National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

Wednesday, January 26, 2022, 11:00 AM – 4:30 PM (ET)

National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Keynote:
The Protection Perspective

Michael J. Orlando
(National Counterintelligence and Security Center [NCSC])

National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Threats to Genomic Data

Michael J. Orlando, Senior Official Performing the Duties of Director,

National Counterintelligence and Security Center (NCSC)

National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

Wednesday, January 26, 2022, 11:00 AM (EST)

NIST National Institute of Standards and Technology U.S. Department of Commerce

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

National Counterintelligence and Security Center

Mission

- **Lead and support the U.S. Government's counterintelligence and security activities critical to protecting our nation.**
- **Provide counterintelligence outreach to U.S. private sector entities at risk of foreign intelligence penetration.**
- **Issue public warnings regarding intelligence threats to the United States.**

Global Threat Picture

Expanding Array of Adversaries

Improving Capabilities & Tradecraft

Expanded Range of Targets & Operations

Emerging Technologies: A Key Focus of Strategic Competitors

- U.S. leadership in emerging technology sectors -- such as biotech, AI, & quantum -- faces growing challenges from strategic competitors.
- China, Russia, and other nations recognize the economic & military benefits of these technologies & have enacted comprehensive national strategies to achieve leadership in these areas.
- To achieve their strategic goals, strategic competitors are using a wide variety of legal, quasi-legal, and illegal methods to acquire technology, talent, and know-how from the U.S. and other nations.

Foreign Exploitation of Genomic Data is Already Occurring

Bottom photo source: https://baijiahao.baidu.com/s?id=1564669932542581

Exploitation of DNA for Societal Control & Repression by the People's Republic of China (PRC)

- The PRC has conducted large-scale collection of DNA and other biometric data from residents of Xinjiang ages 12 to 65.
- DNA samples, fingerprints, iris scans, and blood types are linked to ID numbers and centralized in searchable database used by PRC authorities to carry out surveillance and detentions.
- Since 2017, between 1 million and 1.8 million Uyghurs & other minorities in Xinjiang have been placed in "re-education" centers.
- Multiple Chinese entities & companies, including two subsidiaries of BGI (the world's largest genomics company based in China), have been sanctioned by the U.S. Government for their roles in the PRC repression of Uyghurs in Xinjiang.

PRC Ambitions and Genomic Data

NATIONAL POLICIES: PRC has enacted national policies prioritizing the collection of healthcare data, including genetic data, both at home and abroad to achieve its goal of becoming a global biotech leader.

- **Precision Medicine Initiative:** In 2016, the PRC announced a $9 billion, 15-year project to collect, analyze, and sequence genomic data to become global leader in precision medicine under the "Healthy China 2030" initiative.
- **14th Five Year Plan:** In 2021, the PRC unveiled its 14th Five Year Plan, which listed genetics and biotechnology as among the cutting-edge science and technology research areas the PRC seeks to dominate in the years 2021-2025.
- **China Standards 2035:** The PRC's national strategy to set global rules and standards in emerging technologies, including those critical for future precision healthcare.

ECONOMIC ADVANTAGE: The PRC understands the collection and analysis of large genomic data sets from diverse populations helps foster new medical discoveries that can advance its AI, pharmaceutical, and precision medicine industries.

MILITARY / SECURITY ADVANTAGE: PRC has used genetic analysis for state surveillance, societal control, and has been conducting genetic research for military purposes and biodefense.

PRC Vectors to Access U.S. Genomic Data

INVESTMENTS

- China's largest genomics company, BGI, purchased U.S. genomic sequencing firm Complete Genomics in 2013.
- In 2015 WuXi Pharma Tech acquired U.S. genetic sequencing company NextCODE. WuXi NextCODE later received accreditation to perform molecular diagnostic and genetic testing in the U.S.

PARTNERSHIPS

- China's BGI has partnered with health institutions across America to provide low-cost genomic testing and sequencing services, while also gaining access to genetic data on persons in the U.S.
- According to a 2019 report prepared by Gryphon Scientific, 23 companies associated with China are certified to perform genetic testing in the U.S., giving them access to genetic data on patients in the US.

COMPELLED ACCESS

- All Chinese companies are subject to PRC laws requiring them to share data they acquire with the PRC government.

CYBER INTRUSIONS

- PRC has conducted cyber attacks on U.S. healthcare institutions and companies (such as Anthem and others) to acquire personal health information.

Risks Associated with PRC Access to U.S. Genomic Data

PRIVACY: Your genetic data could end up in the hands of the PRC and used for purposes you never intended. The loss of your DNA is permanent and not only affects you, but your relatives, and potentially, generations to come.

INTELLIGENCE: By combining genetic data with other PII and healthcare / lifestyle data the PRC has acquired through cyberattacks and other means, the PRC could use this information to target U.S. personnel, dissidents, journalists, and others around the world for potential surveillance, manipulation, or extortion.

ECONOMIC COMPETITION: Large, diverse sets of genetic and health data from around the world can help the PRC enhance its AI, pharmaceutical, healthcare, and precision medicine industries at the expense of U.S. biotech industry.

- **No Reciprocity:** The PRC severely restricts U.S. and other foreign access to Chinese genetic data, putting America's biotech industry at a disadvantage.

MILITARY & SECURITY CAPABILITIES: There is growing concern over PRC research and exploitation of genetic data for bioweapons and biodefense, including to enhance the performance of soldiers in combat and more effectively support force readiness.

Direct-to-Consumer (DTC) Genetic Testing Companies

DESIRABLE TARGETS: DTC genetic testing companies, information exchanges, and data libraries are desirable targets for foreign adversaries, cyber criminals, and insider threats.

HUGE GENETIC HOLDINGS: DTC genetic testing companies hold large quantities of human genetic data and other personal information. Last year, the American Medical Association (AMA) projected that as many as 100 million individuals would undergo DTC genetic tests by the end of 2021.

LESS REGULATED THAN U.S. HEALTHCARE PROVIDERS: Data held by DTC genetic testing companies are not subject to HIPAA / privacy and security requirements that apply to health care providers, as consumers send samples directly to the companies without the involvement of a health care provider.

Cyber Risks and DTC Genetic Testing Companies

Image source: DNA Testing Firm Data Breach Exposed Sensitive Information of More Than 2.1 Million People - CPO Magazine

In November 2021, an Ohio-based DNA testing company reported to regulators that personal information on more than 2.1 million people was acquired in a hacking incident. No genetic data reported stolen.

