Genetic privacy and the case of the Golden State Killer—diving into the science

Consumer DNA testing companies are rushing to reassure customers about the security of their genetic information following news that DNA data from a genealogy website was used by police to arrest the man they believe is the Golden State Killer, responsible for at least 12 murders and more than 50 rapes from 1976 to 1986.

But are DNA data ever really private?

The companies say they will not provide DNA information unless compelled to by court order—but consumers can post it. And it’s now possible to identify someone by comparing the short DNA repeats tracked in forensic genotyping to the genome-wide SNP (single nucleotide polymorphism) markers that form the basis of many health and trait tests, with mathematical correlations that arise from matching flanking DNA sequences. A remarkable study published exactly a year ago revealed the strong correlations among different types of genetic markers that are closely linked on their chromosome. (I’ll return to it.)

The different contexts of DNA testing are finally colliding: basic research, health and traits, clinical testing, ancestry and forensics. Technology meets human nature. While people with genetic evidence of a disease risk often want to keep it private, those using DNA data to find relatives want to spread the word. It’s easy to see how Ancestry.com’s “You may provide people with varying levels of access to your DNA results” could lead to unintentional dissemination of private information.

That’s what seems to have led to the capture of Joseph James DeAngelo, 72. Familial DNA searches that nabbed criminals through their imprisoned relatives were controversial enough; now the strategy has spread to consumer DNA testing. And that has many people worried.

The leaks apparently aren’t with the testing companies.

“We do not allow our customer data to be shared or processed by any third party without our customers’ explicit approval,” said Mehdi Maghsoodnia, CEO of Vitagene, whose tests focus on wellness and lifestyle. Prominent direct-to-consumer genetic testing company 23andme issued a statement: “It’s our policy to resist law enforcement inquiries to protect customer privacy.”

The capture of the suspected Golden State Killer was so dramatic that some media accounts glossed over the science. But the story of DNA profiling has been unfolding for decades.

Brief history of DNA profiling
Sir Alec Jeffreys at Leicester University developed DNA typing to solve crimes in 1985. His original technique followed DNA sequences 10 to 80 bases long called “variable number of tandem repeats” (VNTRs). The gene variant (allele) was the number of repeats.

The first famous conviction was that of Colin Pitchfork, a 27-year-old baker and father who raped and strangled two 15-year-old girls in the Leicester countryside in 1983 and 1986. After testing DNA from blood and saliva samples from about 5,000 local men, with no matches, a man in a bar overheard boasting about donating DNA for a friend led to finding the real criminal, and his DNA matched that in semen found in the victims.

The first exoneration from DNA profiling (then called fingerprinting) came in 1993, in the US, when death row inmate Kirk Bloodsworth read about Colin Pitchfork and requested DNA testing of a stain on the panties of a young girl he was convicted of raping and murdering. Another early use of DNA fingerprinting was less disturbing – ensuring that Dolly the sheep was really a clone.

And here’s an interesting coincidence: The accused DeAngelo was a police officer, putting him in position to have insights into DNA fingerprinting in 1986, when the Golden State Killer’s crime spree ended. (In June 1988, I wrote the cover story for Discover magazine, “DNA Fingerprints: New Witness for the
Prosecution, about a rapist who struck near Disney World and was done in by his DNA. He kept a copy of the magazine in prison and wrote to the editor, wanting to contact me.)

Forensic scientists began to use shorter VNTRs 2 to 10 bases long, called short tandem repeats (STRs), in the early 1990s. The smaller DNA pieces were more likely to persist in the extreme conditions of a disaster like a fire or bomb blast and were more accurate in distinguishing individuals. The FBI developed a panel of 13 STRs, generating 26 data points because a person inherits one from each parent, which may be identical or not. In addition, forensic genotyping determines sex from a gene, amelogenin, on the X and Y chromosomes. It is 6 bases shorter in a female.

The DNA Identification Act of 1994 declared DNA data confidential and restricted access to criminal justice agencies and defendants. The act also called for stripping of information that could identify an individual in other ways, like a distinctive mutation or genotype for an observable trait.

