Journal of Orthopedic Oncology
ISSN: 2472-016X

Barbosa, J Orthop Oncol 2016, 2:2
DOI: 10.4172/2472-016X.1000112

Short Communication Open Access

Challenges with Big Data in Oncology

Carmela Dantas Barbosa*

Centre Leon Berard, CRCL, UMR INSERM 1052-CNRS 5286, Team Targeting the Tumor and its Immune Environment, Cheney D, 3rd Floor, 28 Rue Laennec69373, Lyon Cedex 08, France

Abstract

We are in the era of large-scale science. In oncology there is a huge volume of data sets grouping information on cancer genome, transciptome, clinical data and more. The challenge of big data in cancer is to integrate all this diversity of data collections into a unique platform that can be analyzed leading to the generation of readable files. The possibility of harnessing information from all the accumulated data lead to an improvement in cancer patient treatment and outcome. Few examples of successful implementation of bioinformatics platforms do prove the challenge being as big as the data, but possible to be overcome.

Keywords: Big data in oncology; Cancer; Bioinformatics; Proteomics; Biomarkers; Breast cancer

It is largely accepted between scientist and physicians that improvement in cancer treatment depends on a better knowledge about cancer biology. The development of so-called "omics" approaches, including genomics, transcriptomics, proteomics, epigenomics, just to name few, are improving our understanding about tumors biology, and at the same time, are generating a huge volume of data in cancer. The generation of big data in cancer comes mainly from the high-throughput technologies used to study the "omics" sciences. The genomics concern genome sequence, structures, mutations, repeat contents and evolution. The next-generation sequencing (NGS), or high-throughput sequencing, is a recent technology that allows DNA and RNA sequencing much faster and cheaply than the previously used method [1]. The NGS FASTQ files are very big. A project containing about 10 - 20 whole genome sequencing (WGS) samples can generate approximately 4TB of raw data. It is the "data deluge" problem. The transcriptomic analyses the totality of RNA transcripts produced by a genome. Since transcription is regulated, the transcriptome can be modulated under specific circumstances, allowing the study of genes differentially expressed in different populations of cells, or under different treatments. These kinds of studies have been largely applied in the field of oncology. Transcriptomic analyses are made by high-throughput methods, like real-time quantitative PCR (qPCR), microarrays and RNA-Seq (NGS sequencing). Proteomics is the study of a specific proteome, including information on protein structure and function, protein expression profiling, their variations and modifications, aiming to understand cellular processes [2]. Proteomics studies have been particularly applied on cancer studies to identify specific biomarkers linked to diagnosis, prognosis and therapeutic prediction. The high-throughput proteomics analysis is based on mass spectrometry (MS). The epigenomics studies the epigenetics modifications of the genetic contents of a cell, known as epigenome. Tumorigenesis is part of the cellular processes regulated by epigenetic modifications. NGS has allowed the development in genome characterization, including the identification of epigenetic modifications. Along with clinical databases, containing the diagnosis, treatments and patient outcomes and clinical trials information, they constitute the big data collections of cancer patients.

The concept of "big data" is relatively new; it appeared in the early 2000s. Big data was defined, at first, as the three Vs: *for volume, velocity and variety of information.* To complete the definition a 4th V was added, it refers to veracity, i.e., reliability of the ac-cumulated data [3].

The challenge for cancer research is to better explore all data sets, from tumor biology and clinical information about patients. One difficulty is to compose a readable file joining all these multiple varieties of data types in a centralized platform to allow interfacing with each other. The volume of existing data is enormous and growing very fast. It is necessary to be able to harness then, asking complex questions to identify new knowledge in existing data. There is a growing need for new types of computing analytics. Actually, there are many databases enabling data sharing to make possible for anyone interested to improve our knowledge of the field, using the data collection. Following are few examples of existing databases to illustrate the new era of large-scale science.

The tumor suppressor gene *TP53,* which encodes the p53 protein, is the most commonly altered gene in human neoplasms [4]. For more than 30 years this gene have been widely studied in oncology. The International Agency for Research on Cancer (IARC), in Lyon, France disposes of a TP53 database containing exclusively TP53 mutations associated with human cancers [5]. The database was built from TP53 sequences published in the peer-reviewed literature or compiled in mutation data repositories, since 1989. Beyond TP53 sequences, there is also information about gene function, clinicopathologic features of tumors and patient information. The database can be fully downloaded and is available for scientists and clinicians. This database is a depository of TP53 accumulated knowledge. The latest issue, R18, released in April 2016, compiles data on over 29,000 somatic mutations and is an invaluable tool helping investigators to better understand the function of p53 protein in oncology.

The National Institutes of Health (NIH) created a complex project in 2006. The Cancer Genome Atlas (TCGA) Data Portal started as a three-year pilot project. Their objective was to create a vast and comprehensive data collection of mutations that occur in specific cancer types. The challenge was to create a national network pooling the results obtained by different research teams working in related projects.