In July 2019, a California-based DNA testing company accidently exposed the personal data of 3,000 customers online, including some 300 files containing genetic data.

In June 2018, an Israel-based DNA testing company, announced it had been breached and the email addresses of more than 92 million users were compromised. No genetic data reported stolen.

Key Takeaways

GREAT PROMISE, BUT KEY RISKS: **The collection and analysis of genomic data holds great promise for medical breakthroughs, but with it comes important risks to privacy as well as economic and national security.**

- **Large genetic databases that allow people's ancestry to be revealed and crimes to be solved also can be misused for surveillance and societal repression.**
- **Genomic technology used to design disease therapies tailored to an individual also can be used to identify genetic vulnerabilities in a population that potentially could be targeted.**

ADVERSARIES ALREADY EXPLOITING GENOMIC DATA: **Adversaries are already exploiting genomic data and have national plans to acquire and harness genomic data at home and abroad for their economic advantage and national security.**

- **Foreign companies and authoritarian regimes have already gained significant access to U.S. genomic data and related healthcare data through investments, research partnerships, contractual agreements, and other means.**

LEGAL / REGULATORY GAPS ON GENETIC DATA: **U.S. laws currently do not treat genetic data as a national security asset, but primarily focus on privacy and IP protection. Few restrictions prevent a U.S. company from selling genetic data to parties outside the U.S.**

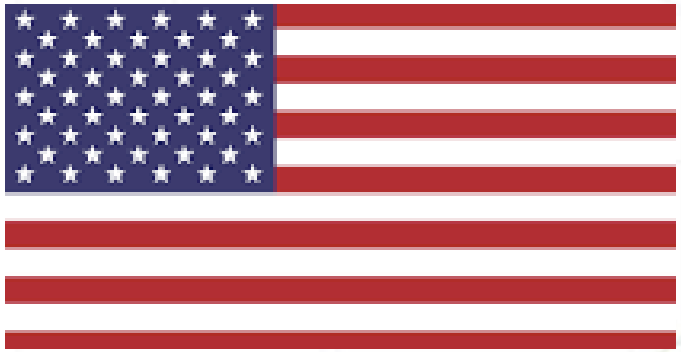
Keynote:
The Enabling Perspective

Yaniv Erlich (Eleven Therapeutics)

National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Free-for-all genetic surveillance nation



