EJHG Open

European Journal of Human Genetics (2015) 23, 1271–1278

www.nature.com/ejhg

POLICY

Genomic cloud computing: legal and ethical points to consider

Edward S Dove*[,1], Yann Joly[1], Anne-Marie Tassé[2], Public Population Project in Genomics and Society (P3G) International Steering Committee, International Cancer Genome Consortium (ICGC) Ethics and Policy Committee and Bartha M Knoppers[1,2]

The biggest challenge in twenty-first century data-intensive genomic science, is developing vast computer infrastructure and advanced software tools to perform comprehensive analyses of genomic data sets for biomedical research and clinical practice. Researchers are increasingly turning to cloud computing both as a solution to integrate data from genomics, systems biology and biomedical data mining and as an approach to analyze data to solve biomedical problems. Although cloud computing provides several benefits such as lower costs and greater efficiency, it also raises legal and ethical issues. In this article, we discuss three key 'points to consider' (data control; data security, confidentiality and transfer; and accountability) based on a preliminary review of several publicly available cloud service providers' Terms of Service. These 'points to consider' should be borne in mind by genomic research organizations when negotiating legal arrangements to store genomic data on a large commercial cloud service provider's servers. Diligent genomic cloud computing means leveraging security standards and evaluation processes as a means to protect data and entails many of the same good practices that researchers should always consider in securing their local infrastructure.

European Journal of Human Genetics (2015) **23,** 1271–1278; doi:10.1038/ejhg.2014.196; published online 24 September 2014

INTRODUCTION

The genomic research community is facing a big data challenge. The cost of sequencing is falling faster than that of storage and bandwidth, and thanks to advanced technologies for genetic sequencing and analysis, more data have been generated than ever before. Consider that the average human whole-genome sequence contains approximately three billion data points (ie, 'base pairs') and generates roughly 100 gigabytes of data, and that the whole genome of a tumor and a matching normal tissue sample consumes 1 terabyte of uncompressed data. A project utilizing thousands of genomes (not to mention phenotypic data and the linking of local data with online public data) for disease research would quickly generate petabytes of data.

Yet, current sequencers are incapable of generating a single string of properly organized nucleotides. Instead, they produce shorter, fragmented and unordered sections. Researchers must rely on technicians and computers to properly organize them. At the same time, the current reality is that the amount of genomic data and associated clinical data needed to procure the statistical power required to advance biomedical research and clinical practice exceeds the technical capacity of any single site or server.

For instance, the International Cancer Genome Consortium (ICGC) and its member project, The Cancer Genome Atlas (TCGA), have analyzed and released the data from over 10 000 donors, generating >1.5 petabytes of raw and interpreted data. It is estimated that when the ICGC project is complete in 2018, it will comprise $>50\,000$ individual genomes with an estimated 10–15 petabytes of data.[1] Although the interpreted results are available for browsing, mining and downloading at a site-specific data portal (in Toronto), and the raw sequencing data are archived at two sites (the European Genome-Phenome Archive (EGA) in the UK and at cgHub in UC Santa Cruz in the United States), there are many compelling scientific reasons for researchers to have remote access to the raw sequencing data. Foremost, pan-cancer analyses can be performed to identify commonalities and differences among the various cancer types.

Certain projects have been successfully designed to handle specific legal and ethical issues surrounding identifiability for large-scale community projects. For example, Bio-PIN allows a biospecimen donor to be registered without any identity data, and a distinguishing biological PIN code (called Bio-PIN) is produced based on that individual's unique biological characteristics (eg, nucleotides that are not part of any genotype) in a way that the resulting PIN code cannot be linked back to the individual.[2] This not only ensures anonymity but also enables a secure, two-way, individual-controlled, web-based communication with the research platform, such as a biobank. Similarly, DataSHIELD enables analysis of pooled (but not shared) data based on parallel local processing, distributed computing and advanced asymmetric encryption algorithmic methods.[3,4]

Yet, the biggest challenge in twenty-first century data-intensive science is more fundamental: comprehensive analyses of genomic data sets to advance biomedical research and clinical practice cannot be done without greater collaboration, a vast computer infrastructure and advanced software tools. Simply buying more servers for a local research site is no longer an optimal or even feasible solution to handle the data deluge. As a result, researchers are increasingly turning to cloud computing both as a solution to integrate data from genomics,

[1]Centre of Genomics and Policy, McGill University, Montreal, QC, Canada; [2]Public Population Project in Genomics and Society, McGill University, Montreal, QC, Canada
*Correspondence: ES Dove, Centre of Genomics and Policy, McGill University, 740 Dr. Penfield Avenue, Suite 5200, Montreal, QC H3A 0G1, Canada. Tel: +1 514 398 8187; Fax: +1 514 398 8954; E-mail: edward.dove@mcgill.ca
Received 13 May 2014; revised 4 August 2014; accepted 19 August 2014; published online 24 September 2014

systems biology and biomedical data mining and as an approach to mine and analyze data to solve biomedical problems.[5,6]

GENOMIC CLOUD COMPUTING

Though an evolving paradigm, genomic cloud computing can be defined as a scalable service where genetic sequence information is stored and processed virtually (ie, in the 'cloud') usually via networked, large-scale data centers accessible remotely through various clients and platforms over the Internet.[7] Rather than buying more servers for the local research site, as was done in the past, genomic cloud computing allows researchers to use technologies, such as application programming interfaces (APIs) to launch servers (Figure 1). Various cloud computing platforms have emerged for genomic researchers, including Galaxy,[8] Bionimbus[9] and DNAnexus,[10] which allow researchers to perform genomic analyses using only a web browser. These platforms in turn may run on specific clouds provided by cloud service providers (CSPs).