***Corresponding author:** Carmela Dantas Barbosa, Centre Leon Berard, CRCL, UMR INSERM 1052-CNRS 5286, Team Targeting the Tumor and its Immune Environment, Cheney D, 3rd Floor, 28 Rue Laennec69373, Lyon Cedex 08, France, Tel: +33 610 455 397; E-mail: carmeladantas@hotmail.com

Received April 22, 2016; **Accepted** May 16, 2016; **Published** May 23, 2016

Citation: Barbosa CD (2016) Challenges with Big Data in Oncology. J Orthop Oncol 2: 112. doi: 10.4172/2472-016X.1000112

The success of the pilot encouraged NIH to add more investment, the project lasted 10 years. Actually they finalized the tissue collection. The data is public and freely available which enabled researchers all over the world to make and validate important discoveries. Indeed, thanks to this project, a big contribution was made in the understanding of cancer genome, for example in lung adenocarcinoma [6], or breast cancer [7]. The database contains samples of more than 11,000 patients and 33 tumor types, including rare diseases. The study of cancer genomics advances the personalized medicine. The target therapy is based on the utilization of a drug that targets specifically a genomic mutation. TCGA provided researchers with comprehensive catalogs of the key genomic changes in many major cancer types supporting clinicians with valuable information that enables a more effective diagnose and treatment. TCGA will close in 2016 but it paved the way for new cancer genomic projects based on its model.

Another important project of NIH is the Roadmap Epigenomics Mapping Consortium. The main objective is to produce a public resource of human epigenomic data. The consortium uses NGS technologies to study DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in different kinds of normal cells and tissues. The database provides a framework or reference of normal epigenomes that could be compared of tissues and organ systems frequently involved in human disease. The Roadmap allows scientists to make revelant discoveries and clinicians to improve medicine, based on epigenetics features of tumors. Important discoveries in cancer have been made using this database, for example an epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemia [8].

The big-data project CancerLinQ, from ASCO, will allow patients and physicians to share information about treatments and outcomes. It creates a continuous cycle of learning that begins and ends with the patient. One very important point is that patients are anonymized in databases to protect their identity. Patients and their doctors can contribute and gain from the accumulated information present in the database. Even if it seems obvious nowadays, they traveled a long way to get there. The CancerLinQ concept was established in 2010, two years later they created their first prototype to demonstrate the feasibility of the project concentrating only in breast cancer. By 2013, more then 170 000 medical records of breast cancer patients were gathered. So far, twelve sites, including cancer treatment centers, will share data through CancerLinQ. ASCO signed an agreement with SAP, a software company, to create the big data software platform that will allow CancerLinQ, using SAP's HANA technology [9]. CancerLinQ database can help physicians to improve patient treatment and outcome. For example, a type of cancer with a particular genetic mutation is found to develop resistance to a targeted therapy. The shared information can avoid the physician to use the same drug in another patient with the same mutation. For patients suffering from rare diseases, such as sarcomas, this kind of database can be particularly helpful. ASCO expects the first tool to be online during this year.

Not only consortiums, grouping different cancer centers, can produce big datasets. Important cancer centers, such as the MD Anderson Cancer Center are also concerned about the big data. To handle this problem, the big data and and APOLLO (Adaptive Patient-Oriented Longitudinal Learning and Optimization) platforms were created in the MD Anderson Cancer Center. The big data is an adaptive learning environment composed of the Institutional Longitudinal atient Disease Registry which securely houses clinical and omics data plus a suite of massive data analytics that can be interrogated and provide end users with understandable and actionable answers to their clinical or research questions. The APOLLO platform aims to create a more cohesive system for standardizing long-term collection of patients, clinical history and data derived from their biological samples. Both platforms are intended to enable physicians to practice science-based medicine, to improve even more the level of patient care in the MD Anderson and help physicians located worldwide to practice with the MD Anderson standard.

All the data that compose the big-data projects comes from patients and already translates into benefits to them. Solving the big data problem in oncology has multiple facets. We see in the various existing platforms the need for collaboration. One of the main challenges is how fast data can be analyzed when there is so much. How can we make sense of data? While there are a lot of initiatives, there is still room for improvement. Having great ways to collaborate is already a good start. The future is plenty of hopes. Serious works are being done and we shall think on how much we can contribute to one of those initiatives.

References

1. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456 3123: 53-59.
2. Anderson NL, Anderson NG (1998) Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis 19: 1853-1861.
3. Zhou Z, Chawla N, Jin Y, Williams G (2014) Big data opportunities and challenges: Discussions from data analytics perspectives. Computational Intelligence Magazine IEEE 9: 62-74.
4. Vogelstein B, Sur S, Prives C (2010) The Most Frequently Altered Gene in Human Cancers. Nature Education 3: 6.
5. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, et al. (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum Mutat 28: 622-629.
6. The Cancer Genome Atlas Research Network (2014) Comprehensive molecular profiling of lung adenocarninoma. Nature 511: 543-550.
7. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, et al. (2015) Comprehensive molecular portraits of invasive lobular breast cancer. Cell 163: 506-519.
8. Knoechel B, Roderick JE, Williamson KE, Zhu J, Lohr JG, et al. (2014) An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemia. Nature Genetics 46: 364–370.
9. Shah A, Stewart AK, Kolacevski A, Michels D, Miller R (2016) Building a Rapid Learning Health Care System for Oncology: Why CancerLinQ Collects Identifiable Health Information to Achieve Its Vision. J Clin Oncol 34: 756-763.