

Genomic Data & Privacy

Risks & opportunities

Michael Baudis | UZH BIO390 HS22 | December 2022

Genomic Data & Privacy

Risks & opportunities

- Why do we need a lot of data for understanding genomic variation in health and disease?
- Data sharing protocols ...
 - ▶ GA4GH Beacon
- Breaking data privacy
 - ▶ Different types of (genomic) privacy attacks
 - ▶ Beacon attacks and mitigation
 - ▶ DTC and Longe-range familial attacks
- Regulation of genome data production & access in Switzerland
- Some strategies for enabling genomic data sharing & re-use

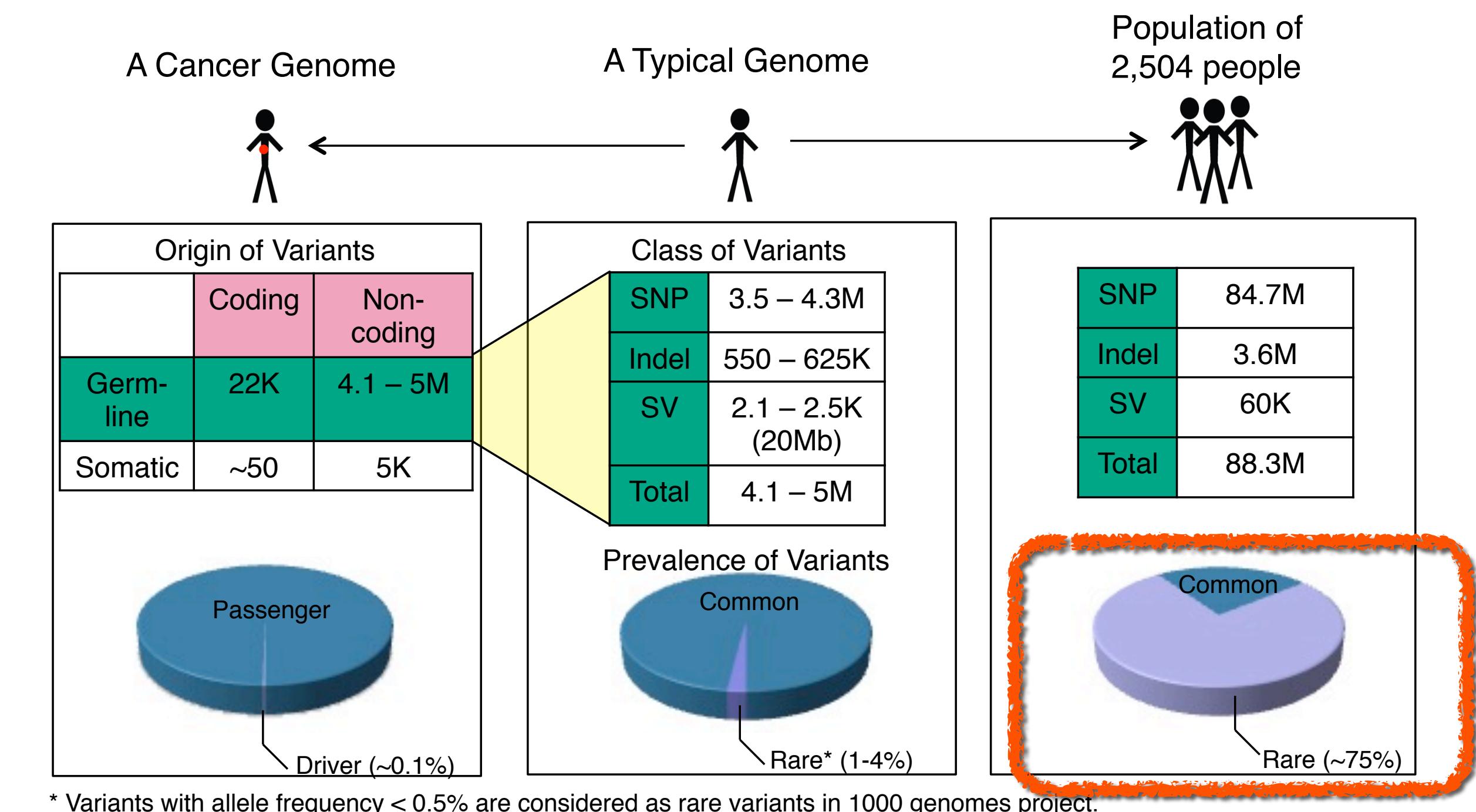
The trouble with human genome variation



Finding Somatic Mutations In Cancer

Many Needles in a Large Haystack

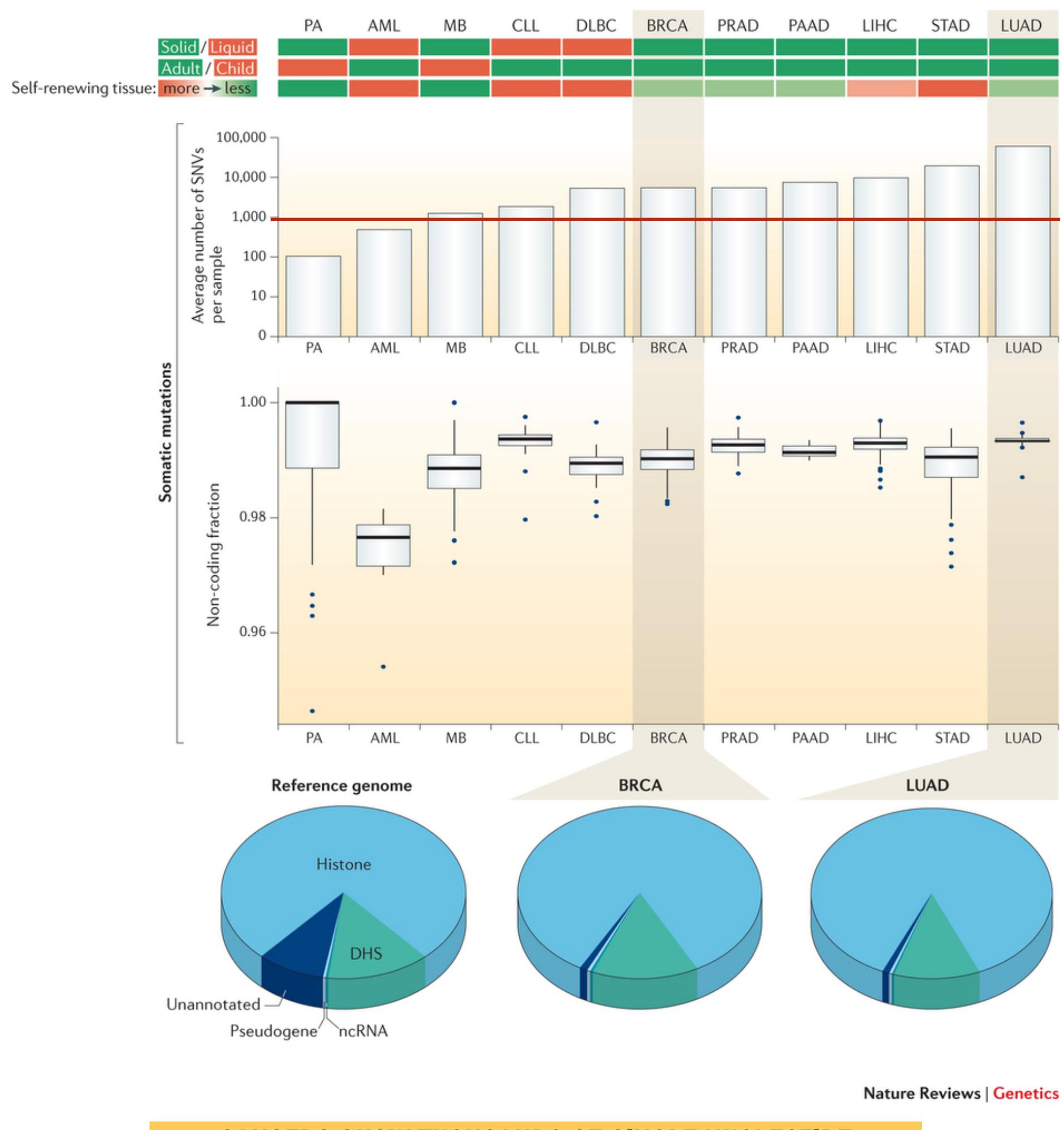
- a typical human genome (~3 billion base pairs) has ~5 million variants
- most of them are "**rare**"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease



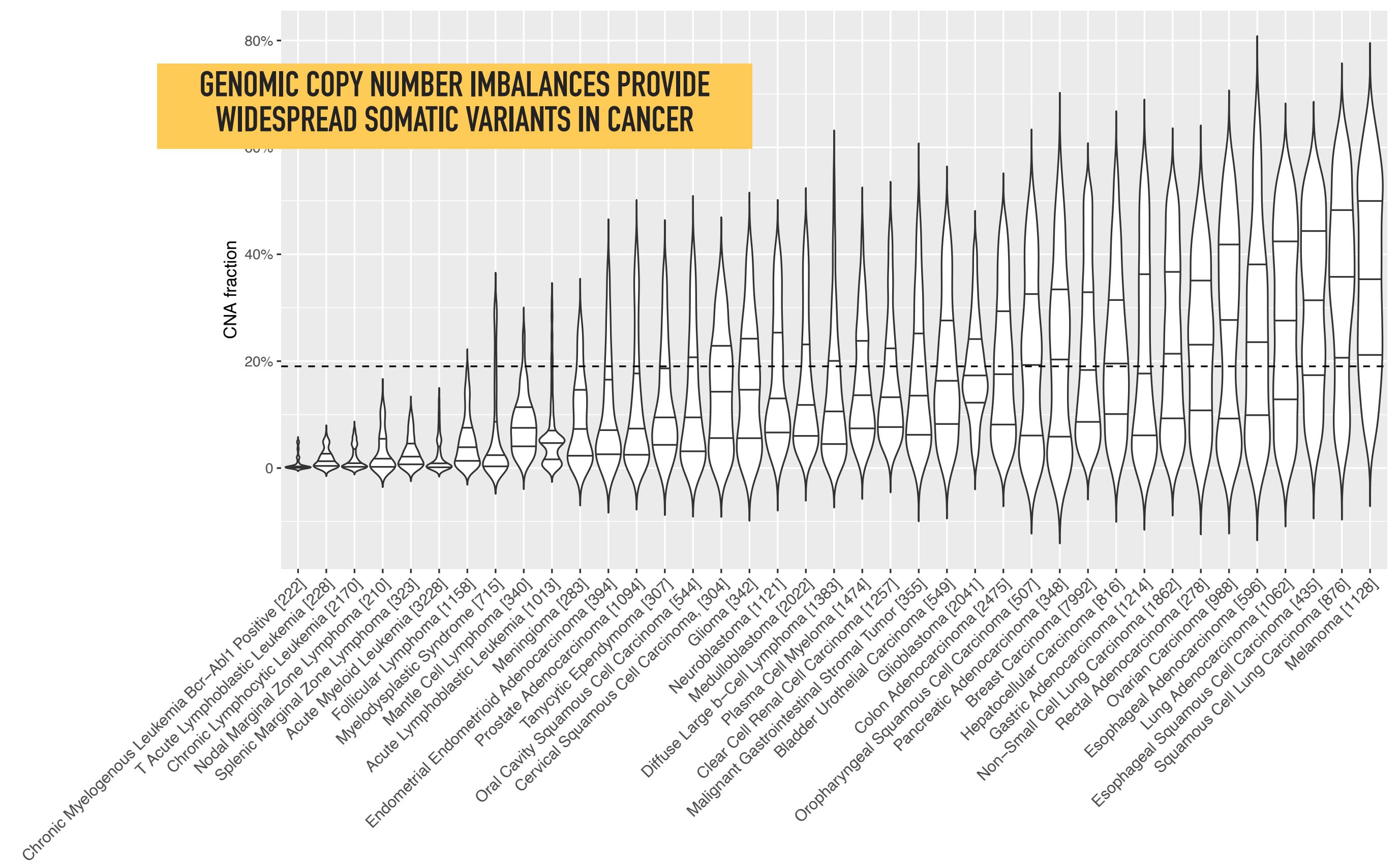
The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

Quantifying Somatic Mutations In Cancer



Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016))



On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from progenetix.org

Comparison of *PIK3CA* Mutation Prevalence in Breast Cancer Across Predicted Ancestry Populations

Jessica W. Chen, PhD¹; Karthikeyan Murugesan, MS²; Justin Y. Newberg, PhD²; Ethan S. Sokol, PhD²; Heidi M. Savage, BA¹; Thomas J. Stout, PhD³; Sophia L. Maund, PhD¹; and Katherine E. Hutchinson, PhD¹

JCO Precision Oncology no. 6 (2022) e2200341. Published online November 29, 2022.
PMID: 36446041 | DOI: 10.1200/PO.22.00341

PURPOSE Understanding the differences in biomarker prevalence that may exist among diverse populations is invaluable to accurately forecast biomarker-driven clinical trial enrollment metrics and to advance inclusive research and health equity. This study evaluated the frequency and types of *PIK3CA* mutations (*PIK3CA*mut) detected in predicted genetic ancestry subgroups across breast cancer (BC) subtypes.

METHODS Analyses were conducted using real-world genomic data from adult patients with BC treated in an academic or community setting in the United States and whose tumor tissue was submitted for comprehensive genomic profiling.

RESULTS Of 36,151 patients with BC (median age, 58 years; 99% female), the breakdown by predicted genetic ancestry was 75% European, 14% African, 6% Central/South American, 3% East Asian, and 1% South Asian. We demonstrated that patients of African ancestry are less likely to have tumors that harbor *PIK3CA*mut compared with patients of European ancestry with estrogen receptor-positive/human epidermal growth factor receptor 2-negative (ER+/HER2-) BC (37% [949/2,593] v 44% [7,706/17,637]; $p = 4.39E-11$) and triple-negative breast cancer (8% [179/2,199] v 14% [991/7,072]; $p = 6.07E-13$). Moreover, we found that *PIK3CA*mut were predominantly composed of hotspot mutations, of which mutations at H1047 were the most prevalent across BC subtypes (35%-41% ER+/HER2- BC; 43%-61% HER2+ BC; 40%-59% triple-negative breast cancer).

CONCLUSION This analysis established that tumor *PIK3CA*mut prevalence can differ among predicted genetic ancestries across BC subtypes on the basis of the largest comprehensive genomic profiling data set of patients with cancer treated in the United States. This study highlights the need for equitable representation in research studies, which is imperative to ensuring better health outcomes for all.

JCO Precis Oncol 6:e2200341. © 2022 by American Society of Clinical Oncology

Creative Commons Attribution Non-Commercial No Derivatives 4.0 License

Key Objective

As both biomarker-driven precision medicine trials and calls for diversity in clinical trials become increasingly common, accurate assessment of biomarker prevalence is critical for informing study enrollment metrics. In this study, we investigated the variation in the frequency and spectrum of *PIK3CA* mutations in breast cancer (BC) across predicted genetic ancestry subgroups.

