National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

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

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

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

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

This webinar is being recorded.

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

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

AGENDA

Segment	***Time***
Segment 1: Workshop Overview and Background	11:00 AM – 11:40 AM
Segment 2: Keynotes	11:40 AM – 12:20 PM
Segment 3: Challenges from the Field	12:20 PM – 12:50 PM
Intermission	12:50 PM – 1:30 PM
Segment 4: Challenges Sessions	1:30 PM – 2:25 PM
Break	2:25 PM – 2:35 PM
Segment 4 (Continued): Challenges Sessions	2:35 PM – 3:45 PM
Break	3:45 PM – 3:50 PM
Segment 5: Open Lightning Round	3:50 PM – 4:20 PM
Segment 6: Next Steps	4:20 PM – 4:30 PM

Welcome to the NCCoE Virtual Workshop on Cybersecurity of Genomic Data

Natalia Martin, NIST

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

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Virtual Workshop on the Cybersecurity of Genomic Data

Natalia Martin, Acting Director for National Cybersecurity Center of Excellence (NCCoE), NIST

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

NCCoE
NATIONAL CYBERSECURITY
CENTER OF EXCELLENCE

About the NCCoE

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

WHO WE ARE

A **solution-driven**, **collaborative** hub addressing complex cybersecurity problems

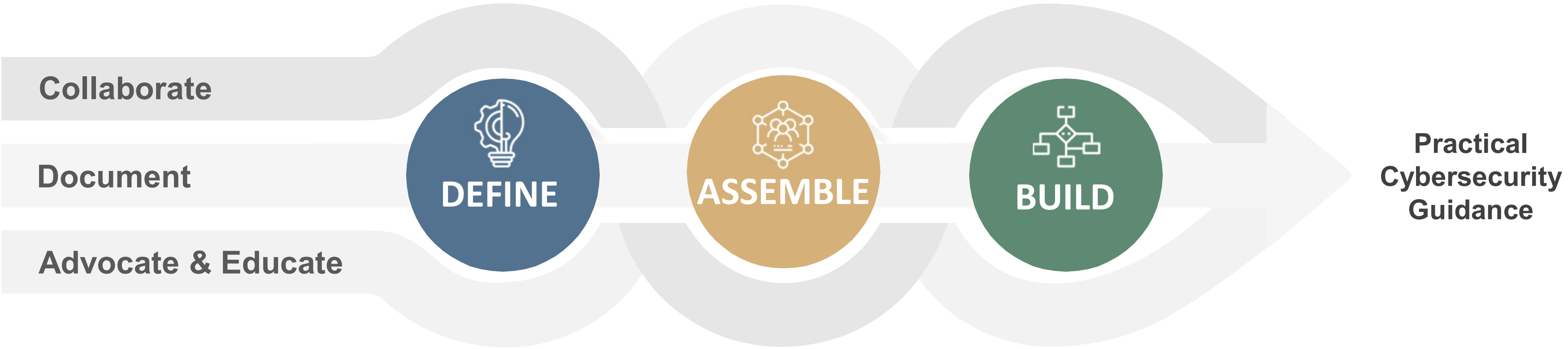
OUR GOALS

Improve cybersecurity for businesses and commerce

Lower the learning curve for cybersecurity

Spark innovation in secure technology

OUR APPROACH



Collaborate

Document

Advocate & Educate

DEFINE

ASSEMBLE

BUILD

Practical Cybersecurity Guidance

Define a scope of work with industry to solve a pressing cybersecurity challenge

Assemble teams to address all aspects of the cybersecurity challenge

Build a practical, usable, repeatable implementation to address the cybersecurity challenge

Virtual Workshop on the Cybersecurity of Genomic Data

Workshop Overview

Ron Pulivarti, NIST

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Housekeeping

- We support the health and well being for all.
  - We are supporting virtual collaboration.
  - We have three breaks planned for the day.

- We want audience engagement.
  - Please pose your questions for today's workshop using the Q&A window.
  - Please voice your insights in the Open Lightening Round from 3:50 – 4:30 PM.

- We intend to share our learnings today.
  - We are recording this session for future post on the NCCoE Website.
  - We will summarize key insights.

NIST Experiences in Genomics, Cybersecurity, and Privacy

Samantha Maragh (NIST)
Naomi Lefkovitz (NIST)
Ron Pulivarti (NIST)

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

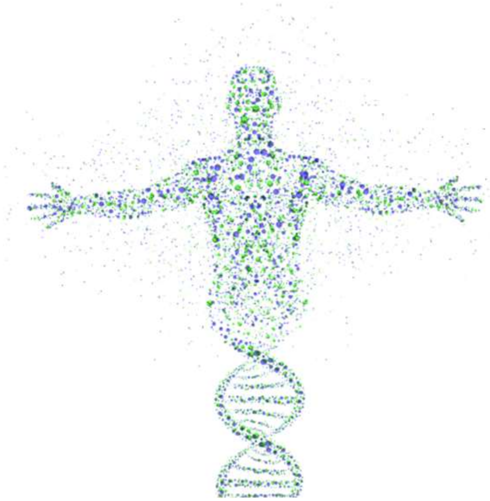
NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Human Genomics at NIST

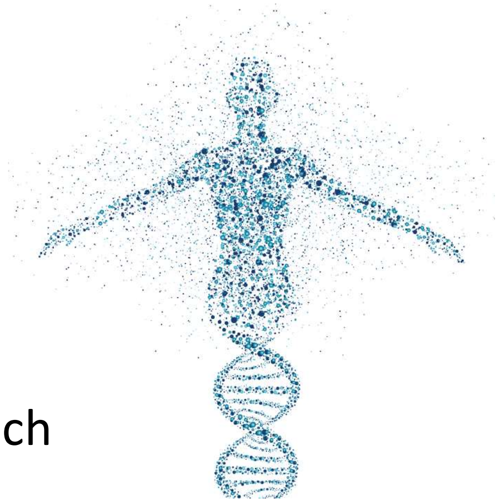
Samantha Maragh
Leader, Genome Editing Program

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

The Human Genome



- The instruction code for humans
- ~6.4 billion letters long
- Present in each cell of a person
- ~ Half inherited from each biological parent
- Code is highly similar between people, but each person has a unique identifiable sequence



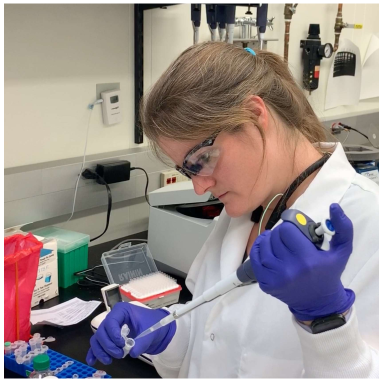
(Credit: Elymas/Shutterstock)

Uses of human genomic information

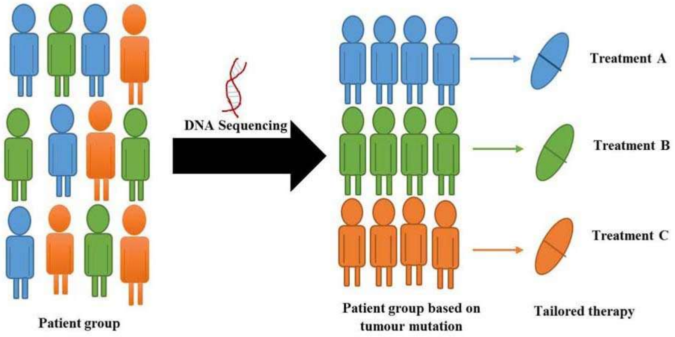
Human Identification

Population Diversity and Ancestry

Human Health: Diagnostics

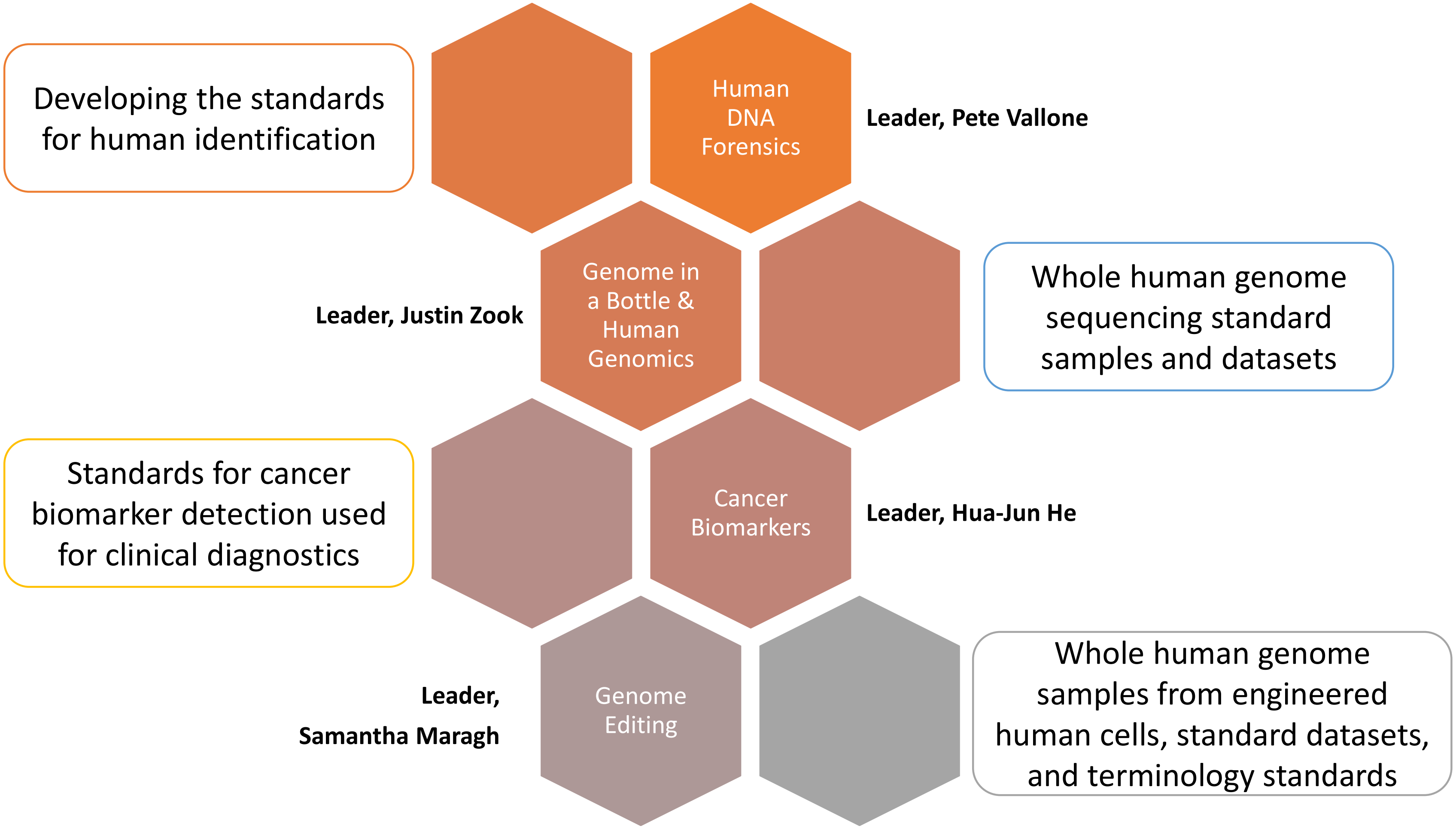


Scientific Research



Human Health: Treatment

NIST expertise with human genomics



Developing the standards for human identification

Human DNA Forensics

Leader, Pete Vallone

Leader, Justin Zook

Genome in a Bottle & Human Genomics

Whole human genome sequencing standard samples and datasets

Standards for cancer biomarker detection used for clinical diagnostics

Cancer Biomarkers

Leader, Hua-Jun He

Leader, Samantha Maragh

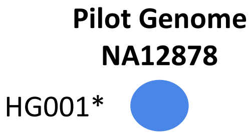
Genome Editing

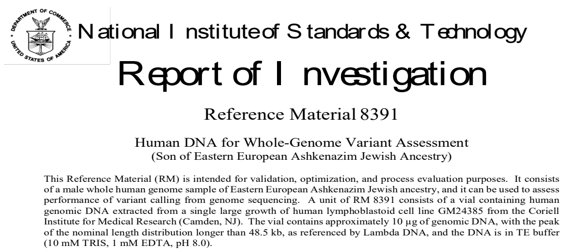
Whole human genome samples from engineered human cells, standard datasets, and terminology standards

NIST formed the Genome in a Bottle Consortium in 2012

NIST

GIAB has characterized variants in 7 human genomes and released NIST whole genome DNA standards



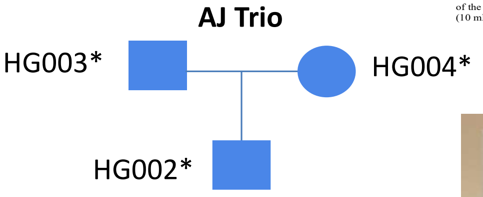

National Institute of Standards & Technology

Report of Investigation

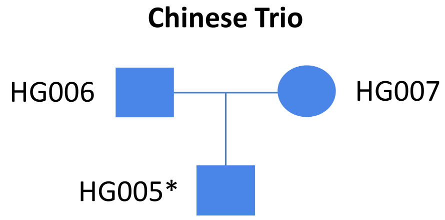
Reference Material 8391

Human DNA for Whole-Genome Variant Assessment
(Son of Eastern European Ashkenazim Jewish Ancestry)

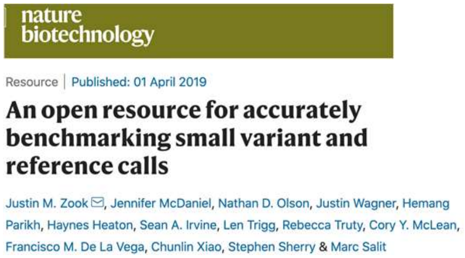
This Reference Material (RM) is intended for validation, optimization, and process evaluation purposes. It consists of a male whole human genome sample of Eastern European Ashkenazim Jewish ancestry, and it can be used to assess performance of variant calling from genome sequencing. A unit of RM 8391 consists of a vial containing human genomic DNA extracted from a single large growth of human lymphoblastoid cell line GM24385 from the Coriell Institute for Medical Research (Camden, NJ). The vial contains approximately 10 µg of genomic DNA, with the peak of the nominal length distribution longer than 48.5 kb, as referenced by Lambda DNA, and the DNA is in TE buffer (10 mM TRIS, 1 mM EDTA, pH 8.0).



REFERENCE MATERIAL
8392
Human DNA for Whole-Genome Variant Assessment
(Family Trio of Eastern European Ashkenazim Jewish Ancestry)
(HG-002, HG-003, HG-004)
NIST
National Institute of Standards and Technology
U.S. Department of Commerce
100 Bureau Drive
Gaithersburg, MD 20899
NOTICE AND WARNING:
Intended for research use. This is a human-source material. Handle as a Biosafety Level 1 material potentially capable of transmitting infectious disease.
Store at –20 °C. See report for more information.
http://www.nist.gov/srm
Tel. 301-975-2200
ID160809
NIST
Variant Assessment
Sample: HG-002
Sample: HG-003
Sample: HG-004



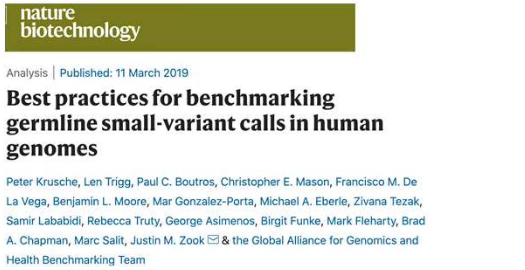
*NIST RMs developed from large batches of DNA


nature biotechnology

Resource | Published: 01 April 2019

An open resource for accurately benchmarking small variant and reference calls

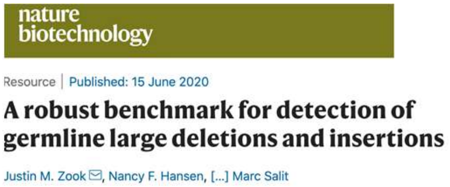
Justin M. Zook, Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, Len Trigg, Rebecca Truty, Cory Y. McLean, Francisco M. De La Vega, Chunlin Xiao, Stephen Sherry & Marc Salit


nature biotechnology

Analysis | Published: 11 March 2019

Best practices for benchmarking germline small-variant calls in human genomes

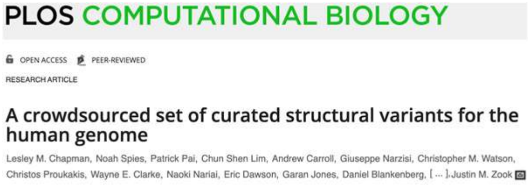
Peter Krusche, Len Trigg, Paul C. Boutros, Christopher E. Mason, Francisco M. De La Vega, Benjamin L. Moore, Mar Gonzalez-Porta, Michael A. Eberle, Zivana Tezak, Samir Lababidi, Rebecca Truty, George Asimenos, Birgit Funke, Mark Fleharty, Brad A. Chapman, Marc Salit, Justin M. Zook & the Global Alliance for Genomics and Health Benchmarking Team


nature biotechnology

Resource | Published: 15 June 2020

A robust benchmark for detection of germline large deletions and insertions

Justin M. Zook, Nancy F. Hansen, [...] Marc Salit


PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

A crowdsourced set of curated structural variants for the human genome

Lesley M. Chapman, Noah Spies, Patrick Pai, Chun Shen Lim, Andrew Carroll, Giuseppe Narzisi, Christopher M. Watson, Christos Proukakis, Wayne E. Clarke, Naoki Nariai, Eric Dawson, Garan Jones, Daniel Blankenberg, [...],Justin M. Zook

Reference samples

The Harvard Personal Genome Project — we are open for science
seracare — Passion. Innovation. Precision.
NIST
Coriell Institute for Medical Research — Decoding the Genome
horizon — Inspired Cell Solutions
ThermoFisher Scientific

Clinical Laboratories

INVITAE
AMP — Association for Molecular Pathology
Personalis — Pioneering Genome-Guided Medicine

Academic Laboratories

Baylor College of Medicine
Stanford Medicine
Johns Hopkins University
Harvard T.H. Chan School of Public Health

NGS technology developers

ThermoFisher Scientific
Oxford Nanopore Technologies
illumina
bionano Genomics
Complete Genomics — A BGI Company
PACBIO
10x Genomics
Nabsys

Government

National Institutes of Health
FDA*
NIST*
CDC — Centers for Disease Control and Prevention

* Funders

Bioinformatics developers

Google
SevenBridges
Sentieon — Precision Data for Precision Medicine
Arvados
verily
Roche
DNAnexus

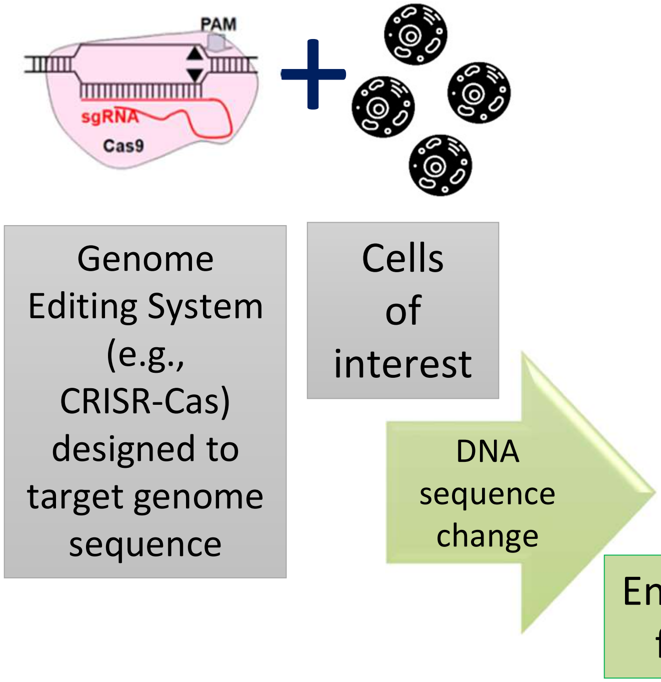
CRISPR technologies and uses

CRISPR-Cas

CRISPR - **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats

Cas - **C**RISPR **as**sociated protein

CRISPR-Cas system were identified in nature as bacterial immune systems and have been pivoted to enable modification of the genetic code within cells at designed target positions (**genome editing)**

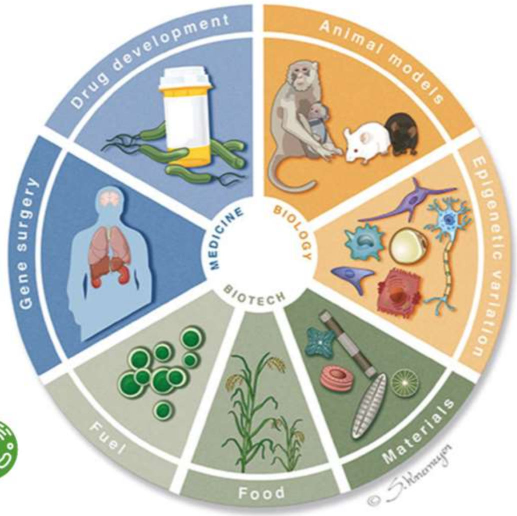


Genome Editing System (e.g., CRISR-Cas) designed to target genome sequence

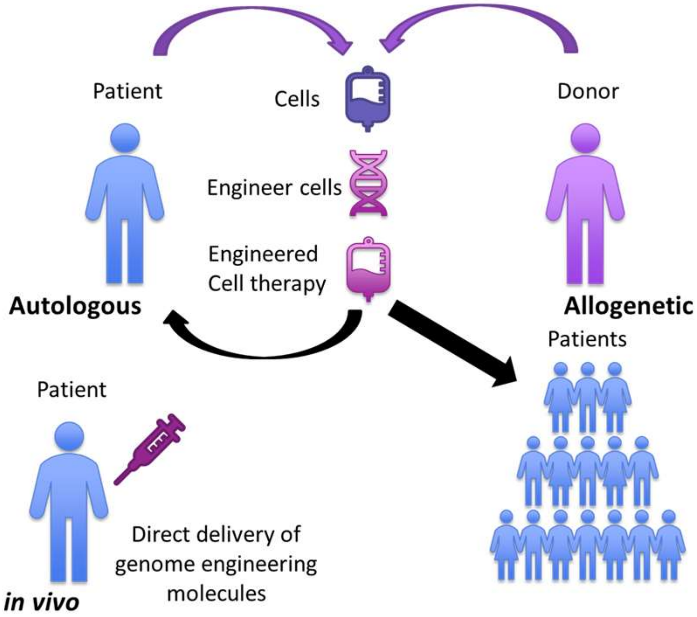
Cells of interest

DNA sequence change

Engineered/edited functional cells



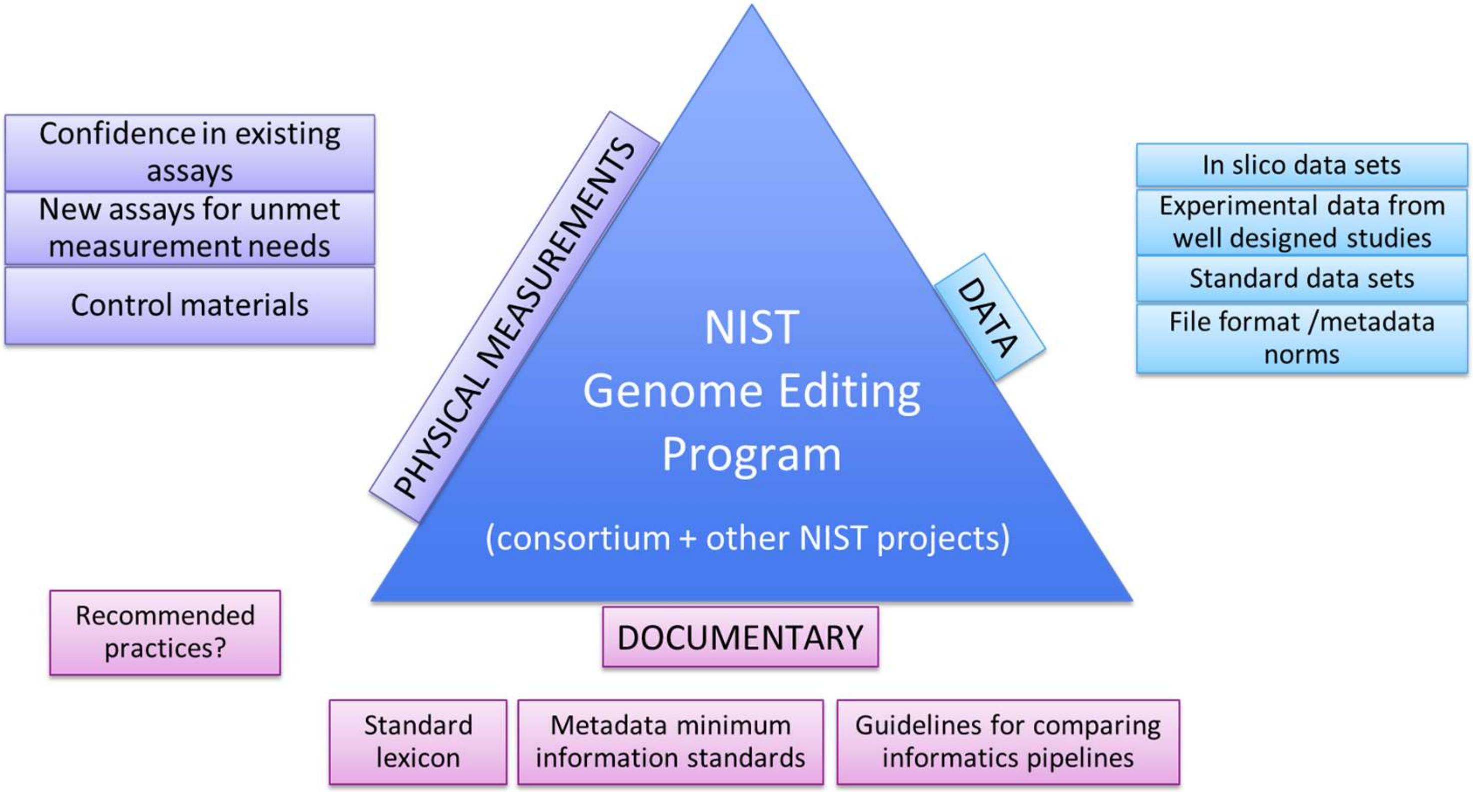
Human Gene Therapy applications



NIST experience with CRISPR technologies

Vision: Support quality in measurements for translating genome edited product to market

Goal: Develop measurement tools standards to increase the confidence of utilizing genome editing technologies in research and commercial products.



NIST Genome Editing Program

(consortium + other NIST projects)

PHYSICAL MEASUREMENTS

- Confidence in existing assays
- New assays for unmet measurement needs
- Control materials

DATA

- In slico data sets
- Experimental data from well designed studies
- Standard data sets
- File format /metadata norms

DOCUMENTARY

- Recommended practices?
- Standard lexicon
- Metadata minimum information standards
- Guidelines for comparing informatics pipelines

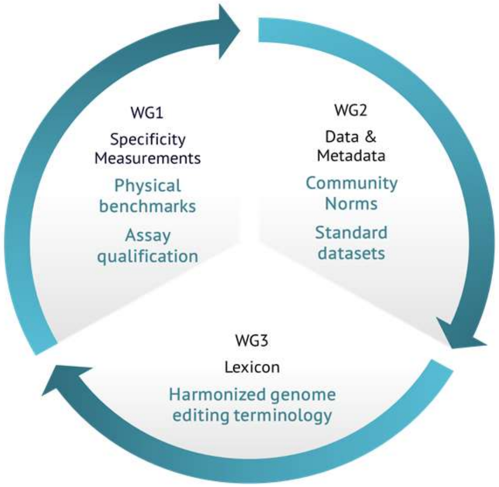
NIST Genome Editing Consortium
(launched October 2018)

NIST

MISSION

Convene experts across academia, industry, non-profit & government to addresses the measurements and standards needed to increase confidence of utilizing genome editing technologies in research and commercial products

ORGANIZATION



MEMBER BENEFITS

- Access to a neutral forum for addressing pre-competitive needs
- Participation in the development of experimental benchmarks, guidelines and terminology
- Access to tools developed by the consortium ahead of public release

MEMBERS

- Agilent
- Aldevron
- Applied StemCell
- AstraZeneca
- Bionano Genomics
- Bio-Rad
- Bluebird bio*
- Caribou Biosciences
- Catalytic Data Science
- Cergentis
- COBO Technologies
- **College of American Pathologists (CAP)**
- CRISPR Therapeutics
- DARPA
- DowDuPont Agroscience (Corteva)*
- Editas Medicine
- EMBL-EBI
- **FDA CBER**
- Genomic Vision
- Horizon Discovery
- Illumina
- Inscripta
- Integrated DNA Technologies
- Intellia Therapeutics
- KromaTiD
- Lonza
- Macrogen
- Mass General Hospital
- Mission Bio
- Novartis
- New England Biolabs
- NIH/NINDS
- NIH SCGE
- Precision Biosciences
- Sangamo Therapeutics
- SeQure Dx
- St. Jude Children's Research Hospital
- Synthego
- ThermoFisher Scientific
- Twinstrand Biosciences
- UCSC
- WhiteLab Genomics

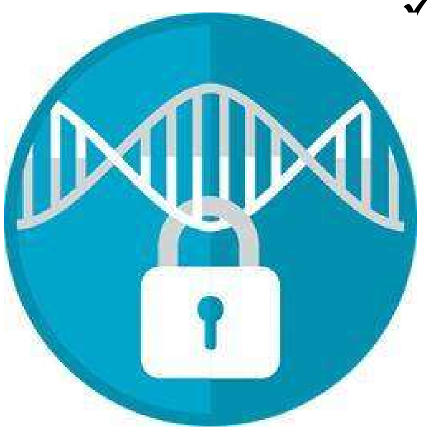
NIST coordinates with FDA and Center for Veterinary Medicine (CVM)

Cost sharing model. All members contribute $20,000 annually or *in-kind*

* Former members

Importance of securing human genomics data

- ✓ Data integrity
- ✓ Data reliability
- ✓ Maintain limited access to individual genomic information
- ✓ Protect knowledge / intellectual property
- ✓ Data reusability / prevent loss
- ✓ Prevent against nefarious use or misuse
- ✓ Privacy



Credit: mcmurryjulie

Contact Us

Thank You!

Contact

Samantha Maragh
Leader, Genome Editing Program

Email address

samantha@nist.gov

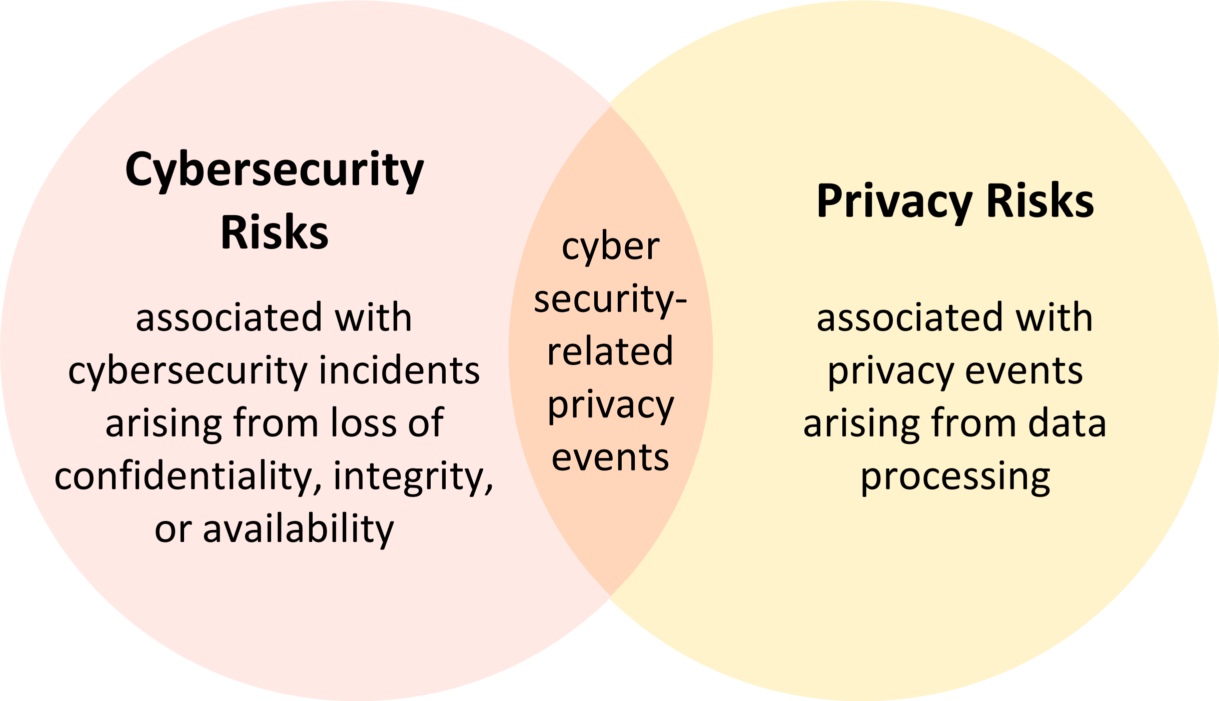
National Institute of
Standards and Technology
U.S. Department of Commerce

Privacy at NIST

Naomi Lefkovitz

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

Relationship Between Cybersecurity and Privacy Risk



Cybersecurity Risks

associated with cybersecurity incidents arising from loss of confidentiality, integrity, or availability

cyber security-related privacy events

Privacy Risks

associated with privacy events arising from data processing

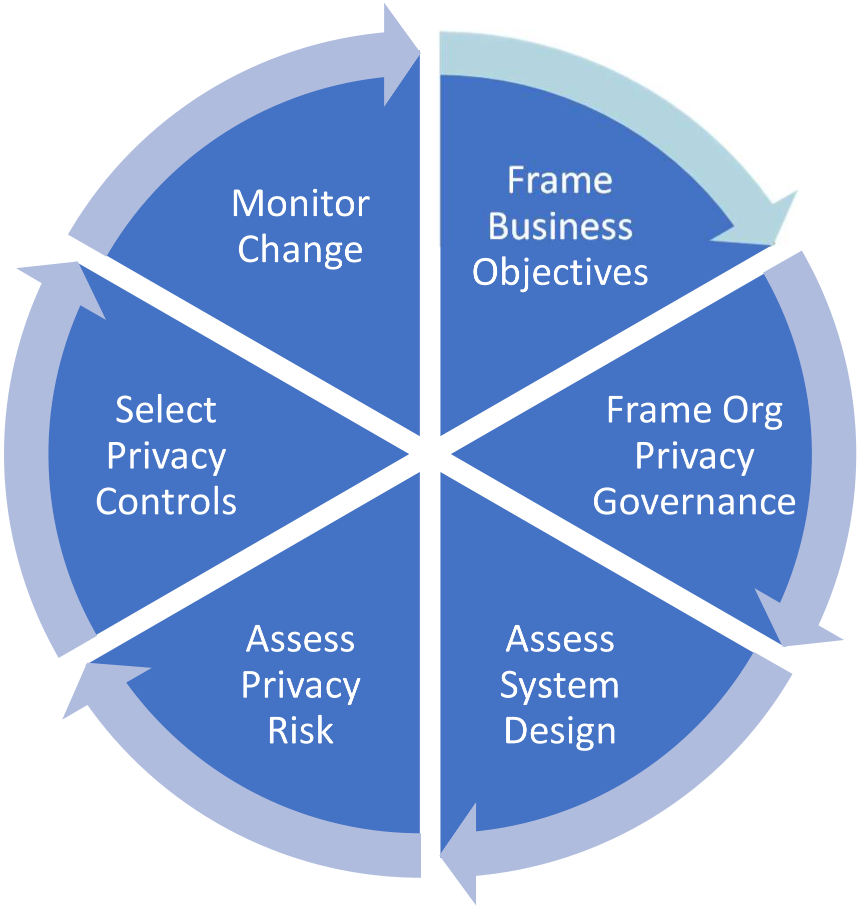
Data: A representation of information, including digital and non-digital formats

Privacy Event: The occurrence or potential occurrence of problematic data actions

Data Processing: The collective set of data actions (i.e., the complete data life cycle, including, but not limited to collection, retention, logging, generation, transformation, use, disclosure, sharing, transmission, and disposal)

Privacy Risk: The likelihood that individuals will experience problems resulting from data processing, and the impact should they occur

NIST Privacy Risk Assessment Methodology (PRAM)



Catalog of Problematic Data Actions
and Problems

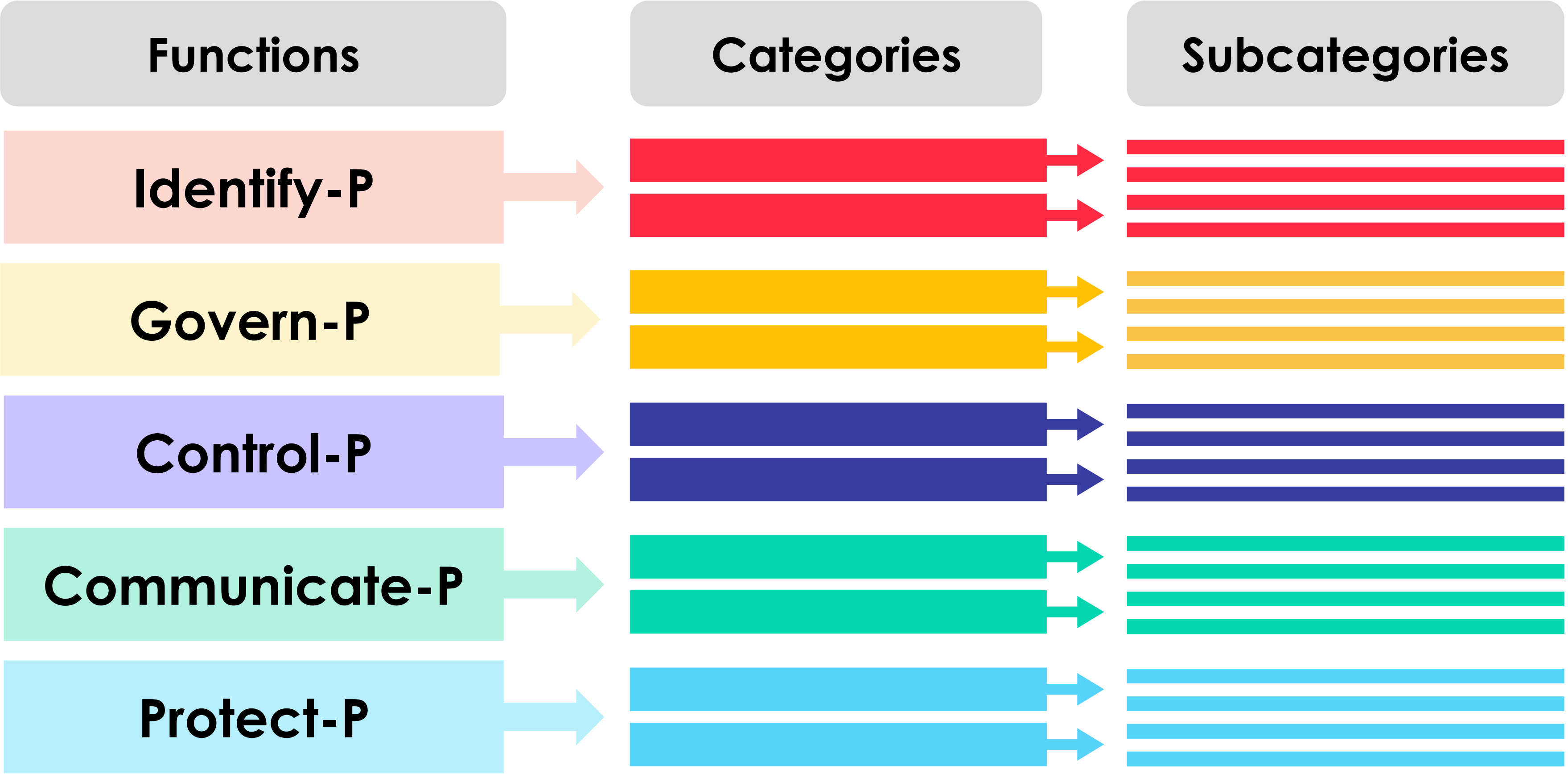
NIST Privacy Engineering Objectives

Predictability: enabling reliable assumptions by individuals, owners, and operators about data and their processing by a system, product, or service.

Manageability: providing the capability for granular administration of data, including alteration, deletion, and selective disclosure.

Disassociability: enabling the processing of data or events without association to individuals or devices beyond the operational requirements of the system.