Four deployment models of cloud computing have emerged in recent years. Each carries different technical, legal and ethical considerations for researchers.[11]

- *Commercial* cloud infrastructure (eg, Google and Amazon) is provisioned for open use by the general public and may be owned, managed and operated by a business, academic or government organization or some combination of them. Commercial cloud infrastructure allow customers to build, manage and scale an infrastructure based on their needs.
- *Community* cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns.
- *Hybrid* cloud infrastructure comprises two or more distinct cloud infrastructures (private, community or commercial) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability.
- *Private* cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers.

In this article, we focus on commercial cloud infrastructures that typically process data transnationally, have built-in security mechanisms and can handle the large-scale data generated by international genomic research projects. Large commercial CSPs have the advantage of often already having public genomic data sets on their cloud infrastructure, which save researchers time and effort in organizing and paying for the transfer of common data such as reference genomes.

In addition to these categories, cloud computing can be organized into different types of service categories:[12]

- *Infrastructure as a Service* provides raw computing resources, including processing power (known as 'compute') and storage to the user. Often this service allows users to install their own operating systems and applications on the provider's infrastructure (ie, an ability to rent the compute space) and mount bespoke research tools for genomic analysis on the cloud. Examples of Infrastructure as a Service include Amazon Elastic Compute Cloud (http://aws.amazon.com/ec2/) and Google Compute Engine (https://cloud.google.com/products/compute-engine/).
- *Platform as a Service* provides platforms for developing and launching software applications (eg, Google's App Engine).

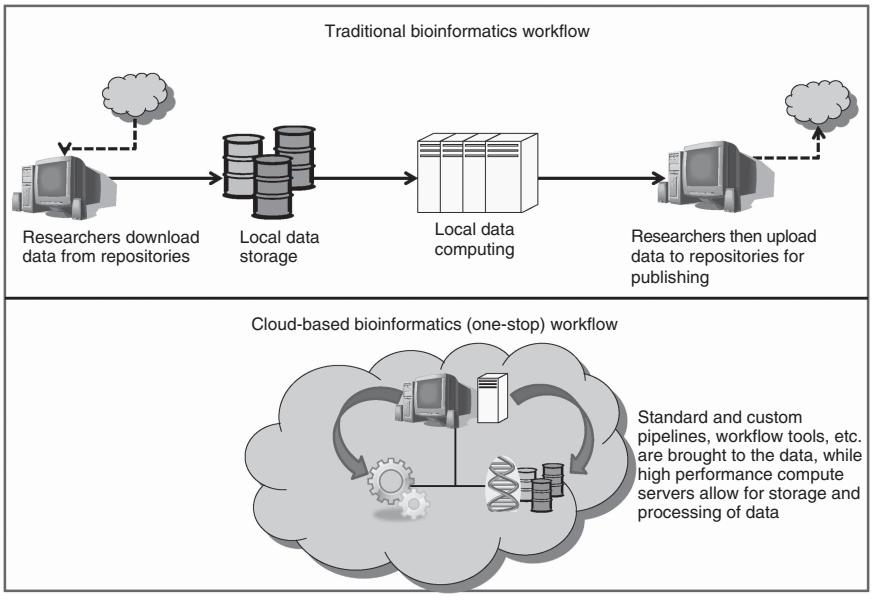

Figure 1 Contrast between traditional bioinformatics workflow and new cloud-based workflow. The traditional bioinformatics workflow is characterized by researchers downloading or uploading genomic and health-related data to local on-site storage (eg, computers) for processing, analysis and obtaining of results. The results are then uploaded to repositories for publishing. This process is typically slower, redundant and necessitates high IT capital expenditure. Indeed, the traditional practice of genome analysis requires researchers to spend weeks to months downloading hundreds of terabytes of data from a central repository before computations can begin. By contrast, the new cloud computing bioinformatics model eliminates the need for researchers to download the data to their own computers. Instead, it is characterized by a one-stop workflow where the compute (eg, standard and custom pipelines, workflow tools) is brought to the data. Particularly in Infrastructure as a Service cloud computing, researchers can upload their analytic software into a cloud, run the software, and download the compiled results in a secure fashion. Platform as a Service and Software as a Service cloud computing can also provide a one-stop bioinformatics workflow, albeit with less raw computing resources made available to researchers.

- *Software as a Service* provides end-user applications, such as Dropbox or Google Docs.

In this article, we focus on Infrastructure as a Service cloud computing.

Cloud contracts can be either non-negotiable 'standard form' contracts or negotiated contracts tailored to fit the specific requirements of the cloud service customer. Typically for scalability reasons, however, commercial CSPs provide non-negotiable, standard-form contracts that apply to all types of data. Both types of contracts are generally called 'Terms of Service' and cover terms and conditions, service level agreements, acceptable use policies and general security and privacy policies—the latter of which can often be quite lengthy and complex. Sometimes these additional documents may be folded into the Terms of Service, other times they are incorporated by reference.

