JIM BLOCK PHOTOGRAPHY

Be prepared for the big genome leak

It is only a matter of time until idealism sees the release of confidential genetic data on study participants, says **Steven E. Brenner**.

Most people in the United States could soon know someone whose genome is held in a research database. Concerns are growing about our ability to properly control access to that information. Also growing among some scientists is the feeling that restricting access to genomic data fetters research. How long will it be until an idealistic and technically literate researcher deliberately releases genome and trait information publicly in the name of open science?

Both the open-access literature and the open-source software movements began with idealists. It seems inevitable that there will be a major leak of genome information in the near future. Individual scientists, institutions and funders should consider now how they will react when this happens.

Some studies already gather the genetic data of more than 50,000 individuals in a single analysis. Although this information is supposed to be highly protected, it is disseminated to various institutions that have inconsistent security and privacy standards. In practice, data protection often comes down to individual scientists. Once leaked, these data would be virtually impossible to contain.

What harm would come from a leak of personal and genomic data? The consent form for the Personal Genome Project (PGP) — which makes no attempt to keep genetic information secret — offers a guide. It lists a range of adverse consequences, from revealing non-paternity to being framed with synthesized DNA planted at a crime scene.

Most research genome data are de-identified, but given progress in re-identification and commercial genetic databases, will they stay that way? De-anonymized genomic data would be most likely to reveal health conditions relevant to the study for which they were collected. The effects might be uncomfortable but would probably reveal less than a typical Google search history. So far, no PGP participant who released genomes and traits has experienced adverse consequences that have been reported to the Institutional Review Board. In the longer term, the risk of harm may rise as our understanding of genetic variation increases.

Then there is the public outcry a genome breach might incite. The public often has an exaggerated perception of the links between genes and personal traits. Lacking contextual information, research participants could wonder whether their own genomes had been leaked and dread implausibly dire consequences.

Thus a genome leak might lead to a backlash. Volunteers might withdraw from research studies and refuse to join new ones. Research might even be subject to moratoriums and prohibitive restrictions. The harm to genetic research could be great, and study participants could be unsettled.

NATURE.COM
Discuss this article online at:
go.nature.com/oybzqm

THE QUESTION IS NOT HOW TO PREVENT **A LEAK** BUT HOW TO MITIGATE THE **FALL-OUT.**

What can be done? Two extreme options offer appealing simplicity. One is for research projects to incorporate unrestricted data release from the outset. This option should be offered more broadly owing to the certainty and research benefits it offers. However, would enough people be willing to share so openly? The second option would be to lock down genomes so tightly that they are virtually impossible to steal, for example by only allowing analyses on central computers through restricted interfaces. Although useful as an alternative, this system would stymie research were it to become the exclusive means of access to data, but it would still remain vulnerable to ingenious ways of eliciting inappropriate genetic information.

Neither option is comprehensively workable, which means that the question is not how to prevent a leak but how to mitigate the fall-out. This requires some specific steps, as well as progress in adapting concepts already used elsewhere in biological research and in applying principles proposed by groups such as the Presidential Commission for the Study of Bioethical Issues in Washington DC.

Funders should develop rapid mechanisms for notifying study participants, governments and the media when breaches occur and provide informed guidance about scope and probable consequences for those affected. This would require recontacting research participants to warn those whose data were leaked and, implicitly, to calm others whose data remain secure. More research is needed about the possible harm of such leaks to better inform and protect research participants before and after leaks occur.

We should also take steps to minimize the frequency and extent of future genome leaks. Institutions could establish uniform protocols and reviews to ensure the safety of protected genomic data. All researchers using restricted genomic data should be trained regarding the ethics of and the technologies involved in protecting human data. Technical and legal strategies should be proactively deployed to help limit dissemination of leaked data to those who furtively hunt for them.

Augmented legal protections could reduce the harm from inappropriate use of such data. In the meantime, we need to address a quandary: research with leaked data would undoubtedly speed immediate scientific progress, but should scientists exploit them?