NIST Privacy Framework Core



Functions

Categories

Subcategories

Identify-P

Govern-P

Control-P

Communicate-P

Protect-P

Resources

Websites

https://www.nist.gov/privacyframework

Mailing List

List.nist.gov/privacyframework

Contact Us

PrivacyFramework@nist.gov

National Institute of Standards and Technology - Cybersecurity

Ron Pulivarti, Senior Cybersecurity Engineer for the Healthcare Sector at the National Cybersecurity Center of Excellence (NCCoE), which is part of NIST

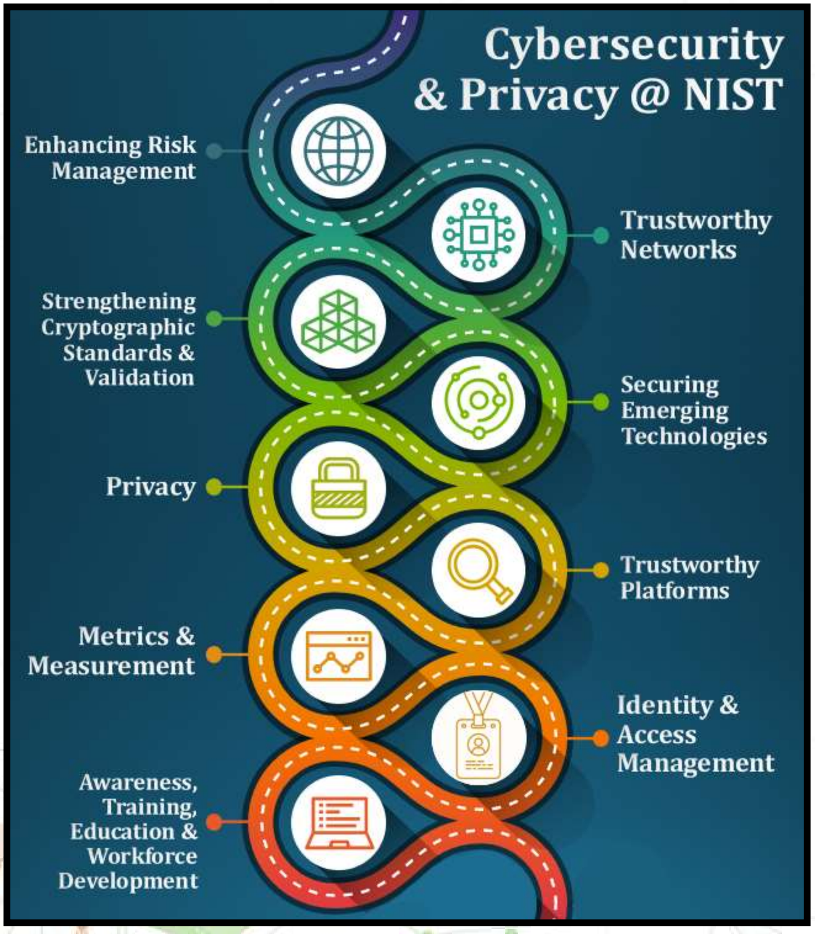
NIST National Institute of Standards and Technology U.S. Department of Commerce

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

NIST
National Institute of Standards and Technology
U.S. Department of Commerce
Celebrating 50 years of Cybersecurity at NIST

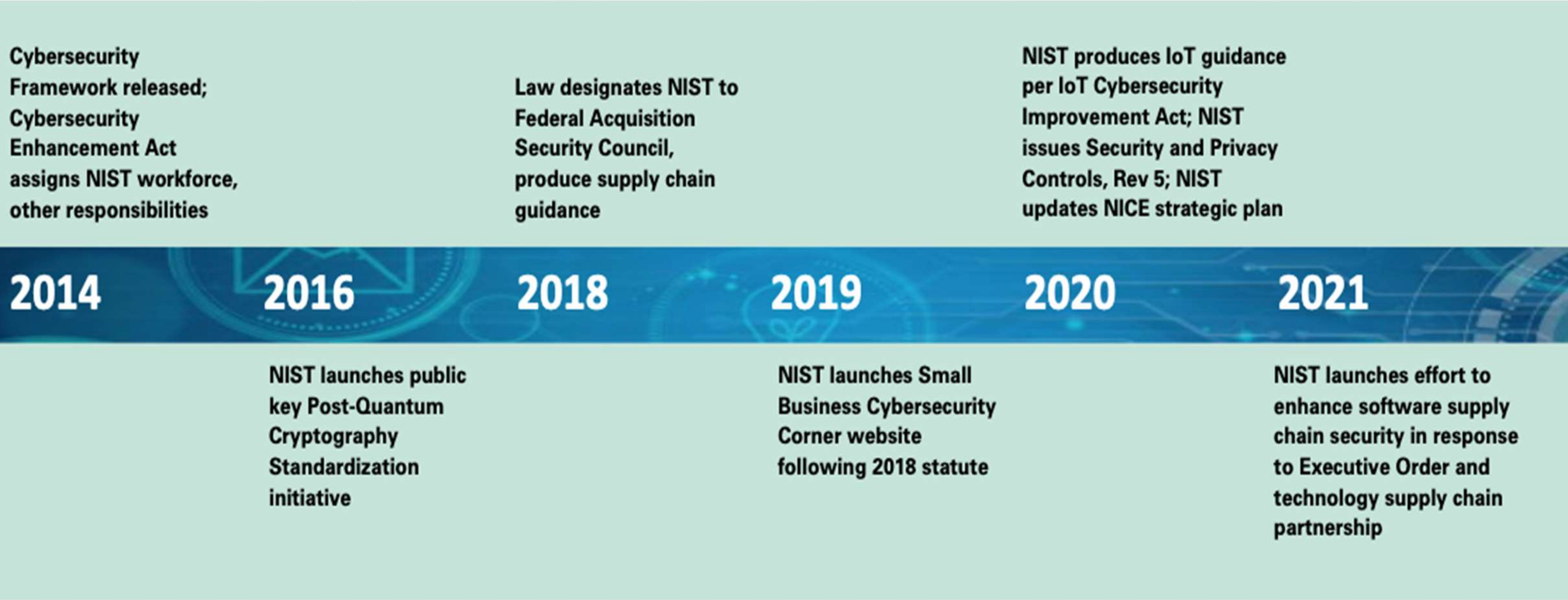
NIST National Institute of Standards and Technology U.S. Department of Commerce

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE


Cybersecurity & Privacy @ NIST
Enhancing Risk Management
Trustworthy Networks
Strengthening Cryptographic Standards & Validation
Securing Emerging Technologies
Privacy
Trustworthy Platforms
Metrics & Measurement
Identity & Access Management
Awareness, Training, Education & Workforce Development

NINE PRIORITY AREAS:

Enhancing Risk Management
Trustworthy Networks
Strengthening Cryptographic Standards & Validation
Securing Emerging Technologies
Privacy
Trustworthy Platforms
Metrics & Measurement
Identity & Access Management
Awareness, Training, Education & Workforce Development



2014

Cybersecurity Framework released; Cybersecurity Enhancement Act assigns NIST workforce, other responsibilities

2016

NIST launches public key Post-Quantum Cryptography Standardization initiative

2018

Law designates NIST to Federal Acquisition Security Council, produce supply chain guidance

2019

NIST launches Small Business Cybersecurity Corner website following 2018 statute

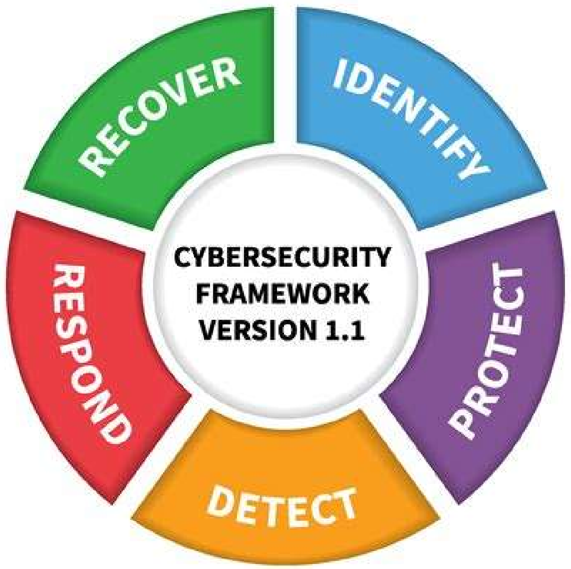
2020

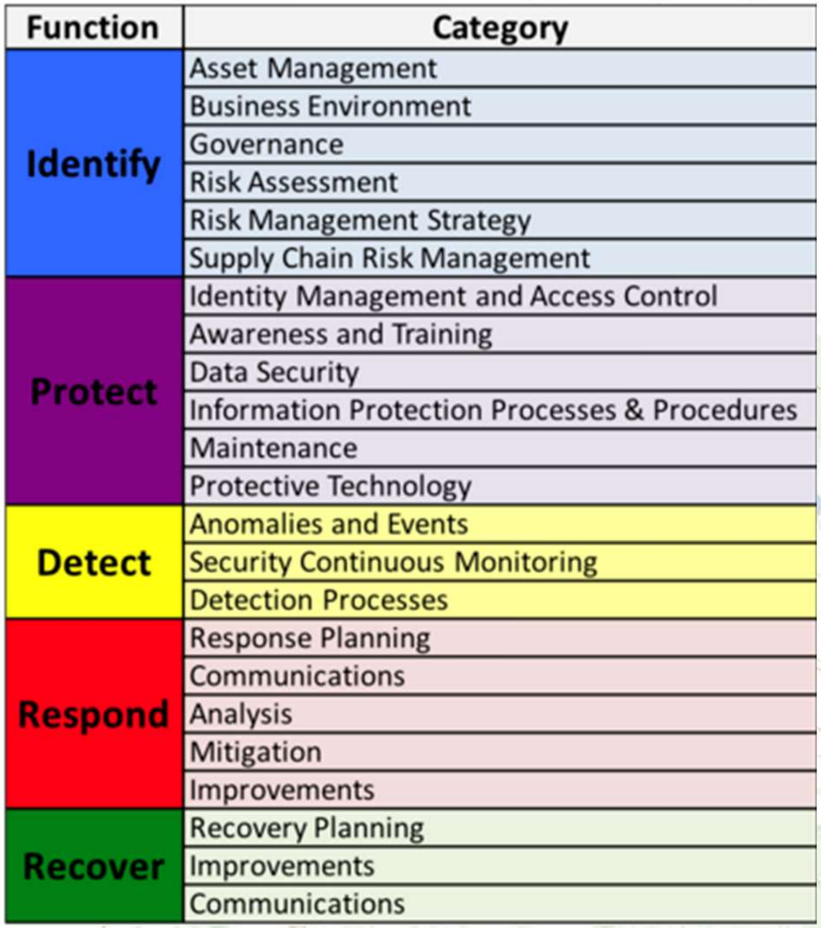
NIST produces IoT guidance per IoT Cybersecurity Improvement Act; NIST issues Security and Privacy Controls, Rev 5; NIST updates NICE strategic plan

2021

NIST launches effort to enhance software supply chain security in response to Executive Order and technology supply chain partnership

NIST Cybersecurity Framework 1.1

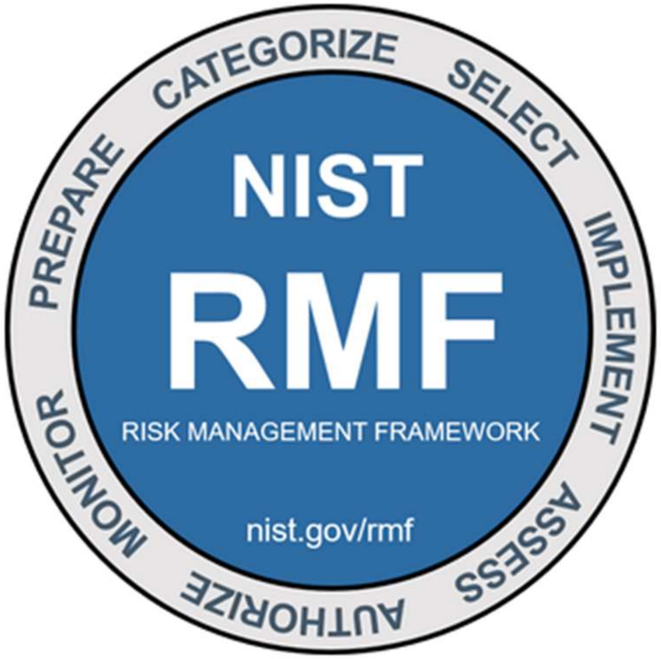


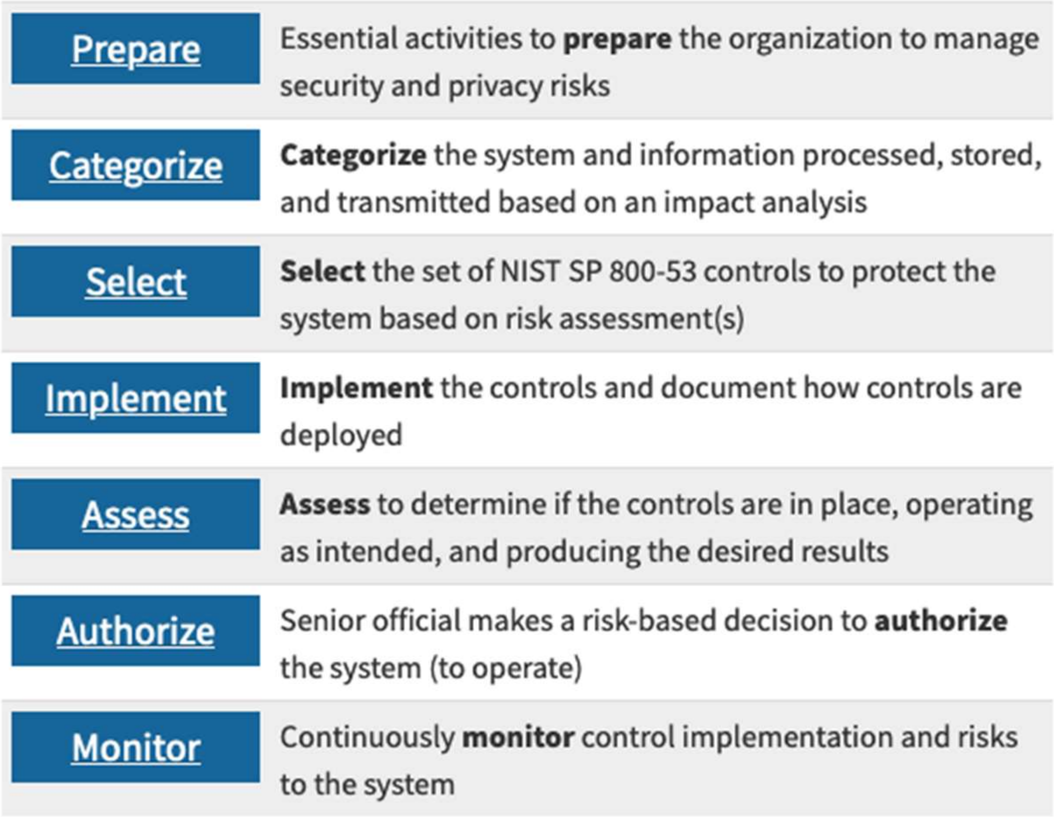


Function	Category
Identify	Asset Management
	Business Environment
	Governance
	Risk Assessment
	Risk Management Strategy
	Supply Chain Risk Management
Protect	Identity Management and Access Control
	Awareness and Training
	Data Security
	Information Protection Processes & Procedures
	Maintenance
	Protective Technology
Detect	Anomalies and Events
	Security Continuous Monitoring
	Detection Processes
Respond	Response Planning
	Communications
	Analysis
	Mitigation
	Improvements
Recover	Recovery Planning
	Improvements
	Communications

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

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE





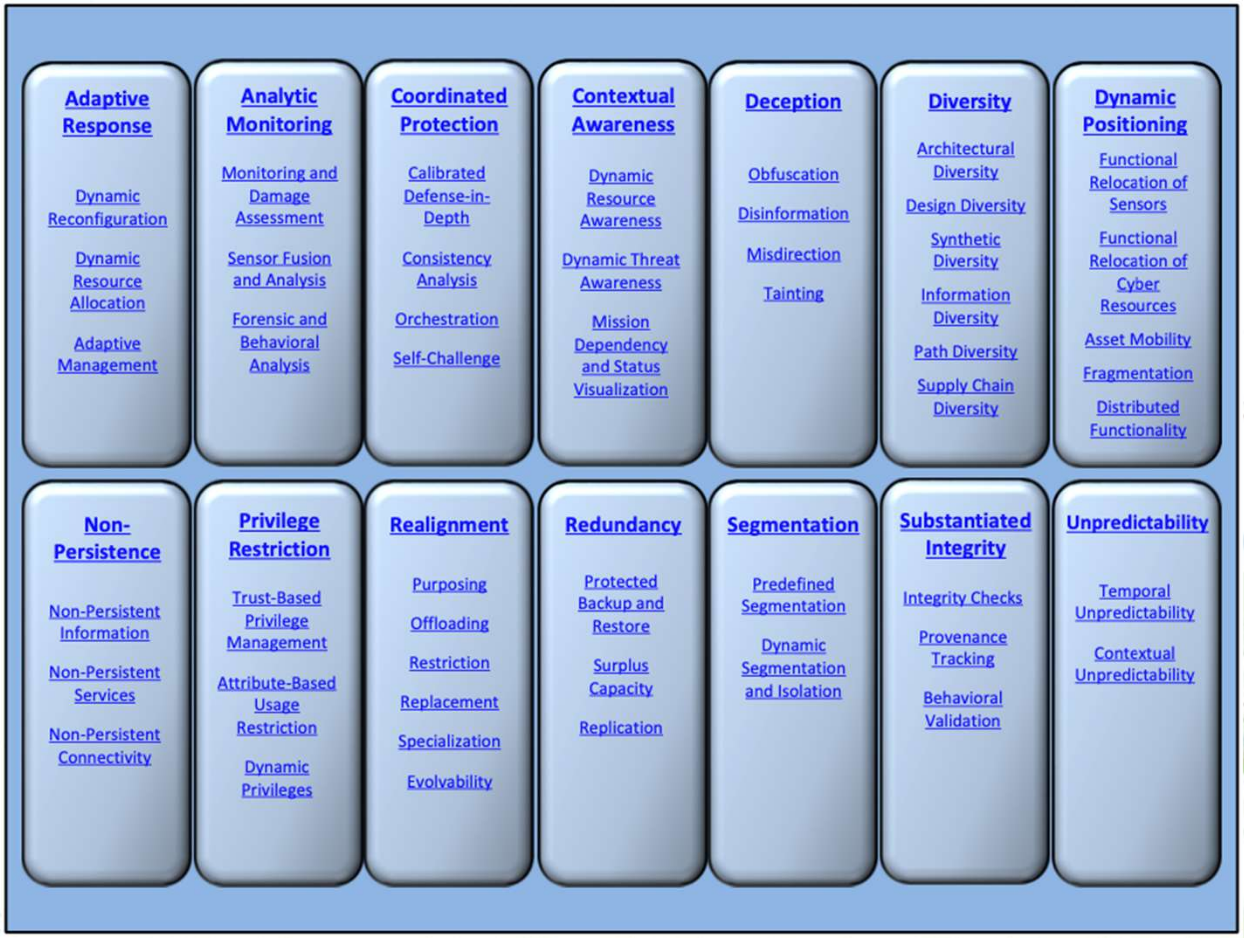
Prepare	Essential activities to **prepare** the organization to manage security and privacy risks
Categorize	**Categorize** the system and information processed, stored, and transmitted based on an impact analysis
Select	**Select** the set of NIST SP 800-53 controls to protect the system based on risk assessment(s)
Implement	**Implement** the controls and document how controls are deployed
Assess	**Assess** to determine if the controls are in place, operating as intended, and producing the desired results
Authorize	Senior official makes a risk-based decision to **authorize** the system (to operate)
Monitor	Continuously **monitor** control implementation and risks to the system

NIST Special Publication 800-160, Volume 2
Revision 1

Developing Cyber-Resilient Systems:

A Systems Security Engineering Approach

DEPARTMENT OF COMMERCE
UNITED STATES OF AMERICA



Adaptive Response
- Dynamic Reconfiguration
- Dynamic Resource Allocation
- Adaptive Management

Analytic Monitoring
- Monitoring and Damage Assessment
- Sensor Fusion and Analysis
- Forensic and Behavioral Analysis

Coordinated Protection
- Calibrated Defense-in-Depth
- Consistency Analysis
- Orchestration
- Self-Challenge

Contextual Awareness
- Dynamic Resource Awareness
- Dynamic Threat Awareness
- Mission Dependency and Status Visualization

Deception
- Obfuscation
- Disinformation
- Misdirection
- Tainting

Diversity
- Architectural Diversity
- Design Diversity
- Synthetic Diversity
- Information Diversity
- Path Diversity
- Supply Chain Diversity

Dynamic Positioning
- Functional Relocation of Sensors
- Functional Relocation of Cyber Resources
- Asset Mobility
- Fragmentation
- Distributed Functionality

Non-Persistence
- Non-Persistent Information
- Non-Persistent Services
- Non-Persistent Connectivity

Privilege Restriction
- Trust-Based Privilege Management
- Attribute-Based Usage Restriction
- Dynamic Privileges

Realignment
- Purposing
- Offloading
- Restriction
- Replacement
- Specialization
- Evolvability

Redundancy
- Protected Backup and Restore
- Surplus Capacity
- Replication

Segmentation
- Predefined Segmentation
- Dynamic Segmentation and Isolation

Substantiated Integrity
- Integrity Checks
- Provenance Tracking
- Behavioral Validation

Unpredictability
- Temporal Unpredictability
- Contextual Unpredictability

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

NCCoE
NATIONAL CYBERSECURITY
CENTER OF EXCELLENCE

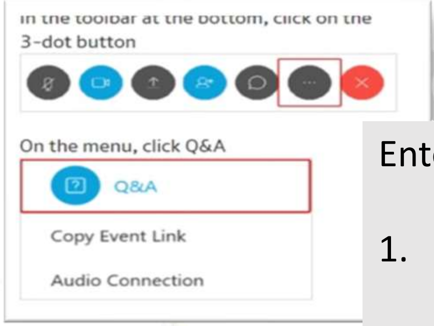
Thank You.

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

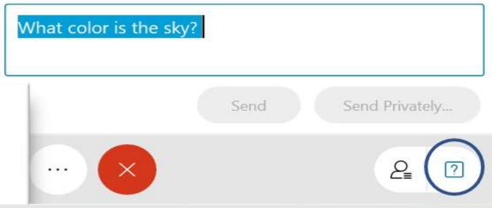
NIST Experiences in Genomics, Cybersecurity, and Privacy

Moderated Questions and Answers



In the toolbar at the bottom, click on the 3-dot button

On the menu, click Q&A



What color is the sky?

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

National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

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

NIST 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)

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.

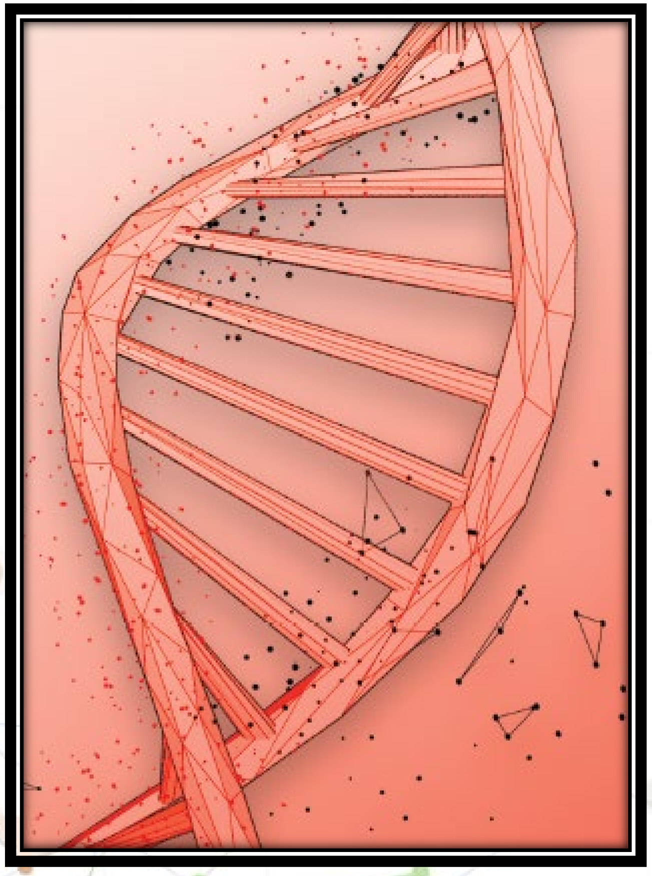
INTELLIGENCE: By combining genetic data with other PII and healthcare / lifestyle data the PRC has acquired through cyberattacks and other means, the PRC could use this information to target U.S. personnel, dissidents, journalists, and others around the world for potential surveillance, manipulation,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

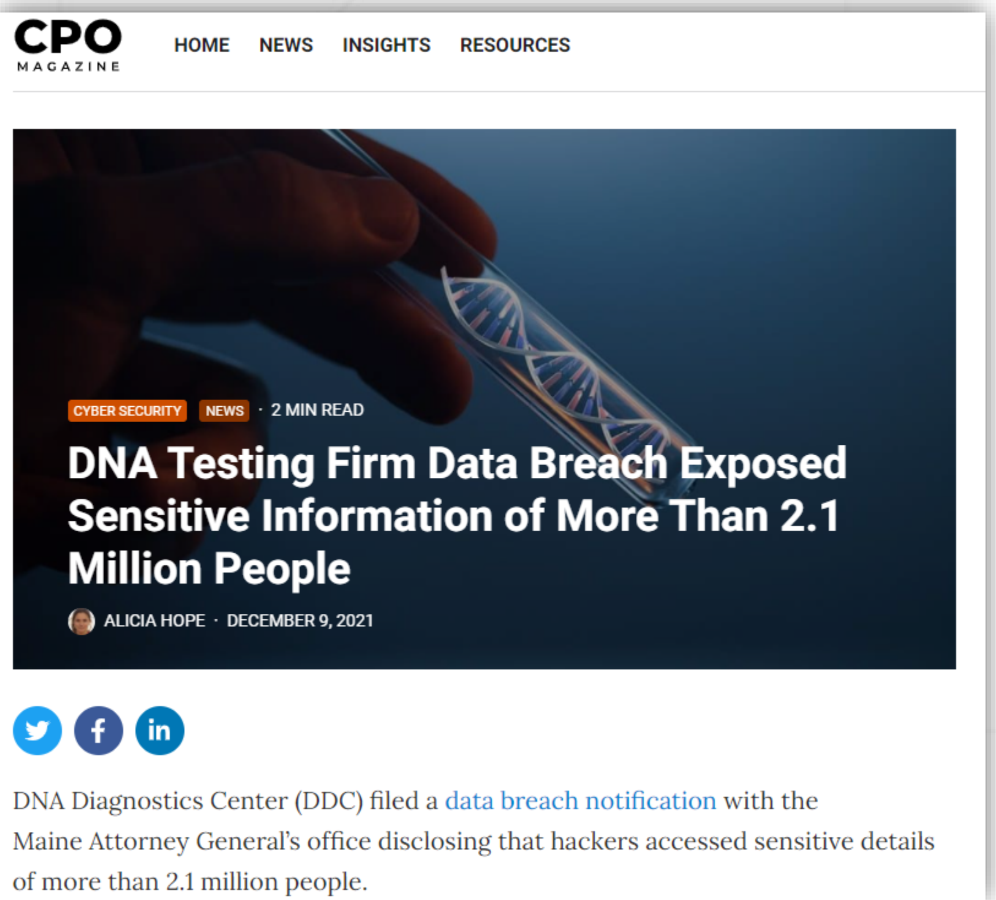


DESIRABLE TARGETS: DTC genetic testing companies, information exchanges, and data libraries are desirable targets for foreign adversaries, cyber criminals, and insider threats.

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



CPO MAGAZINE
HOME NEWS INSIGHTS RESOURCES
CYBER SECURITY NEWS · 2 MIN READ

DNA Testing Firm Data Breach Exposed Sensitive Information of More Than 2.1 Million People

ALICIA HOPE · DECEMBER 9, 2021

DNA Diagnostics Center (DDC) filed a data breach notification with the Maine Attorney General's office disclosing that hackers accessed sensitive details of more than 2.1 million people.

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.

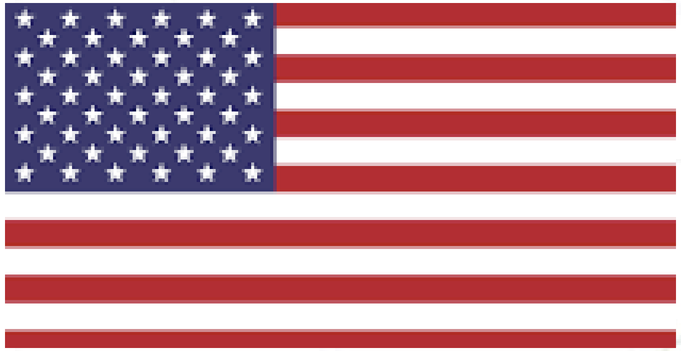
- Foreign companies and authoritarian regimes have already gained significant access to U.S. genomic data and related healthcare data through investments, research partnerships, contractual agreements, and other means.

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)

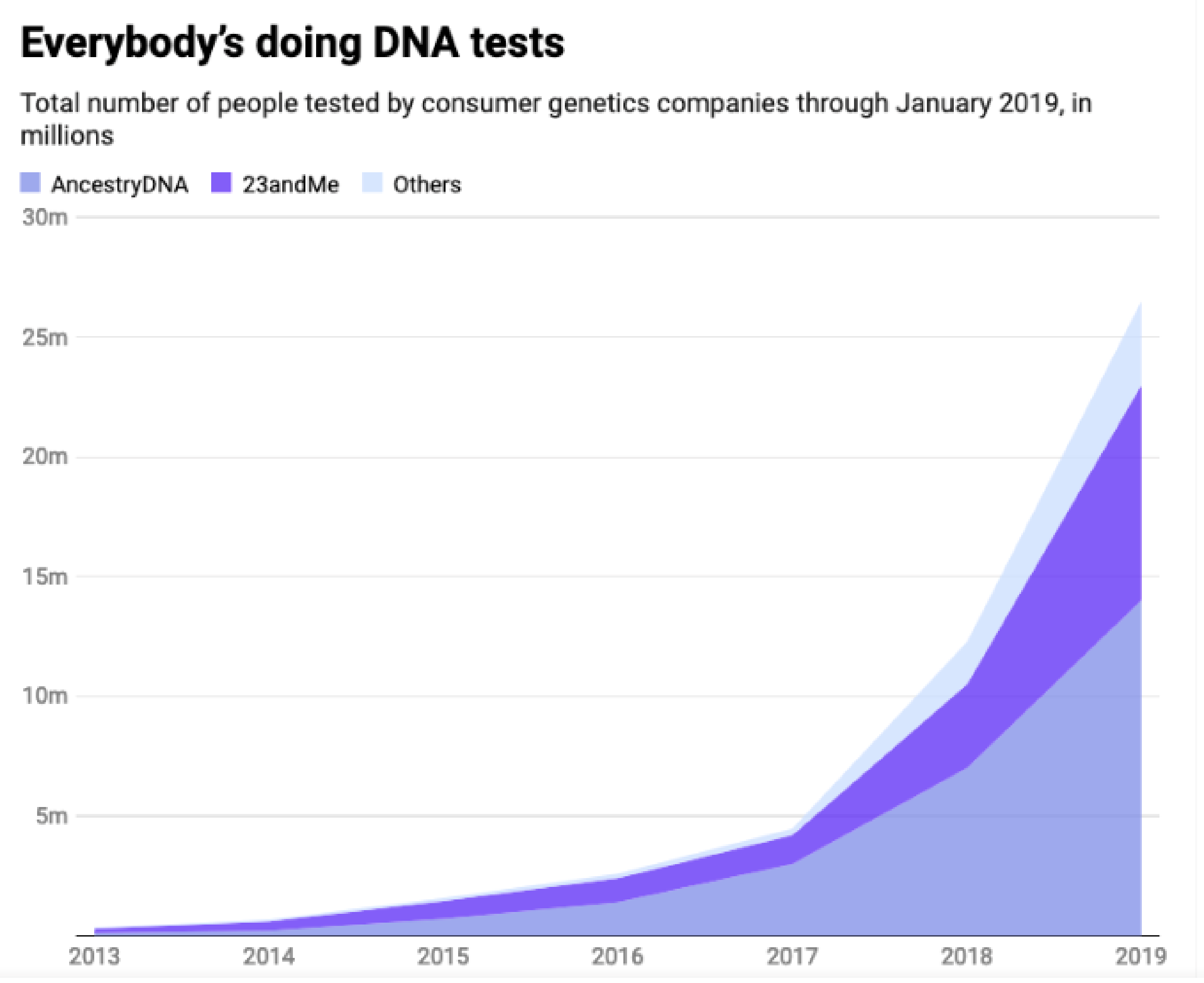
Free-for-all genetic surveillance nation



