NATURE | NEWS

Privacy loophole found in genetic databases

DNA donors' identities can be determined from publicly available records.

Erika Check Hayden

17 January 2013

A potentially serious loophole could allow anyone to unmask the identities of people who contribute their DNA sequences to some research projects, researchers report today.

Greg Pease/Getty

Sifting through DNA databases can lead to identify some male subjects that were supposed to be anonymous.

This is the latest in a series of findings over the past five years that have highlighted privacy vulnerabilities in public databases containing genetic data. The US National Institute of General Medical Sciences (NIGMS), part of the National Institutes of Health (NIH) in Bethesda, Maryland, reacted to the study by removing some data from public view. Some geneticists however question that step, although they acknowledged that the research community must respond to the genetic privacy issue.

“I don’t think removing data from the public domain is any kind of answer,” says computational biologist Eric Schadt at Mount Sinai Hospital in New York city, who was not involved in the latest study. “Rather, we should be up front with participants that we can’t protect their privacy completely, and we should ensure that the most appropriate legislation is in place to protect participants from being exploited in any way.”

Revealing sequences

The new study, which is published today in *Science*[1], was led by Yaniv Erlich, a human geneticist at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts. Other researchers, including Schadt, had already shown that it is possible to confirm the identity of a study participant from public genetic data, if one already knows the person’s genetic makeup. Erlich’s team shows that it is also possible to discover a study participant’s identity by cross-referencing research data about him and his DNA sequence (the particular method used is only applicable to men) to information posted on genetic genealogy and public-records databases (see box: Unmasking a genome).

Unmasking a genome

Erlich’s team discovered the identity of genome donors by cross-referencing their genetic markers with demographic information in public databases.

Erika Check Hayden

1. A program called lobSTR extracts information about the haplotypes of genetic markers called short tandem repeats on the donor’s Y chromosome from his genome.
2. The haplotypes are entered into genealogical databases to find possible surnames of the donor.
3. These surnames, along with information about the age and location of a DNA donor, can be entered into demographic databases to pinpoint his identity.

Erlich’s team used this cross-referencing technique to discover the identities of five men whose genomes were sequenced and released as part of the 1,000 Genomes Project, and who had also participated in a project that studied Mormon families from Utah. He was also able to discover the identities of their male and female relatives.

Erlich did not reveal the names of these anonymous participants or their relatives. He notified the NIH, which funds public databases containing data from the 1,000 Genomes and Utah studies, about the potential for a privacy breach. NIGMS decided to remove the ages of participants in the Utah study from public view to make it more difficult to cross-reference their records to information in public databases.

NIH took a similar step in 2008, when it removed some data on genome-wide association studies (which link DNA sequence features with traits such as disease susceptibility), from public view. The move came after a team led by David Craig at the Translational Genomics Research Institute in Phoenix, Arizona, showed that it was possible to confirm the identities of individual participants in such studies.

Low risk

Erlich says that studies like his and Craig's should compel researchers to ensure that genome research participants understand that their identities may be discovered. Craig says that the risk of a privacy breach for any single research participant is "phenomenally low": "You have a better chance of winning the lottery" than suffering harm from such a breach, he says. Still, he says, policy makers have to consider even the most unlikely risks, as the volume and breadth of the data collected in research studies grows.

"Among all these studies taken together, there may be one case where something bad happens and ends up being debated in Congress. And you don't want that to happen," Craig says.

Nature | doi:10.1038/nature.2013.12237

References

1. Gymrek, M. *et al*. *Science* **339**, 321–324 (2013).