Knowledge Generated

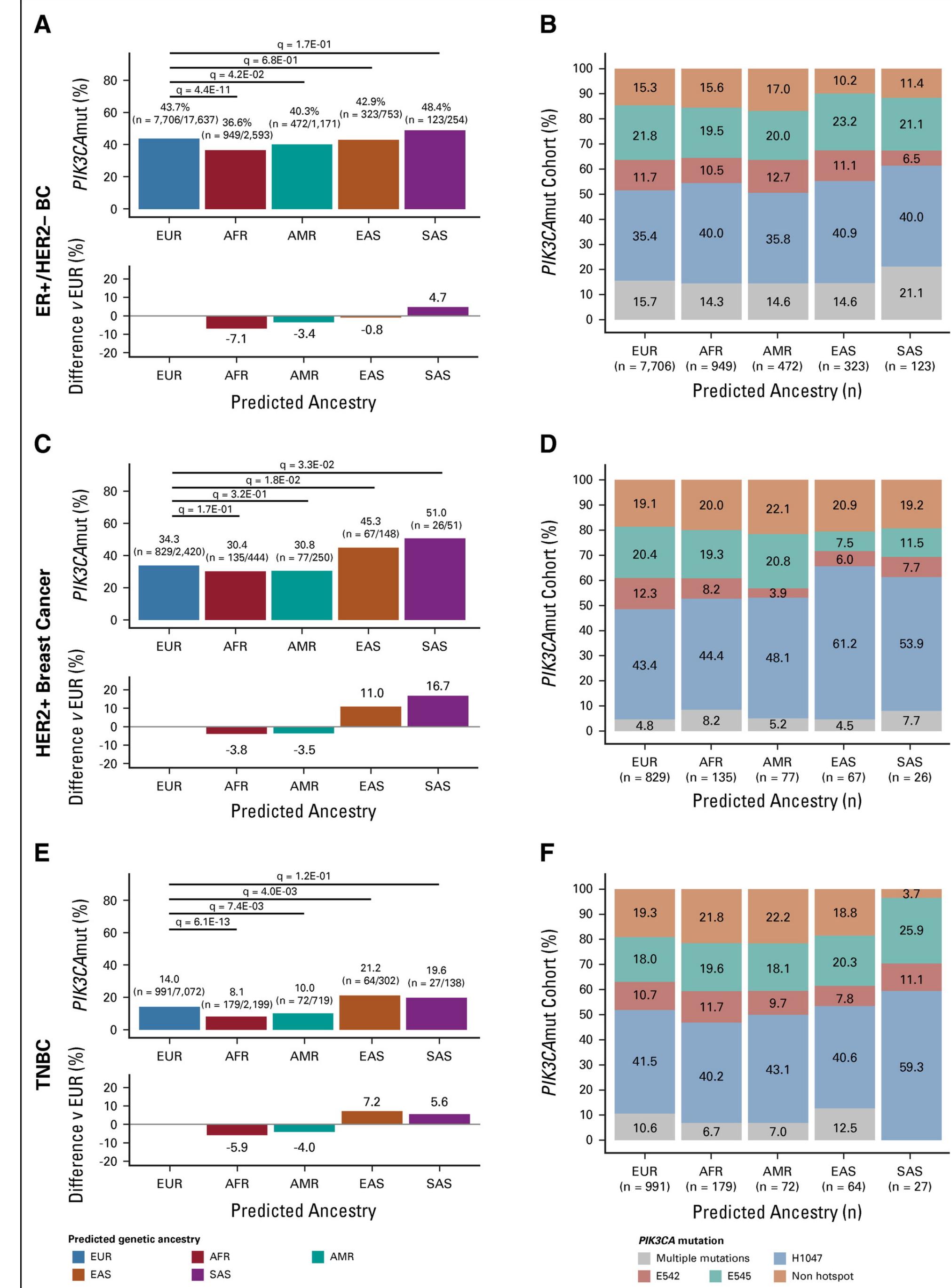
Patients of African ancestry are less likely to have tumors that harbor *PIK3CA* mutations compared with patients of European ancestry with estrogen receptor-positive/human epidermal growth factor receptor 2-negative BC and triple-negative BC. However, across predicted genetic ancestry groups, the most frequently observed *PIK3CA* mutations were generally similar and most, but not all, are able to be identified using commercially available polymerase chain reaction-based assays.

Relevance

This study highlights the need to systematically assess biomarker prevalence in historically under-represented populations to increase confidence in the generalizability and translatability of clinical trial outcomes to the population at large.

Why is population representation needed?

- types and prevalences of somatic variations may vary on different ancestral backgrounds
- relevant e.g. for design of variant panels, drug selection and clinical trial statistics