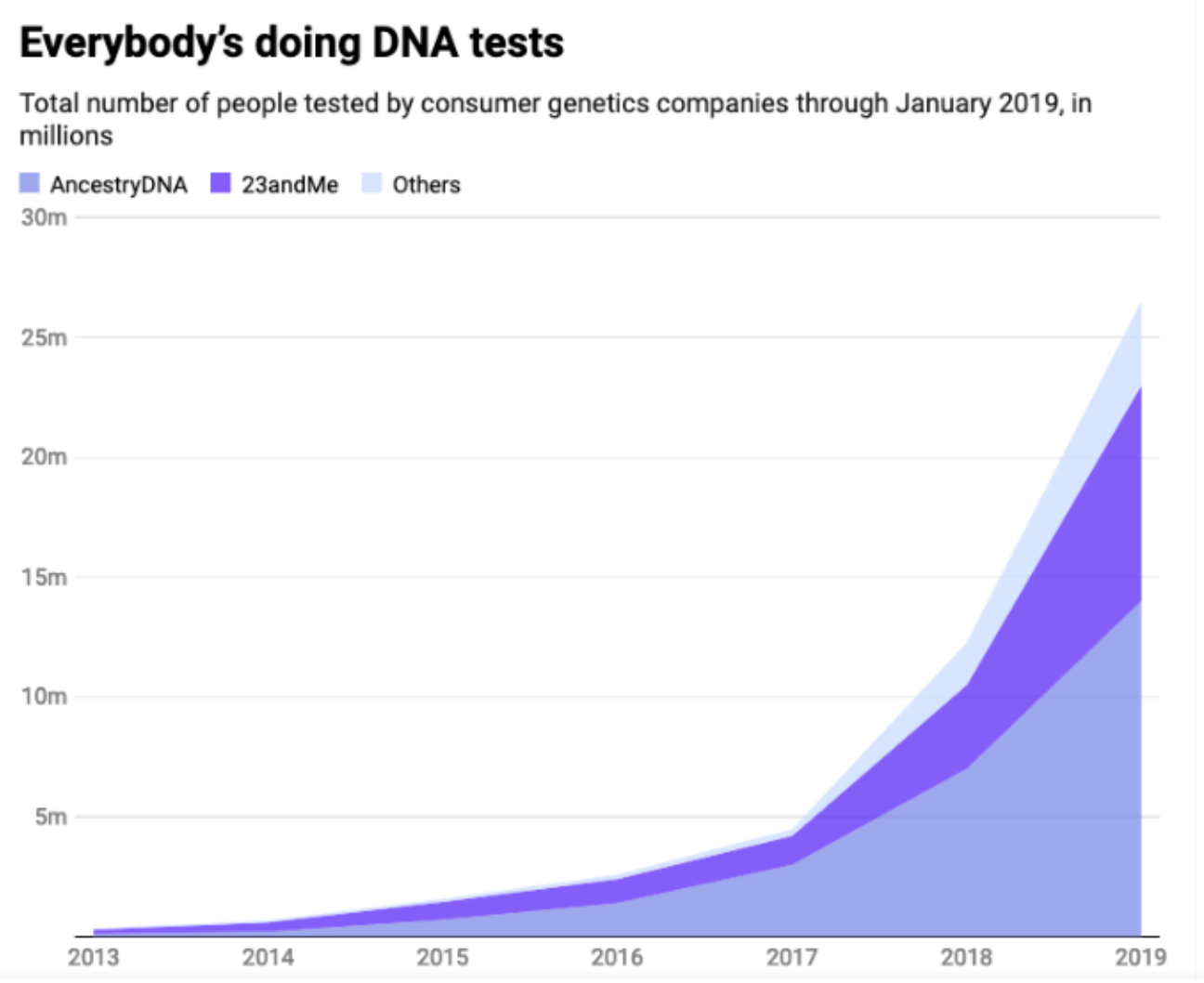
- Dr. Yaniv Erlich

erlichya

National Institute of Standards and Technology
U.S. Department of Commerce

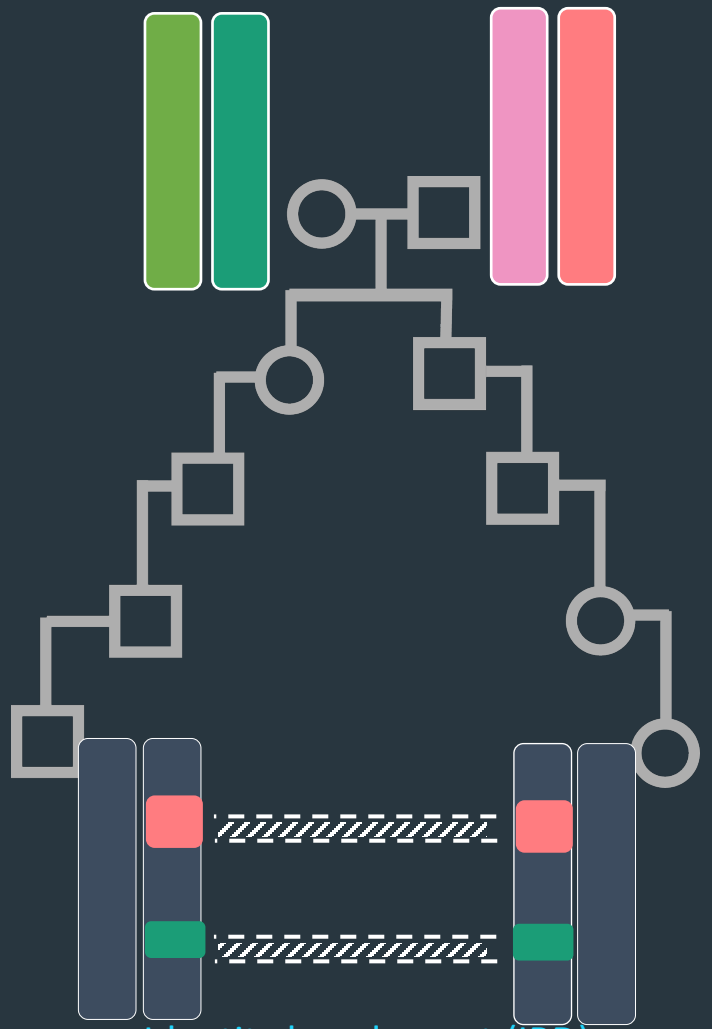
NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

