

National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

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

Session 1: Cybersecurity Challenges Affecting Genomic Sequencing

Charles Fracchia (BioBright)
Phillip Whitlow (HudsonAlpha)

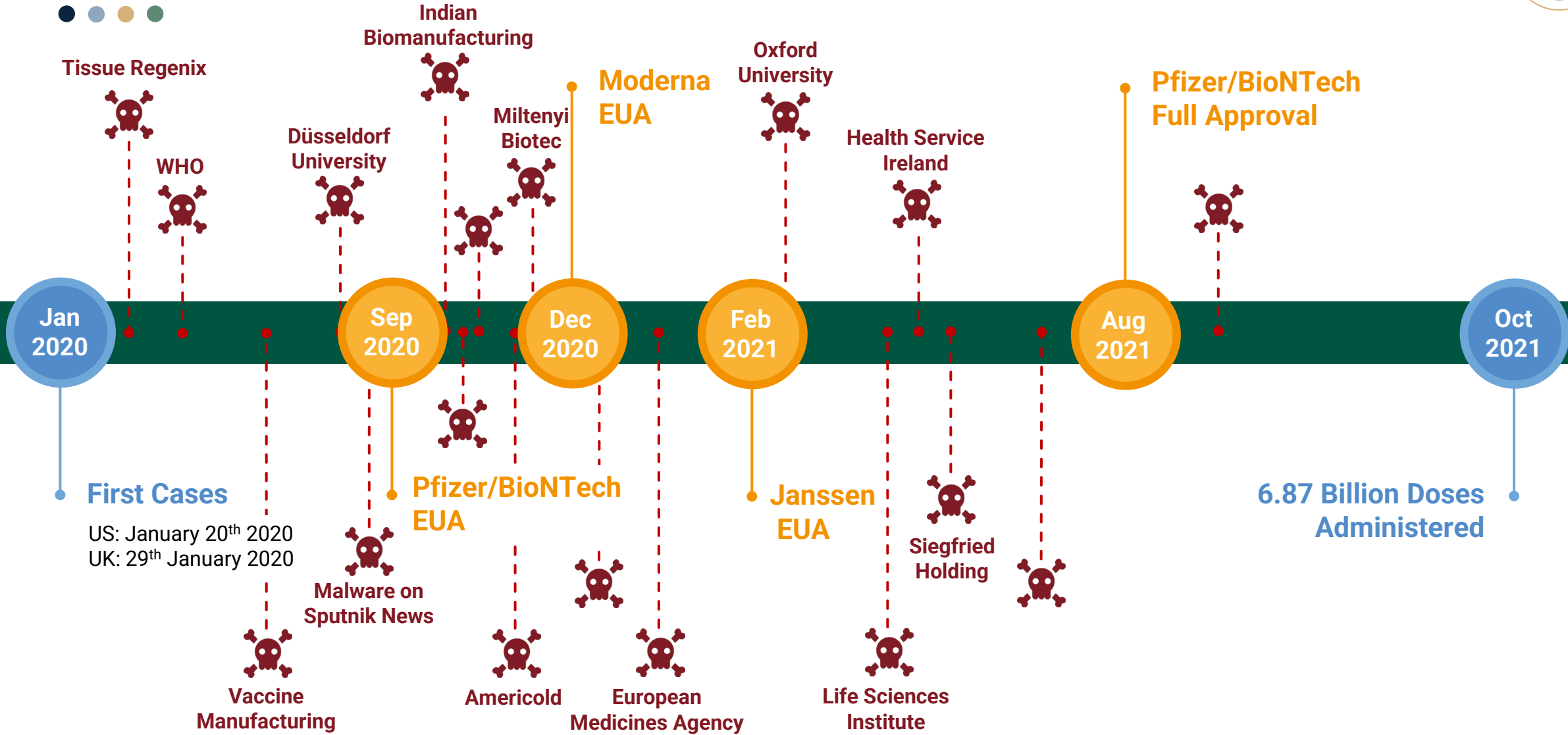
Digital Biosecurity in a Modern Context

Charles Fracchia <charlesfracchia@gmail.com>

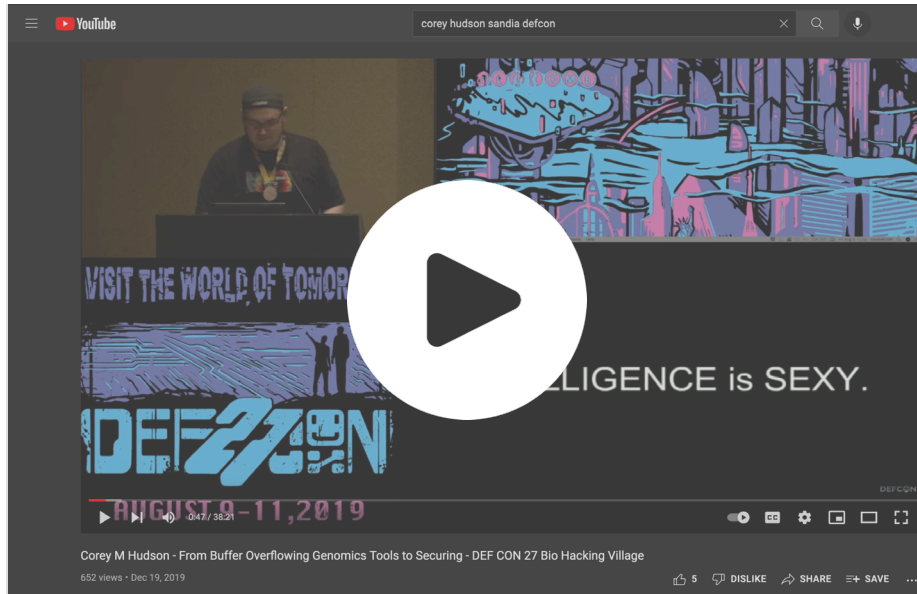
PGP: 4A06 3D3A B157 C3DE C31C 91B0 6E76 3F6A DA35 06C4

CEO BioBright / VP Data Dotmatics

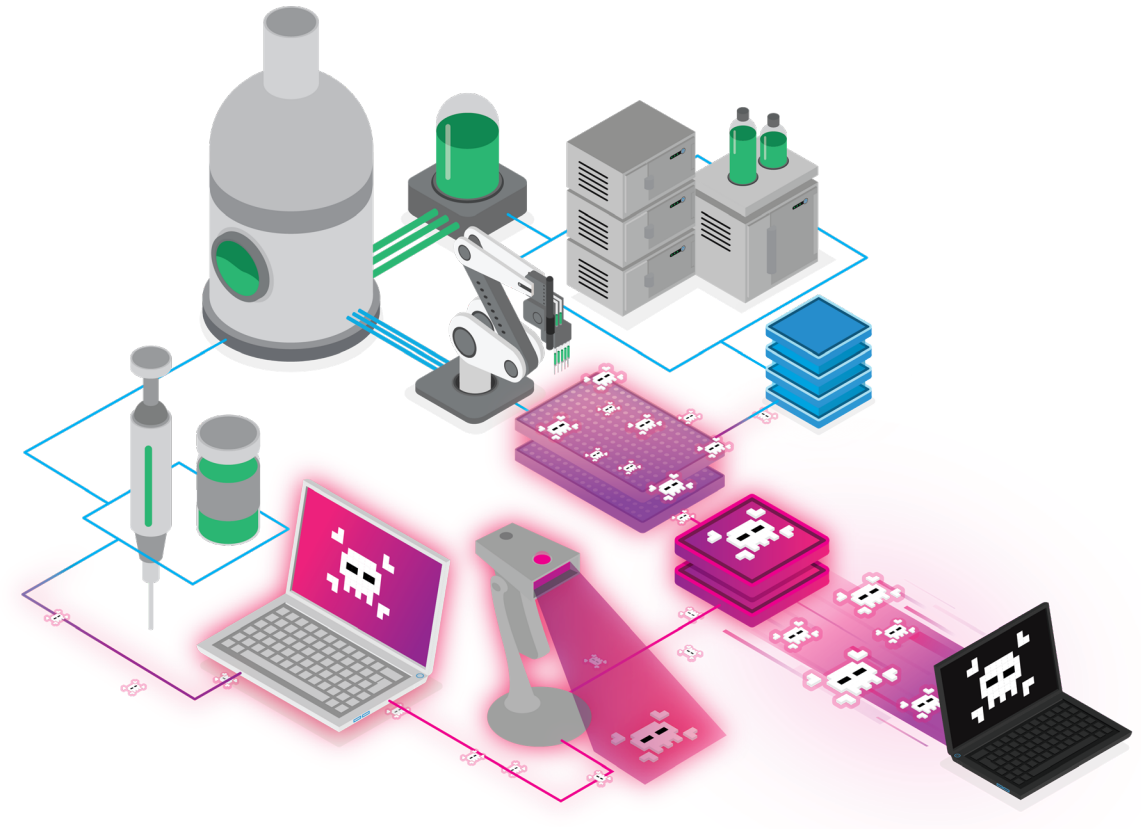
PUBLIC DOMAIN ATTACKS ON BIOINFRASTRUCTURE



