ELSEVIER

Big data in biomedicine

Fabricio F. Costa[1,2,3,4]

[1] Genomic Enterprise, Chicago, IL 60614, USA
[2] DataGenno Interactive Research Ltda, Rua Gastão Machado 66, Edifício CME, Salas 503/504, Campos dos Goytacazes, Rio de Janeiro, RJ 28035-120, Brazil
[3] 1871: DataGenno Interactive Research, 222 W. Merchandise Mart Plaza, 12th Floor, Suite 1212, Chicago, IL 60654, USA
[4] Cancer Biology and Epigenomics Program, Ann & Robert H. Lurie Children's Hospital of Chicago Research Center and Department of Pediatrics, Northwestern University's Feinberg School of Medicine, Chicago, IL 60614, USA

The increasing availability and growth rate of biomedical information, also known as 'big data', provides an opportunity for future personalized medicine programs that will significantly improve patient care. Recent advances in information technology (IT) applied to biomedicine are changing the landscape of privacy and personal information, with patients getting more control of their health information. Conceivably, big data analytics is already impacting health decisions and patient care; however, specific challenges need to be addressed to integrate current discoveries into medical practice. In this article, I will discuss the major breakthroughs achieved in combining omics and clinical health data in terms of their application to personalized medicine. I will also review the challenges associated with using big data in biomedicine and translational science

Introduction

A series of breakthroughs in medical science and IT are triggering a convergence between the healthcare industry and the life sciences industry that will quickly lead to more intimate and interactive relations among patients, their doctors and biopharmaceutical companies [1]. Big data analytics has an indispensable role in fostering those enhanced relations because it vastly enriches the remarkable but isolated wonder of the genome-on-a-thumb drive. Healthcare providers and drug makers now have the ability to explore and analyze omics data not only for an individual, but also in an aggregate from an increasing number of patients in specific population studies [2].

With rapid improvements in computer power, the cost of genome sequencing has plunged from millions of dollars per genome to thousands of dollars (and the cost will keep dropping). With advances in technology, patients will see a shift from population-based healthcare to personalized medicine that includes targeted diagnostics and treatment based on each patient's history, ancestry and genetic profile [3]. Online tools, such as the General Practice Research Database (GPRD), applied to clinical studies in drug discovery and assessment exemplify how IT is impacting biomedicine [4]. Although complex, this trend could one day revolutionize life sciences, biomedicine and what it means to be a healthcare professional or a researcher. The main benefits of applying big data analytics in personalized medicine include saving time while improving the overall quality and efficacy of treating disease.

Big data in biomedicine is driven by the single premise of one day having personalized medicine programs that will significantly improve patient care. Constant advances in understanding of different omics information are providing the footholds into establishing, for the first time, the causal genetic factors that could help manage the golden triangle of treatment: the right target, the right chemistry and the right patient. Solutions to deal with this overload of information are becoming a reality. However, challenges ahead include funneling clinical data, omics data, administrative data and also financial information securely into an unified system [5] to achieve better patient outcomes, advance research and continually improve the quality of patient care while reducing costs.

E-mail addresses: fcosta@luriechildrens.org, fcosta@genomicenteprise.com, fcosta@datagenno.com.

