THE MYTH OF ANONYMITY

It may not be possible to protect the identity of genomic data. But how much of a problem is that?

ANDREW BAKER

BY NEIL SAVAGE

Misha Angrist is not worried about strangers discovering his personal genetic information, even though it was made public in 2007 and has his name attached. Angrist was the fourth person to submit his genetic sequence to the Personal Genome Project, an effort led by George Church, a geneticist at Harvard Medical School in Boston, Massachusetts, to advance medicine by publicly sharing genomic and health data.

"It was kind of a political statement," says Angrist, a geneticist who studies bioethics and science policy at Duke University's Social Science Research Institute in Durham, North Carolina. He had become frustrated that privacy considerations prohibited scientists involved in genetic studies from interacting with the people those genes belonged to. "We were not allowed to talk to the people we studied, and that always struck me as silly and wrong-headed," he says. The restrictions prevented researchers from gathering additional information, such as recent medical histories or health-related habits, that

might give them more insight into disease risk — and stopped them developing a trusting relationship with the DNA donors.

The Personal Genome Project aims to share DNA sequences, medical histories and other personal information with researchers looking to link gene variants, environment and lifestyle habits to disease risk. The project explicitly does not promise anonymity, and warns that the data will be shared publicly. Each participant is put through an online, questionnaire-based screening process to ensure that they understand both the benefits and the risks of making such information available.

The US Precision Medicine Initiative, meanwhile, is seeking to collect the genomic information and medical records of 1 million participants, and the UK 100,000 Genomes Project is gathering similar data through the National Health Service, raising concerns among privacy advocates that too much personal information could become public. Both projects promise to remove information that identifies participants from the data, and store the data on secure servers that are accessible only to authorized personnel, and they prohibit people from re-identifying the sequences. They concede, however, that anonymity cannot be absolutely guaranteed, and computer scientists have shown that at least some participants can be re-identified fairly easily. Scientists and policymakers are trying to work out exactly what the harm of such disclosures could be, and how they can reduce the risks, but any solutions are more likely to be policy-based than technological.

> "With some knowledge and some dedicated effort, you can identify people from genomic data."

WHAT'S IN A NAME?

Anonymous data are not as unidentifiable as the term suggests. Not all participants in the Personal Genome Project are identified by name like Angrist, but the project does not guarantee anonymity. In 2013, Latanya Sweeney, a computer scientist who heads Harvard's Data Privacy Lab, was able to put names to many of the profiles simply by comparing them with available public records. More than half of the nameless profiles available at the time contained the person's date of birth, gender and postal zip code. By cross-checking against public records such as voter registrations, she was able to attach a name and address to 241 of the 579 profiles. Staff at the Personal Genome Project confirmed that she was correct in all but 7 cases.

The Personal Genome Project is not the only database that is vulnerable to re-identification. Yaniv Erlich, a computer scientist at Columbia University in New York City looked at repeating patterns of nucleotides, known as short tandem repeats (STRs), on the Y chromosomes of men whose DNA had been made publicly available by the international 1000 Genomes Project. He then compared them with data found on two public genealogy databases. The project had not collected names or other identifying information, such as birth date or social security number, and because it stored more samples than were used, there was no way to tell if a given sample was even part of the database. As the project's consent form reassuringly put it: "Because of these measures, it will be very hard for anyone who looks at any of the scientific databases to know which information came from you, or even that any information in the scientific databases came from you."

Despite that promise, however, Erlich was able to put names to nearly 50 people who had donated their DNA. Because the Y chromosome is inherited only by males, it is often linked to family surnames. This means that even if participants in the genome study had not also given their DNA to a genealogy website, people with matching STRs were probably relatives, allowing the researchers to infer more surnames. When his study was published in 2013, Erlich estimated that 12% of US males were vulnerable to this kind of breach. Three years later, with genome databases growing and algorithms for comparing data improving, that figure could be as high as 20%. "It definitely gets easier and easier," he says. "With some knowledge and some dedicated effort, you can identify people from genomic data."

➲ NATURE.COM
Read about a recent effort to ensure privacy:
go.nature.com/2bdwigp

Even those who agree to make their data public may have some information that they would rather keep from other people — or even from themselves. One participant in the Public Genome Project was James Watson, co-discoverer of the double helix structure of DNA. Watson asked that information about his apolipoprotein E gene be redacted — a variant of that gene can indicate a heightened risk for developing Alzheimer's disease, and he did not want to know his risk.