Most importantly, we must ensure that the necessary discussion about the risks of a genome leak is balanced with information about the tremendous benefits that collected genetic information has for all of us. Although the acceleration and promise of genomics makes a leak inevitable, it also guarantees medical progress. SEE EDITORIAL P.137

Steven E. Brenner *is a Professor at the University of California, Berkeley.*
e-mail: brenner@compbio.berkeley.edu

0

Note that the title, subtitle, and pull-quote were picked entirely by the editor; they were not my choice.

1

Leaks happen. A soldier with a security clearance smuggled 750,000 classified and protected military and diplomatic reports from a secure workstation in Iraq[1,2]. Conveyed to WikiLeaks, these became a sensation. Disclosure of 2.5 million documents about 120,000 offshore companies to the International Consortium of Investigative Journalists has led to red faces, tax investigations, and a firing of a Parliamentary deputy speaker[3,4]. A spreadsheet with 20,000 Stanford Hospital emergency room admissions, including names and diagnosis codes, showed up on Student of Fortune, a homework help site[5]. Aaron Swartz downloaded millions of journal articles, possibly intending to redistribute them; the New Yorker just deployed DeadDrop, a digital leak collection service that he developed[6]. LulzSec showed how a small cabal of hackers could steal millions accounts from a multinational, "for the lulz" (perverse entertainment)[7,8,9]. PrivacyRights.org records more than one serious data breach per day; in California alone, the Attorney General averages more than 10 notices of major breaches per month[10].

And, of course, there are Edward Snowden's recent leaks of Top Secret Foreign Intelligence Surveillance Court (FISC) and NSA PRISM documents. (Those leaks occurred after this piece was drafted.)

[1] http://www.washingtonpost.com/world/national-security/judge-accepts-terms-under-which-manning-would-plead-guilty-to-lesser-charges/2012/11/29/e1d2ecae-3a41-11e2-b01f-5f55b193f58f_story.html
[2] http://www.wired.com/threatlevel/2013/02/bradley-manning/
[3] http://www.washingtonpost.com/investigations/piercing-the-secrecy-of-offshore-tax-havens/2013/04/06/1551806c-7d50-11e2-a044-676856536b40_story.html
[4] http://www.icij.org/offshore/secret-files-expose-offshores-global-impact
[5] http://www.nytimes.com/2011/09/09/us/09breach.html
[6] http://www.newyorker.com/online/blogs/closeread/2013/05/introducing-strongbox-anonymous-document-sharing-tool.html
[7] http://www.wired.com/threatlevel/2013/05/lulzsec-sony-hackers-sentenced/
[8] http://arstechnica.com/tech-policy/2011/06/lulzsec-heres-why-we-hack-you-bitches/
[9] https://twitter.com/LulzSec – "high-quality entertainment at your expense."
[10] http://oag.ca.gov/ecrime/databreach/list

2

Several people have asked whether anyone would actually commit such a leak, as it would be career suicide. Obviously this would be a deterrent, but there are several reasons I think a leak could still occur. First, the person may believe that they can make the leak anonymously. Increasingly sophisticated tools exist for sharing data with degrees of anonymity. As a sufficiently high number people have access to genomic data on systems that have sufficiently limited security, it would be plausible to believe that this could be done with impunity. Second, a committed idealist may be willing to risk the ensuing sanctions. The concept of career suicide seems to be a modest concern when considering that Edward Snowden's and Bradley Manning's leaks could lead to extensive jail time. Third, academics (perhaps especially graduate students) are a community particularly enriched in idealists—anyone capable of leaking genomes could easily have chosen more lucrative employment, and has chosen research deliberately. Their goals are usually to promote open discovery. This is noteworthy when contrasting with those who have security clearances and thus have been specifically vetted for their ability to keep information secret. Finally, the recent newsworthy rehabilitations of people such as Mark Sanford, Adrian Lamo, Henry Blodgett, and Jonah Lehrer show that second acts are hardly uncommon.