- Dr. Yaniv Erlich

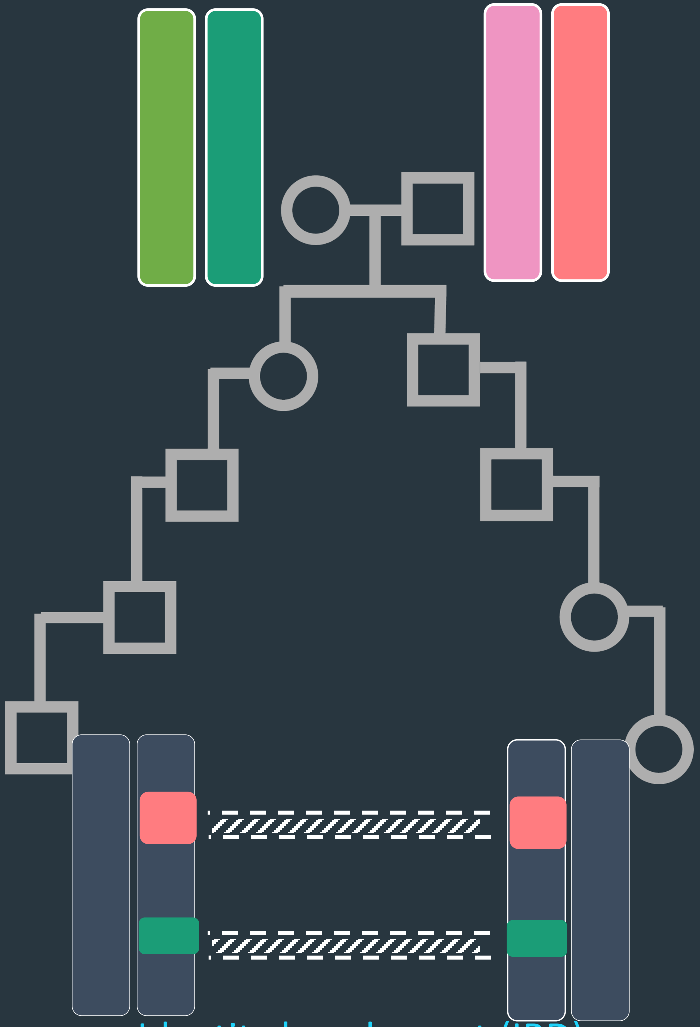
@erlichya

The advent of consumer genomics

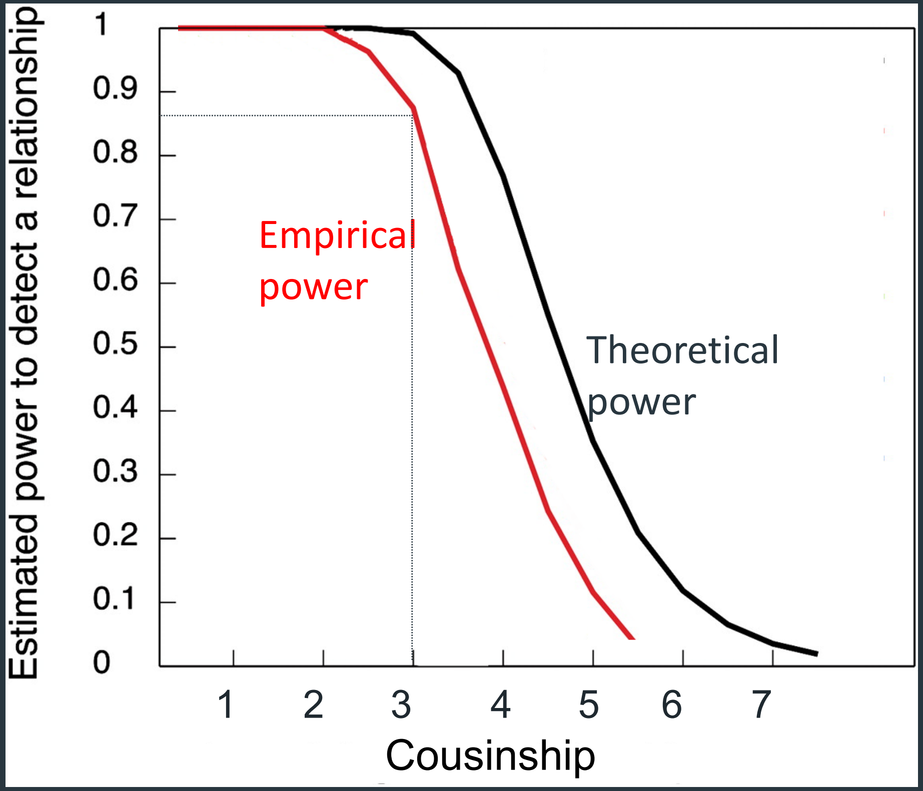


MIT Technology review

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

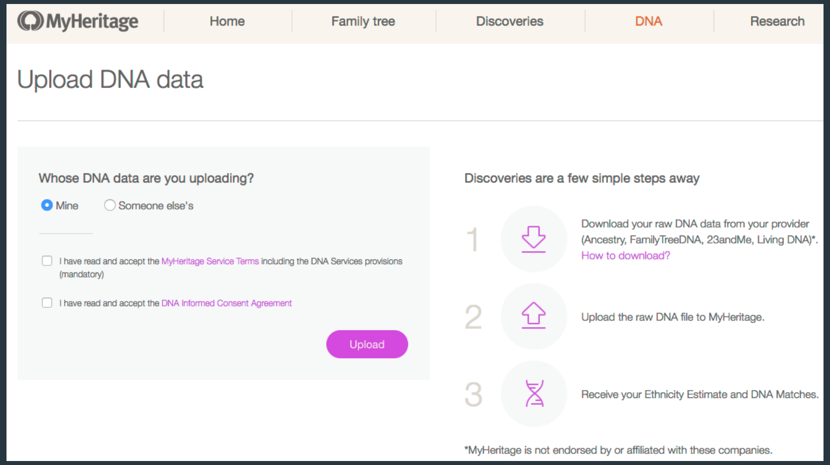
Your raw genetic data

```
# 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)



Upload DNA data

Whose DNA data are you uploading?

● Mine ○ Someone else's

□ I have read and accept the MyHeritage Service Terms including the DNA Services provisions (mandatory)

□ I have read and accept the DNA Informed Consent Agreement

Upload

Discoveries are a few simple steps away

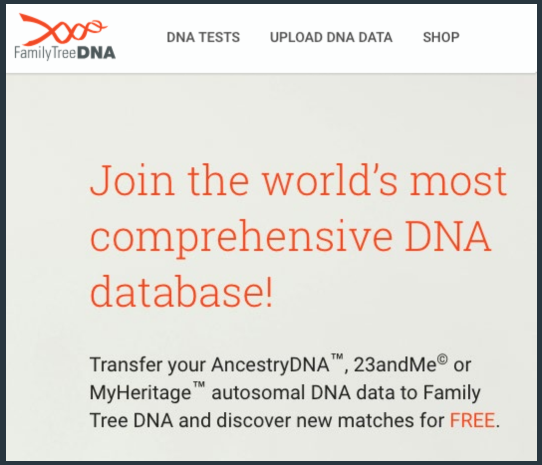
1 Download your raw DNA data from your provider (Ancestry, FamilyTreeDNA, 23andMe, Living DNA)*. How to download?

2 Upload the raw DNA file to MyHeritage.

3 Receive your Ethnicity Estimate and DNA Matches.

*MyHeritage is not endorsed by or affiliated with these companies.

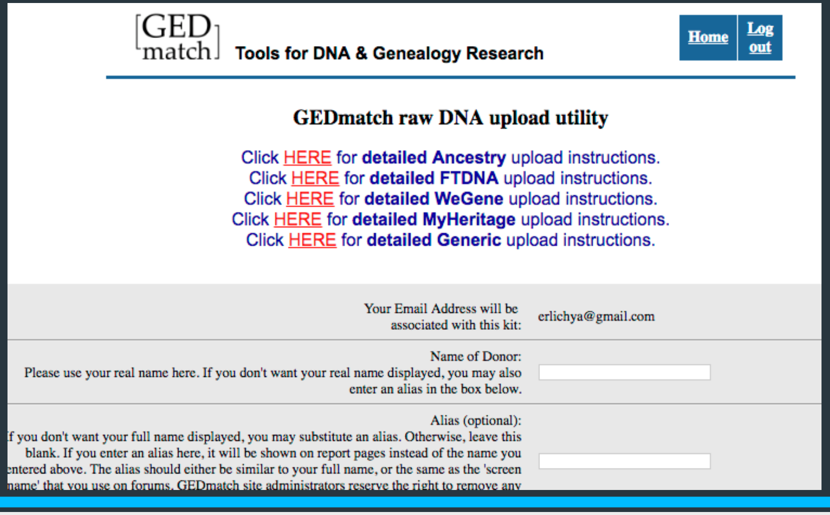
FTDNA (users: 1M)



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)



GEDmatch raw DNA upload utility

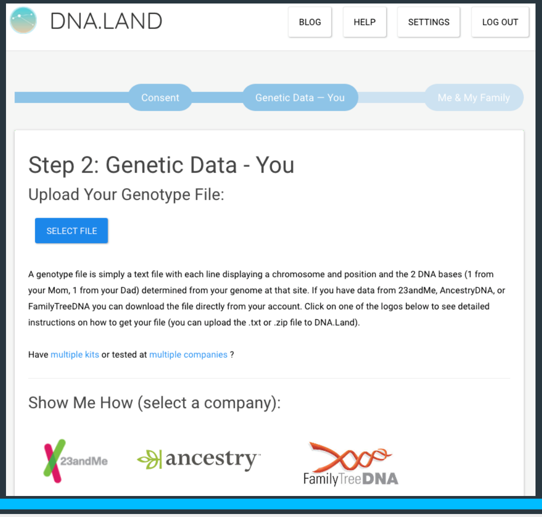
Click HERE for detailed Ancestry upload instructions.
Click HERE for detailed FTDNA upload instructions.
Click HERE for detailed WeGene upload instructions.
Click HERE for detailed MyHeritage upload instructions.
Click HERE for detailed Generic upload instructions.

Your Email Address will be associated with this kit: erlichya@gmail.com

Name of Donor: Please use your real name here. If you don't want your real name displayed, you may also enter an alias in the box below.

Alias (optional): f you don't want your full name displayed, you may substitute an alias. Otherwise, leave this blank. If you enter an alias here, it will be shown on report pages instead of the name you ntered above. The alias should either be similar to your full name, or the same as the 'screen name' that you use on forums. GEDmatch site administrators reserve the right to remove any

DNA.Land (users: 150K)



Step 2: Genetic Data - You

Upload Your Genotype File:

SELECT FILE

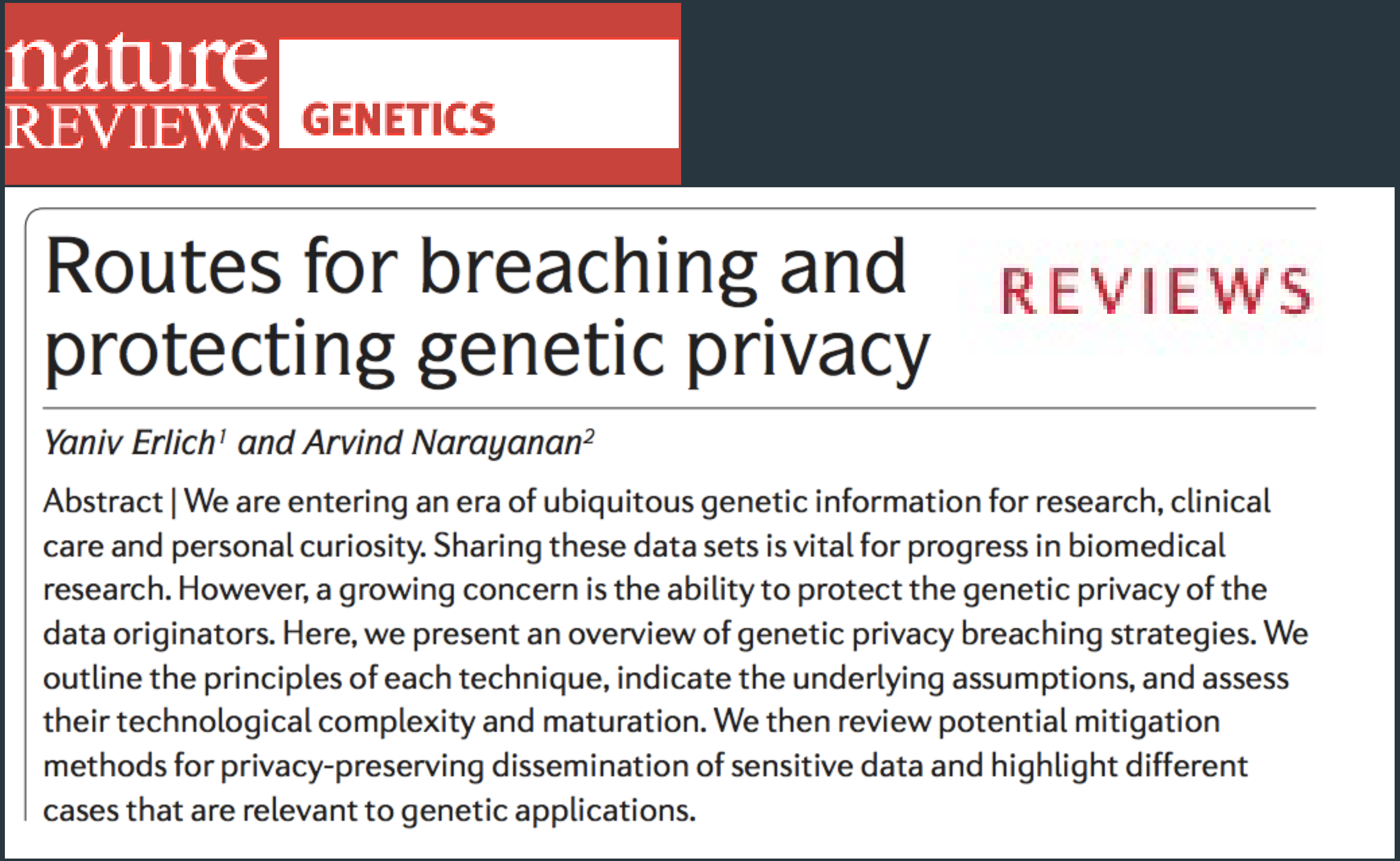
A genotype file is simply a text file with each line displaying a chromosome and position and the 2 DNA bases (1 from your Mom, 1 from your Dad) determined from your genome at that site. If you have data from 23andMe, AncestryDNA, or FamilyTreeDNA you can download the file directly from your account. Click on one of the logos below to see detailed instructions on how to get your file (you can upload the .txt or .zip file to DNA.Land).

Have multiple kits or tested at multiple companies ?

Show Me How (select a company):

3^{rd} party uploads are highly important

Using genetic genealogy for forensic is not a new idea



nature REVIEWS GENETICS

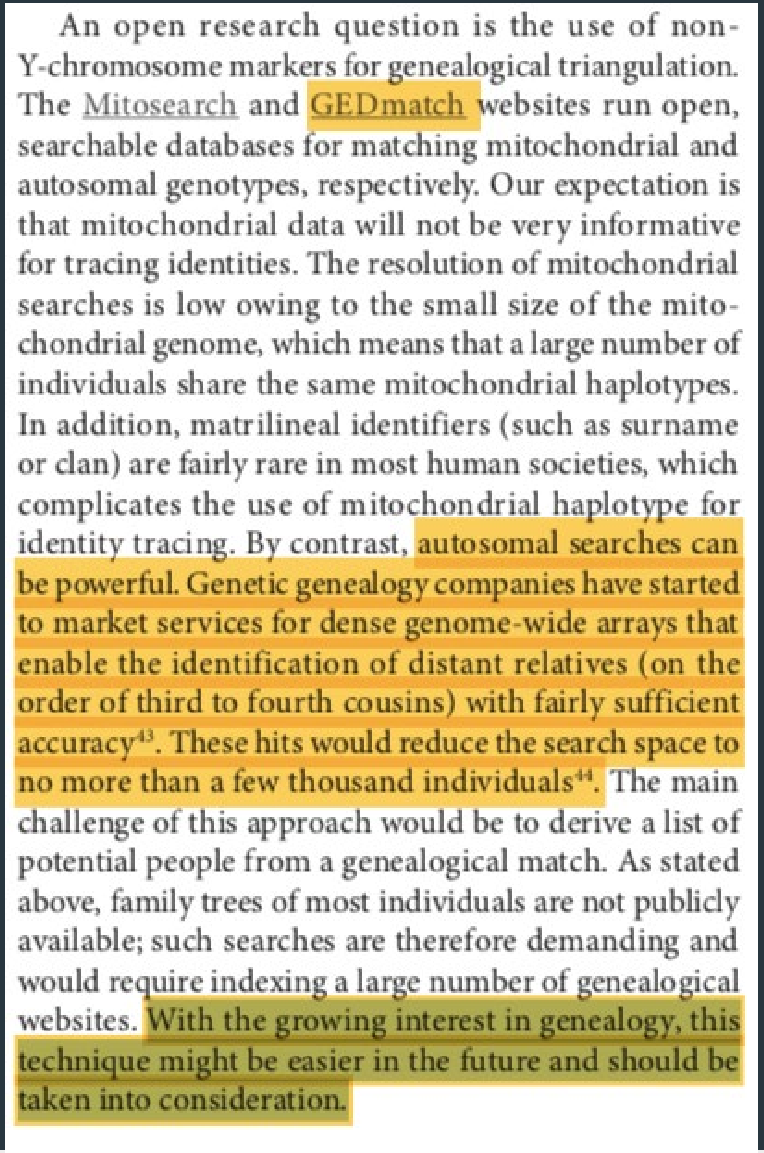
REVIEWS

Routes for breaching and protecting genetic privacy

Yaniv Erlich[1] *and Arvind Narayanan*[2]

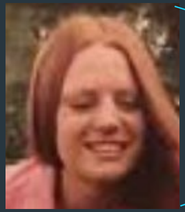
Abstract | We are entering an era of ubiquitous genetic information for research, clinical care and personal curiosity. Sharing these data sets is vital for progress in biomedical research. However, a growing concern is the ability to protect the genetic privacy of the data originators. Here, we present 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.

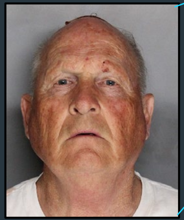
Erlich and Narayanan, 2014

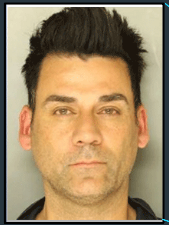


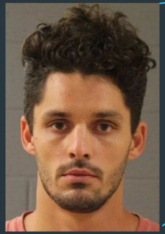
An open research question is the use of non-Y-chromosome markers for genealogical triangulation. The Mitosearch and GEDmatch websites run open, searchable databases for matching mitochondrial and autosomal genotypes, respectively. Our expectation is that mitochondrial data will not be very informative 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.

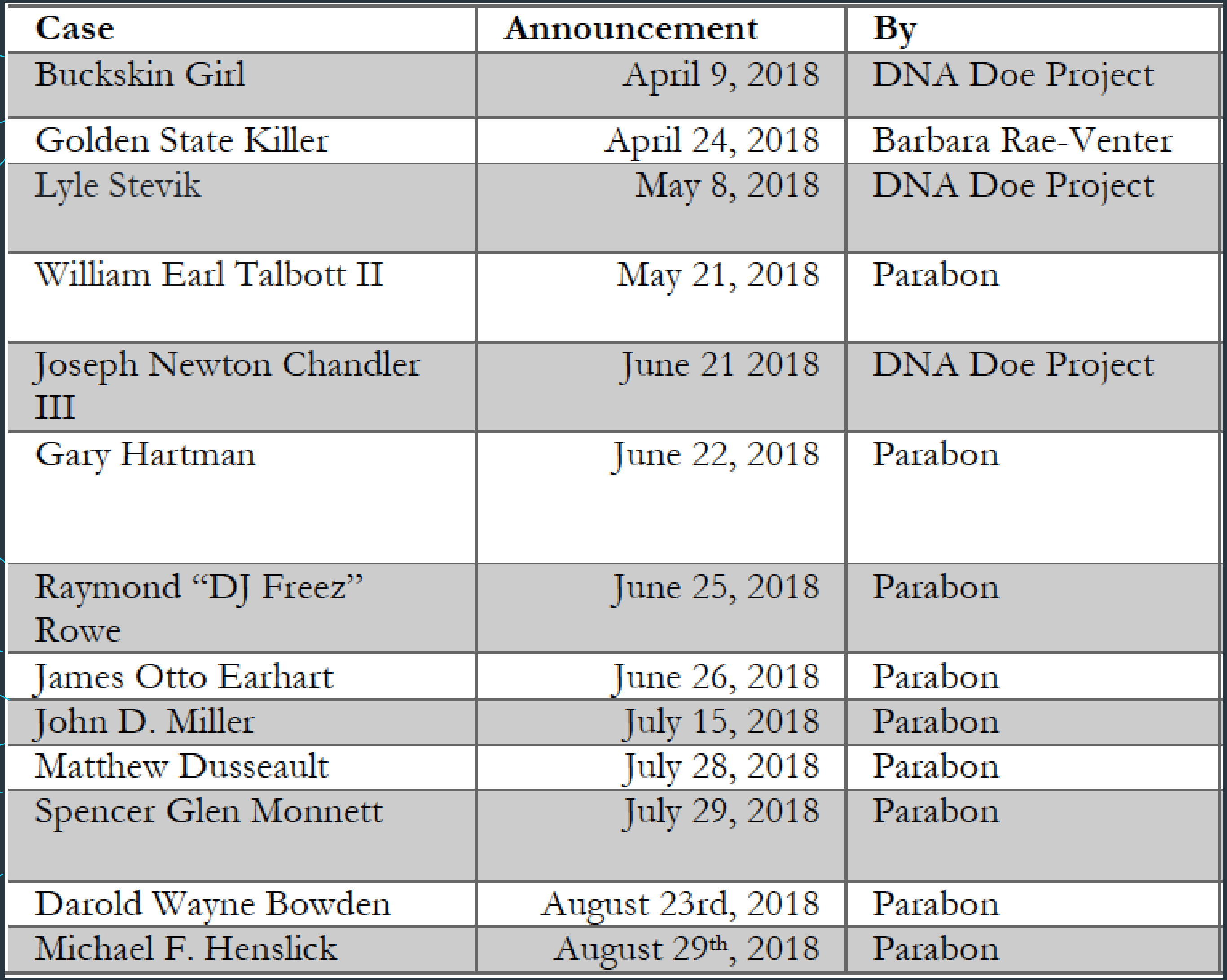
Long range familial searches





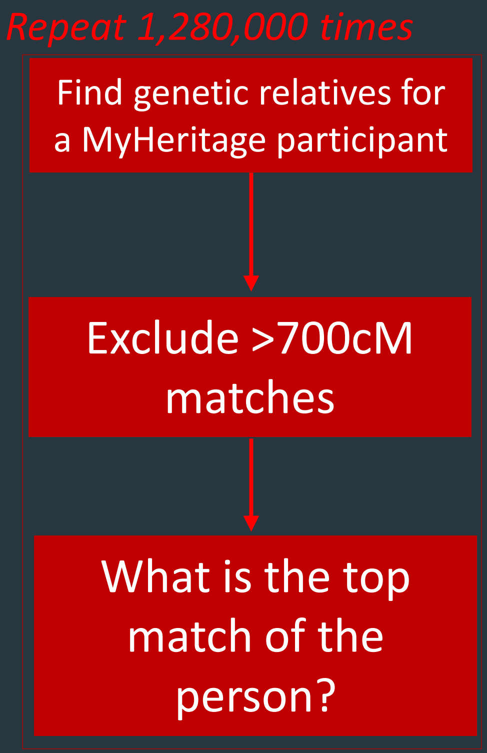






Case	Announcement	By
Buckskin Girl	April 9, 2018	DNA Doe Project
Golden State Killer	April 24, 2018	Barbara Rae-Venter
Lyle Stevik	May 8, 2018	DNA Doe Project
William Earl Talbott II	May 21, 2018	Parabon
Joseph Newton Chandler III	June 21 2018	DNA Doe Project
Gary Hartman	June 22, 2018	Parabon
Raymond "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?

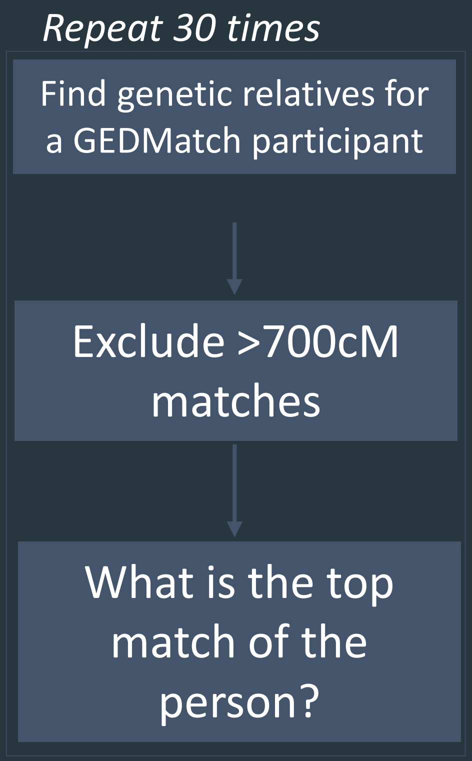


Repeat 1,280,000 times

Find genetic relatives for a MyHeritage participant

Exclude >700cM matches

What is the top match of the person?

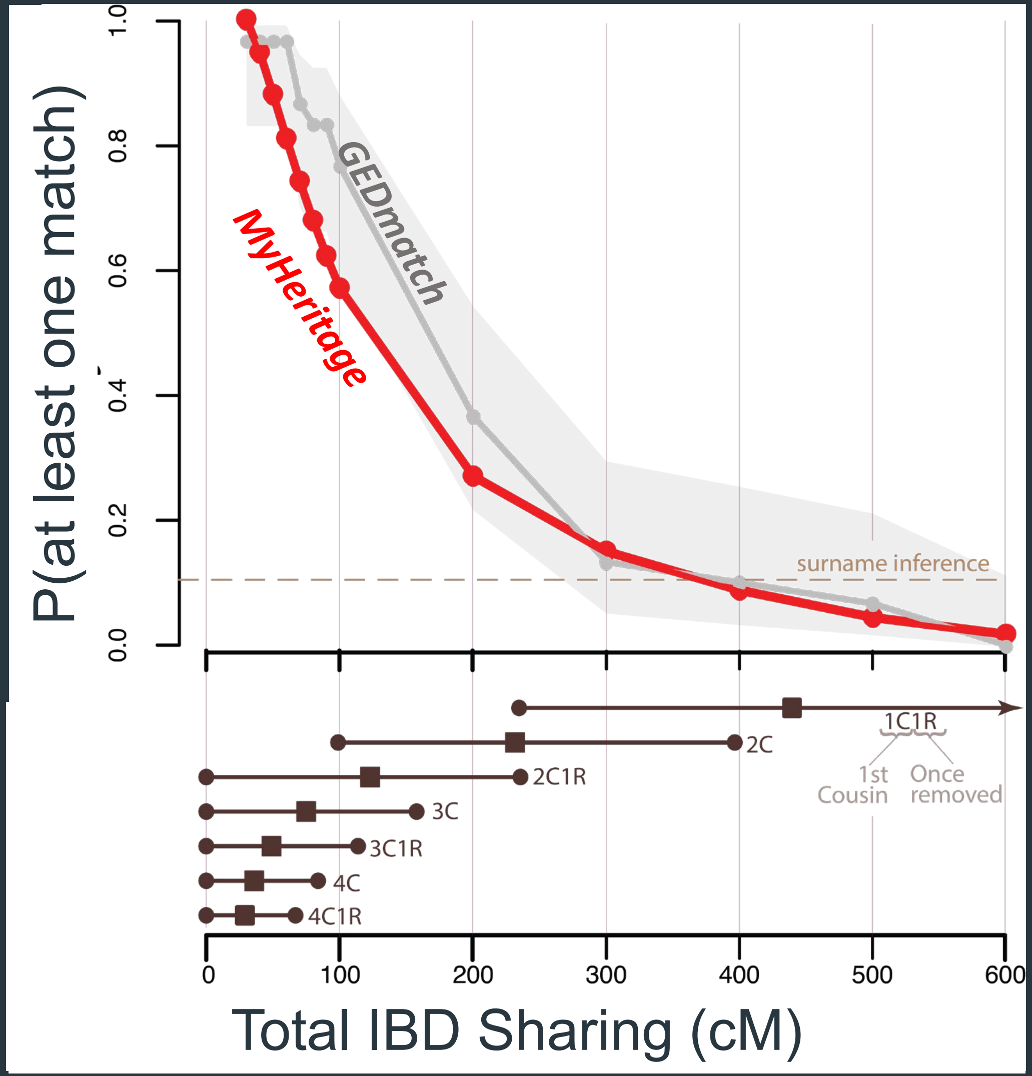


Repeat 30 times

Find genetic relatives for a GEDMatch participant

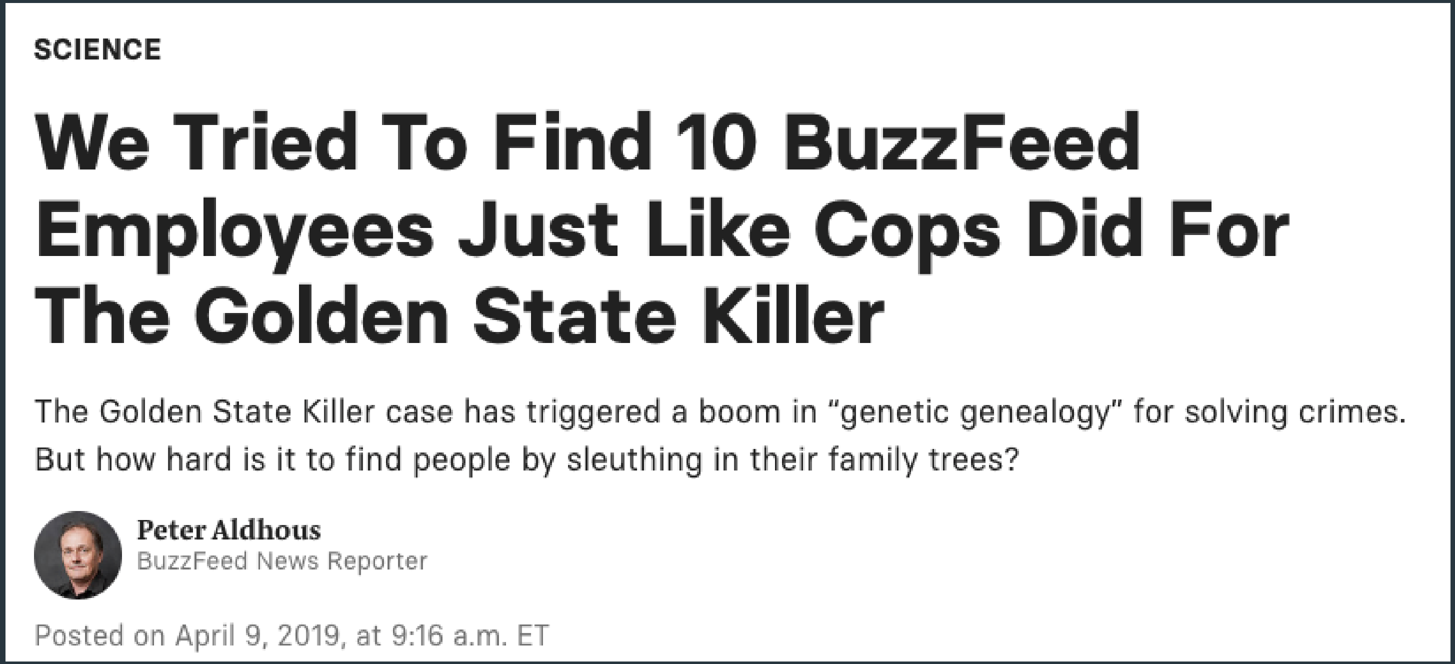
Exclude >700cM matches

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



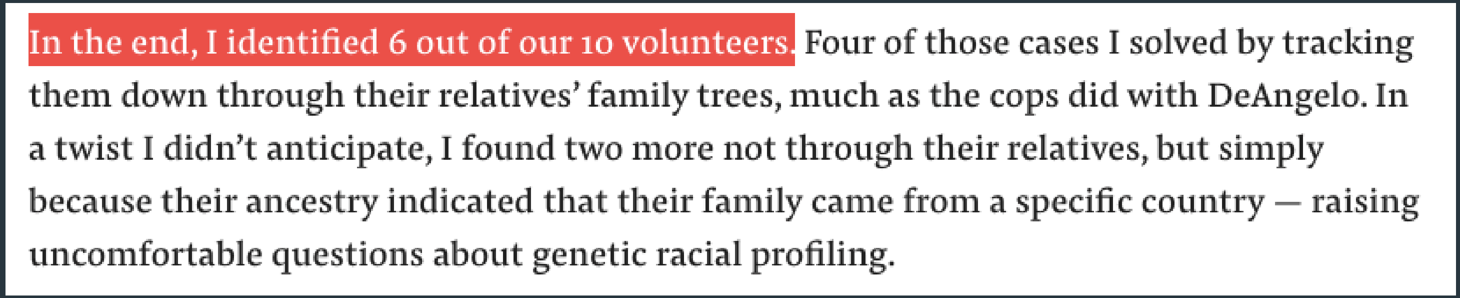
SCIENCE

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

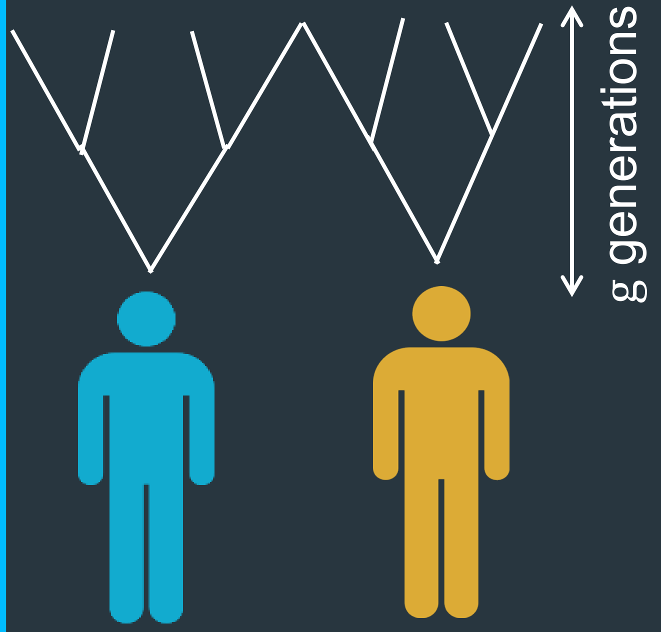
Posted on April 9, 2019, at 9:16 a.m. ET



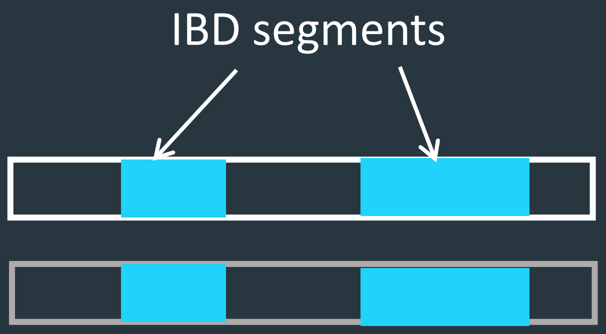
In the end, I identified 6 out of our 10 volunteers. Four of those cases I solved by tracking them down through their relatives' family trees, much as the cops did with DeAngelo. In a twist I didn't anticipate, I found two more not through their relatives, but simply because their ancestry indicated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

Modeling the probability of finding relatives

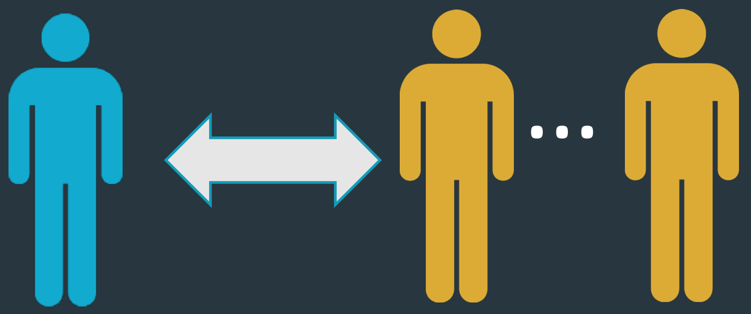
Are these individuals **genealogically** related?



Are these individuals also **genetically** related?

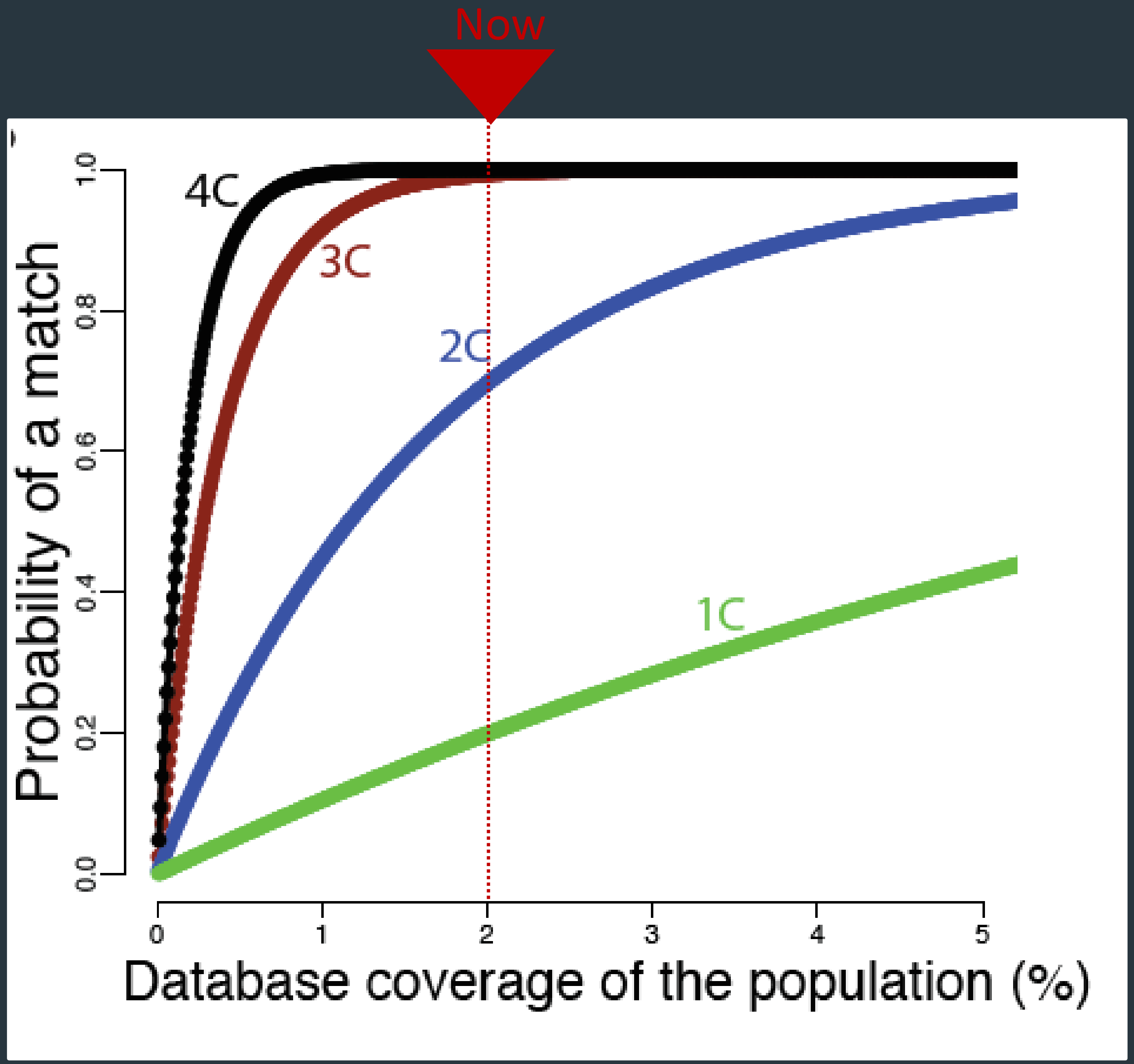


Is there at least one match in a database that covers X% of the population?



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

The probability of a match in the future

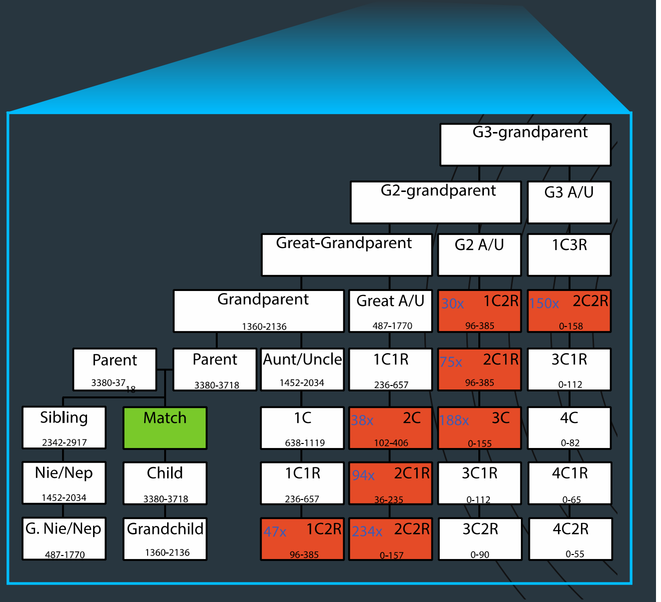


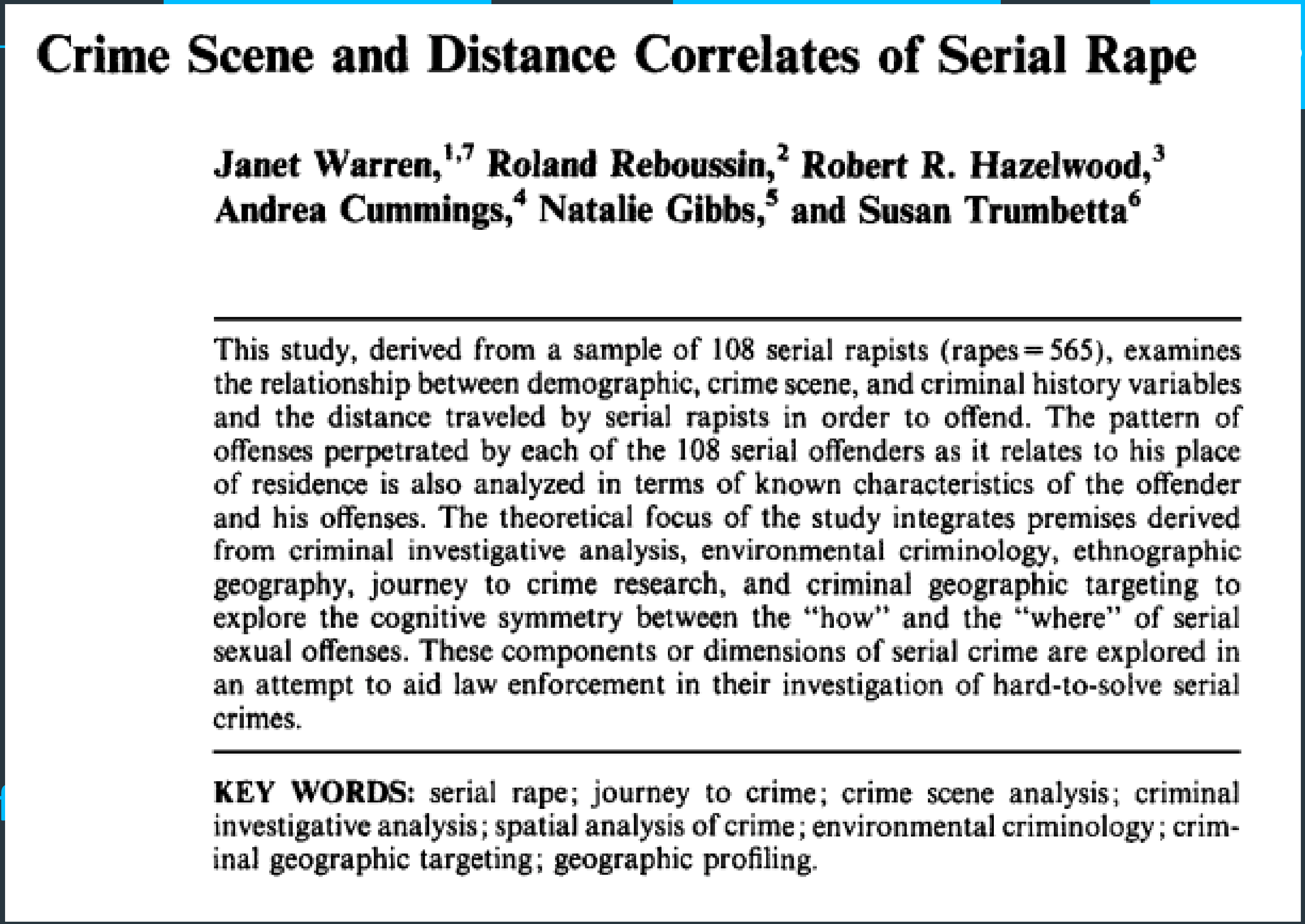
Virtually any US person of European heritage will have a 3rd cousin in these databases.

Can we get to a single person?

325M → 3rd cousin match

ex → 16.5





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.

Even three simple pieces of

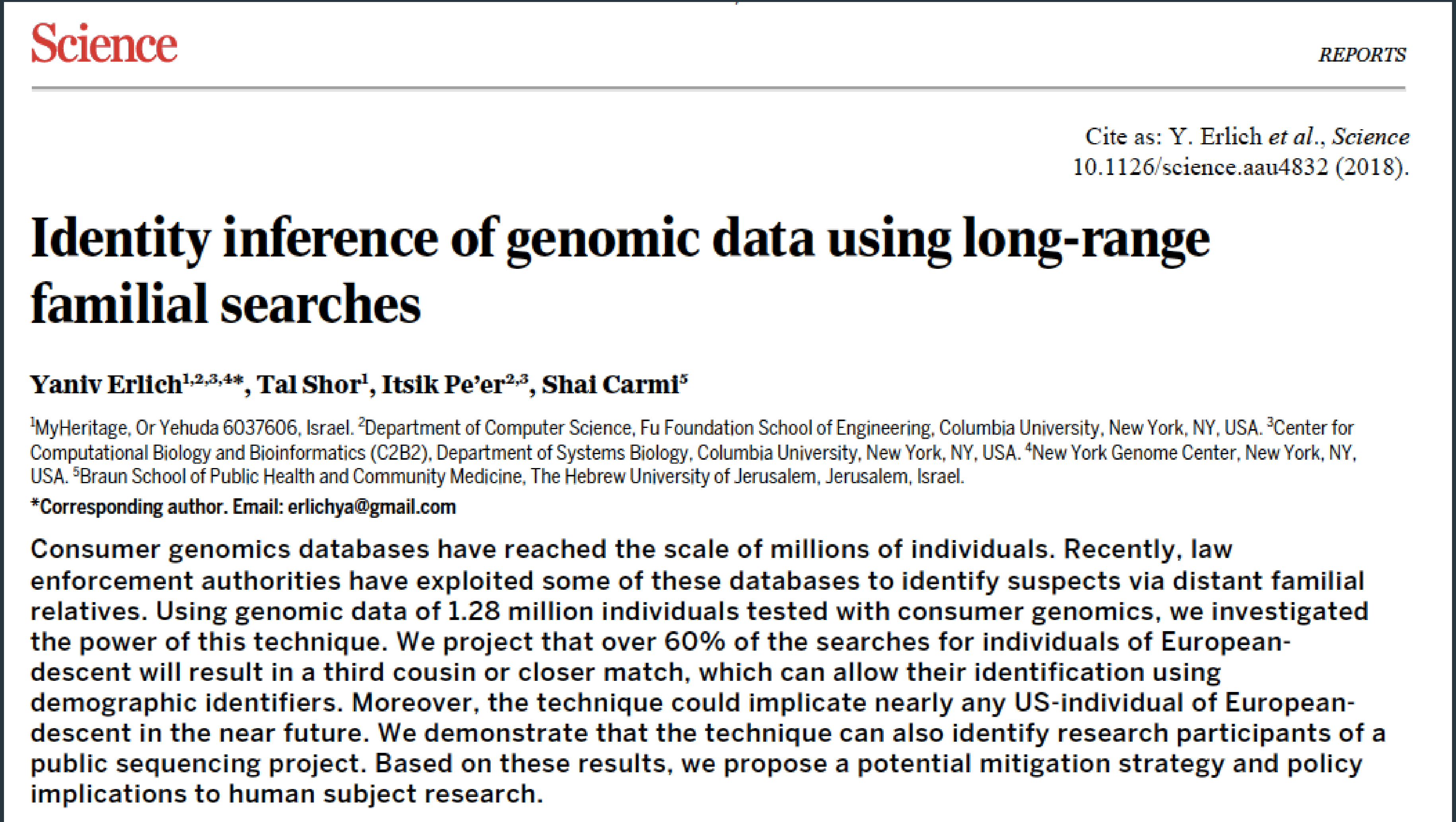
Summary so far

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

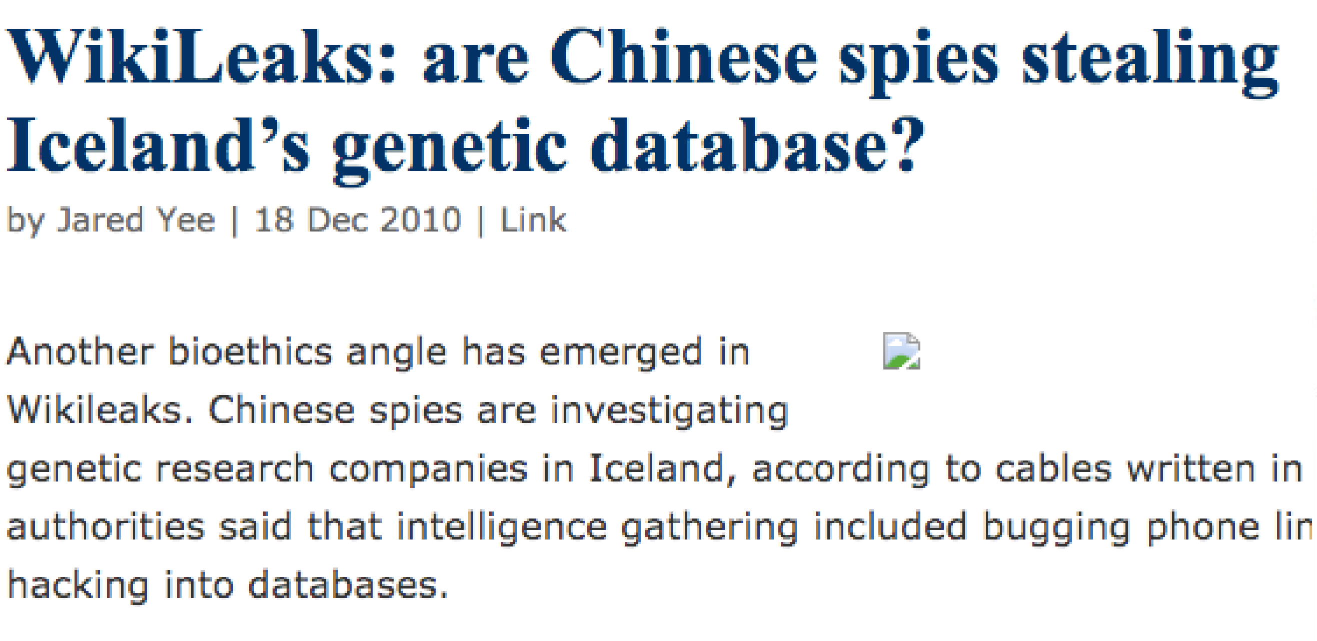
1. Everyone can uploads data to GEDmatch/FTDNA/etc…

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)

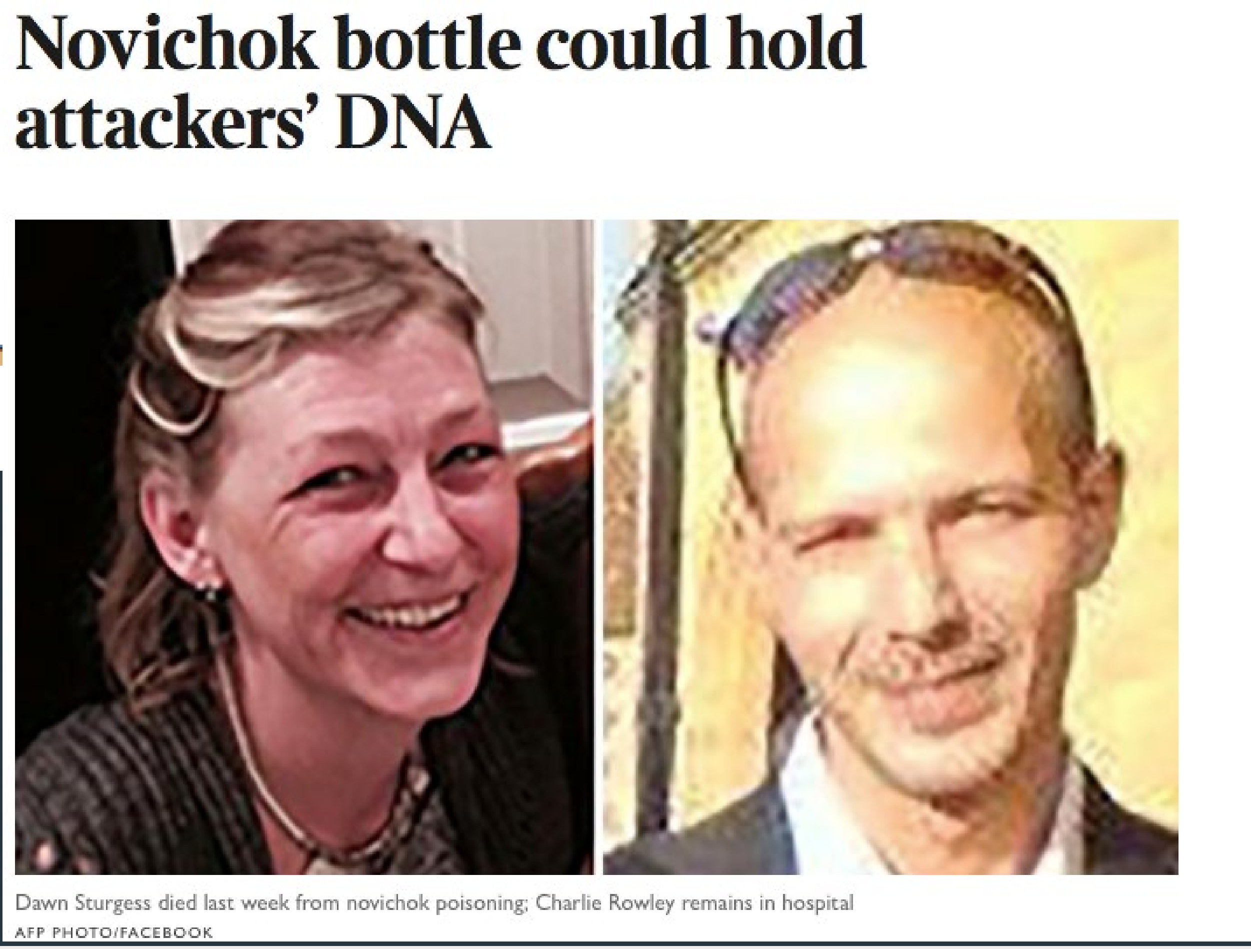
Intelligence services are interested in DNA



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

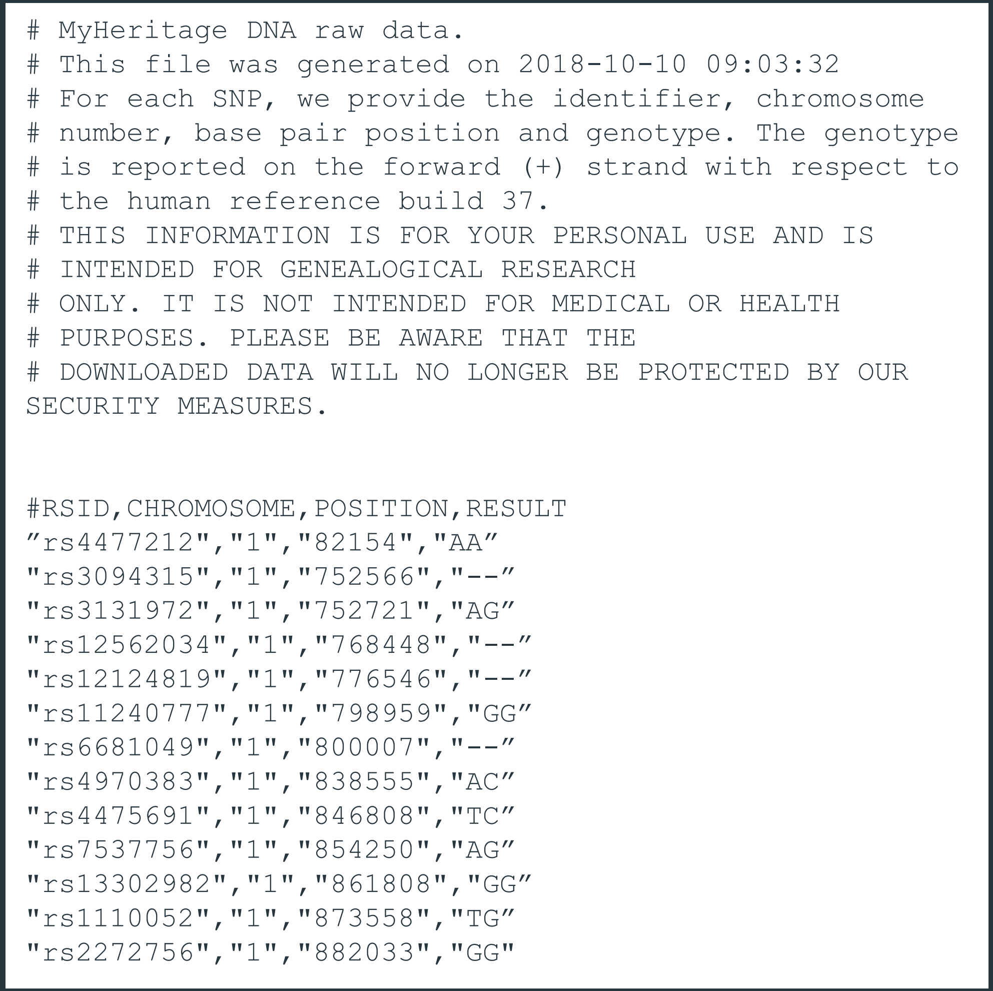
Differentiate good vs. bad actors

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?

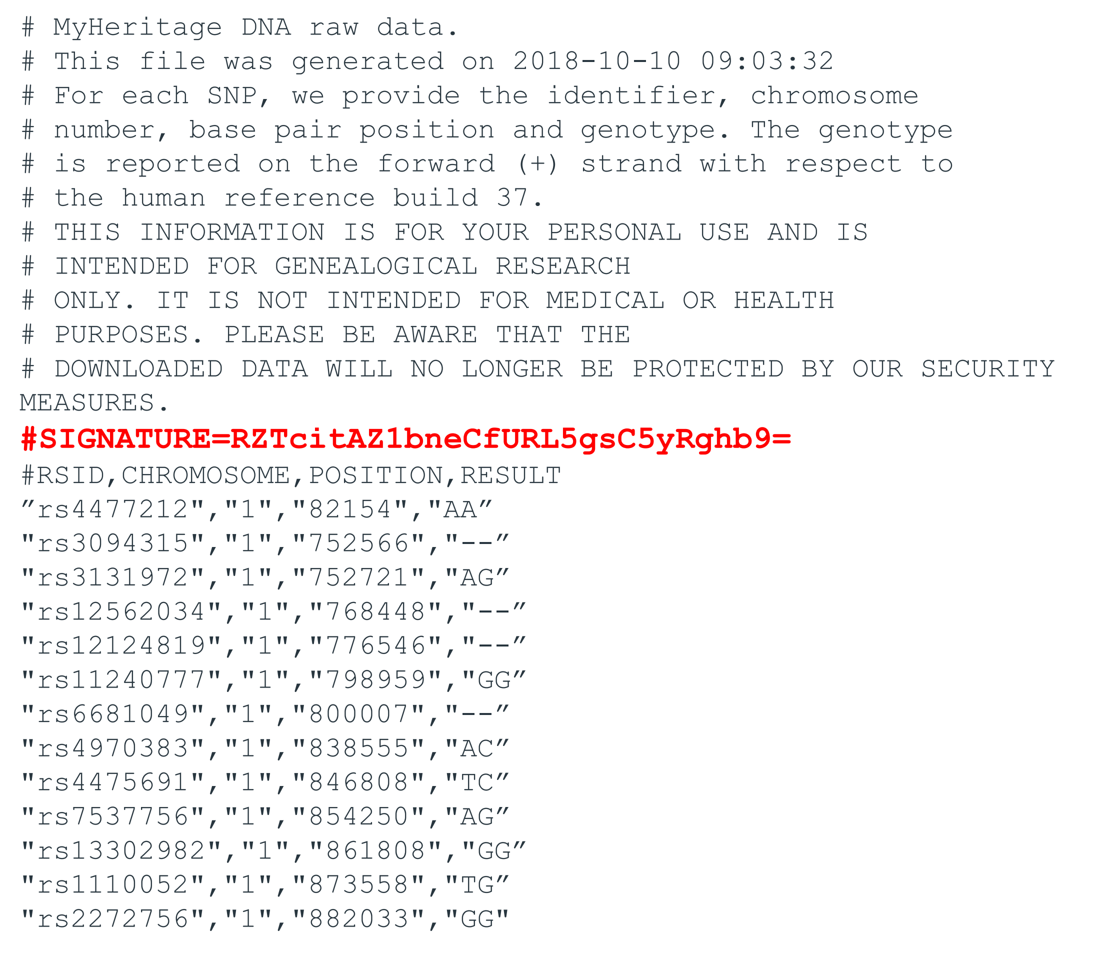
Before



