Law Enforcement Access to Recreational Genomics Databases

Bradley Malin, Ph.D.

Vanderbilt University Medical Center
Co-Director, Center for Health Data Science
Co-Director, Center for Genetic Privacy & Identity in Community Settings

Vanderbilt University
Accenture Professor of Biomedical Informatics, Biostatistics, & CS
Co-Director, Big Biomedical Data Science Training Program

September 27, 2021

Recreational Genomics

23andMe SHOP LEARN SIGN IN REGISTER KIT HELP Shop

Exclusive Offer: Bu

80% receive a report with
meaningful genetic varia

Ancestry Service

Know your world from the inside.

AncestryDNA®

Discover where your family is from without even leaving your living room.

Only $99

Buy now

DNA + Family Tree Bundle

Uncover your whole family story with an AncestryDNA® kit and a World Explorer Membership.

Only $100*

Buy now

AncestryDNA® Traits

All that AncestryDNA® offers, plus unlock personal traits your genes could influence.

Only $119

Buy now

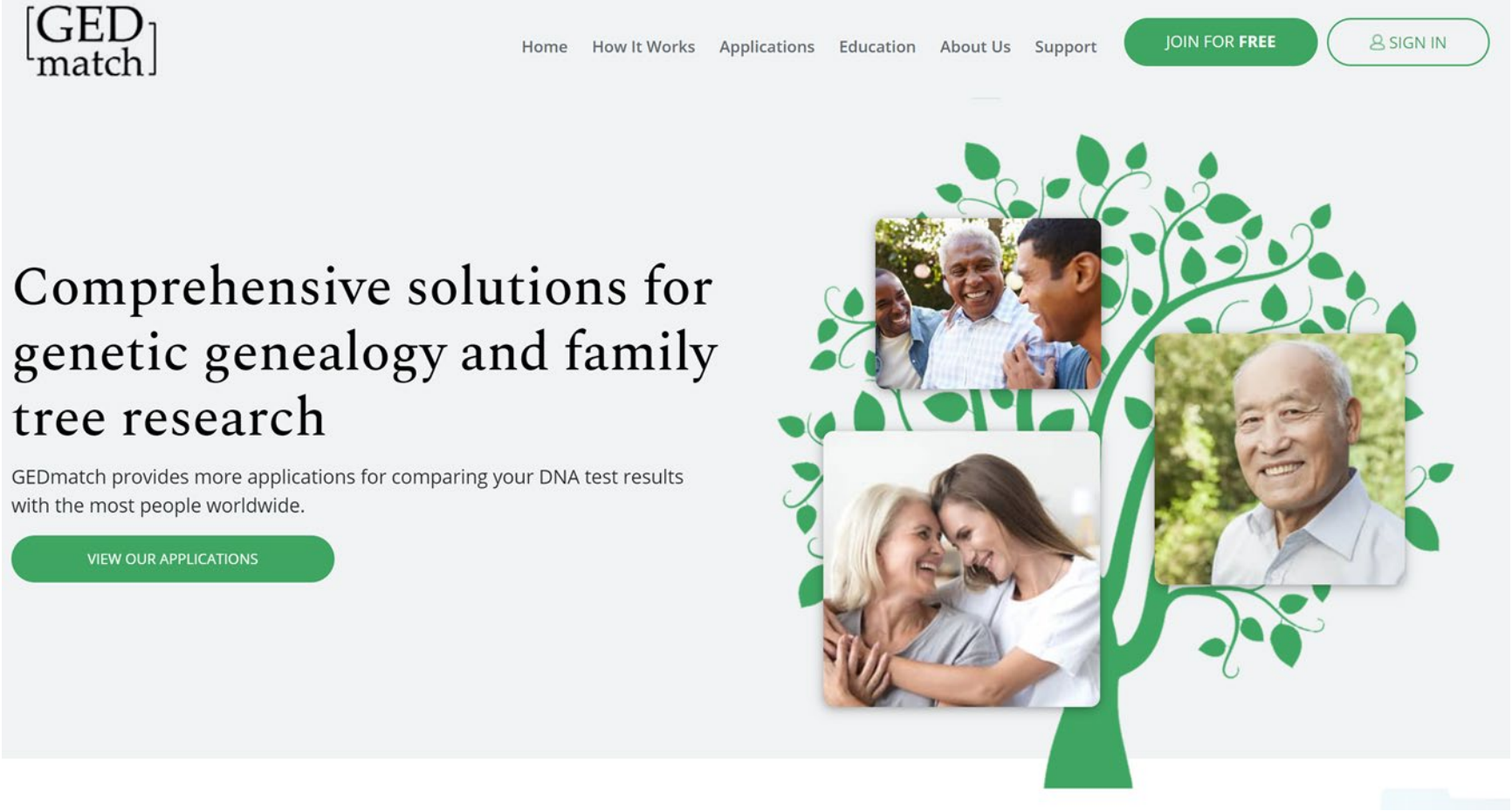
Learn more

*Offer is for new and returning subscribers only and not for renewal of current subscriptions. Your subscription will automatically renew at $79.95 every three months after the introductory 3-month offer. If you don't want to renew, cancel at least two days before your renewal date by visiting the My Account section or by contacting us. See our Renewal and Cancellation Terms for further details.

AncestryDNA® ethnicity estimates are now more precise than ever.

We've applied our latest cutting-edge science to AncestryDNA® ethnicity results. Take a look at your updated results and get a more precise picture of just where your ancestors came from.

Share to Discover



True Crime

The ingenious and 'dystopian' DNA technique police used to hunt the 'Golden State Killer' suspect

Police used DNA information on a genealogy website to track down the Golden State Killer suspect. (Reuters)

