



Social networks, web-based tools and diseases: implications for biomedical research

Fabricio F. Costa^{1,2,3}

¹DataGenno Interactive Research Ltda, Rua Gastão Machado no 6, Edifício CME, Salas 503/504, Campos dos Goytacazes, Rio de Janeiro 28035-120, Brazil

²Genomic Enterprise, Chicago, IL 60614, USA

³Cancer Biology and Epigenomics Program, Department of Pediatrics, Children's Hospital of Chicago Research Center, Northwestern University Feinberg School of Medicine, 225 E. Chicago Avenue, Box 220, Chicago, IL 60611, USA

Advances in information technology have improved our ability to gather, collect and analyze information from individuals online. Social networks can be seen as a nonlinear superposition of a multitude of complex connections between people where the nodes represent individuals and the links between them capture a variety of different social interactions. The emergence of different types of social networks has fostered connections between individuals, thus facilitating data exchange in a variety of fields. Therefore, the question posed now is “can these same tools be applied to life sciences in order to improve scientific and medical research?” In this article, I will review how social networks and other web-based tools are changing the way we approach and track diseases in biomedical research.

The first generation of the Internet was termed Web 1.0, because it was the primary online interface connecting individuals (Fig. 1) [1]. In the beginning of the 21st century, search engines and social media gave rise to a more organized environment in the web by retrieving web pages as a list of results after a search for terms. In the late 1990s, Google Inc. emerged as a powerful web resource in the search business and has dominated ever since.

The terms ‘social media’ and ‘social networks’ are relatively new in the Internet. They describe a virtual space in the web where individuals are able to participate and share information, ideas, images, and other contents. This is considered the second generation of the Internet or Web 2.0 (Fig. 1). The emergence of social networks represented a change from a ‘static’ Internet to a more ‘dynamic’ web as a platform based on the individual feedback and collective intelligence with users generating and sharing contents.

The ability to share ideas has changed completely in the past 5 years with the emergence of both search engines and social media, especially with companies such as Google, Facebook, Twitter and philanthropic websites such as ‘Causes’ (Table 1). In life sciences, novel web technologies have been used to track diseases in more

efficient ways [2–4]. Today, millions of people around the world that have access to the Internet will engage in Internet-enabled self-diagnosis for diseases by using search engines and other web tools.

Search engines and social media can be helpful in predicting epidemic outbreaks of a disease caused by any viruses or bacteria. Using this new concept, Google Inc. was successful in using their search engine to detect areas that influenza epidemics are eminent [2]. In addition, a study has also shown that structural diversity in platforms such as Facebook emerges as an important predictor to evaluate social contagion processes [5] and this feature could be applied to track and manage diseases with the creation of disease communities. There are already groups of discussion on Facebook for people with multiple sclerosis (MS), HIV and Asperger’s Syndrome. These communities enable people affected with specific diseases and their relatives to share information about symptoms and new treatments.

Web-based tools are becoming powerful platforms not just for the general public, but also for science and medicine. Social media and search engines already have the potential to help accelerate the discovery of new and undiagnosed diseases, evaluate new treatments and facilitate information exchange between health-care professionals and patients. Understanding how companies,

E-mail addresses: fcosta@datagenno.com, fcosta@genomicenterprise.com, fcosta@childrensmemorial.org.

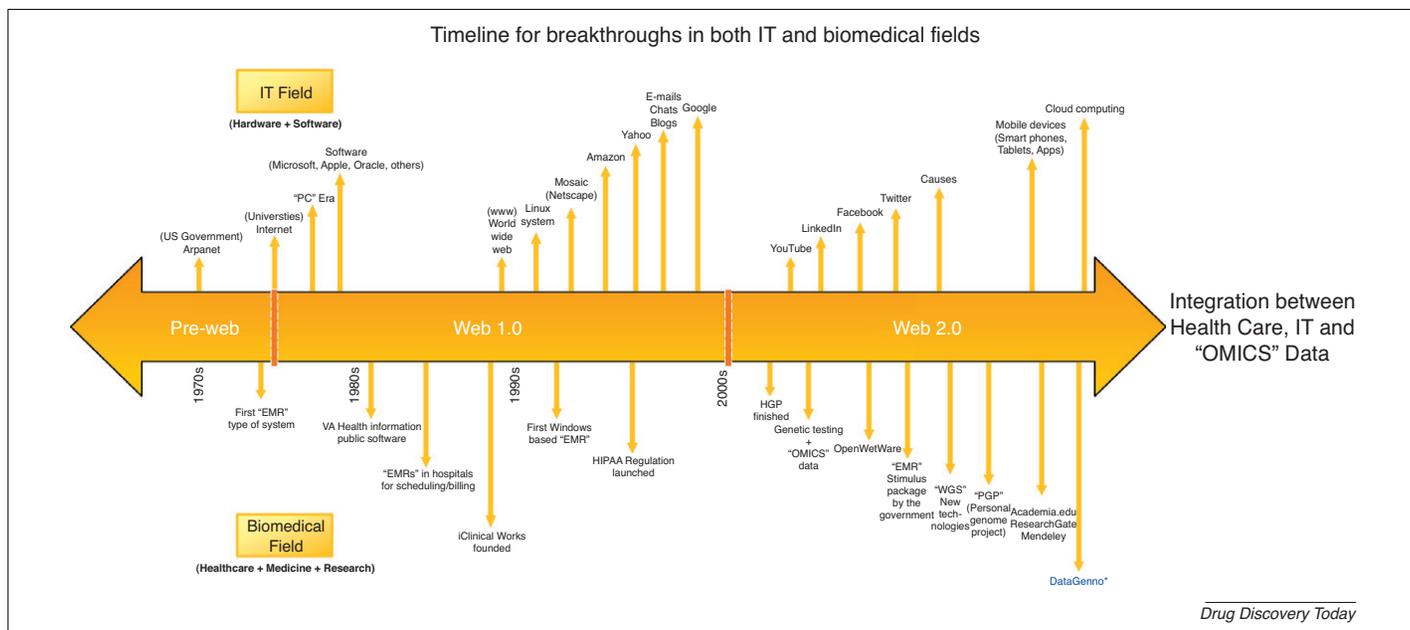


FIGURE 1

Historical timeline for the evolution of the Internet from the pre-web era, through the Web 1.0 and the beginning of the Web 2.0 eras in both Information Technology (IT) and Biomedical fields. The orange arrow represents the historical timeline. On the upper side of the figure, breakthroughs in the Information Technology sector are shown. On the lower side of the figure, landmarks in the biomedical fields including healthcare, medicine and science are shown. Specific vertical arrows coming from the main timeline indicate examples of innovations or milestones for both fields. The social network DataGenno has an asterisk to point out that it can integrate both biomedical discoveries with 'omics' data from patients and samples. The end of the arrow indicates that both fields will need to converge in order to handle the integration between clinical and molecular data.