Genomic cloud computing provides several benefits. First, it is, relatively speaking, low-cost in terms of allowing access to resources due to its 'elasticity'—an on-demand service wherein one pays for what one needs.[8] This shifts the need from purchasing many information technology resources in-house (ie, capital expenditure) to 'renting' such resources from third parties when needed (ie, operating expenditure). For genomic researchers, this tends to mean paying for computing time and transfer. Moreover, large CSPs are able to buy data transit in bulk to increase network connectivity and translational bandwidth, reducing internal bandwidth costs and passing on the savings to their customers. Second, cloud computing may afford greater data security, as large-scale cloud-based infrastructure typically have the capacity to invest in and implement state-of-the-art encryption, firewalls and auditing capabilities. Third, genomic cloud computing also offers increased data storage capacity and efficient processing, and 'scaled up' genomic analysis through increased computing power, which can accelerate discovery and innovation and avoid the 'researchers' bottleneck' where researchers are forced to size their work to the infrastructure their organization built. Finally, with efficiency and economies of scale, cloud computing services are becoming not only a cheaper solution but a much more environmentally friendly one to build and deploy IT services. Cloud computing is seen to deliver energy savings through external data storage and data bundling on powerful mainframe computers, though the growing demand of cloud infrastructure has drastically increased the energy consumption of data centers.

However, as recognized by governments and scholars alike, these benefits do not come without some concerns.[12–16] In addition to data security concerns, reliability can plague even the biggest names in cloud computing. In 2008, 2011 and 2012, for example, Amazon's data centers suffered multiple outages, including once due to heavy thunderstorms, bringing down a number of popular websites and services and preventing users from logging in to access their data.[17–20] One outage led a technology commentator to remark: 'The duration of the outage has surprised many, as Amazon has a lot of backup computing infrastructure. If Amazon can't safeguard the cloud, how can we rely on it?'[21] As well, closure of CSPs can cause concern about data control and migration. In 2013, for example, Nirvanix, a well-funded and established CSP, suddenly ended its cloud service and gave its customers only 2 weeks to save and migrate their data.[22] Similarly, in January 2012, to the surprise of many, Google discontinued Google Health and gave users a year within which to make alternative arrangements for their data.[23]

As genomic researchers rely on data contributed by patients and participants and are bound to abide by laws and ethical guidelines, it is critical to respect the privacy and autonomy of patients and participants by proactively assessing the full range of legal and ethical issues surrounding genomic cloud computing. This risk assessment is all the more important given that the Terms of Service of CSPs are generic documents that have not been developed with sensitive health or genomic data specifically in mind.

In this article, we discuss three key 'points to consider' (data control; data security, confidentiality and transfer; and accountability). These points to consider should be borne in mind by genomic research organizations when negotiating legal arrangements to store genomic data on a large commercial CSP's servers (see Box 1 for examples of large CSPs). Interspersed with these points to consider are recommendations for researchers and their organizations. Depending on the nature of the data, researchers must determine whether specialized secure private or hybrid clouds (which large CSPs may offer) are required to ensure sensitive data (eg, clinical data) remains within an organization and follows appropriate regulations, or whether commercial clouds may be used to analyze and distribute publicly available data,[24,25] perhaps subject to approval by a data access committee that authorizes scientific researchers and other end users. The points raised in this article are based on a preliminary review of several publicly available CSP Terms of Service. Critical reviews of privacy legislation and academic literature on cloud privacy were also undertaken. As there are other types of CSPs, and this is a quickly evolving area, other issues can arise in the course of genomic cloud computing. Genomic researchers are therefore encouraged to seek ethical and legal advice before concluding agreements with cloud providers.

POINTS TO CONSIDER

Data control

Cloud computing with commercial CSPs entails the outsourcing (or off-shoring) of data and services to third party providers. Any genomic or health-related data that used to be stored locally may thereafter be stored in the cloud, including in a 'cloud stack' format whereby multiple layers of services are provided by separate CSPs. Attributable to a multi-tenant environment (ie, where a single instance of software runs on a server and serves multiple client organizations), and possibly geographically dispersed data centers, genomic researchers place their computation and data on machines they cannot directly control. To a large extent, control over computation and data is thereby relinquished. Among the risks associated with cloud computing are unauthorized access (or reuses for which consent has not been obtained from researchers, patients or participants), data corruption, infrastructure failure or unavailability. In case something goes wrong, it can be difficult to discern who has caused the problem, and, in the absence of solid evidence, it is nearly impossible for the parties

Box 1 Examples of several large commercial cloud service providers (CSPs)

Amazon Web Services (http://aws.amazon.com/) and Elastic Compute Cloud (Amazon EC2) (http://aws.amazon.com/ec2/)
Google Cloud Platform (https://cloud.google.com/)
Microsoft Azure (http://www.azure.microsoft.com/)
IBM Cloud (http://www.ibm.com/cloud-computing)
HP Public Cloud (http://www.hpcloud.com/)
Citrix CloudPlatform (https://www.citrix.com/products/cloudplatform/overview.html)
Rackspace Cloud (http://www.rackspace.com/cloud/)

involved to hold each other responsible for the problem if a dispute arises.