Limited Population Diversity in Cancer Studies

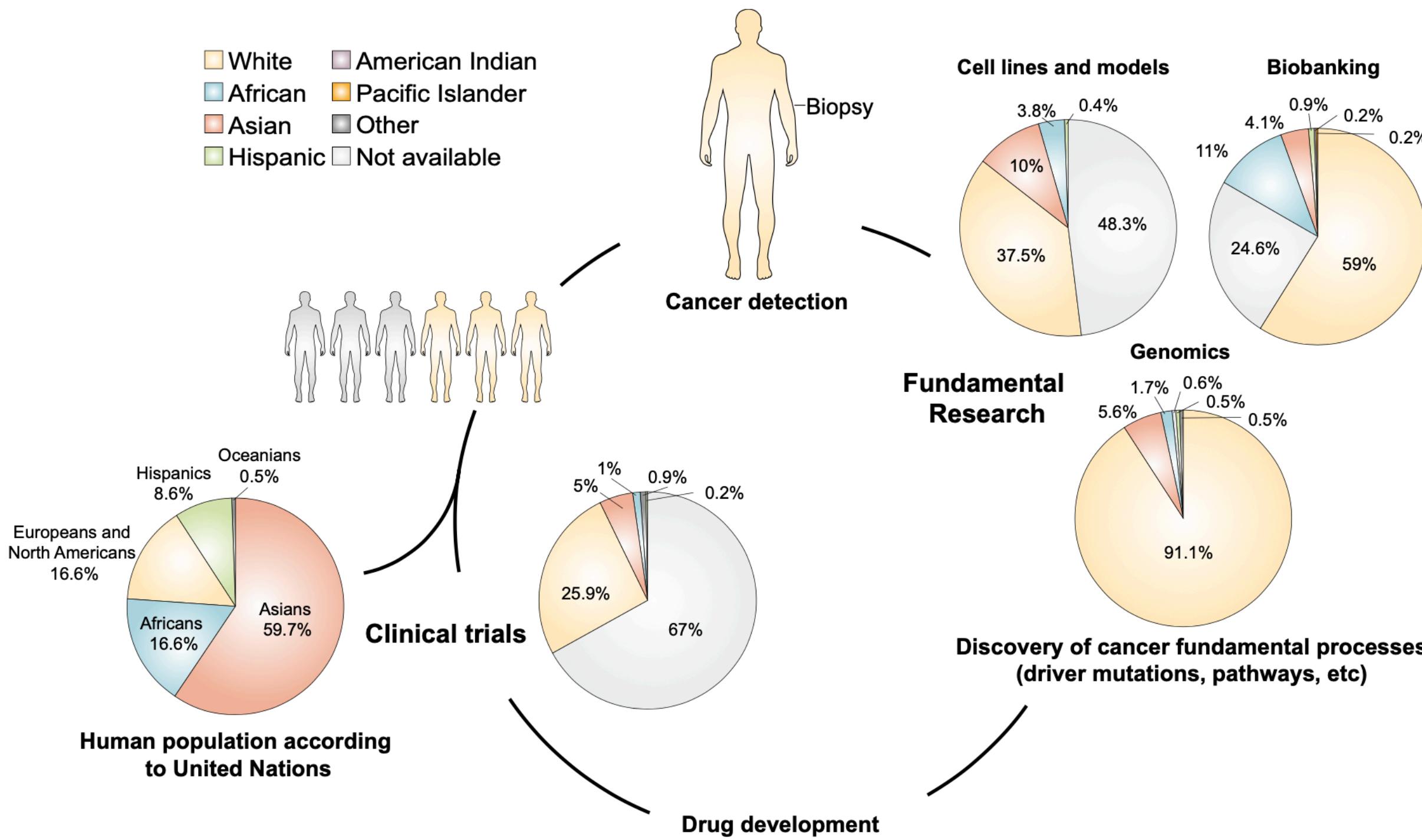


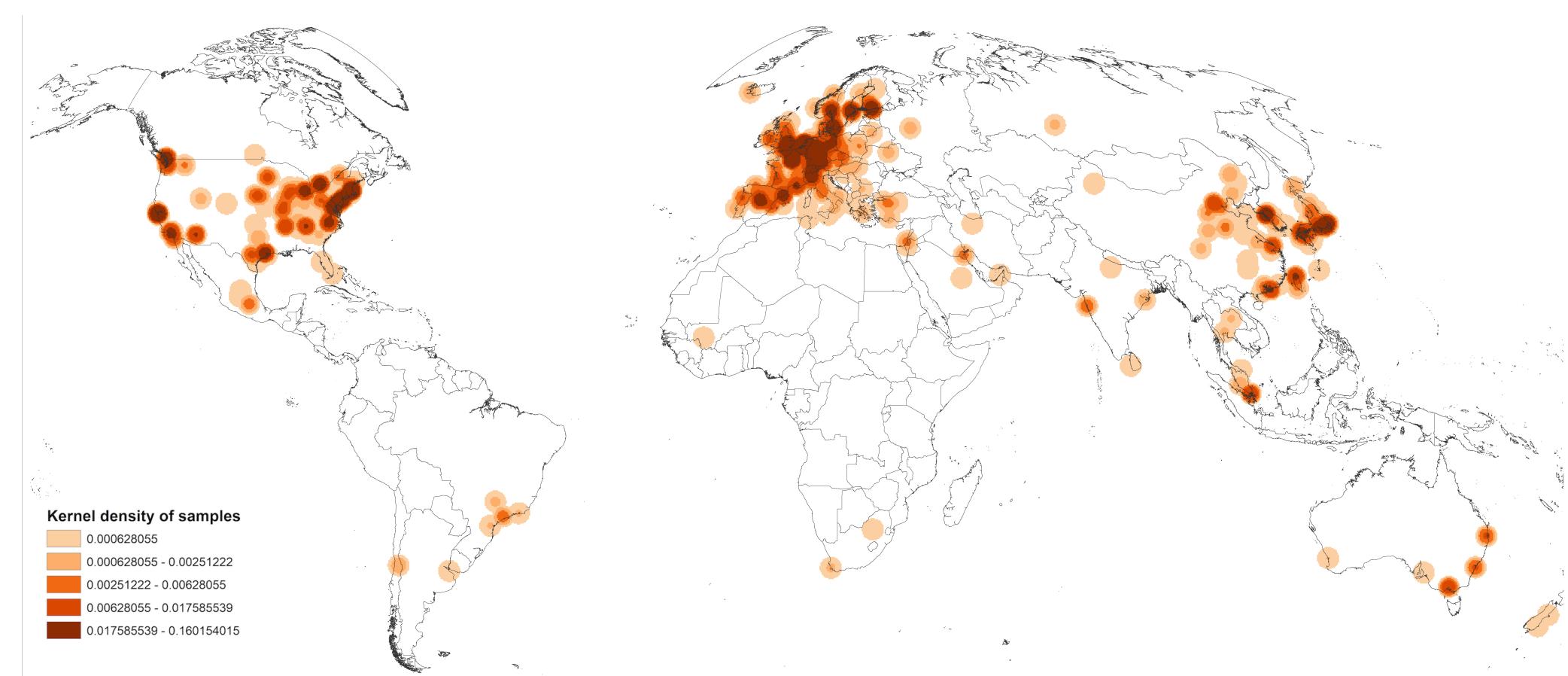
Figure 1. Racial/Ethnic disparities in cancer research. Racial/ethnic inclusion was studied in several aspects of oncological research, from cell lines and patient-derived xenografts to biobanking, genomics and clinical trials.

Guerrero S, López-Cortés A, Indacochea A, et al. Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. *Sci Rep.* 2018;8(1):13978.

Publication Landscape of Cancer CNV Profiling

Publication statistics for cancer genome screening studies. The graphic shows our assessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

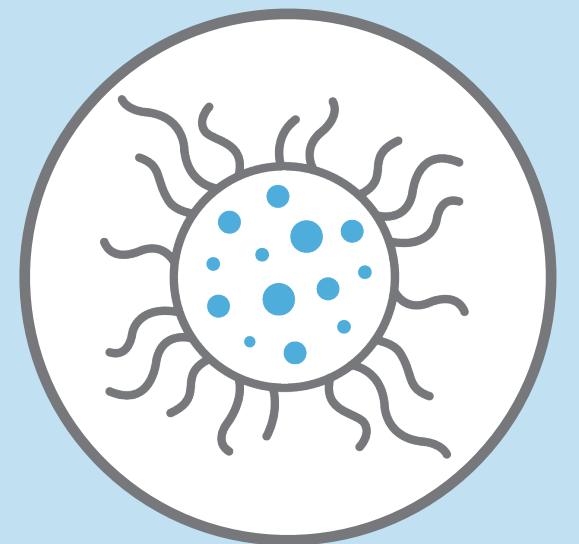
For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.



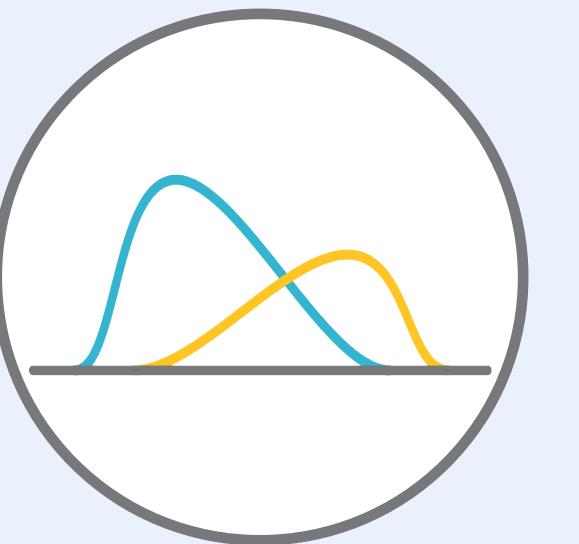
Global Genomic Data Sharing Can...



Global Alliance
for Genomics & Health



Demonstrate
patterns in health
& disease



Increase statistical
significance of
analyses



Lead to
“stronger” variant
interpretations



Increase
accurate
diagnosis



Advance
precision
medicine

200+ Genomic Data Initiatives Globally

Clinical/Genomic
Medicine



Research



National



Cohorts



WHOLE GENOME SEQUENCING DATA ON 200,000 UK BIOBANK PARTICIPANTS ARE NOW AVAILABLE FOR RESEARCH USE



This dataset represents the
world's largest single release of
Whole Genome Sequencing data

5 PETABYTES OF WGS DATA

When combined with the extensive amount of lifestyle,
biochemical and health outcome data already held for
the participants in UK Biobank, it will enable researchers
to better understand the role of genetics for health
outcomes and to advance drug discovery and development

How Many Genomes?



RESEARCH



HEALTHCARE

60M individuals
132.5M sequences



CLINICAL TRIALS

2.7-3M individuals



COHORTS

140M individuals

Direct to Consumer DNA Analyses

Population Background, Family Trees, Traits & Disease Risks...

MyHeritage

Enorme Ersparnisse
Letzte Chance DNA-Weihnachtsaktion
Nur 39 CHF 89 CHF

Jetzt bestellen
Angebot endet bald
KOSTENLOSER Versand bei der Bestellung von mind. 2 Kits

Starten Sie Ihre Testversion
Jacqueline 100%
European 50.1%
British & Irish 39.7%
French & German 7.0%
Broadly Northwestern European 3.2%
Scandinavian 0.2%
East Asian & Indigenous American 49.9%
Vietnamese 46.3%
Indonesian, Thai & Myanma 1.5%
Chinese 0.5%

Entdecken Sie Ihre Wurzeln
Erweitern Sie Ihren Stammbaum, entdecken Sie Verwandte, und durchforsten Sie historische Dokumente mit einer 14-tägigen KOSTENLOSEN Testversion

23andMe

YOUR DNA

Behind every data point is a human being.

We heart DNA.
Deciphering the human genome is the most exciting scientific discovery of our lifetime.

And you should be able to access, understand and benefit from the endlessly interesting and diverse things your genetics can tell you. How you choose to explore your DNA is up to you. With that said, we'd like to be the first to say, "Welcome to you®."

AncestryDNA

What would you like to learn about your family history?
Select all that apply

Details about my ancestors
My origins
I'm not really sure what I can discover on Ancestry

Skip Next
Dismiss

ing the generations together with a gift from Ancestry®.

HOLIDAY SALE

Inherited from Parent 1
Inherited from Parent 2

*Ends 31 Dec 2022. Terms apply.
Pricing for U.S. customers only.

By the numbers

Year	Number	Description
2006	12M+	The year we set out to make DNA more accessible and meaningful for all
2006	55+	The number of health reports that meet FDA requirements

The vision for genomic research: **Federation** of data





Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.