especially in the pharmaceutical and biotechnology sectors, can use social networks will be paramount to facilitate the achievement of higher productivity in these sectors. Recent reports have shown that life sciences companies have already started to use social networks effectively to interact with external communities and collect data. Specific surveys have already demonstrated that two thirds of the life sciences companies plan to adopt any type of social network on a daily basis, but one third do not have plans to use it in any capacity, mainly because of regulatory uncertainty and privacy issues [6], especially when dealing with patient data.

Even with lack of regulatory clarity as the largest barrier, a trend towards social media tools and their use for clinical studies by pharmaceutical companies, information exchange between physicians and scientists and other applications in biomedical research are starting to become a reality. The same way social

media is now part of our daily life, it could also become an essential tool for life sciences and biomedical research.

Social media in science and medicine – more openness needed

Science is generally a closed society organized into little circles of highly trained specialists. One of the many definitions of science is ‘a systematic knowledge of the physical or material world gained through observation and experimentation’ [7]. The larger challenge is that most scientific data is proprietary. Harnessing non-scientists to help in specific scientific problems requires sending data out into the world, something that scientists are afraid to do. Many scientists remain skeptical of scientific openness, especially in the hyper-competitive biomedical fields, where patents, promotions and tenure positions can hinge on being the first to publish a

TABLE 1

General social networks and web-based tools from Internet companies that have been used to track and manage diseases.		
Company	Tool(s)	Website
Google	Google Trends (GT) is a tool that uses aggregated Google search data to estimate current epidemic activity of different infectious diseases around the world in real-time.	www.google.com/trends
Facebook	Facebook uses no specific tool but with almost 1 billion people engaged, it could be utilized to track trends for diseases and outbreaks similarly to GT. Facebook recently added a feature of organ donor status for its users that will facilitate organ donation and transplants in the medical field.	www.facebook.com
Twitter	Twitter uses no specific tool but with more than 350 million people engaged, the trends and hashtags used in the website could help in the identification of infectious disease outbreaks in real-time.	www.twitter.com
Causes	Causes represent the world's largest platform for philanthropy, especially related to diseases. This social network can be a valuable tool to raise awareness and resources to help patients and disease communities among other philanthropic applications.	www.causes.com

new discovery. Examples of successful open-source projects, especially in software research, include the development of the widely used open source operation system Linux, launched by Linus Torvalds in the 1990s [Notes for linux release 0.01 kernel.org: <http://ftp.funet.fi/pub/linux/historical/kernel/old-versions/RELNOTES-0.01>] and SourceForge that helps millions of people to collaborate in software development [8].

The scientist's interest in keeping things private and getting credit is directly opposed to society's interest in tackling some problems with the help of researchers. Exceptions exist, such as large astronomical and biological data sets that are freely available online [9,10]. For example, particle accelerators have produced a new collaborative working model for science, with hundreds of specialists all over the globe working to solve a specific problem in physics [9]. Biological databases that are supported by the United States government, such as the National Center for Biotechnology Information (NCBI), store data from whole genomes of several organisms (including our genome) and are opened to the general public [11,12]. In the past 10 years, however, there has been an improvement in technology that allows large numbers of people to

do cooperative things with information as exemplified by social media platforms, but the scientific establishment has taken only baby steps toward figuring out ways to share information productively [7].

The act of sharing in science will become something more common than we expect, especially now that social networks are reshaping society [7]. To encourage this shift, the federal government through the National Institutes of Health (NIH), which funds the United States research, has been pressuring scientists to work more cooperatively, in groups, and share more of what they discover. In addition, there are nascent efforts within academia to identify ways that researchers from different fields might be recognized for their contributions to the community as a whole, beyond the publication of their individual discoveries in scientific journals.

Social networks such as ResearchGate was recently launched and is growing fast [Cracking Open the Scientific Process: <http://www.nytimes.com/2012/01/17/science/open-science-challenges-journal-tradition-with-web-collaboration.html?pagewanted=all>]. ResearchGate has already reached more than 1.5 million users