WE CANNOT FIX THE ISSUE WITHOUT TACKLING ... DEVICE INSECURITY



<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: intserres644.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: EB2A1A4A094A88BE07 EBA96258BFEE2E95C7FC

contributed by: **BioBright**
a dynamilis company

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


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



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



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



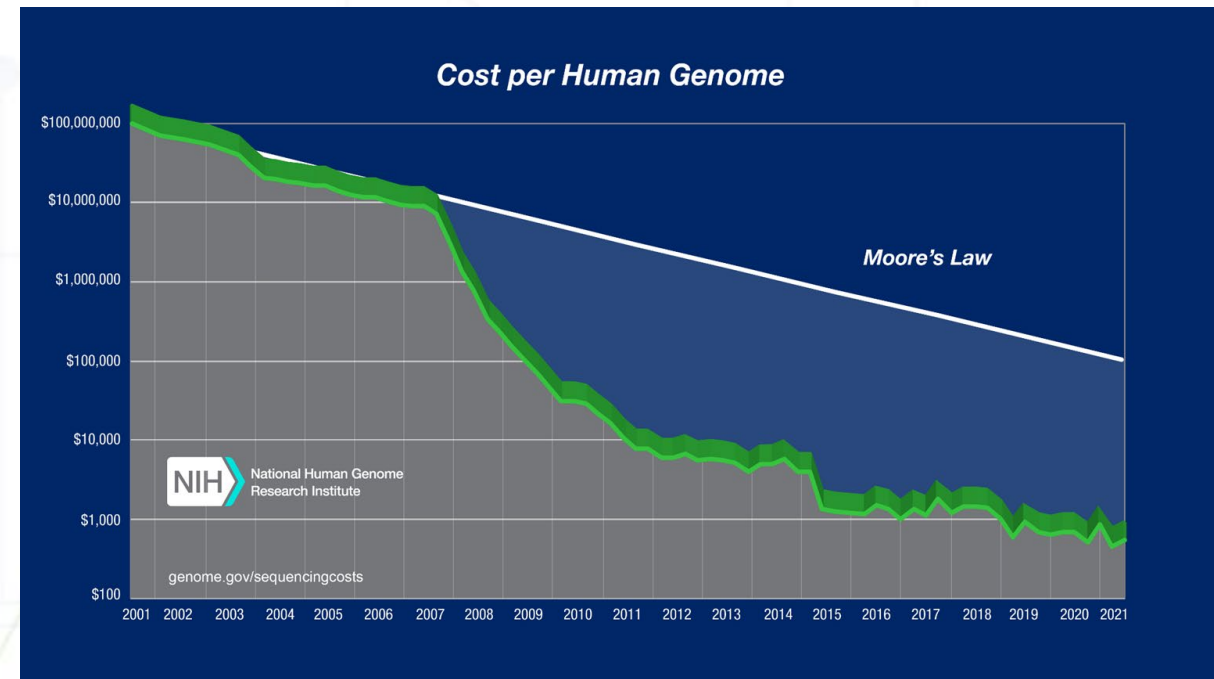
Cybersecurity Challenges

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 Cybersecurity Center of Excellence

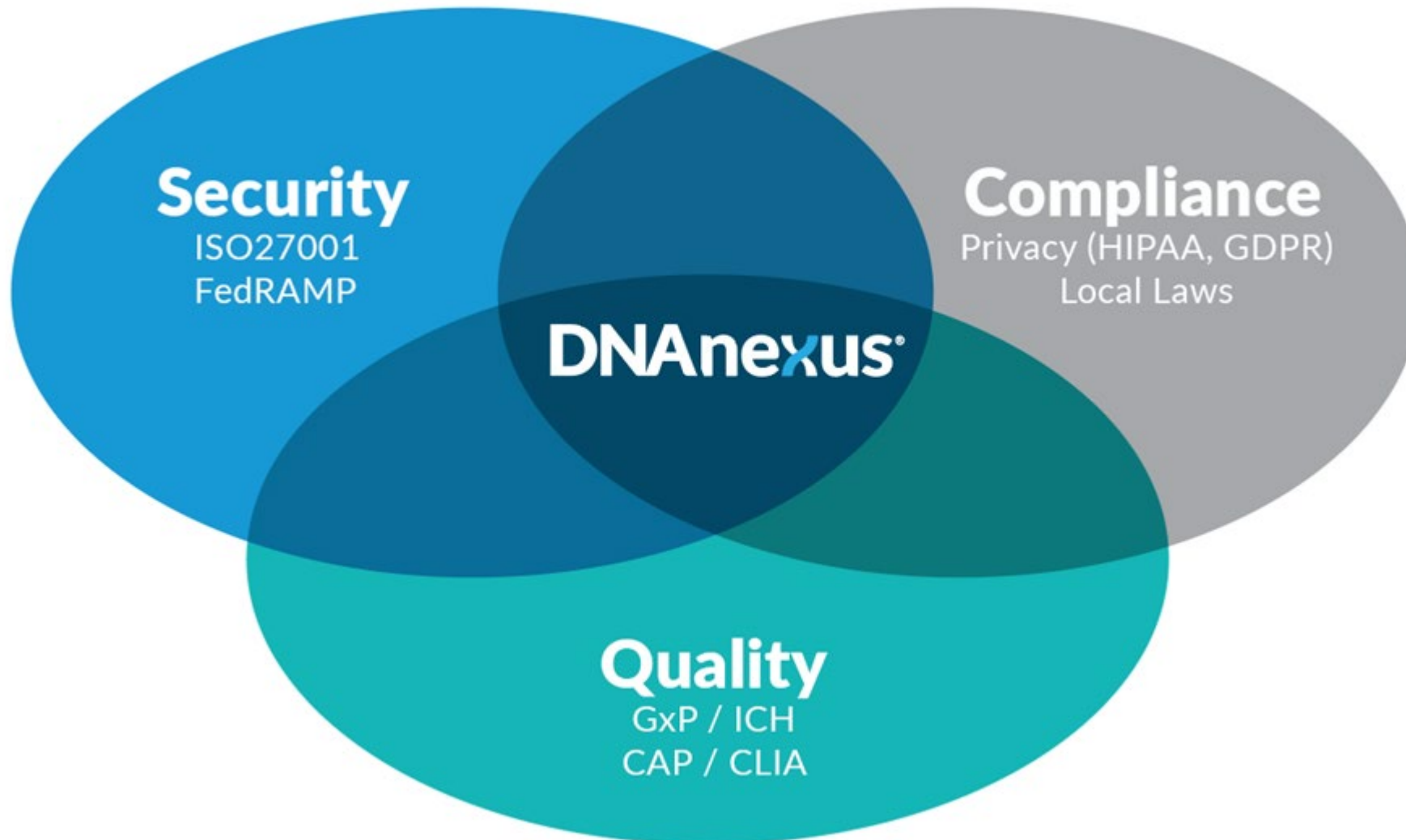
Session 2: Cybersecurity Challenges for Genomic Software

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

E. Loren Buhle

VP of Security, Quality and Compliance at DNAnexus

CONSIDER A HOLISTIC APPROACH



SECURITY REQUIREMENTS

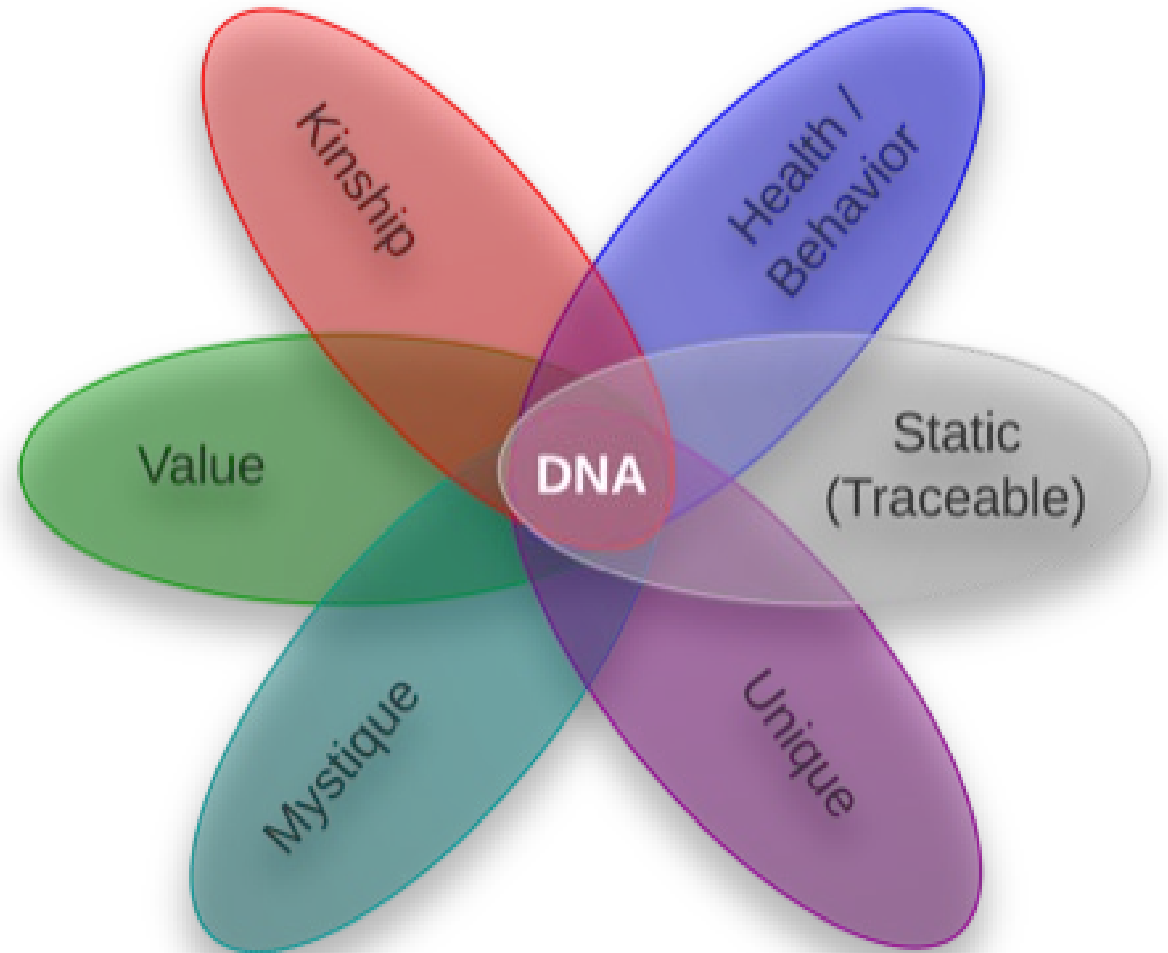


- Architecture
 - Consider zero-trust architecture
- Robust and auditable authentication and authorization
- Continuous Monitoring
 - (24x7x365) with a high degree of automation
 - Automate the Incident Response Procedure as much as possible
- Accountability to industry standards:
 - FedRAMP “Moderate,” “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