The advent of consumer genomics

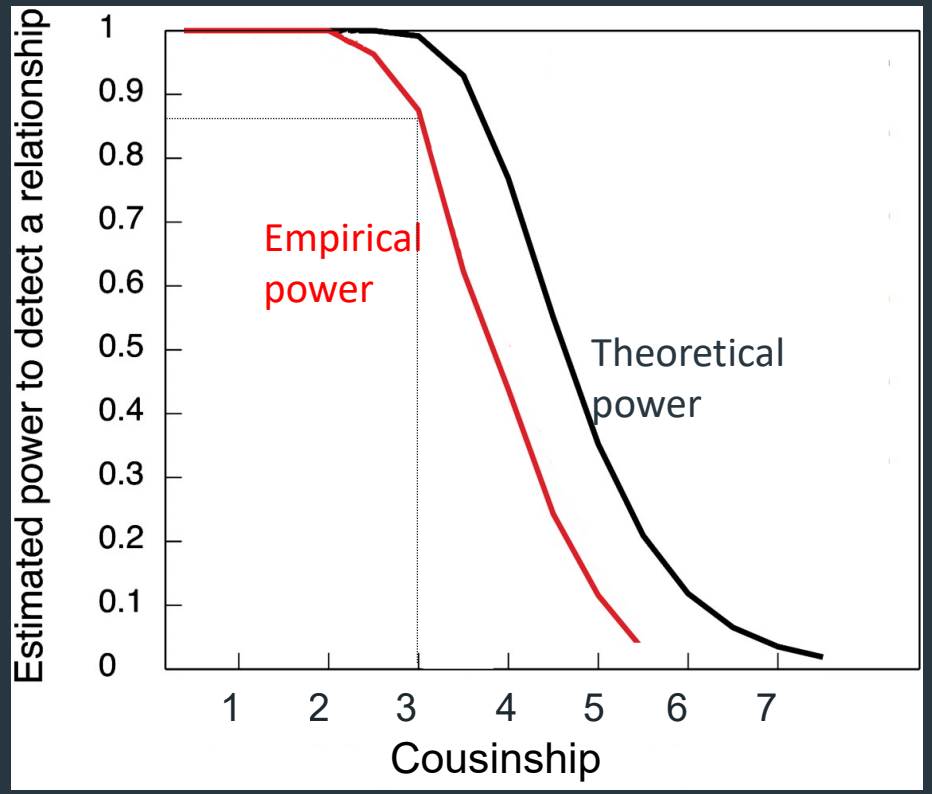


MIT Technology review

Relative matching via shared IBD



Identity by –descent (IBD) segments



Modified from Huff et al., Genome Research, 2011

Relative matching is the core of genetic genealogy

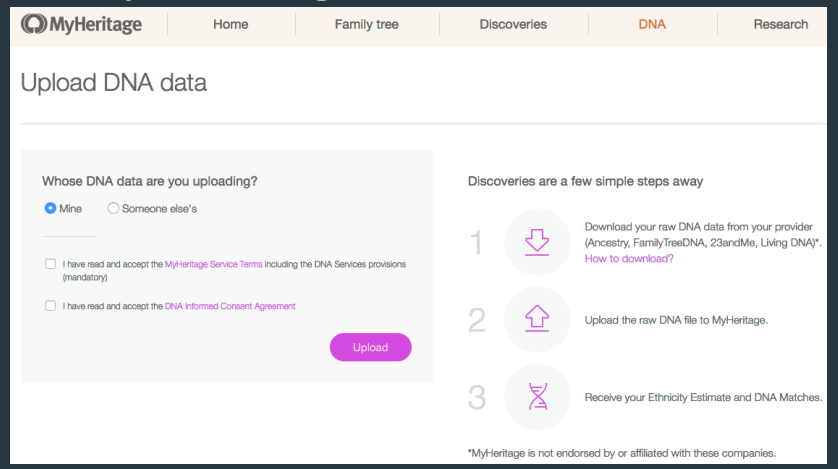
3rd party support for relative matching

Your raw genetic data

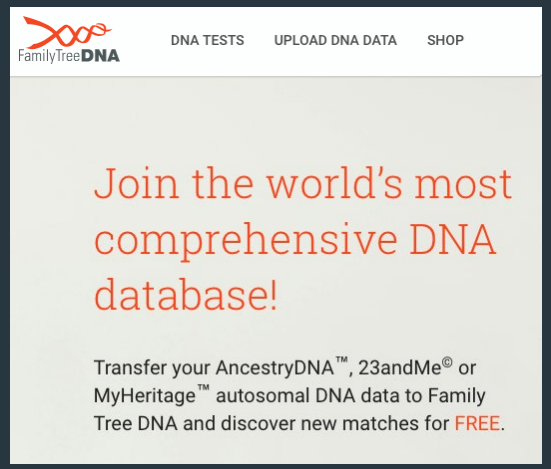
```
# MyHeritage DNA raw data.
# This file was generated on 2018-10-10 09:03:32
# For each SNP, we provide the identifier, chromosome
# number, base pair position and genotype. The genotype
# is reported on the forward (+) strand with respect to
# the human reference build 37.
# THIS INFORMATION IS FOR YOUR PERSONAL USE AND IS
# INTENDED FOR GENEALOGICAL RESEARCH
# ONLY. IT IS NOT INTENDED FOR MEDICAL OR HEALTH
# PURPOSES. PLEASE BE AWARE THAT THE
# DOWNLOADED DATA WILL NO LONGER BE PROTECTED BY OUR
SECURITY MEASURES.


#RSID,CHROMOSOME,POSITION,RESULT
"rs4477212","1","82154","AA"
"rs3094315","1","752566","--"
"rs3131972","1","752721","AG"
"rs12562034","1","768448","--"
"rs12124819","1","776546","--"
"rs11240777","1","798959","GG"
"rs6681049","1","800007","--"
"rs4970383","1","838555","AC"
"rs4475691","1","846808","TC"
"rs7537756","1","854250","AG"
"rs13302982","1","861808","GG"
"rs1110052","1","873558","TG"
"rs2272756","1","882033","GG"
```

Upload

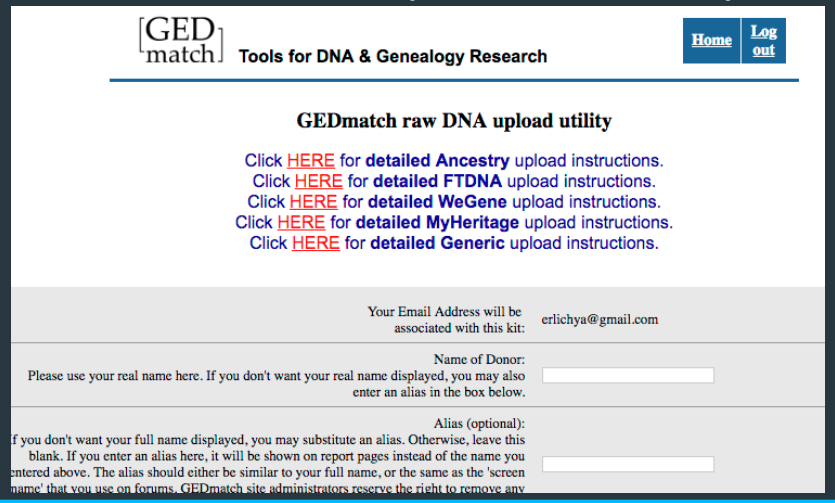
MyHeritage (users: 3M)



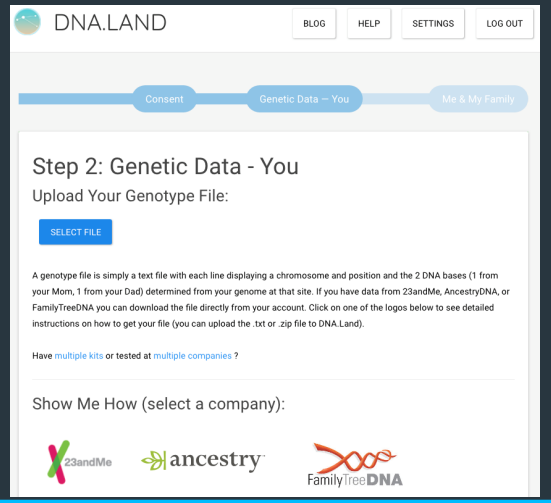
FTDNA (users: 1M)



GEDmatch (users: 1.4M)



DNA.Land (users: 150K)



3rd party uploads are highly important

Using genetic genealogy for forensic is not a new idea

nature REVIEWS GENETICS

Routes for breaching and protecting genetic privacy

REVIEWS

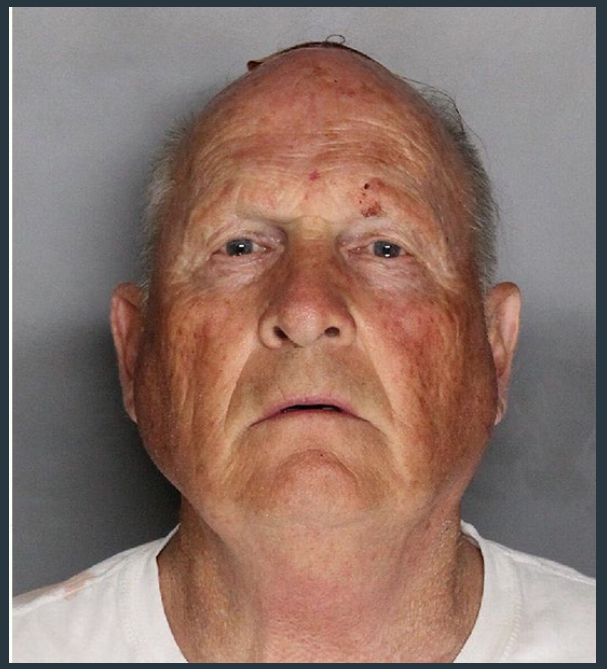
Yaniv Erlich[1] and Arvind Narayanan[2]

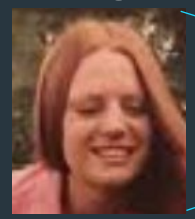
Abstract | We are entering an era of ubiquitous genetic information for research, clinical care and personal curiosity. Sharing these data sets is vital for progress in biomedical research. However, a growing concern is the ability to protect the genetic privacy of the data originators. Here, we present an overview of genetic privacy breaching strategies. We outline the principles of each technique, indicate the underlying assumptions, and assess their technological complexity and maturation. We then review potential mitigation methods for privacy-preserving dissemination of sensitive data and highlight different cases that are relevant to genetic applications.