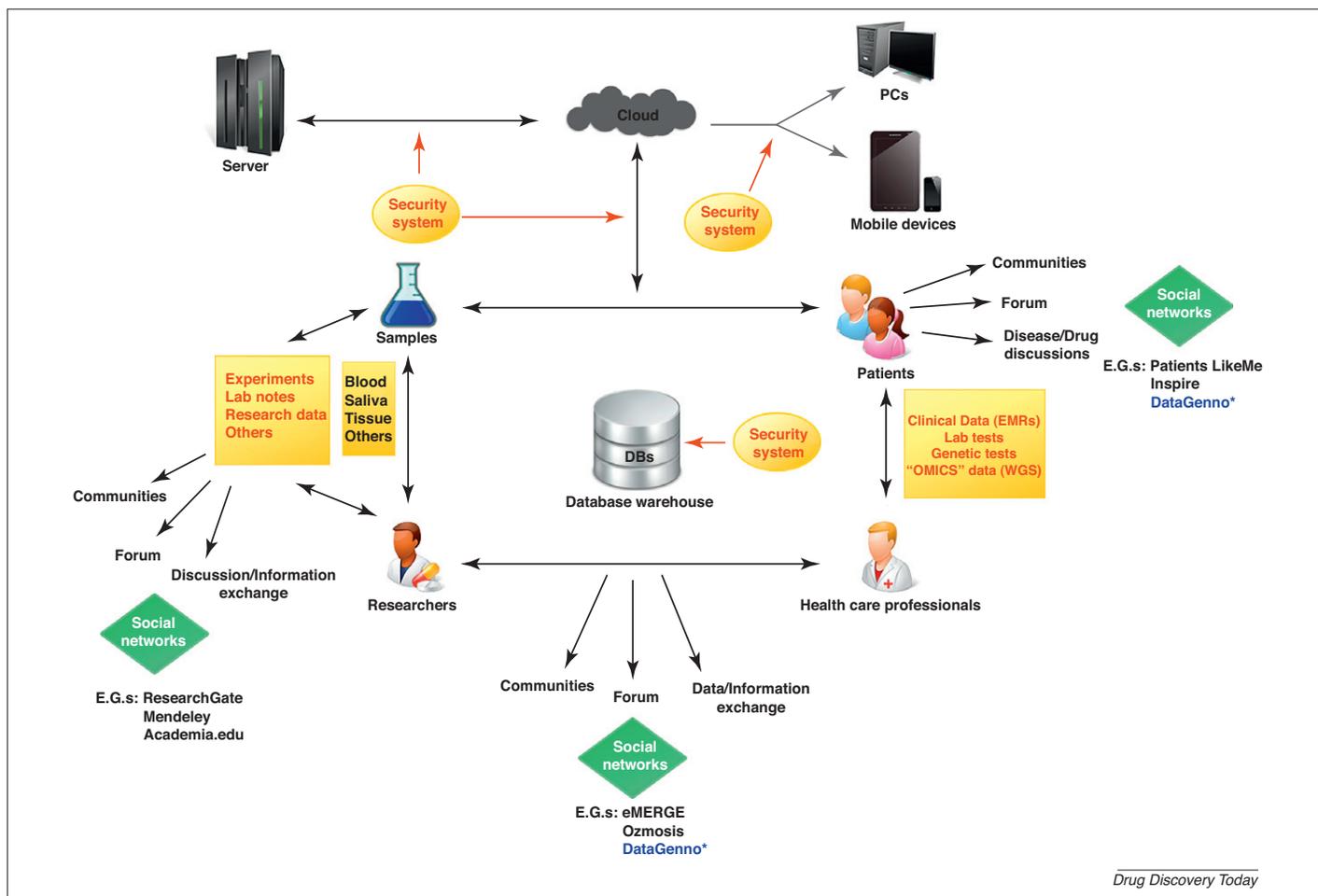


FIGURE 2 Overview of social networks in science and medicine. All the components involved and some specific examples of social networks are shown. Green diamonds represent social networks and a grey cylinder represents big databases. The security system indicates the encryption and/or algorithms used for data privacy and they are shown in yellow circles. Social Networks for scientists (left lower side), for a combination between clinical and research data (center) and for patients (right upper side) to exchange information/data are represented with some examples of portals and websites. DataGenno is shown in blue with an asterisk and it overlaps data from all types of social networks combining clinical and molecular data. DBs, DataBases; EMRs, Electronic Medical Records; PCs, Personal Computers; WGS, Whole Genome Sequencing.

TABLE 2

Social network platforms applied to the scientific field.

Network	Description	Website
Academia.edu	Academia.edu is a platform for academics to share research papers with the mission to accelerate research worldwide.	www.academia.edu
Frontiers Research Network	Frontiers Research Network is a social networking platform for scientists, researchers, academics and anyone that wants to know more about the latest scientific discoveries.	www.frontiersin.org
Mendeley	Mendeley is a free reference manager and academic social network that can help you organize your research, collaborate with others online, and connect to updated scientific literature.	www.mendeley.com
OpeWetWare	OpenWetWare is an effort to promote the sharing of information, know-how, and wisdom among researchers and groups who are working in biology and biological engineering.	www.openwetware.org
ResearchGate	ResearchGate is a social media for scientists in a variety of fields to exchange information and collaborate.	www.researchgate.net
SageBioNetworks	SageBionetworks has the aim to guide the development of an open source community where computational biologists can develop and test models in biological sciences.	www.sagebase.org

around the world that share scientific information in a variety of fields [Cracking Open the Scientific Process: <http://www.nytimes.com/2012/01/17/science/open-science-challenges-journal-tradition-with-web-collaboration.html?pagewanted=all>]. Another successful example is the open source website OpenWetWare, a self-sustaining community created by the Massachusetts Institute of Technology (MIT) students [13]. OpenWetWare has grown quickly into a social network for biologists and biological engineers (Fig. 2). Other portals that are used for instant distribution of scientific ideas include Academia.edu and Mendeley (Table 2). The last social network platform launched for scientists was 'The Frontiers Research Network', designed by the Frontier Publisher [Research Networking: <http://www.frontiersin.org/about/researchnetworking>]. There are also scientific group discussions that are formed in general social media platforms such as Twitter and Facebook; but these are diluted between other group discussions of no interest to the scientific community. In addition, journals have their own accounts in Twitter and Facebook to advertise new articles and scientific material. Despite growing interest by researchers for social media, there is no specific platform well established (except for the ones described above that are very new) for scientists to openly share data and information of their discoveries in the biomedical field.

In medicine, patients are individual data stores, and as online social networks link all of them, people can become part of this dynamic data warehouse (Fig. 2). The application of social networking in medicine proves to be more complicated with the privacy of the patient's health records and data as an issue. Networking between doctors and healthcare professionals has increased in the recent years along with the emergence of social networks. Examples of social media websites that foster information exchange between healthcare professionals include Sermo and Ozmosis, and they facilitate access to physicians' knowledge about current medical practices. Patient networks such as Inspire and PatientsLikeMe are rich sources of patients' insights and clinical data (Table 3). Recent reports have shown how social networks of patients are able to accelerate clinical discovery for complex diseases such as Parkinson's [14], amyotrophic lateral sclerosis (ALS) [15] and rheumatoid arthritis [16]. PatientsLikeMe is an example of a web-based social networking site that captures patient-reported data from people with life-changing diseases [17]. The website is opening up new ways of testing treatments and speeding up patient recruitment into clinical trials for new drugs [17]. PatientsLikeMe has more than 16 disease communities formed by more than 50,000 patients that participate by anonymously sharing treatment options, symptoms, progression and

TABLE 3

Social network platforms and web-based tools applied to medicine.