I think the most likely source of a leak is from an idealist (broadly defined to include, for example, hactivists). However, as hinted in Note 1 above, I think a leak is inevitable because there is a panoply of means by which a breach could occur—ranging from malevolent hacking to carelessness.

3
Databases are getting larger, to enable yet more effective research. After this piece was drafted, a whitepaper was released about a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data. http://www.broadinstitute.org/news/5046-0

4
It is often said that one cannot delete information from the Internet, and the Streisand Effect posits that the very attempt to do so garners yet greater unwonted (and unwanted) attention.
http://www.kulfoto.com/funny-pictures/39387/delete-something-from-the-internet

5
The consent form for example also lists "contact from the press" amongst the risks.
http://www.personalgenomes.org/consent/PGP_Consent_Approved_02212013.pdf

6
For more information about how this works, see Extended Note 6 on page **7.**

7
Google knows a lot about you. For a little more information, see Extended Note 7 on page 8.

8
Relatives also may be concerned, for by dint of genetic relationship, portions of their genomes are also being released in these cases. But the shared ownership of genomic information of relatives is a larger problem that in general has not been comprehensively addressed.

9
Much of the discomfort to participants could be needless (because their genome was not amongst those leaked) or excessive (because they lack context to understand the likely short-term and long-term harm).

10
Facebook notwithstanding, I believed that making study data entirely open would cripplingly reduce enrollment in for biomedical research studies. However, some people I spoke with disagreed. Perhaps many people don't want to be part of a clinical study—but those who are willing to participate may be also the people who are willing to share their personal information. Indeed, I learned that when dozens of volunteers in a recent study were offered the additional option of extensive genetic disclosure, all availed themselves of it.

11
Such an alternative has been widely explored, such as at the NIH meeting about Establishing a Central Resource of Data from Genome Sequencing Projects
http://www.genome.gov/27549169

The Global Alliance also seems to primarily envision some similar sort of mechanism. http://www.broadinstitute.org/news/5046-0

Such resources would allow people access to data without having to set up local security protocols and computational resources, which could be a boon to research. Programmatic restrictions on data access would reduce risks of inadvertent data leakage, and thwart casual attempts at inappropriate data sharing. However, we continually learn of more ways in which information about individuals can be elicited from summary information. See dx.doi.org/10.1371/journal.pgen.1000167 for one example. Moreover, at some point individual variation information does become critical for specific studies: it may be that a specific mutation that nearly uniquely identifies an individual is specifically causative of the trait of interest.

12
Such steps should be taken now, before large publicized leaks occur.

13
See http://bioethics.gov/node/764

14
Patients also should be warned of the prospect of data breaches in consent forms, and some consent forms are silent about possibility of data leakage. In such cases, participants would have had no real warning about this outcome. While providing such warning may seem mainly to just offer legal protections for institutions responsible for securing the data, the individuals whose data were leaked may be slightly mollified by the fact that this possibility was not wholly unforeseen. Reducing liability for institutions might seem to give them less motivation to secure data, but it will also reduce the incentive for hackers to steal data for ransom and blackmail.

15
Often recontact is not permitted in legacy consent forms to protect subjects, but such notification could help protect them in the event of a leak.

16
Additionally we should evaluate whether any steps can be taken to specifically protect the affected individuals, analogous to the credit monitoring offered to those whose social security numbers are disclosed.

17
Biology research laboratories must have chemical hygiene procedures, submit protocols for biological uses for institutional review, and have inspections to ensure they are compliant. Institutions could enforce analogous safeguards for biological data. Additionally, they could make more broadly available the types of safeguards already available in clinical environments for managing HIPAA data—though in many cases security in such systems are designed specifically to thwart (inappropriate) research.

18

Researchers who study human subjects undergo substantial training. However, they have little training about techniques for maintaining data security. Further, users of much "de-identified" human genetic information are not required to have any human subjects training at present. Researchers should be educated about how to protect data, just as they receive laboratory safety training; they should also be taught why these protections are in place and how in this instance the protections promote research.

