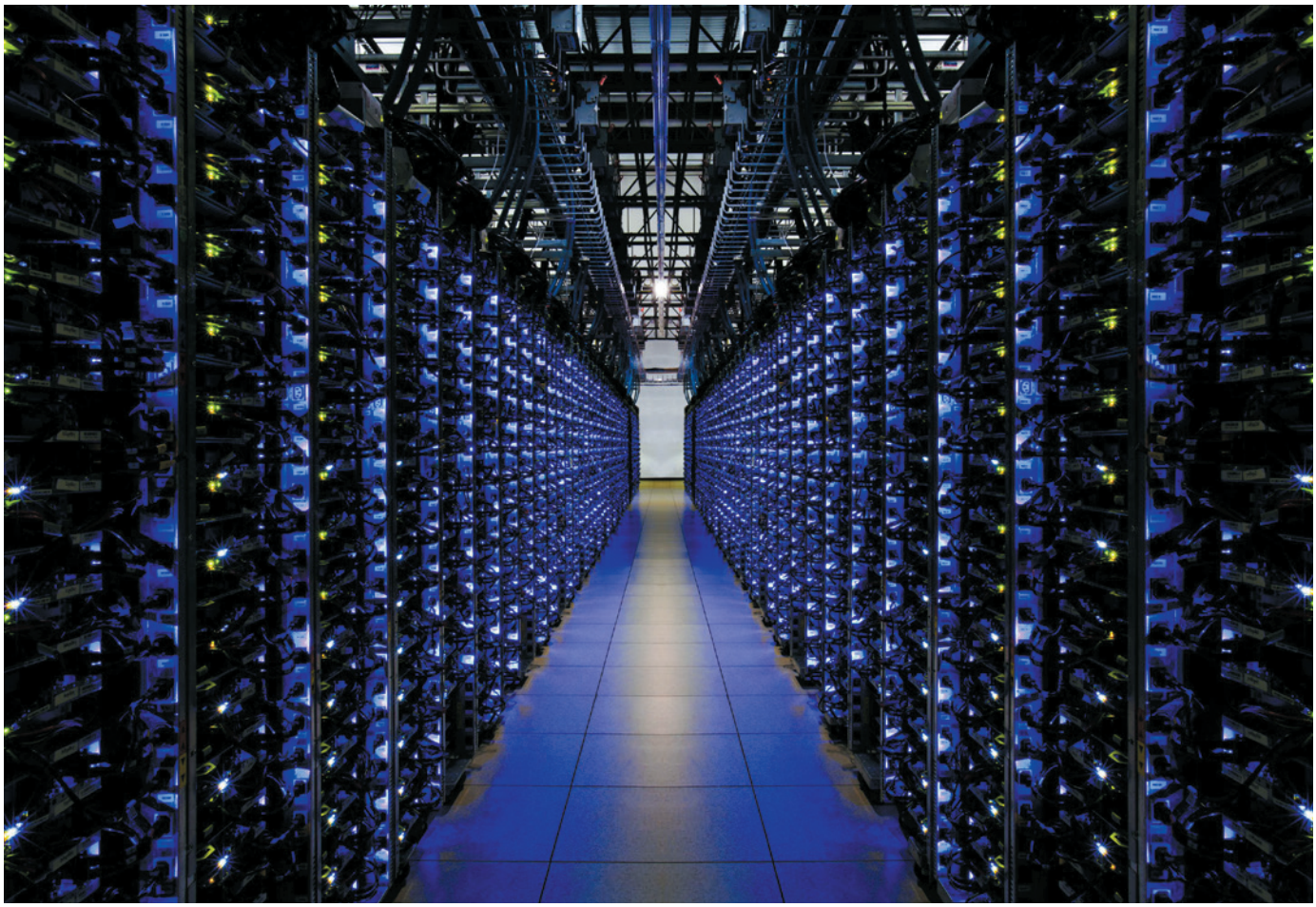
COMMENT

KEYSTONEUSA-ZUMA/REX



Google's cloud services are among those increasingly being used by researchers who want to analyse large genomics data sets.

Create a cloud commons

Major funding agencies should ensure that large biological data sets are stored in cloud services to enable easy access and fast analysis, say **Lincoln D. Stein** and colleagues.

There was a collective cheer in the human genomics community earlier this year, as researchers — ever more stymied by the challenges of accessing vast data sets — saw a major roadblock disappear. In March, the US National Institutes of Health (NIH) lifted its 2007 restriction on the use of cloud computing to store and analyse the tens of thousands of genomes and other genetic information held in its repository, the database of Genotypes and Phenotypes (dbGaP)[1].