- **Confidentiality, 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
- **Governance**
 - Establishing the appropriate authorizations on code/data
 - What must be retained? Where? For how long?
- **Provenance**
 - Privacy concerns, alignment with consent, ability to track and prove compliance
 - Metadata – source of sequence, sequencer, processing steps
- **Quality**
 - 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)

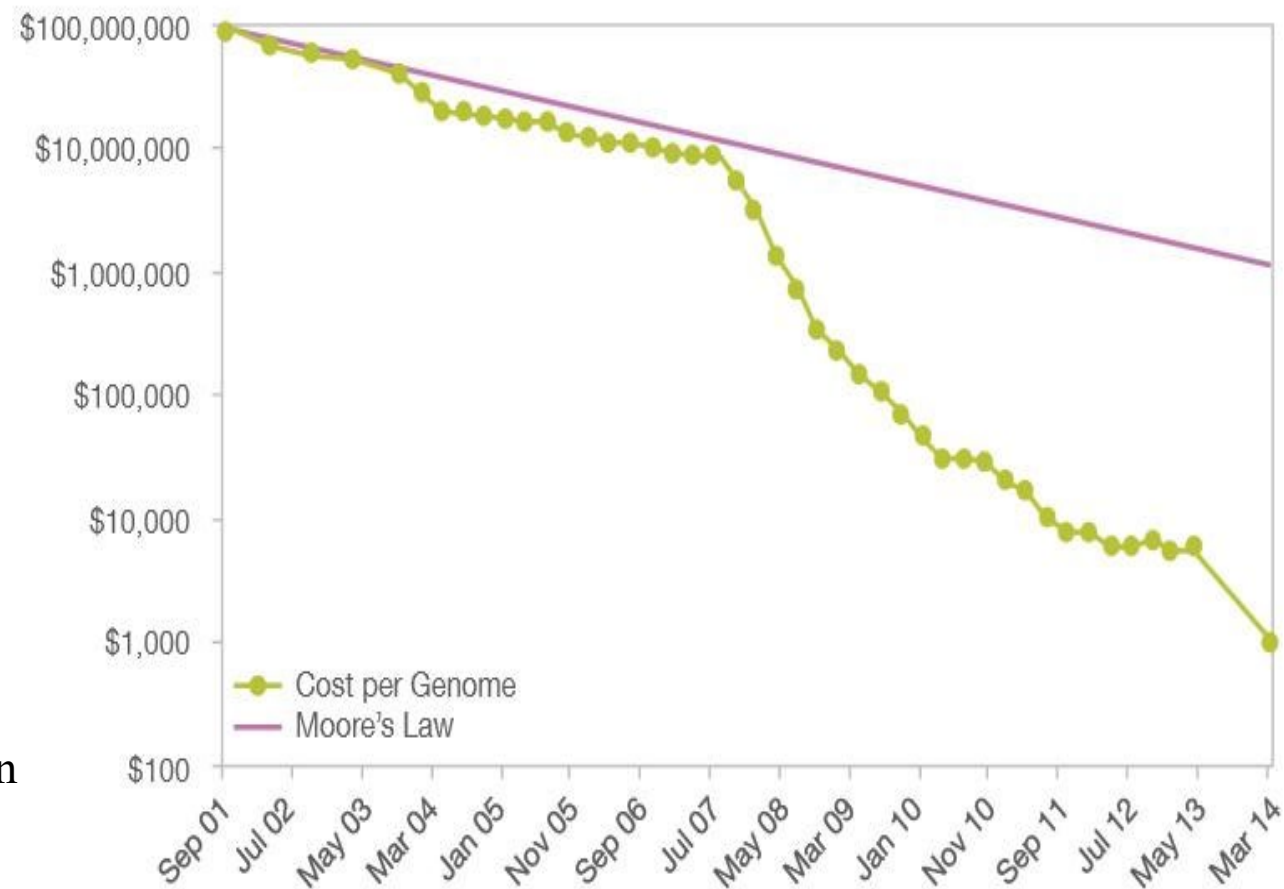
Privacy in the Genomic Era

XiaoFeng Wang, IEEE Fellow, Rudy Professor at IUB

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

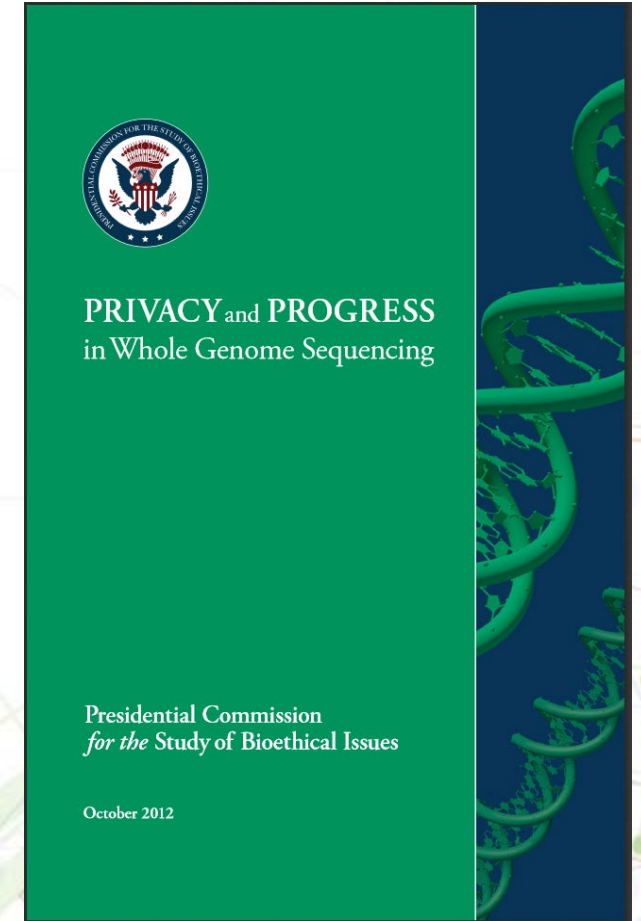
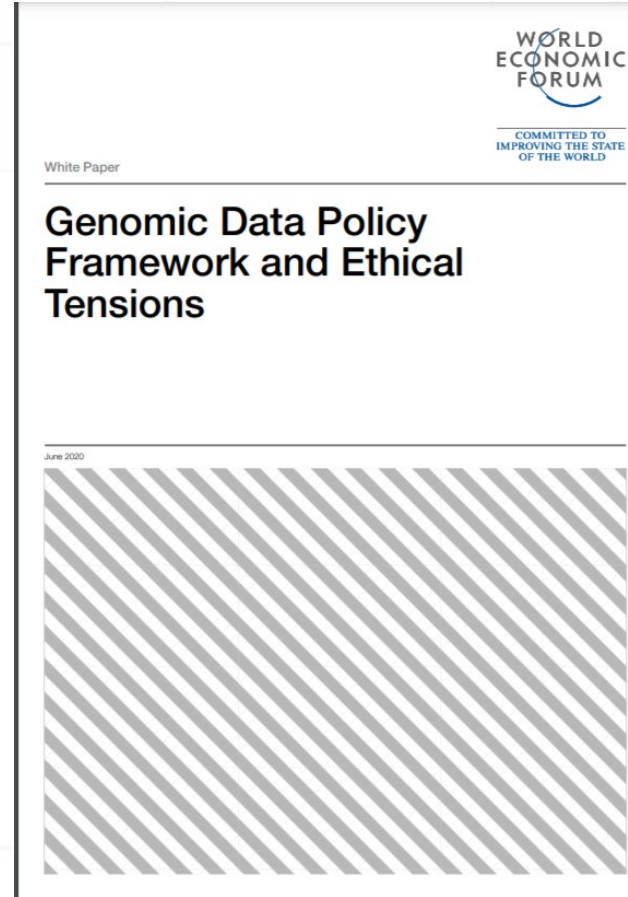
Genomic Revolution

- 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



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?