Enabling genomic data sharing for the benefit of human health

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a **human rights framework**



**Genomic Data
Toolkit**



**Regulatory & Ethics
Toolkit**



**Data Security
Toolkit**



[VIEW OUR LEADERSHIP](#)

[MORE ABOUT US](#)

[BECOME A MEMBER](#)

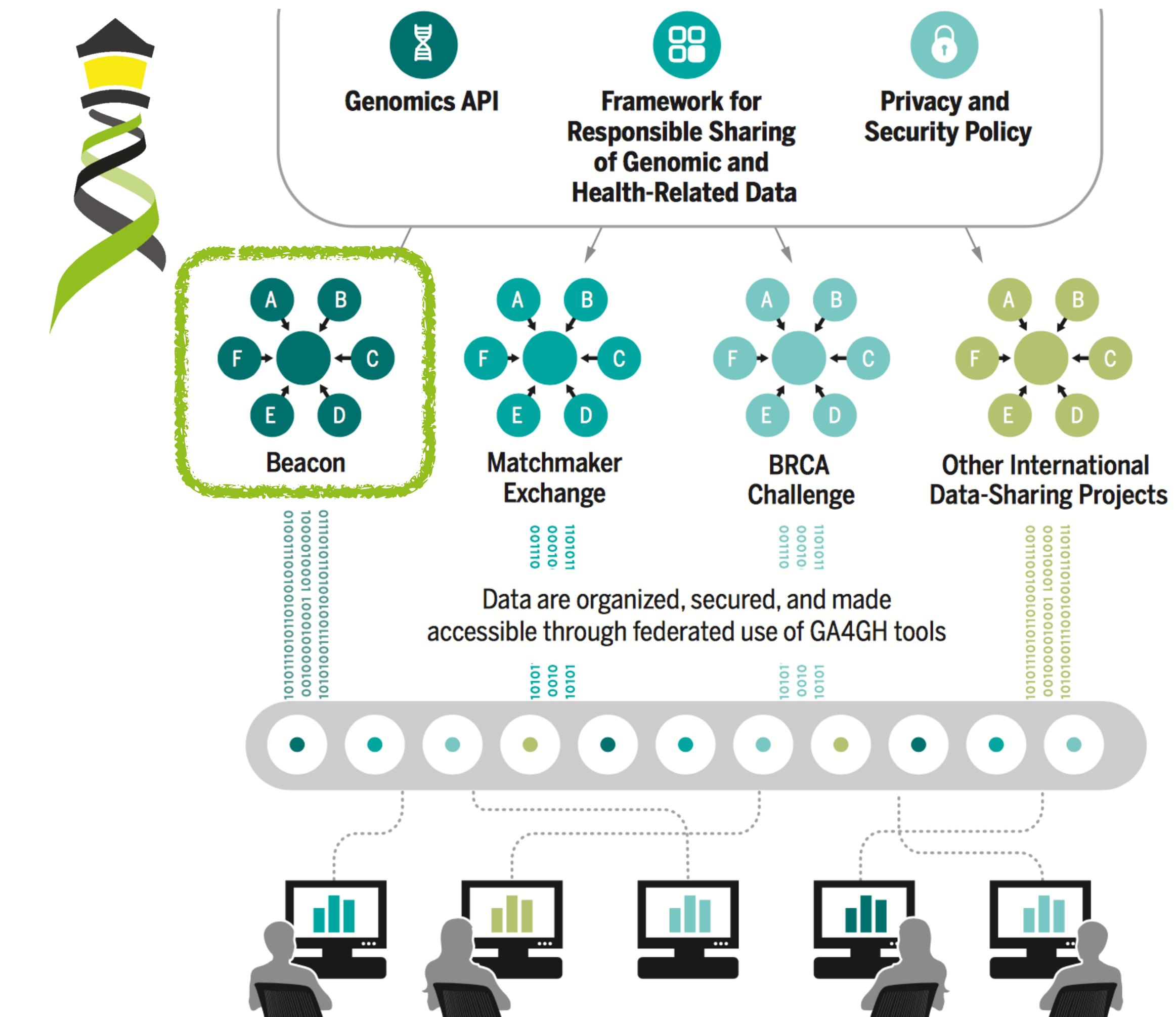


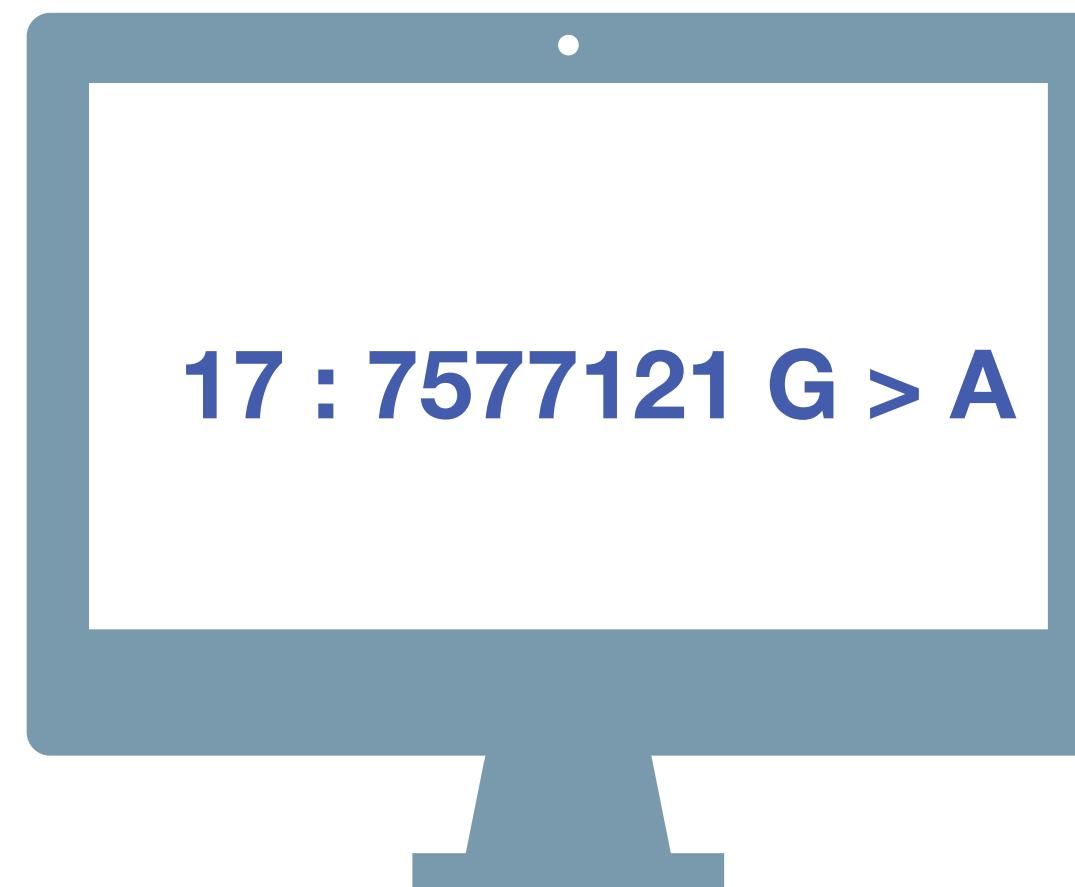
GENOMICS

A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems

A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.