```
# 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

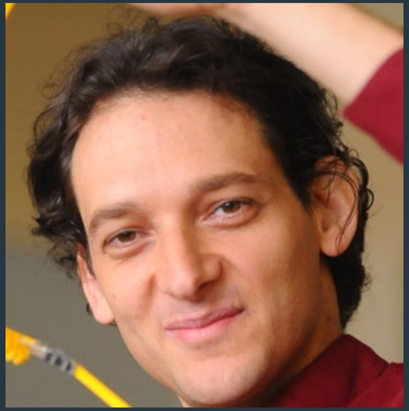


```
# MyHeritage DNA raw data.
# This file was generated on 2018-10-10 09:03:32
# For each SNP, we provide the identifier, chromosome
# number, base pair position and genotype. The genotype
# is reported on the forward (+) strand with respect to
# the human reference build 37.
# THIS INFORMATION IS FOR 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

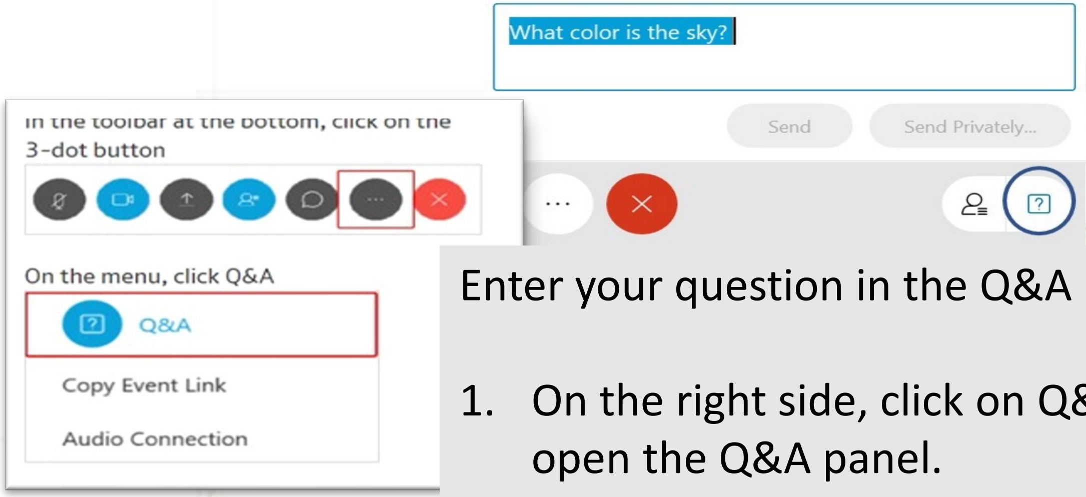


Itsik Pe’er
Columbia University

Shai Carmi
Hebrew University

Keynotes: The Protection and Enabling Perspectives

Moderated Questions and Answers



In the toolbar at the bottom, click on the 3-dot button

On the menu, click Q&A

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 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

Challenges from the Field: Research Perspective

Jean-Pierre Hubaux
(EPFL/ Global Alliance for Genomics and Health [GA4GH])

Protecting and Sharing Genomic Data: a Swiss/European Perspective

Prof. Jean-Pierre Hubaux, EPFL

Co-Founder of Tune Insight SA

Work done in close collaboration with Lausanne University Hospital (CHUV)

With gratitude to all the colleagues I have had the privilege to work with

About Switzerland

8.5 inhabitants

26 cantons (states), each with its own laws

Most of the health system managed by the cantons, not the federal government; the latter defines the overall policy and strategy

Data protection laws very similar to EU GDPR

Very strong political decentralization

One of highest GDP/capita in the world

Strong pharma: Roche, Novartis,…

5 university hospitals

2 federal institutes of technology: EPFL (Lausanne), ETH (Zurich)

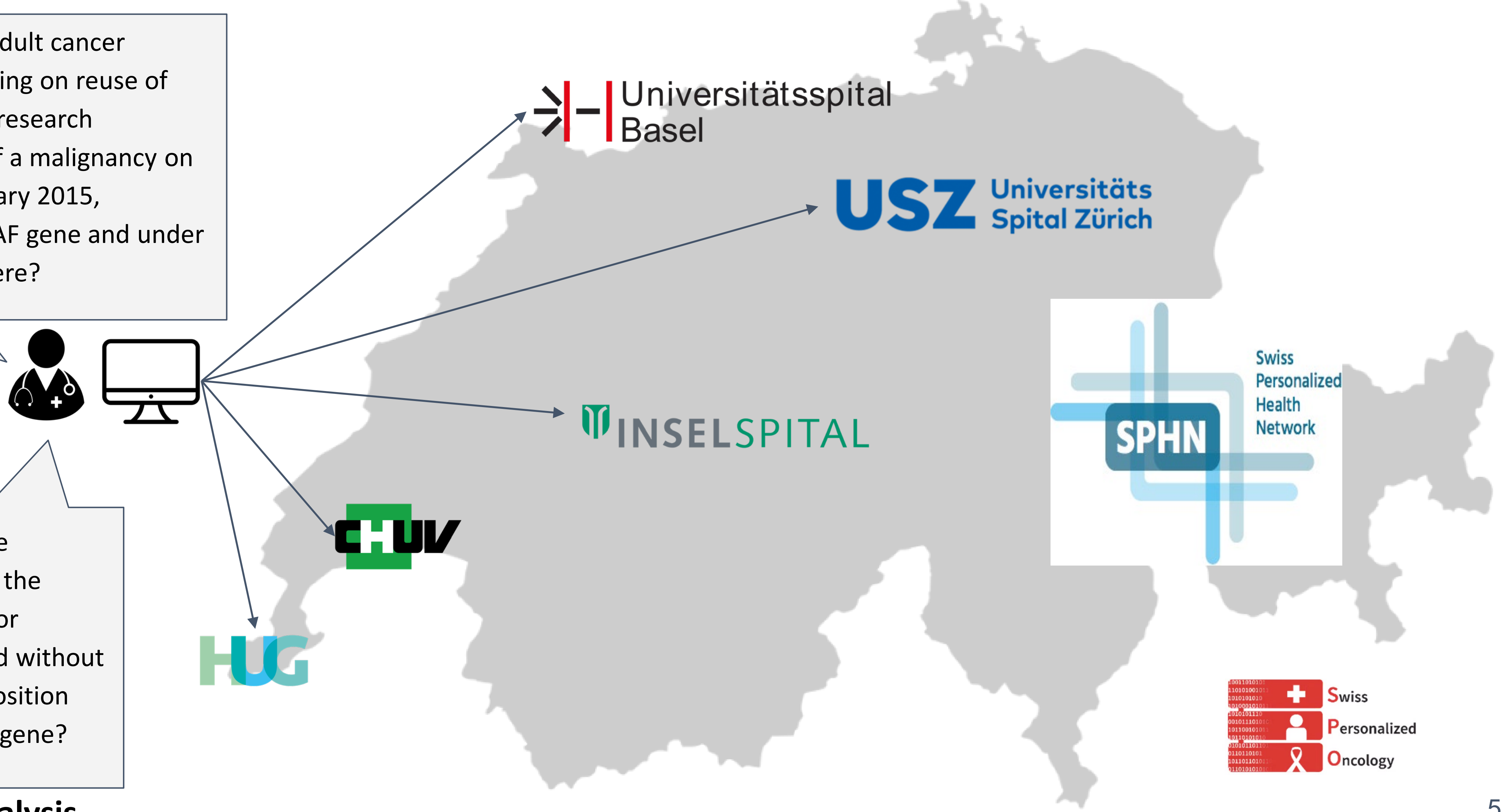
Use case for Swiss Personalized Oncology Project: federated analytics platform for research and molecular tumor board

Q1: How many adult cancer patients consenting on reuse of routine data for research with diagnosis of a malignancy on or after 1st January 2015, mutations in BRAF gene and under anti-PD-1 are there?

Explore

Q2: Among these patients, what is the overall survival for patients with and without a mutation on position 600 of the BRAF gene?

Analysis

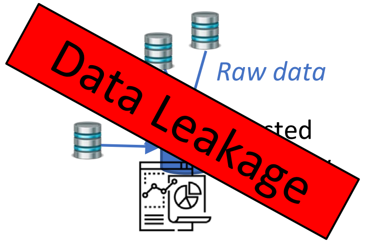

Universitätsspital Basel
USZ Universitäts Spital Zürich
INSELSPITAL
CHUV
HUG
SPHN Swiss Personalized Health Network
Swiss Personalized Oncology

The Main Challenges we Faced

- Multi-disciplinary nature of the problem: bio-informaticians, clinicians, geneticists, hospital IT specialists, hospital lawyers, data protection authorities, ethicists, computer scientists
- Mess of the health data
- Financial sustainability of the solution

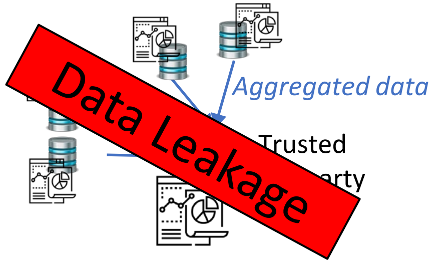
Distributed Learning - Current Approaches

a) Fully centralized



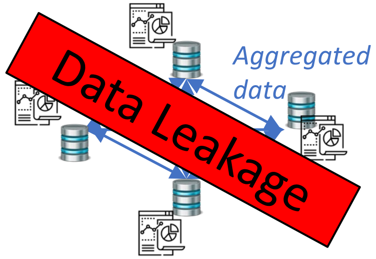
Examples:
- All of Us
- EGA
- Genomics England

Meta-analysis



https://covidclinical.net/

Decentralized

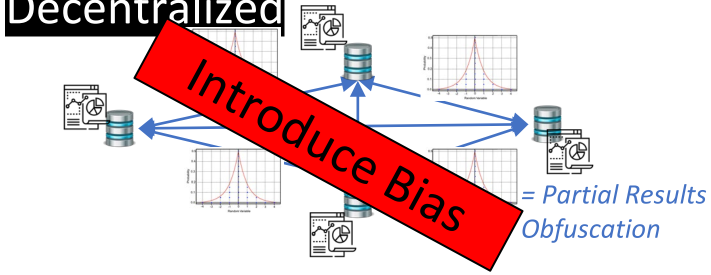


http://www.datashield.ac.uk

Personalized Health Train (PHT)

Differential Privacy

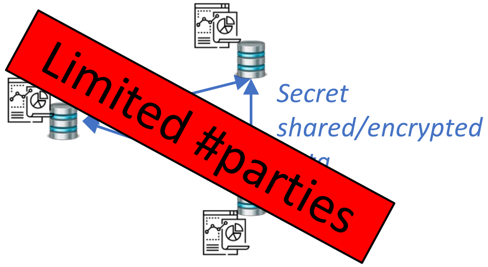
Decentralized



Examples:
- M. Kim et al. "Secure and Differentially Private Logistic Regression for Horizontally Distributed Data," TIFS 2019
- M. Abadi et al. Deep learning with differential privacy. In ACM CCS, 2016.
- Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In NIPS, 2009.

(e) Cryptographic (SMC, HE)

Decentralized



Examples:
- A. Gascón et al.. Privacy-preserving distributed linear regression on high-dimensional data. PETS, 2017.
- P. Mohassel and Y. Zhang. SecureML: A system for scalable privacy-preserving machine learning. In IEEE S&P, 2017.

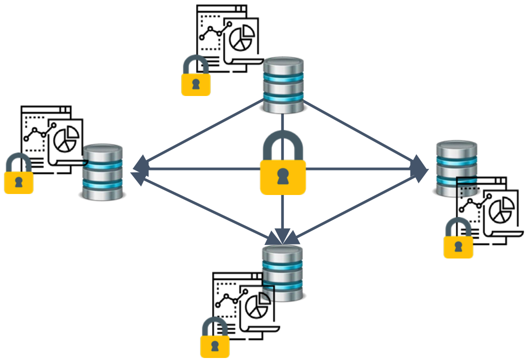
Our approach

→ **Data + Model Confidentiality as long as 1 entity is honest**

→ No data outsourcing

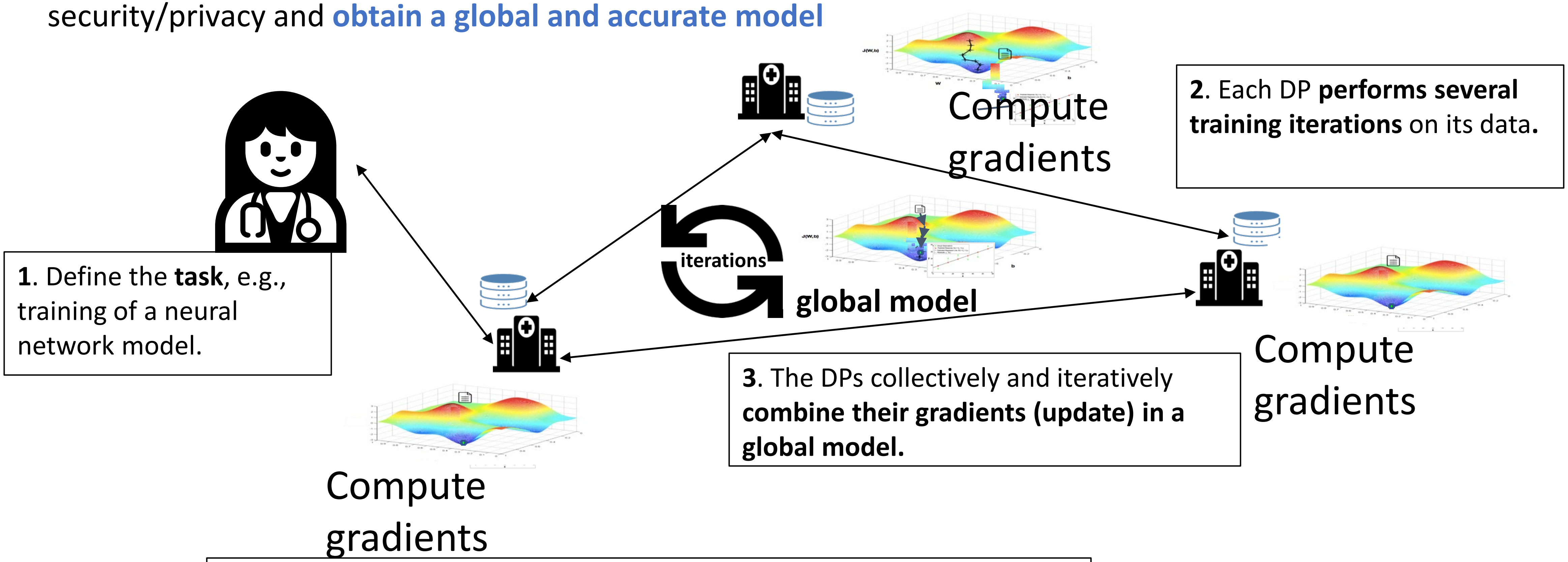
→ Scales with #parties

→ Exact results



Privacy-Preserving Federated Neural Network Learning

Solution: The data providers (DPs) collaborate to enable a joint gradient descent while protecting their security/privacy and **obtain a global and accurate model**



1. Define the **task**, e.g., training of a neural network model.

2. Each DP **performs several training iterations** on its data**.**

3. The DPs collectively and iteratively **combine their gradients (update) in a global model.**

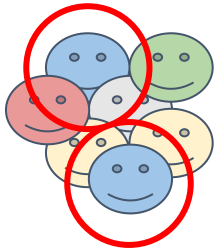
4. After the training, **the model** is **kept secret for oblivious predictions**

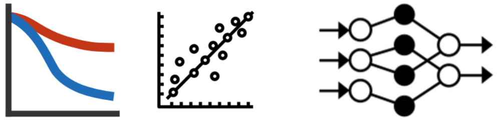
S.Sav, A. Pyrgelis, J.R. Troncoso-Pastoriza, D. Froelicher, J.P. Bossuat , J.S. Sousa and J.P. Hubaux,

POSEIDON: Privacy-Preserving Federated Neural Network Learning. NDSS, 2021

MedCo

- **Distributed software platform** for federated cohort exploration and analytics of clinical and genomic data
- Co-developed by EPFL and CHUV
- Built on top of the i2b2 cohort explorer (i2b2 is used by 250+ hospitals worldwide)
- Relies on **advanced cryptographic techniques**
  → Multi-party homomorphic encryption (MHE)
- Code-reviewed and pen-tested by third-party industrial companies, compliant with hospitals' information security policies
- Main functionalities
  - **MedCo-Explore: cohort exploration**
    - Obtaining cohort sizes for clinical research studies based on inclusion/exclusion criteria
  - **MedCo-Analysis: federated analytics**
    - Survival analysis
    - ML training and testing





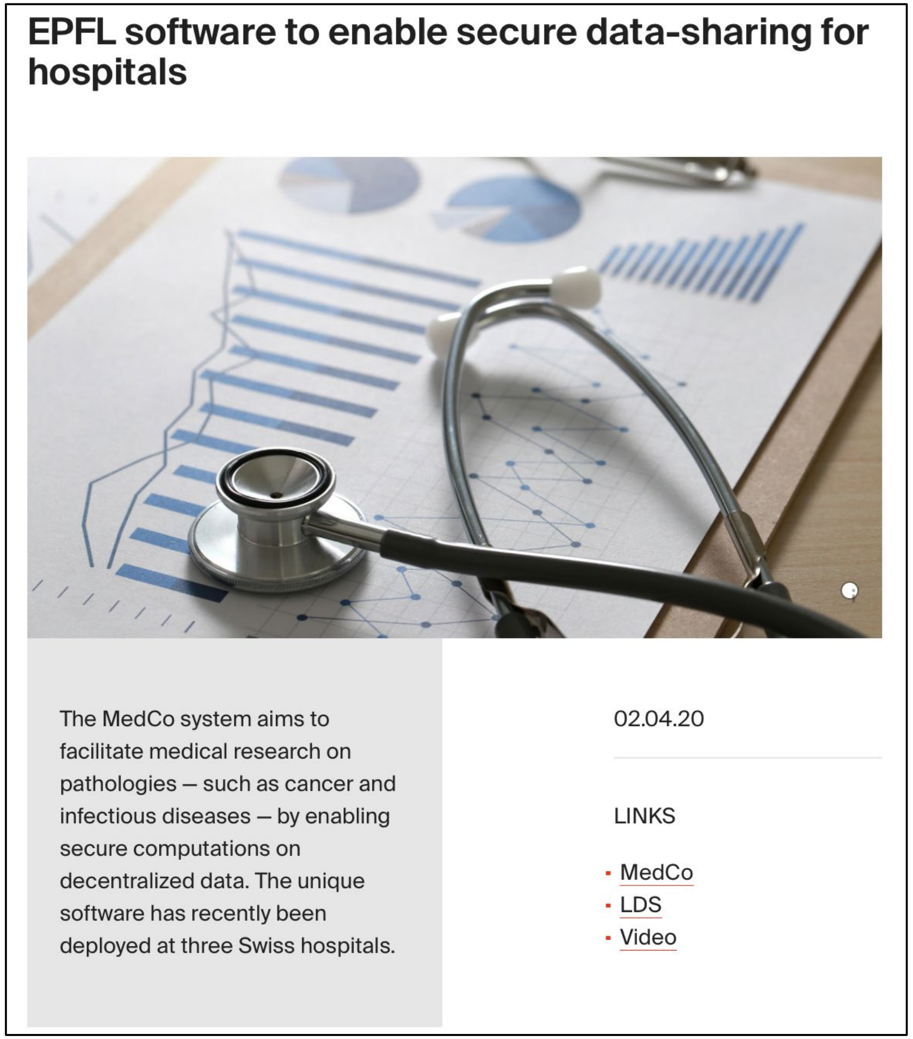
April 2020: MedCo deployed at 3 hospitals

MedCo
Collective protection of medical data

SPHN Swiss Personalized Health Network

Strategic Focus Area
Personalized Health and Related Technologies

HUG CHUV INSEL



EPFL software to enable secure data-sharing for hospitals

The MedCo system aims to facilitate medical research on pathologies – such as cancer and infectious diseases – by enabling secure computations on decentralized data. The unique software has recently been deployed at three Swiss hospitals.

02.04.20

LINKS

- MedCo
- LDS
- Video

- First application:
  Swiss Personalized Oncology project:
  → melanoma data and beyond
- Planned deployment at Zurich University Hospital
- Ongoing international deployments: USA, NL, Italy, France

Data Protection Impact Assessment (DPIA) for multisite medical data analysis (June 2021)

Centralized approach with standard pseudonymization

Threat	Threat likelihood	Threat impact	Risk	Risk level
Unlawful access to the system	Unlikely	High	Loss of data confidentiality	Moderate
Malicious use of the system	Possible	High	Loss of data confidentiality	High
Loss of data	Unlikely	Minor	Loss of data integrity, data unavailability	Minor
Data leak of host/cloud	Possible	High	Loss of data confidentiality	High
Collusion of host/cloud	Possible	High	Loss of data confidentiality	High
Corrupted or malicious host/cloud	Possible	High	Data unavailability, loss of data integrity, loss of data confidentiality, loss of data correctness	High
Unavailability of host/cloud	Possible	Minor	Data unavailability, loss of data correctness	Moderate
Re-identification/attribute inference	Possible	High	Loss of data confidentiality	High

Federated approach enhanced **with MedCo**

Threat	Measure introduced with MedCo	Threat likelihood	Threat Impact	Risk	Risk level
Unlawful access to the system	1	Unlikely	Minor	Loss of data confidentiality	Low
Malicious use of the system	1, 2, 4, 10	Possible	Minor	Loss of data confidentiality	Low
Loss of data	3, 5	Unlikely	Minor	Loss of data integrity, data unavailability	Low
Data leak	4, 5, 8, 9, 10	Unlikely	Minor	Loss of data confidentiality	Low
Collusion between nodes	4, 9	Unlikely	Moderate	Loss of data confidentiality	Moderate
Corrupted or malicious nodes	2, 5, 6, 7, 8, 9	Unlikely	Moderate	Data unavailability, loss of data integrity, loss of data confidentiality, loss of data correctness	Moderate
Unavailability of of nodes	6, 7	Possible	Minor	Data unavailability, loss of data correctness	Moderate
Re-identification or attribute inference	1, 2, 4, 9, 10	Unlikely	Minor	Loss of data confidentiality	Low

Feedback from Swiss authorities on MedCo DPIA

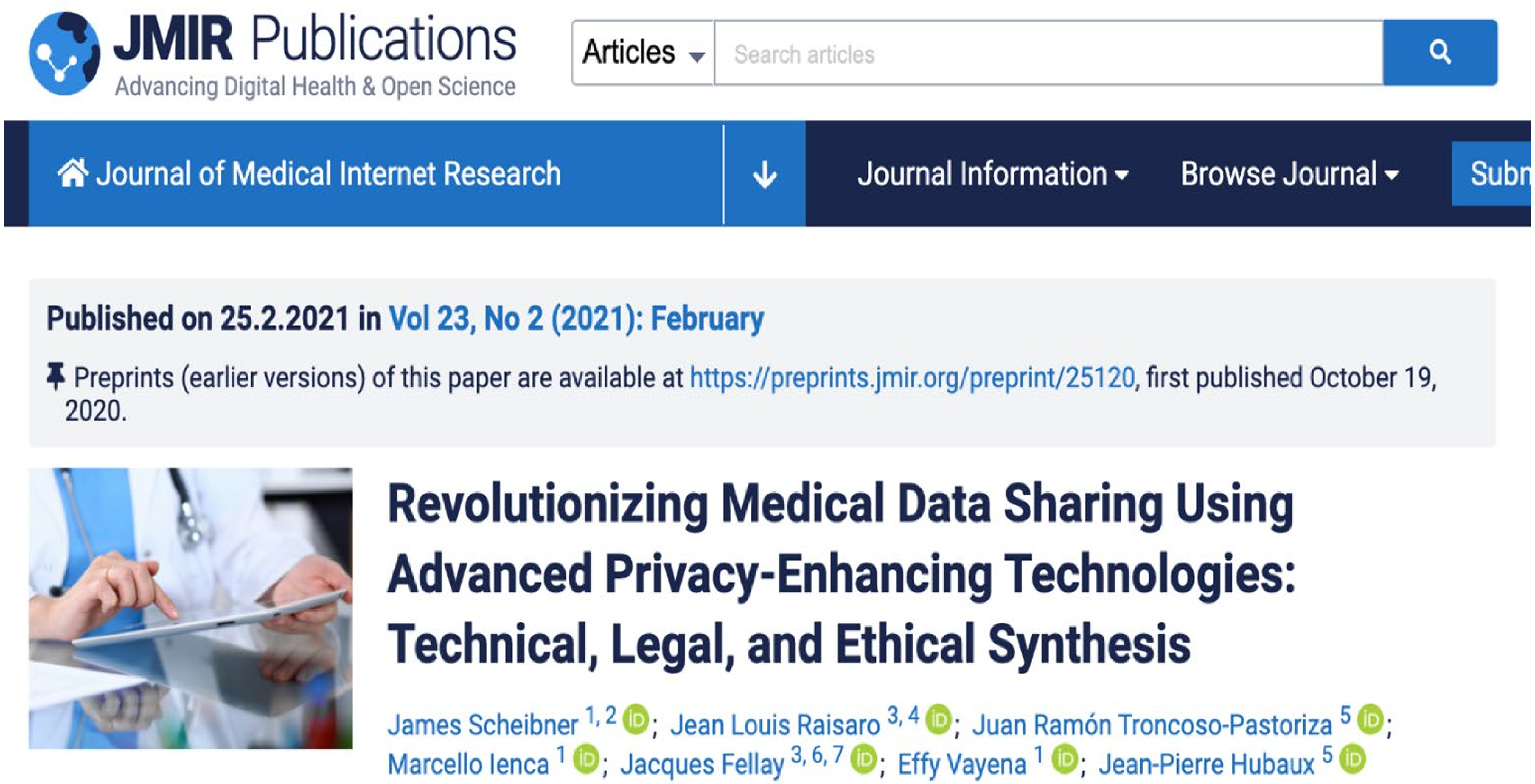
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Data Protection and Information Commissioner

"... the threat impact of most risks with the MedCo system shows to be clearly lower than with traditional systems. Since data processed within the Medco framework remain encrypted during computation, an attacker would cause little damage. **As no entity has the full decryption key, it seems indeed unlikely that he could decrypt and abuse the stolen data**. ..."

13 September 2021

GDPR legal compliance: partial aggregates are not personal data anymore, they are anonymous

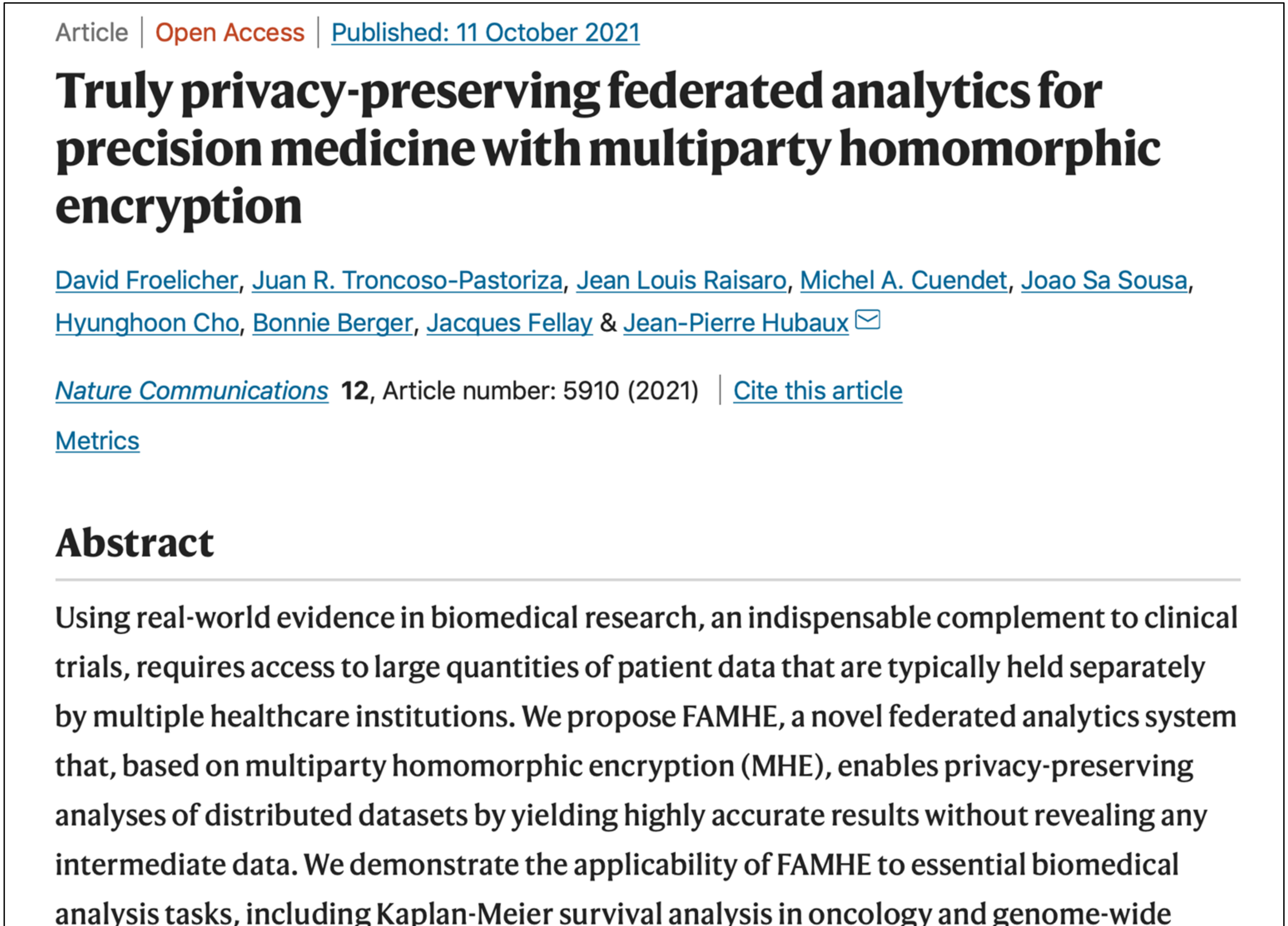


Published on 25.2.2021 in Vol 23, No 2 (2021): February

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/25120, first published October 19, 2020.

Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis

James Scheibner [1, 2]; Jean Louis Raisaro [3, 4]; Juan Ramón Troncoso-Pastoriza [5]; Marcello Ienca [1]; Jacques Fellay [3, 6, 7]; Effy Vayena [1]; Jean-Pierre Hubaux [5]



Article | Open Access | Published: 11 October 2021

Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption

David Froelicher, Juan R. Troncoso-Pastoriza, Jean Louis Raisaro, Michel A. Cuendet, Joao Sa Sousa, Hyunghoon Cho, Bonnie Berger, Jacques Fellay & Jean-Pierre Hubaux ✉

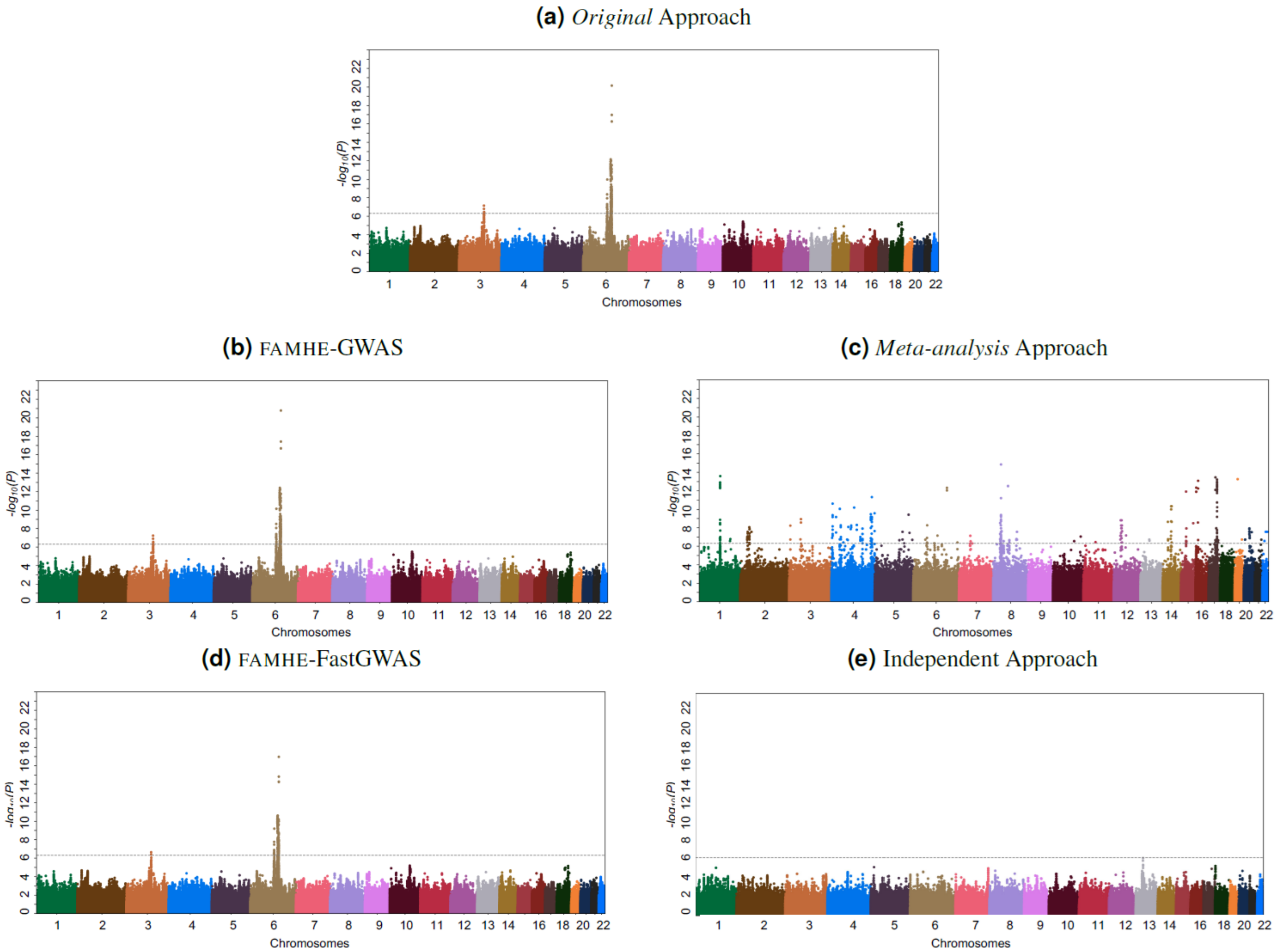
Nature Communications **12**, Article number: 5910 (2021) | Cite this article

Metrics

Abstract

Using real-world evidence in biomedical research, an indispensable complement to clinical trials, requires access to large quantities of patient data that are typically held separately by multiple healthcare institutions. We propose FAMHE, a novel federated analytics system that, based on multiparty homomorphic encryption (MHE), enables privacy-preserving analyses of distributed datasets by yielding highly accurate results without revealing any intermediate data. We demonstrate the applicability of FAMHE to essential biomedical analysis tasks, including Kaplan-Meier survival analysis in oncology and genome-wide

FAMHE: Privacy-Preserving Federated Analytics for Precision Medicine with MHE - GWAS



(a) *Original* Approach

(b) FAMHE-GWAS

(c) *Meta-analysis* Approach

(d) FAMHE-FastGWAS

(e) Independent Approach

[Original approach] McLaren, P. J. et al. Polymorphisms of Large Effect Explain the Majority of the Host Genetic Contribution to Variation of HIV-1 Virus Load. Proc. Natl. Acad. Sci. 112, 14658–14663 (2015).

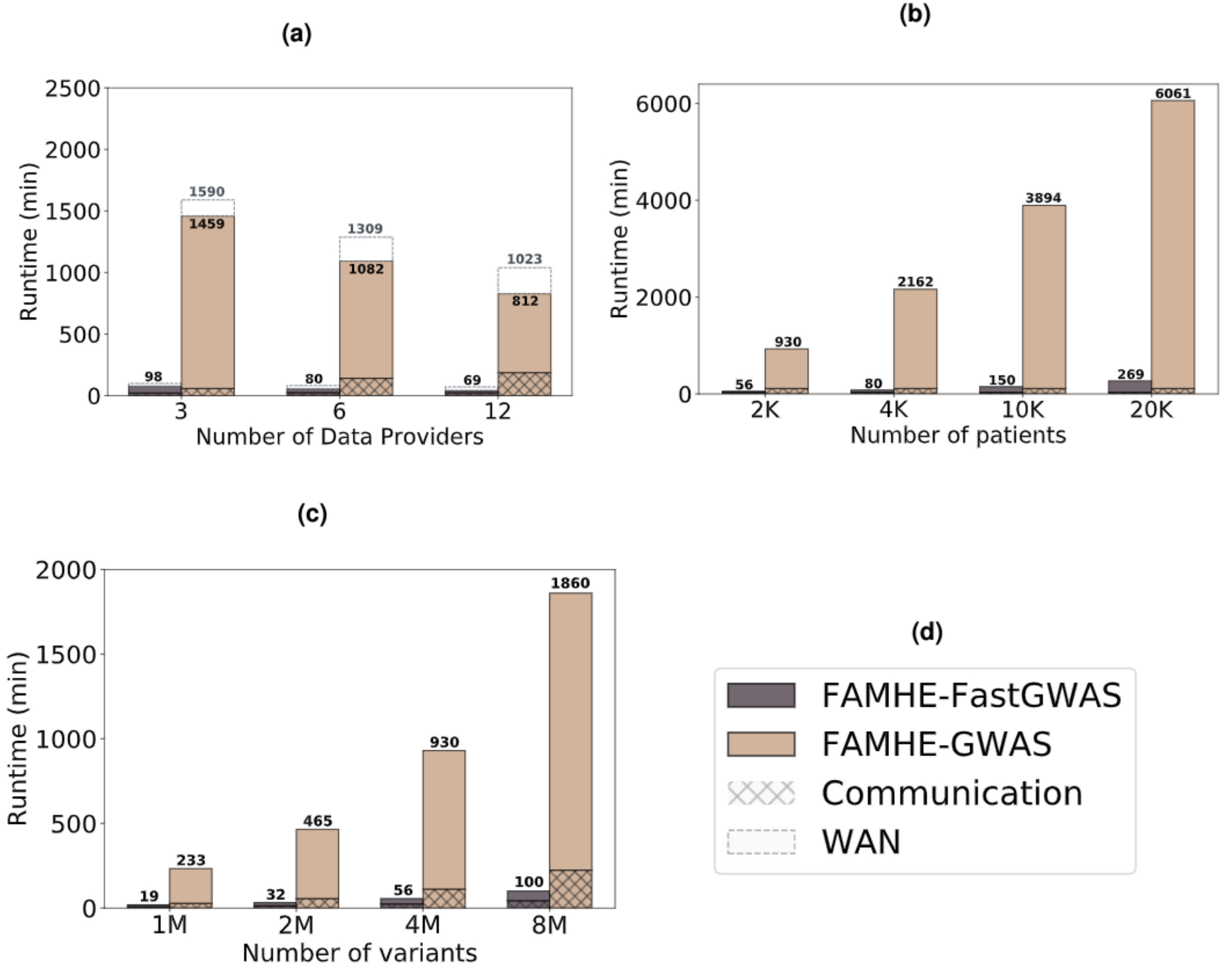
[FAMHE] Froelicher et al. Truly Privacy-Preserving Federated Analytics for Precision Medicine with Multiparty Homomorphic Encryption.

FAMHE: Genome-wide association study

Default: 1857 patients spread among 12 data providers.

→ scale in all dimensions

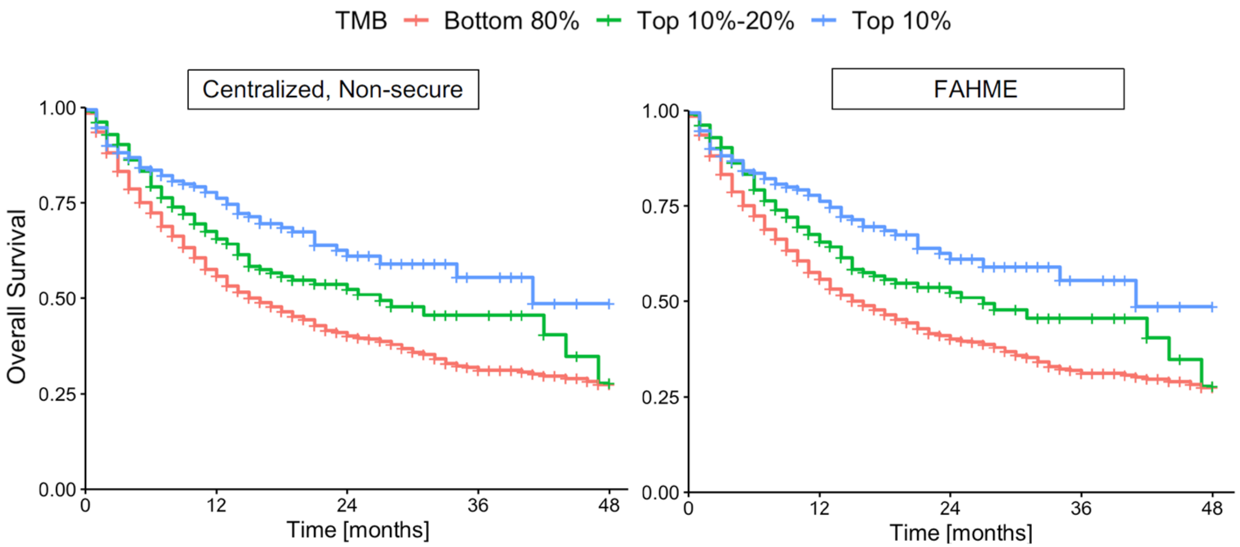
a. *With the number of data providers*
b. *With the number of patients*
c. *With the number of variants*



[Centralized] McLaren, P. J. et al. Polymorphisms of Large Effect Explain the Majority of the Host Genetic Contribution to Variation of HIV-1 Virus Load. Proc. Natl. Acad. Sci. 112, 14658–14663 (2015).
[FAMHE] Froelicher et al. Truly Privacy-Preserving Federated Analytics for Precision Medicine with Multiparty Homomorphic Encryption. Submitted to Nat. comms. https://doi.org/10.1101/2021.02.24.432489

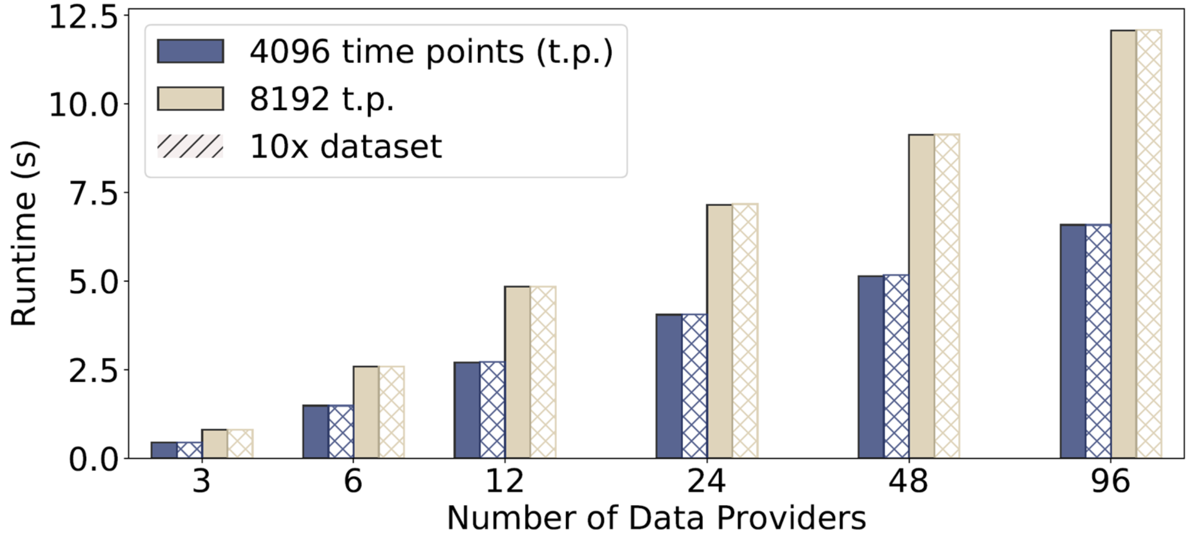
FAMHE: Privacy-Preserving Federated Analytics for Precision Medicine with MHE - Survival curves (Kaplan-Meier)

Data split among 3
data providers:



Number at risk

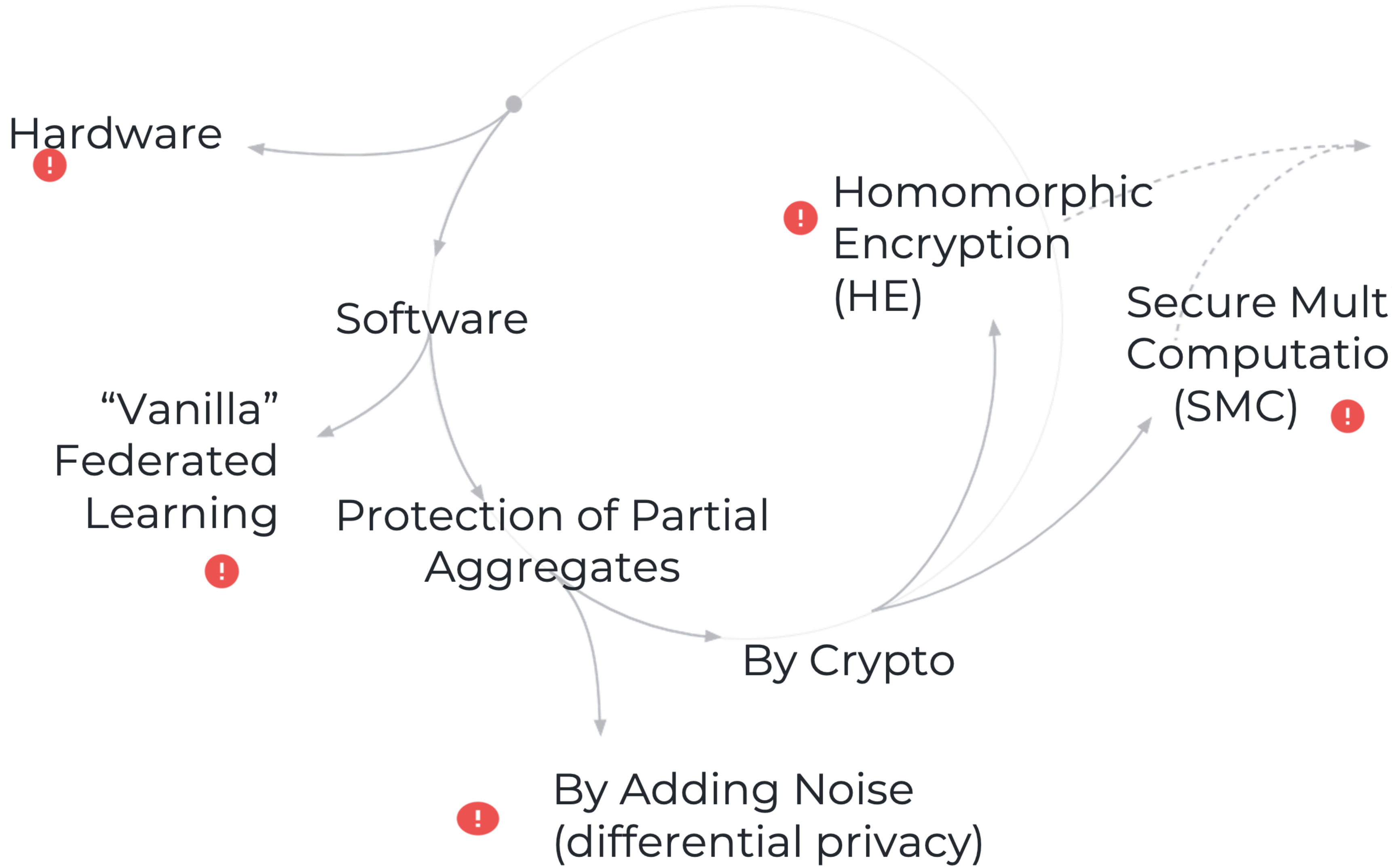
Time [months]	0	12	24	36	48
Bottom 80%	1302	584	230	85	33
Top 10%-20%	186	103	41	17	4
Top 10%	172	99	41	13	4



[Centralized] Samstein, R. M. et al. Tumor Mutational Load Predicts Survival after Immunotherapy across Multiple Cancer Types. Nat. genetics 51, 202–206 (2019).

[FAMHE] Froelicher et al. Truly Privacy-Preserving Federated Analytics for Precision Medicine with Multiparty Homomorphic Encryption.

Share without Sharing: Available Options



Hardware !

Software

“Vanilla” Federated Learning !

Protection of Partial Aggregates

By Adding Noise (differential privacy) !

By Crypto

! Homomorphic Encryption (HE)

Secure Multiparty Computation (SMC) !

Multiparty Homomorphic Encryption

HE + SMC ✓

TUNE INSIGHT

Enterprise Data & Analytics

$100B Market

(GlobalData, Enterprise Data and Analytics Market, 2020))

56%
say own data is not enough, expanding ability to source external data

(Forrester, The Insights Professional's Guide To External Data Sourcing, 2019)

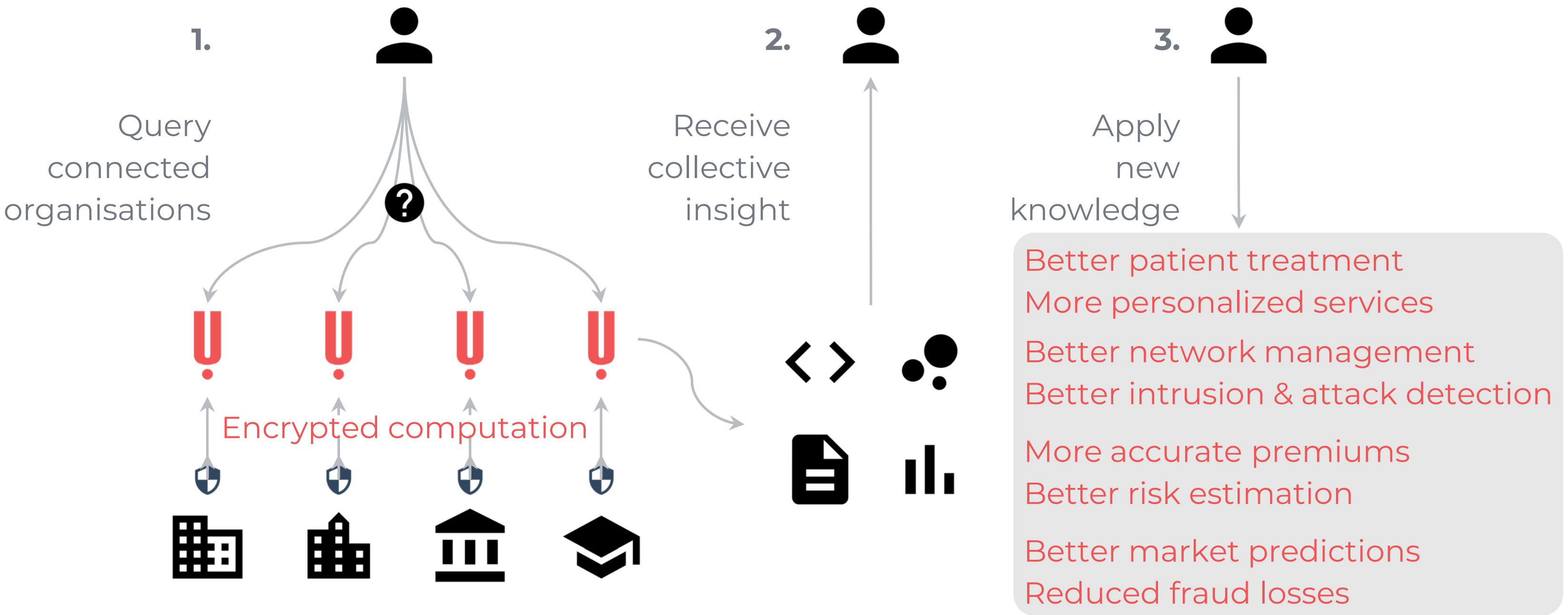
However, organizations are **prevented** to enter valuable data collaborations due to fear of **data leaks** and **data protection regulations**

TUNE INSIGHT

juan@tuneinsight.com

Cross-vertical enterprise SaaS enabling organizations to make better decisions, together, by orchestrating secure collaborations around their sensitive data.

- CHF400k in customer-paid projects including with Swiss Re, Armasuisse
- Pilot deployed at Swiss hospitals
- CHF100k EPFL Innogrant
- State-of-the-art post-quantum encryption technology
- Raised pre-seed with Wingman Ventures



Better patient treatment
More personalized services

Better network management
Better intrusion & attack detection

More accurate premiums
Better risk estimation

Better market predictions
Reduced fraud losses

Access to insights
Personalization

Immediacy
Scalability

Compliance
Control

MHE: mathematical proofs instead of vendor lock-in and side-channel attacks

	Software-based solutions (MHE)	Hardware-based solutions (e.g., Intel SGX)
System and trust model	**Decentralized** (federated computing, edge computing) or **centralized** (outsourced) systems	**Only centralized** systems (data has to be transferred to the TEE)
