RSAConference2020
San Francisco | February 24 – 28 | Moscone Center

HUMAN ELEMENT

SESSION ID: PDAC-F01

Securing the Genome: The Intersection of Genomics, Cloud and Big Data

Brandi Davis-Dusenbery PhD
Chief Scientific Officer
Seven Bridges

Brian Castagna
Chief Information Security Officer
Seven Bridges

Is My Dad Really 100% Italian?

79%

9%

8%

2%

1%

Agenda

- Genomics 101
  - Complex Problem and Complex Ecosystem
  - Your DNA
  - Precision Medicine, Participants, and Privacy
  - Big Data + Cloud
- Genomic Threat Models & Security Controls
  - Threat Actors & Models
  - Apply: Cloud Security Controls
- Genomic Compliance Standards & Regulations
- Genomics, Cloud & Pharma
  - Genomics and Cloud Shared Responsibility Model
- Case Study: Cancer Genomics Cloud
- Apply Your Learning
  - How You Can Help Secure the Genome

RSAConference2020

Genomics 101

How did we get to the intersection of genomics, big data and cloud?

Complex Problem and Complex Ecosystem

- Biology is extremely complex!
- To do GOOD, we need GOOD data & lots of it.
- Push & Pull: Protecting the data humanity shares.
- A breakthrough is needed: New algorithms, ways of working.

NIH NATIONAL CANCER INSTITUTE

23andMe®

aws

SANOFI

gsk GlaxoSmithKline

MERCK

Johnson&Johnson

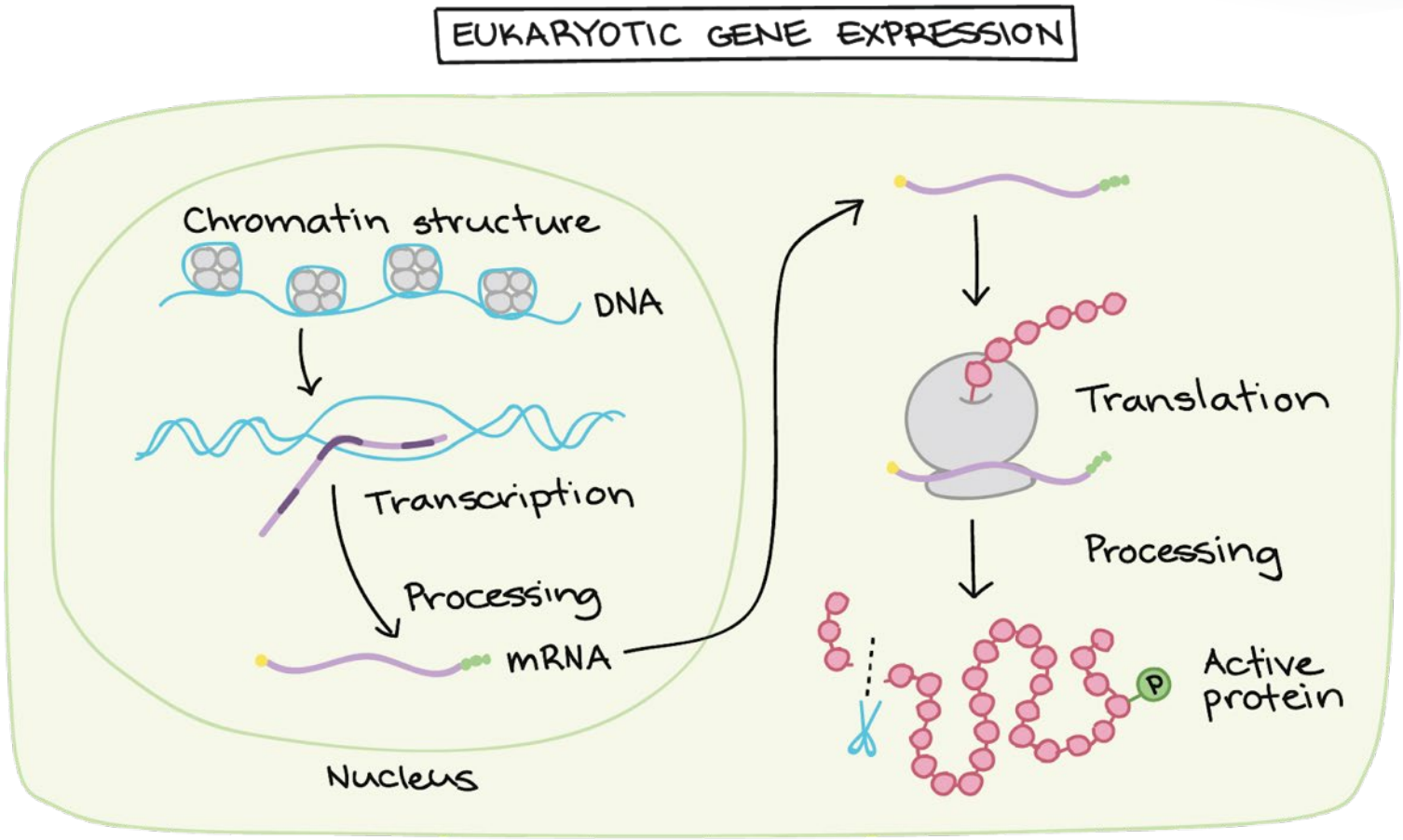
NOVARTIS

Lilly Pfizer

Roche

astellas

DNA Contains the Fundamental Directions For Life



- Your **genome** contains **2.3B nucleotides** of DNA from your mother & father.
- The intersection of these sequences + environment = **phenotypes** (such as response to a drug).
- Uncovering these massively complex relationships requires lots of data.

Combining Genomics + Phenotypic Data Has Already Led to Enormous Advances in Precision Medicine

Single dose gene replacement for leading genetic cause of infant mortality.

Cancer therapies based on molecular signatures of tumor instead of tissue of origin.

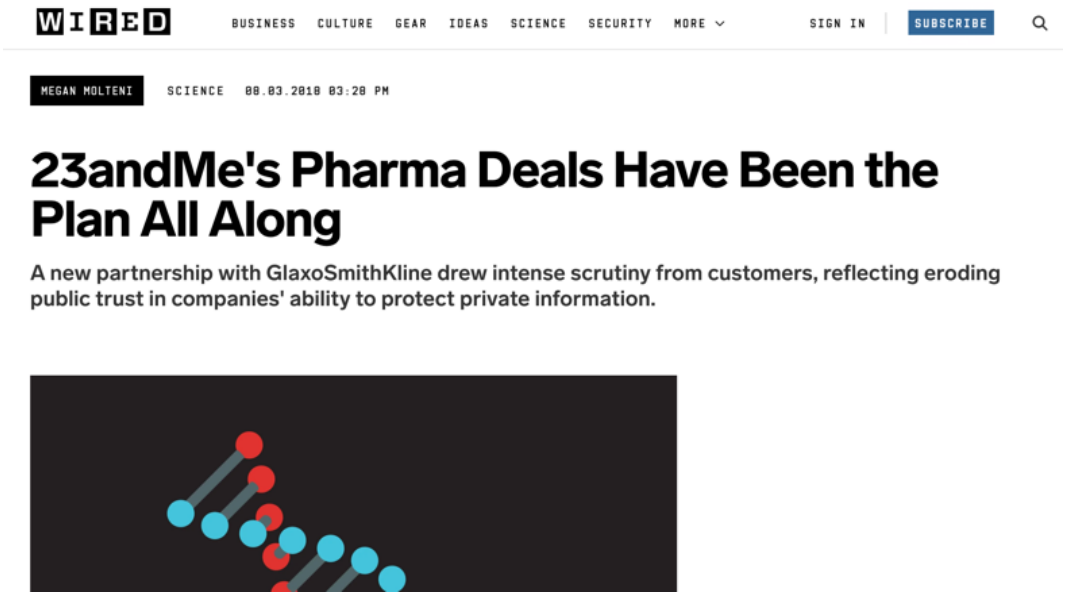
Every Medical Advance Depends on Study Participants

Sometimes these are transparent:

- Research Cohorts (All of Us, UK BioBank)
- Clinical Studies

Use & collection of data is regulated by study consents specifically signed by the participant & approved by Internal Review Boards.

