National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

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

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

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.
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.

Bottom photo source: https://baijiahao.baidu.com/s?id=1564669932542581
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.
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)
Free-for-all genetic surveillance nation

• Dr. Yaniv Erlich
  @erlichya
The advent of consumer genomics

**Everybody’s doing DNA tests**

- Total number of people tested by consumer genetics companies through January 2019, in millions
  - AncestryDNA
  - 23andMe
  - Others

*MIT Technology review*
Relative matching via shared IBD

Probability of a match
From match to a person
Concerns
Solutions

Introduction

Probability of a match

From match to a person

Concerns

Solutions

Identity by-descent (IBD) segments

Empirical power
Theoretical power

Modified from Huff et al., Genome Research, 2011
Relative matching is the core of genetic genealogy
3\textsuperscript{rd} party support for relative matching

### MyHeritage (users: 3M)

- **Upload DNA data**

### FTDNA (users: 1M)

- **Join the world’s most comprehensive DNA database!**
  - Transfer your AncestryDNA™, 23andMe® or MyHeritage™ autosomal DNA data to Family Tree DNA and discover new matches for **FREE**.

### GEDmatch (users: 1.4M)

### DNA.Land (users: 150K)

- **Free-for-all genetic surveillance nation**

---

**Your raw genetic data**

<table>
<thead>
<tr>
<th>RSID, CHROMOSOME, POSITION, RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4477212, '1', '82154', 'AA'</td>
</tr>
<tr>
<td>rs3094315, '1', '75256', '--'</td>
</tr>
<tr>
<td>rs3131972, '1', '75272', 'AG'</td>
</tr>
<tr>
<td>rs12562034, '1', '76844', '--'</td>
</tr>
<tr>
<td>rs12124815, '1', '77654', '--'</td>
</tr>
<tr>
<td>rs11240777, '1', '79859', 'GG'</td>
</tr>
<tr>
<td>rs6681049, '1', '80007', '--'</td>
</tr>
<tr>
<td>rs4970383, '1', '83955', 'AC'</td>
</tr>
<tr>
<td>rs4475691, '1', '84808', 'TC'</td>
</tr>
<tr>
<td>rs7537756, '1', '85425', 'AG'</td>
</tr>
<tr>
<td>rs13302982, '1', '86180', 'GGG'</td>
</tr>
<tr>
<td>rs11100327, '1', '87355', 'TG'</td>
</tr>
<tr>
<td>rs12397256, '1', '88023', 'GGG'</td>
</tr>
</tbody>
</table>

---

**Introduction**

<table>
<thead>
<tr>
<th>Probability of a match</th>
<th>Concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td><strong>Concerns</strong></td>
</tr>
<tr>
<td><strong>Probability of a match</strong></td>
<td><strong>From match to a person</strong></td>
</tr>
<tr>
<td><strong>From match to a person</strong></td>
<td><strong>Concerns</strong></td>
</tr>
<tr>
<td><strong>Concerns</strong></td>
<td><strong>Solutions</strong></td>
</tr>
</tbody>
</table>

---

**# MyHeritage DNA raw data.**

<table>
<thead>
<tr>
<th>RSID, CHROMOSOME, POSITION, RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4477212, '1', '82154', 'AA'</td>
</tr>
<tr>
<td>rs3094315, '1', '75256', '--'</td>
</tr>
<tr>
<td>rs3131972, '1', '75272', 'AG'</td>
</tr>
<tr>
<td>rs12562034, '1', '76844', '--'</td>
</tr>
<tr>
<td>rs12124815, '1', '77654', '--'</td>
</tr>
<tr>
<td>rs11240777, '1', '79859', 'GG'</td>
</tr>
<tr>
<td>rs6681049, '1', '80007', '--'</td>
</tr>
<tr>
<td>rs4970383, '1', '83955', 'AC'</td>
</tr>
<tr>
<td>rs4475691, '1', '84808', 'TC'</td>
</tr>
<tr>
<td>rs7537756, '1', '85425', 'AG'</td>
</tr>
<tr>
<td>rs13302982, '1', '86180', 'GGG'</td>
</tr>
<tr>
<td>rs11100327, '1', '87355', 'TG'</td>
</tr>
<tr>
<td>rs12397256, '1', '88023', 'GGG'</td>
</tr>
</tbody>
</table>
3rd party uploads are highly important
Using genetic genealogy for forensic is not a new idea

Erlich and Narayanan, 2014

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.
### Long range familial searches

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

Introduction

Probability of a match

From match to a person

Concerns

Solutions
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?