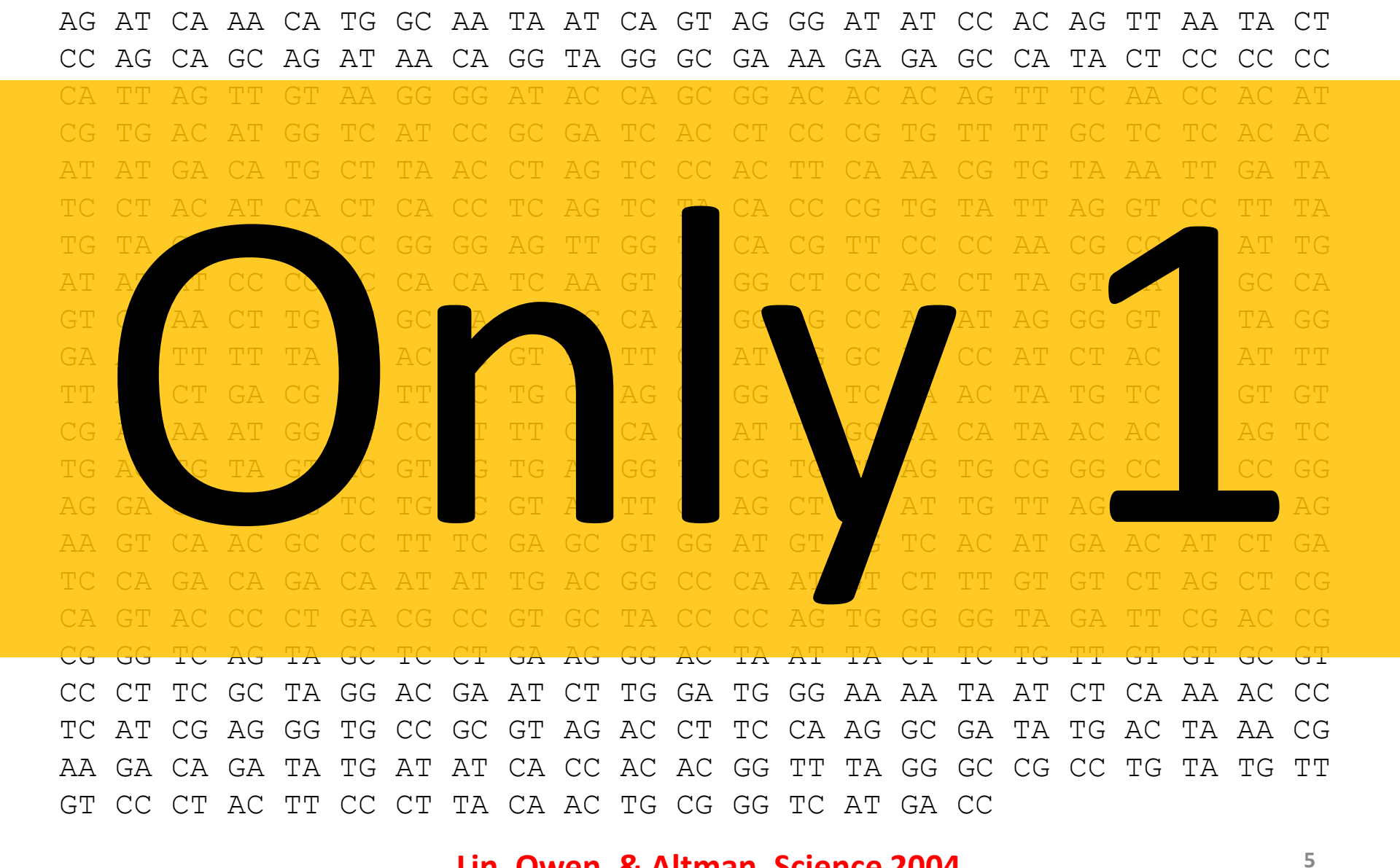
By Avi Selk

April 28, 2018

For decades, police say, the DNA of the "Golden State Killer" sat in evidenc
genetic fingerprint that could identify definitively the man who killed 12 people and raped 45 women across California between 1976 and 1986.