Every Medical Advance Depends on Study Participants



Sometimes these are not so transparent:

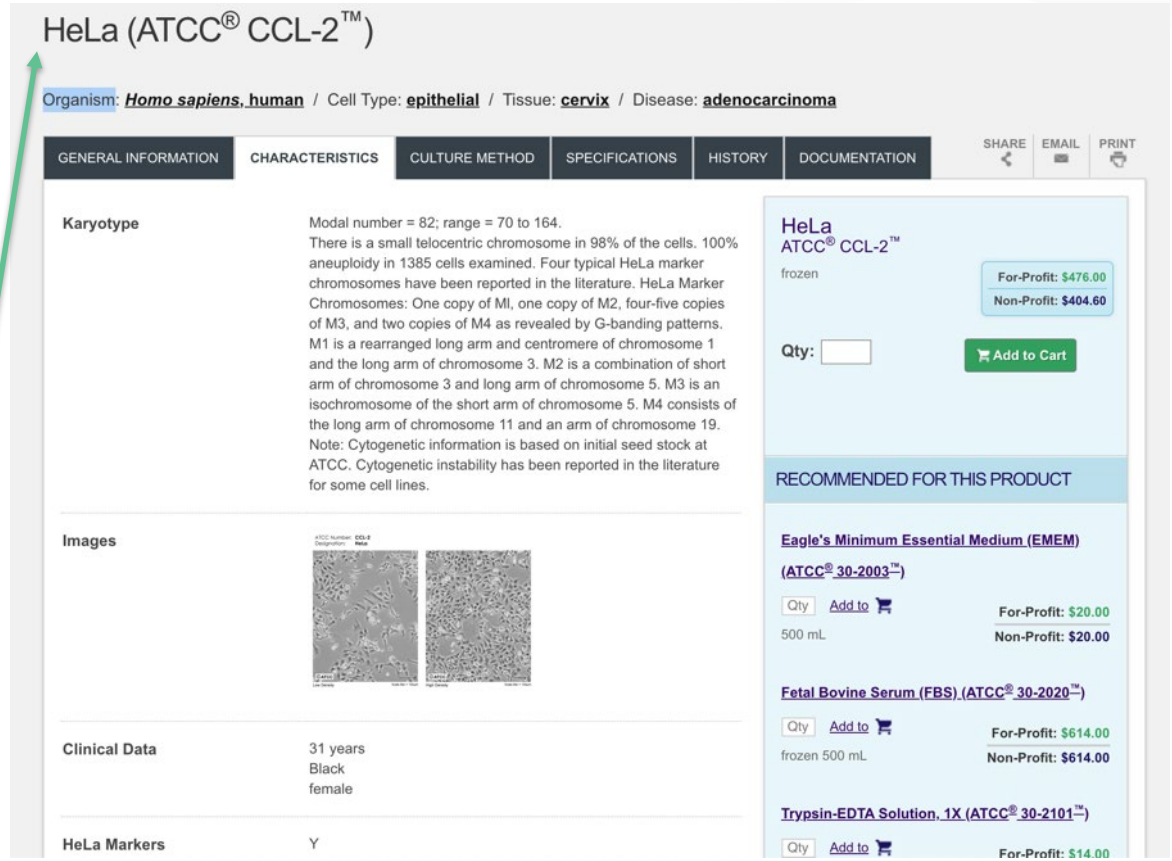
- Direct to consumer genetic tests
- Data collected during health care

Data use regulated by Terms & Conditions of DTC tests – usually need to 'consent in' to allowing your data being used in research studies.

Health Insurance **Portability** & Accountability Act (HIPAA) governs medical data in the US.

Keeping Participant Data Safe & Used for the Intended Purpose is Critical to the Continued Advancement of Medicine

1951 Henrietta Lacks diagnosed with cervical cancer at Johns Hopkins.



HeLa cells were and continue to be fundamentally important to human health.

But its hard to imagine a more complete privacy FAIL.

Genomic Data Presents Unique Privacy Concerns

The genetic sequence of HeLa cells disclose information about Henrietta's descendants.

(note there are specific cancer driving mutations only in the 'somatic' or cancer cells that are not present in the 'germline' and thus not inheritable)

- Unchangeable
- Inherently identifiable
- Discloses information about yourself & relatives

The Scale, Complexity, Velocity of Genomic Data Requires New Analytical Methods, Infrastructure & Frameworks

PLOS | BIOLOGY

PERSPECTIVE

Big Data: Astronomical or Genomical?

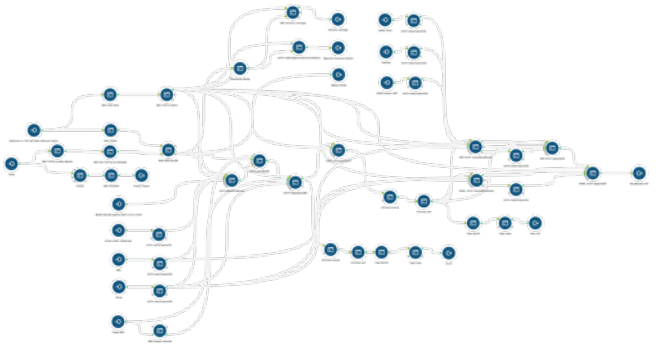
Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

1 Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **2** Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **3** Carl R. Woese Institute for Genomic Biology & Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **4** School of Library and Information Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **5** Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **6** Carl R. Woese Institute for Genomic Biology, Department of Entomology, and Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America

* mschatz@cshl.edu (MCS); sinhas@illinois.edu (SS); generobi@illinois.edu (GER)

CrossMark click for updates

500GB text file per genome



Complex & computationally demanding algorithms

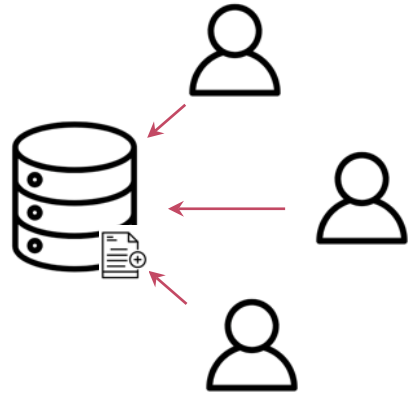
Actionable insights


Cloud is the most economically reasonable way to store and analyze our growing health data corpus.

Cloud Provides Significant Benefits for Health Data Analysis at Scale

- Immediate Scaling
- Levels the Playing Field
  - Researchers at institutions can access powerful data and compute resources.
- Extreme Durability
  - Eliminates or reduces need for backup copies.
- Multi-tenancy of Data
  - Many researchers can access a single dataset.