Assumptions	Protection against passive adversaries with quantum computing power: **processing infrastructure (including side-channels) and other data providers**	Protection against passive adversaries (other tenants); **limited protection against the processing infrastructure**; protection against side-channels is implementation-dependent
Implementation cost	**Tailored solution**; application-specific design; composition of cryptographic building blocks; limited range of efficient functionalities	**Available SDKs**; relatively easy conversion to secure enclave; general-purpose solutions; limited libraries and memory inside the enclave
Performance and overhead	**Less than 10x** overhead when full packing capacity is utilized (federated training of GLMs and NNs). Up to 4-5 orders of magnitude overhead for non-optimized or non-packed solutions	**Negligible** overhead for **regular instructions**; **4x overhead for memory** copy operations; **35x overhead for syscalls** to/from enclave
Response to newly discovered vulnerabilities	**Software patch** with protocol update; usually, no re-encryption of the data is needed	**Firmware** patch with variable **performance impact** (1x to 20x slow-down); **architecture change and hardware replacement**; **enclave code update** (update signatures, keys, and require new attestation)

GLM	: Generalized Linear Model	SDK	: Software Development Kit
MHE	: Multi-party homomorphic encryption	SGX	: Software Guard eXtensions
NN	: Neural Network	TEE	: Trusted Execution Environment

International collaborations

- Prof. Xiaoqian Jiang, UT Health
- GA4GH Data Security Work Stream
- MedCo now part of the i2b2 official community projects
- Prof. Shawn Murphy, HMS, and the ACT Network
- Broad Inst. + MIT
- Cancer Institute of the Netherlands
- …

Events devoted to the topic

- GenoPri.org: International workshop on genome privacy and security
    - Yearly workshop, typically co-located with GA4GH main annual event
- iDash - http://www.humangenomeprivacy.org/2021
    - Annual event with technical challenges on genome data protection and sharing

Conclusion

- We have solved the problem of GDPR-compliant federated learning for medical data, including genomic data
- Solution: Multi-party homomorphic encryption (MHE)
  - Perform computations without “seeing” the data
  - Rely on decentralized trust and mathematical proofs
  - No need to transfer the data
- Scalability with the number of data providers and the size of the datasets
- Green light from the Swiss federal data protection authority
- Support and development of new features: provided by Tune Insight

Contact me at jean-pierre.hubaux@epfl.ch

More information at https://medco.epfl.ch

Challenges from the Field: Individual's Perspective

John Verdi (Future of Privacy Forum)

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Privacy Risks and Challenges from the Perspective of Individual Rights

John Verdi, Senior Vice President of Policy at the Future of Privacy Forum

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Privacy Risks and Challenges from the Perspective of Individual Rights

FPF Work:

- In July 2018, the Future of Privacy Forum released Privacy Best Practices for Consumer Genetic Testing Services
- FPF developed the Best Practices following consultation with technical experts, regulators, leading consumer genetic and personal genomic testing companies, and civil society
- On January 1, 2022, California's Genetic Information Privacy Act (GIPA) became effective, codifying many of FPF's best practices

Privacy Risks and Challenges from the Perspective of Individual Rights

FPF Requirements:

- Express consent for collection, use, and retention of genetic data;
- Separate express consent for transfer of to third parties and for incompatible uses;
- Informed consent for research;
- Educational resources about the basics, risks, benefits, and limitations of genetic and personal genomic testing;
- Access, correction, and deletion rights;
- Valid legal process for disclosure to the government and transparency reports;
- Ban on sharing genetic data with third parties (such as employers, insurance companies, educational institutions, and government agencies) without consent or as required by law;
- Restrictions on marketing based on genetic data; and
- Strong data security protections and privacy by design

Privacy Risks and Challenges from the Perspective of Individual Rights

Privacy Risks and Challenges of genomic data:

- Unique, immutable biometric
- Potentially reveals information about identity
- Potentially reveals information about heritage
- Potentially reveals information about health
- Potentially reveals information about relatives’ identities, heritage, and health
- Difficult or impossible to de-identify without undermining utility

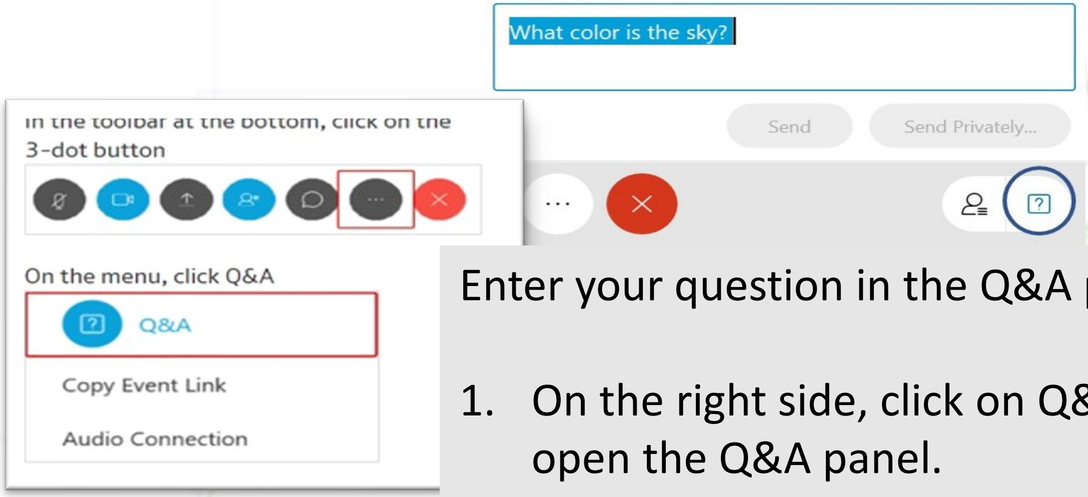
Privacy Risks and Challenges from the Perspective of Individual Rights

Privacy Risks and Challenges of genomic data:

- False identifications in criminal matters (evidence mishandling)
- Unexpected family connections and non-connections
- Dept. of Defense warning re: health tests and readiness reporting
- False identifications in criminal matters (remote relatives)
- Data breaches – e.g. 2020 GED Match law enforcement breach
- Re-identification attacks, e.g. cross-referencing clinical, research, and publicly available data sets

Challenges from the Field: Research and Individual’s Perspectives

Moderated Questions and Answers



in the toolbar at the bottom, click on the 3-dot button

On the menu, click Q&A

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 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

Session 1: Cybersecurity Challenges Affecting Genomic Sequencing

Charles Fracchia (BioBright)
Phillip Whitlow (HudsonAlpha)

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

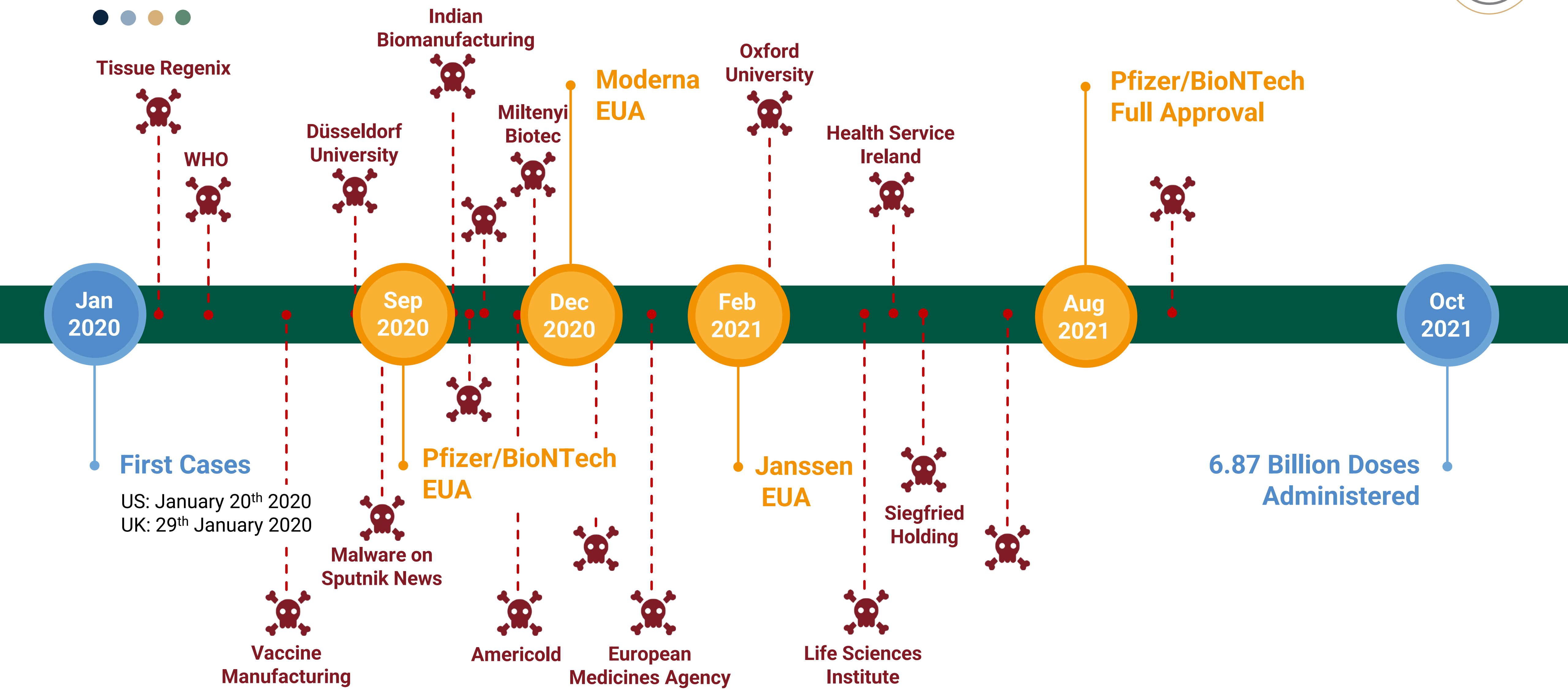
Digital Biosecurity in a Modern Context

Charles Fracchia <charlesfracchia@gmail.com>

PGP: 4A06 3D3A B157 C3DE C31C 91B0 6E76 3F6A DA35 06C4

CEO BioBright / VP Data Dotmatics

PUBLIC DOMAIN ATTACKS ON BIOINFRASTRUCTURE



Jan 2020

First Cases

US: January 20th 2020
UK: 29th January 2020

Tissue Regenix

WHO

Vaccine Manufacturing

Düsseldorf University

Sep 2020

Malware on Sputnik News

Pfizer/BioNTech EUA

Indian Biomanufacturing

Miltenyi Biotec

Americold

Dec 2020

Moderna EUA

European Medicines Agency

Feb 2021

Oxford University

Janssen EUA

Life Sciences Institute

Health Service Ireland

Siegfried Holding

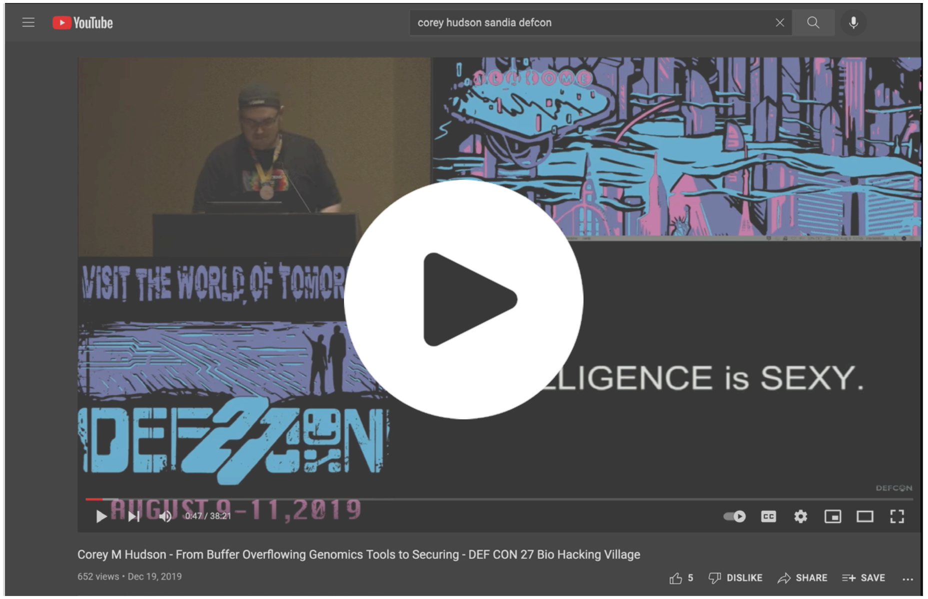
Aug 2021

Pfizer/BioNTech Full Approval

Oct 2021

6.87 Billion Doses Administered

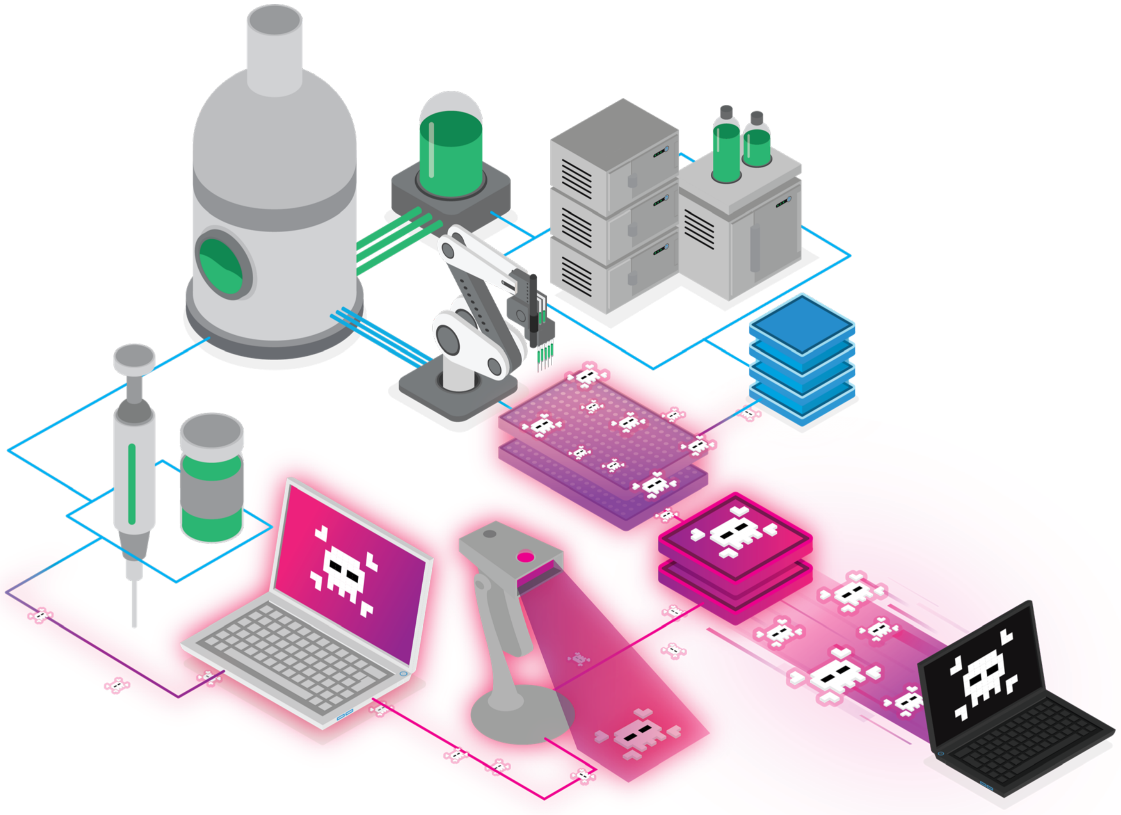
WE CANNOT FIX THE ISSUE WITHOUT TACKLING DEVICE INSECURITY



Corey M Hudson - From Buffer Overflowing Genomics Tools to Securing - DEF CON 27 Bio Hacking Village

652 views • Dec 19, 2019

https://www.youtube.com/watch?v=7du1TltZOJg



Devices are embedded at every step of the biological process.

TARDIGRADE APT ON THE BIOECONOMY

bio-isac

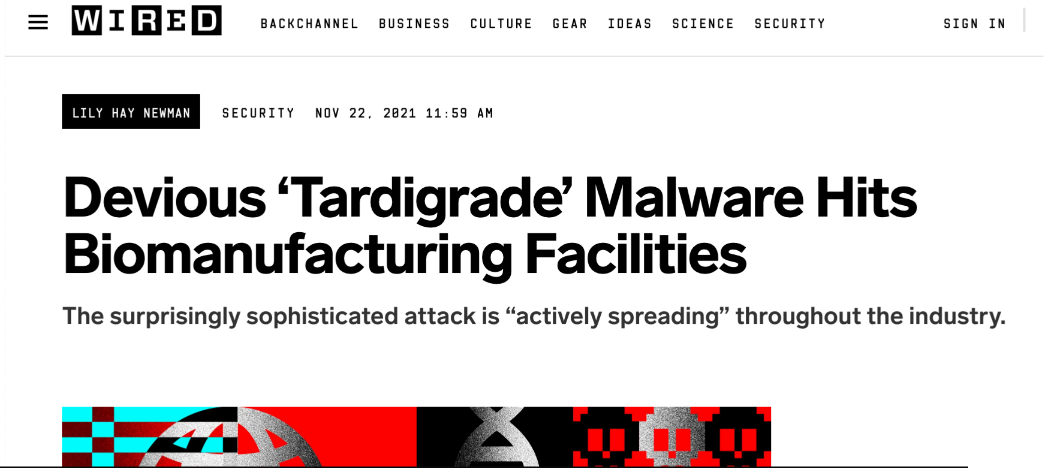
Creating Collaborative Threat Intelligence for the Bioeconomy

Tardigrade:
APT Attack on the Bioeconomy
(Bulz.253748 Variant Overview: intserrs644.dll)

Callie Churchwell – Senior Digital Biosecurity Analyst
Charles Fracchia – VP of Data & CEO BioBright
Ed Chung, MD – Digital Biosecurity Lead & CMO BioBright

Contact: tips@isac.bio
PGP: EB2AA1AA4AD94A0BBE07 E8A9625BBFEE2E95C7FC

contributed by: BioBright a dotmatics company



WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN

LILY HAY NEWMAN SECURITY NOV 22, 2021 11:59 AM

Devious ‘Tardigrade’ Malware Hits Biomanufacturing Facilities

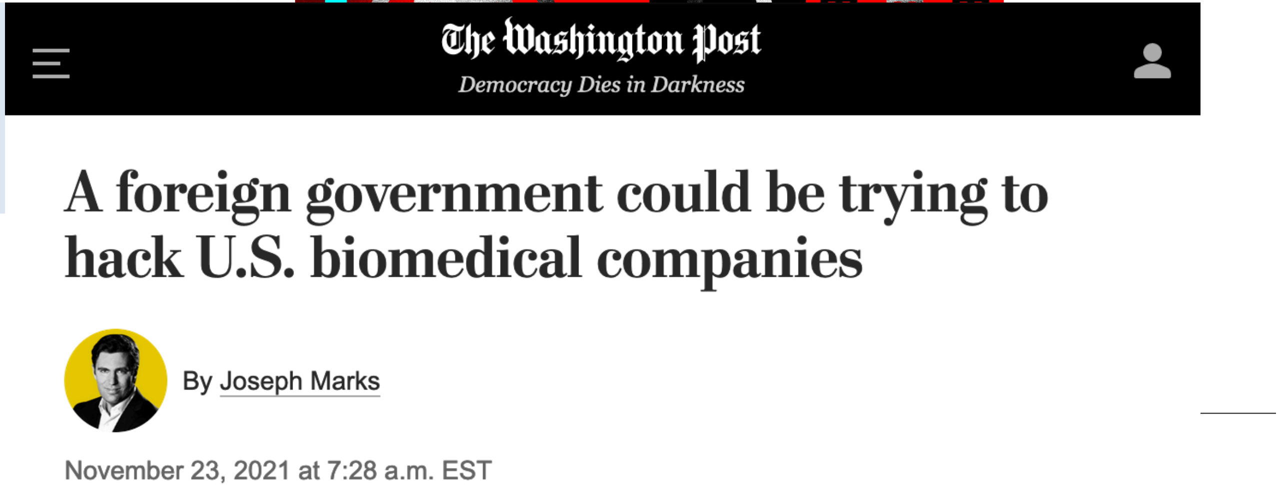
The surprisingly sophisticated attack is “actively spreading” throughout the industry.

LEADERSHIP FOR IT SECURITY & PRIVACY ACROSS HHS
HHS CYBERSECURITY PROGRAM
OFFICE OF INFORMATION SECURITY

HC3: Alert

November 23, 2021 TLP: White Report: 202111231300

BIO-ISAC Tardigrade Amplify Alert



The Washington Post
Democracy Dies in Darkness

A foreign government could be trying to hack U.S. biomedical companies

By Joseph Marks

November 23, 2021 at 7:28 a.m. EST

THE HILL

Biomanufacturing companies getting hit by hackers potentially linked to Russia

BY MAGGIE MILLER - 11/22/21 01:14 PM EST 37 COMMENTS

BIO-ISAC:

BIOECONOMY-FOCUSED INFORMATION SHARING AND ANALYSIS CENTER

bio-isac

Creating Collaborative Threat Intelligence for the Bioeconomy

- Provides members tailored and actionable threat intelligence information
- Leverages a coordinated disclosure process to simplify information sharing from small, medium and large enterprises of digital biosecurity issues
- Establishes and shares best practices and standards to improve digital biosecurity in the bioeconomy, including biomanufacturing
- Educates members and partners in digital biosecurity by creating and teaching content directly
- Promotes the creation of a skilled workforce for digital biosecurity with industry and government partners
- Interacts with lawmakers and policy stakeholders to further the development of a resilient infrastructure for the bioeconomy
- Acts as a convening place for trusted international partners to collaborate on digital biosecurity and biological supply chain security issues

https://isac.bio

Genomic Sequencing Instrument Security

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

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

Phillip Whitlow
Security Architect
HudsonAlpha Institute for Biotechnology
pwhitlow@hudsonalpha.org

Who is HudsonAlpha?

HudsonAlpha Institute for Biotechnology is a nonprofit institute founded in 2008 specializing in genetics and genomics research and biotech education.

Tens of thousands of genomes (human and non-human) sequenced per year on campus

Sequencing use cases include:

- Genetic testing
- Clinical genome sequencing
- Genomic screening programs
- Original plant genome sequencing

We also host more than 50 associate companies on our campus - all of them involved in bioscience and many performing genomic sequencing in their own labs

HUDSONALPHA
INSTITUTE FOR BIOTECHNOLOGY

Cybersecurity Challenges

Campus and Entrepreneurial Mission

- Associate companies provide their own sequencers
- Combines all the challenges of IoT and BYOD

Cybersecurity Challenges

Sequencers are essentially IoT devices

- Internet connection required
- Dedicated PC
- Unknown software
- Software firewalls

Cybersecurity Challenges

Availability vs. Security

- Security takes a back seat to availability and accuracy

- Software updates can be problematic

Software update available

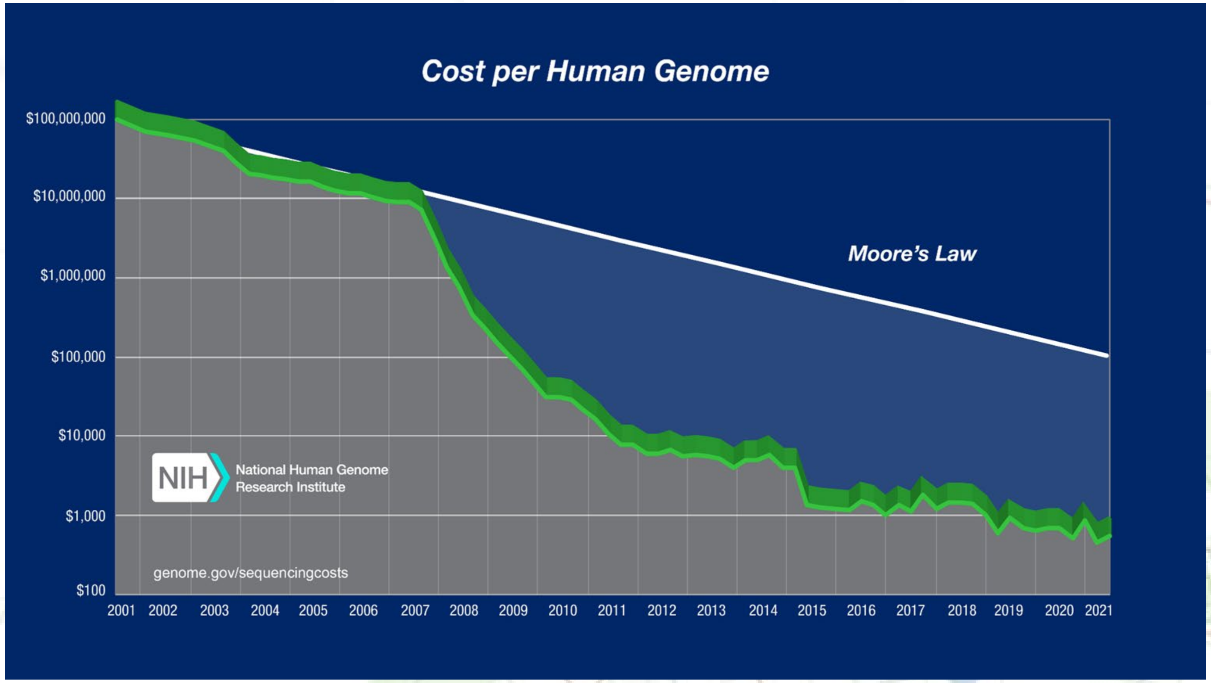
Cybersecurity Challenges

Lack of Security Standards

- No guidelines or standards (e.g., STIG)

Anticipated proliferation

- Decreasing cost and more widespread use will lead to more attacks



Session 2: Cybersecurity Challenges for Genomic Software

E. Loren Buhle (DNAnexus)

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

National Cybersecurity Center of Excellence

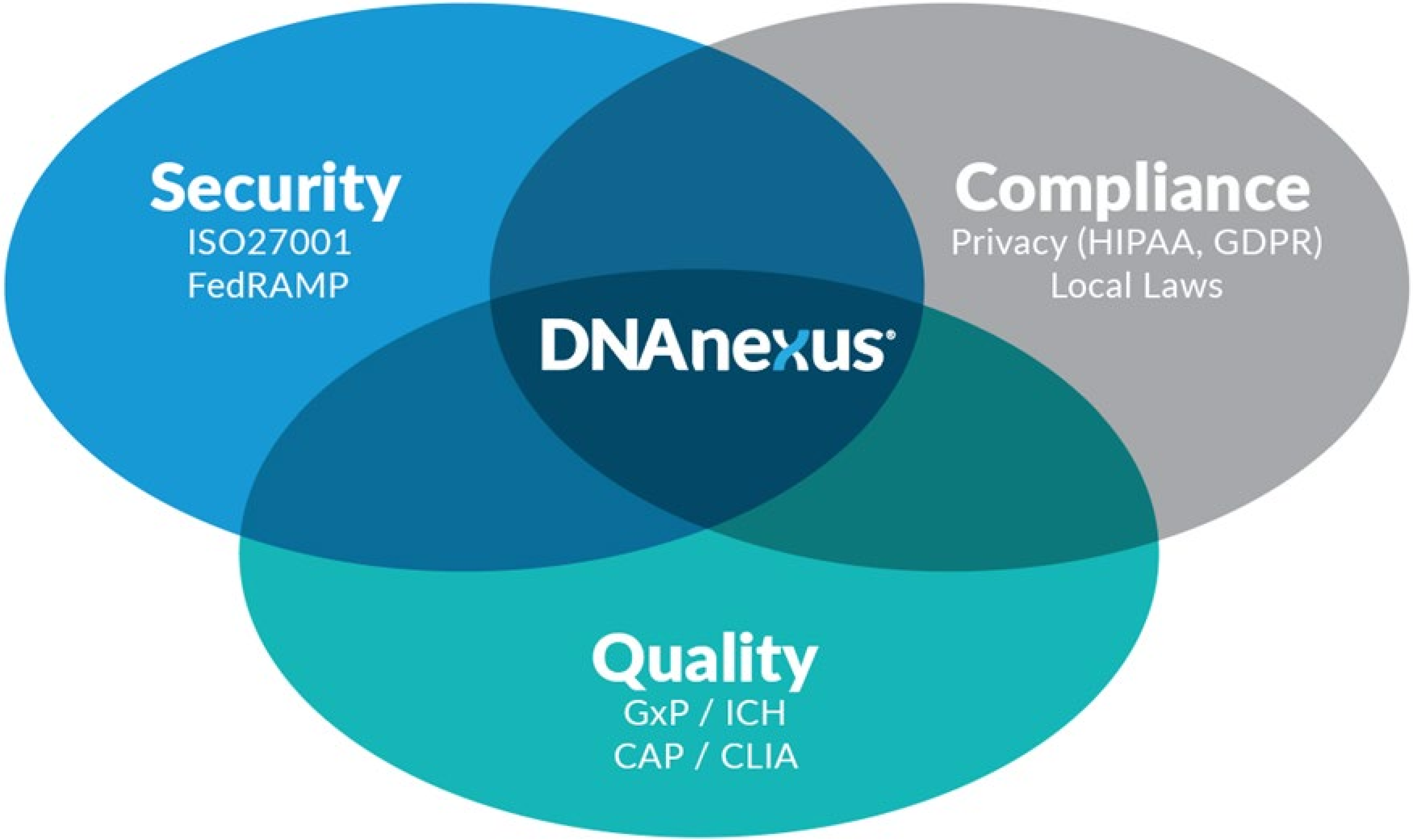
Session 2: Cybersecurity Challenges for Genomic Software

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

E. Loren Buhle

VP of Security, Quality and Compliance at DNAnexus

CONSIDER A HOLISTIC APPROACH



Security
ISO27001
FedRAMP

Compliance
Privacy (HIPAA, GDPR)
Local Laws

Quality
GxP / ICH
CAP / CLIA

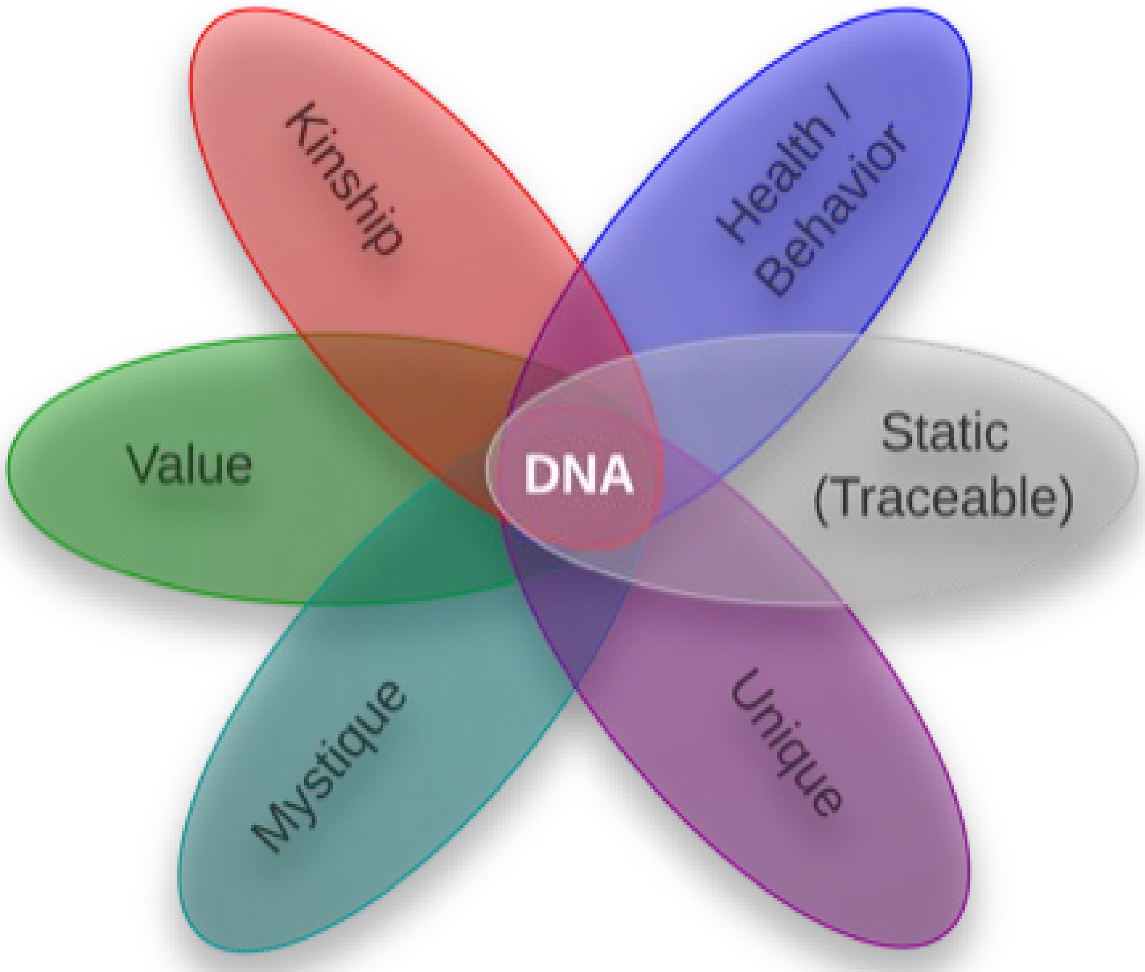
DNAnexus

SECURITY REQUIREMENTS

- Architecture
  - Consider zero-trust architecture
- Robust and auditable authentication and authorization
- Continuous Monitoring
  - (24x7x365) with a high degree of automation
  - Automate the Incident Response Procedure as much as possible
- Accountability to industry standards:
  - FedRAMP “Moderate,” “High” or similar standards (NIST 800-family)
  - ISO27000 family of standards
  - Discovery & sampling audit at frequent intervals (trust & verify)

PRIVACY REQUIREMENTS

- Compliant with the applicable privacy regulations
  - GDPR for citizens of the EU
  - Federal, state, and local regulations regarding PII and PHI
- Consent management
  - Ongoing management of contributor's consent with audit trails
  - Managing familial scans (law enforcement)
- Automated data privacy scans
  - Support for DSARs, etc.



Are long read genomic sequences inherently identifiable?

DATA REQUIREMENTS

- **C**onfidentiality, Integrity and Availability (CIA)
  - Encrypted data in transit and at rest
  - Evidence proving the accuracy and consistency throughout the lifespan of the data
  - Code/data are available to only those who are authorized within time limits
- **G**overnance
  - Establishing the appropriate authorizations on code/data
  - What must be retained? Where? For how long?
- **P**rovenance
  - Privacy concerns, alignment with consent, ability to track and prove compliance
  - Metadata – source of sequence, sequencer, processing steps
- **Q**uality
  - How do you measure quality per use case?

SOFTWARE REQUIREMENTS

- Unconstrained by the size of the sequences and “-omics” type
- Horizontal and vertical scaling – application, metadata processing, etc.
- Walled Garden vs Open Federation
  - Users add their own software, which could contain malware, crypto mining, unsupported and vulnerable supporting libraries (log4j)
- License Management in leveraged software (Asset Mgmt)
  - Open Source Software – permissive, viral light, highly viral, Affero
  - Commercial Software – terms need to balance with usage
- Auditability
  - Immutable logs showing *every action performed on every object*

NETWORK REQUIREMENTS

- Volume and velocity
  - Increasing at 40+% per year
- Centralized repository or Federated repository
  - Network usage varies on the overall design
- Enforcement of data localization regulations and agreements
  - Privacy regulations require data to stay within a country/region’s jurisdictions and how the data can be used.
  - Commercial agreements control where the data resides, how it can be accessed and used.

CLOSING THOUGHTS