Network	Description	Website
DataGenno	Datagenno is an interactive database containing molecular and clinical genetic information from diseases targeted to healthcare professionals, research scientists and patients.	www.datagenno.com
eMERGE	The eMERGE Network brings together researchers with a wide range of expertise in genomics, statistics, ethics, informatics, and clinical medicine from leading medical research institutions across the United States.	www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page
Inspire	Inspire is a community of patients that can share information about several different conditions.	www.inspire.com
MediGuard	MediGuard is a patient and consumer network that helps patients to track their medication and exchange information with others.	www.mediguard.org
Ozmosis	Ozmosis is a provider of solutions for physicians to share their knowledge and clinical experiences with each other.	www.ozmosis.com
PatientsLikeMe	PatientsLikeMe is a website for patients that suffer from the same conditions to share their experiences, progress and get input from others in their network.	www.patientslikeme.com
Sermo	Sermo is an online network for physicians with panel discussions about specific topics.	www.sermo.com

outcome data of complex diseases with entire communities [17]. Rather than disseminating medical advice, PatientsLikeMe serves as an online platform for peers to interact with one another in data-driven contexts [17]. The system has worked successfully for ALS because data available for hundreds of patients enrolled in clinical studies and trials was used [15]. The platform can also be used to identify shifts in patients' perceptions and behaviors in response to a safety issue. For example, their opinions on the MS drug Tysabri (natalizumab) was linked to cases of progressive leukoencephalopathy and this information was accurately accessed in PatientsLikeMe forums and discussions [Did the July 2008 PML announcement affect the Tysabri brand?: http://s3.amazonaws.com/patientslikeme_bizdev/TysabriReport.pdf]. It is thus possible that this social network, and many others that have similar features, may be able to detect drug or treatment problems that have not shown up in clinical trials. In fact, this is an example of how social networks can generate epidemiological and clinical data sets that were previously dispersed between the physicians' charts, electronic medical records (EMRs) and clinical histories. Patients engaged in these platforms are able to share information on new drugs being developed for specific conditions, drugs that they are using, side effects and also allow engagement between patients with the same signs and symptoms. It will be of great interest for both pharmaceutical and biotechnology companies that are developing and testing new drugs for specific diseases to have a data warehouse of patients giving real-time feedback on clinical trials and also newly approved drugs. The ultimate goals will be to directly partner with industry in a 'Public-Private Partnership' or PPP in the discovery process to accelerate the development of biomarkers, drugs and new treatments for diseases. This could speed the process for drug approval by the FDA with the direct participation of the patient community.

A new social network-type of solution that could help promote better engagement between healthcare professionals to share information about diseases is DataGenno (<http://www.datagenno.com/>) [18]. This new platform is a database with high storage capacity that can be used by healthcare professionals and researchers to access patient and sample information in a real-time fashion [18]. DataGenno has been designed to bridge the gap between healthcare professionals and scientists, combining clinical, genetic and genomic information from patients and samples because it provides not just clinical data but also all genetic information available from specific diseases and patients [18]. This association of features could help improve patient care and reduce costs related to diagnosis, drug administration and side effects associated to specific drugs.

The basis of every patient–physician interaction is trust, because shared personal health information is kept confidential and only used exclusively for the patient's benefit. Thus, the main barrier for the adoption of social networks in health care and medicine will be privacy and security of the data (this issue is discussed in a different section below). Interestingly, privacy issues always drive innovation – one example is the constant fear that social networks such as Facebook and Twitter have about the privacy of their users. In the case of life sciences, the privacy discussions will be even deeper and more complex. There is growing sense among data privacy experts that data de-identification and encryption are major standards for privacy protection. Improvements in privacy will be very important

to foster the use of social networks and web-based tools by both the scientific and medical communities.

Web-based tools to predict and track disease outbreaks

Millions of people worldwide search online for health-related information every day, which makes web search queries a valuable source of information on collective health trends [19]. This type of research was mainly done by the Center for Diseases Control (CDC), which is an organization from the United States government. Even though the CDC was always important to track outbreaks of diseases, not just in the United States but worldwide, it had limitations. Recently, Google Inc. did a partnership with the CDC in a study to evaluate if their search engine was able to detect areas that an influenza epidemic could be eminent [2,20,21]. The study evaluated queries that were made on the Google website using keywords containing anything related to influenza. Using this approach, they were able to track areas that could be having an epidemic of influenza. Their results were compared with the ones got from the CDC traditional methods in a parallel study to determine which of them could detect areas affected earlier and Google's algorithm was faster and more accurate [2]. Google web search logs could provide one of the most timely, broad-reaching disease monitoring systems to date [2]. Whereas traditional monitoring systems used by the CDC and the government require 1–2 weeks to gather and process the data, Google's estimates were available on a daily basis [2]. Now both Google and the CDC work together in any disease outbreak proving web tools that can help governments in tracking disease outbreaks.

Google searches were also used in a non-seasonal flu epidemic to track the H1N1 virus in the outbreak of 2009 and this model has been used regularly for health-seek behavior [20]. This new service provided by Google was named 'Google Trends' or GT and it uses anonymized, aggregated Internet search activity to provide real-time estimates on any type of influenza activity worldwide (Table 1) [21]. Google works closely with both public and private healthcare practitioners to improve tools such as Google Trends to use it as a blueprint to track infectious diseases [21]. However, Google Trends is in its early stage of development and needs improvements, such as standardization by grouping search queries into syndromes. For example, surveillance of diseases using web-based tools has been successfully established in emergency departments with medical records in hospitals and in educational institutions [22,23].