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

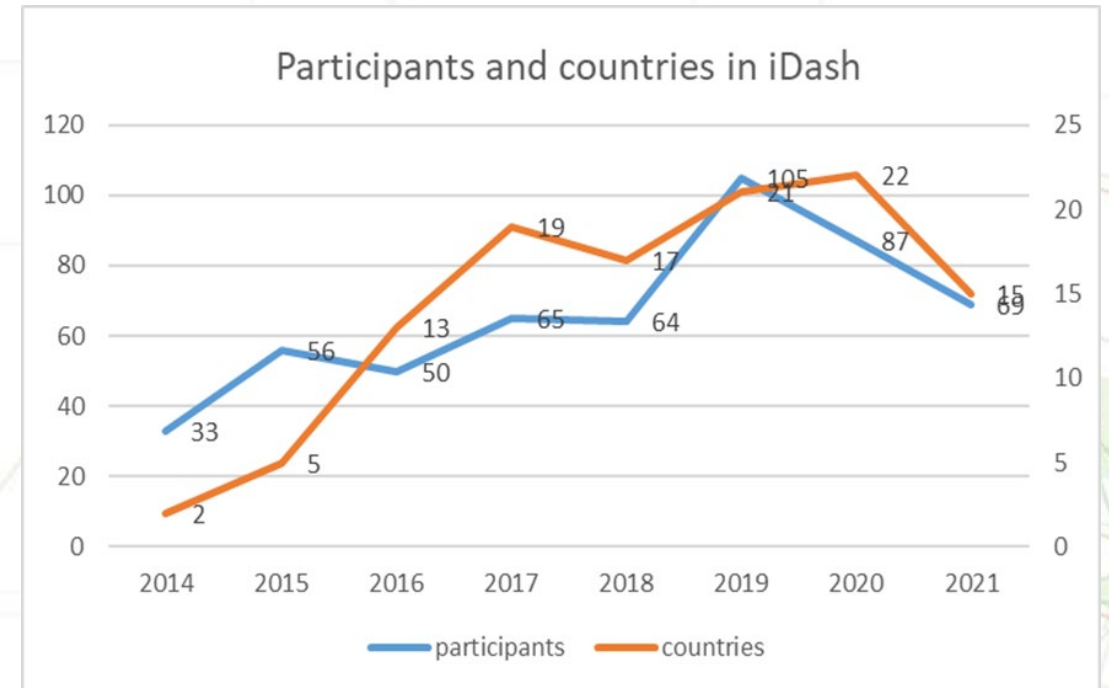
Demonstrate the feasibility of secure genome analysis and dissemination using DP, MPC, HE, TEE

Reported in the media (e.g., Nature News)

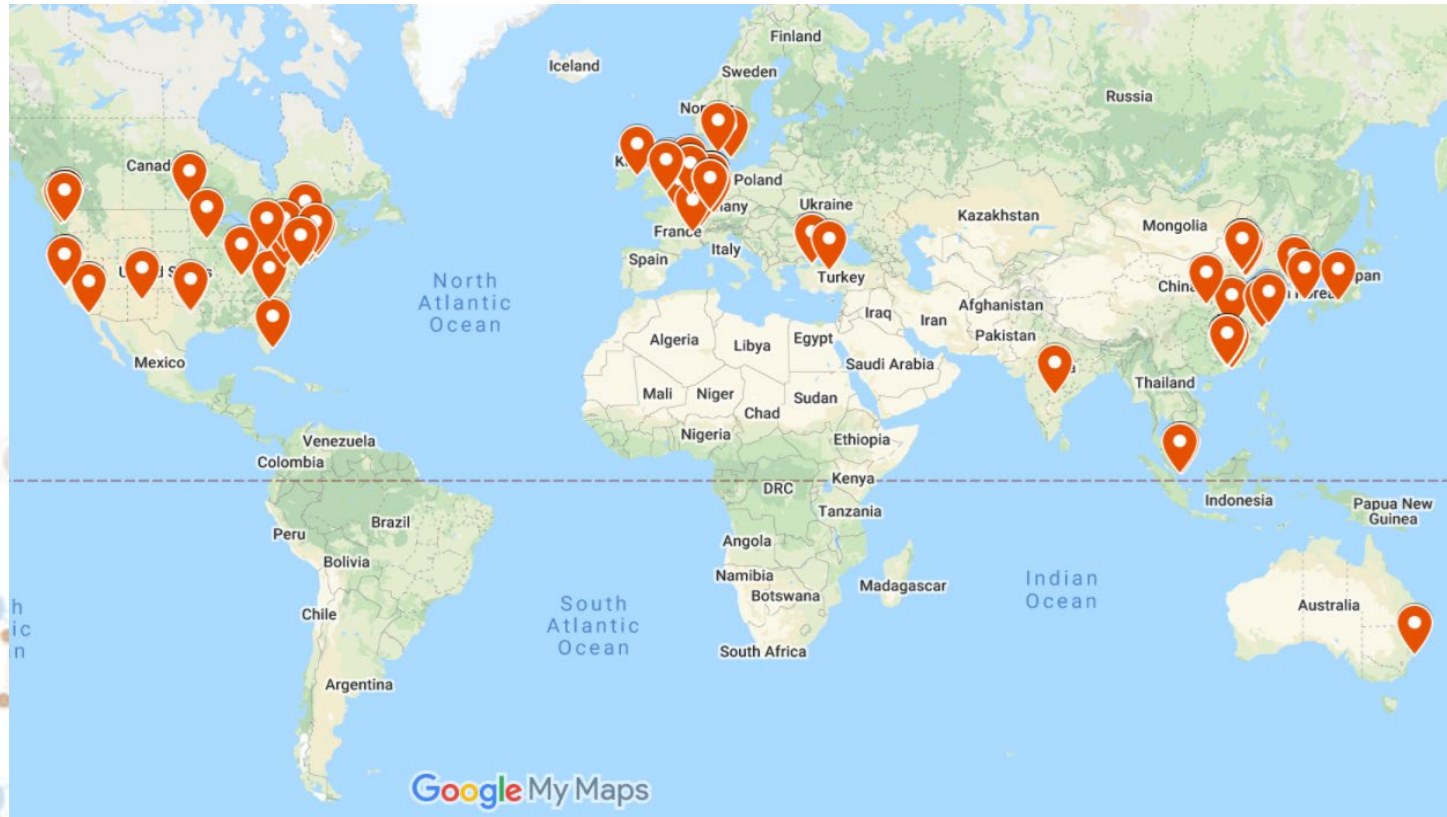


Topics and Trend from 2014 to 2021

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

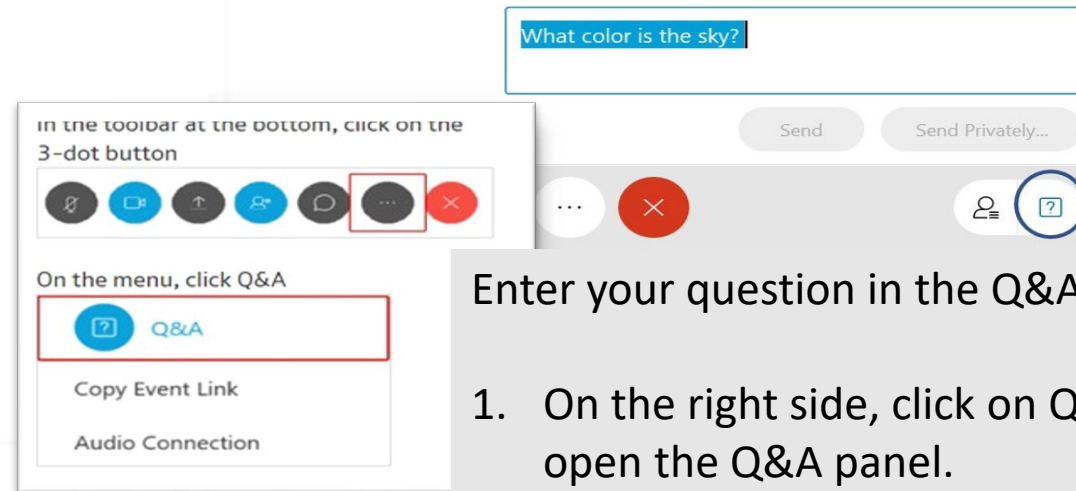
- NIH R01HG010798: “Secure and Privacy-preserving Genome-wide and Phenome-wide Association Studies via Intel Software Guard Extensions (SGX)”
- NIH R01HG007078: “Privacy Preserving Technologies for Human Genome Data Analysis and Dissemination”
- NSF-CNS-1408874: “Broker Leads for Privacy-Preserving Discovery 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



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 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

