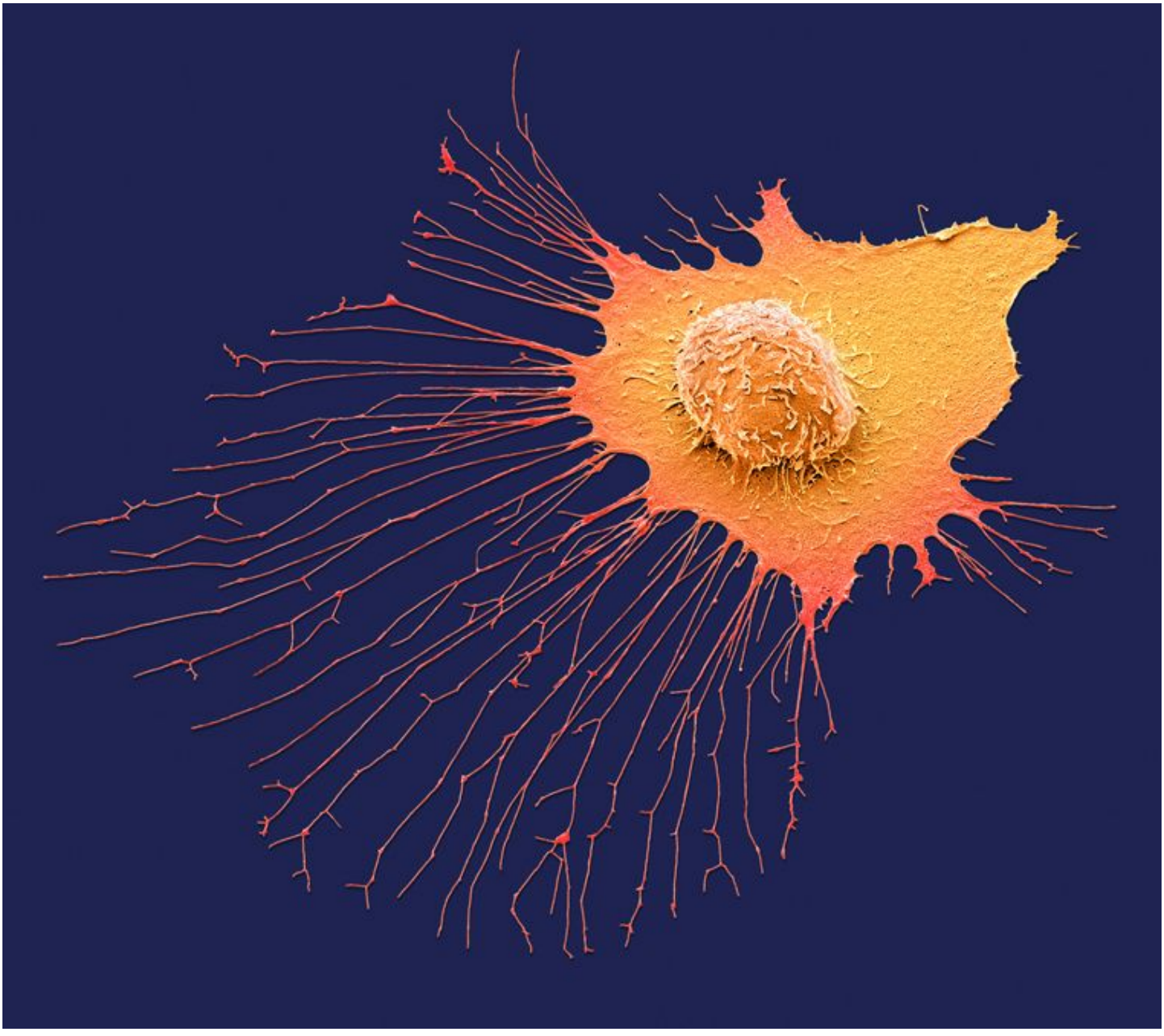
COMMENT · 05 FEBRUARY 2020

Genomics: data sharing needs an international code of conduct

Efforts to protect people's privacy in a massive international cancer project offer lessons for data sharing.