The second tool that is improving the health sector is social media. Two recent studies were able to show, using the Twitter platform, that we can monitor trends based on the posts of users. PacSocial, which is a corporation that is focused on the development of technologies that enable large-scale shaping of social communities online was able to use 'socialbots' (which represent virtual individuals who infiltrate social media platforms) to influence connections and interactions between two users [PacSocial: Field Test Report: http://www.pacsocial.com/files/pacsocial_field_test_report_2011-11-15.pdf]. They were able, for the first time, to show how tools such as 'socialbots' can influence human connections online [The Socialbot Network: When Bots Socialize for Fame and Money: http://www.acsac.org/2011/openconf/modules/request.php?module=oc_program&action=summary.php&id=175]. In a different study, professors from MIT focused on Twitter users to examine the effects of social network structure innovation adoption

creating a model based on geography and mass media [24]. The authors showed that mass media was responsible for increasing Twitter users with time [24]. Interestingly, the authors also showed that the mouth-to-mouth spread of these medium between people is powerful, especially young individuals [24]. The social media Twitter was also used to track diseases such as the influenza A H1N1 virus in the USA recently [25].

Social networks such as Facebook and Twitter could be used to improve surveillance, track diseases, symptoms and health issues in specific groups of people and geographic regions (Table 1). Social media adoption by public health agencies is in the early stage, but it can reach and foster interactivity and engagement between its users [22]. These reports on new technologies influencing society and decisions, especially in epidemic outbreaks, show how empowering search engines and social media have become for healthcare in our society.

The future of personalized medicine – patients in control of their data

The term 'P4 Medicine' was first introduced more than 5 years ago by the researcher Leroy Hood in an attempt to change medicine from a reactive to a proactive perspective [26]. The ultimate goal of this approach is to maximize wellness for each individual rather than just treat the causes of the disease. P4 stands for the letter 'P' in four words: Predictive, Preventive, Personalized and Participatory [27]. Emerging from debates related to healthcare system reform and translational research, this new way of thinking about medicine has put the focus on the patients, other than the physicians. In this case, the patients have control on their own health data. One exciting prospect for this new type of medicine is the opportunity for patients, especially those who do not have years to wait for a clinical trial, to monitor and participate in real-world natural experiments.

P4 medicine uses four interconnected aspects to improve both the effectiveness and efficiency of health management at the individual level. The first 'P' that stands for Predictive medicine studies the risk factors, predispositions, and genetic susceptibilities that exist before issues arise by the use of genetic testing and whole-genome sequencing, technologies that have evolved fast in the past decade [26,27]. This gives the patient and healthcare team a warning sign long before a disease is diagnosed. It also gives information that can influence changes in the diet and lifestyle of individuals. The second 'P' stands for Preventive medicine and uses the knowledge of the risk factors learned from predictive medicine to change a person's behavior. It also helps in predisease treatments with early screening to cut off the development of more serious conditions. The third 'P' is for Personalized medicine in which each individual will respond differently to treatments. This helps the physician understand which drugs works and the ones that do not work. Personalized or individualized medicine studies and treats each individual as a unity and the current system of health care attempts to treat patients in a universal fashion, which is clearly wrong. Patients having this knowledge before a treatment empower them to take different actions depending on their clinical options. Better predictions on what works more effectively will help patients and the healthcare team responsible in selecting the treatment to best meet the patient's needs. The fourth and last 'P' is for Participatory medicine in which the

patient and doctor relationship will become more interactive, helping in better diagnosis and effective treatments. Participation of the patient in health care decisions will empower the patient and can encourage innovation among health care teams.

The emergence of new technologies in the Internet such as search engines and social media is leveraging these interactions. Patients already come to the physician's office with all the information on their symptoms or disease and in some cases with more information than the health care professional about their own case. Social media is empowering individuals to share their data with others that may have the same conditions (such as the PatientsLikeMe website mentioned earlier). This new era in medicine has been shaped by technology breakthroughs such as increasing computer storage capacities and speed, the emergence of the cloud and systems biology (Fig. 2). This new concept that enables data and models to be shared by patient communities in an open-source manner will change the style and speed of discovery (Fig. 2). This revolution in medicine will also facilitate clinical trials that need to recruit a very large number of patients to test new drugs and therapies [28].

The future of personalized medicine 'P4' can be glimpsed in a recent report of an individual's genomic sequence with RNA, protein, metabolic and autoantibody profiles taken 20 times over a two-year period [29]. This new concept was coined 'integrative Personal Omics Profile' or iPOP and it was created by merging all 'omics' from this individual to uncover his predisposition to disease and his body's response to viral infections [29]. As a proof-of-principle example of personalized genomic medicine, it is distinct from other studies because it applies whole-genome diagnostics to a healthy person rather than to individuals with disease, before the disease appears [29]. The results of this project revealed that the sequenced individual is genetically predisposed to type 2 diabetes, despite no family history or other risk factors [29]. During the study, his blood glucose levels escalated following the second of two viral infections, and he was subsequently diagnosed with diabetes. Dietary and lifestyle changes were made following the study to manage his blood sugar levels and the glucose levels went back to normal showing the power of such studies [29]. Thus, it is already possible to be in charge of your own health and make decisions based on this report. This will indeed change the way we approach and treat diseases in the future.

Cloud computing, big data analysis and biomedical research

Biomedical research in the post-genome era is intensely data-driven and increasingly more integrative as new technologies are introduced, such as second- and third-generation DNA sequencing, mass spectrometry for proteomics analysis, and different imaging systems to identify novel biological insights in cells and organisms. The volume and complexity of biomedical data is increasing exponentially as faster high-throughput machines are introduced. As a result, many research institutes, biotechnology companies, pharmaceutical companies, and computational laboratories are considering cloud computing as a cost-effective alternative to process and store this vast amount of data. Cloud computing is a term used to describe a system that allows easy access to a shared pool of resources (e.g. applications, servers, storage, networks) that can be quickly allocated and released with

minimal effort by an administrator. The 'cloud' acts like a virtual supercomputer that can pull together a cluster of other computers to perform certain tasks, especially when dealing with big amounts of data (Fig. 2) [30].

Efforts in next-generation sequencing or NGS [31], comparative genomics [32], and proteomics [33] have already successfully incorporated the cloud to expedite their amount of data processing (Fig. 2). However, biomedical infrastructure to deal with big data analytics lags well behind the curve and the prospect is that it will become worse with exponential growth of data points. Despite these limitations, it is becoming clear that methods that are used by social media such as Facebook, will prove as revolutionary for science and healthcare as they have been for communications and retail. The main solution to this issue is the ability to create an infrastructure that allows researchers and physicians to harvest big data points in an efficient way, similarly to the way social media is now able to generate maps and networks of people, and gather their data in seconds.