The language is DNA, the power is statistics

STRs are non-coding DNA sequences, which means that they do not encode protein, although some STRs are within protein-encoding genes. The STRs used in forensics are 3 or 4 bases long, like CATCCATCCATC. In contrast, a SNP alters one base, such as CATCCATC compared to CGTCCATC.

px Codis profile

The power of DNA profiling lies in its combinatorial nature. Each STR harbors a variable number of the repeats – a suspect might have 5 copies of the STR on chromosome 2 from his father and 7 from his mother. The allele tables for the STRs in this FBI document indicate the many possible combinations.

The FBI also considers the frequency of each allele — the number of copies at each STR site — in distinct population groups, such as Trinidadians and Apache. In its early days the technology used too-broad groupings. By upping variability while restricting the populations, analysis of a bit of DNA left on a coffee cup or under a fingernail can go a long way towards eliminating suspects and narrowing the search for the culprit.
In 2017, the FBI’s Combined DNA Index System, or CODIS, which shares DNA profiles among local, state, and federal crime laboratories, added seven more STR “core loci,” making an already formidable forensic tool even more so. (A locus is the site of a DNA sequence as part of a chromosome.)

CODIS maintains a database of more than half a million unidentified DNA samples from crime scenes, which is where investigators initially turn if they don’t find a match in samples that already have a name attached. If that doesn’t work, a controversial strategy is a familial search, looking for incarcerated individuals who share many of the 20 STRs with the crime scene sample. In California, a familial match requires at least 15 of the 20 CODIS core loci.

Follow the latest news and policy debates on sustainable agriculture, biomedicine, and other ‘disruptive’ innovations. Subscribe to our newsletter.

**SIGN UP**

**Unusual uses of DNA profiles**

Until now, the most famous familial DNA match case was that of another notorious California criminal, the “Grim Sleeper,” who killed nine women and one girl, over many years. Finally, a DNA sample from a victim closely matched DNA of an imprisoned weapons trafficker whose father turned out to be the Grim Sleeper, Lonnie David Franklin Jr. DNA extracted from a tossed piece of pizza nailed him — discarded DNA raises no privacy issues. More victims’ identities appeared in the man’s computer files. Familial searches may lead to false accusations and much _angst_. But familial searches usually zero in on one individual.

Another important precedent was the 2013 case of Alonzo Jay King, who was arrested in Wicomico County, Maryland, for “menacing a group of people with a shotgun.” State law compelled taking a cheek swab for DNA testing. When the sample matched DNA evidence from an unsolved 2003 rape case, King was tried and convicted for the older crime based on new DNA evidence.
case, was a new situation, and the Supreme Court eventually overruled an appeals court determination of unlawful seizure and unreasonable search. The court called the taking and use of a DNA profile a “legitimate police booking procedure” equivalent to obtaining conventional fingerprints.

But the Golden State Killer case goes even beyond the exceptional circumstances of Maryland v. King: The DNA that points to DeAngelo didn’t come from a forensic DNA database, but from a consumer website.

The prosecution found a distant relative of the killer at GEDmatch.com, a free website where people in search of relatives upload their autosomal DNA ancestry data from companies such as AncestryDNA, 23andMe, Family Tree DNA, MyHeritage DNA, Living DNA, or for health traits from WeGene, GenetiConcept or Genes for Good. No warrant was necessary, said lead investigator Paul Holes.

Once registered at GEDmatch.com, DNA from a Golden State Killer crime scene matched DNA from a distant relative of the man who would become the suspect. Checking more STR sites led to closer male relatives. Investigators considered practical clues among the relatives of the GEDmatch, like DeAngelo’s experience in law enforcement. The noose was tightening. Finally, DNA on a discarded object found near DeAngelo’s home in Citrus Heights matched. An answer emerged from the shared STR genotypes of amateur genealogist relatives, crime scene evidence, and recently discarded DNA.

**Will DNA database crosstalk kill genetic privacy?**

Identifying individuals from their DNA sequences stored in databases isn’t a new concern. A prescient study published in Science, “Identifying Personal Genomes by Surname Inference,” from 2013, led by then-grad student Melissa Gymrek, now at UCSD, interrogated the 1000 Genomes Project database, compiled from 2008-2015. The informed consent for the project read, “...it will be hard for anyone to find out anything about you personally from any of this research.”

Well, no.

Gymrek was a student of Yaniv Erlich, a researcher at the Whitehead Institute who had worked with databases at financial banks. She cross-referenced STR profiles and surnames from genetic genealogy websites, public information such as state of residence and birth year, mutation databases naming children with genetic diseases along with their hometowns and dates of birth, and DNA sequence information stored at a cell bank.

