Computational model links family members using genealogical and law-enforcement databases

October 11 2018

Credit: CC0 Public Domain

The notion of using genetic ancestry databases to solve crimes recently crossed from hypothetical into credible when police used an online genealogical database to track down the alleged Golden State Killer, a serial criminal who terrorized much of California in the 1970s and

1980s. Now, in a study published October 11 in *Cell*, researchers are reporting ways in which that type of inquiry could potentially be expanded.

Specifically, they have published a computational method for linking individuals in ancestry databases to those in law-enforcement databases. These two databases use completely different systems of genetic markers. The investigators report in a proof of principle with 872 people that for close relatives—either sibling or parent-offspring pairs—more than 30% can be accurately matched with the correct relative using nonoverlapping genetic markers from the two different databases.

"There's a legacy problem in that so many DNA profiles have been collected with this older genetic marker system that's been used by law enforcement since the 1990s. The system is not designed for the more challenging queries that are currently of interest, such as identifying people represented in a DNA mixture or identifying relatives of the contributor of a DNA sample," says senior author Noah Rosenberg, a biology professor at Stanford University. "In this study, we were trying to pose the question of whether a newer, more modern system of genetic markers could be tested against the old system and still get matches and find relatives."

The database used by the FBI and other law-enforcement agencies is known as the Combined DNA Index System (CODIS). It relies on short tandem repeat (STR) markers, a type of copy-number variation, in noncoding regions of the DNA. (The system originally used 13 markers; it recently was updated and now includes 20.) By contrast, ancestry databases look for differences in single-nucleotide polymorphisms (SNPs) across hundreds of thousands of sites in the genome.

In a study published last year, Rosenberg's team reported that software could match individuals who appeared in both databases even with

genotype datasets that had no shared markers. They matched more than 90% of people using the 13-marker version of CODIS and up to 99% with 20 markers. The key idea is that each STR marker is surrounded by SNPs that are typically inherited together with the STR. As a result, a person's genotypes for those SNPs can partially predict the genotype of the neighboring STR and vice versa. When these subtle correlations are accumulated across many STRs, it becomes possible to match an SNP profile with an STR profile.

The new paper built on that research by looking at whether the same approach would work in connecting close family members. They found that when one individual had been analyzed for STR markers and the other for SNP markers, about 30%-32% of parent-offspring pairs and 35%-36% of sibling pairs could be linked.

In the Golden State Killer case, law enforcement submitted DNA collected from one of the crime scenes for SNP genotyping, then used an open-source ancestry database to link that profile with other individuals who were present in the database. But the technique reported in the new paper suggests that familial searches might be possible to perform linking people in CODIS to relatives in an ancestry database or vice versa.

The study was intended to provide data for discussing many of the issues surrounding forensic genetics and genomic privacy, Rosenberg explains. "We wanted to examine to what extent these different types of databases can communicate with each other," he says. "It's important for the public to be aware that information between these two types of genetic data can be connected, often in unexpected ways."

When current policies surrounding DNA evidence were established, it wasn't possible to make this connection. "We have shown that the investigative reach of forensic STR profiles might be possible to expand

beyond what was previously believed to be the limit," he adds.

In the paper, the researchers note other policy-relevant issues surrounding this expanded capability. For example, certain populations are overrepresented in law-enforcement STR databases. Expanding the use of database searches could change the calculation about who is accessible to investigators from the profiles in those databases. "There has already been a lot of legal analysis on how STR databases are used," Rosenberg says. "With this study, we suggest that SNP databases and their links to STR databases should also be considered in that analysis."

The new findings have applications for other areas of study beyond law enforcement. For example, ecologists studying organisms in the field could use this approach to determine whether animals living in a particular geographic site descended from animals whose DNA had been collected on a previous sampling trip even if only STR data is available from the older samples. The linkage tools also could potentially be used to link DNA fragments from ancient humans with each other—for example, when multiple samples are tested from an ancient burial site.

More information: *Cell*, Kim et al. "Statistical detection of relatives typed with disjoint forensic and biomedical loci." www.cell.com/cell/fulltext/S0092-8674(18)31180-2 , DOI: 10.1016/j.cell.2018.09.008

Provided by Cell Press