The healthcare system is dispersed and disjointed with lack of capacity to store and process the amounts of data that are coming from widespread whole-genome sequencing together with all the clinical data from patients and samples in the case of researchers. According to Katharine Miller [Big Data Analytics In Biomedical Research: <http://biomedicalcomputationreview.org/content/big-data-analytics-biomedical-research>], there are six important milestones that should be considered when combining big data (clinical and 'omics' information) analysis for biomedical research with cloud computing: (i) define big data in biomedicine – big data on biomedicine are coming from two distinct ends which are the genomics-driven end and the medical records of patients. We should be able to integrate these data sets in a more efficient manner; (ii) select the right tools for big data analysis – dealing with multiple layers of data is not trivial so we need to be able to integrate. In addition to machine learning and Artificial Intelligence (AI), there is also a need for approaches that are able to scale up to the interpretation of the data; (iii) clinical data is already abundant but genomics is a newcomer – large healthcare organizations already have big collections of data on individuals in their computers or files and these records are becoming digitalized, however, genomics data is somewhat new. Most hospitals and research institutes get big data on the scales of Facebook, Amazon, Google and others but do not know how to correlate the clinical data with the molecular information that genomics and the other 'omics' are generating; (iv) hypothesis-free approach – in this case rather than starting with a question or hypothesis the idea is to let the technology of big data run and give you the correlations. For example, Medco (<http://www.express-scripts.com/>) is using this approach for a study with the drug Plavix and they were able to take the strongest signals from data and avoid wasting time on hypothesis that do not lead to any answer [34]. Google also uses hypothesis-free approaches, especially when tracking disease outbreaks with the queries that were searched using their system [35]; (v) adding genomics to this mix – examples that combined clinical data with genomics and successfully identified novel therapeutic intervention points include the joint study between GNS Healthcare (<http://www.gnshealthcare.com/>) and Biogen (<http://www.biogenidec.com/>) [36]. They were able to find that one-third of arthritis patients do not respond to a specific therapeutic

regimen [36]. Another example trying to index genomics data in the mix of the clinical data already available is the eMERGE Network, which is a NIH funded collaboration linking medical records data with genomics data [37]. There is also the company Explorys (<https://www.explorys.com/>), which is a big data bioinformatics warehouse, and will implement the type of data architecture that supports Google, Yahoo and Facebook in biomedicine and (vi) making big data actionable – this will be the ultimate goal in biomedical research, which is to implement an Amazon or Facebook style in biomedicine. One successful example was a study that Medco did in collaboration with the Mayo Clinic using cloud infrastructure to increase the success for a genotyping test [38].

Although very ambitious, big data analysis in biomedicine will be a stepwise process affecting the time taken for action in clinical practice. In the end, both physicians and researchers along with their patients will gain from these technologies. Big data analytics in biomedicine lags finance and commerce because it has not taken advantage of commercial methods of handling large data sets such as parallel computing. For example, Facebook released a tool that enables users to give information on their organ donor status providing significant contribution for patients that are in need of an organ because this will facilitate finding better matches for organ donations (Table 1). The practices and experience from big corporations with large amounts of data such as Amazon, Google, Yahoo, Facebook will need to propagate back to the academic and research arenas to help accelerate the process of acquiring and organizing both clinical and scientific data [Privacy and Biomedical Research: Building a Trust Infrastructure: <http://biomedicalcomputationreview.org/content/privacy-and-biomedical-research-building-trust-infrastructure>].

Drug discovery and new treatments based on online social networks

Social networking is beginning to make an impact in the drug discovery process. A convergence of different commercial and publicly accessible chemical informatics, databases and social networking tools is positioned to change the way that scientific and medical collaborations are initiated, maintained and expanded, particularly in the realm of complex, rare and orphan diseases [4, Privacy and Biomedical Research: Building a Trust Infrastructure: <http://biomedicalcomputationreview.org/content/privacy-and-biomedical-research-building-trust-infrastructure>]. A community-based platform that combines traditional drug discovery informatics with Web 2.0 platforms and strong privacy is believed to be the key to facilitate richer and instantaneous collaborations involving health care professionals with the same interests [39]. This way data from differential diagnosis, experiments and new drugs being tested are archived, mined and then selectively shared between colleagues in the Internet with standardized formats (Fig. 2).

New drugs are subject to exhaustive scrutiny, yet there has never been a corresponding effort to collect reports of drugs delivering unexpected benefits [39]. If open innovation can lead to the creation of the world's most complete encyclopedia such as Wikipedia, that same approach could be used to capture the exceptional untapped value associated with existing drugs, and to power the discovery of important new medicines. In the Facebook and

Google era, there might be a better, more efficient and systematic way of harnessing this communal wisdom for drug discovery. One example is the Collaborative Drug Discovery (CDD) network that has a growing community of hundreds of scientists collaborating in a web-based platform to advance drug candidates more efficiently [39]. The platform supports a wide range of collaborations and the current community includes leading researchers working on neglected diseases in the developing world such as malaria, tuberculosis, Chagas disease, among others [39]. Other examples include the ChEMBL database [40] that is an open database containing information for a large number of drug-like bioactive compounds; Chemspider [41], an Internet database of known chemical compounds that could be tested as new drugs for diseases and OCHEM [42], a web platform for data storage, which offers collaborative development of models and prediction tools as well as data sharing in the Internet. In light of these initiatives, large pharmaceutical organizations are in the process of transitioning from a fully integrated pharmaceutical company to fully integrated pharmaceutical networks with a social networking component connecting members with similar interests as well as project tracking and planning.

To foster better communication and data sharing between professionals in biomedical research, a platform that is freely available at datagenno.com has been developed with the objective of providing a web space for healthcare professionals, researchers, and even patients with their relatives, to exchange information about rare genetic and complex diseases (Table 2, Fig. 2) [18]. There are numerous features that differentiate DataGenno from the websites already available online. The portal has a dynamic database with patient, sample and disease information with specific images for each sign or symptom, a search engine for differential diagnosis and features for information exchange between healthcare professionals and patients (i.e. messaging, media tools, among others). It is also the first portal that provides a strong connection between clinical and molecular genetics (especially genomics data) in the same environment on the web [18]. DataGenno's platform will be mainly targeted to physicians, scientists, physician scientists, genetic counselors, nurses and other professionals that have interests in knowing more about specific diseases.