The strategy identified 50 people so easily that Dr. Erlich, alarmed, notified the NIH, the sponsor of the 1000 Genomes Project, which then stepped up de-identifying efforts.
Then a year ago came the study from Noah Rosenberg, Professor of Population Genetics and Society at Stanford and his team. They looked at the original 13 CODIS core loci and 642,563 genomewide SNPs for 872 people, and for 98% of them, could correlate one type of test result to the other. When they included the additional 7 STRs in the updated set of 20 used in forensic DNA typing, matching accuracy neared 100%.

A genome-wide association study considers patterns of SNPs.

STRs and SNPs weren’t intended to correspond, like translating from French to Spanish instantaneously online. But over the years, genome sequencing has revealed the DNA information surrounding the STRs – a little like being able to distinguish Los Angeles from Chicago by identifying landmarks in their suburbs. A physical explanation for the correlations is based on classical genetic linkage: Areas bordering STRs can enter the real estate of the SNP maps.

Gymrek’s clever experiment in a way foreshadowed the Stanford team’s finding. In short, the new CODIS panel, plus more specific markers added once a familial match appears, “uncovers new risks to privacy,” they write. Matching CODIS and SNP genotypes indicate the same person, and the problem is that the SNP profiles that underlie so many DTC tests can reveal much about a person’s health risks and traits.

The testing websites shy away from the accurate term for these SNP profiles, and understandably so. “Genomewide association studies (GWAS) based on single nucleotide polymorphisms” isn’t going to sell test kits. And it takes scrolling through the pretty graphics and emotional success stories to get to a vague phrase like 23andMe’s “specific locations in your DNA” or Vitagene’s “700,000 markers.”

Family Tree DNA allows customers to select the number of autosomal SNP markers, combined with the Y chromosome and mitochondrial DNA sequences used to trace geographical origins, and offers the ability to upload data from other testing companies. Will the data-sharing promiscuity of the companies, welcome in that it helps find relatives, at the same time add to the privacy problem?
Mother’s Day or graduation gift. The arrest of the suspected Golden State Killer has rendered the New Yorker’s flippant description of “spit parties” to collect DNA, from a decade ago, no longer funny.

Crosstalk between consumer DNA test sites and forensic investigations can go both ways. “The genotypes of the genetic markers most commonly used by law enforcement are correlated enough with genotypes of markers used in personal genomics, genealogical, and biomedical studies that it will sometimes be possible to connect genotypes with these different markers to the same person,” explained Dr. Rosenberg. 23andMe will have to update their blog post from March 16, 2016, which stated that their “tests are of little use to law enforcement or government because they cannot technically be matched against the information in CODIS or other governmental databases.” Now they can.

Could someone sending in a spit sample to test for ear wax consistency, Neanderthal genes and cancer risk discover that her DNA is being used to out a creepy old uncle? Or will prisoners be informed of a late-onset inherited disease like Huntington’s looming? The possible scenarios are endless.

In the meantime, let’s inject some common sense. Caveat emptor: Let the buyer (or uploader) beware. A company’s pledge to require a court order before providing DNA data is useless if people talk and their genetic info ricochets all over the Internet. One company’s promise to “destroy your physical DNA saliva sample after it has been analyzed,” hardly helps if the DNA sequence information is in their database. The capture of Colin Pitchfork at the dawn of the DNA profiling era, thanks to a blabbermouth friend in a bar, may have set a precedent for how not to maintain genetic privacy.

Using GEDmatch to close in on and capture suspect Joseph James DeAngelo was genius. But consumers concerned about DNA-based identification should recognize two things: data on a public platform are no longer private, and DNA sequences by their nature provide information on relatives.

“The privacy discussion surrounding this recent case is helpful in giving the public a clearer understanding of the limits of DNA privacy. Because of correlations between genotypes at different markers, DNA databases on different sets of genetic markers can ‘talk’ to one another more than has been appreciated,” Dr. Rosenberg told Genetic Literacy Project.

So think twice, maybe three times, before sending off that precious DNA sample.

Ricki Lewis is the GLP’s senior contributing writer focusing on gene therapy and gene editing. She has a PhD in genetics and is a genetic counselor, science writer and author of The Forever Fix: Gene Therapy and the Boy Who Saved It, the only popular book about gene therapy. Follow her at her website or Twitter @rickilewis.