Big Data in Genomics: Challenges and Solutions

Is Life Sciences prepared for a big data revolution?

Every era has its technological breakthroughs. The widespread use of computers and the internet in the beginning of the 21st century has impacted the way we approach and search for information [1]. The emergence of social networks (examples include Facebook, Twitter, LinkedIn and others) and "cloud" solutions for data storage, with computer processor speed increasing at a fast pace has changed the way we generate information [1].

Fabrício F. Costa, Ph.D., Northwestern University's Feinberg School of Medicine, Chicago, IL, USA

Life Sciences have been highly affected by the generation of large data sets, specifically by overloads of omics information (genomes, transcriptomes, epigenomes and other omics data from cells, tissues and organisms). The use of DNA sequencing machines, which are smaller in size but capable of generating piles of data faster and at a lower cost, have changed science and medicine in ways never seen before [1]. The current era is beginning to look like the era of "big data"; a term that refers to the explosion of available information, which is a byproduct of the digital revolution [2]. However, with biomedical data accumulating in computers and servers around the world [1], concerns over privacy and security of patient data are emerging.

Next-Generation Sequencing (NGS) platforms that use semiconductors [3] or nanotechnology [4] have exponentially increased the rate of biological data generation in the last two years. While the first human genome was a $3 billion dollar project requiring over a decade to complete in 2002, we are now close to being able to sequence and analyze an entire genome in a few hours for less than a thousand dollars. The decrease in costs has enabled the generation of information at the petabyte (1012 bytes) scale. However, even though both computers and the internet have become faster, we have a lack of computational infrastructure that is needed to securely generate, maintain, transfer, and analyze large-scale information in life sciences and to integrate omics data with other data sets, such as clinical data from patients (mainly from Electronic Medical Records or EMRs). In this article, a short overview of the challenges faced by big data production, transfer and analysis will be given. In addition, the changing landscape of privacy and personal information in the era of big data will be discussed.

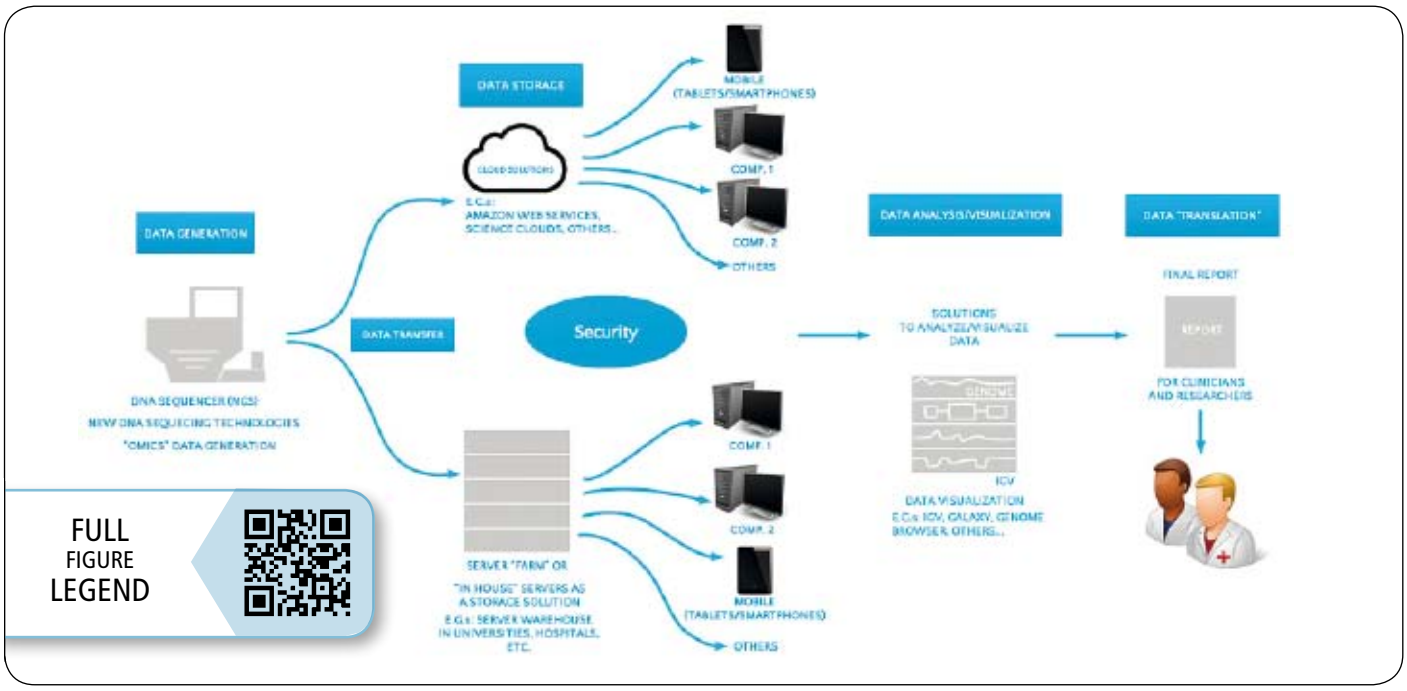

Fig. 1: Big Data in Genomics. Schematic representation of a pipeline from data generated using NGS to data "translation" for clinicians and researchers. Image Design: Eduardo Braga Ferreira Junior. For complete figure description see http://goo.gl/UxSSz

How did Big Data Become so Big?

