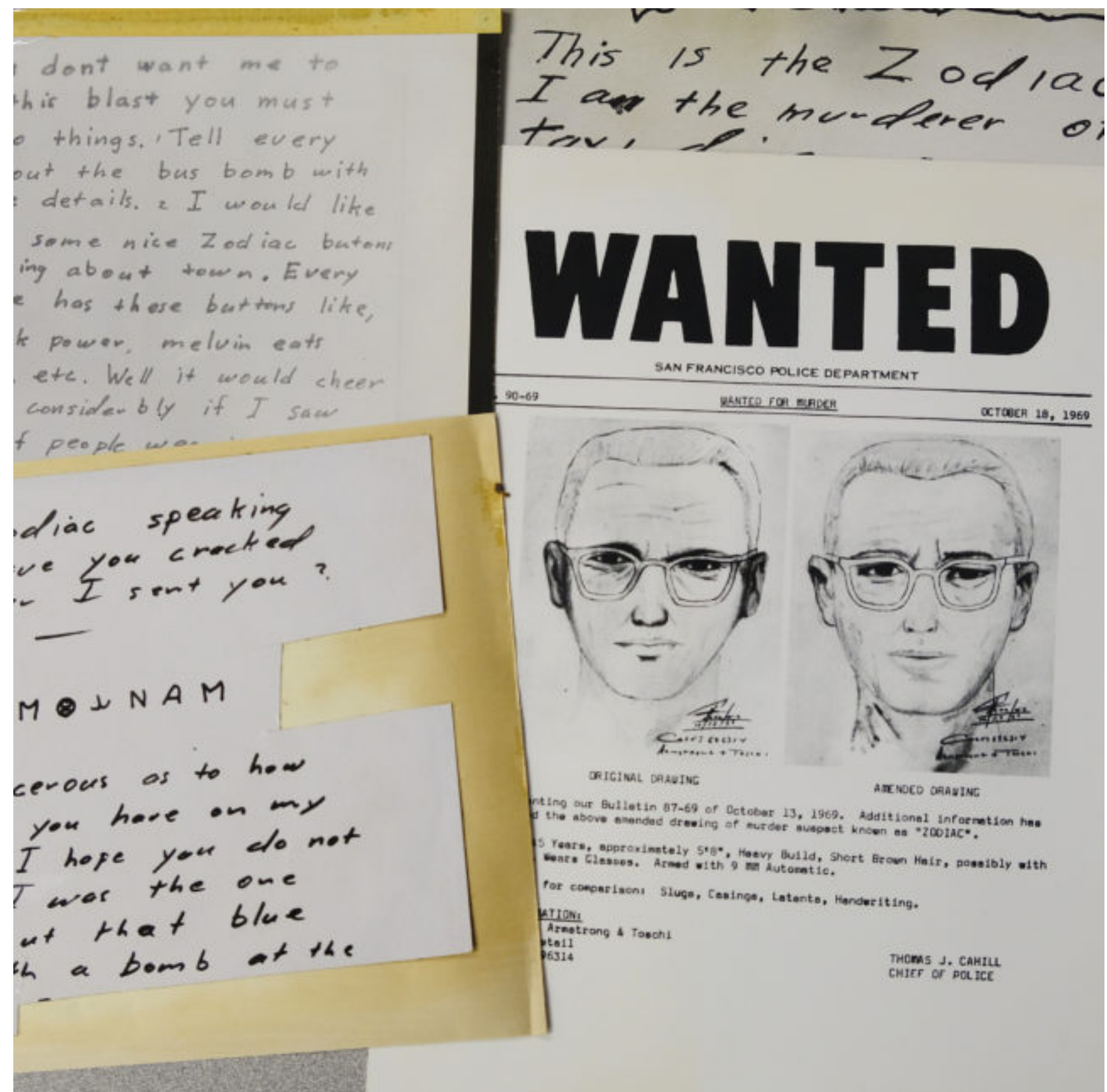# 'We are increasingly exposed': New studies show how easy it is to identify people using genetic databases

By Rebecca Robbins @rebeccadrobbins

October 11, 2018

A San Francisco Police Department wanted bulletin and copies of letters sent to the San Francisco Chronicle by a man who called himself Zodiac. Detectives in Northern California are trying to get a

In recent months, consumer genealogy websites have unleashed a revolution in forensics, allowing law enforcement to use family trees to track down the notorious Golden State Killer in California and solve other cold cases across the country. But while the technique has put alleged killers behind bars, it has also raised questions[11] about the implications for genetic privacy.