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2013.10.012

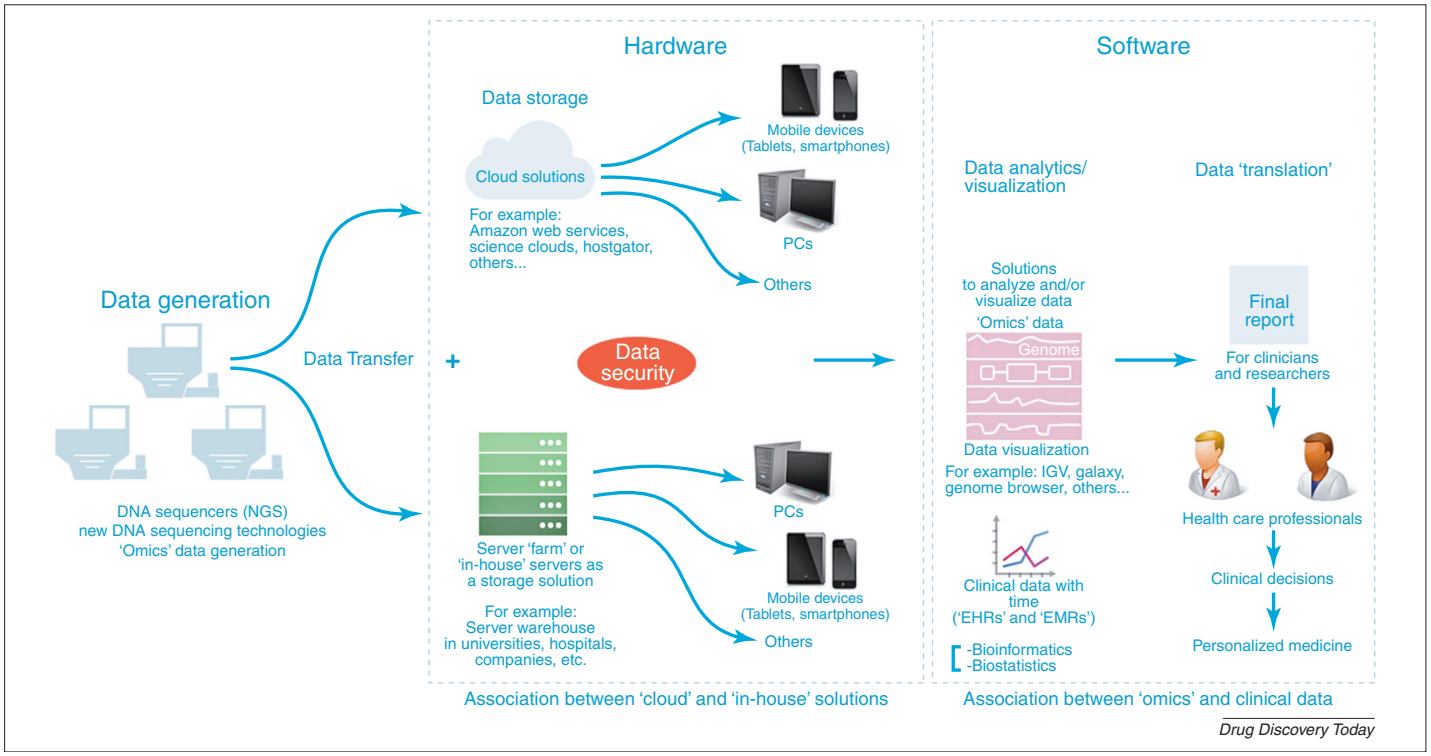

FIGURE 1

Big data in biomedicine. Schematic representation and depiction of a pipeline starting with data produced using next-generation sequencing (NGS), to data 'translation', and the generation of a 'final report' for clinicians and researchers. Personal health information and data generated by next-generation DNA sequencers (i.e. omics data such as genomes, transcriptomes, exomes, epigenomes and other types of similar information) are correlated, transferred to the 'cloud' or internal servers, analyzed and visualized using different solutions and tools that are available for big data analytics. Finally, data is translated as a short report to clinicians and researchers after a deep analysis for biomarkers and drug targets associated with specific disease phenotypes and after comparisons with public or private databases. Genome variants could be identified when comparing different samples, thus generating high-quality interpretation based on current knowledge and literature. This type of pipeline will ease the implementation and application of personalized medicine for clinicians and for research purposes. Between data transfer, storage and visualization, patient data needs to be secured by encryption of the information. Some solutions for both medical and scientific data security have been developed recently, but since this is a new area of study in biomedical informatics, big challenges lie ahead creating increasing opportunities in the market. Abbreviations: PCs, personal computers; EHRs, electronic health records; EMRs, electronic medical records; IGV, integrated genome viewer. Image designed by Eduardo Braga Ferreira Junior.

Although both computers and the internet have become faster, there is a lack of computational infrastructure that is needed to generate, maintain, transfer and analyze large-scale information securely in biomedicine and to integrate omics data with other data sets, such as clinical data from patients (Fig. 1). Indeed, it might now be less expensive to generate the data than it is to analyze and store it [6]. Another challenge is to transfer data from one location to another, because it is mainly done by mailing external drives with the information inside [6]. The security and privacy of the data from individuals are also a concern before and during data transfer [6]. Possible solutions to these issues include the use of better security systems with advanced encryption and de-identification algorithms, such as those used by banks in the financial sector to secure their clients' privacy [6]. The future of big data in life sciences is full of insecurities and challenges, but changes in several sectors are occurring to deal with it. Importantly, making sense of accumulating data in life sciences requires improved computational infrastructure, new methods to interpret the information and unique collaborative approaches.

In this article, I will discuss some of the major improvements in combining omics and clinical health data applied to personalized medicine. Moreover, an overview of the challenges faced by Big Data generation, transfer and analytics will be addressed. This article will also exemplify some of the major improvements needed to bridge the current technological gaps to address these challenges. Computational strategies, instrumentation and the current knowledge to interpret Big Data in order to make clinical decisions with a positive impact in biomedicine will also be discussed.

Big data, big impacts

Big data describe a new generation of technologies and architectures, designed to extract value from large volumes of a wide variety of data by enabling high-velocity capture, discovery and analysis [7]. This world of big data requires a shift in computing architecture so that researchers can handle both the data storage requirements and the heavy server processing needed to analyze large volumes of data in a secure manner [8]. Most of the big data surge is unstructured information and is not typically easy for traditional databases to analyze it. Therefore, the predictive power of big data has been explored recently in fields such as public health, science and medicine.