- Zero-Trust - verify and corroborate
- Track-and-Trace – anything you say, be prepared to prove it!
- Governance

Balancing Security, Privacy, Quality, and Science

Session 3: Cybersecurity Challenges for Genomic Data Storage

Xiaofeng Wang (Indiana University)

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Privacy in the Genomic Era

XiaoFeng Wang, IEEE Fellow, Rudy Professor at IUB

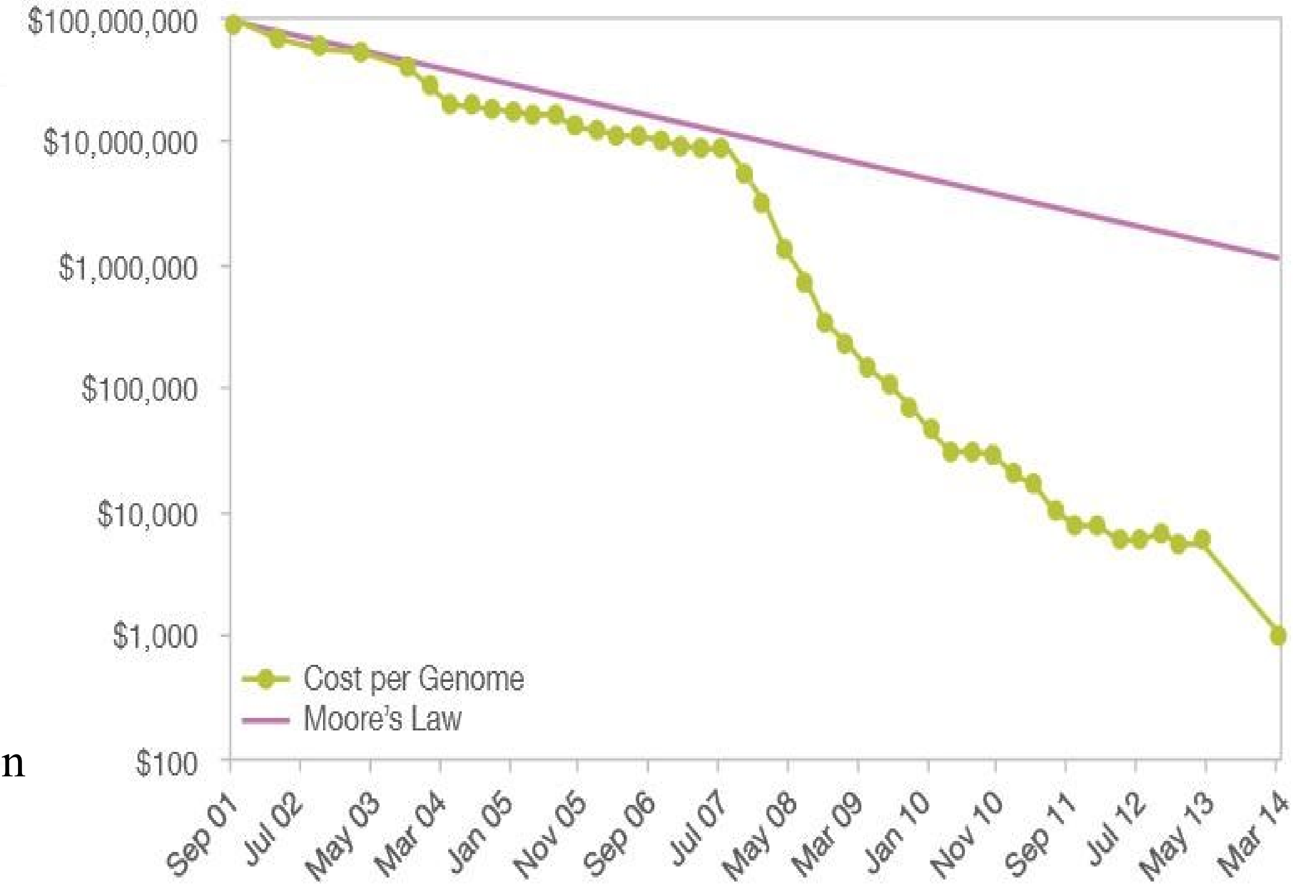
http://www.informatics.indiana.edu/xw7

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Genomic Revolution

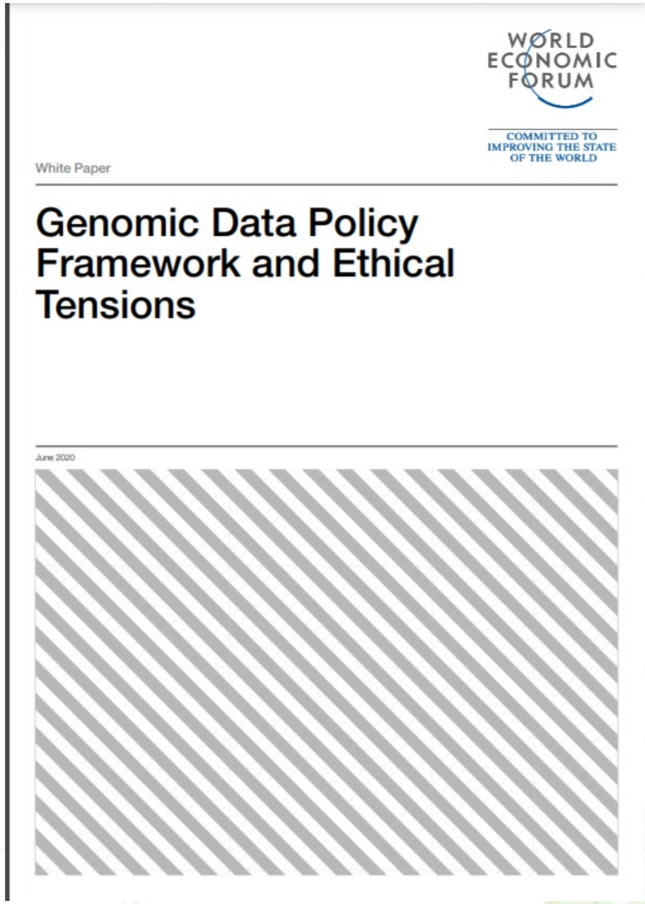
- Fast drop in the cost of genome-sequencing
  - 2000: $3 billion
  - Mar. 2021: $800 - $1,000
  - Genotyping 1M variations: below $200



- Unleashing the potential of the technology
  - Healthcare: e.g., disease risk detection, personalized medicine
  - Biomedical research: e.g., geno-phono association
  - Legal and forensic
  - DTC: e.g., ancestry test, paternity test
  - ……

Genome Privacy

- Privacy risks
  - ➢ Genetic disease disclosure
  - ➢ Collateral damage
  - ➢ Genetic discrimination …
- Protection
  - ➢ Clear access policies
  - ➢ Accountability
  - ➢ Data anonymization
  - ➢ Best practice for data privacy
  - ➢ Privacy awareness ……



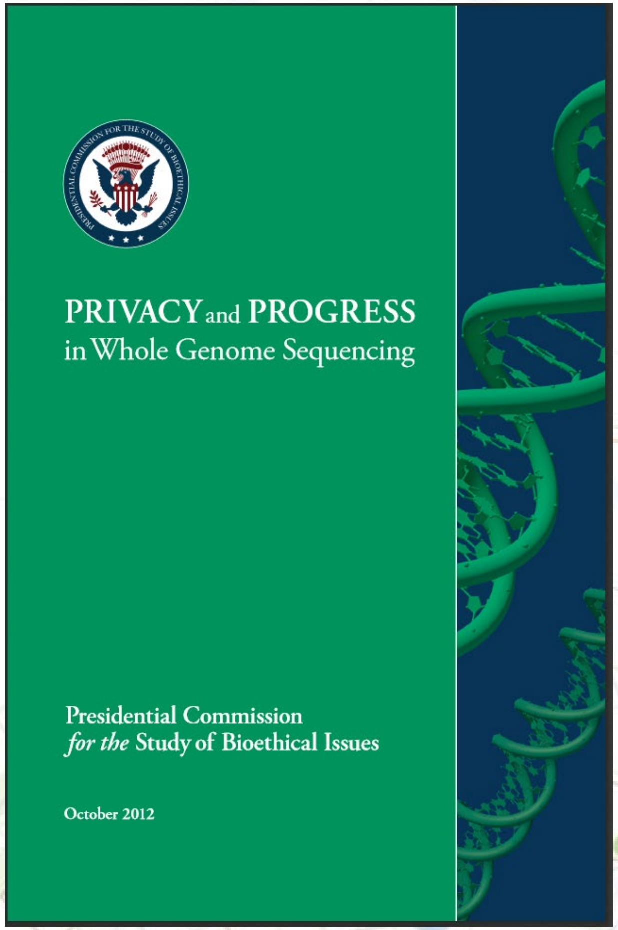
WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

White Paper

Genomic Data Policy Framework and Ethical Tensions

June 2020



PRIVACY and PROGRESS in Whole Genome Sequencing

Presidential Commission *for the* Study of Bioethical Issues

October 2012

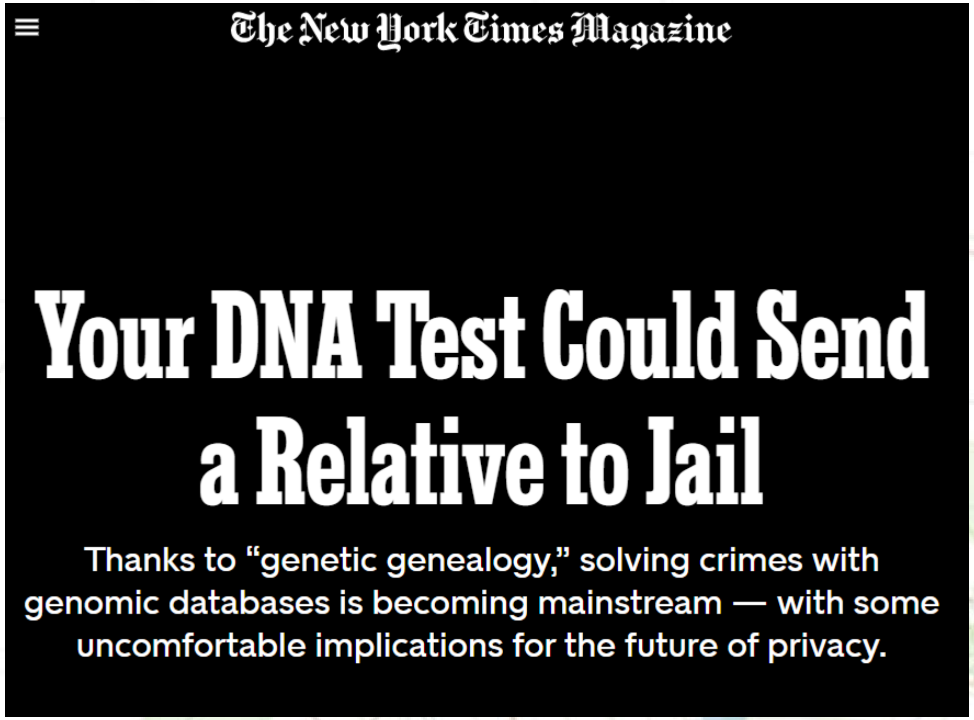
For more information: Privacy and Security in the Genomic Era by Naveed, E. Ayday, E. Clayton, J. Fellay, C. Gunter, JP Hubaux, B. Malin and X. Wang

Available at http://arxiv.org/pdf/1405.1891v1.pdf

Disclosed Genomic Data can be Abused

- Hawasupai case (1989): use of Indian tribe genome data without proper informed consent, with impacts on NIH's All of Us project (2020)

- Genomic data for solving crimes (with privacy implications)
  - ➢ E.g., Capture of the Golden State Killer (through GEDmatch)
  - ➢ But privacy concern is raised: how one's individual choice affects others?



The New York Times Magazine

Your DNA Test Could Send a Relative to Jail

Thanks to "genetic genealogy," solving crimes with genomic databases is becoming mainstream — with some uncomfortable implications for the future of privacy.

Unauthorized Disclosure of DNA/Meta Data Continue to Happen

- DNA Diagnostics Center (DDC), breach more than 2.1 million people (2021)

- GEDmatch hack causes email addresses from its users to be used in a phishing attack on another leading genealogy site (2020)

- Veritas Genetics claim a data breach resulted in unauthorized access of some customer information (2019)

……

Genomic Privacy: Technical Challenges

- Dissemination: privacy protection is difficult !
  - Anonymization is hard: genotype to phenotype mapping
  - Impact of genetic genealogy
  - Extremely high dimensions: hard to balance between privacy and utility

- Computing: big data analysis
  - Beyond the capability of existing secure computing technologies
    - NIH originally disallows reads with human DNA to be given to the public Cloud
    - Now, use the cloud at your own risk

Challenge in Privacy-preserving Genomic Data Sharing

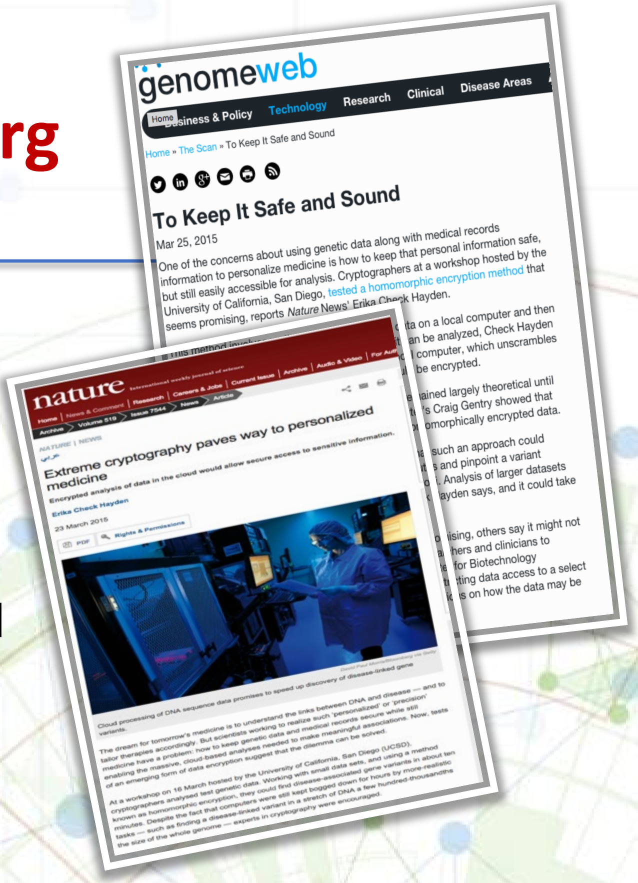
- Old problems:
  - Statistical inference control, access control, query auditing…
- However, genome data are special:
  - Special structures, e.g. linkage disequilibrium
  - Existence of reference genomic data that are publicly available (e.g. large population studies as HapMap, WTCCC, 1000 Genome)
- Examples:
  - Homer's attack and NIH's responses (2008)
  - Our analysis on test statistics released by GWAS papers (2009)
  - Shringarpure and Bustamante's attack on beacons (2015)

iDASH Genomic Data Privacy and Security Protection Competition

Since 2014, http://www.humangenomeprivacy.org

An interdisciplinary challenge on genomic privacy research

- Motivated by real world biomedical applications and with participation of privacy technology experts, Biomedical and ELSI researchers (academia and industry)
- Develop practical solutions for privacy preserving genomic data sharing and analysis
- Demonstrate the feasibility of secure genome analysis and dissemination using DP, MPC, HE, TEE
- Reported in the media (e.g., Nature News)



Home » The Scan » To Keep It Safe and Sound

To Keep It Safe and Sound

Mar 25, 2015

One of the concerns about using genetic data along with medical records information to personalize medicine is how to keep that personal information safe, but still easily accessible for analysis. Cryptographers at a workshop hosted by the University of California, San Diego, tested a homomorphic encryption method that seems promising, reports *Nature* News' Erika Check Hayden.

This method invol… ata on a local computer and then … an be analyzed, Check Hayden … l computer, which unscrambles … be encrypted.

… ained largely theoretical until … 's Craig Gentry showed that … omorphically encrypted data.

… such an approach could … s and pinpoint a variant … . Analysis of larger datasets … ayden says, and it could take

… ising, others say it might not … hers and clinicians to … for Biotechnology … ting data access to a select … s on how the data may be

NATURE | NEWS

Extreme cryptography paves way to personalized medicine

Encrypted analysis of data in the cloud would allow secure access to sensitive information.

Erika Check Hayden

23 March 2015

PDF Rights & Permissions

David Paul Morris/Bloomberg via Getty

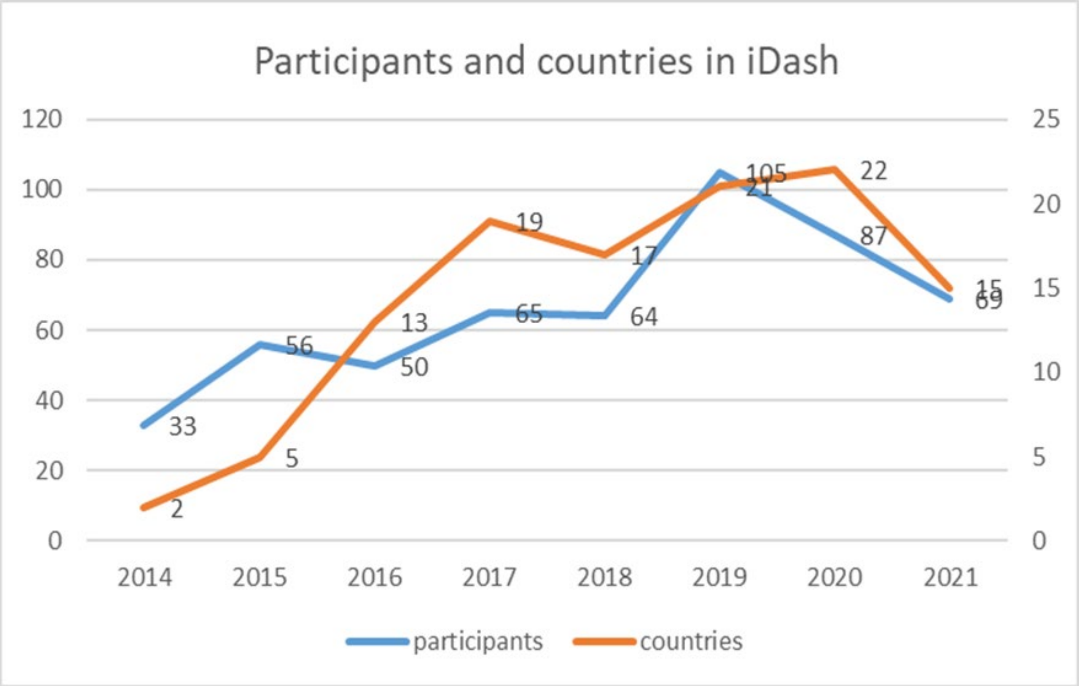
Cloud processing of DNA sequence data promises to speed up discovery of disease-linked gene variants.

The dream for tomorrow's medicine is to understand the links between DNA and disease — and to tailor therapies accordingly. But scientists working to realize such 'personalized' or 'precision' medicine have a problem: how to keep genetic data and medical records secure while still enabling the massive, cloud-based analyses needed to make meaningful associations. Now, tests of an emerging form of data encryption suggest that the dilemma can be solved.

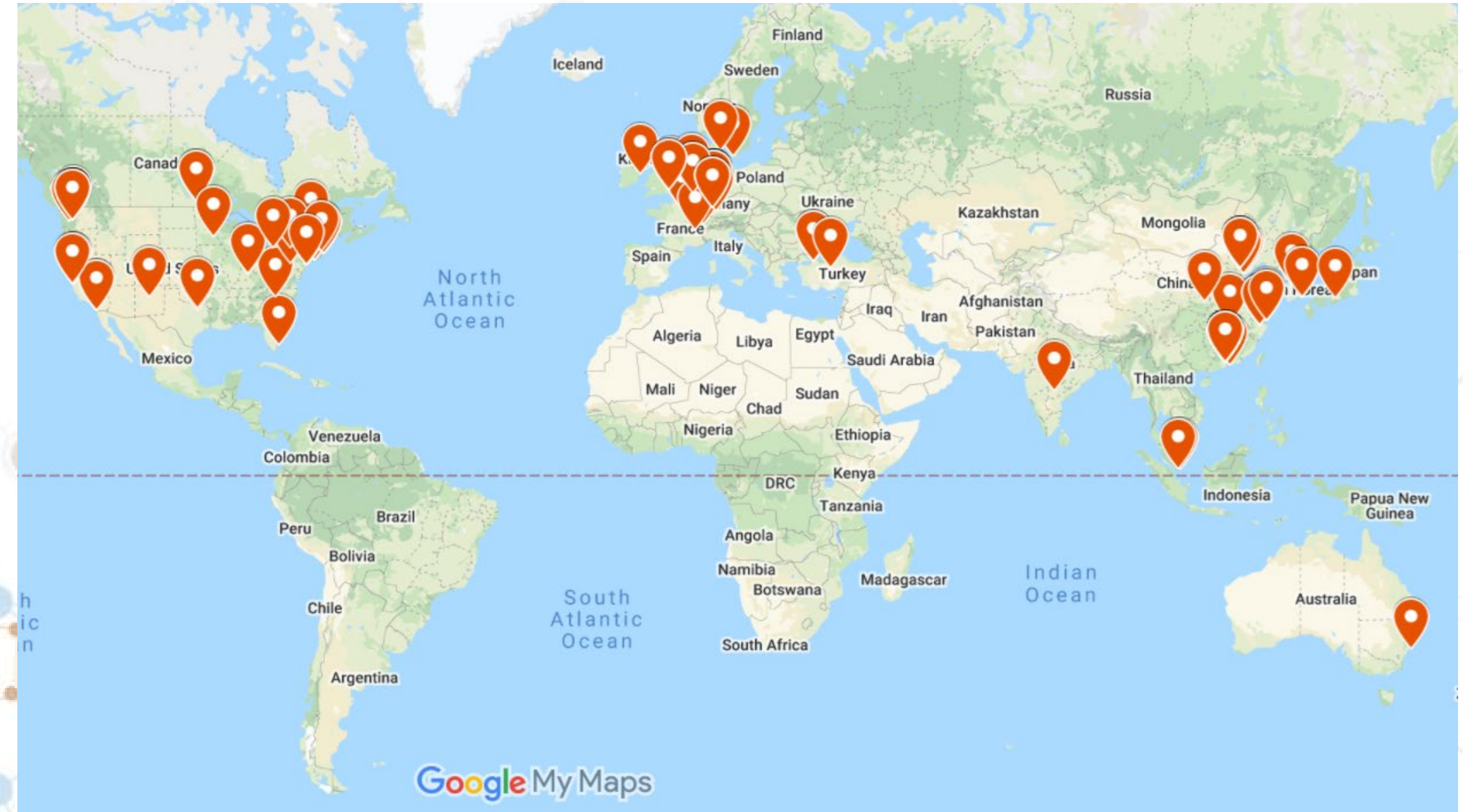
At a workshop on 16 March hosted by the University of California, San Diego (UCSD), cryptographers analysed test genetic data. Working with small data sets, and using a method known as homomorphic encryption, they could find disease-associated gene variants in about ten minutes. Despite the fact that computers were still kept bogged down for hours by more-realistic tasks — such as finding a disease-linked variant in a stretch of DNA a few hundred-thousandths the size of the whole genome — experts in cryptography were encouraged.

Topics and Trend from 2014 to 2021

Privacy-preserving Data Sharing	Encryption Testing
Secure Release	De-duplication
Secure Outsourcing	Software Guard Extensions
Homomorphic Encryption	Secure Search
Secure Collaboration	Blockchain and Smart Contract
Secure Multiparty Computation	Secure Machine Learning
Beacon Service	Privacy-preserving Machine Learning
Privacy-preserving Search	



Participation Around the World



- Academia: Cornell, MIT, UTHealth, UCSD, Yale, Purdue, Vanderbilt, EPFL, SNU , CUHK, Manitoba …
- Industry: IBM, MSR, Samsung, Alibaba, Tencent, Baidu …
- Government: Sandia National Lab, French Alternative Energies and Atomic Energy Commission …

Contributions to the Progress in Genome Privacy

For the task secure multi-label tumor classification using Homomorphic Encryption in 2020, most teams are utilizing linear/logistic regression models to implement cancer classification. These models have been improved significantly over the past few years in the HE competition, which is quite scalable and efficient now. The top solutions achieved a Micro-AUC of 0.97 to classify 11 cancer types from encrypted genetic variants of 909 samples within 5 minutes.

For the task differentially private federated learning for the cancer prediction model in 2020, the submitted solutions achieved almost perfect model accuracies while enforcing a high differential privacy standard (privacy budget of 3.0 or lower). The training process of the best-performing solution is very fast, comparable with the efficiency of training a machine learning model with all data by a single party.

For the task of data sharing consent for health-related data using contracts on blockchain in 2021, it is feasible to store patient consent sharing preference records for seven categories for a given clinical/genomic study on blockchain up to ~6,800 records per hour (or ~1.889 records per second).

Acknowledgement

- NIH R01HG010798: “Secure and Privacy-preserving Genome-wide and Phenome-wide Association Studies via Intel Software Guard Extensions (SGX)”

- NIH R01HG007078: “Privacy Preserving Technologies for Human Genome Data Analysis and Dissemination”

- NSF-CNS-1408874: “Broker Leads for Privacy-Preserving Discovery in Health Information Exchange”

More Information:

1. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study, 2009, ACM CCS

2. Addressing Beacon re-identification attacks: Quantification and mitigation of privacy risks, 2017, JAMIA

3. Real-time Protection of Genomic Data Sharing in Beacon Services, 2018, AMIA

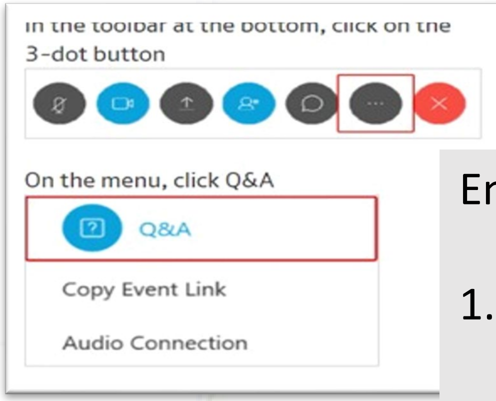
4. A Secure Alignment Algorithm for Mapping Short Reads to Human Genome, 2018, RECOMB

5. MBeacon: Privacy-Preserving Beacons for DNA Methylation Data, 2019, NDSS

6. Haplotype-based membership inference from summary genomic data, 2021 Bioinformatics

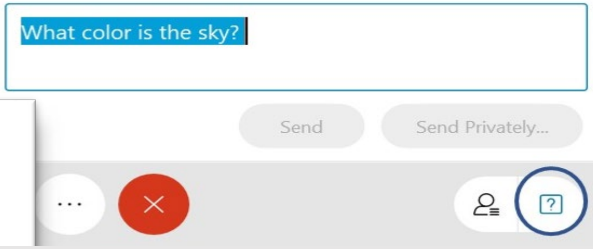
Cybersecurity Challenges

Moderated Questions and Answers



in the toolbar at the bottom, click on the 3-dot button

On the menu, click Q&A



What color is the sky?

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.

Session 4: Privacy Challenges for Genomic Data

Sumitra Muralidhar (Department of Veterans Affairs)

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

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

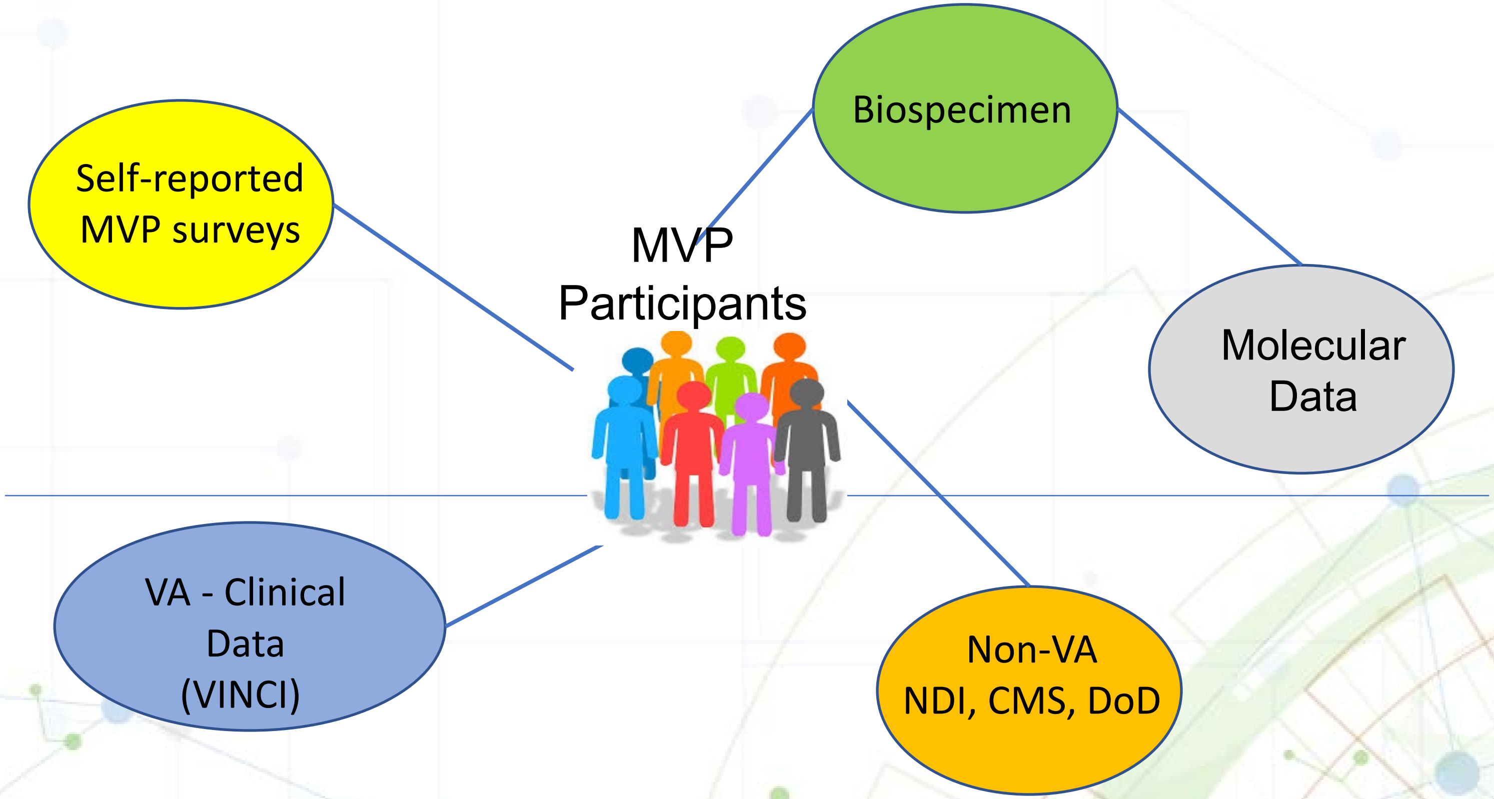
Million Veteran Program (MVP)
Department of Veterans Affairs

Sumitra Muralidhar, PhD
Director, MVP
VA Office of Research and Development

Million Veteran Program (MVP) Overview

- MVP is a **national VA research program**, launched in 2011, designed to advance precision health care by learning how genes, lifestyle, and military experiences and exposures affect health and illness
  - Establish a comprehensive, diverse cohort of at least one million Veterans
  - Provide broad access to the data for scientific discovery
  - Establish pipelines to translate discoveries to the clinic to improve the health of Veterans
- MVP is one of the world's largest healthcare system-based research programs of its kind with **over 864,000 Veterans enrolled (as of Dec. 2021)**

MVP Data Universe



MVP Biospecimen Data Overview

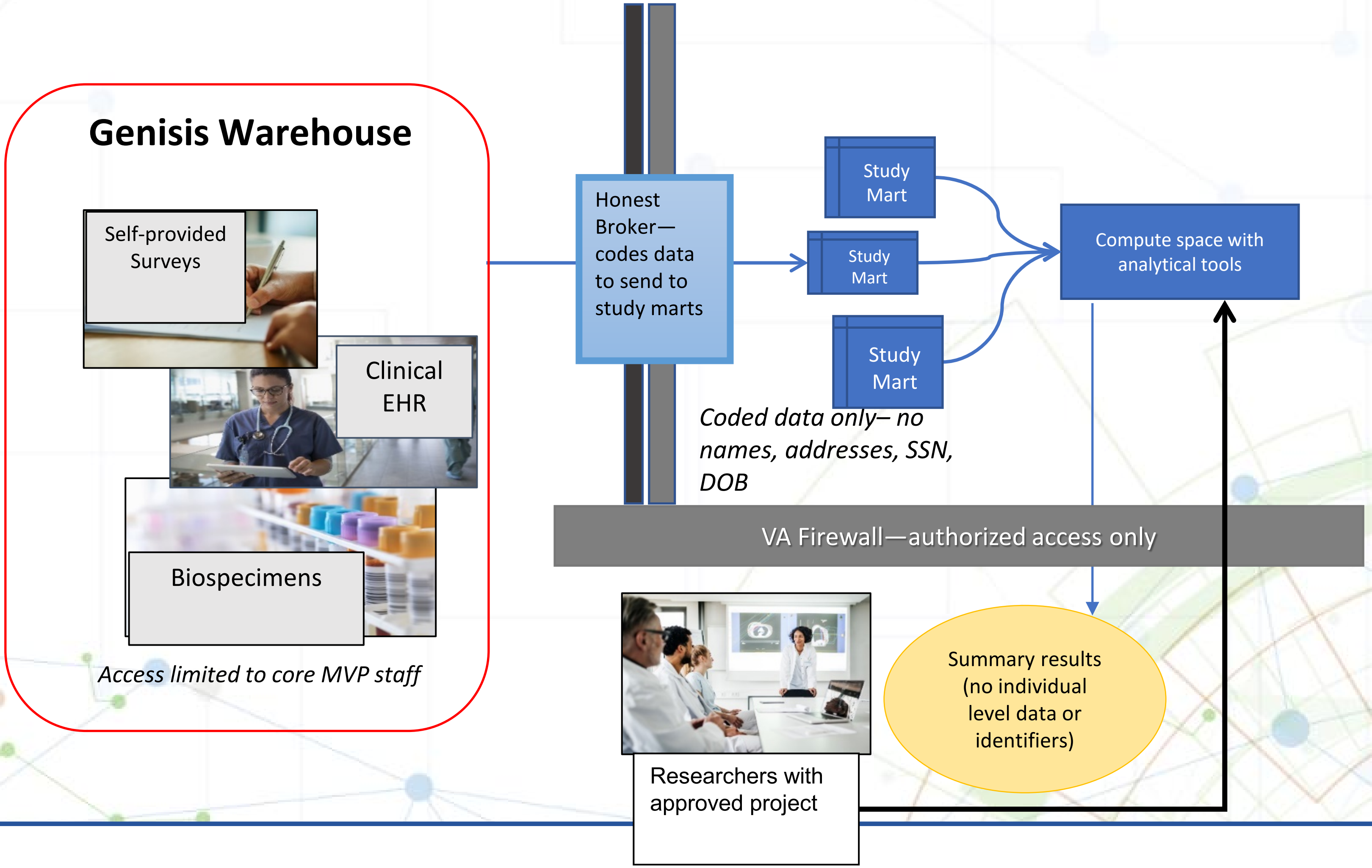
GOAL: Generate the maximum amount of data from biospecimens to enhance scientific discovery

- Baseline genetic data profile (genotype) generated for all participants
  - Data from 650,000 samples are currently provided to approved researchers
- Genetic data for specific ethnic groups (Blacks, Hispanics and Asians) using a customized analytic tool currently underway for ~ 200K participants
- Whole genome sequences have been generated on ~ 140,000 samples
  - Processing underway
- Other data such as proteomics and metabolomics are being piloted

Balancing Data Privacy/Security and Access

- Bring researchers to the data in a central secure scientific computing platform
  - Computing infrastructure within the VA meets VA IT Privacy and Security requirements; DOE and the University of Chicago VA Data Commons have an approved VA authority to operate (ATO)
- Biospecimen (blood sample) and data (surveys) collected are labeled using a code instead of identifiable information
- New ship ID created for sample send-outs to vendors
- Crosswalk to identity of participant is held by few authorized core staff
- Researchers access only coded data (no direct identifiers such as SSN, name, date of birth, street address)
- Researchers sign rules of behavior and analyze data in a central, secure computing system
- No data leaves the system; only summary results can be taken out

MVP Data Access Model



Genisis Warehouse

Self-provided Surveys

Clinical EHR

Biospecimens

Access limited to core MVP staff

Honest Broker—codes data to send to study marts

Study Mart

Study Mart

Study Mart

Compute space with analytical tools

Coded data only– no names, addresses, SSN, DOB

VA Firewall—authorized access only

Researchers with approved project

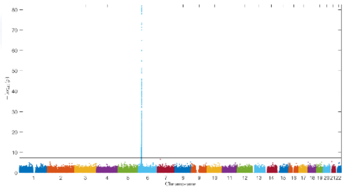
Summary results (no individual level data or identifiers)

External access to MVP Data

- VA Data Commons : allow data access broadly to investigators within and outside the VA
- Contract with the University of Chicago (UoC)
- Data deidentified at the VA and moved to UoC
  - Safe harbor method plus
  - Formal expert statistical determination
- Data will be migrated to a cloud compute infrastructure for many simultaneous approved users
- Beta-testing in FY 2022 ; piloting in FY23

MVP Summary Data Access in dbGaP

15 publications



96 analyses

7 phenotype categories

166 authorized requests for access

https://www.ncbi.nlm.nih.gov/gap/

Search "VA Million Veteran Program"

Reidentification Risk

- **Re-Identification:** the ability to determine whether an individual is included in a pooled sample, based on the allele frequencies in the pool -- without the need to access individual-level genotype data of that pooled data set
- All the published references discussing re-identification are theoretical, not actual case reports of participant re-identification
- In order for re-identification to occur, the user must already have access to that person’s genetic information from another source