Mark Phillips , Fruzsina Molnár-Gábor , Jan O. Korbel, Adrian Thorogood, Yann Joly, Don Chalmers, David Townend & Bartha M. Knoppers



A migrating breast cancer cell. Credit: Steve Gschmeissner/SPL

More than 800 terabytes of genomic data are available to investigators all over the world, thanks to a major international project to identify the genetic traits associated with various types of cancer. Researchers involved have just published six papers in *Nature*. (Another 16 papers have been published elsewhere.)

All eight of us were involved in the six-year endeavour. And four of us helped put in place safeguards to protect the privacy of the thousands of patients and volunteers who consented to have their data used in the research. Here, we reflect on some of the lessons learnt for researchers sharing vast amounts of genomic data.

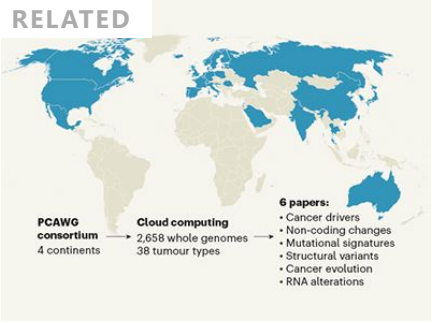
Genomics researchers worldwide are increasingly dealing with vast data sets gathered by consortia spanning many countries. Most are unclear on what to do to protect people's privacy and to comply with international and national data-protection laws, especially given recent and ongoing changes in legislation.

An international code of conduct for genomic data is now crucial. Built by the genomics community, it could be updated as technologies and knowledge evolve more easily than is possible for national and international legislation.

In the clouds

Between 2013 and 2019, 468 institutions from 34 countries in Asia, Australasia, Europe and North America amassed 2,658 cancer genomes — each paired with a non-cancerous sequence from the same person. The effort was led by the International Cancer Genome Consortium (ICGC).

The combined data were made available to investigators largely thanks to cloud computing. The project — the Pan-Cancer Analysis of Whole Genomes (PCAWG) — is the first to try to aggregate so many subprojects across different jurisdictions and make the entire data set available across the world.



Much of the PCAWG data (and the tools for analysing them) were made available through the Cancer Genome Collaboratory, a cloud service built for the genomic research community. (The commercial cloud-service provider Amazon Web Services was also used.) But the data were first processed in high-performance computer centres and the clouds of academic institutions in Germany, the United Kingdom, the

Global genomics project unravels cancer's complexity at unprecedented scale

United States, Canada, Spain, Japan and South Korea. Some were also processed using commercial clouds (Amazon Web Services, Microsoft Azure and Seven Bridges).

Since PCAWG began, several other international genomics projects have turned to the cloud (see 'Cashing in on clouds') including The Human Cell Atlas, an international project to create a reference map of all human cells[1], and the European Open Science Cloud, which is for researchers and professionals in science, technology, the humanities and social sciences[2].

CASHING IN ON CLOUDS

Cloud services have been transformative in enabling large-scale genomic analysis.

Conventionally, any research team wanting to analyse an aggregate data set collected by a consortium would first have to seek authorization from each project partner's research ethics or data-access compliance office. It would then have to download the data from each subproject over the Internet or — more likely — have the hard drives containing the data sent by post. In the case of the Pan-Cancer Analysis of Whole Genomes, which comprises 800 terabytes of raw data, investigators were able to save months, and thousands of dollars, by immediately accessing the data they needed, and experimenting with and customizing the analytical tools developed by the community. They could also obtain authorization to access most of the data from one place — the International Cancer Genome Consortium's Data Access Compliance Office.