Academic and private sectors need user-friendly and efficient tools for information exchange and to share data points from patients. Pharmaceutical companies have interdepartmental databases of patients and samples for research, clinical trials and for patient follow-up, but there is no communication between these databases, even within databases housed in the same company [Social Media in Science and Medicine: <http://www.laboratory-journal.com/science/information-technology-it/social-media-science-and-medicine>]. To date, there are no social networks in the market that offer useful solutions for science and medicine in the clinical genetics field [Social Media in Science and Medicine: <http://www.laboratory-journal.com/science/information-technology-it/social-media-science-and-medicine>]. It is paramount for physicians to have reliable tools that will facilitate diagnosis and identify better treatments for diseases.

As in the medical field, there is an increasing need for solutions in the scientific field for networking and communication between professionals with common interests. ResearchGate, Academia.edu and Mendeley (Table 2, Fig. 2) are already providing some

tools to achieve this goal but more specialized open-source media tools are needed in science and medicine. Despite existing competition in science, the best way for science to have a major impact in society and change the way we approach diseases is through collaboration and networking. In the end, the outcome will be the development of better drugs and faster translation of basic science to the patient's bedside.

Finding solutions for the privacy issue – the biggest challenge

Privacy is the basis of every patient–physician interaction with personal health information always kept confidential and used only for the patient's benefit. Most patient privacy of health records is codified by HIPAA, which stands for the Federal Health Insurance Portability and Accountability Act. Until recently, most health records were stored in paper inside filing cabinets. However, this has been changing since 2009, when the Obama administration and the American Congress made billions of dollars available to fulfill the needs of health care to digitalize all clinical and scientific data [43]. At that time, EMRs were implemented and used by approximately 11% of hospitals and institutions and less than 5% of physicians [43] in the USA. In the past three years, the adoption of EMRs by American doctors has roughly doubled to approximately 30% [43]. On top of that, 80% of hospitals are also developing their own digital medical records [43].

Increased personal health data accrue in larger databases and concerns over privacy increase. While informatics experts have raised many of the vulnerabilities of the privacy-protection schemes currently used, a discussion for implementation of more powerful technologies in the near future have begun. Based on this, we are entering the 'golden age' of privacy research applied to biomedicine. By contrast, great challenges lie ahead, not just in legislation but also technology development to deal with increasing amounts of scientific and health data. Interestingly, the progress occurring in privacy research in the biomedical field is driven by fear; for example, the fear that the biomedical data now being assembled could be vulnerable to the exact same kinds of privacy problems that have recently affected information aggregators such as the search engine Google and the social network Facebook [Facebook's Privacy Issues Are Even Deeper Than We Knew: <http://www.forbes.com/sites/chunkamui/2011/08/08/facebook-privacy-issues-are-even-deeper-than-we-knew/>].

The evidence for breaches in the medical records, clinical data and genomic information remains very low. Most of the incidents to date have involved simple failures of security, or of access control, such as the theft or loss of unsecured computers containing EMRs, rather than unintentional leakage of information from large biomedical databases. Academic researchers who were trying to find weakness in the systems to identify the medical records of particular individuals in a hospital system have perpetrated the worst privacy problems. Also, identifying participants in large-scale genomics studies such as Genome Wide Association Studies (GWAS) designed to link specific diseases to genetic variants had privacy problems with leakage of health data from participants [44,45]. There is a rising concern over the privacy risks associated with genomic data in particular, especially now that whole genomes of individuals can be sequenced faster with lower costs. Genomic data is distinct when compared with morphological

and clinical data, the data is extremely stable and can be used to predict the likelihood that an individual might be more susceptible to diseases (as exemplified by the study of personalized 'omics' in the sections above). The main fear is that this information could be easily used to deny health coverage or even employment opportunities for an individual.

The privacy rights of individuals has prompted a great deal of innovation in the field, especially among mathematicians, cryptographers and computer scientists who are working together to develop better mechanisms that will allow researchers to analyze biomedical data without compromising privacy. Some important points that need to be taken into account to help overcome the privacy issue include: (i) to integrate privacy and security approaches in a single solution; (ii) to make minimum controls and standards combined with a strong encryption for the data being stored – the less complexity in the system the better; (iii) educate people about the privacy issue before applying a solution – literacy is crucial and privacy needs to be as important as the solution provided by any company in this sector. Thus, it is clear that the privacy problem will require finding an integrated solution with trustful infrastructure that introduces not only technical solutions but also legal frameworks to efficiently combine technology with education and legislation.

Concluding remarks

The web has impacted our society, but over the next few years, the web is going to entirely re-invent the way that scientists and physicians interact, especially with patient data. Social media networks will have increased value in the scientific and medical communities. There is indeed a conundrum faced by life sciences because both pharmaceutical and biotechnology industries need information from people (mostly patients), patients need

information from these companies and the life sciences industry also need to collaborate outside their walls building connections with other industries and institutions. Science needs to be more open and scientists have to be able to share data to speed up the process of transition between basic to applied research. As regulatory clarity emerges with better algorithms and encryption to ease privacy issues, more companies will adopt social media tools and these technologies will have the potential to revolutionize healthcare, clinical trials and research the same way that they are impacting our society.

Developing solutions to allow better ways to share information will add value to the life sciences and medical communities. In fact, science and medicine are undergoing one of the most exciting changes in history. The full adoption of online platforms such as social networks by scientists and physicians will transform medicine. Healthcare professionals will start to interact and communicate in ways that will have an enormous impact on scientific progress, drug discovery and in the development of new treatments for diseases.

Financial disclosure

The author is a consultant for Genomic Enterprise and the co-founder of DataGenno Interactive Research Ltda.

Acknowledgements

I would like to thank DataGenno Interactive Research's Team including Dr Marcelo P. Coutinho, Julio C.B. Araujo, Eduardo B.F. Junior, Lucas F. Correa, Italo M. Soares and Maria Celeste L.A.M. Cabral for their inputs and technical assistance. I am really thankful to Eduardo B.F. Junior that designed both figures for this article, Tristan Gill for giving important inputs to improve the article and Kelly Arndt for English corrections.