And for decades, those samples were basically useless to investigators, who ran into the same wall that has frustrated police since the invention of DNA forensics: A genetic fingerprint is not much

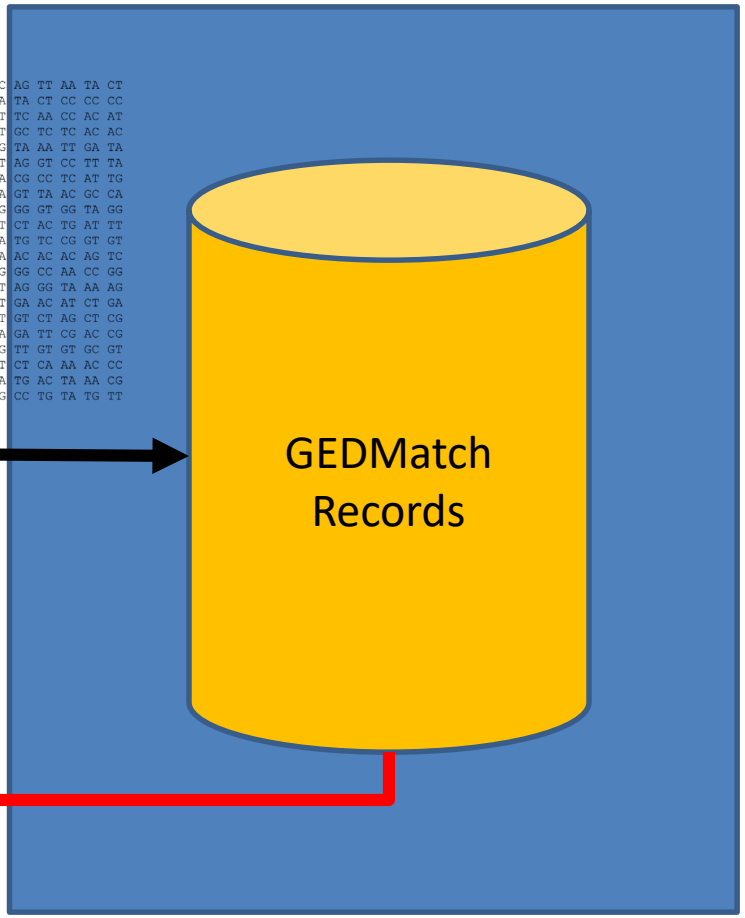
*omics Data is High-Dimensional!



Lin, Owen, & Altman. Science 2004.