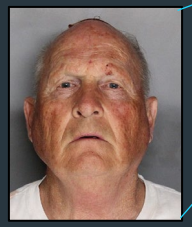
Erlich and Narayanan, 2014

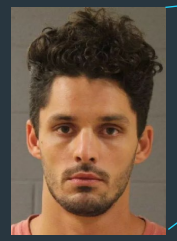
An open research question is the use of non-Y-chromosome markers for genealogical triangulation. The Mitosearch and GEDmatch websites run open, searchable databases for matching mitochondrial and autosomal genotypes, respectively. Our expectation is that mitochondrial data will not be very informative for tracing identities. The resolution of mitochondrial searches is low owing to the small size of the mitochondrial genome, which means that a large number of individuals share the same mitochondrial haplotypes. In addition, matrilineal identifiers (such as surname or clan) are fairly rare in most human societies, which complicates the use of mitochondrial haplotype for identity tracing. By contrast, autosomal searches can be powerful. Genetic genealogy companies have started to market services for dense genome-wide arrays that enable the identification of distant relatives (on the order of third to fourth cousins) with fairly sufficient accuracy[43]. These hits would reduce the search space to no more than a few thousand individuals[44]. The main challenge of this approach would be to derive a list of potential people from a genealogical match. As stated above, family trees of most individuals are not publicly available; such searches are therefore demanding and would require indexing a large number of genealogical websites. With the growing interest in genealogy, this technique might be easier in the future and should be taken into consideration.

Long range familial searches









Case	Announcement	By
Buckskin Girl	April 9, 2018	DNA Doe Project
Golden State Killer	April 24, 2018	Barbara Rae-Venter
Lyle Stevik	May 8, 2018	DNA Doe Project
William Earl Talbott II	May 21, 2018	Parabon
Joseph Newton Chandler III	June 21 2018	DNA Doe Project
Gary Hartman	June 22, 2018	Parabon
Raymond "DJ Freez" Rowe	June 25, 2018	Parabon
James Otto Earhart	June 26, 2018	Parabon
John D. Miller	July 15, 2018	Parabon
Matthew Dusseault	July 28, 2018	Parabon
Spencer Glen Monnett	July 29, 2018	Parabon
Darold Wayne Bowden	August 23rd, 2018	Parabon
Michael F. Henslick	August 29th, 2018	Parabon

The probability of finding a relative?

Repeat 1,280,000 times

Find genetic relatives for a MyHeritage participant

↓

Exclude >700cM matches

↓

What is the top match of the person?

Repeat 30 times

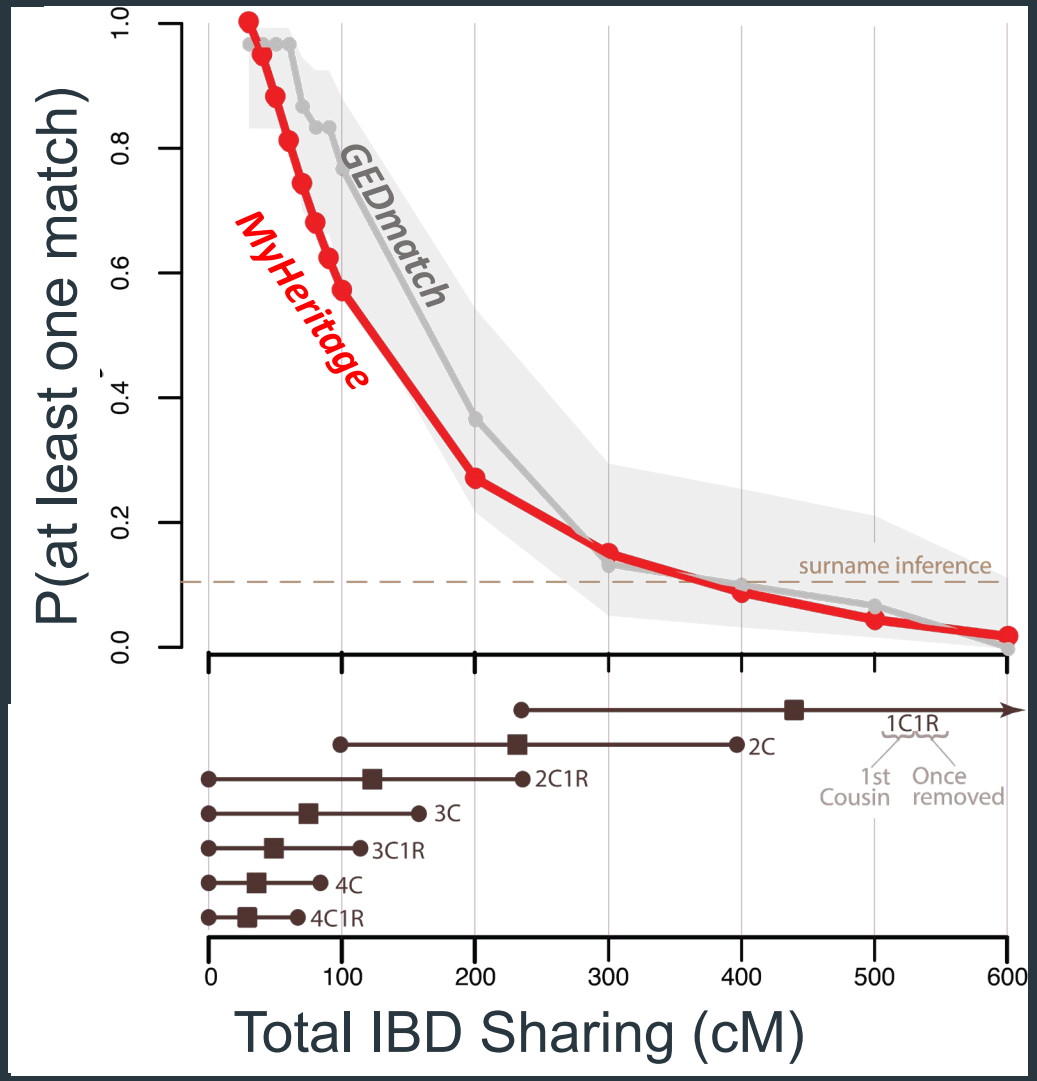
Find genetic relatives for a GEDMatch participant

↓

Exclude >700cM matches

↓

What is the top match of the person?