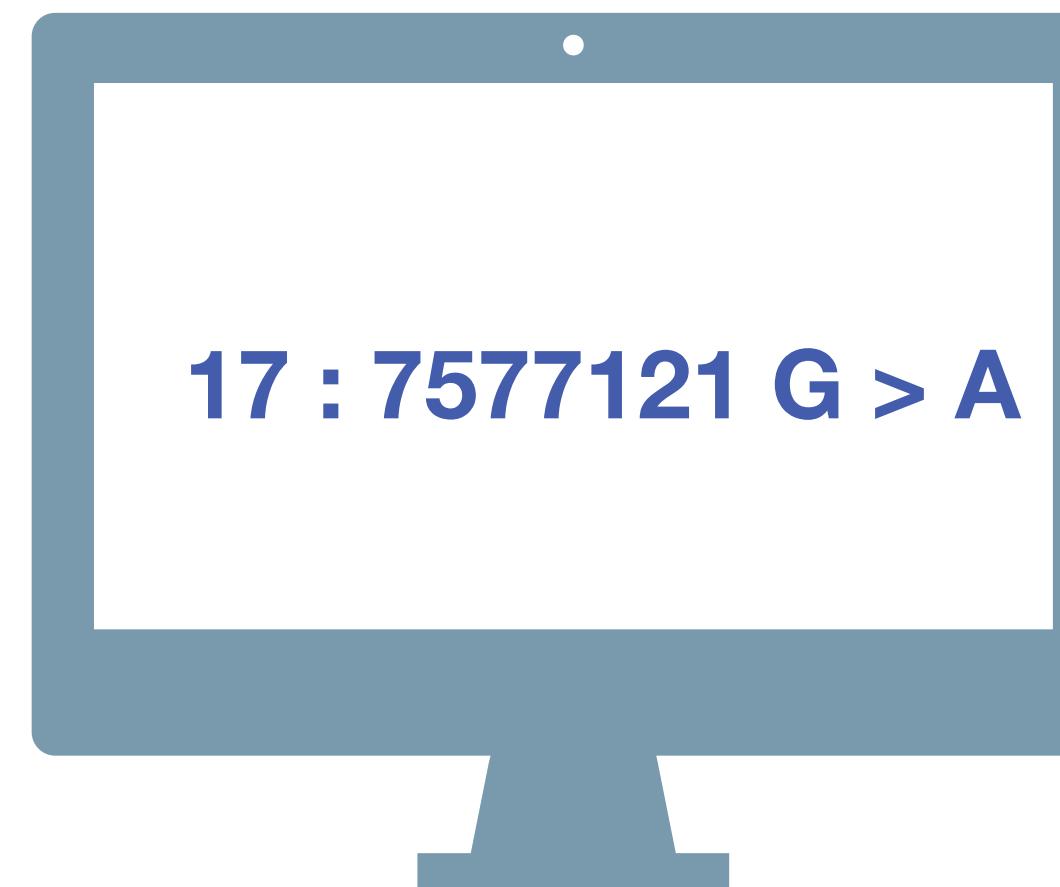




Beacon

A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections

YES | NO | \0



Have you seen this variant?
It came up in my patient
and we don't know if this is
a common SNP or worth
following up.

A Beacon network federates
genome variant queries
across databases that
support the **Beacon API**

Here: The variant has
been found in **few**
resources, and those
are from **disease**
specific **collections**.

Introduction

... I proposed a challenge application for all those wishing to seriously engage in **international** data sharing for human genomics. ...

1. Provide a public web service
2. Which accepts a query of the form “Do you have any genomes with an “A” at position 100,735 on chromosome 3?”
3. And responds with one of “Yes” or “No” ...

“Beacon” because ... people have been scanning the universe of human research for **signs of willing participants in far reaching data sharing**, but ... it has remained a dark and quiet place. The hope of this challenge is to 1) **trigger the issues** blocking groups ... in way that isn’t masked by the ... complexities of the science, fully functional interfaces, and real issues of privacy, and to 2) in **short order** ... see **real beacons of measurable signal** ... from **at least some sites** ... Once your “GABeacon” is shining, you can start to take the **next steps to add functionality** to it, and **finding the other groups** ... following their GABeacons.

Utility

Some have argued that this simple example is not “useful” so nobody would build it. Of course it is not the first priority for this application to be scientifically useful. ...intended to provide a **low bar for the first step of real ... engagement**. ... there is some utility in ...locating a rare allele in your data, ... not zero.

A number of more useful first versions have been suggested.

1. Provide **frequencies of all alleles** at that point
2. Ask for all alleles seen in a gene **region** (and more elaborate versions of this)
3. Other more complicated queries

“I would personally recommend all those be held for version 2, when the beacon becomes a service.”
Jim Ostell, 2014

Implementation

1. Specifying the chromosome ... The interface needs to specify the **accession.version** of a chromosome, or **build number**...
2. Return values ... right to **refuse** to answer without it being an error ... DOS **attack** ... or because ...especially **sensitive**...
3. Real time response ... Some sites suggest that it would be necessary to have a “**phone home**” **response** ...

Beacon Project in 2016

An open web service that tests the willingness of international sites to share genetic data.



Beacon Network

Search Beacons

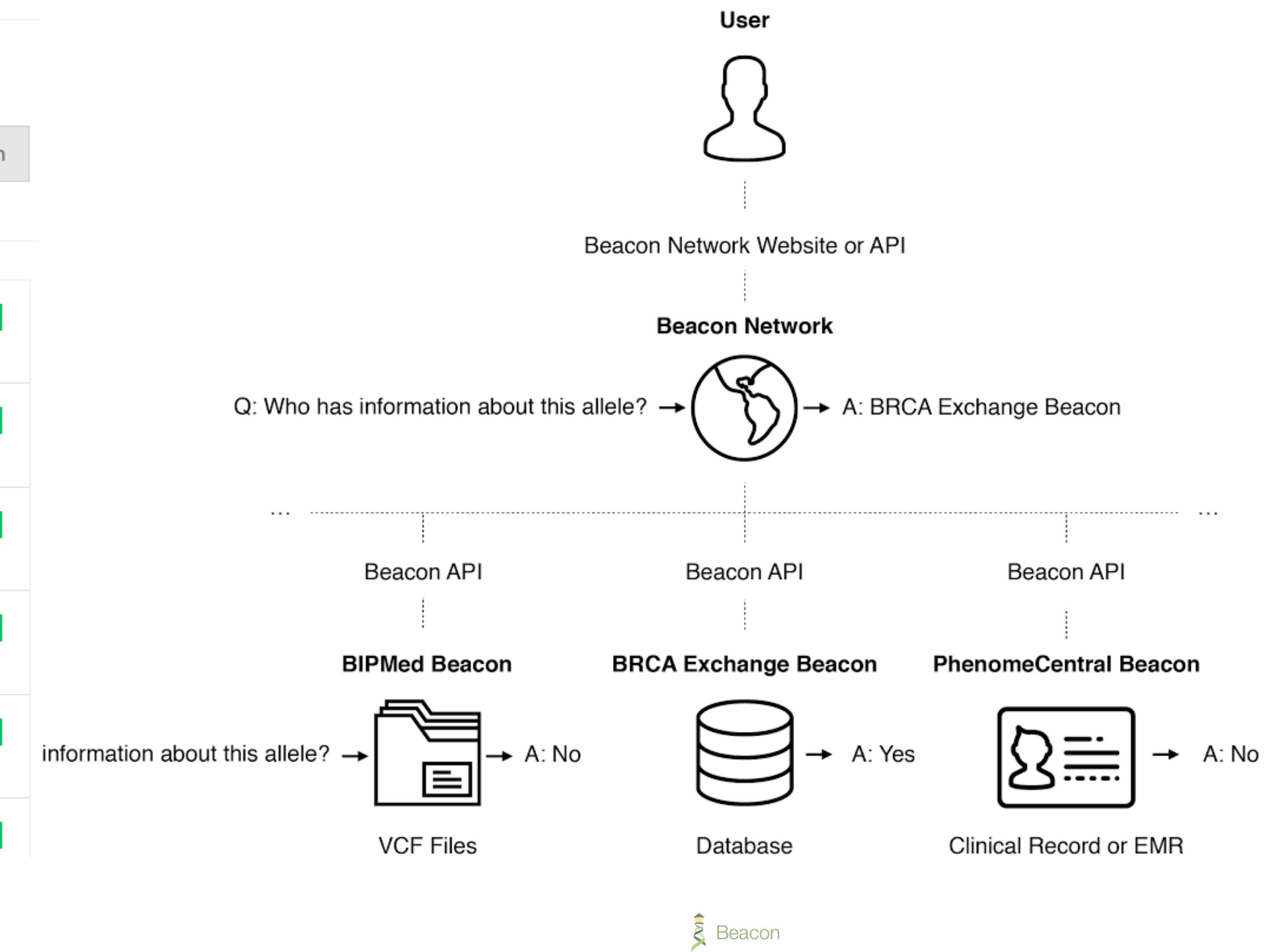
Search all beacons for allele

GRCh37 ▾ 10:118969015 C / CT Search

Response All None
 Found 16
 Not Found 27
 Not Applicable 22

Organization All None
 AMPLab, UC Berkeley
 BGI
 BioReference Laborato...
 Brazilian Initiative on ...
 BRCA Exchange
 Broad Institute
 Centre for Genomic R...
 Centro Nacional de A...
 Curoverse
 EMBL European Bio...
 Global Alliance for G...
 Google
 Institute for Systems ...
 Instituto Nacional de ...