19

Similarly, just as studios routinely scan the internet for pirated movies and issue legal takedown notices, so funders should establish surveillance for leaked genomic information and develop systems to help have them taken down when possible. While this risks the Streisand Effect, legal protections could make these approaches more effective.

20

These include broad legal protections like the Genetic Information Nondiscrimination Act http://www.genome.gov/24519851 and enhanced successors.

21

Because these leaked data are now "public," would it be appropriate to use them in biomedical research, where they would actually lead to more rapid discoveries? Or should these data be verboten because of their origin?

This can lead to paradoxes, such as the situation where those with security clearances are generally not allowed to read the secret WikiLeaks collection—even though the general public can sift through it as they like. http://www.nytimes.com/2010/12/05/world/05restrict.html

Similarly in 2006 AOL released millions of search queries to support research, but it quickly became clear that these allowed re-identification of some users who made the queries. http://www.nytimes.com/2006/08/09/technology/09aol.html These data are readily found in web searches and seem to used by researchers to develop new methods, but it seems that the community treats these as verboten in publications.

In biology, a longstanding problem is the HeLa cell line derived from cervical cancer cells taken from a woman named Henrietta Lacks without permission. These cells continue to be widely used to this day. This story is described in detail in The Immortal Life of Henrietta Lacks by Rebecca Skloot 2010 (Crown, ISBN: 1400052173). The challenges associated with these unconsented cells recently came into focus when their genome sequence was publicly released by a group unaware of the troublesome history. http://www.nytimes.com/2013/03/24/opinion/sunday/the-immortal-life-of-henrietta-lacks-the-sequel.html

22

In particular, genetic research will often have the greatest beneficial effect for those in a research study (and their families), because that provides the best opportunity for their own genetic features to be associated with diseases afflicting them. The discoveries can lead to biological understanding and thence hopefully to effective ways to maintain health, as well as accurate diagnosis and treatment.

Frequently asked questions

What is this document?

Collected jetsam. I submitted a rough piece to *Nature*, and the editor there helped whip it into shape and cut it down to size. Along the way, bits of text were left on the cutting room floor. I've collected some of those elisions that help motivate or explicate my points, as well as expanded on some queries I received. This document does not purport to be a full-fledged support of every detail in the piece.

Why are you providing this annotation in such a weird form?

Nature owns the copyright on my piece, and they were firm that I could not post the text anywhere else for 6 months, even for the sole purpose of providing the annotation. This print overlay was designed as a method that could be used by anyone to make an annotated version, without any special technology beyond a PDF viewer and a printer. Let me know if you have better ideas for how to legally distribute the annotations.

So do you think *Nature*'s copyright agreement is unfair?

No. I would not have signed the agreement if I thought it was unfair. *Nature* provided a platform for me to share these ideas. They made the article freely available on their website—and printed tens of thousands of copies of it mailed around the globe. You wouldn't be reading this piece had they not provided me this opportunity.

Also, *Nature* added value to the piece. The editor heroically rewrote my rough draft and developed it into a piece with suitable style and length. Even the photo researcher had a useful insight, delicately pointing out that perhaps it wasn't ideal to use a picture of me grinning maniacally as I contemplated widespread harm to individuals and research. All of the people I have interacted with at *Nature* in the past decade, from senior managers to graphic artists, have been thoughtful and done a quality job (though too often they fail to perceive my manuscripts' brilliance—and reject them)

That said, I think *Nature* was silly not to grant me rights to redistribute the work with this commentary, especially given the piece's content—and I told them so. I think providing this commentary increases the value of the contribution, which benefits everyone. Had this work been more open, that would have been better. I think *Nature* could and should offer more open access options.

Does this conflict with your longstanding support of open science?

I don't think so. Good science involves sharing the most important discoveries with the most people who can benefit from them. Open access literature and open source software are powerful tools for this, and I strongly advocate them. (I was a founder of the Open Bioinformatics Foundation and the open access journal *PLOS Computational Biology*, and my lab has a special agreement with the University to produce open source software.) But in this case, I think *Nature* provided the best means for disseminating the ideas. Also, this is a policy opinion piece, not a work of scientific discovery.