But researchers from the Queensland Institute of Medical Research in Australia and the University of Washington School of Medicine pointed out that merely removing the gene from the database would not hide the information. Other changes to the genome, some in fairly distant parts of the DNA, are correlated with the higher-risk mutation. Watson responded by deleting an even larger swathe of his genome from the database. But that could be a losing battle, the researchers warned. As our understanding of the genome improves, it will be easier to estimate risks for various diseases from different points along the genome.

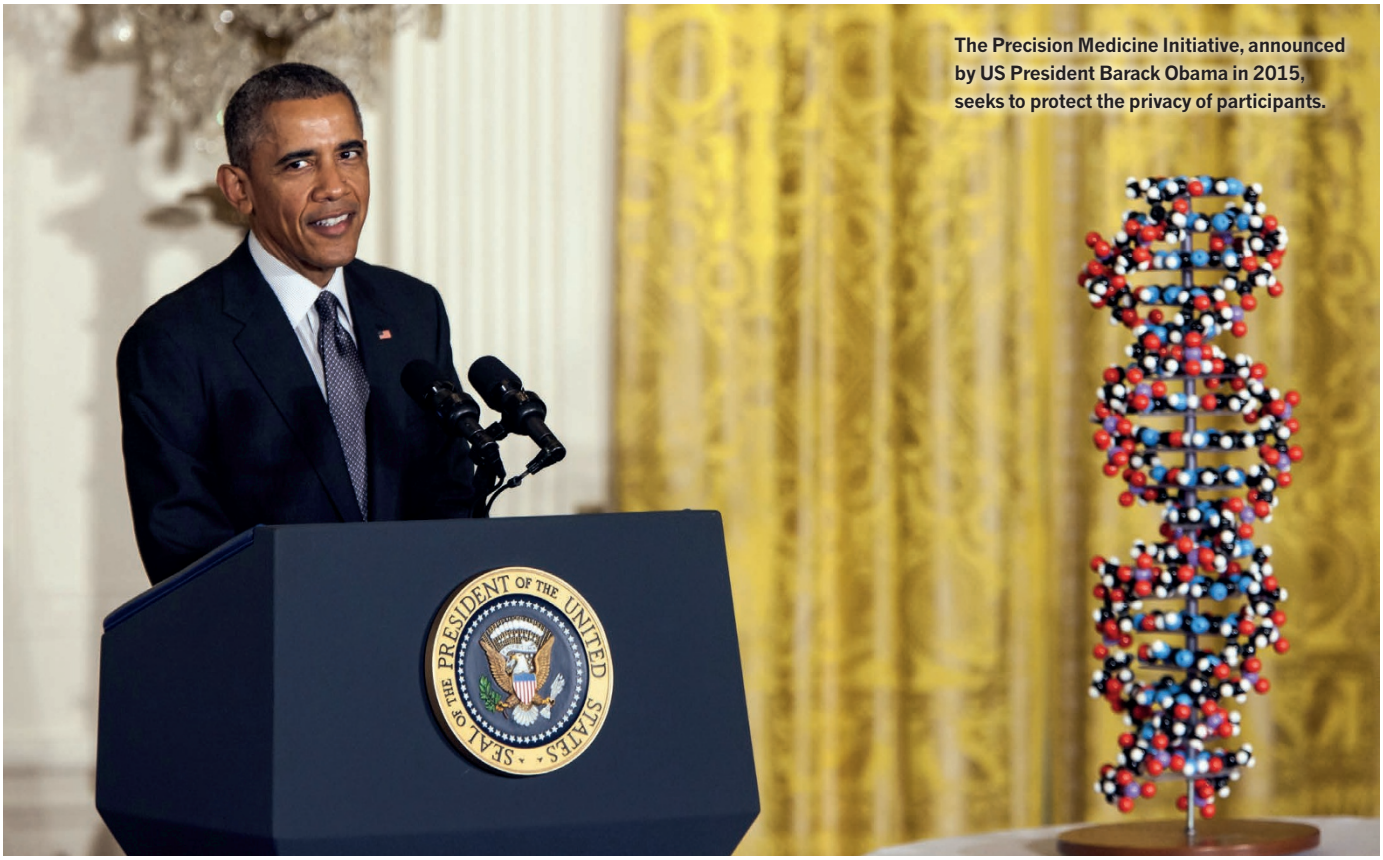
RELATIVE RISKS

If privacy cannot be guaranteed, the next question is whether this is a problem. Some risks seem relatively minor, such as the potential embarrassment of having people find out that you participated in a particular study. But some adoptees have used genetic data to find birth parents who had not expected their identity to be revealed. Others might discover that someone they thought to be a parent or grandparent is not actually related to them.

Include someone's medical history and the potential for awkward revelations grows. If a name can be attached to a genome, and the genome is attached to medical records, then treatments for sexually transmitted diseases, alcoholism or mental illness could be revealed. Some people worry that they may face job discrimination — or health-insurance discrimination in the United States — if a risk of debilitating and expensive diseases is made public.