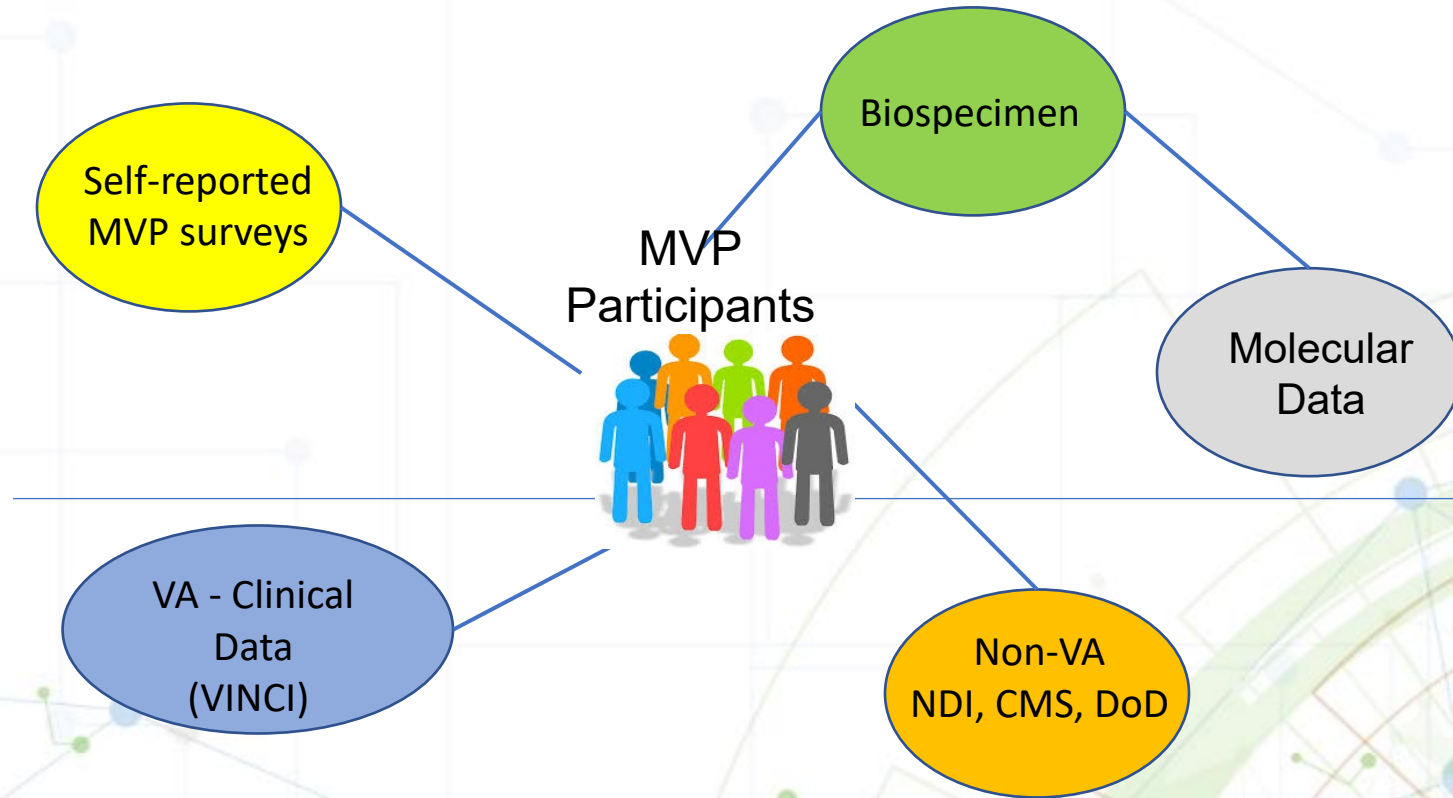
Director, MVP

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

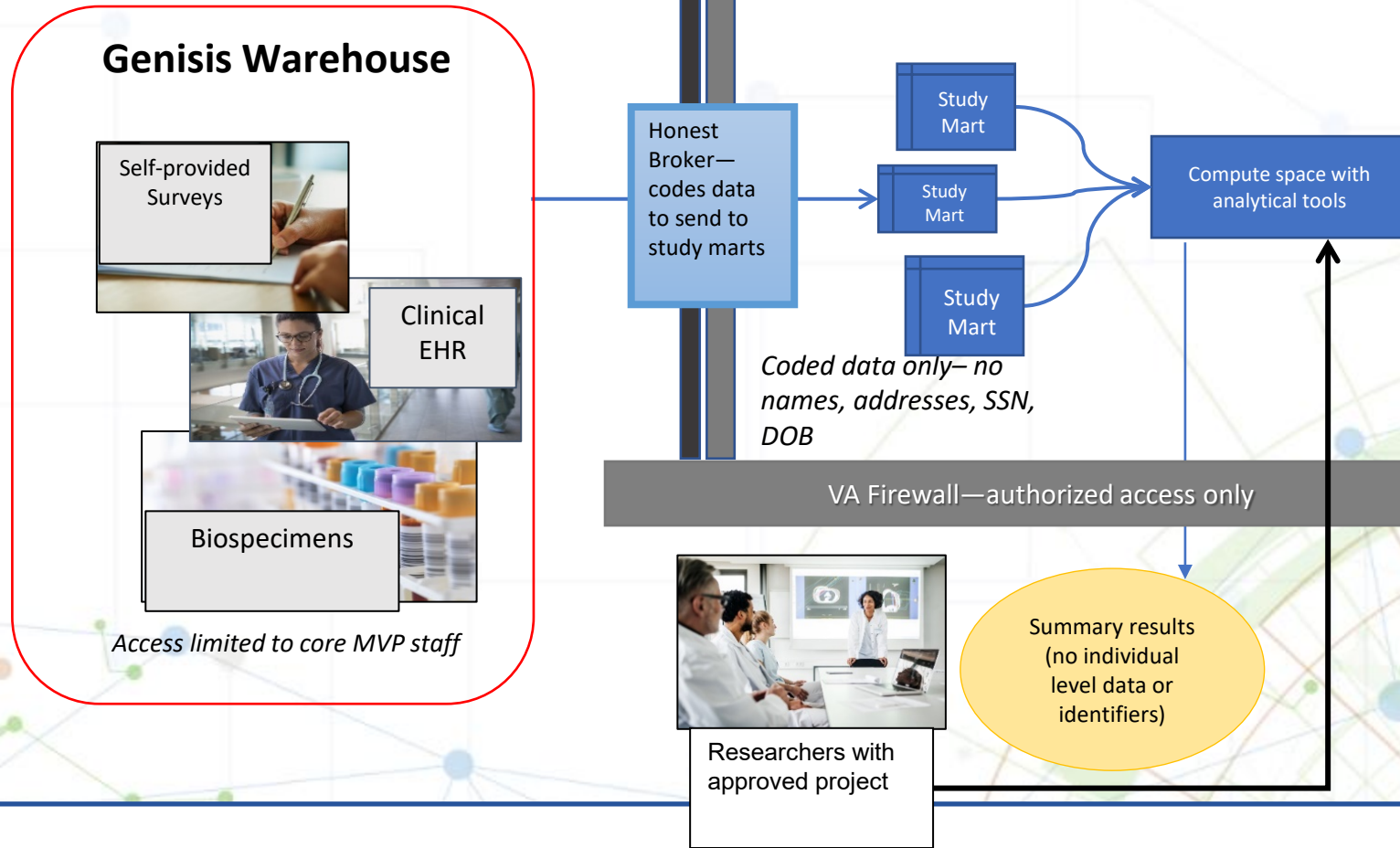
GOAL: Generate the maximum amount of data from biospecimens to enhance scientific discovery

- 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



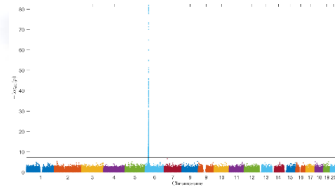
External access to MVP Data

- 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)

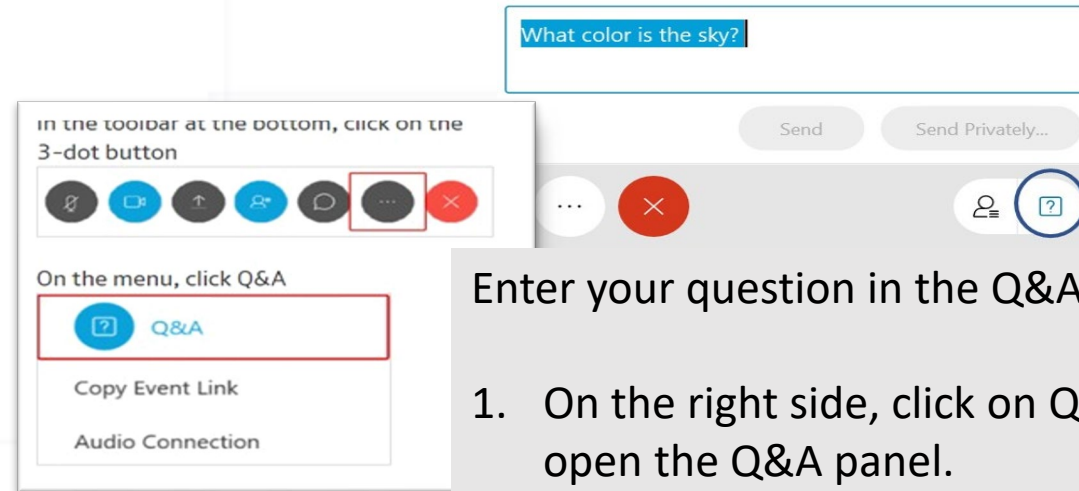
MVP Risk Mitigation Strategies

- 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!

Privacy Challenges for Genomic Data

Moderated Questions and Answers



Enter your question in the Q&A panel.

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

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

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*



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. While each

Success 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

Success Stories

Human Genetics 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

Success Stories

Human Genetics and 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

Success Stories

Human Genetics and Genomics Research



NEUROGENETICS

Federally funded basic research is driving progress toward understanding the genes

Data-sharing Fuels Progress in Human Genetics & Genomics Research

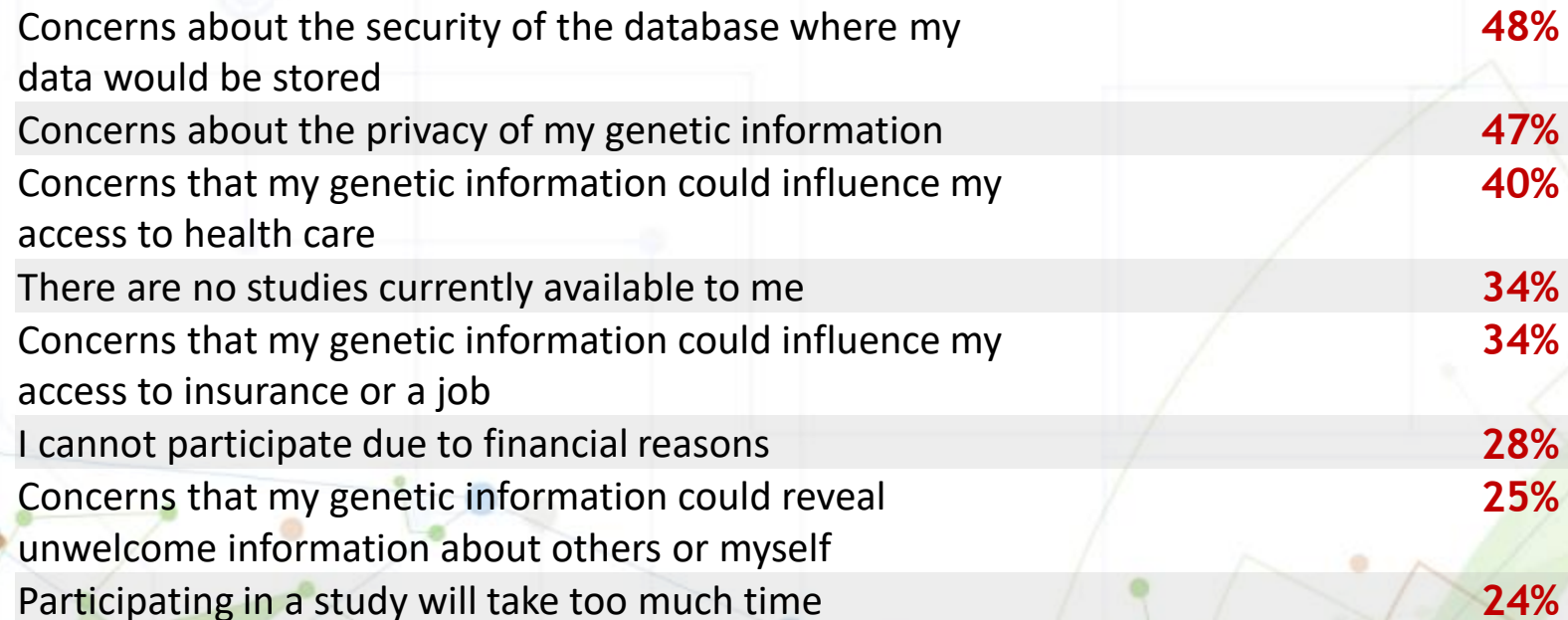
- Broad data-sharing a hallmark of the human genetics community
 - Essential for completion of Human Genome Project
- Data-sharing fundamental for continued advances in research & medicine