References

- Cerf, V.G. and Kahn, R.E. (1974) A protocol for packet network interconnection. *IEEE Trans. Comm. Tech.*, vol. COM-22, V 5 pp. 627–641
- Ginsberg, J. *et al.* (2009) Detecting influenza epidemics using search engine query data. *Nature* 457, 1012–1014
- Schmidt, C.W. (2012) Trending now: using social media to predict and track disease outbreaks. *Environ. Health Perspect.* 120, a30–a33
- Bailey, D.S. and Zanders, E.D. (2008) Drug discovery in the era of Facebook – new tools for scientific networking. *Drug Discov. Today* 13, 863–868
- Ugander, J. *et al.* (2012) Structural diversity in social contagion. *Proc. Natl. Acad. Sci. U. S. A.* 109, 5962–5966
- van der Linden, H. *et al.* (2009) Inter-organizational future proof EHR systems. A review of the security and privacy related issues. *Int. J. Med. Inform.* 78, 141–160
- Nielson, M. (2011) *Reinventing Discovery: The New Era of Networked Science*. Princeton University Press
- Munos, B. (2006) Can open-source R&D reinvigorate drug research? *Nat. Rev. Drug Discov.* 5, 723–729
- Ginsparg, P. (2011) ArXiv at 20. *Nature* 476, 145–147
- Rohde, H. *et al.* (2011) Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* 365, 718–724
- Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
- Sansone, S.A. *et al.* (2012) Toward interoperable bioscience data. *Nat. Genet.* 44, 121–126
- Butler, D. (2005) Science in the web age: joint efforts. *Nature* 438, 548–549
- Wicks, P. and MacPhee, G.J. (2009) Pathological gambling amongst Parkinson's disease and ALS patients in an online community (PatientsLikeMe.com). *Mov. Disord.* 24, 1085–1088
- Wicks, P. *et al.* (2011) Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nat. Biotechnol.* 411–414
- Specker, C. *et al.* (1998) RheumaNet – a novel Internet-based rheumatology information network in Germany. *Br. J. Rheumatol.* 37, 1015–1019
- Brownstein, C.A. *et al.* (2009) The power of social networking in medicine. *Nat. Biotechnol.* 27, 888–890
- Costa, F.F. *et al.* (2011) DataGenno: building a new tool to bridge molecular and clinical genetics. *Appl. Clin. Genet.* 4, 45–54
- Hersher, R. (2012) Internet data miners strike disease detection gold. *Nat. Med.* 18, 185
- Cook, S. *et al.* (2011) Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE* 6, E23610
- 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
- Thackeray, R. *et al.* (2012) Adoption and use of social media among public health departments. *BMC Publ. Health* 12, 242
- Zhang, Y. *et al.* (2011) Evaluating syndromic surveillance systems at institutions of higher education (IHEs): a retrospective analysis of the 2009 H1N1 influenza pandemic at two universities. *BMC Publ. Health* 26, 591
- Toole, J.L. *et al.* (2012) Modeling the adoption of innovations in the presence of geographic and media influences. *PLoS ONE* 7, E29528
- Signorini, A. *et al.* (2011) The use of Twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic. *PLoS ONE* 6, E19467
- Auffray, C. *et al.* (2010) Predictive, preventive, personalized and participatory medicine: back to the future. *Genome Med.* 2, 57
- Hood, L. and Flores, M. (2012) A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *Nat. Biotechnol.* 29, 613–624

- 28 Kaye, J. *et al.* (2012) From patients to partners: participant-centric initiatives in biomedical research. *Nat. Rev. Genet.* 13, 371–376
- 29 Chen, R. *et al.* (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293–1307
- 30 Fusaro, V.A. *et al.* (2011) Biomedical cloud computing with Amazon Web Services. *PLoS Comput. Biol.* 7, E1002147
- 31 Kohlmann, A. *et al.* (2012) Integration of next-generation sequencing into clinical practice: are we there yet? *Semin. Oncol.* 39, 26–36
- 32 Wall, D.P. *et al.* (2010) Cloud computing for comparative genomics. *BMC Bioinform.* 11, 259
- 33 Halligan, B.D. *et al.* (2009) Low cost, scalable proteomics data analysis using Amazon's cloud computing services and open source search algorithms. *J. Proteome Res.* 8, 3148–3153
- 34 Mega, J.L. *et al.* (2010) Reduced-function CYP2C19 genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. *J. Am. Med. Assoc.* 304, 1821–1830
- 35 Dugas, A.F. *et al.* (2012) Google Flu trends: correlation with emergency department influenza rates and crowding metrics. *Clin. Infect. Dis.* 54, 463–469
- 36 Xing, H. *et al.* (2011) Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. *PLoS Comput. Biol.* 7, E1001105
- 37 Kho, A.N. *et al.* (2011) Electronic medical records for genetic research: results of the eMERGE consortium. *Sci. Transl. Med.* 3, 79re1
- 38 Wang, L. *et al.* (2011) Genomics and drug response. *N. Engl. J. Med.* 364, 1144–1153
- 39 Hohman, M. *et al.* (2009) Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov. Today* 14, 261–270
- 40 Wassermann, A.M. and Bajorath, J. (2011) BindingDB and ChEMBL: online compound databases for drug discovery. *Expert Opin. Drug Discov.* 683–687
- 41 Little, J.L. *et al.* (2012) Identification of known unknowns utilizing accurate mass data and ChemSpider. *J. Am. Soc. Mass Spectrom.* 179–185
- 42 Sushko, I. *et al.* (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* 533–554
- 43 Costa, F.F. (2011) Basic research, applied medicine and EHRS – are we on the right track? *J. Cancer Sci. Ther.* <http://dx.doi.org/10.4172/1948-5956.1000e102>
- 44 Craig, D.W. *et al.* (2011) Assessing and managing risk when sharing aggregate genetic variant data. *Nat. Rev. Genet.* 12, 730–736
- 45 Im, H.K. *et al.* (2012) On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet.* 90, 591–598