Cloud services offer customers large amounts of storage and computing power on a pay-as-you-go basis. Because these services are available through the Internet, and multiple users share hardware, numerous funding agencies have been concerned that their use in genomics could threaten the privacy of people who supply samples[2].

The NIH turnaround is part of a growing suite of efforts aimed at addressing the fact that in the human genomics research community, the challenges of accessing big data sets are now blocking scientists' ability to do research, and especially to replicate and build on previous work (see go.nature.com/h9jgs1).

To take full advantage of the possibilities that cloud computing offers, we ▶

urge the NIH and other agencies to pay for the storage of major genomic data sets in the most popular cloud services. This way, instead of thousands of researchers wasting time and money by independently transferring data from a repository to the cloud of their choice, authorized scientists would be able to tap easily and cheaply into a global commons as and when they need to.

BIG DATA

Thanks to improvements in sequencing technology, the volume of genomic data submitted to public archives is now well into the multi-petabyte range (1 petabyte is 10^{15} bytes). In the International Cancer Genome Consortium (ICGC)[3], for instance, groups from 17 countries have amassed a data set in excess of two petabytes — roughly 500,000 DVDs-worth — in just five years.

Using a typical university Internet connection, it would take more than 15 months to transfer a data set this size from its repository into a researcher's local network of connected computers. And the hardware needed to store, let alone process the data, would cost around US$1 million.

Cloud services provide 'elasticity', meaning that a researcher can use as many computers as needed to complete an analysis quickly, and pay for only the computing time used. Several researchers can work in parallel, sharing their data and methods with ease by performing their analyses within cloud-based virtual computers that they control from their desktops. Thus the analysis of a big genome data set that might have previously taken months can be executed in days in or weeks (see 'Express lane').

These days, cloud services are just as secure as most academic data centres, often more so. They are now offered by major commercial companies including Amazon, Google and Microsoft, as well as smaller companies focused on genomics research, such as California-based Annai Systems, and several academic institutions such as the European Bioinformatics Institute in Hinxton, UK. These providers use strong encryption for data, have systems, such as firewalls and keychain fobs, for controlling who has access to the data, and provide tools to the owners of the data that allow them to monitor use closely.

A few major funders of human-genomics research are being cautious — for instance, some European funding agencies recommend that researchers keep genomic data within the agencies' jurisdiction to comply with European law on privacy[4]. But the cheapness, flexibility, reliability and security of cloud computing is such that we anticipate a wholesale shift to cloud services over the coming months (see 'Reaching for the cloud'). And we welcome the NIH's decision in hastening this transition.

"The human genomics community could pave the way for other researchers grappling with data overload."

Now is the time to establish mechanisms and practices that maximize the efficiency and usability of cloud computing while minimizing costs.

ACCESS CONTROL

To gain access to much of the human genomic and other data held in central repositories such as the dbGaP or its counterpart, the European Genome-phenome Archive (EGA), a researcher must obtain approval from a data-access committee, or DAC. Currently, if two independent research groups wish to work on the same data set in a private or commercial cloud, they will each need to get approval from the relevant DAC, copy the data across the Internet and store it in their cloud of choice.

Both groups have to wait while the data are copied, and each has to pay for storage while the data are being copied and for as long as they need the data. As hundreds of groups start to do the same thing, this process could collectively waste years of researchers' time and tens of millions of taxpayer dollars. Even with unfettered access to cloud services, it is currently impractical for most groups to work with the largest public genomic data sets because of the time and costs involved in transferring the data from its repository into a cloud.

A better approach would be for the relevant funding agencies to request that every major genomic data set be uploaded into the most popular academic and commercial clouds available, and to pay for the long-term storage of the data in the clouds. This way, the data would need to be copied only once and researchers would have to pay only for the temporary storage they use while their analysis is in progress.

Currently, several commercial providers of cloud services are offering to store research data sets for free or at heavily subsidized rates to prompt more researchers to use their services. Amazon Web Services, for example, levies no charge for hosting the sequences released by the 1,000 Genomes Project (now totalling more than 200 terabytes of data), an international effort to catalogue human genetic variation. And Annai Systems hosts a growing subset of the ICGC data set.

