt, «Toward a Shared Vision for Cancer Genomic Data,» *New England Journal of Medicine ,* vol. 375, n. 12, pp. 1109-1112, 2016.

[2] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty e M. Wong-Erasmus, «International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data,» *Database,* 2011.

[3] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart e Cancer Genome Atlas Research Network, «The cancer genome atlas pan-cancer analysis project,» *Nature Genetics,* vol. 45, n. 10, p. 1113, 2013.

[4] TARGET, «NIH-TARGET,» [Online]. Available: https://ocg.cancer.gov/programs/target.

[5] GTEx Consortium, «The genotype-tissue expression (GTEx) project.,» *Nature Genetics, 45(6),* p. 580, 2013.

[6] M. Mounir, M. Lucchetta, T. C. Silva, C. Olsen, G. Bontempi, X. Chen, H. Noushmehr, A. Colaprico e E. Papaleo, «New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx,» *PLoS computational biology,* vol. 15, n. 3, 2019.

[7] B. F. Ganzfried, M. Riester, B. Haibe-Kains, T. Risch, S. Tyekucheva, I. Jazic, X. V. Wang, M. Ahmadifar, M. J. Birrer, G. Parmigiani e C. Huttenhower, «curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome,» *Database,* 2013.

[8] S. Mei, Q. Qin, Q. Wu, H. Sun, R. Zheng, C. Zang, M. Zhu, J. Wu, X. Shi, L. Taing e T. Liu, «Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse.,» *Nucleic acids research ,* vol. 25, 2016.

[9] A. Lachmann, D. Torre, A. B. Keenan, K. M. Jagodnik, H. J. Lee, L. Wang, M. C. Silverstein e A. Ma'ayan, «Massive mining of publicly available RNA-seq data from human and mouse,» *Nature communications,* vol. 9, n. 1, p. 1366, 2018.

[10] ENCODE Project Consortium, «The ENCODE (ENCyclopedia of DNA elements) project,» *Science 306(5696),* pp. 636-640, 2004.

[11] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, A. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker e P. J. Farnham, «The NIH roadmap epigenomics mapping consortium,» *Nature biotechnology ,* vol. 28, n. 10, p. 1045, 2014.

[12] M. Masseroli, A. Kaitoua, P. Pinoli e S. Ceri, «Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying,» *Methods,* n. 111, pp. 3-11, 2016.

[13] F. Cumbo, G. Fiscon, S. Ceri e M. Masseroli, «TCGA2BED: extracting, extending, integrating, and querying The Cancer Genome Atlas,» *BMC bioinformatics,* p. 6, 2017, 18(1).

[14] A. Bernasconi, S. Ceri, A. Campi e M. Masseroli, «Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data,» in *ER*, 2017.

[15] M. Masseroli, A. Canakoglu, P. Pinoli, A. Kaitoua, A. Gulino, O. Horlova, L. Nanni, A. Bernasconi, S. Perna, E. Stamoulakatou e S. Ceri, «Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data,» *Bioinformatics,* 2018.

[16] M. Bertoni, S. Ceri, A. Kaitoua e P. Pinoli, «Evaluating cloud frameworks on genomic applications,» in *EEE International Conference on Big Data (Big Data)*, Santa Clara (CA), US, 2015.

[17] S. Perna, P. Pinoli, S. Ceri e L. Wong, «TICA: Transcriptional Interaction and Coregulation Analyzer,» *Genomics, proteomics & bioinformatics ,* vol. 16, n. 5, pp. 343-353, 2018.

[18] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha e G. E. Robinson, «Big data: astronomical or genomical?,» *PLoS biology ,* vol. 13, n. 7, 2015.