Knocking on the Front Door

AG AT CA AA CA TG GC AA TA AT CA GT AG GG AT AT CC AC AG TT AA TA CT
CC AG CA GC AG AT AA CA GG TA GG GC GA AA GA GC CA TA CT CC CC CC
CA TT AG TT GT AA GG GG AT AC CA GC GG AC AC AC AG TT TC AA CC AC AT
CG TG AC AT GG TC AT CC GC GA TC AC CT CC CG TG TT TT GC TC TC AC AC
AT AT GA CA TG CT TA AC CT AG TC CC AC TT CA AA CG TG TA AA TT GA TA
TC CT AC AT CA CT CA CC TC AG TC TA CA CC CG TG TA TT AG GT CC TT TA
TG TA CC GA GG CC GG GG AG TT GG TT CA CG TT CC CC AA CG CC TC AT TG
AT AT AT CC CC GC CA CA TC AA GT CT GG CT CC AC CT TA GT TA AC GC CA
GT GA AA CT TG GC GC CA TT CC CA AG GG AG CC AC AT AG GG GT GG TA GG
GA AG TT TT TA CG AC AC GT TG TT CG AT CG GC AC CC AT CT AC TG AT TT
TT AG CT GA CG CA TT GC TG CG AG GC GG TG TC AA AC TA TG TC CG GT GT
CG AC AA AT GG GT CC GT TT CG CA GA AT TT GC TA CA TA AC AC AC AG TC
TG AC TG TA GT AC GT TG TG AA GG TG CG TG CT AG TG CG GG CC AA CC GG
AG GA GG GT CG TC TG CC GT AC TT CC AG CT AT AT TG TT AG GG TA AA AG
AA GT CA AC GC CC TT TC GA GC GT GG AT GT CG TC AC AT GA AC AT CT GA
TC CA GA CA GA CA AT AT TG AC GG CC CA AT GT CT TT GT GT CT AG CT CG
CA GT AC CC CT GA CG CC GT GC TA CC CC AG TG GG GG TA GA TT CG AC CG
CG GG TC AG TA GC TC CT GA AG GG AC TA AT TA CT TC TG TT GT GT GC GT
CC CT TC GC TA GG AC GA AT CT TG GA TG GG AA AA TA AT CT CA AA AC CC
TC AT CG AG GG TG CC GC GT AG AC CT TC CA AG GC GA TA TG AC TA AA CG
AA GA CA GA TA TG AT AT CA CC AC AC GG TT TA GG GC CG CC TG TA TG TT
GT CC CT AC TT CC CT TA CA AC TG CG GG TC AT GA CC



John Smith – likely 1st cousin

…

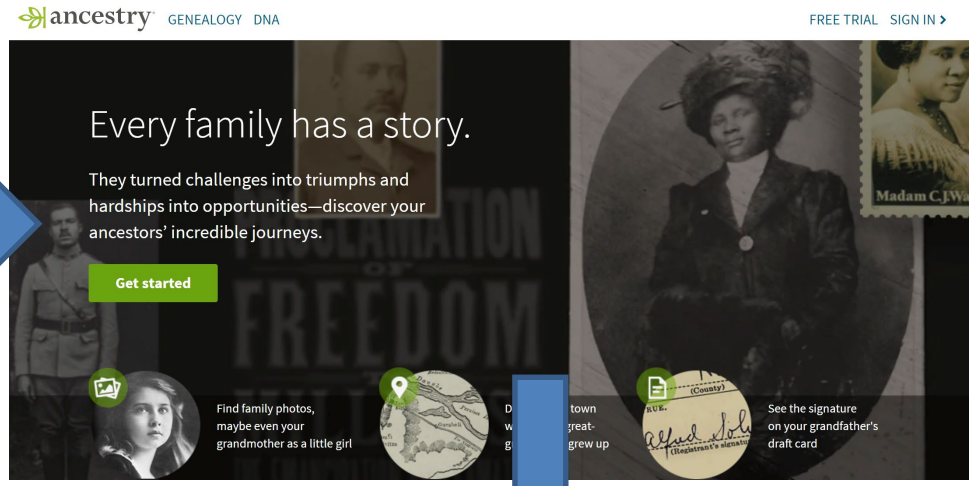
Jane Speith - likely 3rd cousin

The Search is On

John Smith – likely 1st cousin

…

Jane Speith - likely 3rd cousin



Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Surnames are paternally inherited in most human societies, resulting in their co-segregation with Y-chromosome haplotypes (*1*–*5*). Based on this observation, multiple genetic genealogy companies offer services to reunite distant patrilineal relatives by genotyping a few dozen highly polymorphic short tandem repeats across the Y chromosome (Y-STRs). The association between surnames and haplotypes can be confounded by nonpaternity events, mutations, and adoption of the same surname by multiple founders (*5*). The genetic genealogy community addresses these barriers with massive databases that list the test results of Y-STR haplotypes along with their corresponding surnames. Currently, there are at least eight databases and numerous surname project Web sites that collectively contain hundreds of thousands of surname-haplotype records (table S1).