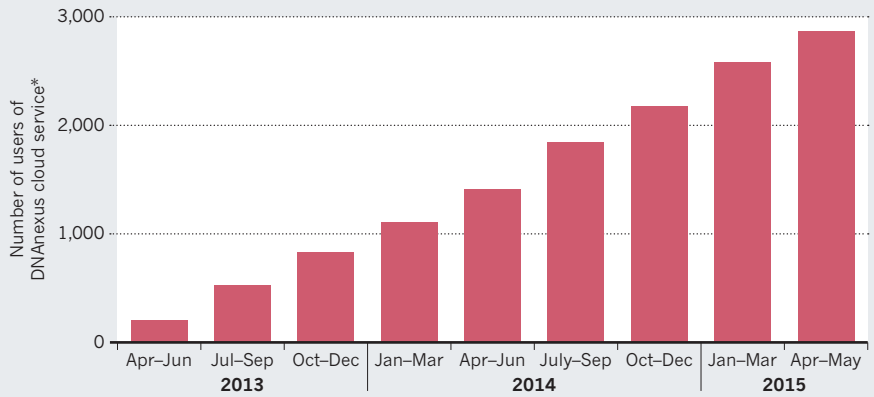
We envisage that entities such as the dbGaP or the EGA[5] would continue to be the primary custodians of the data and that their DACs would still review and authorize data use within the cloud. In this way, genomic cloud computing could even give rise to a micro-economy. For instance, a genome biologist who contributes a valuable data set to a cloud could receive credits for processing time. Similarly, a computer scientist who contributes a software package that enables other geneticists to find cancer variants more efficiently, say, could receive credits every time someone runs their package.

Over time, a virtuous cycle would result. Being able to merge large data sets would enable researchers to link rare genetic variations to diseases, and such successes would encourage others to deposit more data sets and the development of yet more powerful software. Such mechanisms could work in conjunction with requests from funding agencies that certain data sets be deposited in certain clouds.

One possible risk is that, by rising to dominance, a single provider of cloud services could come to control pricing,

REACHING FOR THE CLOUD

Internet cloud services, which provide large amounts of data storage and computing power, are becoming increasingly popular with geneticists grappling with vast data sets.



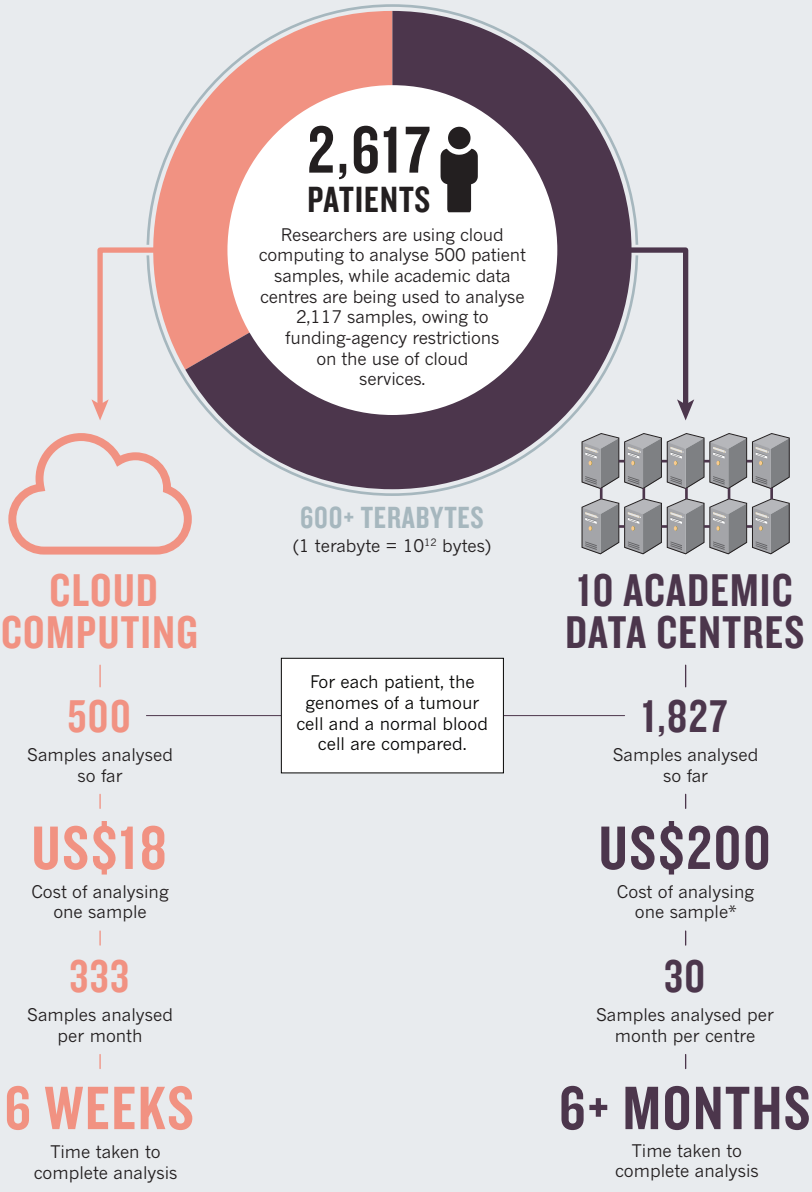
*Data from DNAnexus, a cloud-based genome informatics and data-management platform.