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)



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

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."

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

"Any additional protections or restrictions ... should be commensurate with the actual risk."

Association of American Medical Colleges

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

Thank you!

Contact: gjarvik@medicine.washington.edu

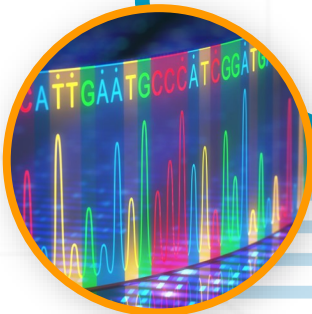
AWS for Genomics

Solving Challenges in Genomic Data Sharing

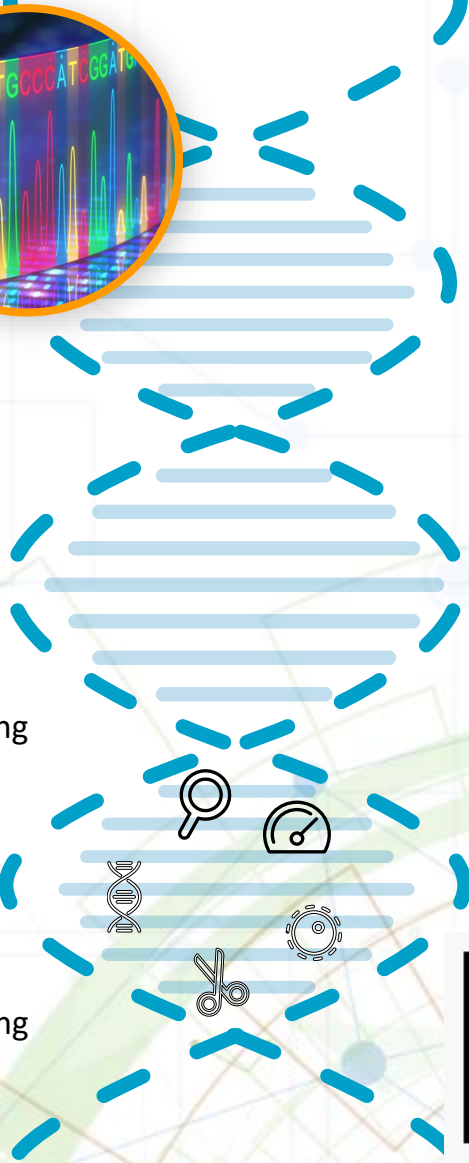
Ankit Malhotra, Ph.D.
Genomics Lead,
AWS Worldwide Public Sector Health

The precision medicine revolution

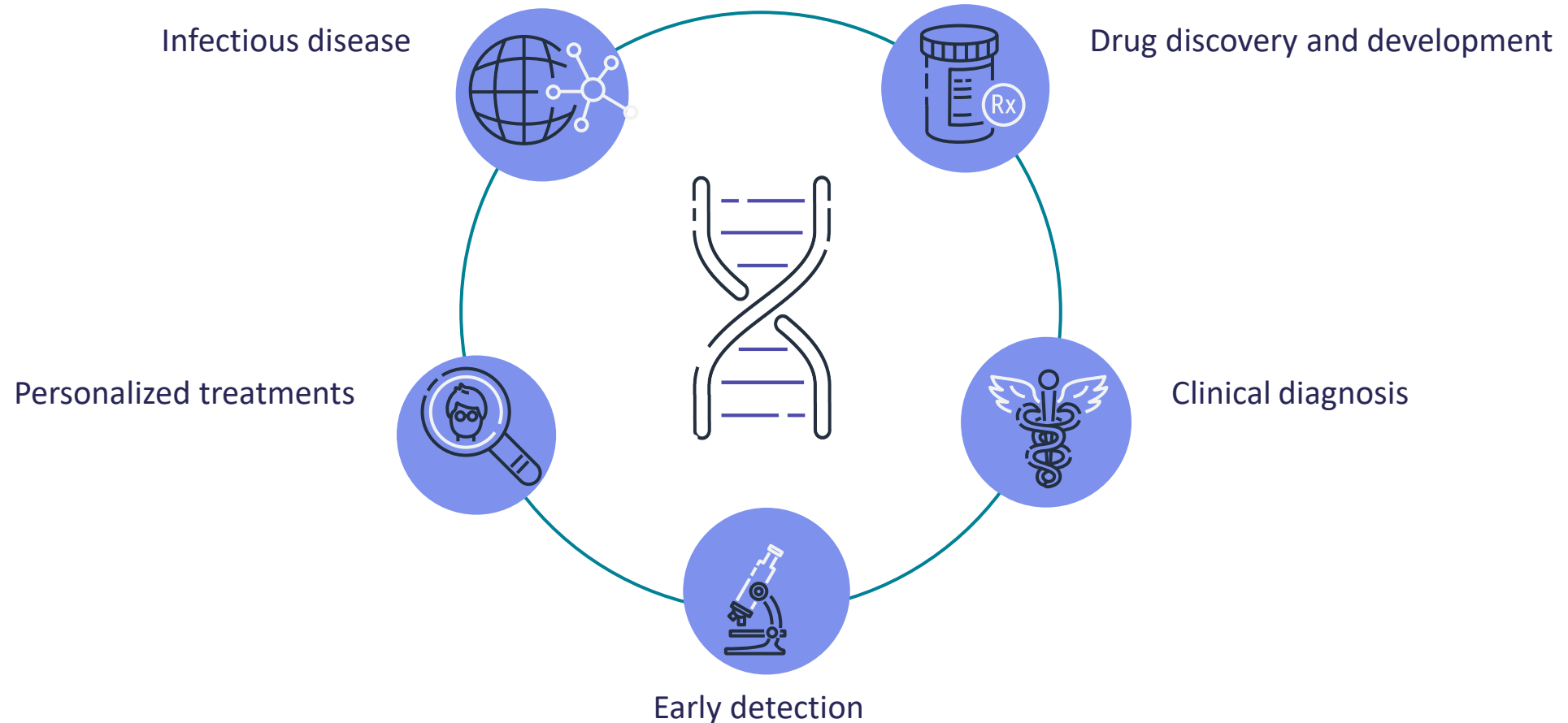
Transformative technologies in sequencing and computing is driving innovation across healthcare and enabling precision medicine.



- Clinical Genome Sequencing
- Genetic Risk Scores
- Targeted Therapeutics
- Induced Pluripotent Stem Cells
- CRISPR Genome Editing



