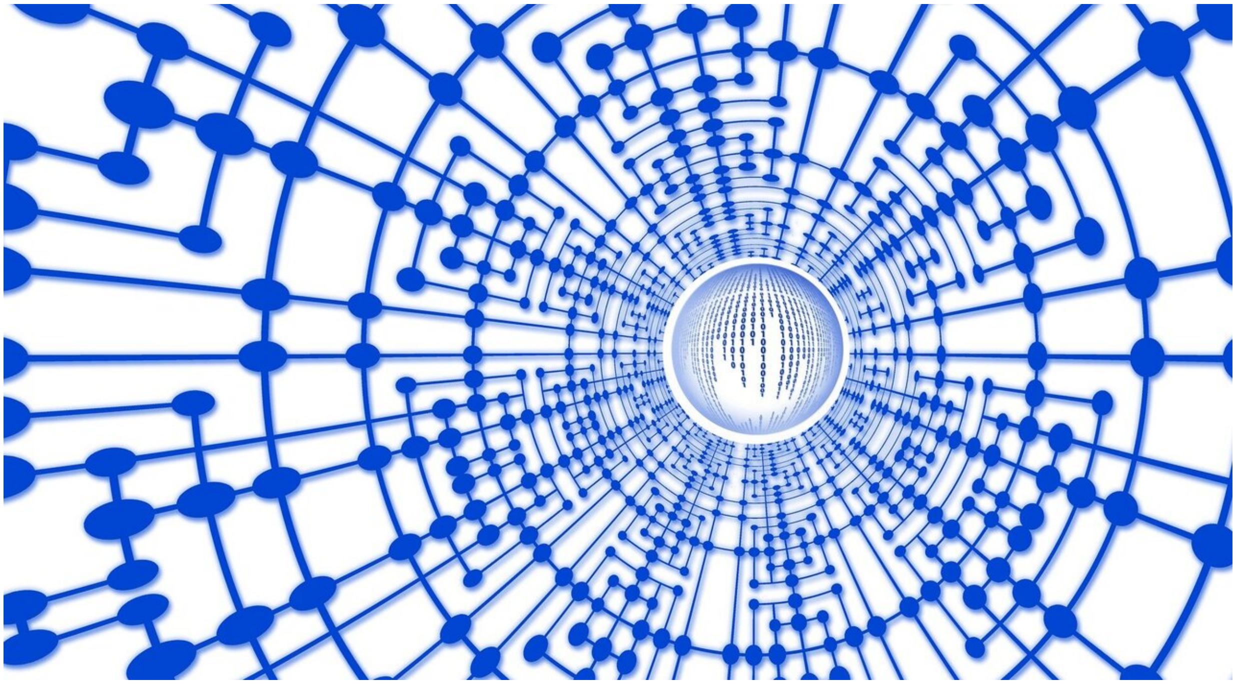
Medical Xpress

Researchers simulate privacy leaks in functional genomics studies

November 12 2020



Credit: CC0 Public Domain

The functional genomics field, which looks at the activities of the genome and levels of gene expression rather than particular gene mutations, generally relies on aggregating information from many samples for its statistical power. This means that broadly sharing raw data is vital; however, sharing these data currently is challenging because of the privacy concerns of individuals within those datasets, leading to these data being largely inaccessible behind firewalls. In a study publishing November 12 in the journal *Cell*, a team of investigators demonstrates that it's possible to de-identify those data to ensure patient privacy. They also demonstrate how these raw data could be linked back to specific individuals through their gene variants by something as simple as an abandoned coffee cup if these sanitation measures are not put in place.

"The purpose of this study is to come up with practical ways to broadly share the raw data without creating undue privacy concerns," says senior author Mark Gerstein, a professor of bioinformatics at Yale University.

Functional genomics research is frequently tied to a specific disease. For example, an investigation into a particular psychiatric condition might look at the expression of certain genes in a type of neuron. And, by nature of having their genetic material included in such a study, an individual's medical status with regard to that condition could inadvertently be revealed.

This can happen through what's known as a quasi-identifier. The way a quasi-identifier works is that if someone has enough individual data points about you, even if those data on their own are not sensitive or unique, they can be combined to create an identifier that is unique to you. In a non-genetic setting, this means if someone has your zip code, birthday, the model of car you drive, and other similar data that might not be considered private or sensitive on their own, they might eventually be able to combine them and create a unique profile that would link you to other data that you wouldn't want public—data like financial records that were collected when you applied for a car loan. The same thing could happen if someone were able to obtain some of your genetic variants and link those variants to the presence of your genetic material in a study on a particular disease. This could in turn reveal a diagnosis, such as HIV status or an inherited cancer predisposition, that you would prefer to keep private.

In their study, the researchers constructed a "linkage attack" scenario to demonstrate how someone could make these kinds of connections from functional genomics studies' data by using DNA obtained from a discarded coffee cup. After adding samples from two consenting participants to a functional genomics database, the researchers gathered used coffee cups from the same individuals. They sequenced genetic material left on the cups and were able to successfully match that material to the samples in the database and infer sensitive health information about the participants. The researchers were also able to use DNA information "stolen" from a genotyping database to match the identities of 421 people with phenotypic information found in a test functional-genomics dataset that the researchers constructed for 436 people.

However, the researchers also identified steps that can be taken to thwart these kinds of linkage attacks and safeguard participants' health information when functional genomics datasets are shared. "Functional genomics is special because variants are usually not needed for data processing," says first author Gamze Gürsoy, a postdoctoral researcher at the Gerstein lab. "Because of this, we can sanitize the variants to prevent data being linked back to the private information connected to the phenotypes included in these studies, while still retaining the utility of the data."

To achieve this balance between privacy and data usefulness, the researchers propose a file-format manipulation that will allow raw functional genomics data to be shared while largely reducing sensitive information leakage by generalizing information about phenotypic variants. The file format is based on a widely used standard file-format system, is compatible with a range of software and pipelines, and when tested, showed little loss of utility. The researchers have also developed a framework with which other researchers can tune the level of privacy and utility balance they want to achieve with the file format based on the policies and consents of the donors.

"As more data are released for these kinds of functional genomics studies, concerns about security and privacy shouldn't be lost," Gerstein says. "At the dawn of the Internet, people didn't realize how important their online activities would become. Now that type of digital privacy has become so important to us. If we move into an era where getting your genome sequenced becomes routine, we don't want these worries about health privacy to become dominating."

More information: *Cell*, Gürsoy et al.: "Data sanitization to reduce private information leakage from functional genomics" www.cell.com/cell/fulltext/S0092-8674(20)31233-2 , DOI: 10.1016/j.cell.2020.09.036

Provided by Cell Press

Citation: Researchers simulate privacy leaks in functional genomics studies (2020, November 12) retrieved 27 April 2024 from https://medicalxpress.com/news/2020-11-simulate-privacy-leaks-functional-genomics.html