The ability of genetic genealogy databases to breach anonymity has been demonstrated in the past. In a number of public cases, male adoptees and descendants of anonymous sperm donors used recreational genetic genealogy services to genotype their Y-chromosome haplotypes and to search the companies' databases (*6*–*9*). The genetic matches identified distant patrilineal relatives and pointed to the potential surnames of their biological fathers. By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al.* (*10*) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (*11*) empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in these databases and reported that potential surnames can be detected. [CEU participants are multigenerational families of northern and western European ancestry in Utah who had originally had their samples collected by CEPH (Centre d'Etude du Polymorphisme Humain) and were later reconsented to participate in the HapMap project.] However, these surnames could match thousands of individuals, and the study did not pursue full re-identification at a single-person resolution.

Our goal was to quantitatively approach the question of how readily surname inference might be possible in a more general population, apply this approach to personal genome data sets, and demonstrate end-to-end identification of individuals with only public information. We show that full identities of personal genomes can be exposed via surname inference from recreational genetic genealogy databases followed by Internet searches. In all cases in which individuals were studied who had donated DNA samples, the informed consent statements they had signed stated privacy breach as a potential risk and the data usage terms did not prevent re-identification. Representatives of relevant organizations that funded the original studies were notified and confirmed the compliance of this study with their guidelines (*12*).

As a primary resource for surname inference, we focused on Ysearch (www.ysearch.org) and

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. [2]Harvard–Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4]Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX 77030, USA. [6]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel. [7]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. [8]Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel Aviv 69978, Israel. [9]The International Computer Science Institute, Berkeley, CA 94704, USA.

*To whom correspondence should be addressed. E-mail: yaniv@wi.mit.edu

Most White Americans' DNA Can Be Identified Through Genealogy Databases

Only two percent of the population needs to have done a DNA test to identify nearly everyone else, researchers found. Leonard Lessin/Science Source

By **Heather Murphy**

Oct. 11, 2018

The genetic genealogy industry is booming. In recent years, more than 15 million people have offered up their DNA — a cheek swab, some

Right to be Let Alone?

Personal Preference… is clear

- Empirical study with laypeople to gauge societal privacy expectations in use of genetic databases
- Law enforcement access to genetic information considered as, or more, intrusive searches of bedrooms, text messages, or emails
- Regardless of if data is held by health care providers, direct-to-consumer genetic testing companies, or public genealogy websites
- Suggest law enforcement access to non-governmental genetic databases require judicial authorization, although not necessarily a traditional warrant

https://dlj.law.duke.edu/article/a-world-of-difference-law-enforcement-genetic-data-and-the-fourth-amendment-hazel-vol70-iss4/

Let’s Restrict Access

SIGN IN ABOUT US GOT A TIP? SUPPORT US BUZZFEED.COM SECTIONS

TRENDING Missing People Of Color The View COVID Booster Shots Guardianship Industry Boppy Recall Gabby Petito

SCIENCE

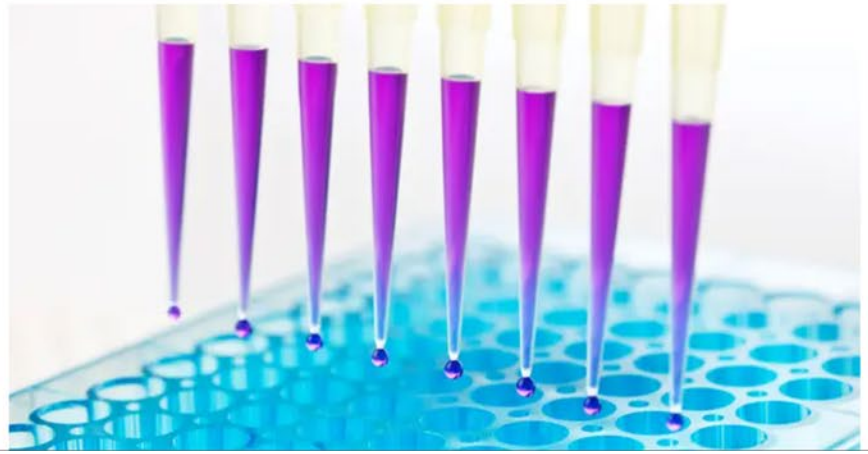
This Genealogy Database Helped Solve Dozens Of Crimes. But Its New Privacy Rules Will Restrict Access By Cops.

GEDmatch’s revamped genetic privacy policy could set off legal battles that go all the way to the US Supreme Court.