Data control issues are manifest in several areas of a CSP's Terms of Service:

Amendments to terms of service. First, control issues arise in the ability for many commercial CSPs to amend the Terms of Service, often without explicit notification to customers. Often included in CSPs' imposed standard-form contracts is the possibility for the CSP to vary terms via unilateral notice. This privilege, which is commonly seen in IT and online contracts, obliges customers to check the CSP website for changes to the Terms of Service, even if the CSP does not mark the changes or indicate the date of last change or review. Some may provide notice only of 'material' changes to the Terms of Service. Further, CSPs will consider continued use of their cloud service as deemed acceptance of the new Terms of Service in the absence of any explicit objection to the amended terms or cessation of use.

Recommendation

- *Researchers should ensure proper notification of amendments to Terms of Service with a reasonable period of time for response and acceptance.*

Data preservation and deletion. Second, control issues are encountered in Terms of Service sections regarding data preservation (what happens to data when the contractual relationship with the CSP ends) and deletion (the removal of data from the cloud). Even though control is potentially problematic, researchers should, to the greatest extent possible, put themselves in a position to control what data are moved to the cloud, as well as to control what data remain in the cloud.

Recommendation

- *Researchers should ensure that they can retrieve genomic data and that a CSP cannot retain it or use it after the contract ends, subject to legal or regulatory requirements and/or authorization of the researchers.*

Regarding preservation, some CSPs will preserve data for a grace period following the end of the agreement (eg, 30 days). Other CSPs may stipulate that they will delete the data immediately upon the end of the agreement, no matter the circumstances, or simply may not discuss at all what will happen, providing neither a grace period nor an undertaking to delete the data.

Recommendations

- *Researchers should ask for clarification of the CSP's data preservation policy regarding how long the grace period lasts (if applicable), commitments to comprehensively delete the data and any costs that may be involved.*
- *Researchers should ensure that data exit and migration strategies are well-planned before importing data into a cloud.*

The Nirvanix experience demonstrates that it is much easier to import data than to recover it or move it to another CSP. Indeed, a cloud exit strategy is as valuable as a cloud deployment strategy.

Regarding deletion, researchers should consider what happens to their genomic data after the relationship with a CSP comes to an end: what is their exit strategy and end-of-contract transition? Researchers should also consider whether they can retrieve their data with relative ease to move it elsewhere. Upon termination of the cloud computing services, does the CSP ensure that the genomic data will be deleted comprehensively (ie, including duplicates or backups) from its servers (and any sub-processors' servers) upon the researchers' retrieval of it, and is there evidence provided of permanent deletion? If this cannot be ensured, a CSP may be seen to assume responsibility, under data protection laws of certain jurisdictions, for the security of any non-deleted personal data. It should be noted that under the proposed European General Data Protection Regulation, a new right to transmit personal data in the same format could require CSPs to allow customers to move their data to competitors offering similar products or services, though upholding this right if the CSP is located outside the EU could be challenging.[26]

Data monitoring. Third, control issues arise in Terms of Service sections pertaining to data monitoring. Can the CSP monitor hosted genomic data, and if so, what form should the monitoring take and what conditions should apply? Even though most commercial CSPs encrypt data while in transit and at rest, researchers should still verify that the data are encrypted (and find out how they are encrypted). Additionally, if it is researchers that encrypt the data, they should query whether they want the CSP to have access to decryption keys. Although monitoring of traffic data or bandwidth consumption may be acceptable, researchers could be concerned with a CSP monitoring personal data or aggregate genomic data uploaded to the cloud, even if such monitoring is to ensure compliance with an accepted use policy.

Recommendations

- *If possible, researchers should endeavor to have the CSP agree to treat any genomic or health-related data obtained from monitoring or support or maintenance activities as subject to confidentiality provisions or to restrict the purposes for which CSPs can monitor data.*
- *Researchers should ensure that their own organization has data encryption capabilities and good management infrastructure for control over data stored on a cloud.*

At the same time, researchers should be aware of their own policies for giving authorization and access privileges to staff to provide login details to CSP employees for certain situations (eg, support).

Data security, confidentiality and transfer

On a structural level, there is a contrast between the nature of cloud computing, built on the idea of 'locationlessness' (or at least disparate localization), and data privacy laws, which are still based on geographic borders and location-specific data processing systems. As cloud computing is largely built on the idea of seamless, borderless sharing and storage of data, it can run into tension with different national jurisdictions governing citizens' rights over privacy and protection of personal data. Indeed, as cloud computing enables personal (health) data to be transferred across national, regional and/or provincial borders, where little consensus exists about which authorities have jurisdiction over the data, cloud clients and providers will each need to understand and comply with the different rules in place—to the extent such rules exist. In an environment where data exchange by researchers is no longer a point-to-point transaction within one country but instead is characterized by transnational, dynamic and decentralized flow, the legal distinction between national and international data use may become less meaningful than in the past.