Old model: send data to compute



New model: send compute to data

RSAConference2020

Genomic Threat Models & Security Controls

It’s the beginning, not the end.

Who are the Likely Threat Actors Targeting Genomic Data?

Nation State

Insider Threat

Criminal

Hacktivist

More Likely

Less Likely

Genomic Threat Background Information

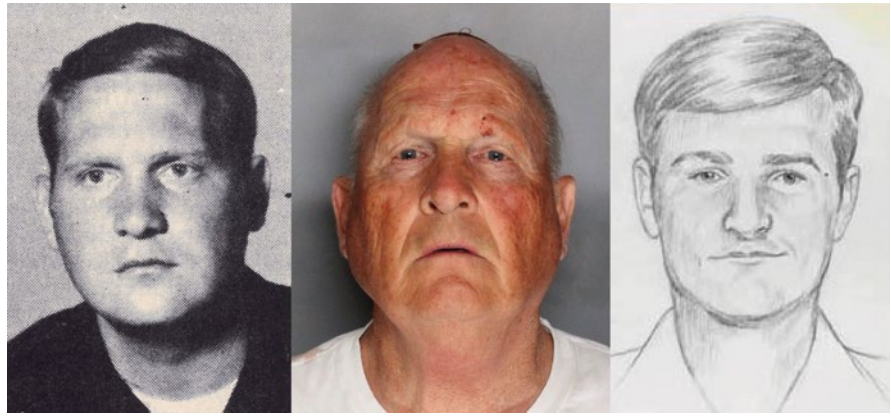
- Your Genome is hacked! Who is at risk?
  - You, Your Children, Grandchildren, Siblings, Cousins, Great Grandchildren, Great Great Grandchildren…..
- “Informed Consent” doesn’t equal understanding what you consent to.
- What did I sign at the hospital?
  - Example: Google’s ‘Project Nightingale’

Genomic Threat Model: Personal & Family Impact

- Finances & Insurance:
  - Life & Disability Insurance
  - Mortgage Loan DENIED due to Alzheimer's gene
- Ransomware Your DNA?
  - Running for political office
  - Blackmail famous or wealthy people – Side Channel Attack
- Over Reliance on Genetic Testing Data: False Sense of Confidence
- Advertising: Monetizing your genes

Genomic Threat Model: Uncle Joe is the Golden State Killer

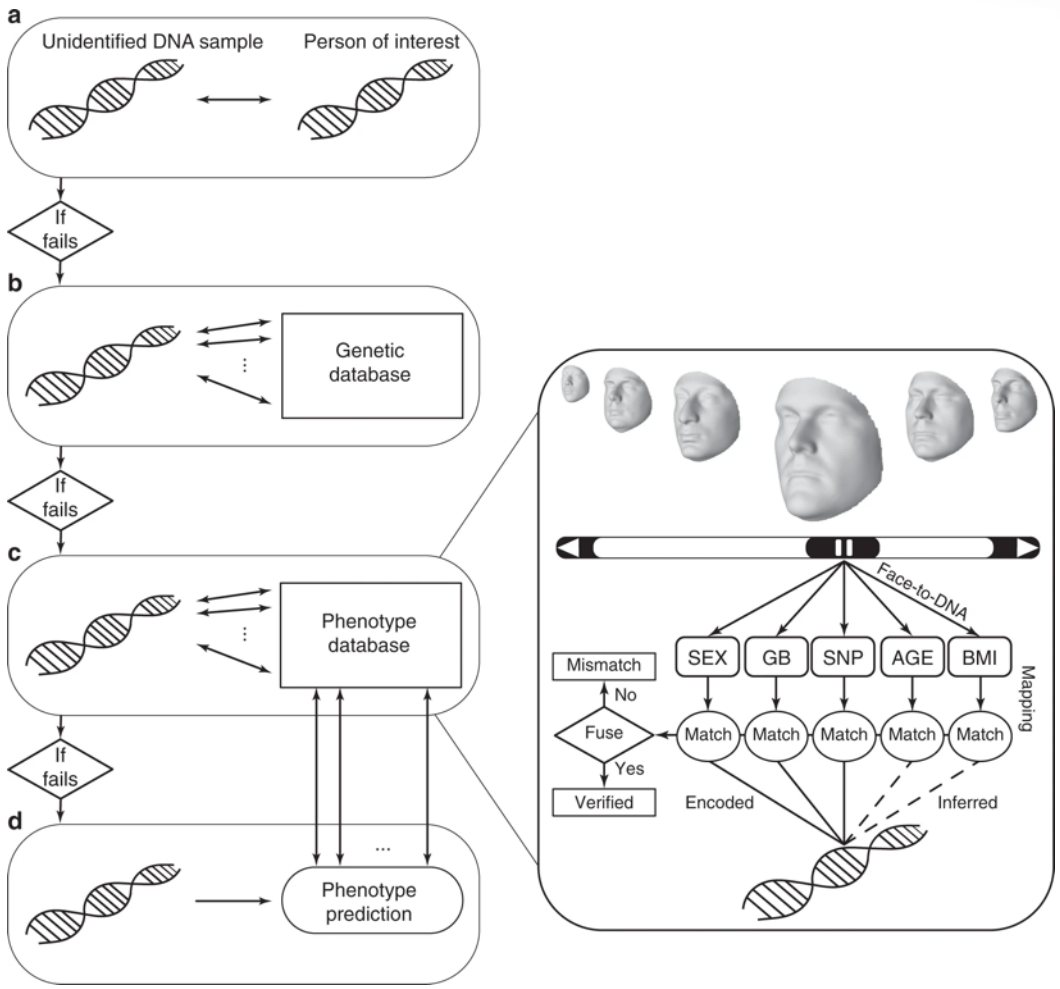
- Golden State Killer committed multiple rapes and murders between 1974 and 1986
- Authorities made a profile on a genealogy website using 37 year old DNA.
- Arrested Joseph James DeAngelo after relative's DNA was a match
- Now a common practice by law enforcement



Genomic Threat Model: Nation State Data Gathering

- Ethnic Discrimination
  - Chinese police purchased 12 DNA sequencers in 2017
  - China collecting DNA from Uyghurs a mostly Muslim ethnic group
- Targeted Bio-Warfare
  - Forced sterilization for selective breeding - Eugenics
  - Ethnic groups susceptible to certain toxins
  - Gene Editing

Genomic Threat Model: Face Predictor



Apply: Cloud Security Controls, Encryption

- In Transit: End to End Encryption TLS 1.2
- At Rest: Database Encryption
- At Rest: Hashing / Encryption of Genomic Data
- In Memory: Limit time decrypted, strong controls where decryption takes place (key management, server hardening, detections, etc.)

Apply: Cloud Security Controls, Core Security Areas

- Security Hardening
- Penetration Testing
- Container Architectures
- Threat Intelligence / Indicators of Compromise (IOCs)
- Data Loss Prevention (DLP)
- Security Tool Chain
  - WAF, Security Groups, IP Tables, Configuration Automation, Patch Automation.
  - Log Mgt, Security Analytics, User Entity Behavior Analysis (UEBA), Security Orchestration, Automation and Response (SOAR).