Peter Aldhous
BuzzFeed News Reporter

Posted on May 19, 2019, at 3:41 p.m. ET

Tweet Share Copy



Nice Try

- GEDMatch provided an opportunity to consent to the use of data for violent crimes investigations
- Many failed to consent
- Orlando detective's request for warrant for use of all records was approved

The New York Times

Your DNA Profile is Private? A Florida Judge Just Said Otherwise

Privacy experts say a warrant granted in Florida could set a precedent, opening up all consumer DNA sites to law enforcement agencies across the country.

Dion MBD

By Kashmir Hill and Heather Murphy

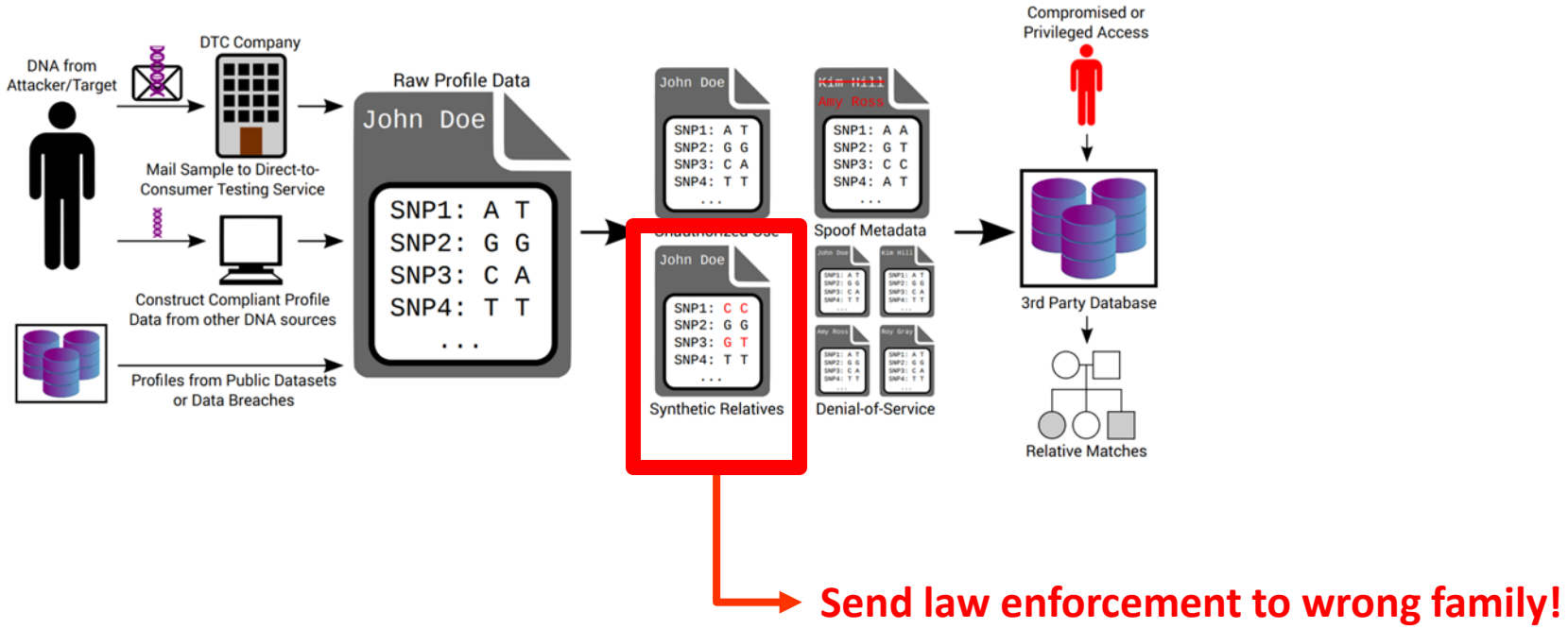
Published Nov. 5, 2019 Updated Dec. 30, 2019

Potential Adversarial Behavior

Computer Security Risks of Distant Relative Matching in Consumer Genetic Databases

Peter M. Ney[1,3,4,5], Luis Ceze[2,4,5], and Tadayoshi Kohno[1,3,4,5]

[1]Security and Privacy Research Lab
[2]Molecular Information Systems Laboratory
[3]Tech Policy Lab



https://arxiv.org/abs/1810.02895

Potential Adversarial Behavior

Computer Security Risks of Distant Relative Matching in Consumer Genetic Databases

Peter M. Ney[1,3,4,5], Luis Ceze[2,4,5], and Tadayoshi Kohno[1,3,4,5]