Genomics—a catalyst for personalized health



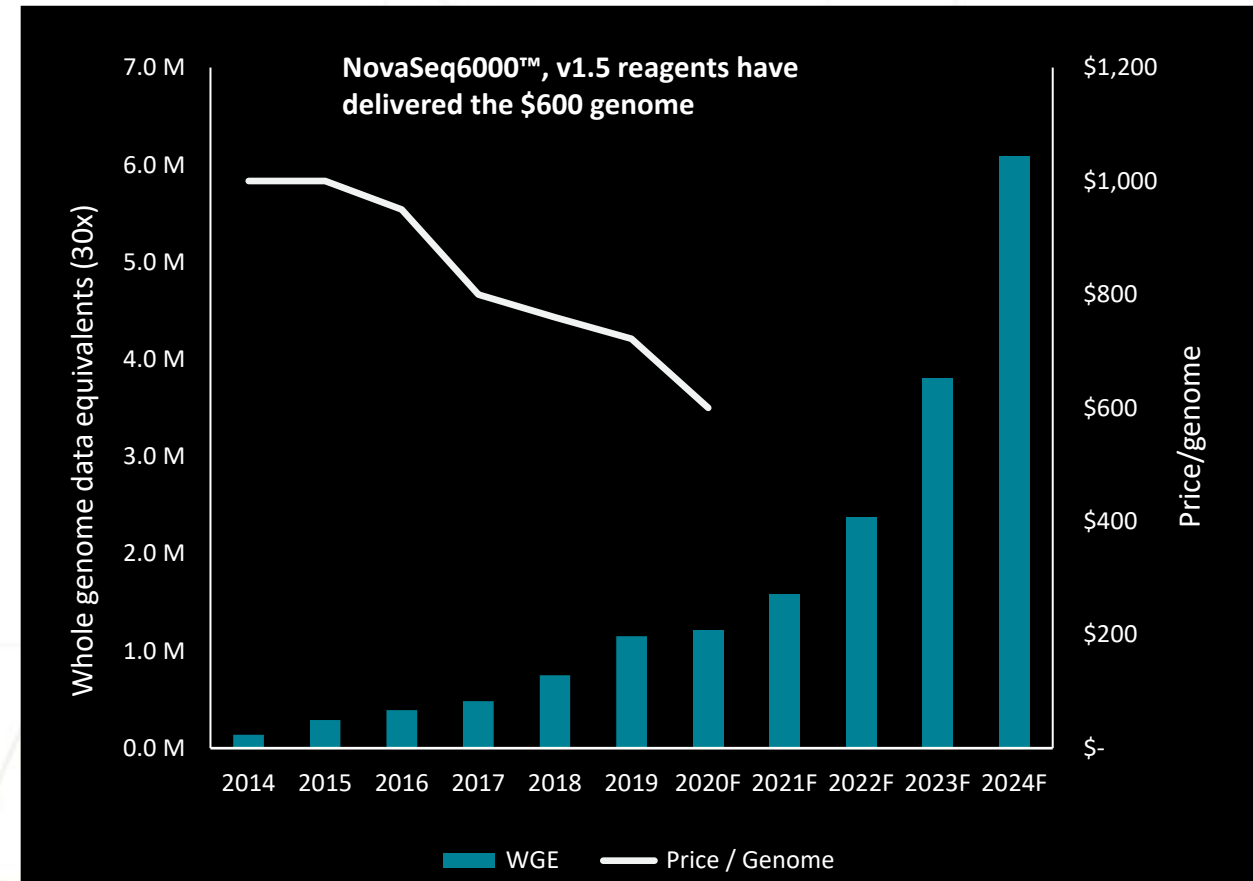
Challenges in leveraging genomics data

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



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

AWS Modern Genomics Data Platform






AWS Genomics CLI



SageMaker, HealthLake, Sequera Labs, Cromwell

BUILD

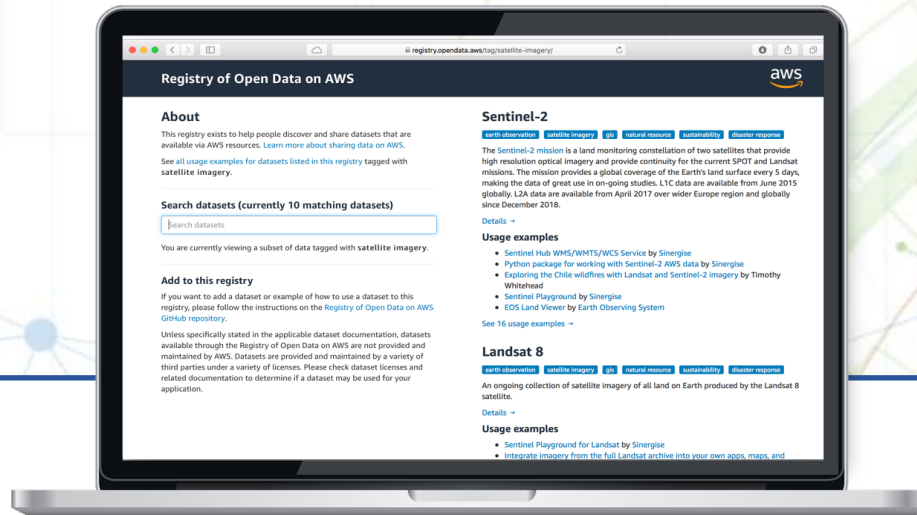






BUY

AWS Open Data Program



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

- 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

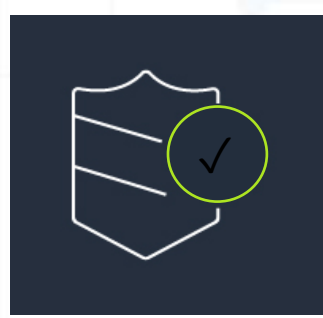
Open Access to Top Genomics Datasets

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....



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

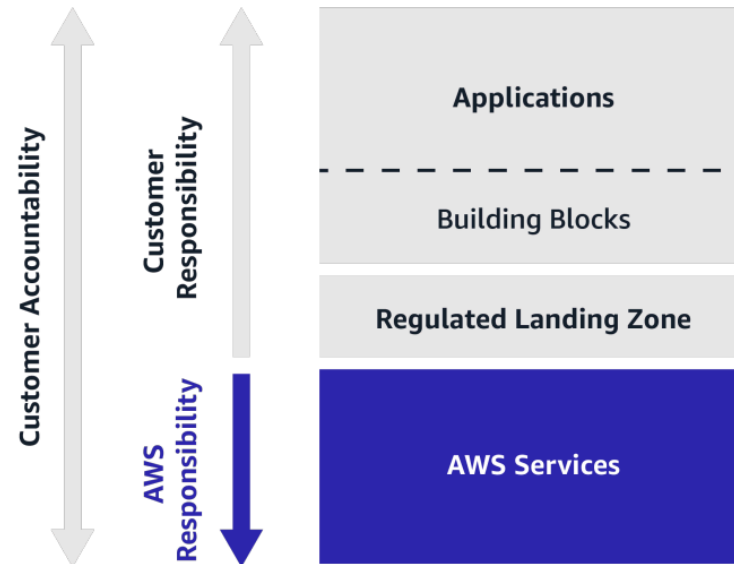
Responsibility for security **"IN" the cloud**

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






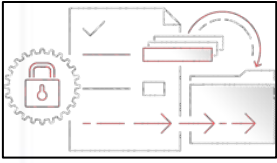
AWS

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
<ul style="list-style-type: none"> AWS Identity and Access Management (IAM) AWS Single Sign-On AWS Organizations AWS Directory Service Amazon Cognito AWS Resource Access Manager 	<ul style="list-style-type: none"> AWS Security Hub Amazon GuardDuty Amazon Inspector Amazon CloudWatch AWS Config AWS CloudTrail VPC Flow Logs AWS IoT Device Defender 	<ul style="list-style-type: none"> AWS Firewall Manager AWS Network Firewall AWS Shield AWS WAF – Web application firewall Amazon Virtual Private Cloud AWS PrivateLink AWS Systems Manager 	<ul style="list-style-type: none"> Amazon Macie AWS Key Management Service (KMS) AWS CloudHSM AWS Certificate Manager AWS Secrets Manager AWS VPN Server-Side Encryption 	<ul style="list-style-type: none"> Amazon Detective Amazon EventBridge AWS Backup AWS Security Hub CloudEndure Disaster Recovery 	<ul style="list-style-type: none"> AWS Artifact AWS Audit Manager



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 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/>

Hong Kong Genome Project (LifeBit)

<https://lifebit.ai/blog/lifebit-awarded-a-four-year-contract-for-hong-kongs-genome-project/>

GISAID

<https://www.gisaid.org/>

UK Biobank (DNANexus)

<https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform>

University of Chicago Biomedical Research Hub (Gen3)