- Accuracy of re-identification is determined by:
  - the size of the population (small sample size = better accuracy)
  - the diversity of the population (homogenous population = better accuracy)
  - the frequency of the genetic variants (rare genetic variants = better accuracy)

MVP Risk Mitigation Strategies

- MVP is sufficiently large and diverse, therefore theoretical re-identification risk is extremely low

  Only aggregate results will be shared, no individual-level data

- Additional steps taken to further reduce risk
  - Results filtered to only include genetic variants with a minor allele count > 30 or minor allele frequency > 0.005, whichever is less (metrics should be based on the subset of the study population actually used for the analysis, not the general population)
  - Total study population used for the analysis must be >3000 participants
  - If a case-control study, there must be >500 cases in the analysis

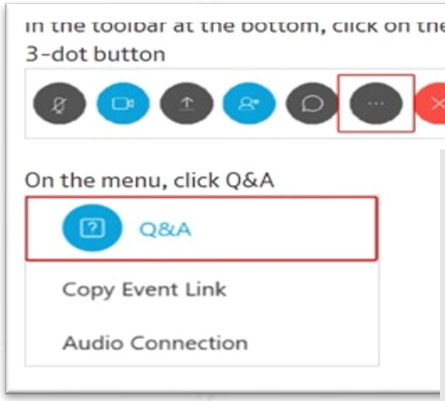
Thank you!

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

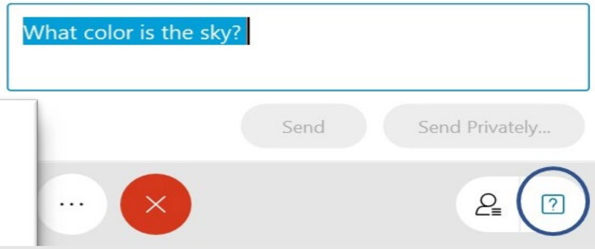
Privacy Challenges for Genomic Data

Moderated Questions and Answers



In the toolbar at the bottom, click on the 3-dot button

On the menu, click Q&A



What color is the sky?

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.

Session 5: Current and Future Genomic Data Use Challenges

Gail Jarvik (American Society of Human Genetics [ASHG])
Ankit Malhotra (AWS)
Heidi Sofia (NIH National Human Genome Research Institute [NHGRI])

Human Genetics & Genomics Research: Data-Sharing & Privacy

Gail Jarvik MD, PhD

2021 President,

American Society of Human Genetics

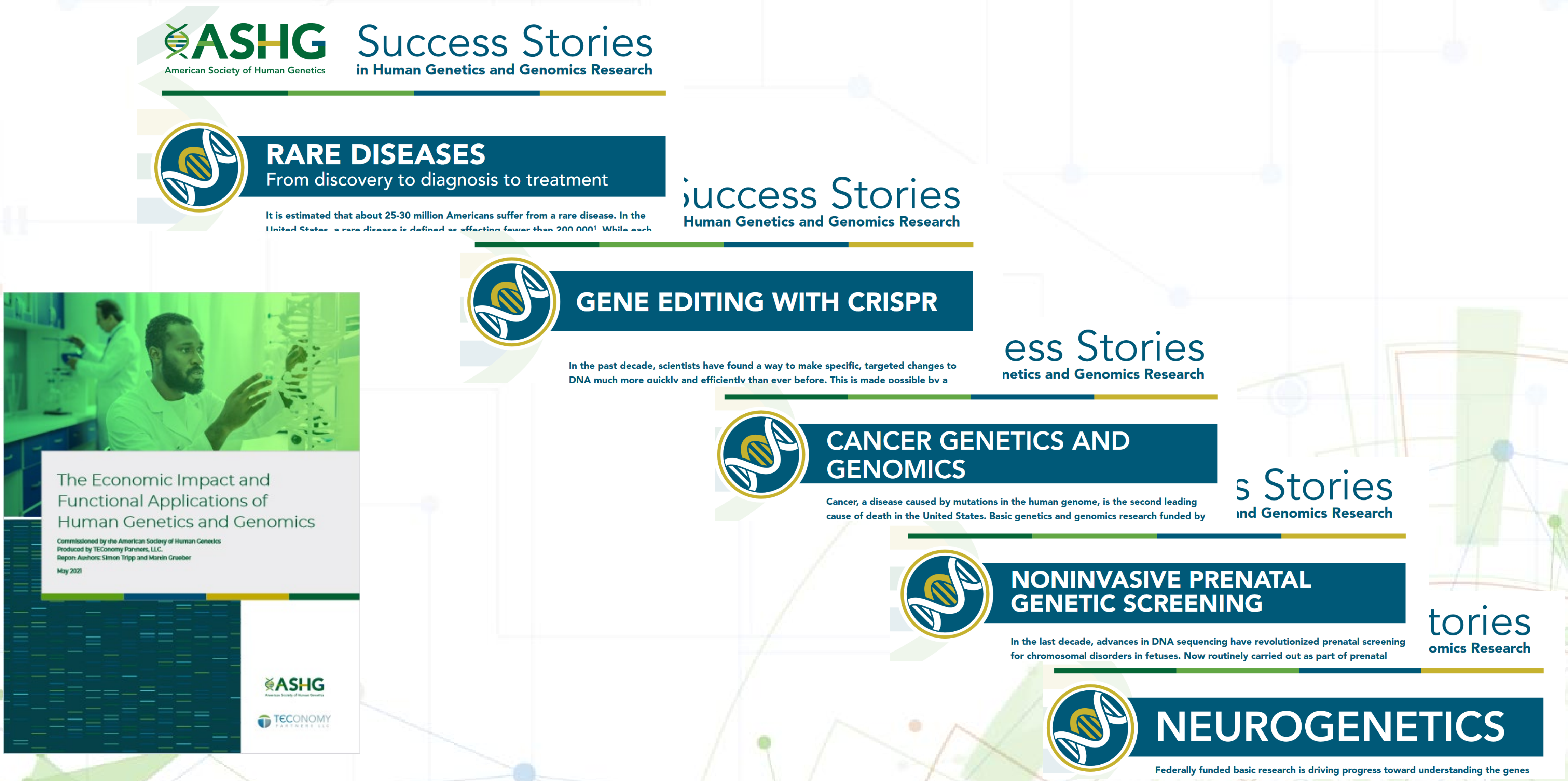
January 26, 2022

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

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

American Society of Human Genetics

- **Mission**: *Advance human genetics and genomics in science, health, and society through excellence in research, education and advocacy*
- **Vision**: *People everywhere realize the benefits of human genetics and genomics research*
- **Annual Meeting**: *Attracts up to 9,000 attendees*
- **Year-Round Scientific Programs**
- **Two Scientific Journals**:
  - *American Journal of Human Genetics*
  - *Human Genetics and Genomics Advances*



ASHG
American Society of Human Genetics

Success Stories
in Human Genetics and Genomics Research

RARE DISEASES
From discovery to diagnosis to treatment

It is estimated that about 25-30 million Americans suffer from a rare disease. In the United States, a rare disease is defined as affecting fewer than 200,000[1]. While each

uccess Stories
Human Genetics and Genomics Research

GENE EDITING WITH CRISPR

In the past decade, scientists have found a way to make specific, targeted changes to DNA much more quickly and efficiently than ever before. This is made possible by a

ess Stories
netics and Genomics Research

CANCER GENETICS AND GENOMICS

Cancer, a disease caused by mutations in the human genome, is the second leading cause of death in the United States. Basic genetics and genomics research funded by

s Stories
nd Genomics Research

NONINVASIVE PRENATAL GENETIC SCREENING

In the last decade, advances in DNA sequencing have revolutionized prenatal screening for chromosomal disorders in fetuses. Now routinely carried out as part of prenatal

tories
omics Research

NEUROGENETICS

Federally funded basic research is driving progress toward understanding the genes

The Economic Impact and Functional Applications of Human Genetics and Genomics

Commissioned by the American Society of Human Genetics
Produced by TEConomy Partners, LLC.
Report Authors: Simon Tripp and Martin Grueber

May 2021

ASHG
American Society of Human Genetics

TEConomy
PARTNERS LLC

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

NCCoE
NATIONAL CYBERSECURITY
CENTER OF EXCELLENCE

Data-sharing Fuels Progress in Human Genetics & Genomics Research

- Broad data-sharing a hallmark of the human genetics community
  - Essential for completion of Human Genome Project
- Data-sharing fundamental for continued advances in research & medicine

Policies and Systems Need to Maintain Privacy of Research Participants

- Acquisition, analysis, sharing of human genetic data, use of genetic tools, need to be conducted responsibly
- ASHG supports policies that strengthen research participant privacy
  - Genetic Information Nondiscrimination Act
  - 21st Century Cures Act
  - Common Rule
  - NIH Genomic Data Sharing Policy

Privacy/Security Essential for Public Participation in Genetics Research

What are the biggest barriers or concerns to your participation in human genetics research? (Choose all that apply)

Concerns about the security of the database where my data would be stored	48%
Concerns about the privacy of my genetic information	47%
Concerns that my genetic information could influence my access to health care	40%
There are no studies currently available to me	34%
Concerns that my genetic information could influence my access to insurance or a job	34%
I cannot participate due to financial reasons	28%
Concerns that my genetic information could reveal unwelcome information about others or myself	25%
Participating in a study will take too much time	24%

Source: A Research!America poll of U.S. adults conducted in partnership with Zogby Analytics in December 2019

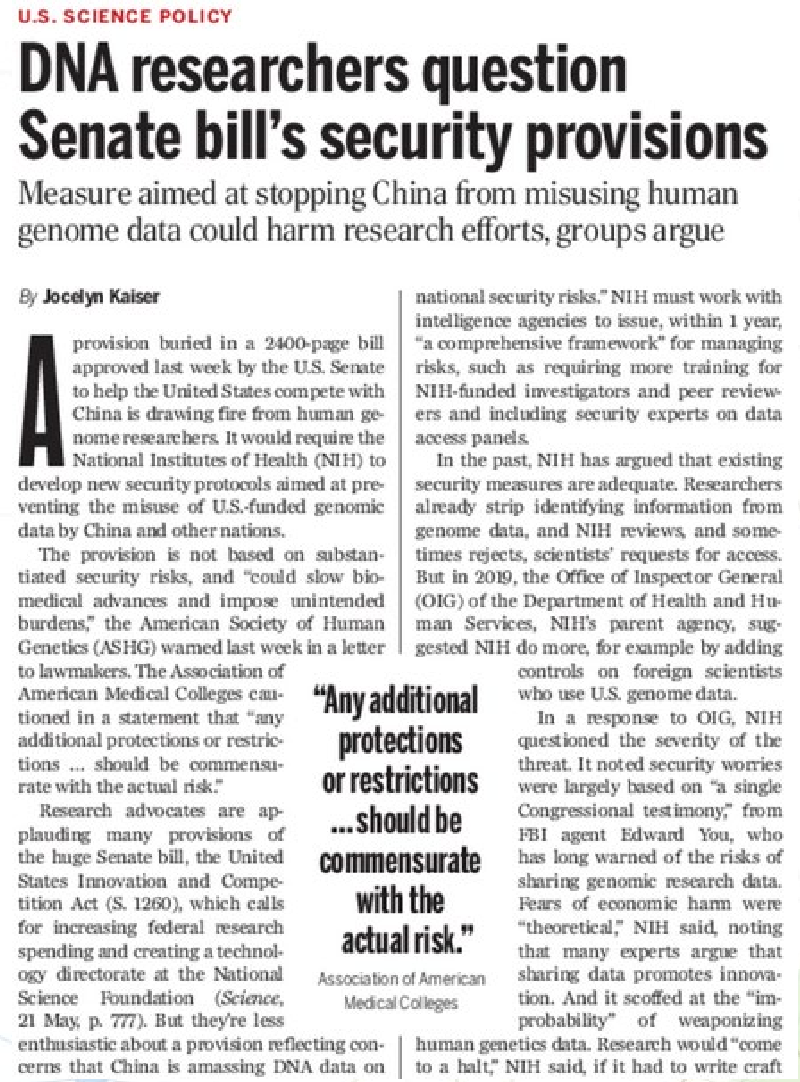
Policies and Systems Must Enable Science, Maintain Privacy

Data-sharing

Advance science

Privacy

Protect participants



U.S. SCIENCE POLICY

DNA researchers question Senate bill's security provisions

Measure aimed at stopping China from misusing human genome data could harm research efforts, groups argue

By **Jocelyn Kaiser**

A provision buried in a 2400-page bill approved last week by the U.S. Senate to help the United States compete with China is drawing fire from human genome researchers. It would require the National Institutes of Health (NIH) to develop new security protocols aimed at preventing the misuse of U.S.-funded genomic data by China and other nations.

The provision is not based on substantiated security risks, and "could slow biomedical advances and impose unintended burdens," the American Society of Human Genetics (ASHG) warned last week in a letter to lawmakers. The Association of American Medical Colleges cautioned in a statement that "any additional protections or restrictions ... should be commensurate with the actual risk."

> "Any additional protections or restrictions ...should be commensurate with the actual risk."
>
> Association of American Medical Colleges

Research advocates are applauding many provisions of the huge Senate bill, the United States Innovation and Competition Act (S. 1260), which calls for increasing federal research spending and creating a technology directorate at the National Science Foundation (*Science*, 21 May, p. 777). But they're less enthusiastic about a provision reflecting concerns that China is amassing DNA data on national security risks." NIH must work with intelligence agencies to issue, within 1 year, "a comprehensive framework" for managing risks, such as requiring more training for NIH-funded investigators and peer reviewers and including security experts on data access panels.

In the past, NIH has argued that existing security measures are adequate. Researchers already strip identifying information from genome data, and NIH reviews, and sometimes rejects, scientists' requests for access. But in 2019, the Office of Inspector General (OIG) of the Department of Health and Human Services, NIH's parent agency, suggested NIH do more, for example by adding controls on foreign scientists who use U.S. genome data.

In a response to OIG, NIH questioned the severity of the threat. It noted security worries were largely based on "a single Congressional testimony," from FBI agent Edward You, who has long warned of the risks of sharing genomic research data. Fears of economic harm were "theoretical," NIH said, noting that many experts argue that sharing data promotes innovation. And it scoffed at the "improbability" of weaponizing human genetics data. Research would "come to a halt," NIH said, if it had to write craft

Science • 18 Jun 2021 • Vol 372, Issue 6548 • p. 1253

Thank you!

Contact: gjarvik@medicine.washington.edu

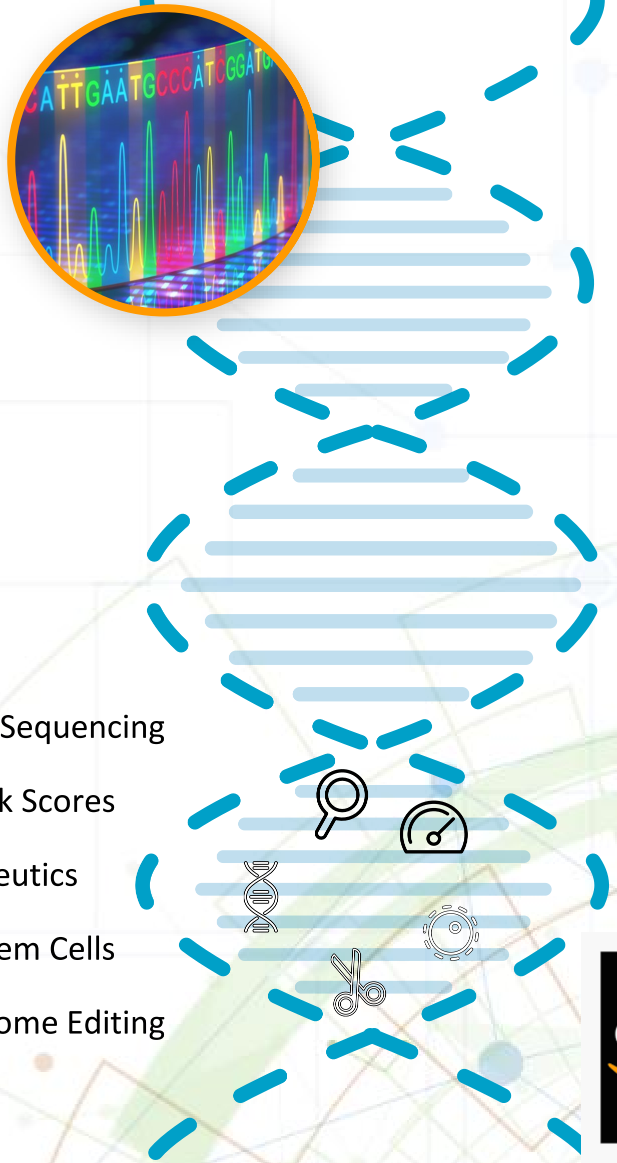
AWS for Genomics

Solving Challenges in Genomic Data Sharing

Ankit Malhotra, Ph.D.
Genomics Lead,
AWS Worldwide Public Sector Health

The precision medicine revolution

Transformative technologies in sequencing and computing is driving innovation across healthcare and enabling precision medicine.



Clinical Genome Sequencing

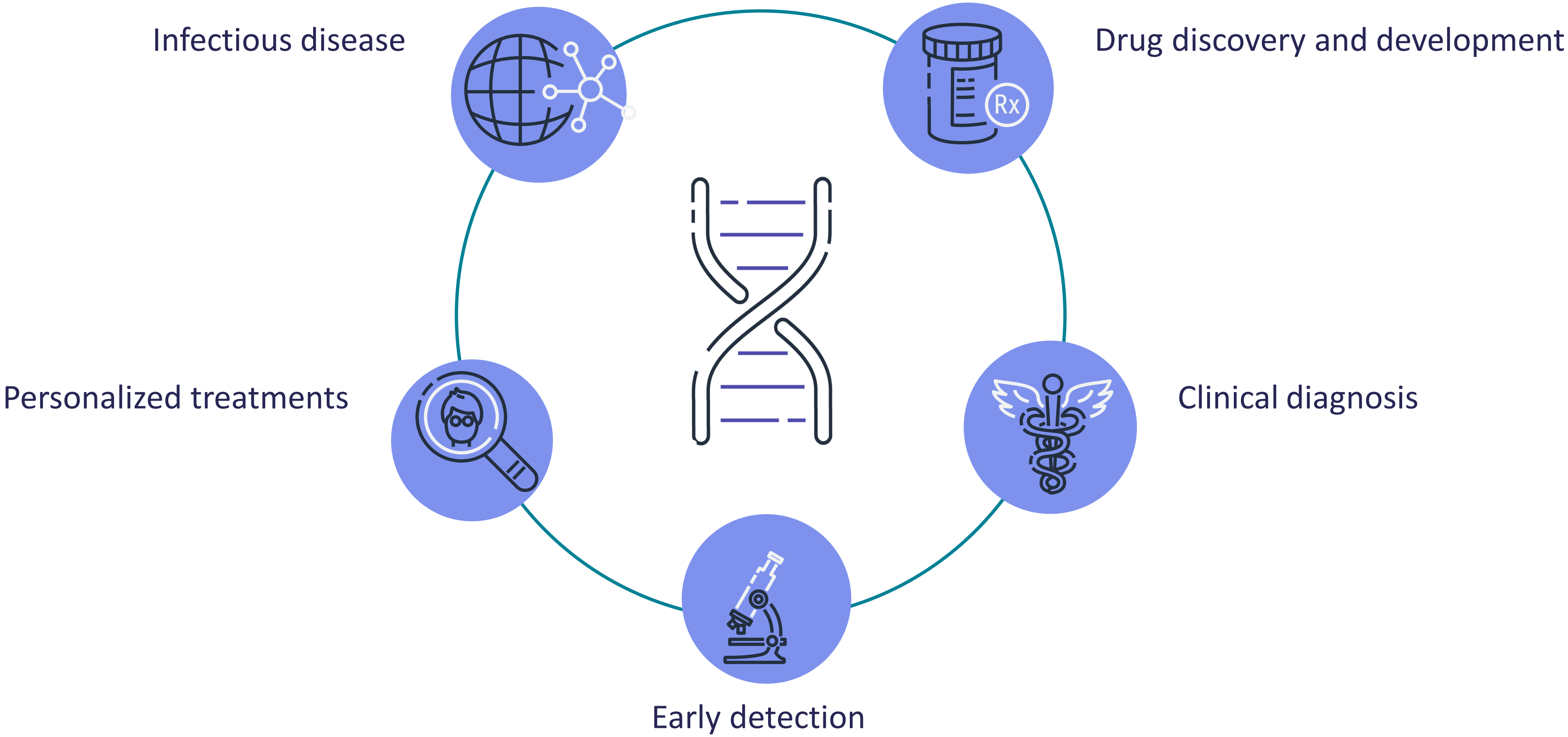
Genetic Risk Scores

Targeted Therapeutics

Induced Pluripotent Stem Cells

CRISPR Genome Editing

Genomics—a catalyst for personalized health



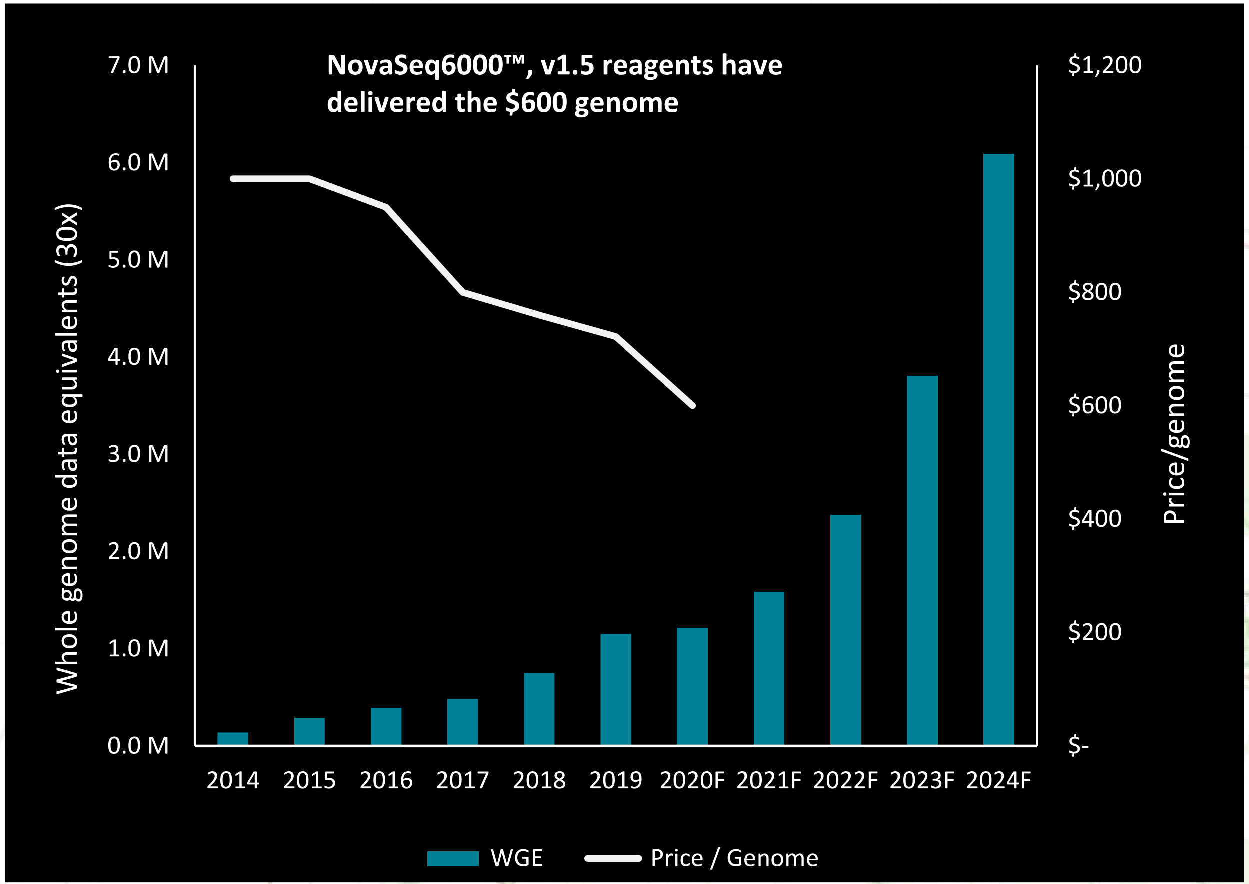
Challenges in leveraging genomics data

Large volumes of data needs to be transfer, stored, analyzed

Sequencing and analysis requires immense processing power, time

Frequently requires integration of multi-modal datasets

Protected health information must be secured


NovaSeq6000™, v1.5 reagents have delivered the $600 genome

Genomics on AWS

Data Transfer & Storage

Trusted partner for secure data transfer, life cycle management, storage cost optimization and digital preservation

Secondary Analysis & Workflow Automation

Manage multiple workflows, accelerate, simplify and scale data analysis with both flexibility and reproducibility

Data Aggregation & Governance

Harmonize multi-omic datasets and govern robust data access controls and permissions across a global infrastructure

Interpretation & Deep Learning

Turn big genomic data into actionable insights with a rich layer of sophisticated solutions and services

BUILD

AWS Modern Genomics Data Platform

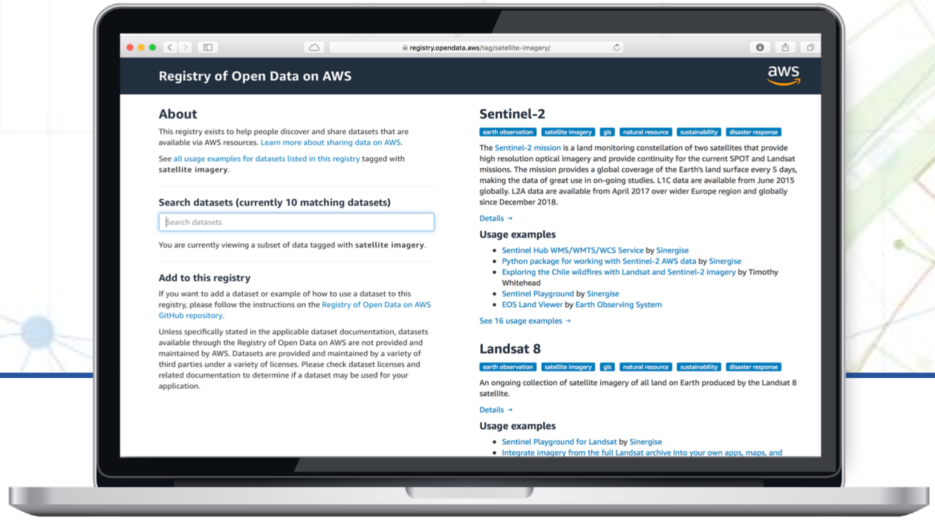
AWS Genomics CLI

SageMaker, HealthLake, Sequera Labs, Cromwell

BUY

databricks · ancestry · genuity science · illumina · paradigm4 · Oxford NANOPORE Technologies · Helix · IQVIA · REALM · Seven Bridges · GRAIL · DNAnexus · LIFEOMIC · lifebit

AWS Open Data Program



registry.opendata.aws/tag/satellite-imagery/

Registry of Open Data on AWS

aws

About

This registry exists to help people discover and share datasets that are available via AWS resources. Learn more about sharing data on AWS.

See all usage examples for datasets listed in this registry tagged with **satellite imagery**.

Search datasets (currently 10 matching datasets)

Search datasets

You are currently viewing a subset of data tagged with **satellite imagery**.

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

Sentinel-2

earth observation · satellite imagery · gis · natural resource · sustainability · disaster response

The Sentinel-2 mission is a land monitoring constellation of two satellites that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission provides a global coverage of the Earth's land surface every 5 days, making the data of great use in on-going studies. L1C data are available from June 2015 globally. L2A data are available from April 2017 over wider Europe region and globally since December 2018.

Details →

Usage examples

- Sentinel Hub WMS/WMTS/WCS Service by Sinergise
- Python package for working with Sentinel-2 AWS data by Sinergise
- Exploring the Chile wildfires with Landsat and Sentinel-2 imagery by Timothy Whitehead
- Sentinel Playground by Sinergise
- EOS Land Viewer by Earth Observing System

See 16 usage examples →

Landsat 8

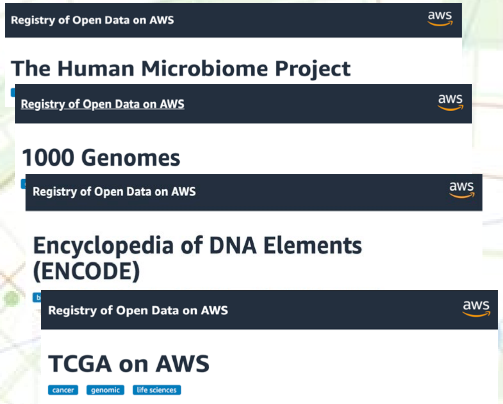
earth observation · satellite imagery · gis · natural resource · sustainability · disaster response

An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.

Details →

Usage examples

- Sentinel Playground for Landsat by Sinergise
- Integrate imagery from the full Landsat archive into your own apps, maps, and



Registry of Open Data on AWS

aws

The Human Microbiome Project

Registry of Open Data on AWS

aws

1000 Genomes

Registry of Open Data on AWS

aws

Encyclopedia of DNA Elements (ENCODE)

Registry of Open Data on AWS

aws

TCGA on AWS

cancer · genomic · life sciences

At last count there were 301 datasets (87 life science) hosted on AWS S3 as part of Registry of open data on AWS

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

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Open Access to Top Genomics Datasets

AWS hosts a variety of public datasets that anyone can access for free. Below are just a few examples

- 1000 Genomes Project
- The Cancer Genome Atlas
- International Cancer Genome Consortium
- 3000 Rice Genome
- Genome in a Bottle (GIAB)
- The Genome Modeling System
- Medicare Drug Spending
- The Human Connectome Project
- The Human Microbiome Project
- OpenNeuro
- Physionet
- Tabula muris
- gnoMAD
- and more….

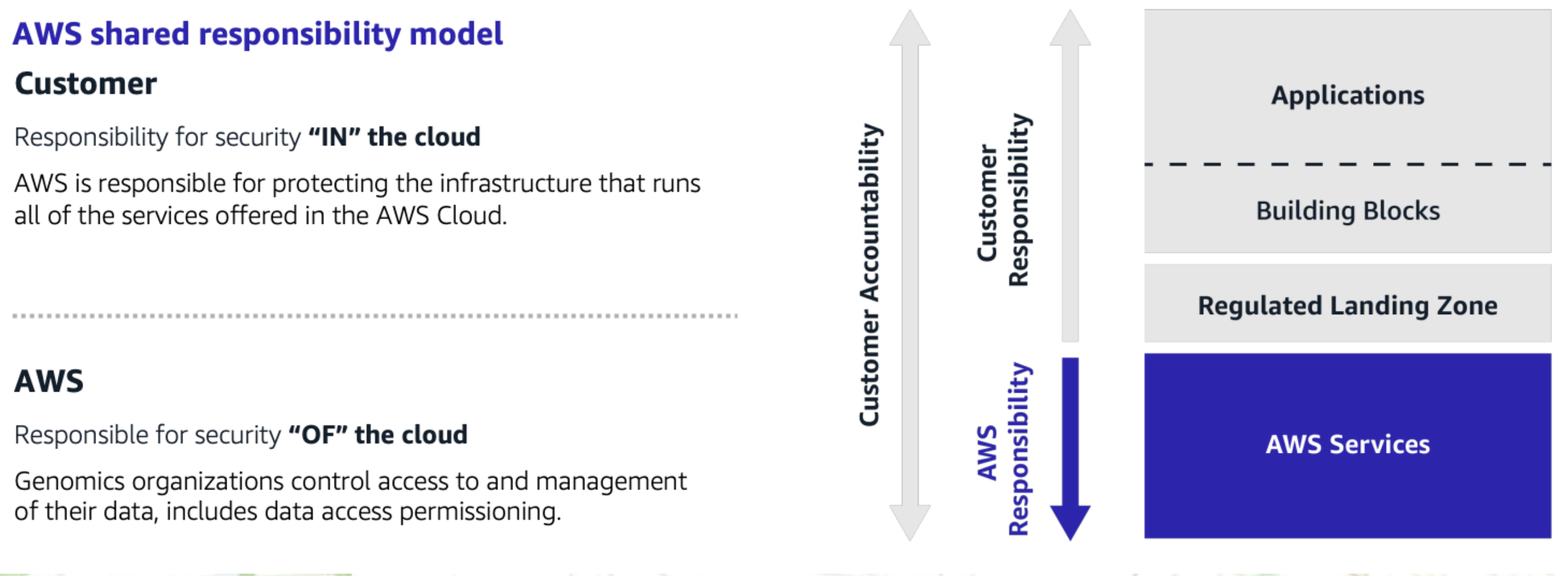
1000 Genomes
A Deep Catalog of Human Genetic Variation
ALLEN BRAIN ATLAS
DATA PORTAL
Beat AML
acute myeloid leukemia
BROAD INSTITUTE
CCLE Cancer Cell Line Encyclopedia
CPTAC
National Cancer Institute
at the National Institutes of Health
ChEMBL
ENCODE
FOUNDATION MEDICINE
Genome in a Bottle Consortium
HCMI
HMP
HUMAN PANGENOME PROJECT
ICGC
HART
IRGSP
INTERNATIONAL RICE GENOME SEQUENCING PROJECT
Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
MMRF CoMMpass Study
NIH
Nucleotide BLAST
nucleotide ▸ nucleotide
Open Targets
Pan-UK Biobank
refgenie
Tabula Muris
TARGET
TCGA
UBC Sunflower Genome Dataset
VGP
VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
VEP/LOFTEE
Genome Refs
GATK
qiime2docs

Security

AWS supports 98 security standards and compliance certifications, including HITRUST, GDPR compliance, FedRAMP, ISO 27001, and HIPAA.

Whitepaper - https://docs.aws.amazon.com/whitepapers/latest/aws-overview/security-and-compliance.html

AWS shared responsibility model



AWS shared responsibility model