RSAConference2020

Genomic Compliance Standards & Regulations

An evolving challenge

A Patchwork of Regulations to Protect Patient Genomic Data

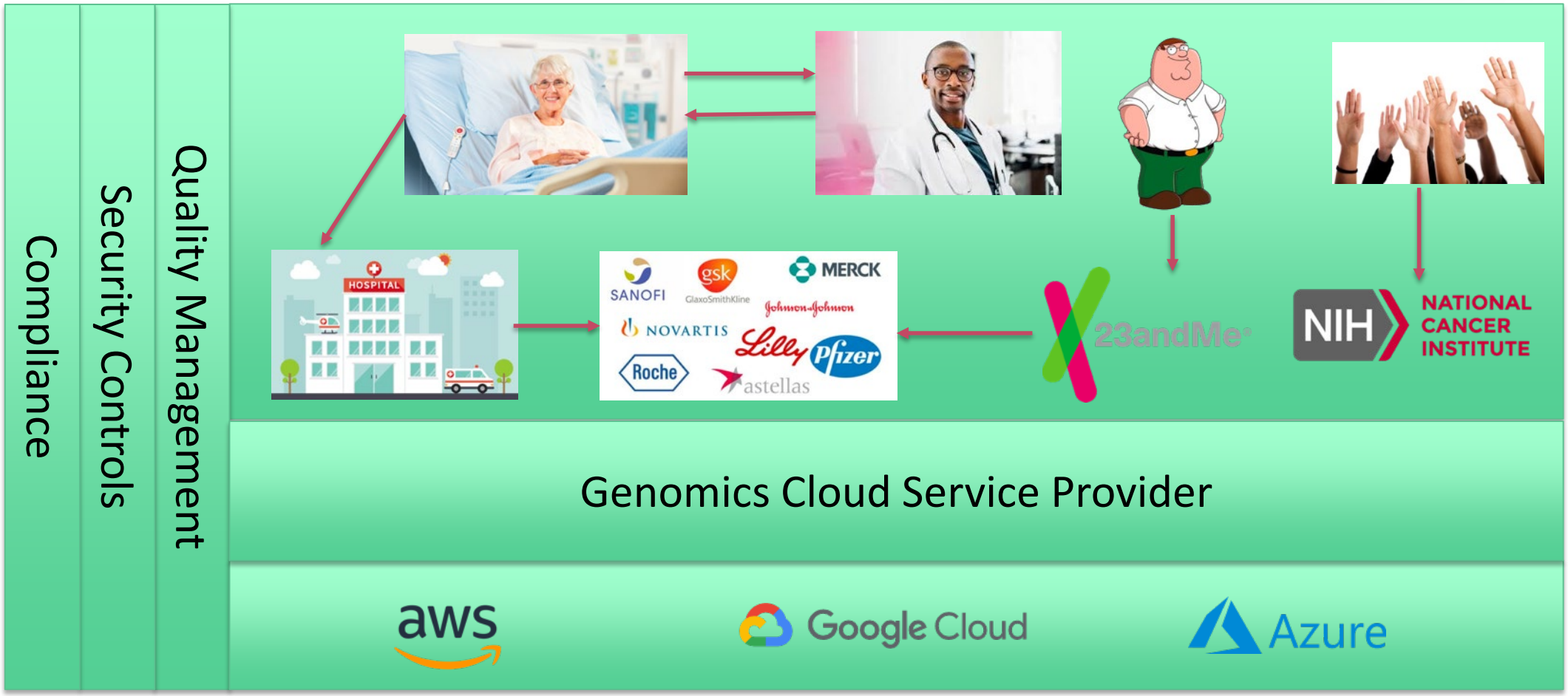
- Large Research Studies
  - Database of Genotypes and Phenotypes (dbGAP)
  - Institutional Review Boards (IRB)
- Direct Consumer Tests
  - Informed Consent
  - Genetic Information Nondiscrimination Act (GINA) (US)
- Data Collected During Healthcare
  - GxP, HIPAA, 21 CFR part 11 (US)
  - GDPR (EU)

RSAConference2020

Genomics & Cloud & Pharma

A New Shared Responsibility Model

Genomics Cloud Shared Responsibility Model



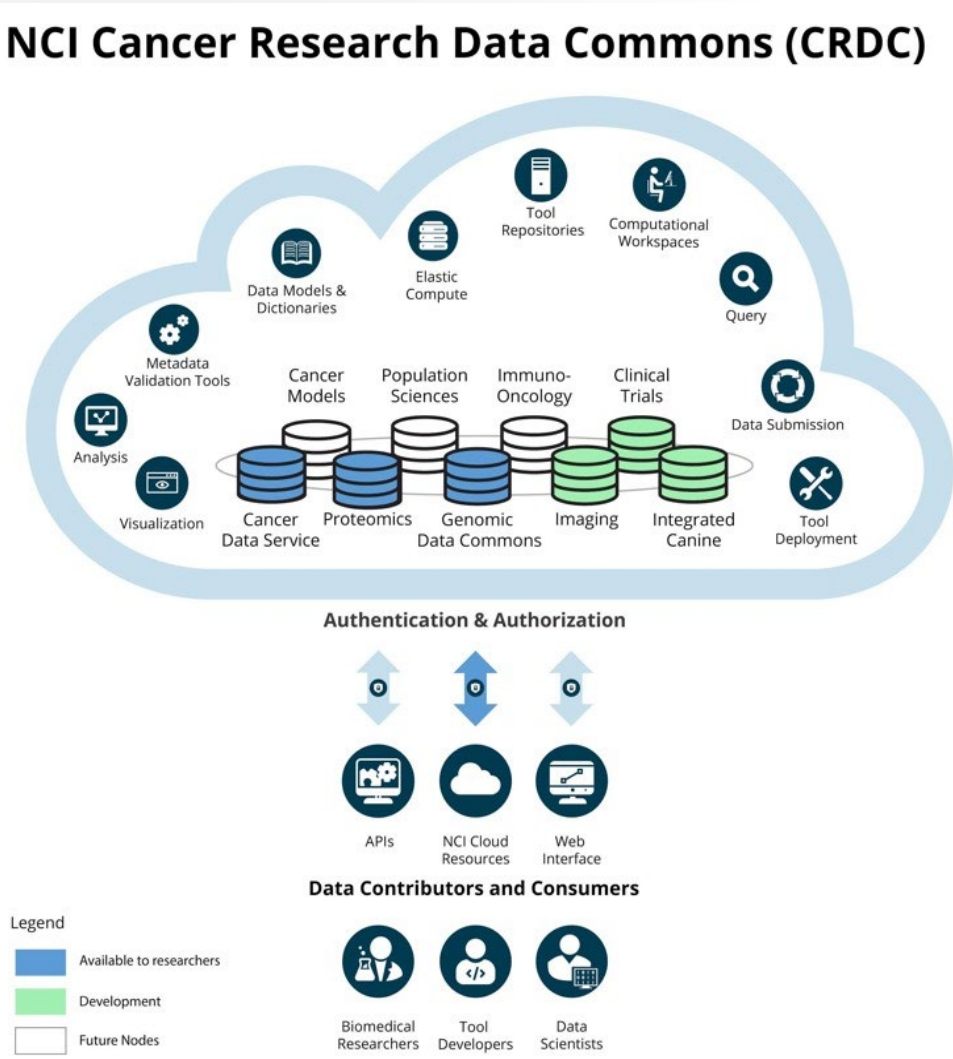
RSAConference2020

Case Study

The Cancer Genomics Cloud for the National Cancer Institute

The Cancer Genomics Cloud (CGC)

A Cloud Resource within the NCI Cancer Research Data Commons for secure storage, sharing & analysis of petabytes of public, multi-omic cancer datasets.



