Open-data project aims to ease the way for genomic research

A revamped approach to informed consent will liberate data for science, developers say.

Erika Check Hayden

25 April 2012

Last summer, biologist Andrew Kasarskis was eager to help decipher the genetic origin of the *Escherichia coli* strain that infected roughly 4,000 people in Germany between May and July. But he knew it that might take days for the lawyers at his company — Pacific Biosciences of Menlo Park, California — to parse the agreements governing how his team could use data collected on the strain. Luckily, one team had released its data under a Creative Commons licence that allowed free use of the data, allowing Kasarskis and his colleagues to join the international research effort and publish their work[1] without wasting time on legal wrangling.

Together with other funders, the Ewing Marion Kauffman Foundation, based in Kansas City, Missouri, is now launching a product that aims to "create the world's largest pool of openly available, user-contributed data about health and genomics" in hopes of easing challenges with informed consent and data ownership that some biomedical researchers say are holding science back in the era of 'big data'. The Portable Legal Consent for Common Genomics Research, developed by the Consent to Research project, is a system through which users can donate their data to databases that remove identifying details, such as name and e-mail address. The databases then assign an identification number to all of the data from each user and deliver the de-identified data to researchers, who must agree to broad conditions designed to prevent harm to the data contributors. Data donors must also undergo a detailed informed consent process, including, for instance, watching a six-and-a-half minute video that cannot be fast-forwarded.

The project received approval from ethics reviewers on 23 April; as soon as May, anyone will be able to sign the consent and begin contributing their own data to the database.

John Wilbanks, head of Consent to Research, says that the project aims to give researchers access to the expanding universe of data being collected through new avenues such as direct-to-consumer genomic testing, devices such as a 'FitBit' that collect data on people's daily habits, and lab tests ordered through medical providers. Through the Portable Legal Consent, all of a user's data from such devices will be assigned the same identifier. The identifier, along with consent to use the data, will then follow the data through any studies for which it is used. Wilbanks hopes that 25,000 people will contribute data to the project by the end of this year. "We want to enable unanticipated computational research on this giant pool of crowdsourced data," Wilbanks says.

Honest approach

The Portable Legal Consent will initially deliver data to Synapse, a computational research environment developed by Sage Bionetworks, a non-profit biomedical research organization based in Seattle, Washington. But the project is also developing tools that will allow researchers outside of Synapse to tap into its databases. The project, part of a groundswell of new consumer- and patient-driven models for conducting science[2], approaches consent in a different way to many genomic studies, by informing donors that it is not possible to guarantee full anonymity of their data, not tying the data to specific studies, and asking researchers to respect broad terms of use — by not attempting to identify data contributors and not sharing the data with others who don't agree to the same terms, for example.

Bioethicists and researchers have argued that informed consent is not truly possible for genomic studies, because it's impossible to predict all the ways that a patient's data might be used at the time that it is collected. And a study[3] published in *Nature Genetics* on 8 April highlighted the issue of re-identification — the prospect that a research subject's identity can be figured out on the basis of data donated to research projects or shared on social media services.

Pilar Ossorio, a bioethicist at the University of Wisconsin Law School in Madison, called Consent to Research's approach to these problems "more honest" than that used by many projects, which, she says, "collect samples, strip them of identifiers, and lump them into a much larger project through which data will be released to a broad enough group that it's almost certainly something that the participants did not anticipate when they signed the initial consent form".

Sharing the wealth

Kasarskis, now co-director of the Institute for Genomics and Multiscale Biology at the Mount Sinai School of Medicine in New York, also

agrees that researchers should stop offering research participants hollow promises about privacy. "We need to move beyond an assumption that you cannot be identified from the data that exist about you and really work to make sure that we're protecting people's rights in ways that allow us to use the data that are out there for individuals' and researchers' benefit," Kasarskis says.

Some researchers say that the approach will have limitations because the study participants are likely to be more educated and healthier than the general population because they have, for instance, paid for a genome scan or bought an expensive device to help monitor their fitness. Others argue researchers who work on most existing big-data studies aren't motivated to share their results.

"For folks who already have well-established cohorts of patients and controls, sharing is not the first thing on their priority list," says Atul Butte, a physician, researcher and entrepreneur at the Stanford University School of Medicine in California.

But, Butte says, Portable Legal Consent could benefit researchers without the means to assemble large data sets. "For the little guy, or the lab that has something unique to offer but has trouble accessing the patients or the data, this is going to be an enormous movement forward," he says.

For Wilbanks, this is exactly the point. "We want to lower the barriers to using this data so much that we can get results that really surprise us," he says.

Nature | doi:10.1038/nature.2012.10507

References

1. Rasko, D. A. *et al. N. Engl. J. Med.* **365**, 709–717 (2011).

2. Kaye, J. *et al. Nature Rev. Genet.* **13**, 371–376 (2012).

3. Schadt, E. E., Woo, S. & Hao, K. *Nature Genet.* http://nature.com/doifinder/10.1038/ng.2248 (2012).