Customer

Responsibility for security **"IN" the cloud**

AWS is responsible for protecting the infrastructure that runs all of the services offered in the AWS Cloud.

AWS

Responsible for security **"OF" the cloud**

Genomics organizations control access to and management of their data, includes data access permissioning.

AWS security, identity, and compliance solutions

Identity and access management	Detective controls	Infrastructure protection	Data protection	Incident response	Compliance
AWS Identity and Access Management (IAM) AWS Single Sign-On AWS Organizations AWS Directory Service Amazon Cognito AWS Resource Access Manager	AWS Security Hub Amazon GuardDuty Amazon Inspector Amazon CloudWatch AWS Config AWS CloudTrail VPC Flow Logs AWS IoT Device Defender	AWS Firewall Manager AWS Network Firewall AWS Shield AWS WAF – Web application firewall Amazon Virtual Private Cloud AWS PrivateLink AWS Systems Manager	Amazon Macie AWS Key Management Service (KMS) AWS CloudHSM AWS Certificate Manager AWS Secrets Manager AWS VPN Server-Side Encryption	Amazon Detective Amazon EventBridge AWS Backup AWS Security Hub CloudEndure Disaster Recovery	AWS Artifact AWS Audit Manager

Genomics england

Genomics England Develops Genomic and Health Information Platform on AWS to Turn Science into Healthcare

Challenge

Through the 100,000 Genomes Project alone, GEL amassed 50 petabytes of data. Seeking to make the data accessible to the research community, GEL is in the process of migrating its data to AWS to enable democratized access.

Solution

GEL is working with AWS to use compression technologies and other advanced tools to optimize cloud storage and analysis of genomic data based on the field's specific needs

Benefits

- To make genomic healthcare a reality, GEL is transitioning from project to platform, using Amazon Web Services (AWS) tools to give researchers reliable, comprehensive, and privacy-compliant access to these massive datasets. Through secure collaboration and analysis, this initiative will inform diagnoses, drive drug development, and unlock the future of precision medicine.

AstraZeneca

AstraZeneca is Raising the Bar with Running its Genome Sequencing Pipeline on AWS

Challenge

AstraZeneca's Centre for Genomic Research (CGR) has a bold target to analyze 2m genomes by 2026. However

- On-premise compute resources, which limit the performance capacity
- Dependency upon 3rd party informatics providers
- Orchestration of bioinformatics pipeline was time-intensive therefore costly and hard to scale

Solution

With support from AWS Professional Services, AstraZeneca built a highly scalable and high performance sequence data processing pipeline on AWS. The solution leveraged FPGA instances for compute and extensive use of Step Functions, Lambda, SQS and AWS Batch. The output is stored in scalable and highly secure AWS managed databases and S3 storage.

Benefits

- The bespoke pipeline was able to increase processing time by 2400%
- The results has been used to provided scientists advanced access to the clinical effects of natural mutations in humans that mimic drug inhibition/suppression

Impact of the pandemic

The SARS-Cov-2 pandemic caused widespread impact for healthcare systems around the world, and brought genomic sequencing and testing into the public eye.

This has also brought forth challenges in global data sharing:

- Privacy concerns - data misuse may lead to infringement of privacy for individuals and their relatives
  - Need for novel approaches to data anonymization for research use

- Compatibility and aggregation
  - accessing and reconciling duplications/differences from distributed data sources hindered meta-analyses

- Real / Near real time data ingestion

Use cases: genomic information in the cloud

NCI Genomic Data Commons
https://aws.amazon.com/solutions/case-studies/university-of-chicago-case-study/

University of Chicago Biomedical Research Hub (Gen3)
https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocab247/6432980

Hong Kong Genome Project (LifeBit)
https://lifebit.ai/blog/lifebit-awarded-a-four-year-contract-for-hong-kongs-genome-project/

CanCOGeN, Genome Canada, Illumina
https://www.genomecanada.ca/en/cancogen

GISAID
https://www.gisaid.org/

Undiagnosed Disease Network, Harvard School of Medicine (Service Workbench)
https://aws.amazon.com/blogs/publicsector/solving-medical-mysteries-aws-cloud-medical-data-sharing-innovation-undiagnosed-diseases-network/

UK Biobank (DNANexus)
https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform

Genomics Research and Responsible Data Sharing at NIH

Heidi Sofia, National Human Genome Research Institute

NIH PRIVACY & DATA SHARING RESEARCH

NHGRI - Mission of responsible Genomics Data Sharing

- ELSI – “Ethical, Legal, and Social Implications” program
- Technical privacy portfolio - research grants and small business
    - Homomorphic encryption, Secure Multiparty Computation, Differential Privacy, Secure Enclaves, Machine learning with privacy, etc.

NIH - Public Trust is the currency of the realm

- Data sharing through cloud commons (one copy instead of many copies)
    - AnVil, BioData Catalyst, CRDC, Kids First etc.
- Federated data sharing
- RAS Researcher Auth System
- DUOS - pilot of data access automation
- Privacy Preserving Record Linkage (PPRL)

ODSS - Catalyze modern computing at NIH

- **Office of Data Science Strategy – Susan Gregurick**
- https://datascience.nih.gov/

GENOMICS PRIVACY PORTFOLIO AT NIH

NCI	NEI	NHLBI
NHGRI	NIA	NIAAA
NIAID	NIAMS	NIBIB
NICHD	NIDCD	NIDCR
NIDDK	NIDA	NIEHS
NIGMS	NIMH	NIMHD
NINDS	NINR	NLM
CC	CIT	CSR
FIC	NCATS	NCCIH
OD		

27 NIH Institutes and Centers

NHGRI National Human Genome Research Institute, NCI National Cancer Institute, NHLBI National Heart, Lung, & Blood Institute

NICHD National Institute of Child Health & Development

- Genomics & health data

NEI National Eye Institute

- Retinal scans

NIDCR National Institute of Dental & Craniofacial Research

- Facial and dental images

NIBIB National Institute of Biomedical Imaging & Bioingineering

- Imaging & signal data

NIEHS National Institute of Environmental Health Sciences

- Geolocation

NCATS National Center for Advancing Translational Sciences

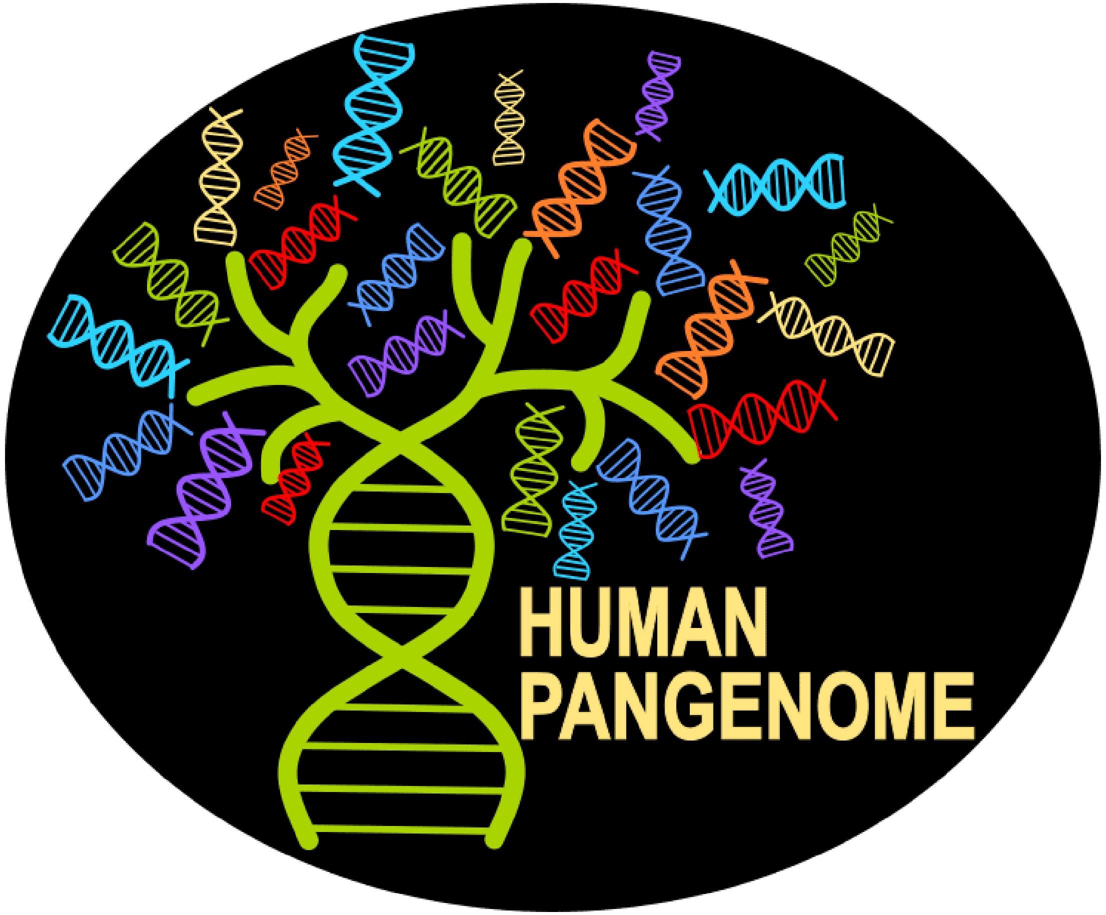
- N3C COVID patient data

HUMAN PANGENOME FAQS

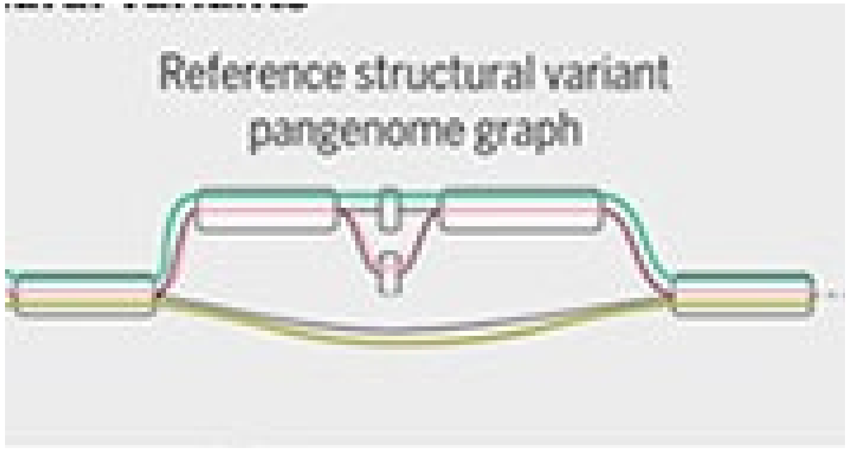
International collaboration at foundation of genomics

- 1’st Human genome $3B
- Now millions at $1000 each
- Simple codes build complexity
  - Genome: 4-letter code (A,C,T,G)
  - Computers: (0,1)
- Need huge numbers to decipher signals and interpret genome data

NHGRI HUMAN PANGENOME REFERENCE



- Pangenome graph replaces linear “single genome” reference
- Represents global human diversity
- Enables population scale analysis
- Graph compresses large number of genomes into compact form
- “Subway map” of human journey



https://humanpangenome.org/

FACIAL RECOGNITION OF GENETIC SYNDROMES

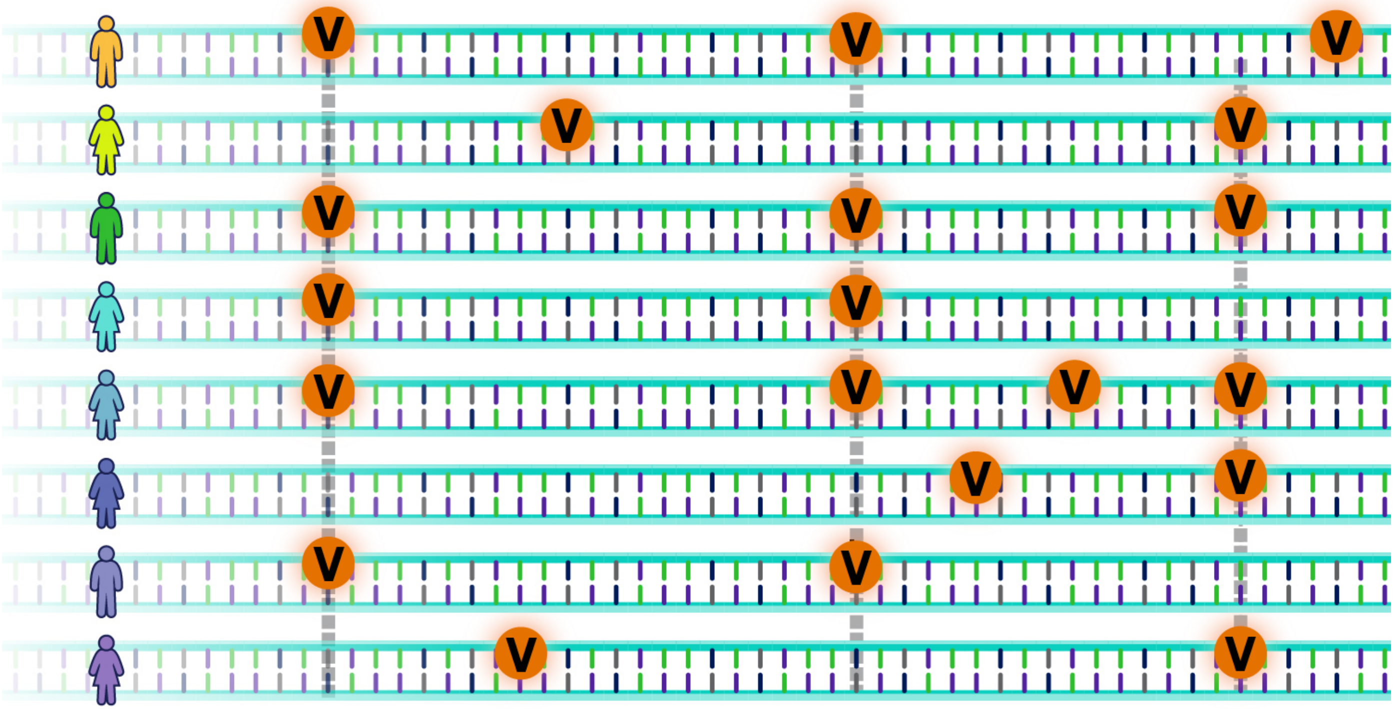
- Many (1000's) genetic syndromes have distinct facial dysmorphisms.
- Sometimes gene defect detected in facial scan of "normal" relative

Hong et al. Genetic syndromes screening by facial recognition technology, *Orphanet J. of Rare Diseases* (2021)

Face2Gene: "…facial recognition software to aid clinical diagnoses of thousands of genetic conditions, such as Sotos syndrome (cerebral gigantism), Kabuki syndrome, intellectual disability, Down syndrome, etc."

POLYGENIC RISK SCORES

NHGRI Consortia

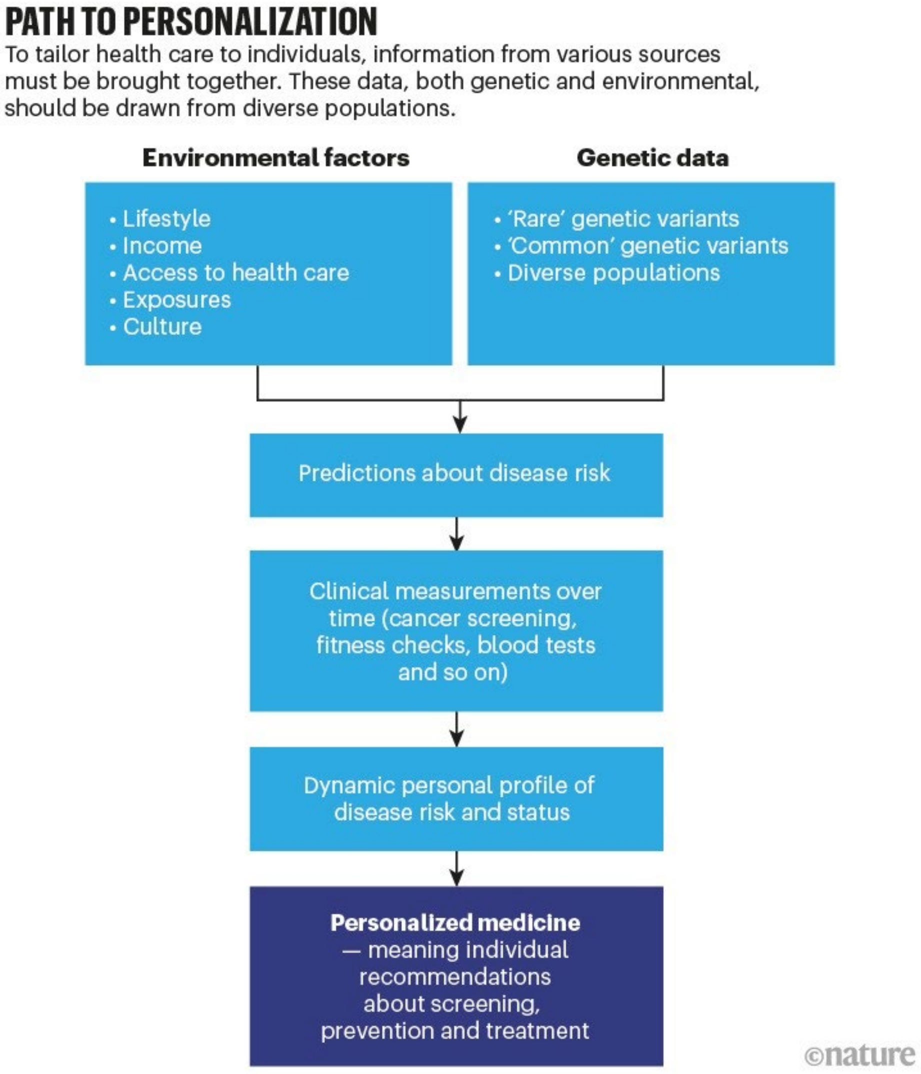


Ali Torkamani et al. “The personal and clinical utility of polygenic risk scores” *Nat. Rev. Genet*. (2018)

GREGoR (previously Mendelian Centers) to find single gene cause of disease (cystic fibrosis, progeria, etc)

PRIMED to study polygenic (“complex”) diseases using advances in Polygenic Risk Scores (diabetes, heart disease, autism, etc.)

PRECISION HEALTH POWERED BY GENOMICS



PATH TO PERSONALIZATION

To tailor health care to individuals, information from various sources must be brought together. These data, both genetic and environmental, should be drawn from diverse populations.

Environmental factors

- Lifestyle
- Income
- Access to health care
- Exposures
- Culture

Genetic data

- 'Rare' genetic variants
- 'Common' genetic variants
- Diverse populations

Predictions about disease risk

Clinical measurements over time (cancer screening, fitness checks, blood tests and so on)

Dynamic personal profile of disease risk and status

Personalized medicine — meaning individual recommendations about screening, prevention and treatment

©nature

Mark McCarthy & Ewan Birney, "Personalized profiles for disease risk must capture all facets of health" *Nature* 597, 175-177 (2021)

- Polygenic & Pangenome with diverse genomes to represent the human family
- Environmental, lifestyle, income, access to health care, exposures, culture
- "This will inevitably bring the realms of research and clinical care together, and will require us to address fundamental questions about data ownership, privacy, equality of access, fairness and social responsibility. Global efforts to create such standards are in place, for example the Global Alliance for Genomics and Health."

GLOBAL DATA STANDARDS

Global Alliance
for Genomics & Health

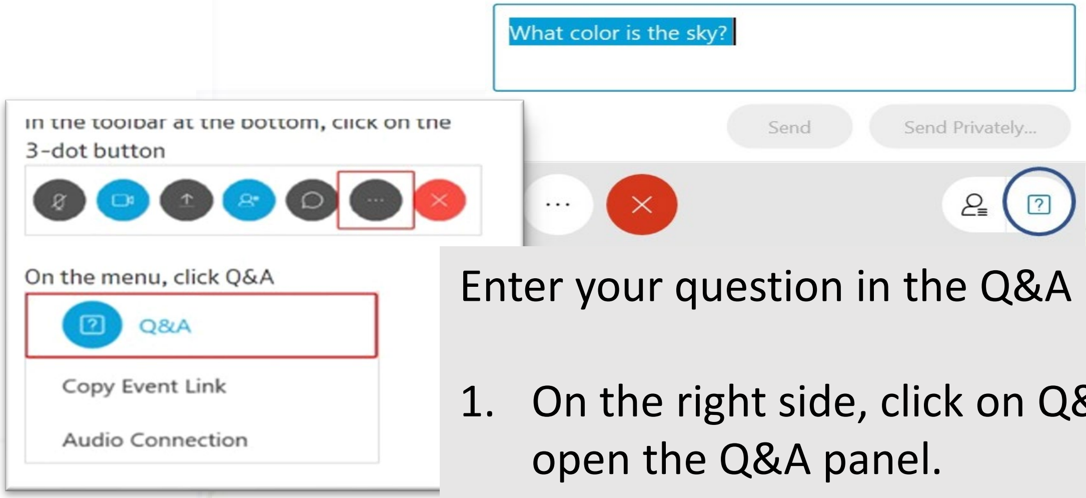
Collaborate. Innovate. Accelerate.

- GA4GH is international collaboration on standards for genomics and health data with human rights foundation
- Many social and technical “onramps” for inclusion and adoption
- Example: GA4GH Passports & Visas
- *Cell Genomics* special issue Nov 2021

GA4GH: https://www.ga4gh.org/

Current and Future Genomic Data Use Challenges

Moderated Questions and Answers



In the toolbar at the bottom, click on the 3-dot button

On the menu, click Q&A

What color is the sky?

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

National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

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

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

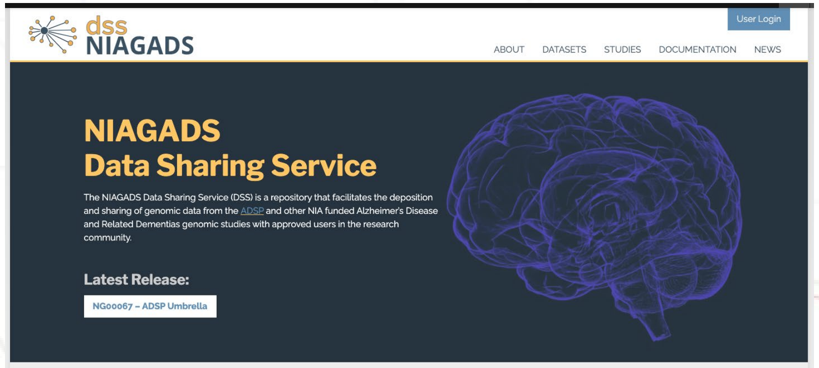
NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Lessons Learned from NIAGADS and ADSP Data Sharing

Wan-Ping Lee & Li-San Wang
University of Pennsylvania Perelman School of Medicine

Alzheimer's Disease Sequencing Project (ADSP) and NIAGADS

- 17k complete genomes released in 2021
- 36k (2022), 75k (2023) planned
- Data releases are managed by NIAGADS
  - National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site at University of Pennsylvania

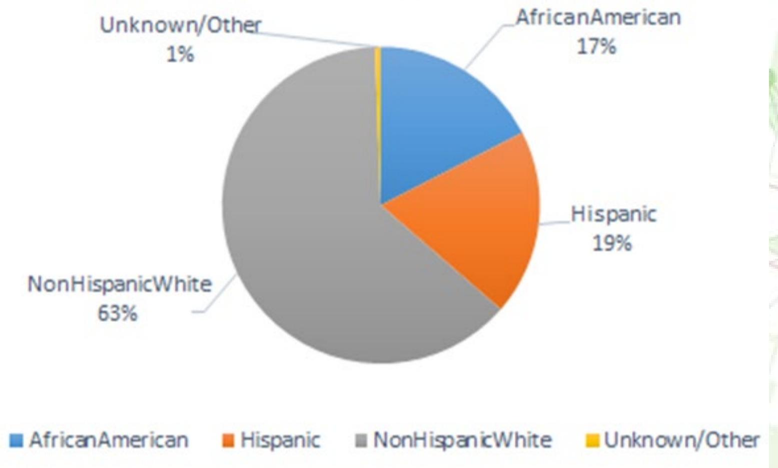


NIAGADS
Data Sharing Service

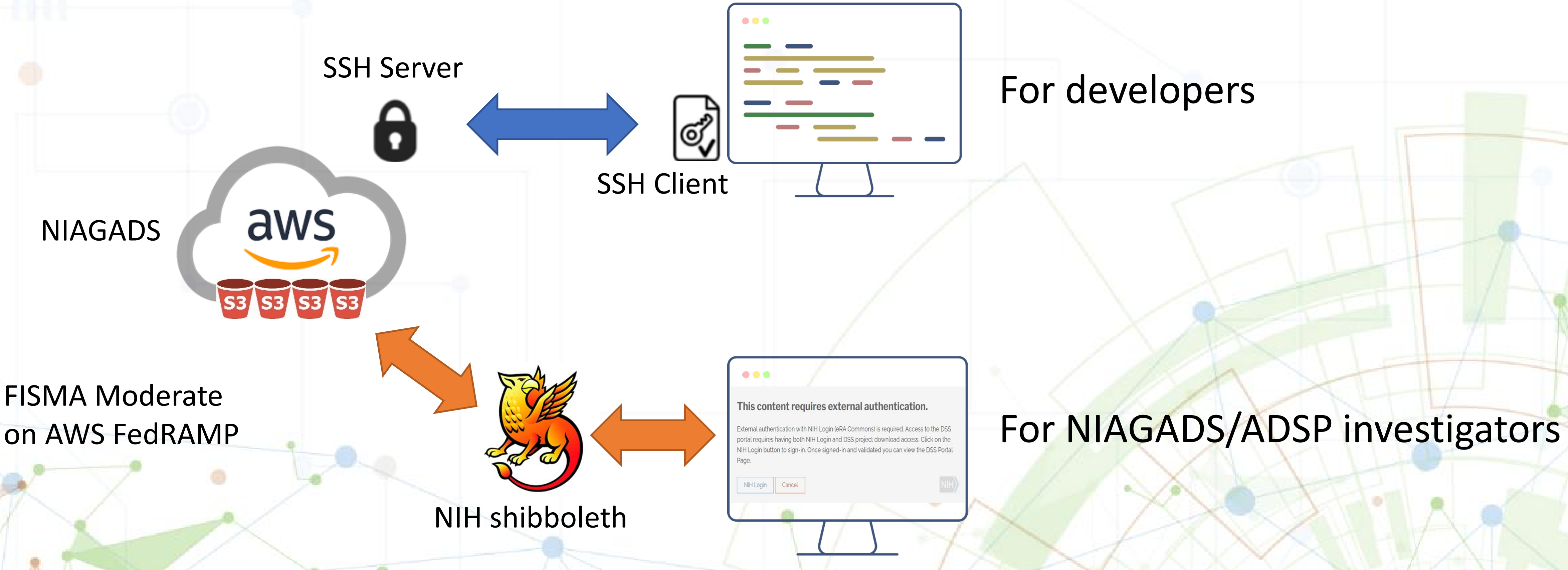
The NIAGADS Data Sharing Service (DSS) is a repository that facilitates the deposition and sharing of genomic data from the ADSP and other NIA funded Alzheimer's Disease and Related Dementias genomic studies with approved users in the research community.

Latest Release:

NG00067 – ADSP Umbrella


R3 ADSP WGS- 16,906 genomes

NIAGADS DSS Security Setup



SSH Server

SSH Client

For developers

NIAGADS

FISMA Moderate
on AWS FedRAMP

NIH shibboleth

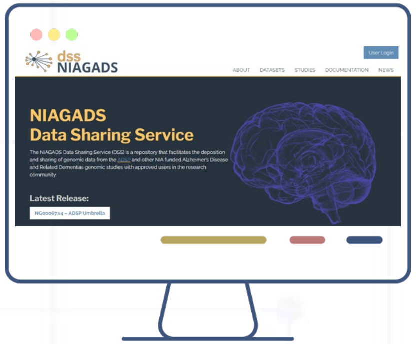
This content requires external authentication.

External authentication with NIH Login (eRA Commons) is required. Access to the DSS portal requires having both NIH Login and DSS project download access. Click on the NIH Login button to sign-in. Once signed-in and validated you can view the DSS Portal Page.

For NIAGADS/ADSP investigators

Authentication of data requesters

Step 1: https://dss.niagads.org



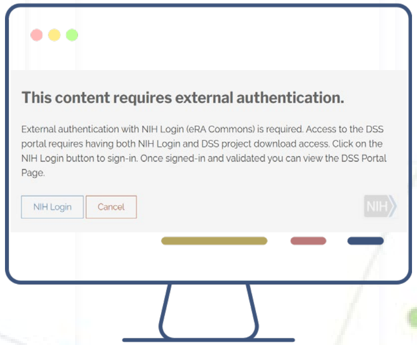
NIAGADS Data Sharing Service

The NIAGADS Data Sharing Service (DSS) is a repository that facilitates the deposition and sharing of genomic data from the ADSP and other NIA funded Alzheimer's Disease and Related Dementias genomic studies with approved users in the research community.

Latest Release:

NG00067.v4 – ADSP Umbrella

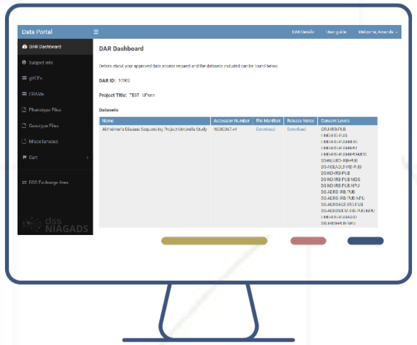
Step 2: NIH eRA Authentication



This content requires external authentication.

External authentication with NIH Login (eRA Commons) is required. Access to the DSS portal requires having both NIH Login and DSS project download access. Click on the NIH Login button to sign-in. Once signed-in and validated you can view the DSS Portal Page.

Step 3: Data manifest browsing on NIAGADS DSS



DAR Dashboard

Details about your approved data access request and the datasets included can be found below.

DAR ID: 10002

Project Title: TEST UPenn

Datasets

Name	Accession Number	File Manifest	Release Notes	Consent Levels
Alzheimer's Disease Sequencing Project Umbrella Study	NG00067.v4	Download	Download	GRU-IRB-PUB HMB-IRB-PUB HMB-IRB-PUB-MDS HMB-IRB-PUB-NPU HMB-IRB-PUB-NPU-MDS DS-NEURO-IRB-PUB DS-AGEADLT-IRB-PUB DS-ND-IRB-PUB DS ND IRB PUB MDS DS ND IRB PUB NPU DS ADRD IRB PUB DS-ADRD-IRB-PUB-NPU DS-ADRDAGE-IRB-PUB DS-ADRDMEM-IRB-PUB-NPU HMB-IRB-PUB-GSO GRU-IRB-PUB-NPU

Step4: Data accessing

Learning curve for FISMA is steep

- Scope
- Amount of work
  - 10 months preparation
  - 1 year for 3 times external assessment
- Regulations to be met
- Cost benefit analysis

Respecting Informed Consent

- NIH Genome Sharing Policy
- Institutional certification to capture informed consent conditions
- Data access committee
- How to split data based on informed consent

De-identification of genomic data

- Genome sequencing is identifiable
- What about functional genomics data? Theoretically RNA-Seq is identifiable because it carries variants. What other types of sequencing data?
- What does it take to de-identify data?

What will be helpful if we do it again

- Guidelines for FISMA requirements that are specific to human genome data
- Tutorials and FAQs on how to set up a FISMA compliant framework: timeline, amount of work, cost (budget and staff)

Workshop Close Out

Ron Pulivarti, NIST

Thank you for joining!

Contribute to the conversation

Email genomic_cybersecurity_nccoe@nist.gov