NATIONAL CANCER INSTITUTE

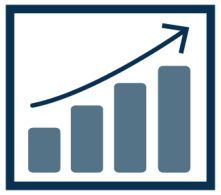
The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 17X053 under Contract No. HHSN261200800001E.

SevenBridges

The CGC Helps Researchers Do More

- A stable, secure, and highly customizable cloud storage and computing platform.
- A user-friendly portal for collaborative analysis of petabytes of public data alongside private data.
- An optimized venue for reproducible data analysis using validated tools and pipelines.

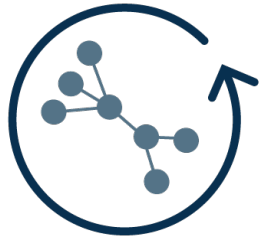
Easy data management



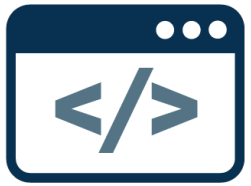
Scalable computation

Optimized bioinformatics algorithms

Secure collaboration

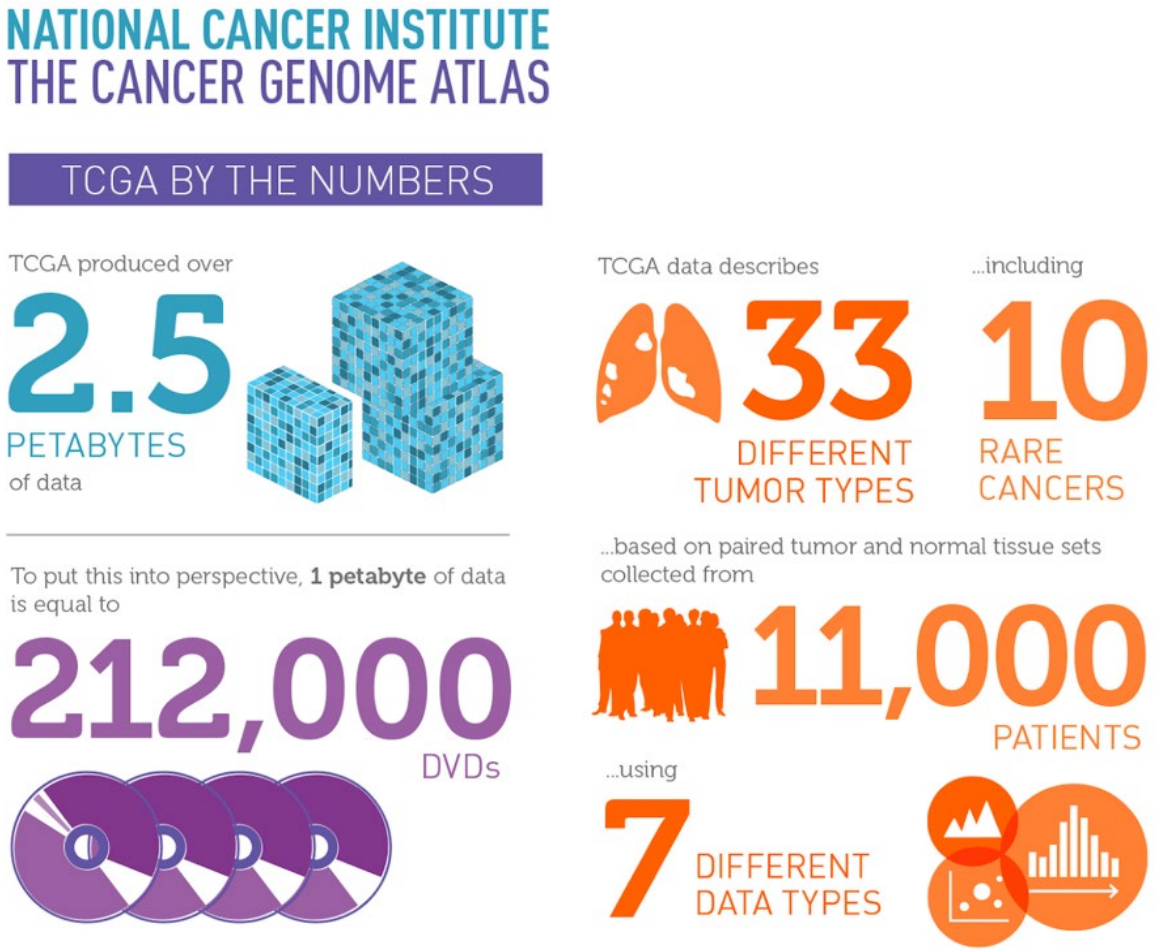


Flexible & fully reproducible methods



Extensible and developer-friendly platform

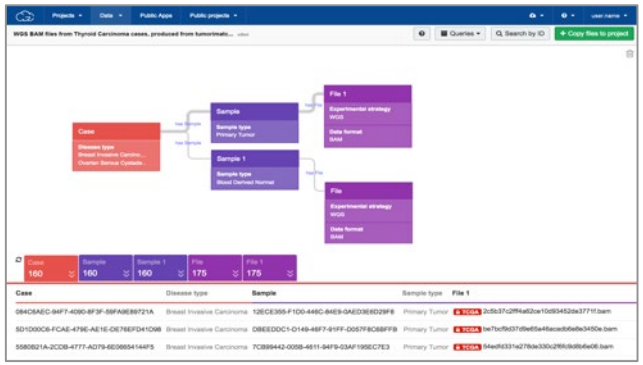
The Challenges of Working With Large Datasets: The Cancer Genome Atlas (TCGA)