Big Data has affected several unrelated sectors in society, including communications, media, medicine, and scientific research among others [2]. In science, for example, in less than 10 years, the time and cost of sequencing genomes was reduced by a factor of 1 million. Today, personal genomes can be mapped faster for a few thousand dollars. Personal genomics is a key enabler for predictive medicine, where a patient's genetic profile can be used to determine the most appropriate medical treatment. The Encode Project offers a nice perspective on Big Data generation and analysis with the participation of different research groups by providing an elaborated framework for personal genomics [5]. Projects such as Encode have product piles of data, illustrating how Big Data is becoming integral for scientific research [5]. Indeed, science today is increasingly "social", especially in fields such as genomics in which huge amounts of data are generated. Encode is a good training tool for researchers in big scientific enterprises that will become increasingly common. In these types of projects, tons of data are generated, stored, transferred and analyzed (see also figure 1 for a complete overview).

With the increased need to store data and information generated by big projects, computational solutions such as cloud-based computing have emerged. Cloud computing is the only storage model that can provide the elastic scale needed for DNA sequencing, whose rate of technology advancement could now exceed Moore's Law. In fact, cloud solutions from different companies have been used, but several challenges are posed by it, particularly related to the security and privacy of personal medical and scientific data, are posed by it (fig. 1, table 1). Perhaps the greatest advantage could be the ability to offer a broad platform for development of new analysis and visualization tools as well as a software service to use these tools on shared data sets in a secure and collaborative workspace [6]. In fact, some companies already offer such solutions (Table 1). There is also an opportunity for a version of an App or Google Play Store, specifically for genomics tools, from which hundreds of specialty applications could be developed [6]. Companies such as Illumina and 23andme already offer an open platform for developers and more companies will implement APIs (Application Programming Interfaces or APIs) in their services. However, solutions to overcome data privacy issues will be crucial.

Pipelines to deal with increasing amounts of genomics data will be needed to store, transfer, analyze, visualize, and generate "short" reports to researchers and clinicians (for more information see figure 1). In fact, an entirely new genomics industry could be made possible by cloud computing, which will transform medicine and life sciences. Cloud computing opens a new world of possibilities for the genomics industry to transform the way we approach research and medicine. However, one of the downsides of cloud computing is keeping the data private.

The Coming Age of Data-driven Science and Medicine

The understanding of how the underlying systems in living organisms operate will require the integration of many layers of biological information that high-throughput technologies are generating. The complexity of the data generated in scientific projects will only increase as we continue to isolate and sequence individual cells and organisms while lowering the costs to generate and analyze this data, such that hundreds of millions of samples can be profiled. Sequencing DNA, RNA, the epigenome and other omics from numerous cells in different individuals will take us to the exabyte (1018 bytes) data scale in the next 5 years or so [7]. Integrating all this data will demand high-performance computational environments like those at genome centers [7]. The integration between hardware and software infrastructures tailored to deal with big data in life sciences will become more common in the years to come.

Importantly, data-driven medicine will enable the discovery of new treatment options based on multi-model molecular measurements on patients and learning from the trends in differential diagnosis, prognosis and prescription side-effects in databases from clinical practice [8]. The combination of omics data with clinical information from patients will enable new scientific knowledge that could be applied in the clinics to help in patient care [8]. In addition, medical informatics, represented by EMRs from patients and personalized therapies will enable the application of targeted treatments for specific diseases. Thus, it is tempting to imagine how both scientific inquiry and patient care would be performed differently when dealing with Big Data repositories if large amounts of genomic and clinical data are collected and shared by health care professionals (fig. 1).

Challenges and Solutions

These revolutionary changes in Big Data generation and acquisition create profound challenges for storage, transfer and security of information. Indeed, it may now be less expensive to generate the data than it is to store it. One example of

Table 1: Examples of companies and Institutions that provide solutions to generate, store, analyze, and visualize omics and clinical data.