SOURCE: DNANEXUS

SOURCE: PCAWG, HTTP://PCAWG.ICGC.ORG

EXPRESS LANE

The Pan Cancer Analysis of Whole Genomes project (in which L.D.S., P.C., G.G. and J.O.K. are involved), an effort to investigate the role of non-coding parts of the genome in cancer, demonstrates how much faster and cheaper it is to use cloud computing than to use conventional academic data centres when analysing vast biological data sets.



and so subtly influence how the science is performed. To prevent this from happening, funding agencies should fund the deposition of the same important data sets in multiple clouds. This would also help to address jurisdictional sticking points. Genomic data originating in Europe, for instance, could be confined to clouds based in Europe.

GENOMIC STANDARDS

Achieving this vision will require work, technical and legal. For example, currently there is no way for a cystic-fibrosis researcher, say, to write software to search the dbGaP database and find the sequences obtained from people with the disease. Systematically tagging the data — specifying the tissue source of the sample, for instance — would help to address this. Since 2001, journal publishers have agreed to accept only RNA microarray studies in which researchers describe their data using the 'minimum information about a microarray experiment (MIAME)' standard[6]. A similar standard is needed for genomic data.

Reliable protocols for authorizing access to sensitive data in the cloud, as well as mechanisms to enable and revoke access will also be needed. Individual project DACs should continue to be gatekeepers in the short term, but ultimately a few 'general-purpose' DACs may be better placed to oversee access to the clouds than the multitude of DACs currently operating.

On the legal side, rules of the road must be established to clarify the roles and responsibilities of the funding agencies, the data custodians, the cloud service providers and the researchers who use cloud-based genomic data. If someone posted an ICGC genome on Facebook, for instance, who among these various players should be held accountable? Fortunately, for the past two years, an international coalition, the Global Alliance for Genomics and Health (genomicsandhealth.org) has prepared a Framework for Responsible Sharing of Genomic and Health-Related Data.

Meanwhile, the US National Cancer Institute has several pilot projects[3] exploring the practicalities of sharing and analysing genomic data on clouds. And the NIH and other funding agencies are already discussing a variety of 'biomedical commons' concepts, which incorporate several of the ideas proposed here.

By taking the right approach to cloud computing, the human genomics community could pave the way for researchers in many other fields, from neuroscience to epidemiology, who are similarly grappling with data overload. ■

Lincoln D. Stein *is director of informatics and biocomputing at the Ontario Institute of Cancer Research, Toronto, Canada.* **Bartha M. Knoppers** *is director of the Centre of Genomics and Policy, McGill University, Montreal, Canada.* **Peter Campbell** *is head of cancer genomics at the Wellcome Trust Sanger Institute, Hinxton, UK.* **Gad Getz** *is director of the Cancer Genome Computational Analysis group at the Broad Institute of MIT and Harvard, Cambridge, Massachusetts, and is the director of bioinformatics in the Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA.* **Jan O. Korbel** *is a group leader in genome biology at the European Molecular Biology Laboratory, Heidelberg, Germany.*
e-mail: lincoln.stein@gmail.com

1. Tryka, K. A. *et al. Nucleic Acid. Res.* **42,** D975– D979 (2014).
2. Gutmann, A. & Wagner, J. W. *Hastings Cent. Rep.* **43,** 15–18 (2013).
3. The International Cancer Genome Consortium. *Nature* **464,** 993–998 (2010).
4. European Commission. *Eleventh Annual Report of the Article 29 Working Party on Data Protection* (European Commission, 2008).
5. Leinonen R. *et al. Nucleic Acid. Res.* **39,** D28–D31 (2011).
6. Brazma, A. *et al. Nature Genet.* **29,** 365–371 (2001).