Estimate: ~60% of US individuals of European heritage have a 3rd cousin match

Small scale study confirms our projection

SCIENCE

We Tried To Find 10 BuzzFeed Employees Just Like Cops Did For The Golden State Killer

The Golden State Killer case has triggered a boom in "genetic genealogy" for solving crimes. But how hard is it to find people by sleuthing in their family trees?

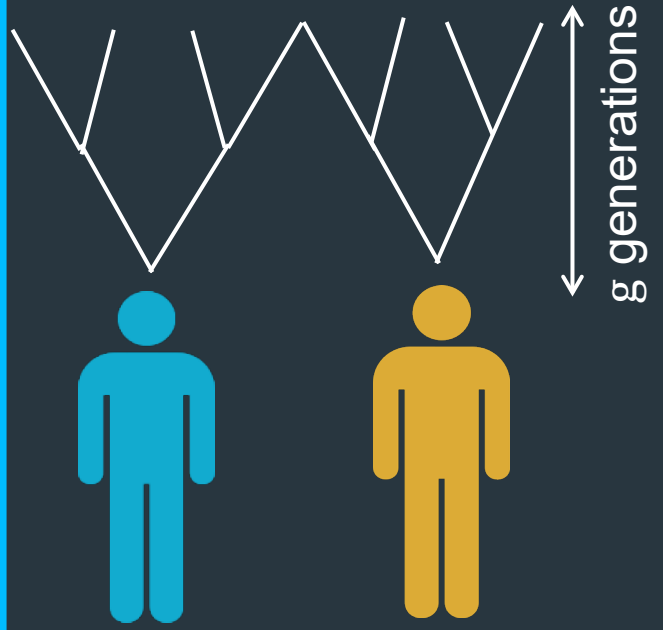
Peter Aldhous
BuzzFeed News Reporter

Posted on April 9, 2019, at 9:16 a.m. ET

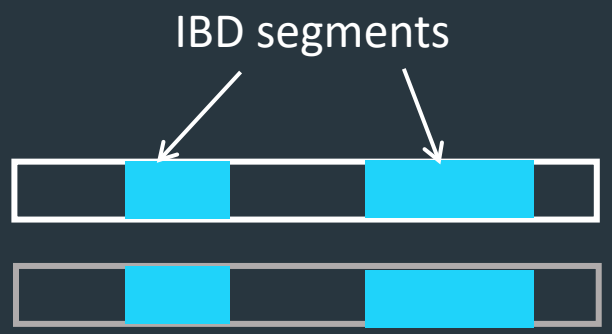
In the end, I identified 6 out of our 10 volunteers. Four of those cases I solved by tracking them down through their relatives' family trees, much as the cops did with DeAngelo. In a twist I didn't anticipate, I found two more not through their relatives, but simply because their ancestry indicated that their family came from a specific country — raising uncomfortable questions about genetic racial profiling.

Modeling the probability of finding relatives

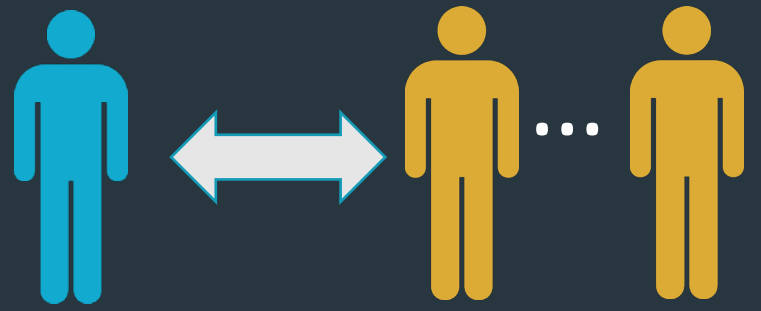
Are these individuals **genealogically** related?



Are these individuals also **genetically** related?