Some privacy advocates worry that despite the general guidelines developed for the Precision Medicine Initiative, the project lacks legal protections. The World Privacy Forum, a non-profit organization based in San Diego, California, says that data collected by the project are not covered by the main US health-privacy law, the Health Insurance Portability and Accountability Act of 1996. It also fears that courts may decide that when participants volunteer information to researchers, they give away their right to doctor–patient confidentiality. Courts have, after all, previously ruled that police do not need a warrant to collect mobile-phone location data because callers have already shared that information with telephone companies.

"People are still worried about discrimination in health insurance and jobs," says Robert Cook-Deegan, a biologist who studies genomics policy at Duke University's Sanford School of Public Policy. In the United States, the Genetic Information Nondiscrimination Act of 2008 is supposed to prohibit that, but it does not cover long-term care or disability insurance, so people who discover that they may need extensive care for a late-onset disease such as Alzheimer's could still face ruinous expenses. The Canadian government recently debated a similar



The Precision Medicine Initiative, announced by US President Barack Obama in 2015, seeks to protect the privacy of participants.

SAMUEL CORUM/ANADOLU AGENCY/GETTY

law, and the European Union has a general mandate against genetic discrimination. There is no specific UK law against it, however, although the Association of British Insurers agreed to a moratorium until 2019 on using predictive genetic tests to inform insurance policies.

Some of the concerns are speculative, such as the possibility that someone's DNA could be planted at a crime scene. Indeed, the trouble with figuring out how to handle privacy, Erlich says, is that "we really don't understand the concept of harm due to privacy loss."

If anything, the risk of personal information being revealed is probably no greater than that from other sources where people willingly provide information, Erlich says. He points to a 2013 study by researchers at the University of Cambridge, UK, and Microsoft Research that identified people's sexual orientation, political affiliation and race with high degrees of accuracy just by examining their 'likes' on Facebook. That is much more information than you could glean from a genome at present. "There is not a single genetic marker in the genome that can predict homosexuality," Erlich says.

Privacy may not even be the right focus, argues Jenny Reardon, a sociologist at the Center for Biomolecular Science and Engineering at the University of California, Santa Cruz, who in May chaired a conference focusing on the fraught issue of personal data in the age of precision medicine. "Privacy doesn't get us to what is more fundamental: what as a society should we be doing with this data," she says. She would like to see more focus on how these large data sets can improve people's lives. But "no one wants to discuss this", she says.

BE CLEAR ON CONSENT

Whatever the problem with privacy, the solution is unlikely to be technological, Erlich says. Techniques to encrypt data or disguise it with statistical noise are of limited value, he explains, because the more they protect privacy, the less useful they make the data. He thinks that a better approach is to rethink how privacy and consent are handled, and to treat the people who hand over their DNA with respect and honesty.

In an example of this approach, Erlich and colleagues at the New York Genome Center, in collaboration with the National Breast Cancer Coalition in Washington DC, have created a project called DNALand to study the genetic risks of breast cancer. People donate the genetic information that they get from DNA-testing companies such as 23andMe, Family Tree DNA and Ancestry.com. In return, DNALand offers users free information about their genome and the possibility of identifying relatives based on genetic matches, as well as the chance to contribute to improving medical knowledge. The consent form spells out the risks and benefits of participating and allows people to withdraw at any time. It also promises to seek further consent before sharing data with a third party.

One problem in obtaining consent is that, once collected, genomic data can be stored indefinitely and used in ways that the original researchers did not foresee. "That's the whole idea of research. You don't know what you're going to find," Cook-Deegan says. The people who set up databases need to take a long view when making promises and asking for consent as they collect the data, he says. The Precision Medicine Initiative has a set of general guidelines about transparency and respect for participants' wishes, and these will be used to inform the future development of more concrete privacy protocols. "The problem we're going to have is to make sure we have a system that respects the rights and interests that were set up at the front end," Cook-Deegan says.

Not being clear about how participation in a study could lead to privacy breaches creates the risk that any problems that arise may make potential donors less willing to have their DNA sequenced. "We can't do research on human beings and look people in the eye and promise them that nothing bad will ever happen," Angrist says. "If we reassure people and something bad happens, then it's that much worse."

Instead, he argues, engaging with donors and spelling out the risks and benefits can change the privacy equation. "If you talk to people who have children with undiagnosed diseases, they would tell you: 'We would gladly forgo privacy in the interest of accelerated research'." ■

Neil Savage *is a freelance writer based in Lowell, Massachusetts.*