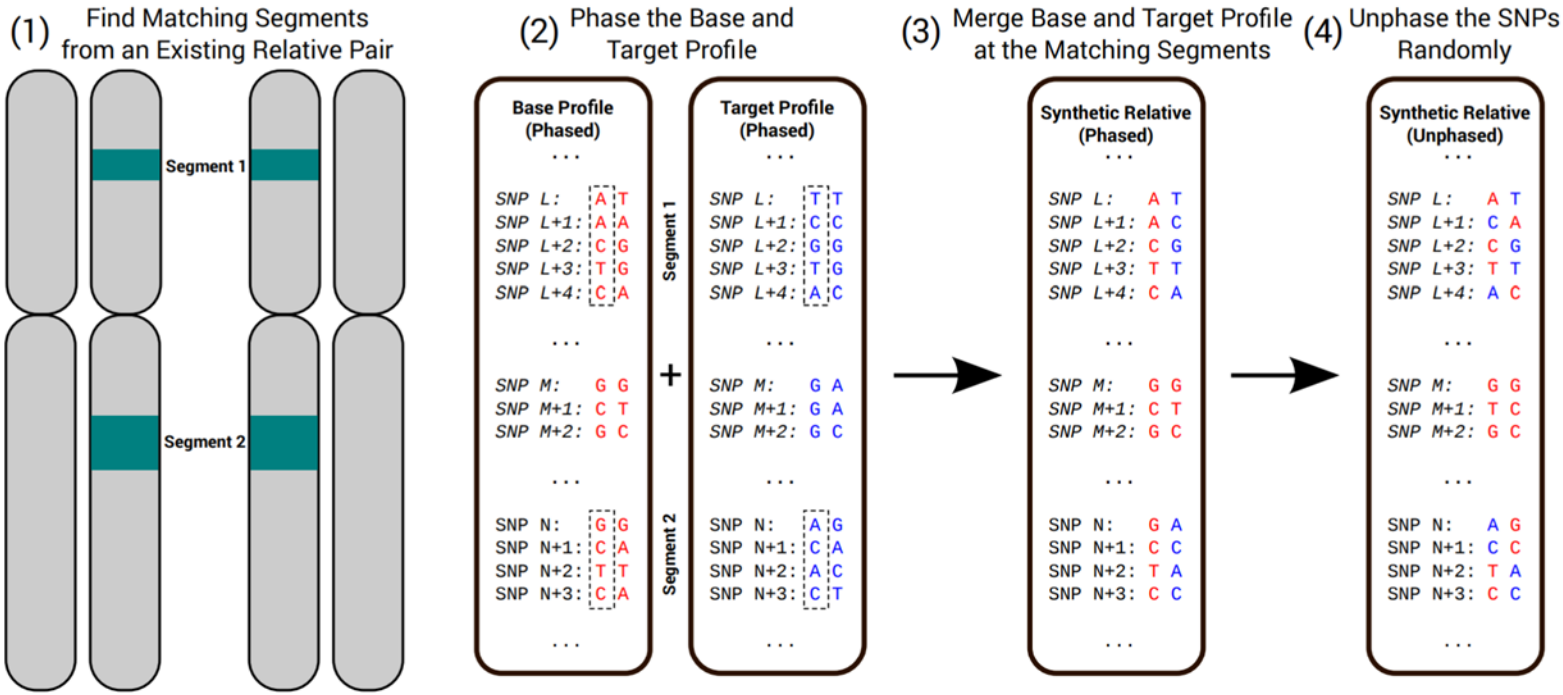


Figure 3: Proposed method to generate synthetic relatives.

https://arxiv.org/abs/1810.02895

So, what do we do?

- States may be leading the way
- MT
  - Requiring government investigators obtain a search warrant before using a consumer DNA database…
  - … unless the consumer has waived the right to privacy
- MD
  - Investigators will need a judge's signoff before using uploading DNA from a crime scene to genealogy websites to find relatives
  - Technique can only for serious crimes (e.g., murder, sexual assault)
  - May only use websites with strict policies around user consent

The New York Times

Two New Laws Restrict Police Use of DNA Search Method

Maryland and Montana have passed the nation's first laws limiting forensic genealogy, the method that found the Golden State Killer.