This seems like a stunt just to draw attention to the value of open access and content remix.

Sorry, is that a question?

6 **Extended Note.** A quick tutorial on genome re-identification. Hopefully this will be more helpful than misleading.

For most Americans, a genome is sort of like a social security number: you get it at birth and you generally can't change it. Both a genome and a SSN are currently mostly not interpretable, but there is meaning embedded in each. For example, the first digits of an SSN reveal where and when it was issued, and of course genome variations in conjunction with environmental exposures are responsible for our individuality. But an isolated SSN or an isolated genome do not presently reveal terribly much. For example, "057-30-2830" does not mean much by itself, though with the aid of a database one can find that it was probably issued in New York between 1953-1955.

Research studies match genomes with traits. So, researchers for a lupus study may have a collection of genomes from people with lupus, which they compare with a collection of genomes from nominally healthy people. Even if these data broke totally into the open today, they wouldn't reveal much immediately. This is because generally we can't match a genome back to a person today. Similarly, in isolation my telling you that the person with SSN 057-30-2830 won a Nobel prize still doesn't leave you being any wiser about who won a Nobel prize. A genome without additional information (like the name of the person it belongs to) would be considered "de-identified."

The problem is that there are ways to re-identify people. One challenging way is to use scattered bits of information collected by the study and included with the genome to try to figure out who the person is. For example, if I told you 057-30-2830 belonged to someone who was 88 years old and died in La Jolla, San Diego on 28 July 2004, it would not be hard to determine (perhaps by looking in a newspaper archive) that I was referring to Francis Crick.

Similar techniques, for example using genealogies and bits of genetic information, can be used to re-identify research participants from various public databases. For example, see http://dx.doi.org/10.1126/science.1229566, http://arep.med.harvard.edu/PGP/Anon.htm, and http://arxiv.org/abs/1304.7605

Such techniques for re-identification from genomes are getting better. Perhaps a greater potential concern is the growth of commercial databases. Any credit bureau would immediately match 057-30-2830 with Francis Crick, as would any other database that had collected his name and SSN together (such as Crick's bank, health insurer, etc.). Thus if I told you 057-30-2830 got a Nobel prize, and you had access to any such database, you would trivially deduce that Francis Crick got a Nobel prize.

Similarly, there are a growing number of commercial databases that have genomic information matched to names. Consider again the collection of genomes in the lupus study. With access to a database that contains any of those genomes, it would trivial to associate that genome with a name. With such a database, then one would then be able to rapidly infer that the person with that name presumably has lupus.

7 **Extended Note.** If you have diabetes, you can bet that Google has implicitly figured it out from the searches you have made—and from the websites you visited (many of which are tracked with Google Analytics and its ad trackers), and from your scanned Gmail account. We willingly give these data to Google in exchange for providing very useful search and other services. These data also help Google provide better services, both in general and for us personally. For example, once Google has effectively deduced that someone is diabetic, they will infer that a search for sugar is more likely to refer to blood glucose than a baking ingredient. However, Google also monetizes this information to sell advertisements targeted at us.

It is a measure of how much personal information is kept in commercial databases at Google, Facebook, Verizon, and their ilk that the NSA has partially outsourced its intelligence collection to those companies. It seems easier for the NSA to learn about targets by subpoenaing information from commercial databases than it is to collect the information firsthand. In short, these companies have extensive surveillance to which we willingly submit.

Risks of genomic information and traits being released should be compared with these types of information we divulge to such commercial databases on a daily basis, and the risks that those databases entail.

Acknowledgements

I am grateful to many people for feedback and thoughts, which greatly improved the piece, including George Church, Stephen P. Ketchpel (http://giving-back.info/), Dorit Berlin, and others who have not yet given their permission to be named. I also thank Robin Peters and Peter Wu for technical help in producing this commentary and overlay, and Annsea Park for the printer image.