BioReference	Hosted by BioReference Laboratories	Found
Catalogue of Somatic Mutations in Cancer	Hosted by Wellcome Trust Sanger Institute	Found
Cell Lines	Hosted by Wellcome Trust Sanger Institute	Found
Conglomerate	Hosted by Global Alliance for Genomics and Health	Found
COSMIC	Hosted by Wellcome Trust Sanger Institute	Found
dbGaP: Combined GRU Catalog and NHLBI Exome Seq...		Found



35+

Organizations

90+

Beacons

200+

Datasets

100K+

Releases
Individuals

Date	Tag	Title
2016-01-24	v0.4.0	Beacon
2016-05-31	v0.3.0	Beacon



Genome Beacons Compromise Security?

Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals in an anonymized genomic data collection

Stanford researchers identify potential security hole in genomic data-sharing network

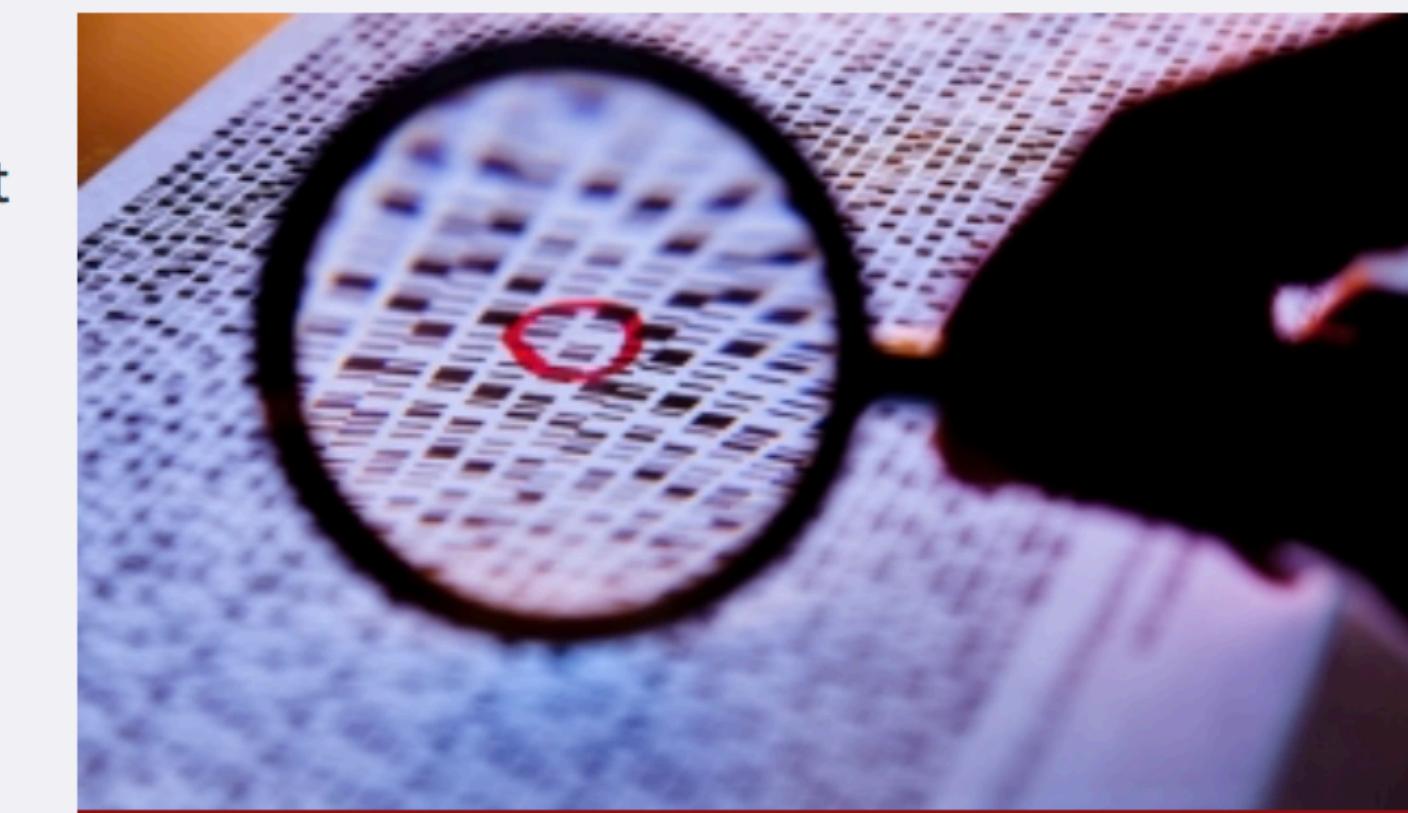
Hackers with access to a person's genome might find out if that genome is in an international network of disease databases.

OCT 29
2015

Sharing genomic information among researchers is critical to the advance of biomedical research. Yet genomic data contains identifiable information and, in the wrong hands, poses a risk to individual privacy. If someone had access to your genome sequence — either directly from your saliva or other tissues, or from a popular genomic information service — they could check to see if you appear in a database of people with certain medical conditions, such as heart disease, lung cancer or autism.

Work by a pair of researchers at the [Stanford University School of Medicine](#) makes that genomic data more secure. [Suyash Shringarpure](#), PhD, a postdoctoral scholar in genetics, and [Carlos Bustamante](#), PhD, a professor of genetics, have demonstrated a technique for hacking a network of global genomic databases and how to prevent it. They are working with investigators from the Global Alliance for Genomics and Health on implementing preventive measures.

The work, published Oct. 29 in *The American Journal of Human Genetics*, also bears importantly on the larger question of how to analyze mixtures of genomes, such as those from different people at a crime scene.



Stanford researchers are working with the Global Alliance for Genomics and Health to make genomic information in the Beacon Project more secure.
Science photo/Shutterstock

IDENTIFICATION OF INDIVIDUALS FROM MIXED COLLECTIONS USING RARE ALLELES

Privacy Risks from Genomic Data-Sharing Beacons

Suyash S. Shringarpure^{1,*} and Carlos D. Bustamante^{1,*}

The human genetics community needs robust protocols that enable secure sharing of genomic data from participants in genetic research. Beacons are web servers that answer allele-presence queries—such as “Do you have a genome that has a specific nucleotide (e.g., A) at a specific genomic position (e.g., position 11,272 on chromosome 1)?”—with either “yes” or “no.” Here, we show that individuals in a beacon are susceptible to re-identification even if the only data shared include presence or absence information about alleles in a beacon. Specifically, we propose a likelihood-ratio test of whether a given individual is present in a given genetic beacon. Our test is not dependent on allele frequencies and is the most powerful test for a specified false-positive rate. Through simulations, we showed that in a beacon with 1,000 individuals, re-identification is possible with just 5,000 queries. Relatives can also be identified in the beacon. Re-identification is possible even in the presence of sequencing errors and variant-calling differences. In a beacon constructed with 65 European individuals from the 1000 Genomes Project, we demonstrated that it is possible to detect membership in the beacon with just 250 SNPs. With just 1,000 SNP queries, we were able to detect the presence of an individual genome from the Personal Genome Project in an existing beacon. Our results show that beacons can disclose membership and implied phenotypic information about participants and do not protect privacy *a priori*. We discuss risk mitigation through policies and standards such as not allowing anonymous pings of genetic beacons and requiring minimum beacon sizes.