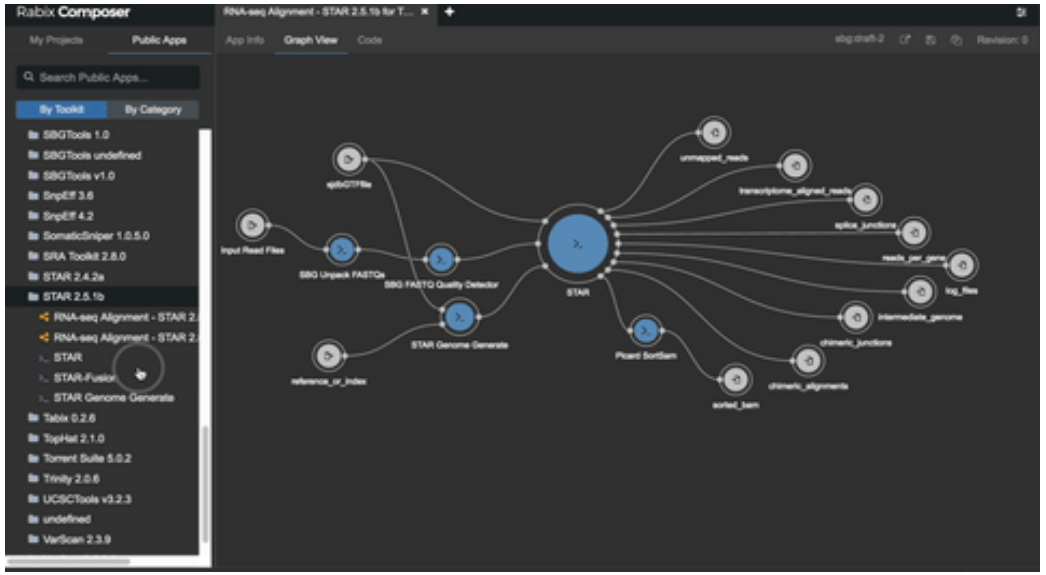


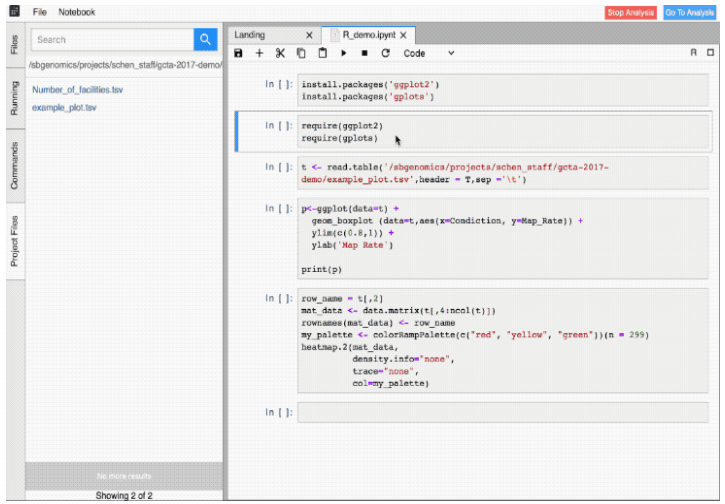
CGC Provides an Easy Way to Find and Analyze Data

Visually explore and access **3+ PB** of multi-omic public data through interactive query tools & APIs.

Use the **400+** cloud- and cost-optimized tools in our Public Apps library OR deploy custom tools using **Rabix Composer**, Jupyter notebooks or R packages



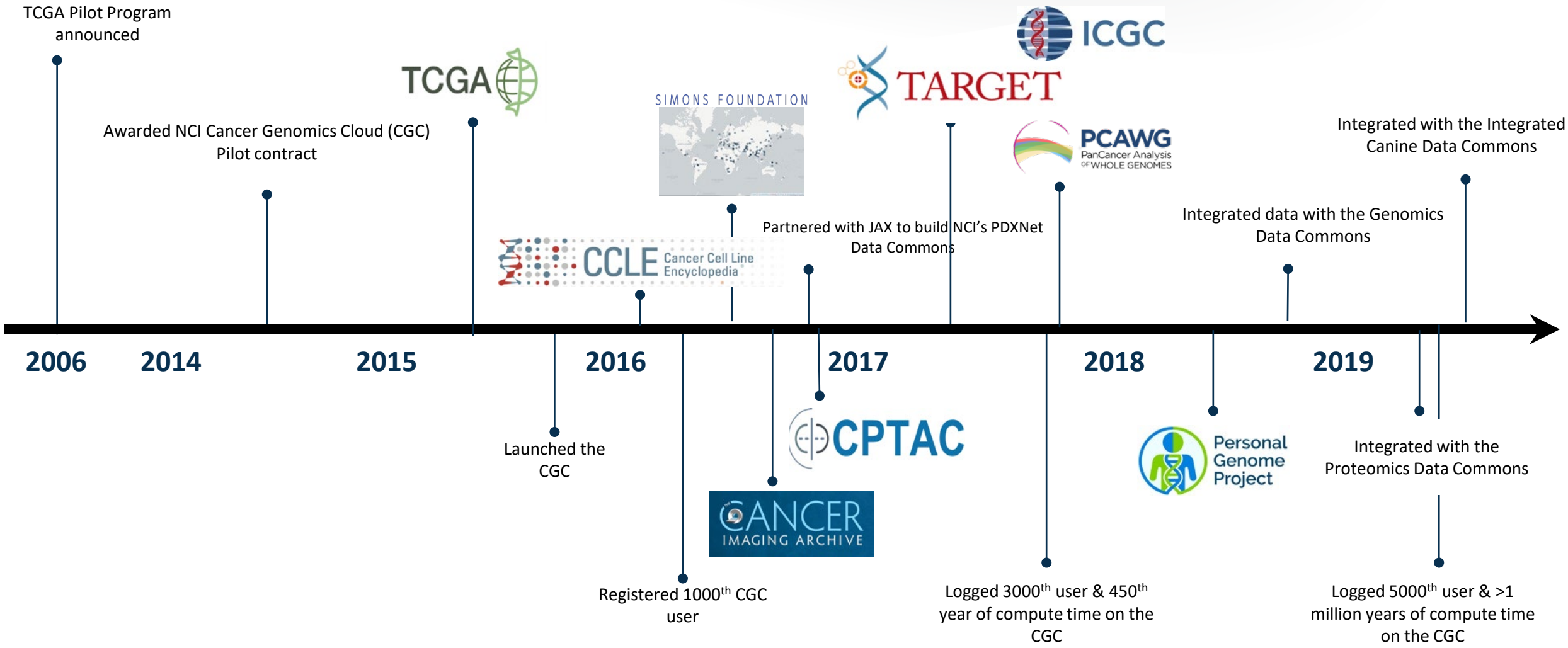




SevenBridges

RSAConference2020

Growth of the Cancer Genomics Cloud Ecosystem



RSAConference2020

Apply Your Learning

How You Can Help Secure the Genome

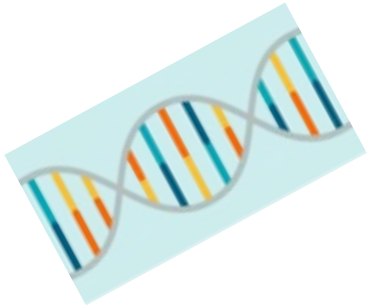
Apply: Genomic Data More Valuable Then Your Bank Account Number

- **At Your Company:**
  - Place higher value in protecting genomic data, use security controls.
- **In Your Life:**
  - Protect your genome!
  - Informed Consent: Consider impacts on other people – Your family, and other families

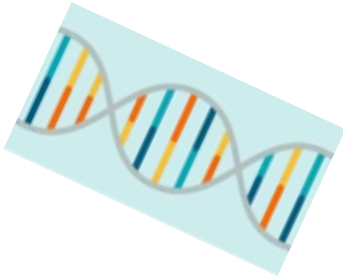
Apply: The Balance of Genomic Data Sharing

Apply: How You Can Help Advance Security & Compliance

- Got Venture Capital Funding? Start a Company…
  - Better encryption schemes needed for genomes
  - A new wave of Tech DLP features for genomes