Company / Institution	Type of Solution	Website
Appistry	Appistry's high-performance big data platform combines self-organizing computational storage with optimized and distributed high-performance computing to provide secure, HIPAA-complaint accurate on-demand analysis of OMICS data in association with clinical information	www.appistry.com
BGI	BGI's solution serves as a solid foundation for large-scale bioinformatics processing. BGI computing platform is an integrated service composed of versatile software and powerful hardware applied to life sciences	www.genomics.cn/en
CLC Bio	CLC Bio bioinformatics has a platform where both desktop and server software are integrated and optimized for best performance. CLC Bio utilize proprietary algorithms, based on published methods, in order to successfully accelerate data calculations to achieve remarkable improvements in big data analytics	www.clcbio.com
DNAnexus	DNAnexus provides solutions for NGS by using cloud computing infrastructure with scalable systems and advanced bioinformatics in a web-based platform to solve data management and the challenges in analysis that are common in unified systems	www.dnanexus.com
Genome International Corporation	Genome International Corporation (GIC) is a research-driven company that provides innovative bioinformatics products and custom research solutions for corporate, government, and academic laboratories in life sciences	www.genome.com
GNS Healthcare	GNS Healthcare is a big data analytics company that has developed a scalable approach to deal with big data solutions that could be applied across the healthcare industry	www.gnshealthcare.com
Foundation Medicine	Foundation Medicine is a molecular information company on the forefront of bringing comprehensive cancer genomic analysis to routine clinical care. Foundation Medicine is pioneering the development of a comprehensive cancer diagnostic test combining OMICS data, clinical information and big data analytics applied to cancer research	www.foundationmedicine.com
Knome	Knome analyzes whole genome data using software-based tests simultaneously to examine and compare many genes, gene networks, and genomes as well as integrate other forms of molecular and non-molecular data. Knome provides a platform and tools to help researchers and doctors develop next generation, software-based tests and make clinical decisions.	www.knome.com
NextBio	NextBio's big data technology enables users to systematically integrate and interpret public and proprietary molecular data and clinical information from individual patients, population studies and model organisms applying genomic data in useful ways both in research and applied to the clinic	www.nextbio.com

*Some Companies/ Institutions provide just software solutions (Ex: Appistry, CLC Bio) and others offer combined hardware and software solutions (Ex: BGI, Knome). For example, Knome is a software company that launched a hardware system to keep data local and address privacy issues (see "Genome interpreter vies for place in clinical market" by Monya Baker. Nature. 490, p. 157, 2012).

this issue is the National Center for Biotechnology Information (NCBI). The NCBI has been leading Big Data efforts in biomedical science since 1988, but neither the NCBI nor anyone in the private sector has a comprehensive, inexpensive, and secure solution to the problem of data storage (even though companies with different solutions are starting to appear as shown in table 1). These capabilities are beyond the reach of small laboratories or institutions, posing several challenges for the future of biomedical research.

Another challenge is to transfer data from one location to another; it is mainly done by shipping external hard disks through the mail. An interesting solution for data transfer is the use of Biotorrents, which will allow open access sharing of scientific data and uses a peer-to-peer file sharing technology [9]. Torrents were primarily designed to facilitate distribution of large amounts of data in the internet and it could be applied to biomedicine [9].

Security and privacy of data from individuals is also a concern. Possible solutions to this issue include the use of better security systems with advanced encryption algorithms, like the ones used by banks in the financial sector to secure their clients' privacy [10]. In addition, a new generation of consent forms that specifically allow study participants or patients to openly share the data generated on them with researchers will be needed [10]. The use of "in house" hardware solutions instead of cloud computing could also ease the implementation of big data with more information protection. One example is the system that Knome is implementing named knoSYS100 (table 1). These are just some of the solutions that could be applied to overcome the challenges to deal with big data privacy, but other solutions will emerge in the near future.

Conclusions

Success in biomedical research dealing with the increasing amounts of omics data combined with clinical information will depend on our ability to interpret high scale data sets that are generated by emerging technologies. Private companies such as Microsoft, Oracle, Amazon, Google, Facebook and Twitter are masters in dealing with petabyte scale data sets. Science and Medicine will need to implement the same type of scalable structure to deal with volumes of data generated by omics technologies. The life sciences will need to adapt to the advances in informatics to successfully address the Big Data problems that will be faced in the next decade.

Acknowledgements

I would like to thank both Kelly Arndt and Steve Iannaccone for their thoughtful inputs and for critically reading this article.

References are available at the author.

Author
Fabricio F. Costa
Cancer Biology and Epigenomics Program, Children's Hospital of Chicago Research Center and Department of Pediatrics, Northwestern University's Feinberg School of Medicine, Chicago, IL, USA;
Datagenno Interactive Research Ltda, Rio de Janeiro, Brazil
Genomic Enterprise, Chicago, IL, USA

Contact
Fabricio F. Costa, Ph.D.
Cancer Biology and Epigenomics Program,
Children's Hospital of Chicago Research Center and Department of Pediatrics, Northwestern University's Feinberg School of Medicine,
Chicago, IL, USA
fcosta@luriechildrens.org