This trend is likely to continue. Cloud services are becoming cheaper and more readily available, and researchers are increasingly reaping the benefits of sharing ever-larger amounts of genomic data with international colleagues[3] (see 'Open data').

OPEN DATA

Over the past three decades, geneticists around the world have been sharing more and more data.

Last year, more than 83,000 researchers from 146 countries downloaded 6.7 petabytes of (mainly human) DNA data from the European Molecular Biology Laboratory's European Bioinformatics Institute. This hosts many biological data sets and makes them accessible worldwide. That is equivalent to around 230 billion whole human genomes.

Such sharing of genomic data will only increase as more data become available. By 2025, more than 60 million patients worldwide are expected to have had their genome or exome (protein-coding regions)sequenced as

part of routine health care[16] – potentially providing a formidable resource for researchers.

Yet cloud services bring fresh challenges with respect to the protection of participants' data – especially given that national governments, law enforcement and private corporations are increasingly showing interest in accessing them. Canadian border authorities, for example, are choosing which country to deport migrants to on the basis of DNA test results from consumer genomic services[4].

Long-term continuity

Organizations such as the Cancer Genome Collaboratory can persist only for as long as they are funded. Even in the case of Amazon and other major tech companies, service outages caused by technical problems, changes to the company's terms of service or even sudden closure of the company could block researchers' access to data at any time. Also, it is often unclear to what extent researchers using cloud services can ensure that their data are not disclosed to third parties, such as those conducting abusive state-level surveillance. Nor is it clear what steps must be taken to protect the data against such breaches of confidentiality.

In the case of PCAWG, the ICGC's Data Access Compliance Office helped to guard against some of these issues. Anyone wanting to use PCAWG data entered into a contract with the project's data-access committee; they had to confirm, for instance, that they would not try to re-identify patients or volunteers once these people's data had been stripped of personal information. No breach of donor confidentiality is known to have occurred.

But even when researchers request the data from an associated data-access committee (as in the case of PCAWG and elsewhere), numerous issues remain unresolved. It is unclear, for instance, what vetting should occur before researchers get access to sensitive genomic data, or what checks should be made before genomic data are shared internationally. Even those involved in PCAWG could not establish a truly international cloud because of restrictions on the transfer of data across borders (caused, in this case, by European regulators having concerns about genomic data from Europeans being held in the United States).

The US component of the project (The Cancer Genome Atlas; TCGA), which contributed one-third of all PCAWG samples, was made available to researchers through the University of Chicago's Protected Data Cloud. Researchers wanting to obtain those data had to abide by an access agreement that was largely compatible with that provided by the ICGC's Data Access Compliance Office. Ultimately, however, TCGA remained conceptually split off from the rest of

the project, because researchers had to follow a different access procedure and to combine the two data sets themselves.

A figure walks past a bank of server racks in a server centre

Server racks in a centre in Berlin. Credit: Thomas Trutschel/Photothek/Getty

Code of conduct

Genomics researchers urgently need clear data-sharing rules that are harmonized across jurisdictions.

An international code of conduct could help investigators to overcome some of the current hurdles, as well as others that might arise as legislation on data protection evolves.

Such a code could outline the steps researchers must take to comply with the European Union's General Data Protection Regulation (GDPR), implemented in 2018, and the US Health Insurance Portability and Accountability Act, among other laws. In fact, the GDPR explicitly encourages the development of sector-specific data-protection codes in its Article 40. And last June, the European Data Protection Board (an independent body tasked with issuing guidance on the GDPR and encouraging the drawing-up of codes of conduct), issued guidelines on the submission, approval and monitoring of such codes for data processing. It also promised further guidance on the use of codes as a potential way to facilitate the transfer of data across borders (see go.nature.com/322nkkv).