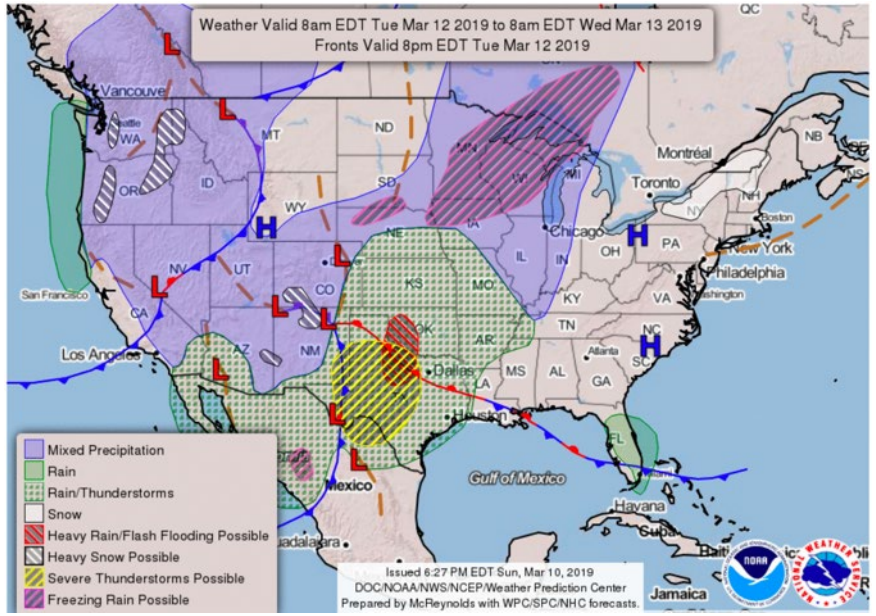


- Regulatory & Standards Updates
  - Advocate standards bodies to align PHI focused standards with the realities of genomic data.



Apply: Secure Collaboration Needed!

- More Like Weather Data….
- Breaking Down Borders
- Government Funding Required!



Questions

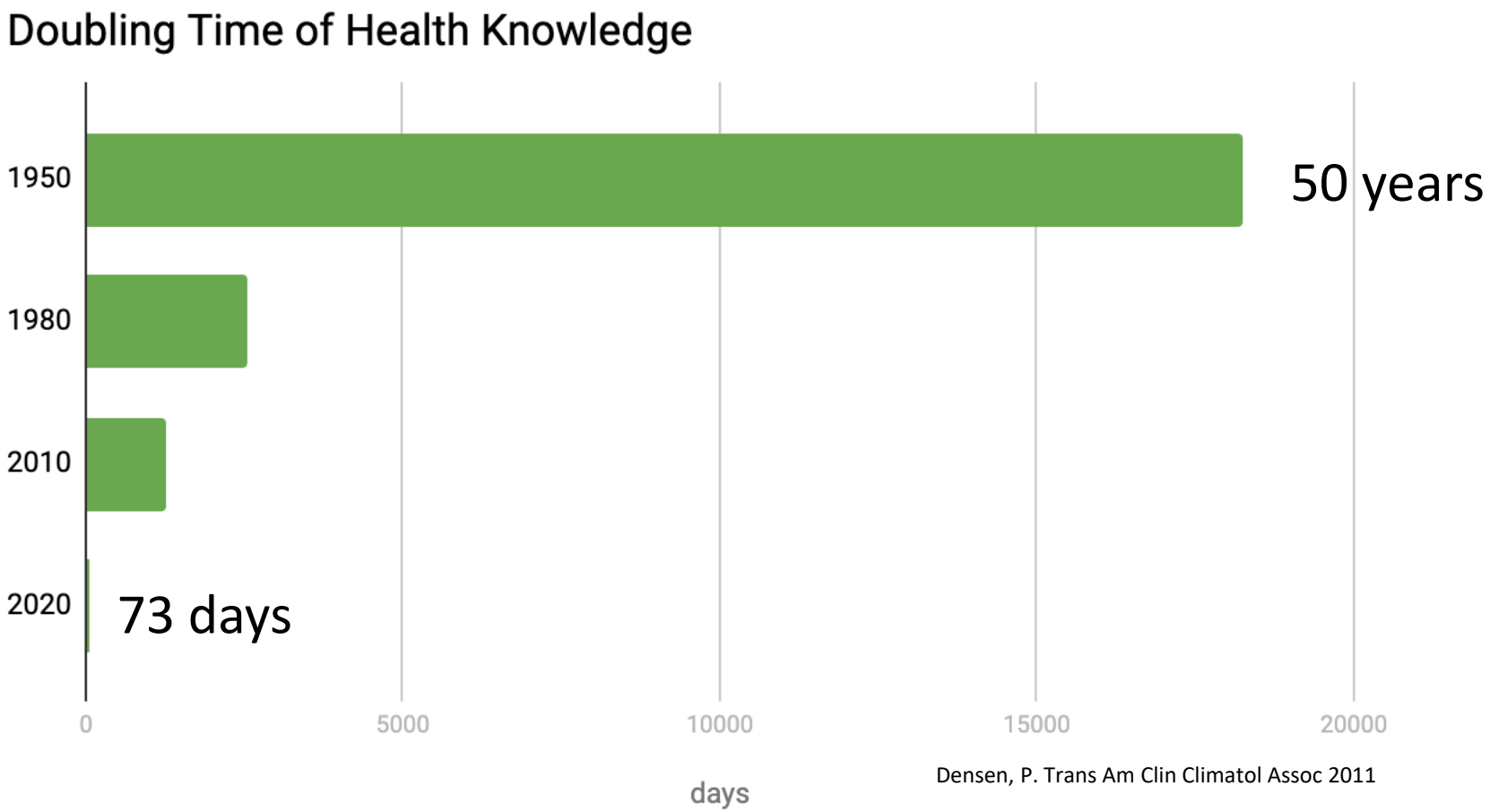
RSAConference2020

Thank You!

RSAConference2020

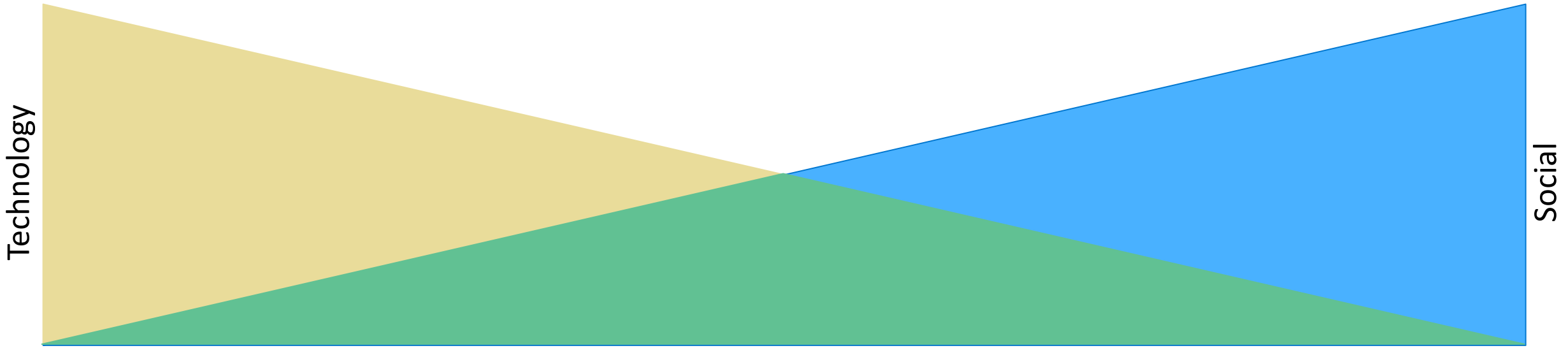
Appendix

The Rate of Data Generation is Increasing Rapidly



Densen, P. Trans Am Clin Climatol Assoc 2011

Using This Information to Improve Patient Outcomes Isn't Just a Technology Challenge