At the same time, the global nature of genomic cloud computing means that it is difficult to know which laws apply, let alone how to ensure compliance with the applicable laws. For example, while national regulatory frameworks such as Clinical Laboratory

Improvement Amendments[27] or the Health Insurance Portability and Accountability Act of 1996 (HIPAA)[28] establish guidelines for clinical data storage and sharing in the United States, there remains no international legal standards on the use and storage of clinical data. In addition, the nascent stage of cloud computing, particularly in the genomics context, leads to uncertainty about how existing laws, especially privacy and data protection laws, will be applied. Though it is now well established that data sets such as genome sequences may uniquely identify an individual,[29] there has not yet been an attempt to reach community consensus or proposed guidelines on safeguarding privacy in cloud computing. Indeed, there is currently an absence of well-defined cloud computing-attuned standards, guidelines or model contractual agreements to inform the practice of genomic researchers considering the use of cloud computing as a solution for their project's data. This said, researchers probably will be more interested in the applicable law for the CSP who is providing the services, rather than where data are stored, as CSPs must comply with and be subject to the laws of their headquartered jurisdiction.

Data security and confidentiality. One of the greatest concerns about storing genomic data in the cloud is whether the data are secure. Researchers may fear that storing data on the cloud will lead to potential unauthorized access to patient data and liability and reputation damage that could result from a mandatory breach notification, such as that stipulated in HIPAA. Even though genomic data stripped of identifiers (including names, addresses, birthdates and the like) may not constitute 'personal health information' for HIPAA or other similar health information privacy law purposes, recent literature suggests that this could well change.[30] Consequently, researchers have reason to seriously consider the security issues of genomic cloud computing and the role of privacy laws.

Such issues arise in Terms of Service sections addressing data security and confidentiality, along with CSP privacy policies, and data location and transfer. Depending on the sensitivity of the data, researchers may want to establish data access committees that oversee the terms of access to cloud-stored data. Similarly, US-based researchers might want CSPs to hold 'trusted partner' status before storing genomic or clinical data in the cloud, or have them sign a HIPAA 'business associate agreement' (BAA). Many commercial CSPs are now able to provide a BAA, which describes what a CSP can and cannot do with 'personal health information', including a prohibition on further disclosing the data to another entity other than those permitted or required by the contract or by law. However, applying such extensive national requirements would not be conducive to the type of global data exchange needed for the development of a healthy, productive genomic research sector. The desire of participants and patients to encourage beneficial research that could eventually lead to the development of a cure for serious afflictions should not be neglected in order to achieve an 'ideal' level of privacy protection.

Recommendations

- *Researchers should verify the data elements to be stored in the cloud, including whether the data constitute sensitive personal data or personal health information. Genomic data should be secured in a way that protects the privacy of everyone whose data are analyzed.*
- *Researchers should consider restricting access to cloud-stored genomic data (individual-level) to bona fide researchers approved by data access committees.*

Often data security and confidentiality questions may be determined by consulting privacy or data protection laws, as well as internal organizational policies and funding requirements. Data protection laws may provide only broad principle-based guidance, requiring the adoption of safeguards that are commensurate with the sensitivity of the data—and data protection laws across the globe frequently treat genomic data as sensitive, thereby requiring strict safeguards.

Many CSPs offer to make 'best efforts' or to take 'reasonable and appropriate measures' to secure data against accidental or unlawful loss, access or disclosure, but this is distinct from a legal representation that the service will be uninterrupted or error free or that data will be fully secure or not otherwise lost or damaged. Indeed, few commercial CSPs will make this latter type of comprehensive representation. At the same time, CSPs themselves must be cognizant of strict privacy and data protection laws in jurisdictions where data may be processed, especially in Europe. For example, section 9 of Germany's Federal Data Protection Act mandates the need for 'necessary technical and organizational measures' to be put in place for a cloud service, and section 9(a) grants audit rights to cloud customers to examine a CSP's 'data protection strategy and their technical facilities.'[31] Commercial CSPs should be in a position to implement security and compliance features that enable compliance with relevant regulations and international guidelines, be it HIPAA,[28] Good Clinical Practice,[32] European data privacy laws or dbGaP Best Practices Requirements.[33]

Researchers should carefully examine a CSP's approach to securing and protecting data, with an understanding that CSPs may not want to fully disclose their security practices, lest doing so compromise their cloud's security itself. Many commercial CSPs offer security 'white papers' that researchers can consult to review the security controls. Important questions include: does data security appear to be a priority? What are the physical, administrative and digital security measures? Is the CSP willing to show specific documentation (eg, ISO27001 or ISO27002 policies and procedures)? How well are web-based applications protected? Are there API access restrictions? Are there multi-factor authentication, automatic session timeouts and logging functions for auditability? Will the CSP provide prompt notification of service interruptions or a potential data compromise? Are cloud customers indemnified for unscheduled downtime? Does the CSP make data backups or are researchers solely responsible for doing so? (Back-ups may be necessary to fulfill the obligations of confidentiality owed to patients and participants.) Do the CSP's security measures accord with the researchers' own security policies or practices regarding the secure storage of genomic data? Given the sensitivity of genomic data, researchers should consider whether a CSP agrees to report any data losses or security breach incidents within a short period of time (eg, <24 h).