3rd cousin match

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 Trumbetta6

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 2\textsuperscript{nd} - 3\textsuperscript{rd} 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
Identity inference of genomic data using long-range familial searches

Yaniv Erlich¹,²,⁵,⁶, Tal Shor⁴, Itsik Pe’er²,³, Shal Carmi³

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

Cite as: Y. Erlich et al., Science 10.1126/science.aau4832 (2018).
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 lines, hacking into databases.

Novichok bottle could hold attackers’ DNA

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

<table>
<thead>
<tr>
<th>RSID</th>
<th>CHROMOSOME</th>
<th>POSITION</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4477212</td>
<td>1</td>
<td>82154</td>
<td>AA</td>
</tr>
<tr>
<td>rs3094315</td>
<td>1</td>
<td>752566</td>
<td>--</td>
</tr>
<tr>
<td>rs3131972</td>
<td>1</td>
<td>752721</td>
<td>AG</td>
</tr>
<tr>
<td>rs12562034</td>
<td>1</td>
<td>768448</td>
<td>--</td>
</tr>
<tr>
<td>rs12124819</td>
<td>1</td>
<td>775646</td>
<td>--</td>
</tr>
<tr>
<td>rs11240777</td>
<td>1</td>
<td>798959</td>
<td>--</td>
</tr>
<tr>
<td>rs6681049</td>
<td>1</td>
<td>800007</td>
<td>--</td>
</tr>
<tr>
<td>rs4970383</td>
<td>1</td>
<td>838555</td>
<td>AC</td>
</tr>
<tr>
<td>rs4475691</td>
<td>1</td>
<td>846808</td>
<td>TC</td>
</tr>
<tr>
<td>rs7537756</td>
<td>1</td>
<td>854250</td>
<td>AG</td>
</tr>
<tr>
<td>rs13302982</td>
<td>1</td>
<td>861808</td>
<td>GG</td>
</tr>
<tr>
<td>rs1110052</td>
<td>1</td>
<td>873558</td>
<td>TG</td>
</tr>
<tr>
<td>rs2272756</td>
<td>1</td>
<td>882033</td>
<td>GG</td>
</tr>
</tbody>
</table>

**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=

<table>
<thead>
<tr>
<th>RSID</th>
<th>CHROMOSOME</th>
<th>POSITION</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4477212</td>
<td>1</td>
<td>82154</td>
<td>AA</td>
</tr>
<tr>
<td>rs3094315</td>
<td>1</td>
<td>752566</td>
<td>--</td>
</tr>
<tr>
<td>rs3131972</td>
<td>1</td>
<td>752721</td>
<td>AG</td>
</tr>
<tr>
<td>rs12562034</td>
<td>1</td>
<td>768448</td>
<td>--</td>
</tr>
<tr>
<td>rs12124819</td>
<td>1</td>
<td>775646</td>
<td>--</td>
</tr>
<tr>
<td>rs11240777</td>
<td>1</td>
<td>798959</td>
<td>--</td>
</tr>
<tr>
<td>rs6681049</td>
<td>1</td>
<td>800007</td>
<td>--</td>
</tr>
<tr>
<td>rs4970383</td>
<td>1</td>
<td>838555</td>
<td>AC</td>
</tr>
<tr>
<td>rs4475691</td>
<td>1</td>
<td>846808</td>
<td>TC</td>
</tr>
<tr>
<td>rs7537756</td>
<td>1</td>
<td>854250</td>
<td>AG</td>
</tr>
<tr>
<td>rs13302982</td>
<td>1</td>
<td>861808</td>
<td>GG</td>
</tr>
<tr>
<td>rs1110052</td>
<td>1</td>
<td>873558</td>
<td>TG</td>
</tr>
<tr>
<td>rs2272756</td>
<td>1</td>
<td>882033</td>
<td>GG</td>
</tr>
</tbody>
</table>

Seamless for the user!
Acknowledgments

Tal Shor
MyHeritage

Itsik Pe’er
Columbia University

Shai Carmi
Hebrew University
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.