Scalable & **Secure** Environments

Data Harmonization & Organization

Data Sharing & Collaboration

Data Analysis Fluency

High Impact Publications on the CGC

Cell Reports
Article

OPEN ACCESS
CellPress

Structural Differences between Pri-miRNA Paralogs Promote Alternative Drosha Cleavage and Expand Target Repertoires

Xavier Bofill-De Ros,[1] Wojciech K. Kasprzak,[2] Yuba Bhandari,[3] Lixin Fan,[4] Quinn Cavanaugh,[1] Minjie Jiang,[1] Lisheng Dai,[1] Acong Yang,[1] Tie-Juan Shao,[1,5] Bruce A. Shapiro,[6] Yun-Xing Wang,[3] and Shuo Gu[1,7,*]
[1]RNA Mediated Gene Regulation Section, RNA Biology Laboratory, Center for Cancer Research, National Cancer Institute, Frederick, MD 21702, USA
[2]Basic Science Program, RNA Biology Laboratory, Frederick National Laboratory for Cancer Research sponsored by the National Cancer Institute, Frederick, MD 21702, USA
[3]Protein-Nucleic Acid Interaction Section, Structural Biophysics Laboratory, National Cancer Institute, Frederick, MD 21702, USA
[4]Small-Angle X-ray Scattering Core Facility, Center for Cancer Research of the National Cancer Institute, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD 21702, USA
[5]School of Basic Medicine, Zhejiang Chinese Medical University, Hangzhou, 310053, China
[6]RNA Structure and Design Section, RNA Biology Laboratory, Center for Cancer Research, National Cancer Institute, Frederick

Molecular Cancer Research

Search... Advanced Search

Home About Articles For Authors Alerts News

Research Article

The Germline Variants rs61757955 and rs34988193 are Predictive of Survival in Lower Grade Glioma Patients

Ajay Chatrath, Manjari Kiran, Pankaj Kumar, Aakrosh Ratan, and Anindya Dutta
DOI: 10.1158/1541-7786.MCR-18-0996 Check for updates

PLOS GENETICS

BROWSE PUBLISH ABO

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Association analysis using somatic mutations

Yang Liu, Qianchan He, Wei Sun

Version 2 Published: November 2, 2018 • https://doi.org/10.1371/journal.pgen.1007746

PNAS Proceedings of the National Academy of Sciences of the United States of America

Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers

Roozbeh Dehghannasiri, Donald E. Freeman, Milos Jordanski, Gillian L. Hsieh, Ana Damljanovic, Erik Lehnert, and Julia Salzman

PNAS first published July 15, 2019 https://doi.org/10.1073/pnas.1900391116

The Cancer Genomics Cloud

5,000+ users from **80+** countries have used the CGC to run **1,000,000+** computational tasks representing **1000+** years of total compute time to:

- Detect aberrant splice junctions and splicing profiles across patient populations
- Identify neoantigens arising from novel gene fusion events
- Profile miRNA expression across patient populations
- Conduct HLA typing to identify neoantigens
- Compare viral infection patterns across patient populations
- Detect novel gene fusions from RNA-Seq data
- Identify cis-regulatory region variants across patient populations
- ...and much more

Image References

- https://en.wikipedia.org/wiki/Peter_Griffin
- http://europe.phillipmartin.info/turkey_dervish.htm
- https://www.clipartwiki.com/downpng/iTJRJx_14-cliparts-for-free-download-italy-clipart-waiter/
- http://europe.phillipmartin.info/croatia_woman.htm
- http://clipart-library.com/clipart/323592.htm
- https://www.ascom.com/products/category/patient-resident-devices/patient-handsets.html
- https://en.wikipedia.org/wiki/23andMe#/media/File:23andMe_logo.svg
- https://www.thebalancecareers.com/how-to-become-a-doctor-525591
- https://www.cancerhorizons.com/innovations/medication/top-20-cancer-drugs/
- https://www.pinclipart.com/downpngs/xiRwJT_free-png-under-construction-barrier-png-images-transparent/
- https://blog.color.com/accelerating-the-pace-of-scientific-research-for-diverse-populations-through-low-coverage-whole-96ae8b934d5
- https://www.nih.gov/news-events/nih-research-matters/beyond-human-genome
- http://clipart-library.com/clip-art/transparent-anonymous-mask-5.htm
- https://www.sanger.ac.uk/science/collaboration/uk-biobank-whole-genome-sequencing-project
- https://www.ukbiobank.ac.uk/
- https://www.forbes.com/sites/kenberman/2019/02/21/genome-sequencing-stocks-on-the-rise/#37decca1f519
- https://genomeasia100k.org/
- https://en.wikipedia.org/wiki/Amazon_Elastic_Compute_Cloud
- https://www.thingforward.io/techblog/2018-11-07-cloud-providers-comparison-for-iot-applications-amazon-vs-microsoft-vs-google.html
- https://www.pinclipart.com/downpngs/bRJTx_clipart-globe-clipart-19-globe-clip-art-transparent/
- https://www.eweek.com/security/research-half-of-enterprises-suffered-insider-attacks-in-last-12-months
- https://www.nature.com/articles/s41467-019-10617-y/figures/1
- http://clipart-library.com/clipart/ziX5aAbXT.htm
- https://www.wired.co.uk/article/encryption-software-app-private-data-safe
- http://www.catherinebruns.com/wp-content/uploads/2011/01/Raised-hands.jpg
- https://mypoppet.com.au/makes/the-patchwork-dress-views-of-hong-kong/
- https://www.weather.gov/news/190204-national-forecast-chart
- https://cloud.google.com/
- https://www.atlassian.com/company
- https://commons.wikimedia.org/wiki/File:Splunk_logo.png
- https://github.com/logos

References

- https://www.theverge.com/2018/6/6/17435166/myheritage-dna-breach-genetic-privacy-bioethics
- https://healthitsecurity.com/news/dna-testing-service-vendor-reports-years-long-consumer-data-breach
- https://www.vox.com/recode/2019/12/13/20978024/genetic-testing-dna-consequences-23andme-ancestry
- https://www.hopkinsmedicine.org/henriettalacks/
- https://en.wikipedia.org/wiki/Henrietta_Lacks
- https://www.theverge.com/2018/4/26/17288532/golden-state-killer-east-area-rapist-genealogy-websites-dna-genetic-investigation
- https://www.nytimes.com/2018/04/27/us/golden-state-killer-case-joseph-deangelo.html
- http://www.cncpunishment.com/forums/showthread.php?12349-Joseph-James-DeAngelo-quot-Golden-State-Killer
- https://leapsmag.com/bad-actors-getting-your-health-data-is-the-fbis-latest-worry/
- https://www.nytimes.com/2019/02/21/business/china-xinjiang-uighur-dna-thermo-fisher.html
- https://www.offthegridnews.com/current-events/politics/bio-warfare-and-terrorism-the-quiet-threat/
- https://www.nature.com/articles/s41467-019-10617-y
- https://www.cnbc.com/2018/06/16/5-biggest-risks-of-sharing-dna-with-consumer-genetic-testing-companies.html
- https://www.washington.edu/news/2019/10/29/genetic-genealogy-site-vulnerable-compromised-data-impersonations/
- https://thenextweb.com/security/2019/11/08/dna-testing-startup-exposes-customer-info-in-data-breach/
- https://www.ucdavis.edu/news/hobbyist-dna-services-may-be-open-genetic-hacking/
- www.cancer.gov/ccg