A European biobanking research infrastructure, known as BBMRI-ERIC, announced in 2017 that it would develop an EU-wide Code of Conduct on Health-Related Data, to submit to the European Commission (see go.nature.com/2j3ihce). When completed, and if approved, such a European code could be beneficial. Meanwhile, we call on the genomics research community to prioritize the establishment of an international code of conduct that lays out how existing ethical and legal obligations can be satisfied in relation to international genomic clouds.

At least five aspects must be considered.

Identifiability. Despite the problems with it[5], de-identification, in which health data are stripped of any information that could be used to identify the participant (such as name, social security number, address) has long been hailed as a way to protect people's privacy in

research[6]. Yet because of conflicting terminology and gaps in understanding, researchers rarely know what standard they must meet for their data to be properly anonymized or 'pseudonymized' (in which a code enables individuals to be re-identified)[7]. What's more, laws are difficult to enforce in practice because it is often unclear how breaches of confidentiality occurred, or which organization or researcher was responsible[8,9].

Data-protection laws, such as the GDPR or the California Consumer Privacy Act, invariably require the identifiability of data to be analysed on a case-by-case basis, in part because the technological tools enabling identification are constantly changing. Even though it is difficult to lay out hard and fast rules in advance, a code could provide some guidance on how to evaluate when it is reasonable to deposit genomic (and health data more broadly) in open-access repositories. This might involve considering, say, whether a set of genomic variants is somatic or present in the germline and so inherited. (Researchers have shown that it is possible to identify an individual using only a few germline variants[10]; no one has yet been able to identify someone on the basis of somatic tumour variants.)

Broad consent. The GDPR explicitly recognizes an exception to 'specific consent', meaning the consent people give for their data to be used in a specific research project. This is to allow participants' data to be used for certain areas of scientific research, in keeping with recognized ethical standards[11]. Guidance is needed on what researchers must do to meet the requirements for broad consent. Furthermore, how should they keep patients and volunteers informed about how their data are ultimately used?

Return of individual findings, portability and access. How to safeguard participants' right to move their data around – by giving them their data in a machine-readable format, rather than as a printed PDF, for example – should be clarified. The code could also lay out what steps are necessary for responsible communication of health data to a patient or volunteer[12,13]. Should people who are being informed about the identification of genomic variants of malignant or unknown significance be offered genetic counselling, for instance?

Withdrawal. Researchers need guidance on how they can meet participants' right to withdraw from research. The GDPR requires that those entrusted with people's data keep records of third parties to whom they have disclosed those data. And when consent is revoked, they must notify the third parties. Yet all sorts of questions remain, such as whether analyses on aggregate data should be revised with the participants' data removed, and so on.

Compelled disclosure. A code of conduct could provide researchers with guidance on how to deal with government requests for personal data, including what legal protections they can

appeal to. In the United States, for example, the National Institutes of Health's Certificates of Confidentiality are designed to shield researchers from such requests[14].

Next steps

The achievements of PCAWG in relation to the sharing and handling of genomic data augur well for the development of an international code that researchers everywhere can refer to.

Genomic research consortia, public and private funding bodies, and those working on existing regional codes (such as the one in Europe) might begin the process of building it. A first step would be to convene a meeting to determine the topics the code would touch on, the best way to consult research participants about their needs and a decision-making process that will allow the text to be finalized in a timely way.

If genomics researchers are instead left in the dark about how to properly address data protection and sharing, they could either be excessively cautious and fail to share as consents allow, or fail to provide participants with appropriate protection[15]. In other words, further regulatory uncertainty risks stalling new genomic analyses and undermining people's faith in scientific collaboration for the public good.

Nature **578**, 31-33 (2020)

doi: 10.1038/d41586-020-00082-9

References

1. Rozenblatt-Rosen, O. *et al. Nature* **550**, 451–453 (2017).

2. *Nature* **523**, 136–137 (2015).

show more

SUPPLEMENTARY INFORMATION

1. Full list of PCAWG consortium working groups and writing committee

Nature ISSN 1476-4687 (online)