HudsonAlpha, a genome sequencing lab in Alabama that has worked on more than 1,000 forensic genealogy cases. Wes Frazer for The New York Times

By Virginia Hughes

May 31, 2021

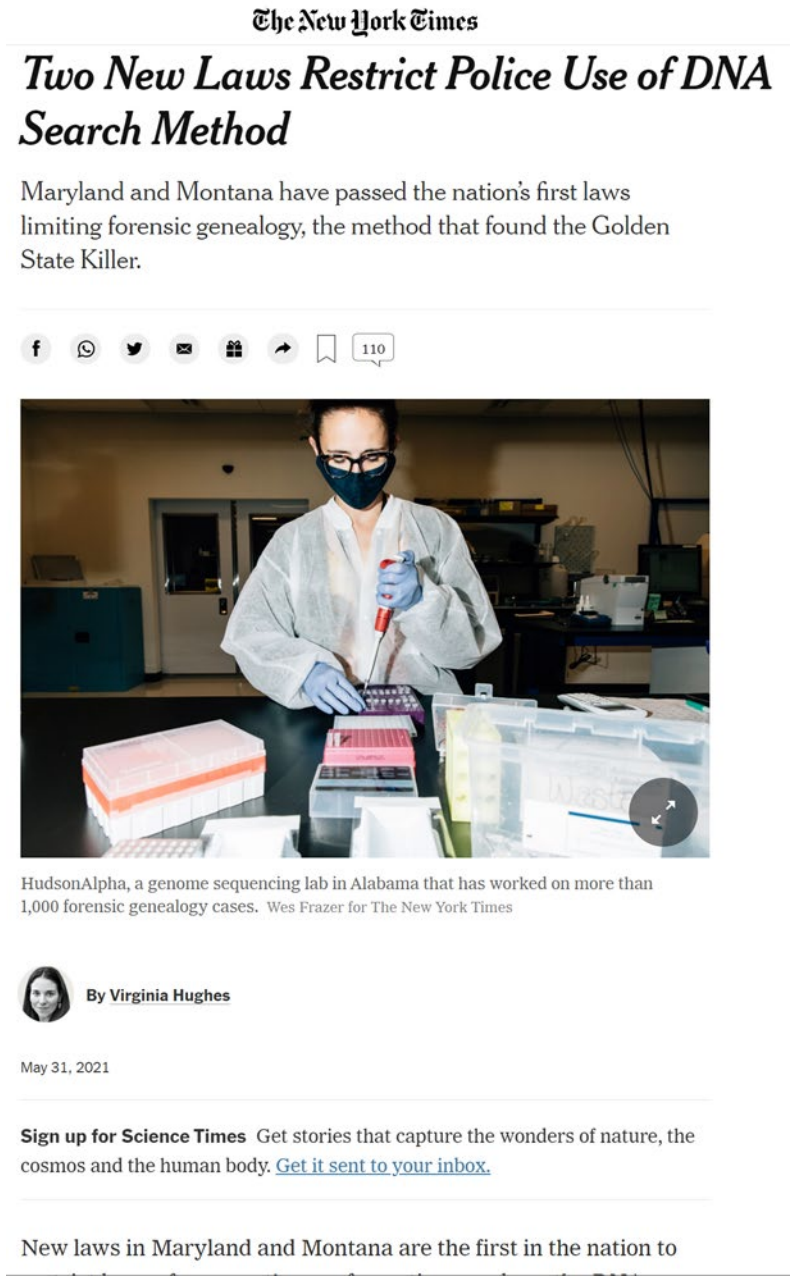
Sign up for Science Times Get stories that capture the wonders of nature, the cosmos and the human body. Get it sent to your inbox.

New laws in Maryland and Montana are the first in the nation to

But Consent?

- There's only so much that consent can support

- As Hazel & Slobogin suggest, may be more appropriate to rely upon
  - Judicial approval
  - Warrant
  - Separate database (from DTC)



The New York Times

Two New Laws Restrict Police Use of DNA Search Method

Maryland and Montana have passed the nation's first laws limiting forensic genealogy, the method that found the Golden State Killer.

110

HudsonAlpha, a genome sequencing lab in Alabama that has worked on more than 1,000 forensic genealogy cases. Wes Frazer for The New York Times

By Virginia Hughes

May 31, 2021

Sign up for Science Times Get stories that capture the wonders of nature, the cosmos and the human body. Get it sent to your inbox.

New laws in Maryland and Montana are the first in the nation to

Questions?

b.malin@vumc.org

Center for Genetic Privacy & Identity in Community Settings

https://www.vumc.org/getprecise

Vanderbilt Center for Health Data Science

https://www.vumc.org/heads