Recommendations

- *Researchers should ensure that CSPs have been independently audited against comprehensive and internationally recognized and respected information security standards, such as those promulgated by the International Organization for Standardization (ISO) and Statement on Standards for Attestation Engagements (SSAE).*
- *Researchers should ensure that the third party audit certifications are current and maintained throughout the duration of the cloud service.*

Researchers should consider how cloud computing can impact data protection and confidentiality, both within the laws of jurisdictions of a CSP and also within the ethical and legal requirements of the research organization and/or funders.

Recommendation

- *If a CSP is unwilling to commit to a general obligation to comply with all applicable data protection and confidentiality laws relating to genomic data, researchers should alternatively attempt to secure a confidentiality or non-disclosure agreement or a commitment by the CSP to be contractually bound to the ethical and legal requirements applicable to the researchers' institution.*

This said, general commitments to laws may be the best researchers can hope for at this time, along with strong adherence to security measures such as data encryption, both when at rest and during transfer.

Data location and transfer. Data location and transfer is a critical issue for researchers, given its intersections with data protection laws, ethical guidelines and consent forms that may (or may not) address data storage and sharing. Researchers should be aware that most commercial CSPs process, store or temporarily move data to any country where the CSP or its agents maintain facilities. This is done for service-related reasons, namely, for security, back-ups, support and cost efficiency. Intra-cloud data transfer (that is, transfer within the same CSP) is permissible under most data protection laws, provided there is compliance with the relevant data export provisions, such as model contractual clauses or Binding Corporate Rules. However, many commercial CSPs may not always tell their customers where the data are going at any given moment. This can cause problems in regions such as Europe, where data protection laws prohibit the transfer of data to third countries without an 'adequate level of protection' for the data.

Recommendation

- *If there are transfers of personal data from one jurisdiction to another, researchers should verify that a CSP is in compliance with data export laws and regulations and that the CSP has adequate oversight mechanisms to monitor ongoing compliance with these laws and regulations.*

At the same time, large commercial CSPs increasingly assure customers with sensitive data that their data will remain in particular countries or regional zones of choice, even for remote access. Researchers concerned about the location of genomic and health-related data storage should be mindful of the specific locations where data are stored on CSP servers, and it may be that researchers want their data stored only in locations providing an equivalent or greater level of protection for genomic and health-related data to that where the data originated. Even when the data are kept in a specifically determined country, special attention should be paid to provision allowing temporary or emergency transfer of data to additional undisclosed locations.

Likewise, researchers should be mindful of the long-arm reach of certain laws, such as the US Patriot Act,[34] which allows the US government to access any data within US territory or within a US-based company (even when there is a subsidiary or operations outside the US) without giving notice and without informing the person concerned of the accessed information. A recent US court case found that US Internet service providers (eg, Microsoft or Google and other companies offering cloud services via the Internet) with EU subsidiaries are required to comply with warrants and subpoenas from US law enforcement agencies relating to data held in the EU.[35] Researchers should be mindful therefore of CSPs in jurisdictions that permit wide surveillance and law enforcement access to data; they should look to structure services, to the extent possible, which can protect data of patients or participants from access requests by law enforcement agencies.

Recommendation

- *Depending on legal/regulatory requirements and the sensitivity of the data, to assuage national security law concerns, researchers should consider requesting CSPs to store data and applications only in designated regions.*

Although not always made public, researchers should attempt to determine in which jurisdictions the CSP maintains servers and which agents (ie, sub-contractors) can access their data. They should investigate as much as possible the trail of data storage and transfer and the ability to have tools that can verify the locations of data storage or transfer. It is important that a CSP will transfer personal and genomic data only in an encrypted manner over secure networks and that the CSP will continue to be responsible for managing its agents' compliance with data security and confidentiality.

Accountability

Finally, researchers should be mindful of what may happen in the event that something goes wrong. What happens when the cloud fails? With more services being built on top of cloud computing infrastructures, a power outage, closure, bankruptcy or breakage/failure can create a domino effect, effectively taking down large amounts of Internet services and Internet-based applications. In cases of failure, what forms of arbitration exist for stakeholders, and what is the responsibility of CSPs?

Liability. Accountability issues appear in the standard clauses in contracts addressing liability. Researchers should be mindful of the breadth of a CSP's waiver of liability. CSPs who have Terms of Service governed by laws of US states rather than European countries may waive all liability for any unauthorized access or use, corruption, deletion, destruction or loss of any data maintained or transmitted though its servers, regardless of who is at fault. Thus any damage caused to a researcher's data, such as losses arising from security breaches, data breach or loss, denial of service, performance failures, inaccuracy of the service, and so on, even if attributed to the CSP or its agents, may excluded from any liability.

Recommendations

- *Researchers should determine the chain of responsibility for preserving the confidentiality and integrity of genomic data.*
- *Researchers should request full indemnity for liability related to privacy and security.*

Given the potential concerns about misuse or damage to genomic data, researchers should be aware that direct liability is typically excluded by CSPs. Researchers should also be aware of the importance of negotiating for a clause that imposes liability on the CSP for, at a minimum, wilful or gross negligence with respect to defined types of breach or loss, such as breach of confidentiality, privacy or data protection laws, data loss/corruption or breach of regulatory or security requirements that could give rise to regulatory sanctions.