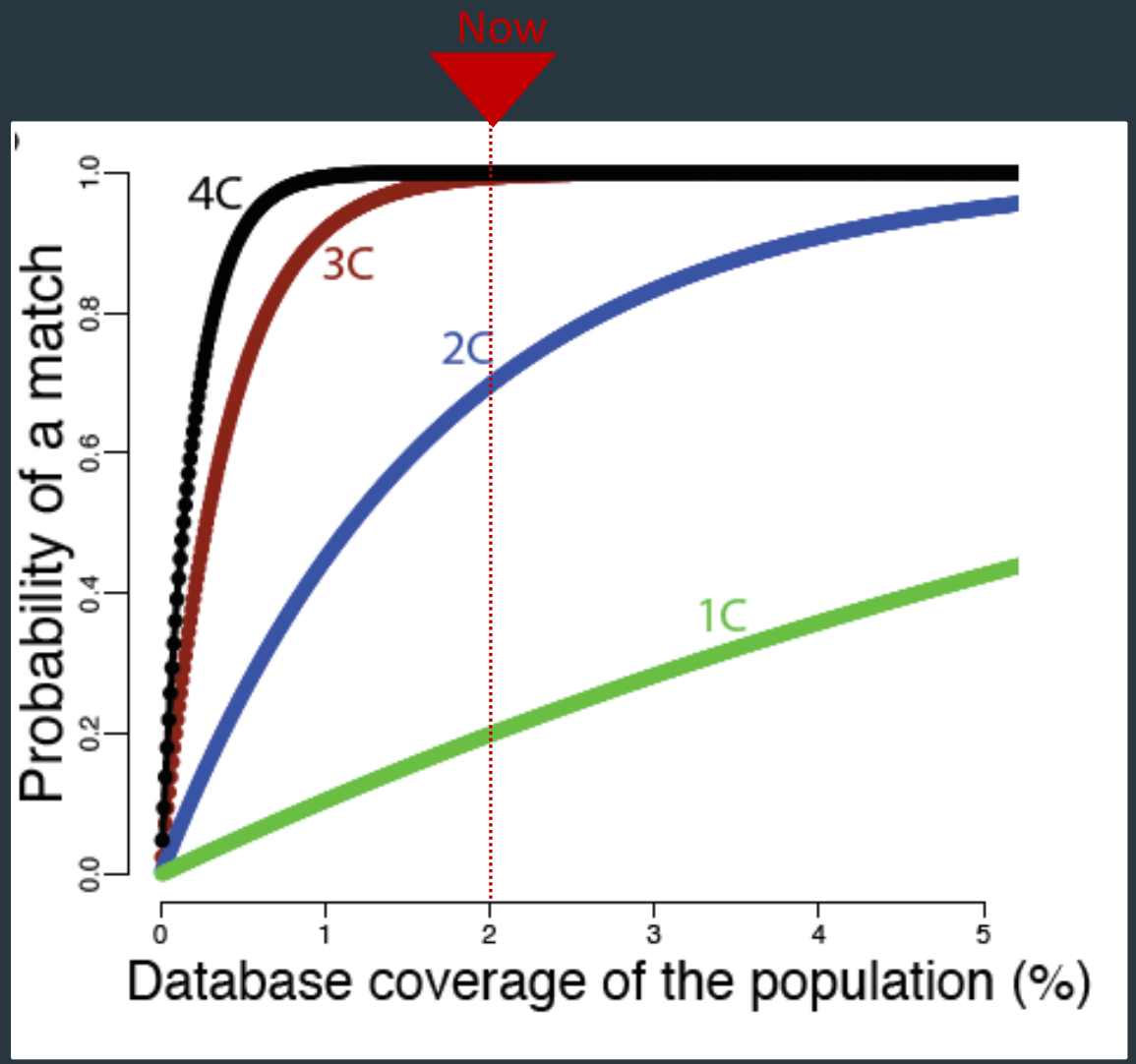


Is there at least one match in a database that covers X% of the population?



Caveats: no-population structure or consanguinity; assumes random samples.

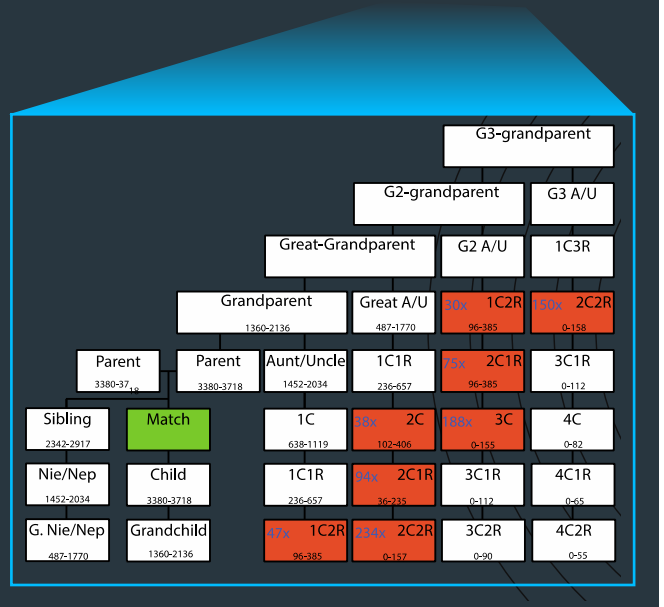
The probability of a match in the future



Virtually any US person of European heritage will have a 3rd cousin in these databases.

Can we get to a single person?

325M → 3rd cousin match → ex → 16.5



Even three simple pieces of

Crime Scene and Distance Correlates of Serial Rape

Janet Warren,[1,7] Roland Reboussin,[2] Robert R. Hazelwood,[3] Andrea Cummings,[4] Natalie Gibbs,[5] and Susan Trumbetta[6]

This study, derived from a sample of 108 serial rapists (rapes = 565), examines the relationship between demographic, crime scene, and criminal history variables and the distance traveled by serial rapists in order to offend. The pattern of offenses perpetrated by each of the 108 serial offenders as it relates to his place of residence is also analyzed in terms of known characteristics of the offender and his offenses. The theoretical focus of the study integrates premises derived from criminal investigative analysis, environmental criminology, ethnographic geography, journey to crime research, and criminal geographic targeting to explore the cognitive symmetry between the "how" and the "where" of serial sexual offenses. These components or dimensions of serial crime are explored in an attempt to aid law enforcement in their investigation of hard-to-solve serial crimes.

KEY WORDS: serial rape; journey to crime; crime scene analysis; criminal investigative analysis; spatial analysis of crime; environmental criminology; criminal geographic targeting; geographic profiling.

Summary so far

We expect 2nd - 3rd cousin for virtually every person in the US with European descent (if access is allowed)

Basic demographic information can substantially narrow the search space to handful of individuals

The method is extremely powerful

Paper

Science REPORTS

Cite as: Y. Erlich *et al.*, *Science* 10.1126/science.aau4832 (2018).

Identity inference of genomic data using long-range familial searches

Yaniv Erlich[1,2,3,4*], Tal Shor[1], Itsik Pe'er[2,3], Shai Carmi[5]

[1]MyHeritage, Or Yehuda 6037606, Israel. [2]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA. [3]Center for Computational Biology and Bioinformatics (C2B2), Department of Systems Biology, Columbia University, New York, NY, USA. [4]New York Genome Center, New York, NY, USA. [5]Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel.

***Corresponding author. Email: erlichya@gmail.com**

Consumer genomics databases have reached the scale of millions of individuals. Recently, law enforcement authorities have exploited some of these databases to identify suspects via distant familial relatives. Using genomic data of 1.28 million individuals tested with consumer genomics, we investigated the power of this technique. We project that over 60% of the searches for individuals of European-descent will result in a third cousin or closer match, which can allow their identification using demographic identifiers. Moreover, the technique could implicate nearly any US-individual of European-descent in the near future. We demonstrate that the technique can also identify research participants of a public sequencing project. Based on these results, we propose a potential mitigation strategy and policy implications to human subject research.

So why am I worried?

Genetic genealogy can be weaponized by counter-intelligence

1. Everyone can uploads data to GEDmatch/FTDNA/etc…
2. Also adversaries of the US (they don’t give a damn to Toc)
3. Counter-intelligence and other players can exploit genetic genealogy to cast population-scale genetic surveillance over the US
4. The risk is asymmetric (US is substantially affected but not other countries)

Intelligence services are interested in DNA

WikiLeaks: are Chinese spies stealing Iceland's genetic database?

by Jared Yee | 18 Dec 2010 | Link

Another bioethics angle has emerged in Wikileaks. Chinese spies are investigating genetic research companies in Iceland, according to cables written in authorities said that intelligence gathering included bugging phone lin hacking into databases.

Novichok bottle could hold attackers' DNA

Dawn Sturgess died last week from novichok poisoning; Charlie Rowley remains in hospital

AFP PHOTO/FACEBOOK

Differentiate good vs. bad actors

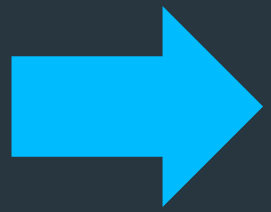
Can we differentiate legitimate searches from illegitimate searches?