Computer tools to collect knowledge and insights from the vast trove of unstructured data available via the Internet are improving

TABLE 1

Examples of companies and institutions that provide solutions to generate, interpret and visualize combined omics and health clinical data

Company or institution	Type of solution	Website
Appistry	High-performance big data platform that 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	http://www.appistry.com
Beijing Genome Institute	This solution serves as a solid foundation for large-scale bioinformatics processing. The computing platform is an integrated service comprising versatile software and powerful hardware applied to life sciences	http://www.genomics.cn/en
CLC Bio	Utilizes proprietary algorithms, based on published methods, to accelerate successfully data calculations to achieve remarkable improvements in big data analytics	http://www.clcbio.com
Context Matters	Provides a comprehensive tool that empowers pharmaceutical and biotechnology companies to make better strategic decisions using web-based applications, and easy-to-use interface and visualization tools to deal with complex data sets	http://www.contextmattersinc.com
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.	http://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	http://www.genome.com
GNS Healthcare	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	http://www.gnshealthcare.com
NextBio	Big data technology that enables users to integrate and interpret systematically public and proprietary molecular data and clinical information from individual patients, population studies and model organisms applying omics data in useful ways both in research and in the clinic	http://www.nextbio.com
Pathfinder	Develops customized software applications, providing solutions in different sectors, including healthcare and omics, offering technologies that enable business breakthroughs and competitive advantages	http://www.pathfindersoftware.com

at this task. At the forefront of the rapidly advancing techniques of artificial intelligence (AI) are natural-language processing [9], pattern recognition [10] and machine learning [11]. Those AI technologies can be applied to many fields, especially in biomedicine and life sciences. One such example is the algorithm used by Google to track diseases that is known as Google Trends (GT) [12]. GT and other strategies to track diseases using geospatial maps is a daunting big data challenge, parsing vast quantities of information and making decisions instantaneously. For example, GT can find spikes in Google search requests for terms such as 'flu symptoms' and 'flu treatments' before there is an increase in flu patients coming to hospital emergency rooms in specific regions [13]. Tools that are able to identify term requests in epidemic areas are just one application of big data analytics in biomedicine. The impacts that these information-driven tools will have in health tracking and disease monitoring are currently immensurable [14].

Computational solutions and the use of the internet are also helping to create tools to manage diseases. For example, data repositories have been created to guide doctors and patients that suffer from diseases such as cancer helping them find the right drug for their disease type, one of the foundations of personalized medicine. One such tool is the portal 'My Cancer Genome' created by researchers at Vanderbilt University in the USA [15]. This solution began 2 years ago and now has more than 50 contributors from 20 institutions worldwide [15]. The portal lists mutations in different cancer types, as well as drug therapies that might or might not be of benefit to patients. Most of the drugs described on the website are in clinical trials and only a few have been approved by the US Food and Drug Administration (FDA). It is important to note that there are significant limitations to this type of practice since the FDA did not approve the majority of the drugs targeting these mutations. However, the portal is free and doctors, researchers, patients, relatives and institutions can access it, easing the translation of the findings in research laboratories to the bedside of patients.

Another tool is exemplified by the solution provided by Context Matters. This company provides a comprehensive solution that leverages targeted biomedical information to pharmaceutical and biotechnology companies using web-based applications with an easy-to-use interface and personalized visualization tools to deal with complex data sets. Some of these tools use crowdsourcing approaches to analyze and make sense of big data. More solutions and tools provided by specific companies are shown in Table 1.

Although these online tools are helpful for physicians, no treatment plan can yet be based on the results provided by them. A barrier that needs to be overcome is the difficult conversation between patients that are empowered by preliminary results provided by these online solutions and their physicians, who sometimes do not know the limitations of applying such tools in their practice. However, based on accumulating examples, it is clear that increasing amounts of information in databases and the use of web solutions by healthcare professionals and patients will have a big

impact on biomedicine by facilitating drug development and disease treatment.

A digital revolution in life sciences

The Hippocratic oath is an oath historically taken by physicians and other healthcare professionals swearing to practice medicine honestly and ethically [16]. Although 25 centuries have passed since Hippocrates' call, physicians have not yet attained the dream of true evidence-based healthcare. The Hippocratic oath is an example of how medicine has to change and, as recently described, how it has to somehow adapt to new technological breakthroughs [17]. Computer-aided medicine, web-based solutions and big data analytics will need to be taken seriously by physicians. Physicians will also need to absorb and incorporate these changes. In the Oath, physicians promise to treat according to their ability and judgment. Evidence-based medicine has to be incorporated in the Oath and in medical schools because it is becoming a reality. For example, large quantities of data about wellness and illness continue to be disconnected, rather than collected and harnessed to optimize the provision of care. I believe that we now stand at the brink of a potential digital revolution in data-centric healthcare, enabled by advances in computer technologies. The digital revolution in life sciences promises to enhance the quality of healthcare while cutting costs and, more generally, enabling physicians and researchers to do their very best with what is available from aligned healthcare resources. Aligning available resources in IT with the core promise that all healthcare professionals make when they raise their hand and recite the Hippocratic oath upon receipt of their medical degree will completely change the life sciences [16]. However, enabling this vision of true evidence-based healthcare based on big data analytics will require crucial investments for translating key methods and insights into working systems, as well as for advances in core computer science research and engineering to address key conceptual bottlenecks and opportunities. The collection and analysis of data available on health and disease promise to enhance the quality and efficacy of healthcare, and to enhance the quality and longevity of life. This can also provide new insights about diseases. In addition, data-centric methods will enable researchers to transform information into predictive models, thus resulting in so-called 'personalized medicine'.

Importantly, data-driven medicine will facilitate the discovery of new treatment options based on multimodel molecular measurements on patients and on learning from the trends in differential diagnosis, prognosis and prescription adverse effects from available clinical databases [18]. In addition, medical informatics, represented by patients' electronic medical records (EMRs) and personalized therapies will enable the application of targeted treatments for specific diseases. Mining of EMRs has the potential for establishing new patient-stratification principles for revealing unknown disease correlations [19]. Integrating EMRs with genetic profiles will also give a finer understanding of genotype–phenotype relations [19]. However, a broad range of ethical, legal and technical reasons hinder the systematic analysis of the data contained in EMRs [19]. Even with several challenges and barriers, there is a data-sharing trend in the web, exemplified by online health resources, such as PatientsLikeMe, which allow patients to share detailed health and treatment information, providing a novel data source for different types of study and analysis [19].

The same way as big enterprises, such as Amazon, Google, Facebook, and other companies in computer technology, are leveraging consumer data to target offers to individuals by offering specific products based on the consumer actions, healthcare providers should be able to leverage the power of analytics to evaluate an individual's medical record and omics data signatures to compare those to insights gained from the analysis of outcomes in large populations. This is just the start for the big data analytics and data-centric models in the digital revolution that researchers are starting to experience. However, technological breakthroughs dealing with clinical and genetic data from patients and populations bring several challenges, such as the security and privacy of this information.

Information-driven technologies applied to biomedical research

A wave of new sequencing technologies, named third- and fourth-generation DNA sequencing, makes it possible to sequence genomes, transcriptomes and epigenomes faster at a lower cost. These new technologies are based on semiconductors [20] and nanopores [21]. With these types of approaches, it is possible to develop, with relative success, large-scale sequencing projects and to analyze this information using big data analytics solutions. Two examples of such projects are the 1000 Genomes Project and the Encyclopedia of DNA Elements (ENCODE).

The international 1000 Genomes Project is a government-backed initiative launched in 2008 that aims to sequence the entire genome of thousands of people from around the world and it is continuing to grow as the largest data set worldwide on human genetic variation [22]. Additionally, data from this project will be combined with expression and genotype data to create a big data repository in biomedicine [23]. Phase one of this project has already generated sequence for more than 1000 genomes [24]. Phase three was recently reported, with exome sequencing of several genomes to extract expression data [25]. Information generated by the 1000 Genomes Project has been widely used by the genetics community, making it one of the most cited studies in biology [26]. The challenge now is to apply the knowledge from these genomes and understand disease phenotypes to facilitate drug discovery.

Another ongoing big data project in biomedicine is ENCODE [27]. The main objective of ENCODE was to map and characterize how the entire human genome function. Members of this Project have already performed 1600 experiments in approximately 150 cell types to deliver an incredible amount of data and information [28] and the main research article was published by almost 500 authors working in 32 institutes worldwide [28]. The data generated by ENCODE highlighted biochemical functions for approximately 80% of the human genome, with a particular focus in regions that are outside the well-studied protein-coding DNA (i.e. protein-coding genes) [29]. In addition, the project showed that 90% of all human genetic variants fall inside a region that has no protein-coding gene annotated, indicating that these regions might be responsible for differences between individual humans and are also likely to be important for a better understanding of complex diseases [30].

ENCODE was also able to provide new insights into the organization and regulation of human genes and genomes and will

TABLE 2

Examples of big corporations offering solutions and pipelines to store, analyze and deal with complex biomedical information

Company	Solution(s)	Website
Amazon Web Services	Provides the necessary computing environment, including CPUs, storage, memory (RAM), networking, and operating system, for a hardware infrastructure as a service in the biomedical and scientific fields	http://aws.amazon.com
Cisco Healthcare Solutions	Offers different types of solution for the life sciences, including specific hardware and cloud computing for reliable and highly secure health data communication and sharing across the healthcare community	http://www.cisco.com/web/strategy/healthcare/index.html
DELL Healthcare Solutions	Connects researchers to the right technology and processes to create information-driven healthcare and accelerate innovation in life sciences with electronic medical record (EMR) solutions	http://www.dell.com/Learn/us/en/70/healthcare-solutions?c=us&l=en&s=hea
GE Healthcare Life Sciences	Provides expertise and tools for a wide range of applications, including basic research of cells and proteins, drug discovery research, as well as tools to support large-scale manufacturing of biopharmaceuticals	http://www3.gehealthcare.com/en/Global_Gateway
IBM Healthcare and Life Sciences	Provides healthcare solutions, technology and consulting that enable organizations to achieve greater efficiency within their operations, and to collaborate to improve outcomes and integrate with new partners for a more sustainable, personalized and patient-centric system	http://www-935.ibm.com/industries/healthcare
Intel Healthcare	Currently builds frameworks with governments, healthcare organizations, and technology innovators worldwide to build the health IT tools and services of tomorrow by combining different types of health information	http://www.intel.com/healthcare
Microsoft Life Sciences	Provides innovative, world-class technologies to help customers nurture innovation, improve decision-making and streamline operations	http://www.microsoft.com/health/en-us/solutions/Pages/life-sciences.aspx
Oracle Life Sciences	Delivers key functionalities built for pharmaceutical, biotechnology, clinical and medical device enterprises. Oracle maximizes the chances of discovering and bringing to market products that will help in treating specific diseases	http://www.oracle.com/us/industries/life-sciences/overview/index.html

serve as an expansive resource of information for biomedical research during the next decade [31]. From an evolutionary perspective, ENCODE ends the myth that most of our genome is 'junk' DNA, given that researchers now know that 75% of the genome is capable of being transcribed in at least one cell type, thus affecting the concept of a 'gene' [32]. Before ENCODE, biologists understood that only a small fraction of the DNA of a gene encodes a protein (approximately 2–3%). ENCODE reported convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including nonprotein-coding transcripts [33]. This paradigm shift in molecular biology, with nonprotein-coding genes or noncoding RNAs pervasively transcribed and with putative functions, has been discussed in my previous articles [34–38].

Both consortia have shown that collaborative projects are possible and they will become more common as scientists tackle big problems, such as the human genome [39]. Big data projects in biomedicine such as these can 'socialize' science and make discoveries faster by accelerating drug development, tests, and approval [40]. Additionally, these projects show that biomedical research is becoming an information-driven science and, as discussed above, researchers will need to use the same approaches that big computer technology companies use to deal with increasing amounts of personal data. A depiction of a step-by-step pipeline used by big projects similar to ENCODE, from omics data generation using next-generation sequencing (NGS) to the production of a final report to clinicians and researchers is illustrated in Fig. 1.

Although these collaborative projects show great promise, the direct clinical impacts of their findings and discoveries have yet to be demonstrated. Large sample sizes, such as the ones from big data projects, improve the ability to detect trivial differences that might have limited the clinical applications [41,42]. In other words, the use of large sample sizes facilitate the identification of small populations of patients that might benefit from specific drugs already approved by the FDA or that are currently in clinical trials.

Solutions for data management and interpretation in personalized medicine

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. Moore's law is the observation that, over the history of computing hardware, the number of transistors on integrated circuits and the speed of computers doubles approximately every 2 years. Although cloud solutions from different companies have been used, several challenges remain, particularly related to the security and privacy of personal medical and scientific data (Fig. 1). Perhaps the greatest advantage could be the ability to offer a broad platform for the 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 [43]. In fact, some companies and big corporations already offer such solutions applied to healthcare and life sciences (Tables 1 and 2). There is

TABLE 3

Examples of companies that offer personalized genetics and omics solutions

Company	Applications and/or services	Website
23andme	A DNA analysis service providing information and educational tools for individuals to learn and explore their DNA through personal genomics	http://www.23andme.com
Counsyl	Offers tests for gene mutations and variations in more than 100 inherited rare genetic disorders using a DNA biochip designed specifically to test for these disorders	http://www.counsyl.com
Foundation Medicine	A molecular information company at the forefront of bringing comprehensive cancer genomic analytics to routine clinical care	http://www.foundationmedicine.com
Knome	Analyzes whole-genome data using software-based tests to examine and compare simultaneously many genes, gene networks and genomes as well as integrate other forms of molecular and nonmolecular data	http://www.knome.com
Pathway Genomics	Incorporates customized and scientifically validated technologies to generate personalized reports, which address a variety of medical issues, including an individual's propensity to develop certain diseases	http://www.pathway.com
Personalis	A genome-scale diagnostics services company pioneering genome-guided medicine focused on producing the most accurate genetic sequence data from each sample, using data analytics and proprietary content to draw accurate and reliable biomedical interpretations	http://www.personalis.com

also an opportunity for the development of applications or apps, specifically for omics tools, from which hundreds of specialty solutions could be developed [44]. Companies such as Illumina and 23andme already offer an open platform for developers and more companies will implement application programming interfaces (APIs) in their services. Therefore, solutions to overcome data privacy issues will be crucial.

Pipelines to deal with increasing amounts of omics data will be needed to store, transfer, analyze, visualize and generate 'short' reports for researchers and clinicians (Fig. 1). In fact, an entirely new genomics industry could result from cloud computing, which will transform medicine and life sciences. Indeed, cloud computing opens a new world of possibilities for the genomics industry to transform the way that it approaches research and medicine.

Other solutions to deal with big data, especially when analyzing complex genomics information, include the use of graphics processing units (GPUs). GPUs have the potential to improve quickly and drastically computational power over conventional processors, even when compared with the cloud [45]. For example, GPUs can be used as a tool to detect gene–gene interactions in genome-wide studies [46]. Compared with the currently used central processing units (CPUs), GPUs are highly parallel hardware providing massive computation resources. GPUs have been recently used for proteomic analysis [47] and metagenomic sequence classification [48], and could be applied to deal with heterogeneous sources of data, such as clinical and genomic information.

Big data analytics is also affecting how both biotechnology and pharmaceutical sectors identify new drug targets. The pharmaceutical industry is partnering with different omics companies and with academia to develop personalized drugs based on a patient's genetic code (Table 3). For example, Vertex Pharmaceuticals developed a collaborative study with more than 200 scientists in a cystic fibrosis (CF) project that aimed to screen >500 000 compounds using computer software. Using this approach, this project virtually screened thousands of compound combinations to narrow the choice to a single drug capable of helping a small group of CF patients with a specific DNA mutation (G551D) that affects 4% of such patients [49,50]. The end product of this collaboration was Kalydeco [50]. This is a clear example of the future of genetically targeted drugs in personalized medicine. The identification of this new drug was a powerful result of combinatory technology using big data analytics and genetics and was the first drug discovered to correct an underlying cause of CF [51].

Successful applications of personalized medicine in cancer include three drugs that have been identified and used in specific groups of patients. Patients with melanoma and the BRAF mutation V600E can be treated with dabrafenib [52], patients with breast cancer and the amplification or overexpression of the gene encoding Her2/Neu can be treated with a targeted therapy using trastuzumab [53] and different types of tumor that contain the fusion protein BCR-ABL can be treated with imatinib [54]. These targeted therapies show how important personalized medicine programs will be to identify novel treatments for rare genetic diseases and for complex diseases, such as cancer. In such cases, the use of big data analytics tools to deal with complex combinations of information simultaneously will be crucial.

Other examples of how personalized computer-aided diagnostics can help save time while improving the overall quality of care for patients is the use of computer algorithms to screen patients for cancer [55]. These informatics tools are just as accurate as trained radiologists, except that the computer algorithms have a lower false positive rate [55]. Computer-aided diagnostics (CAD) can also help in ascertaining responses to the use of specific drugs [56].

Challenges ahead

These revolutionary changes in big data generation and acquisition create profound challenges for the storage, transfer and the security of information. Indeed, it might now be less expensive to generate the data than it is to store, secure and analyze it. In addition, biological and medical data are more heterogeneous than information from any other research field. For example, the National Center for Biotechnology Information (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 Tables 1 and 2). These capabilities are beyond the reach of small laboratories and institutions, posing several challenges for the future of biomedical research.

Another challenge is to transfer data from one location to another, because this is mainly performed by shipping external hard disks containing the information. An interesting solution for data transfer is the use of different types of software to compress the data without losing pieces of information. Another tool that could be used is open-access sharing of scientific data and the use of peer-to-peer file-sharing technology [57]. In addition, a specific solution that became available for data storage and transfer is a type of Dropbox for data scientists named Globus Online, which provides a 'Software as a Service' (SaaS) for the storage and transfer of data [58–61]. In this case, data is generated in one location where large-scale storage is not available. Then, the data produced, especially in genomics, when whole genomes are sequenced, need to be transferred to other locations. Globus Online provides storage capacity and secure solutions to transfer the data [58]. Aspera also offers a service named 'fast' that is a software able to speed up data transfer hundreds of times compared with the other methods available, using a regular internet protocol [61]. However, all transfer protocols have challenges associated with transferring large, unstructured data sets. Finally, tools to speed the process of data transfer and latency have been developed recently; one example is a cloud-based solution that overcomes this problem by processing the data while it is being transferred to another location [62].

The security and privacy of the data from individuals is also a concern. Possible solutions to this issue include the use of better security systems with advanced encryption algorithms, such as those used by the financial sector to secure their clients' privacy [63]. 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 has been proposed and might be implemented soon [64]. A context-specific approach to informed consent for web-based health research can facilitate a dynamic research enterprise and, at the same time, maintain public trust [64]. Furthermore, if privacy concern is an issue, the use of 'in-house' hardware solutions instead of cloud computing could ease the implementation of big data with more information protection. One example is the hardware system that the company Knome is implementing, called 'knoSYS100' [65] (Table 3). These are just some of the solutions that could be applied to overcome the challenges of dealing with big data privacy, but I believe that other tools will emerge in the near future.

Concluding remarks

Success in biomedical research to deal with the increasing amounts of omics data combined with clinical information will depend on the ability to interpret large data sets that are generated by different emerging technologies. Big corporations, such as Microsoft, Apple, Oracle, Amazon, Google, Facebook and Twitter, are masters in dealing with big data sets. The scientific and medical fields will need to implement the same type of scalable structure to deal with the volumes of data generated by different omics technologies and health information. Biomedicine will need to adapt to the advances in informatics to address successfully the big data problems that will be faced in the future, especially in personalized medicine programs, to improve significantly patient care. Additionally, more studies will be needed to demonstrate that personalized medicine and computer-aided diagnostics directly benefit patients.

Financial disclosure

FFC has no financial relation, interest or affiliation with the institutions and companies discussed in this article. FFC is the Founder of Genomic Enterprise (www.genomicenterprise.com), the co-founder and Chief Scientific Officer (CSO) of the company DataGenno Interactive Research (http://www.datagenno.com), a Member of the Digital Technology Incubator 1871 Chicago (http://www.1871.com) and a Member of the Start Up Health Academy (http://www.startuphealth.com/).

Acknowledgements

I thank DataGenno Interactive Research's Team, including Dr. Marcelo P. Coutinho, Julio C.B. Araujo, Alex Moreira, Lucas F. Correa, Italo M. Soares and Maria Celeste L.A.M. Cabral for technical assistance. I also thank Eduardo B.F. Junior for designing Fig. 1. I am grateful to Tristan Gill for his insights and helpful discussions during the preparation of this manuscript, and to Kelly Arndt for editing help.

References

1 Costa, F.F. (2013) Social networks, web-based tools and diseases: implications for biomedical research. *Drug Discov. Today* 18, 272–281
2 Knoppers, B.M. *et al.* (2012) Sampling populations of humans across the world: ELSI issues. *Annu. Rev. Genomics Hum. Genet.* 13, 395–413
3 Costa, F.F. (2009) Genomics, epigenomics and personalized medicine: a bright future for drug development? *BioForum Eur.* 13, 29–31
4 Skow, A. *et al.* (2013) The association between Parkinson's disease and anti-epilepsy drug carbamazepine: a case-control study using the UK General Practice Research Database. *Br. J. Clin. Pharmacol.* http://dx.doi.org/10.1111/bcp.12100
5 Costa, F.F. (2011) Basic research, applied medicine and EHRs: are we on the right track? *J. Cancer Sci. Ther.* 3, 1948–5956
6 Costa, F.F. (2012) Big data in genomics: challenges and solutions. *G.I.T. Lab. J.* 11–12, 1–4
7 Villars, R.L. *et al.* (2011) *Big Data: What It is and Why You Should Care.* IDC
8 Scarpati, J. (2012) *Big Data Analysis in the Cloud: Storage, Network and Server Challenges.* CloudProvider
9 Liu, K. *et al.* (2011) Natural language processing methods and systems for biomedical ontology learning. *J. Biomed. Inform.* 44, 163–179
10 Kemp, C. and Tenenbaum, J.B. (2008) The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.* 105, 10687–10692
11 Sajda, P. (2006) Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.* 8, 537–565
12 Carneiro, H.A. and Mylonakis, E. (2009) Google Trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin. Infect. Dis.* 49, 1557–1564
13 Dugas, A.F. *et al.* (2012) Google Flu trends: correlation with emergency department influenza rates and crowding metrics. *Clin. Infect. Dis.* 54, 463–469
14 Dugas, A.F. *et al.* (2013) Influenza forecasting with Google Flu trends. *PLoS ONE* 8, e56176
15 Van Allen, E.M. *et al.* (2013) Clinical analysis and interpretation of cancer genome data. *J. Clin. Oncol.* 31, 1825–1833
16 Farnell, L.R. (2004) *Greek Hero Cults and Ideas of Immortality.* Kessinger Publishing
17 Topol, E. (2012) *The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care.* Basic Books
18 Shah, N.H. and Tenenbaum, J.D. (2012) The coming age of data-driven medicine: translational bioinformatics' next frontier. *J. Am. Med. Inform. Assoc.* 19, e2–e4

19 Jensen, P.B. *et al.* (2012) Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* 13, 395–405
20 Rothberg, J.M. *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475, 348–352
21 Clarke, J. *et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4, 265–270
22 Kuehn, B.M. (2008) 1000 Genomes Project promises closer look at variation in human genome. *JAMA* 300, 2715
23 Buchanan, C.C. *et al.* (2012) A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *J. Am. Med. Inform. Assoc.* 19, 289–294
24 1000 Genomes Project Consortium, *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65
25 Shen, H. *et al.* (2013) Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty-four Caucasians. *PLoS ONE* 8, e59494
26 Clarke, L. *et al.* (2012) The 1000 Genomes Project: data management and community access. *Nat. Methods* 9, 459–462
27 Maher, B. (2012) ENCODE: the human encyclopaedia free. *Nature* 489, 46–48
28 ENCODE Project Consortium, *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
29 Djebali, S. *et al.* (2012) Landscape of transcription in human cells. *Nature* 489, 101–108
30 Gerstein, M.B. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100
31 Neph, S. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90
32 Gerstein, M.B. *et al.* (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 17, 669–681
33 Pennisi, E. (2012) ENCODE project writes eulogy for junk DNA. *Science* 337, 1159–1161
34 Costa, F.F. (2005) Non-coding RNAs: new players in eukaryotic biology. *Gene* 357, 83–94
35 Costa, F.F. (2007) Non-coding RNAs: lost in translation? *Gene* 386, 1–10
36 Costa, F.F. (2008) Non-coding RNAs, epigenetics and complexity. *Gene* 41, 9–17
37 Costa, F.F. (2009) Non-coding RNAs and new opportunities for the private sector. *Drug Discov. Today* 14, 446–452
38 Costa, F.F. (2010) Non-coding RNAs: meet thy masters. *Bioessays.* 32, 599–608
39 Birney, E. (2012) The making of ENCODE: lessons for big-data projects. *Nature* 489, 49–51
40 Gerstein, M. (2012) ENCODE leads the way on Big Data. *Nature* 489, 208
41 Friston, K. (2012) Ten ironic rules for non-statistical reviewers. *Neuroimage* 61, 1300–1310
42 Lin, C. *et al.* (2013) Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. *PLoS ONE* 8, e69932
43 Vanacek, J. (2012) How cloud and big data are impacting the human genome: touching 7 billion lives. *Forbes* 16 April
44 Anon., (2012) *23andMe and Illumina Forge Consumer Genomics Goliath.* Bio-IT World Magazine
45 Greene, C.S. *et al.* (2010) Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics* 26, 694–695
46 Yung, L.S. *et al.* (2011) GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics* 27, 1309–1310
47 Hussong, R. *et al.* (2009) Highly accelerated feature detection in proteomics data sets using modern graphics processing units. *Bioinformatics* 25, 1937–1943
48 Jia, P. *et al.* (2011) MetaBinG: using GPUs to accelerate metagenomic sequence classification. *PLoS ONE* 6, e25353
49 Smolan, R. and Erwitt, J. (2012) *The Human Face of Big Data.* Against All Odds Productions
50 Ramsey, B.W. *et al.* (2011) A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *N. Engl. J. Med.* 365, 1663–1672
51 Herper, M. (2012) The most important new drug of 2012. *Forbes* 27 December
52 Tsao, H. *et al.* (2012) Melanoma: from mutations to medicine. *Genes Dev.* 26, 1131–1155
53 Incorvati, J.A. *et al.* (2013) Targeted therapy for HER2 positive breast cancer. *J. Hematol. Oncol.* 3, 38
54 Horne, S.D. *et al.* (2013) Why imatinib remains an exception of cancer research. *J. Cell. Physiol.* 228, 665–670
55 Nguyen, T.B. *et al.* (2012) Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology* 262, 824–833
56 Zhang, Z. *et al.* (2013) Automatic diagnosis of pathological myopia from heterogeneous biomedical data. *PLoS ONE* 8, e65736
57 Langille, M.G. and Eisen, J.A. (2010) BioTorrents: a file sharing service for scientific data. *PLoS ONE* 5, e10071
58 Allen, B. *et al.* (2012) Software as a service for data scientists. *Commun. ACM* 55, 81–88
59 Foster, I.T. *et al.* (2012) Campus bridging made easy via globus services. In *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment* http://dx.doi.org/10.1145/2335755.2335847
60 Foster, I. (2011) Globus Online: accelerating and democratizing science through cloud-based services. *IEEE Internet Computing.* 15, 70–73
61 Marx, V. (2013) The big challenges of big data. *Nature* 498, 255–260
62 Issa, S.A. *et al.* (2013) Streaming support for data intensive cloud-based sequence analysis. *Biomed. Res. Int.* 2013, 791051
63 Schadt, E.E. (2012) The changing privacy landscape in the era of Big Data. *Mol. Syst. Biol.* 8, 612
64 Vayena, E. *et al.* (2013) Caught in the web: informed consent for online health research. *Sci. Transl. Med.* 5 173fs6
65 Baker, M. (2012) Genome interpreter vies for place in clinical market. *Nature* 490, 157