NEWS

‘Sanitizing’ functional genomics data may prevent privacy breaches

BY **PETER HESS**

8 JANUARY 2021

A technique for masking portions of a person’s raw genomic data increases shareability **without sacrificing privacy**, a new study shows.

To conduct functional genomics research, scientists must share genetic data about large numbers of participants, which can inadvertently expose a person’s stigmatizing or otherwise private details to unscrupulous parties or employers.

Under U.S. law, employers and health insurers are **forbidden to discriminate** on the basis of genetic test results. But not all countries have such laws in place, and anecdotal reports suggest that employers in the United States **flout the law**. In one Swedish survey, more than half of parents reported fears about how their **autistic children’s genetic data** could be used against them in the future.

Informational labels on genetic data can also expose study participants to privacy breaches, says **Yaniv Erlich**, associate professor of computer science at Columbia University and chief science officer at genetic genealogy company **MyHeritage**, who was not involved in the work. Malicious actors can link genetic database donors to their genetic data by cross-referencing multiple registries to re-identify them, a breach called a ‘linkage attack.’

The new data ‘sanitization’ technique obscures regions of a participant’s genome in a dataset to secure her privacy, and may encourage more people to participate in genetic studies, says lead investigator **Mark Gerstein**, professor of biomedical informatics at Yale University.

“If someone hacks into your email, you can get a new email address; or if someone hacks your credit card, you can get a new credit card,” Gerstein says. “If someone hacks your genome, you can’t get a new one.”

Masking data:

To determine which information and how much of it should remain private to prevent a linkage attack, Gerstein and his colleagues performed linkage attacks on existing genetic datasets. In one sample attack, they compared two publicly available databases and RNA sequencing results to successfully identify 421 individuals.

In another linkage attack, Gerstein's team sequenced the RNA of two volunteers and shuffled these data into a larger dataset. They then obtained DNA samples from the volunteers' used coffee cups and sequenced their genomes. Again, they could link the two individuals to their genomes with a high degree of certainty.

Based on what they learned from the mock linkage attacks, Gerstein's team developed a technique to mask some variants from a person's genetic data while preserving where those variants are located in the genome. To do this, they replace the genetic variant of concern with one from a reference genome; which variants are removed depend on the genetic conditions or predispositions someone's genetic data reveals.

Introducing too many of these privacy-masking variants can decrease the usefulness of the data. But Gerstein's team struck a balance that enables researchers to obtain data on gene-expression values but also enables study participants to dictate how much of their genetic information they wish to keep hidden.

The work appeared in November in *Cell.*

Adoption issues:

This study shows how linkage attacks can reveal sensitive information about research participants, says **Karen Maschke**, a research scholar at The Hastings Center, a nonprofit bioethics research institute in Garrison, New York. "The privacy-preserving data format they developed is another layer of protection."

But more work is needed before genetic-data custodians are likely to adopt it, Erlich says. Changing how genetic data are stored is not as simple as installing a new computer program; researchers who maintain genetic databases will also require more replications of this software, because even though the masking technique looks strong, there may be "weaknesses that we are not aware of," he says.

In addition to using such techniques, genetic researchers should do more to earn the trust of their study participants by explaining how they will protect data and make things right if it is leaked, Erlich adds. "Privacy is not the problem; it's trust. When there is trust, you don't really need privacy."

Solid genetic research requires data from many people, and “to get all those millions of people to participate, we have to give them good assurances that their privacy is protected,” Gerstein says. Trust is especially critical when it comes to studying genetic conditions that are heritable, such as autism, because disclosing that genetic information has the added potential to expose participants’ relatives.

Gerstein and his team plan to offer their software to different labs to get people comfortable with using it, and to determine whether it works at larger scales than those tested in this study.