CONCLUSION

Genomic cloud computing is an emerging technology platform for the biomedical research community. Many cloud computing issues remain unsettled. The 'points to consider' in this article can provide a useful starting point for researchers to consider when negotiating legal arrangements to store genomic data in the cloud. Just as financial institutions and insurers increasingly use the cloud to manage their data and services (and hold our information in an encrypted state),

migrating genomic and other health-related data to the cloud is part of the evolution of existing technologies and large-scale genome science. But it comes with its own challenges. All stakeholders should come to the table and work together toward common solutions that enable trust. Collaboration between large biobanks and genomic research consortia on these legal and ethical issues to present a common position to commercial CSPs will be a major determinant of success.

In any environment, cloud computing or otherwise, there is no such thing as zero risk. Dedicating more resources to secure genomic and health-related data will be critical for researchers harnessing the cloud. But so too will it be critical to have an open discussion with CSPs of the risks and benefits of cloud computing—the latter of which is equally important. Transparent practices will be critical to build trust among participants (government, organizations and individuals) for genomic cloud computing. Transparency prevents abuse by organizations and encourages participants to share their data. The broad access, computing power and speed of cloud computing impels organizations responsible for sensitive data to institute and maintain clearly defined policies on transparency.

We emphasize the need for multi-stakeholder involvement, because introducing genomics into clinical practice through cloud computing 'is not only a linear function of computing capacity but it (also) requires the input from diverse disciplines'.[36] CSPs should consider the specific challenges of genomic and health-related data. Researchers should consider scrutiny of cloud contracts as only one part of the risk assessment process of data migration to the cloud. They should also assess which functions should be migrated to the cloud and how, as well as which internal controls to develop, how data will be encrypted and backed up internally, and, after contractual agreement with a CSP, how monitoring will be conducted. Ultimately, diligent genomic cloud computing means leveraging security standards and evaluation processes as a means to protect data, whether local or remote. More importantly, diligent cloud computing entails many of the same good practices that researchers should always consider in securing their local infrastructure. In itself, there is nothing magical about an outsourced relationship with a CSP.

At the same time, broader and non-technical questions remain. Scholars of ethical, legal and social issues (ELSI) in genomics may need to consider whether consent practices should be refined to specifically address the possibility that genomic and health-related data will be processed in the cloud and any ensuing ethical and legal implications. Further research will be needed to uncover the legal and ethical issues associated with other types of cloud computing. Such research would be best informed if it is evidence-based and seeks the views of various stakeholders. This will be critical to better develop cloud computing policy and 'best practices' for the genomics community. That CSPs and genomic and ELSI researchers consider these issues is not only evidence of due diligence but a sign of ethical conduct and respect for patients and participants whose data are being used to advance biomedical science.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The authors are grateful for the valuable comments provided by Saminda Pathmasiri. P3G acknowledges the funding support of Genome Quebec. ICGC-EPC acknowledges the funding support of the Ontario Institute for Cancer Research (OICR).

P3G International Steering Committee: Anthony J Brookes; Paul Burton; Rex Chisholm; Isabel Fortier; Pat Goodwin; Jennifer Harris; Kristian Hveem; Jane Kaye; Alistair Kent; Bartha Maria Knoppers; Klaus Lindpaintner; Julian Little; Peter Riegman; Samuli Ripatti; and Ronald Stolk. ICGC Ethics and Policy Committee: Martin Bobrow; Anne Cambon-Thomsen; Lynn Dressler; Yann Joly; Kazuto Kato; Bartha Maria Knoppers; Laura Lyman Rodriguez; Treasa McPherson; Pilar Nicolás; Francis Ouellette; Carlos Romeo-Casabona; Rajiv Sarin; Susan Wallace; Georgia Wiesner; Julia Wilson; and Nikolajs Zeps. Howard Simkevitz and Assunta De Rienzo are observers.