<https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocab247/6432980>

CanCOGeN, Genome Canada, Illumina

<https://www.genomecanada.ca/en/cancogen>

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/>

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



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 & Bioengineering

- Imaging & signal data

NIEHS National Institute of Environmental Health Sciences

- Geolocation

NCATS National Center for Advancing Translational Sciences

- N3C COVID patient data

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		

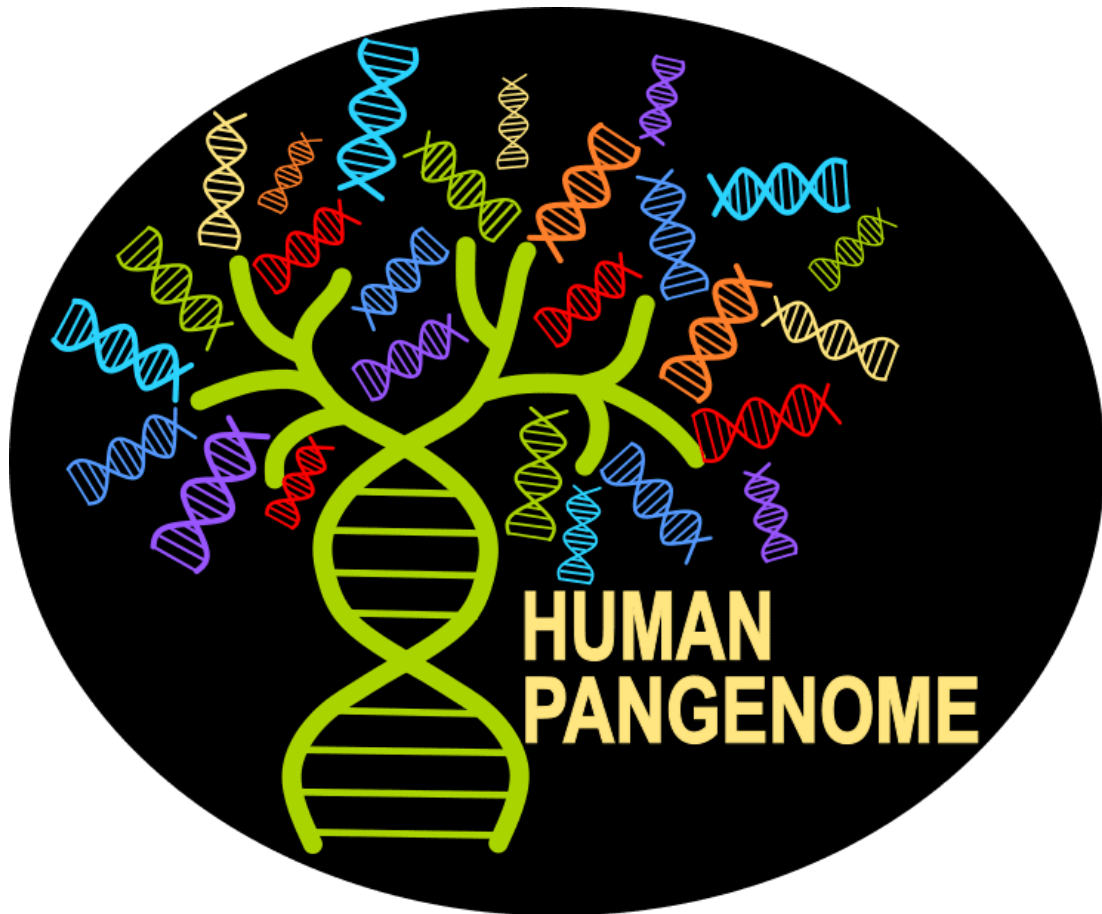
HUMAN PANGENOME FAQs



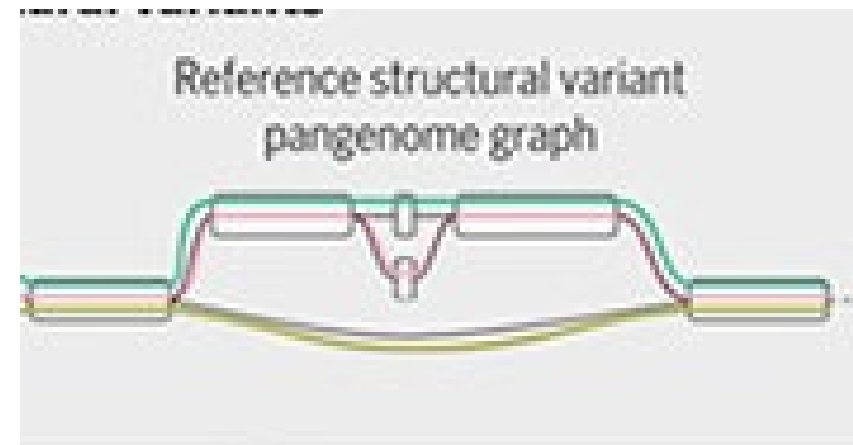
International collaboration at foundation of genomics

- 1st 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

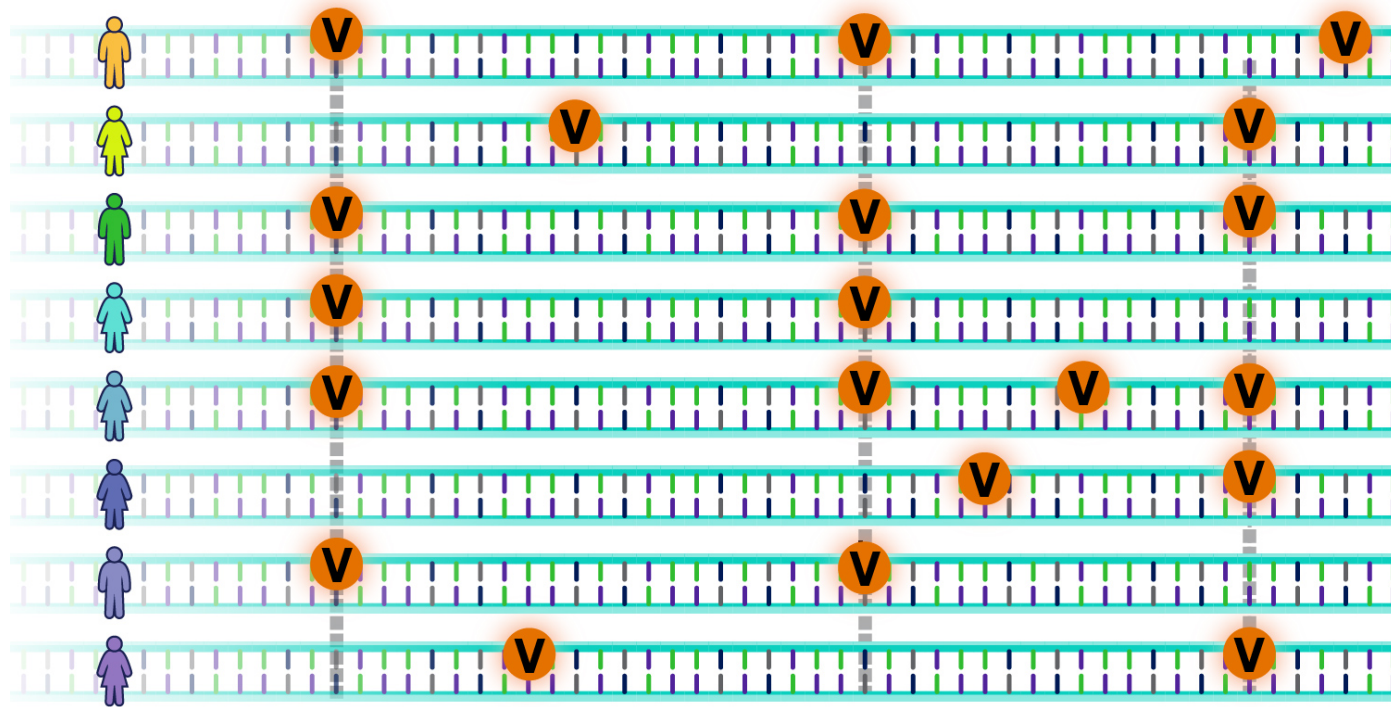


- Many (1000's) genetic syndromes have distinct facial dysmorphisms.
- Sometimes gene defect detected in facial scan of "normal" relative

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



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

NHGRI Consortia



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

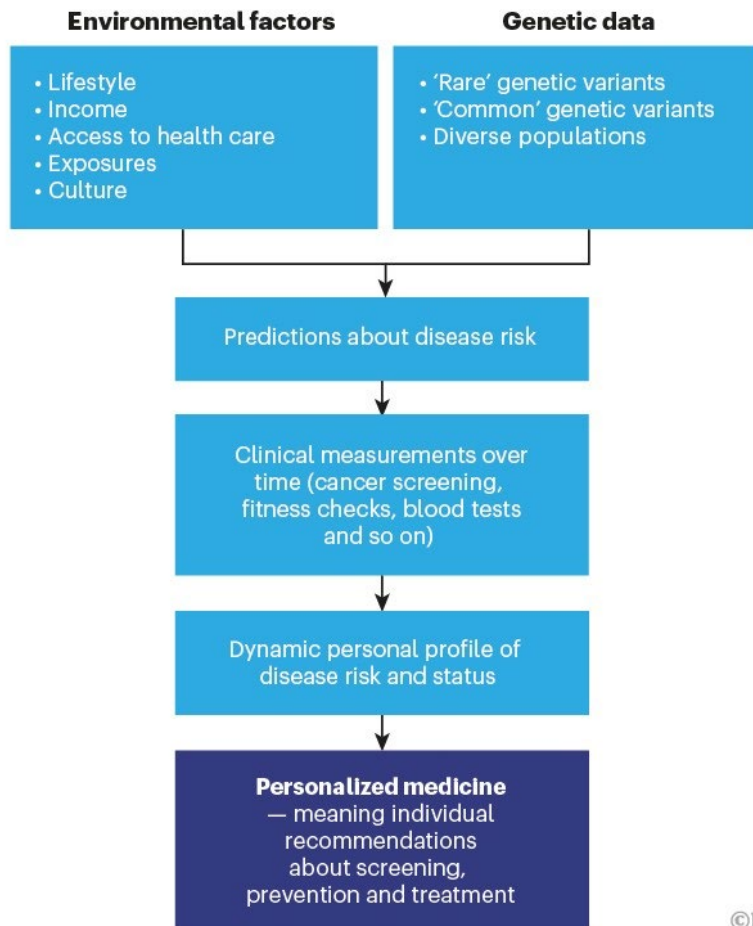
PRIMED to study polygenic (“complex”) diseases using advances in Polygenic Risk Scores (diabetes, heart disease, autism, etc.)

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.



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

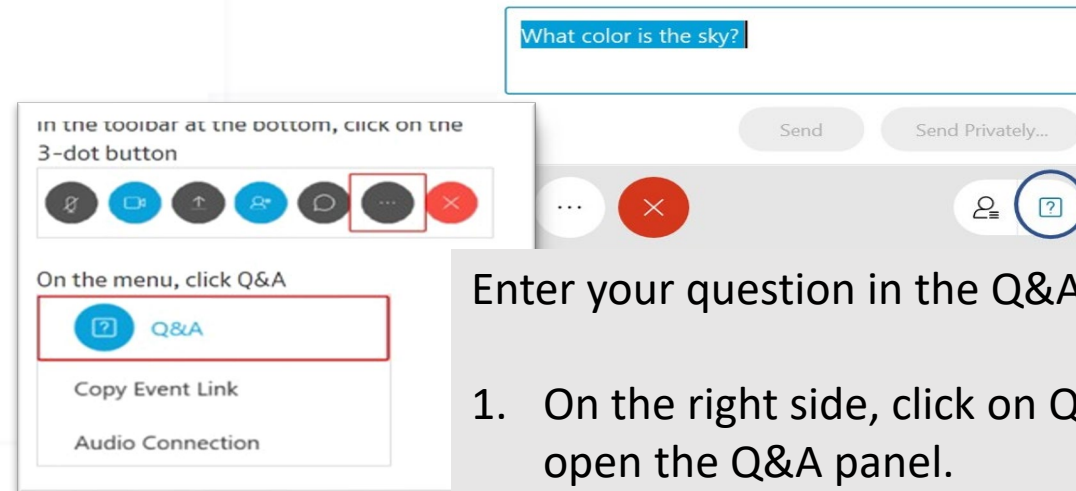
Collaborate. Innovate. Accelerate.

- 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



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.