Scientists explore safeguards for genomic data privacy

1 March 2014, by Nancy Owano

By now the general public has become aware that mobile phone applications, bank security systems and credit card databases are not immune to vulnerabilities; information thefts happen. Some computer scientists now say it's time to recognize that vulnerabilities in genetic databases need recognition too. Are anonymous genetic profiles truly anonymous? Is data de-identifying technically feasible for genetic data?

People having their genomes mapped for research or for private information would prefer to assume the data will be anonymous. After all, as an article in *Inside Science* points out, a person's genome carries information about inherited diseases and physical traits, stored in strands of DNA. "The consequences of being able to search, cross-reference, and analyze this information are profound." The solution is not to stop the research but to find ways where research can continue without privacy abuse.

At a February 16 symposium at the American Association for the Advancement of Science. (AAAS) Annual Meeting, a panel of experts addressed tensions between the need for genetic privacy and the medical community's interest in genetic data. Yaniv Erlich, a fellow at the Whitehead Institute for Biomedical Research, has been a key voice in discussing the genetic privacy issue. At the February meeting, he talked about routes by which genetic data can be breached..

Also, in a review submitted October last year, "Routes for Breaching and Protecting Genetic Privacy," Erlich and Arvind Narayanan, an assistant professor in the Department of Computer Science at Princeton, called attention to the topic of data privacy of genetic information.

"We are entering the era of ubiquitous genetic information for research, clinical care, and personal curiosity," they wrote. "Sharing these datasets is vital for rapid progress in understanding the genetic basis of human diseases. However, one growing concern is the ability to protect the genetic privacy of the data originators. Here, we technically map threats to genetic privacy and discuss potential mitigation strategies for privacy-preserving dissemination of genetic data."

(Narayanan, elsewhere, in talking about his doctoral research on problems with data anonymization, said his thesis, "in a sentence, is that the level of anonymity that consumers expect—and companies claim to provide—in published or outsourced databases is fundamentally unrealizable.")

The two authors suggested that privacy by design algorithms include access control as well as differential privacy and cryptographic techniques. "So far," they said, "data custodians of genetic databases mainly adopted access control as a mitigation strategy." They added that new developments in cryptographic techniques "may usher in an additional arsenal of security by design techniques."

In January 2013, Erlich took part in a study, "Identifying Personal Genomes by Surname Inference," published in *Science*, that clearly demonstrated a potential for breaches of privacy in genomics studies. The authors of that study are Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin and Erlich. They showed how identities of volunteers who donate personal genome sequence data for research may be revealed merely on the basis of publicly available information. The researchers recovered the identities of nearly 50 anonymous participants in the 1000 Genomes Project through free, publicly accessible Internet resources.

"Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases.

We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources."

The findings were shared with officials at the National Human Genome Research Institute (NHGRI) and National Institute of General Medical Sciences (NIGMS). In response, NIGMS and NHGRI moved certain demographic information from the publicly-accessible portion of the NIGMS cell repository to help reduce the risk of future breaches.

More information: *Science* 18 January 2013: Vol. 339 no. 6117 pp. 321-324 DOI: 10.1126/science.1229566