1 International Cancer Genome Consortium (ICGC): Available at http://www.icgc.org. Accessed 4 August 2014.
2 Nietfeld JJ, Sugarman J, Litton JE: The Bio-PIN: a concept to improve biobanking. *Nat Rev Cancer* 2011; **11**: 303–308.
3 Murtagh MJ, Demir I, Jenkings KN *et al*: Securing the data economy: translating privacy and enacting security in the development of DataSHIELD. *Public Health Genomics* 2012; **15**: 243–253.
4 Wallace SE, Gaye A, Shoush O, Burton PR: Protecting personal data in epidemiological research: DataSHIELD and UK law. *Public Health Genomics* 2014; **17**: 149–157.
5 Friend SH, Norman TC: Metcalfe's law and the biology information commons. *Nat Biotechnol* 2013; **31**: 297–303.
6 Marx V: Genomics in the clouds. *Nat Methods* 2013; **10**: 941–945.
7 Stein LD: The case for cloud computing in genome informatics. *Genome Biol* 2010; **11**: 207.
8 Afgan E, Baker D, Coraor N *et al*: Harnessing cloud computing with Galaxy Cloud. *Nat Biotechnol* 2011; **29**: 972–974.
9 Heath AP, Greenway M, Powell R *et al*: Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. *J Am Med Inform Assoc*; e-pub ahead of print 24 January 2014; doi:10.1136/amiajnl-2013-002155.
10 Reid JG, Carroll A, Veeraraghavan N *et al*: Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics* 2014; **15**: 30.
11 Seddon JJM, Currie WL: Cloud computing and trans-border health data: unpacking U. S. and EU healthcare regulation and compliance. *Health Pol Technol* 2013; **2**: 229–241.
12 Schwartz PM: Information privacy in the cloud. *Univ PA Law Rev* 2013; **161**: 1623–1662.
13 Stylianou KK: An evolutionary study of cloud computing services privacy terms. *J Comput Inf Law* 2010; **27**: 593–612.
14 Article 29 Data Protection Working Party: Opinion 05/2012 on Cloud Computing. Adopted 1 July 2012. Available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2012/wp196_en.pdf. Accessed 4 August 2014.
15 European Commission: Unleashing the potential of cloud computing: COM (2012) 529 final. Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0529:FIN:EN:PDF. Accessed 4 August 2014.
16 European Parliament: Resolution of 10 December 2013 on unleashing the potential of cloud computing in Europe (2013/2063(INI)). 10 December 2013. Available at http://www.europarl.europa.eu/sides/getDoc.do?type=TA&language=EN&reference=P7-TA-2013-0535. Accessed 4 August 2014.
17 Gunderloy M: S3 outage: the aftermath. *GigaOM*. 21 July 2008. Available at http://gigaom.com/2008/07/21/s3-outage-aftermath/. Accessed 4 August 2014.
18 Harris D: Cloud platforms Heroku, DotCloud & Engine Yard hit hard by Amazon outage. *GigaOM*. 21 April 2011. Available at http://gigaom.com/2011/04/21/more-than-100-sites-went-down-with-ec2-including-your-paas-provider/. Accessed 4 August 2014.
19 Bilton N: Amazon Web Services knocked offline by storms. *New York Times.com*. 30 June 2012. Available at http://bits.blogs.nytimes.com/2012/06/30/amazon-web-services-knocked-offline-by-storms/. Accessed 4 August 2014.
20 Perlroth N: Amazon cloud service goes down and takes popular sites with it. *New York Times.com*. 22 October 2012. Available at http://bits.blogs.nytimes.com/2012/10/22/amazon-cloud-service-goes-down-and-takes-some-popular-web-sites-with-it/. Accessed 4 August 2014.
21 Takahashi D: Amazon's outage in third day: debate over cloud computing's future begins. VB News. 23 April 2011. Available at http://venturebeat.com/2011/04/23/amazons-outage-in-third-day-debate-over-cloud-computings-future-begins/. Accessed 4 August 2014.
22 Coughlin T: Nirvanix provides cautionary tale for cloud storage. Forbes. 30 September 2013. Available at http://www.forbes.com/sites/tomcoughlin/2013/09/30/nirvanix-provides-cautionary-tail-for-cloud-storage/. Accessed 4 August 2014.
23 Brown A, Weihl B: An update on Google Health and Google PowerMeter. Google Blog. 24 June 2011. Available at http://googleblog.blogspot.ca/2011/06/update-on-google-health-and-google.html. Accessed 4 August 2014.
24 Grossman RL, White KP: A vision for a biomedical cloud. *J Intern Med* 2012; **271**: 122–130.
25 Yang YT, Borg K: Regulatory privacy protection for biomedical cloud computing. *Beijing Law Rev* 2012; **3**: 145–151.
26 Gunasekara G: Paddling in unison or just paddling? International trends in reforming information privacy law. *Int J Law Inf Technol* 2014; **22**: 141–177.

27 Clinical Laboratory Improvement Amendments (CLIA) of 1988: Public Law 100–578, codified at U.S. Code 42, §263a.
28 Centers for Medicare & Medicaid Services: *The Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. Available at http://www.cms.hhs.gov/hipaa/default.asp. Accessed 13 May 2014.
29 Erlich Y, Narayanan A: Routes for breaching and protecting genetic privacy. *Nat Rev Genet* 2014; **15**: 409–421.
30 Rodriguez LL, Brooks LD, Greenberg JH, Green ED: The complexities of genomic identifiability. *Science* 2013; **339**: 275–276.
31 *Federal Data Protection Act* (BDSG) (Germany): Federal Law Gazette I, p. 66 (2003). Available at http://www.bfdi.bund.de/EN/DataProtectionActs/Artikel/BDSG_idFv01092009.pdf?__blob=publicationFile. Accessed 4 August 2014.
32 ICH: *Good Clinical Practice E6*. 1996. Available at http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guideline.pdf. Accessed 4 August 2014.
33 National Institutes of Health: *dbGaP Best Practices Requirements*. Available at http://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf. Accessed 4 August 2014.
34 *USA PATRIOT Act*: Public Law 107–56, 115 Stat. 272 2001.
35 In the Matter of a Warrant: (2014) United States District Court – Southern District of New York, 13 Mag. 2814.
36 Prainsack B, Schicktanz S, Werner-Felmayer G: Geneticising life: a collective endeavor and its challenges; In: Prainsack B, Schicktanz S, Werner-Felmayer G (eds.): *Genetics as Social Practice: Transdisciplinary Views of Science and Culture*. Surrey, UK: Ashgate, 2014.