- Legitimate datasets are produced with a regular DTC lab or authorized crime labs
- Illegitimate datasets are produced by research labs, unauthorized crime labs, etc.
- Idea: ask authorized labs to sign datasets before letting users downloading the data.

How it works?

Before

```
# MyHeritage DNA raw data.
# This file was generated on 2018-10-10 09:03:32
# For each SNP, we provide the identifier, chromosome
# number, base pair position and genotype. The genotype
# is reported on the forward (+) strand with respect to
# the human reference build 37.
# THIS INFORMATION IS FOR YOUR PERSONAL USE AND IS
# INTENDED FOR GENEALOGICAL RESEARCH
# ONLY. IT IS NOT INTENDED FOR MEDICAL OR HEALTH
# PURPOSES. PLEASE BE AWARE THAT THE
# DOWNLOADED DATA WILL NO LONGER BE PROTECTED BY OUR
SECURITY MEASURES.


#RSID,CHROMOSOME,POSITION,RESULT
”rs4477212","1","82154","AA”
"rs3094315","1","752566","--”
"rs3131972","1","752721","AG”
"rs12562034","1","768448","--”
"rs12124819","1","776546","--”
"rs11240777","1","798959","GG”
"rs6681049","1","800007","--”
"rs4970383","1","838555","AC”
"rs4475691","1","846808","TC”
"rs7537756","1","854250","AG”
"rs13302982","1","861808","GG”
"rs1110052","1","873558","TG”
"rs2272756","1","882033","GG"
```

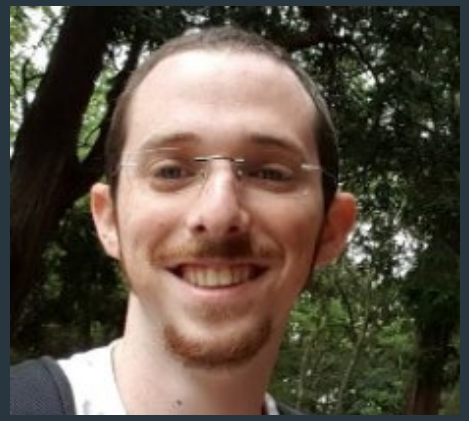


After

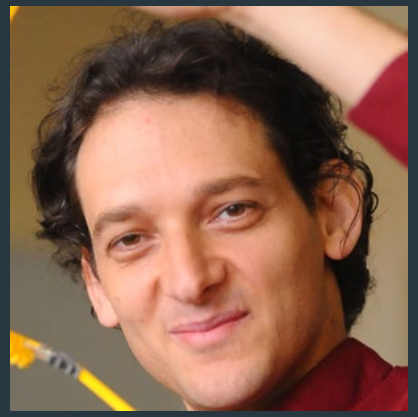
```
# MyHeritage DNA raw data.
# This file was generated on 2018-10-10 09:03:32
# For each SNP, we provide the identifier, chromosome
# number, base pair position and genotype. The genotype
# is reported on the forward (+) strand with respect to
# the human reference build 37.
# THIS INFORMATION IS FOR YOUR PERSONAL USE AND IS
# INTENDED FOR GENEALOGICAL RESEARCH
# ONLY. IT IS NOT INTENDED FOR MEDICAL OR HEALTH
# PURPOSES. PLEASE BE AWARE THAT THE
# DOWNLOADED DATA WILL NO LONGER BE PROTECTED BY OUR SECURITY
MEASURES.
#SIGNATURE=RZTcitAZ1bneCfURL5gsC5yRghb9=
#RSID,CHROMOSOME,POSITION,RESULT
”rs4477212","1","82154","AA”
"rs3094315","1","752566","--”
"rs3131972","1","752721","AG”
"rs12562034","1","768448","--”
"rs12124819","1","776546","--”
"rs11240777","1","798959","GG”
"rs6681049","1","800007","--”
"rs4970383","1","838555","AC”
"rs4475691","1","846808","TC”
"rs7537756","1","854250","AG”
"rs13302982","1","861808","GG”
"rs1110052","1","873558","TG”
"rs2272756","1","882033","GG"
```

Seamless for the user!

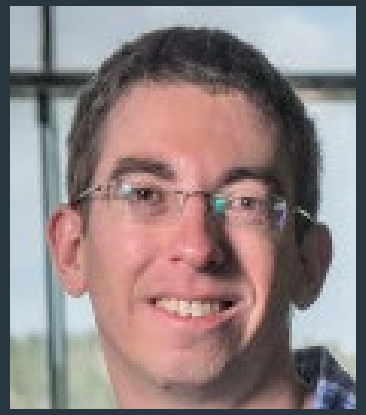
Acknowledgments



Tal Shor
MyHeritage



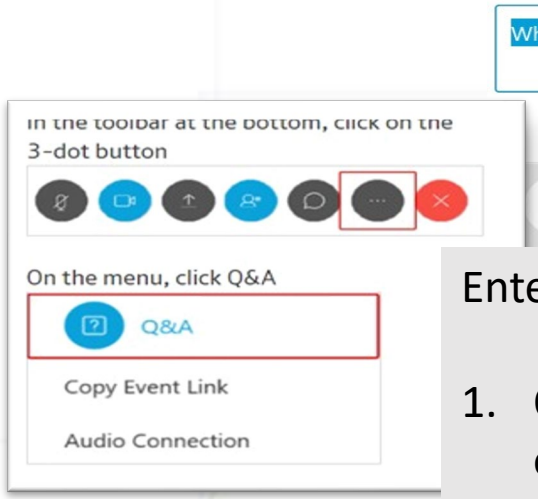
Itsik Pe'er
Columbia University

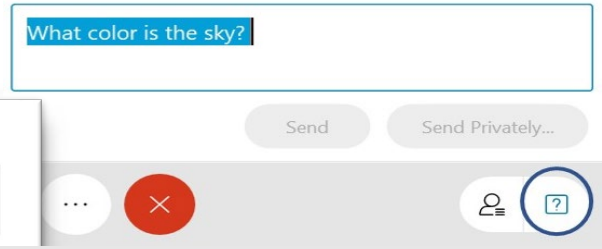


Shai Carmi
Hebrew University

Keynotes: The Protection and Enabling Perspectives

Moderated Questions and Answers





Enter your question in the Q&A panel.

1. On the right side, click on Q&A header to open the Q&A panel.
2. Type in the box **your name, organization and question**.
3. Click send.

National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE