GENETICS

Genealogy Databases Enable Naming Of Anonymous DNA Donors

CAMBRIDGE, MASSACHUSETTS—One afternoon in March last year, Yaniv Erlich sat down at his computer to do an experiment. Before he became a geneticist here at the Whitehead Institute for Biomedical Research, Erlich was a white hat: a hacker hired by banks and credit card companies to break into their computer systems and identify weaknesses. Now he was about to do something similar with genome databases. With little more than the Internet, Erlich wondered, is it possible to identify people who anonymously donate their DNA for research? In other words, could he hack someone's name from their genome data?

Hunched over the computer with him was Massachusetts Institute of Technology undergraduate (and now Ph.D.) student Melissa Gymrek who had helped develop an algorithm to extract genetic markers from DNA sequences. By applying the algorithm to an anonymized genome from a research database and doing some online sleuthing with popular genealogy sites, they came up with a guess for the name of the DNA donor and information about his family. But was it correct?

Erlich and Gymrek did a quick search with the man's name and state of residence using Google, and a family Web site popped right up. Every single detail that they had guessed about an anonymous DNA donor matched up with this man living in Utah. "I kept looking at my notepad to see if we missed anything," says Erlich, who was so shocked that he had to go for a walk.

On page 321 of this issue, Erlich and his collaborators report that they were ultimately able to expose the identity of 50 individuals whose DNA was donated anonymously for scientific study through consortiums such as the 1000 Genomes Project. Those revelations have prompted the National Institutes of Health (NIH) to hide certain data associated with anonymized DNA sequences that it makes public for researchers. "The scientific community needs to have an open discussion about this," says Laura Rodriguez, director of policy, communications, and education at the National Human Genome Research Institute in Bethesda, Maryland, and a co-author of an NIH response to Erlich's study on page 275.

Privacy concerns have been raised about publicly accessible genome data before. A study 5 years ago showed that individuals whose genomes were in seemingly anonymous pools of DNA data could be identified by certain genetic markers, known as single nucleotide polymorphisms, or SNPs (*Science*, 5 September 2008, p. 1278). But this is the first time that people have been identified without needing a sample of their DNA as a reference.

Erlich's team exploited two tricks. The first is that metadata about anonymous DNA donors, such as age at the time of donation and state of residence, is often included with their sequences. Erlich started with the genomes of 32 men of northern and western European ancestry collected in a public database as part of the International HapMap Project (*Science*, 26 May 2006, p. 1131). Based on the metadata, he knew the men's ages and that each resided in Utah when they donated their DNA. But that only narrowed the search down to approximately 10,000 men.

For the next step in Erlich's hack, he turned to a few dozen SNPs on the Y chromosome called Y-STR markers. These are almost certain to remain unchanged between father and son. Taken together, Y-STR markers are like a family crest that distinguishes one patrilineal pedigree from another. That's a powerful tool if you want to know whether a man is a member of a particular family.

That is where the second trick comes in. Cheap DNA-sequencing has made it possible for people to share their genetic markers in databases on recreational genealogy Web sites. To ferret out the donors' identities, Erlich used the two most popular, which provide free access to databases containing nearly 40,000 records matching Y-STR to surnames.

When he plugged the 10 genomes with the most recoverable Y-STR markers into those genealogy databases, eight strongly matched to surnames of Mormon families in Utah. Ultimately, he was confident of his guesses for the surnames of five of the genome donors.

Erlich then gathered more information on each one using online resources such as public record search engines and obituaries. He hit the jackpot with metadata in records from Coriell Cell Repositories, a facility in New Jersey that provides cells from the 1000 Genomes Project donors to researchers. With that, he identified family members who had donated their own genomes to the same project, including women.

"I was surprised but not flabbergasted," Rodriguez says. The managers of the 1000 Genomes Project were aware of the risks posed by the metadata and genealogy Web sites, but, she says, "We didn't realize how easy it was to access this information." They immediately removed donors' ages from the publicly available metadata—critical for Erlich's method—but Rodriguez admits that this is only a short-term fix.

This has "huge implications" for the way that consent is obtained from DNA donors, says George Church, a geneticist at Harvard Medical School in Boston. Church founded the Personal Genome Project, which has a consent form for donors that is "very explicit that their DNA and trait data are identifiable." By contrast, Church points to a phrase from the consent form of the 1000 Genomes project: "… it will be hard for anyone to find out anything about you personally from any of this research."

Deanonymizing genomes could have consequences for DNA donors, Rodriguez says. Federal law prohibits health insurance companies from using a person's genetic data, "but many people worry that the law does not go far enough," she says. For example, there is nothing stopping companies from using genetic data to determine policies for life insurance and long-term disability care.

As genealogy databases and other resources improve, "the reidentification of existing data sets will become easier," Church says. But he and Rodriguez hope that the scientific community will not react by clamping down. "There are enormous benefits to sharing research data," Rodriguez says.

–JOHN BOHANNON

CREDIT: G. GRULLÓN/*SCIENCE*

Published by AAAS