So many people have had their DNA sequenced that they've put other people's privacy in jeopardy

October 16 2018, by Deborah Netburn, Los Angeles Times

Credit: CC0 Public Domain

Everyone's DNA sequence is unique. But for those who wish to maintain their genetic privacy, it may not be unique enough.

A new study argues that more than half of Americans could be identified

by name if all you had to start with was a sample of their DNA and a few basic facts, such as where they live and how about how old they might be.

It wouldn't be simple, and it wouldn't be cheap. But the fact that it has become doable will force all of us to rethink the meaning of privacy in the DNA age, experts said.

There is little time to waste. The researchers behind the new study say that once 3 million Americans have uploaded their genomes to public genealogy websites, nearly everyone in the U.S. would be identifiable by their DNA alone and just a few additional clues.

More than 1 million Americans have already published their genetic information, and dozens more do so every day.

"People have been wondering how long it will be before you can use DNA to detect just about anybody," said Ruth Dickover, director of the forensic science program at the University of California, Davis who was not involved with the study. "The authors are saying it's not going to take that long."

This new reality represents the convergence of two long-standing trends.

One of them is the rise of direct-to-consumer genetic testing. Companies such as Ancestry.com and 23andMe can sequence anyone's DNA for about $100. All you have to do is provide a sample of saliva and drop it in the mail.

The other essential element is the proliferation of publicly searchable genealogy databases like GEDmatch. Anyone can upload a full genome to these sites and powerful computers will crunch through it, looking for stretches of matching DNA sequences that can be used to build out a

family tree.

To test the growing power of these sites, researchers led by Columbia University computer scientist Yaniv Erlich set out to see whether they could find a person's name—and thus, his identity—if all they had to go on was a piece of his DNA and a small amount of biographical information.

They started with a full DNA sequence from a person whose genetic information was published anonymously as part of an unrelated scientific study. (They had actually identified this woman in a previous study, but for the purposes of this work, they pretended they didn't know who she was.)

Erlich and his collaborators uploaded her genetic code to GEDmatch and ran a search to see if she had any relations on the site. They found two: one in North Dakota and one in Wyoming.

The researchers could tell they were all related because they shared a number of single nucleotide polymorphisms, or SNPs. These are single letters in specific spots among the roughly 3 billion A's, Cs, Ts and Gs that make up the human genome.

The more SNPs people share, the more closely related they are.

By comparing the DNA of all three relatives, Erlich's team was able to find a common ancestral couple that were the Utah woman's great-grandparents.

Next, the researchers scoured genealogical websites and other sources for additional descendants of that long-ago couple. They found 10 children and hundreds of grandchildren and great-grandchildren.

Then they started culling their massive list of descendants. They eliminated all the men from the sample, then those who were not alive when the Utah woman's DNA was sequenced. The authors also knew that their subject was married and how many children she had, which helped them zero in on their target.

After a long day of painstaking work, they researchers were able to correctly name the owner of the DNA sample.

The authors said the same process would work for about 60 percent of Americans of European descent, who are the people most likely to use genealogical websites, Erlich said. Though the odds of success would be lower for people from other backgrounds, it would still be expected to work for more than half of all Americans, they said.

To come to this conclusion, the researchers analyzed a different database consisting of 1.28 million anonymous individuals who had their DNA sequenced by MyHeritage, a DNA testing and family history company where Erlich is the chief science officer.

If you can find a person's third cousin in a genealogical database, then you should be able to identify the person with a reasonable amount of sleuthing, Erlich said. So the team checked to see how many relatives on the order of a third cousin or closer they could find for each individual in their data set.

They found plenty: 60 percent of the 1.28 million people were matched with a relative who was at least as close as a third cousin, and 15 percent had a relative who was at least as close as a second cousin.

The findings were published Thursday in the journal *Science*.

So far, 72-year-old Joseph James DeAngelo is the most famous person

to be identified this way. You may know him better as the suspected Golden State Killer, charged with 13 counts of murder and 13 counts of attempted kidnapping.

When law enforcement officials used a publicly accessible DNA database to catch DeAngelo in April, it was only the second time in crime-solving history that the strategy was implemented successfully.

Since then, at least 13 additional suspected criminals have been identified in the same way.

"The solving of the Golden State Killer case opened this method up as a possibility and other crime labs are taking advantage of it," Dickover said. "Clearly a trend has started."

Private citizens are benefiting from the technology as well. Adoptees have found biological parents and siblings, and others have found distant cousins who can shed new light on a family's origins and heritage.

But as more of us upload DNA to publicly searchable databases, the implications can be creepy.

"When the police caught the Golden State Killer, that was a very good day for humanity," Erlich said. "The problem is that the very same strategy can be misused."

Think of foreign governments using this technique to track down American citizens, he said. Or protesters and activists being pursued in this way.

Erlich and his co-authors proposed a mitigation strategy that would make it harder to upload an unknown DNA sequence to a genealogical database and search for a match.

They suggest that direct-to-consumer DNA testing companies put a special code on the raw data files they send to their customers. Genealogy sites could then agree to allow people to upload DNA sequences only if they have a valid code. This would ensure that people could conduct searches related only to their own DNA.

A system like this would not prevent law enforcement from using genealogical databases to search for suspects, Erlich said.

The ultimate goal is to allow people to use their DNA to find out more about their own families without sacrificing their privacy, Erlich said.

Just this year, his adopted cousin found a biological sister who lives halfway around the world, he said.

"This is why we have this technique," he said.

More information: Y. Erlich el al., "Identity inference of genomic data using long-range familial searches," *Science* (2018). science.sciencemag.org/lookup/ … 1126/science.aau4832

Citation: So many people have had their DNA sequenced that they've put other people's privacy in jeopardy (2018, October 16) retrieved 28 June 2024 from https://phys.org/news/2018-10-people-dna-sequenced-theyve-privacy.html