According to a pair of studies published Thursday, your genetic privacy may have already eroded even further than previously realized.

In an analysis published in the journal Science[22], researchers used a database run by the genealogy company MyHeritage to look at the genetic information of nearly 1.3 million anonymized people who've had their DNA analyzed by a direct-to-consumer genomics company. For nearly 60 percent of those people, it was possible to track down someone whose DNA was similar enough to indicate they were third cousins or closer in relation; for another 15 percent of the samples, second cousins or closer could be found.

And when you use narrowing factors like age and geographical location to zoom in on a person of interest, over 60 percent of Americans of European ancestry — the population with the most representation in these databases — can have their identities revealed this way, the researchers concluded.

Yaniv Erlich, the lead author on the Science paper, said his team's findings should prompt regulators and others to reconsider the assumption that genetic information is de-identified. "It's really not the case. At least technically, it seems feasible to identify some significant part of the population" with such investigations, said Erlich, who's a computer scientist at Columbia University and chief science officer at MyHeritage.

The Science paper counted 12 cold cases that were solved between April and August of this year when law enforcement turned to building family trees based on genetic data; a 13th case was an active investigation.

The most famous criminal identified this way: the Golden State Killer, who terrorized California with a series of rapes and murders in the 1970s and 1980s. With the [help of a genetic genealogist](#)[33], investigators uploaded a DNA sample collected from an old crime scene to a public genealogy database, built family trees, and tracked down relatives. They winnowed down their list of potential suspects to one man with blue eyes, and in April, they made the landmark arrest.

To crack that case, the California investigators used GEDmatch, an online database that allows people who got their DNA analyzed by companies like 23andMe and Ancestry to upload their raw genetic data so that they can track down distant relatives. MyHeritage's database — which contains data from 1.75 million people, mostly Americans who've gotten their DNA analyzed by MyHeritage's genetic testing business — works similarly, although it explicitly prohibits forensic searches. (23andMe warns users about the privacy risks of uploading their genetic data to such third party sites.)

Erlich called on the industry to add a cryptographic signature — a string of encrypted characters amounting to gibberish— to these raw genetic data files that consumers can download. The idea would be to use this key to identify genetic data files that were generated properly from companies — and weed out those that have been tampered with, or produced from samples collected at crime scenes.

None of the major companies use such a signature right now, and it would only work if all the companies in the industry agreed to do it, Erlich said.

In the separate study published in the journal Cell[55], researchers demonstrated a computational method to bridge two types of databases that are generally thought to be speaking in different languages. On one hand, there are consumer genomics tests, which analyze genetic markers known as single nucleotide polymorphisms. Then there's the database used by the FBI and other law enforcement agencies, which looks at flags in DNA known as short tandem repeat markers.

But with the right computational tools, the researchers were able to take a pool of 872 people and identify a sibling or a parent or child in more than 30 percent of cases across the two different databases.

That's not necessarily something that a consumers expect when they send away for a spit kit, said Jun Li, a geneticist at the University of

Michigan who was co-author on the Cell study.

When people sign consent forms to take consumer DNA tests or participate in biomedical research, they "sometimes are not fully informed about how this enlarges their public profile footprint," Li said. "Quantitatively, it's not too hard to reveal their identity well beyond the initially intended consent."

Li added: "By degrees, we are increasingly exposed."

But Dr. Robert Green, a medical geneticist at Harvard and Brigham and Women's Hospital, cautioned against overreacting to the new studies.

"For me, these articles are fascinating and important and we shouldn't shy away from the privacy concerns that these articles raise. But at the same time, we should keep in mind the personal and societal value that we believe that we are accruing as we make these large collections," said Green, who was not involved in the new studies and is an adviser for genomics companies including Helix and Veritas Genetics.

He pointed to the potential of genomics not only to reunite family members and put criminals behind bars, but also to predict and prevent heritable diseases and develop new drugs.

As with using social media and paying with credit cards online, reaping the benefits of genetic testing requires accepting a certain level of privacy risk, Green said. "We make these tradeoffs knowing that we're trading some vulnerability for the advantages," he said.

About the Author