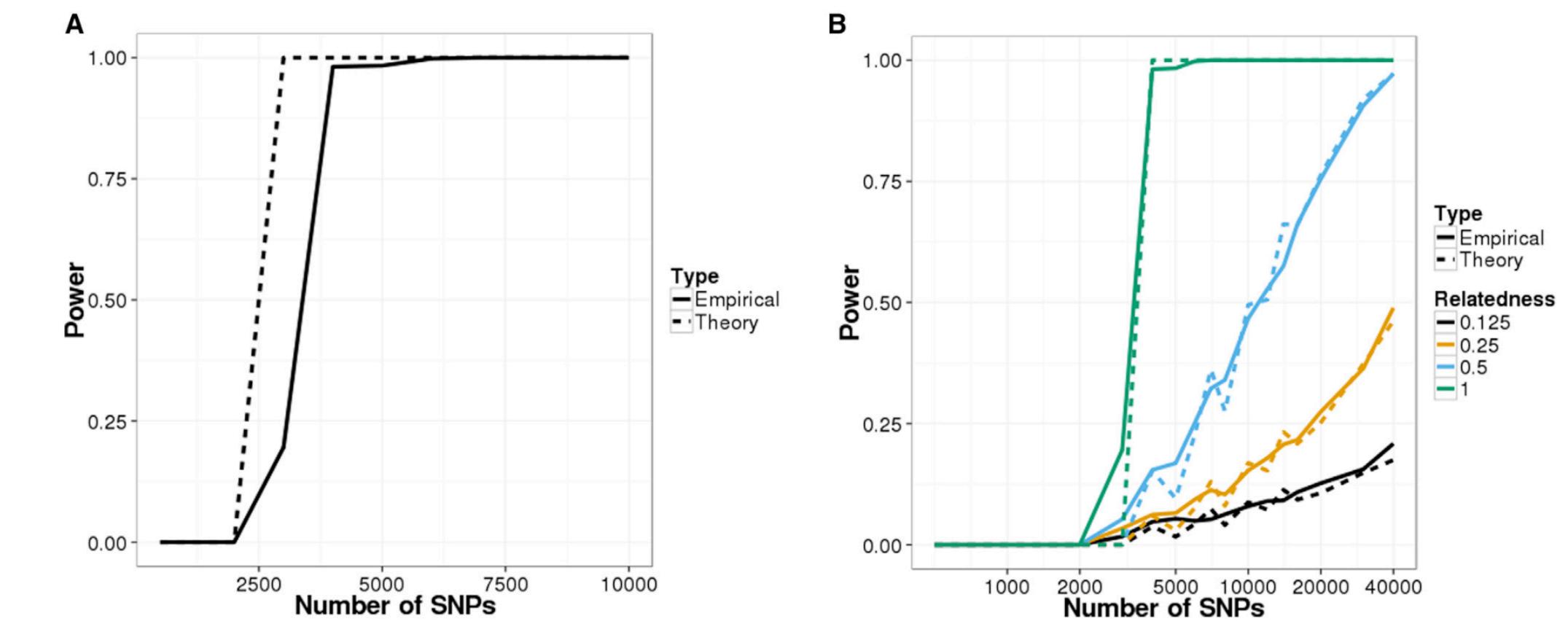


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- ▶ rare allelic variants can be used to identify an individual (or her relatives) in a genome collection without having access to individual datasets
- ▶ however, such an approach requires previous knowledge about the individual's SNPs

Information Leakage from Functional Genomics Data

- many research studies contain "functional" genomics data, e.g. from expression analyses
- such (anonymized) data may have lower protection levels than data from dedicated genotyping studies
- with a non-noisy genome of interest, attackers can generate linkage scores to identify the best match to the genomic profile

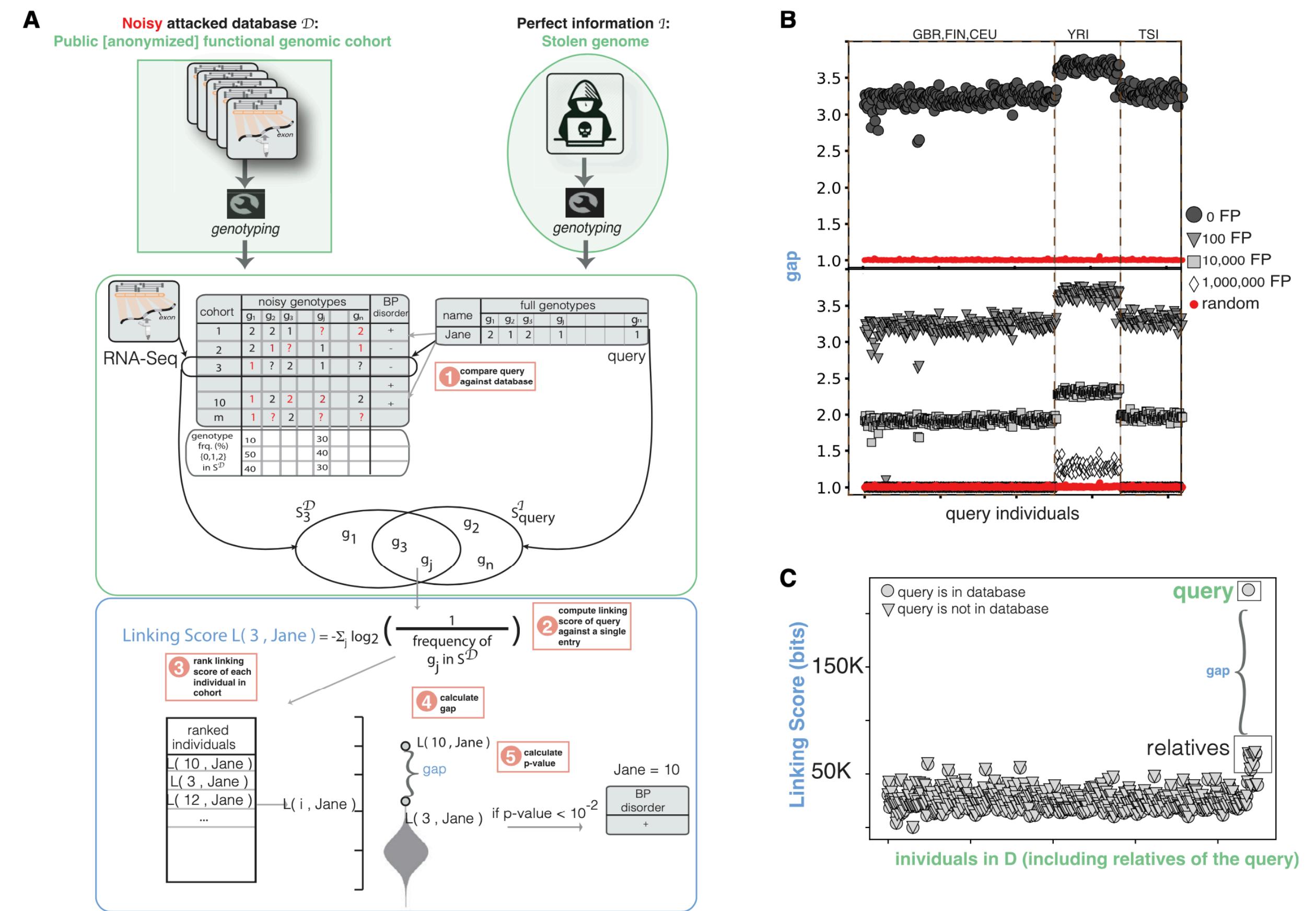


Figure 1. Functional Genomics Data De-anonymization Scheme with Perfect Genomes

(A) Anonymized functional genomics data from a cohort of individuals can be seen as a database \mathcal{D} to be attacked, which contains functional genomics reads and phenotypes for every individual in the cohort. The perfect information I about an individual can be the genome of an individual. After obtaining genotypes from the functional genomics reads, the attacker scores each individual in the cohort based on the overlapping genotypes between the known individual's genome and the noisy genotypes called from functional genomics. These scores are then ranked and the top-ranked individual in the cohort is selected as the known individual. See also [Figure S1](#).

(B) gap values for the 1000 Genomes Project individuals in the gEUVADIS RNA-seq cohort. Red circles are the gap values obtained by linking a random set of genotypes to the RNA-seq panel. gap values are also shown after adding false-positive genotypes to the genotype set of each individual in the database.

(C) The linking scores for each individual in the functional genomics cohort after the addition of genetically related individuals to the query, with and without the query individual present in the database.

But genotyping itself is for professional labs, right?

Rapid re-identification of human samples

...

We developed a rapid, inexpensive, and portable strategy to re-identify human DNA using the MinION. Our strategy requires only ~60 min preparation and 5-30 minutes of MinION sequencing, works with low input DNA, and enables familial searches using Direct-to-Consumer genomic reference datasets. This method can be implemented in a variety of fields:



Forensics

Identification of abandoned material using DNA fingerprinting is a common practice. The main challenge currently being: time. Our method allows rapid sample preparation at the crime scene (see movie). We envision that the method can be adopted in the field for rapid checks, after a mass disaster, and can be adopted in border control to fight human trafficking.



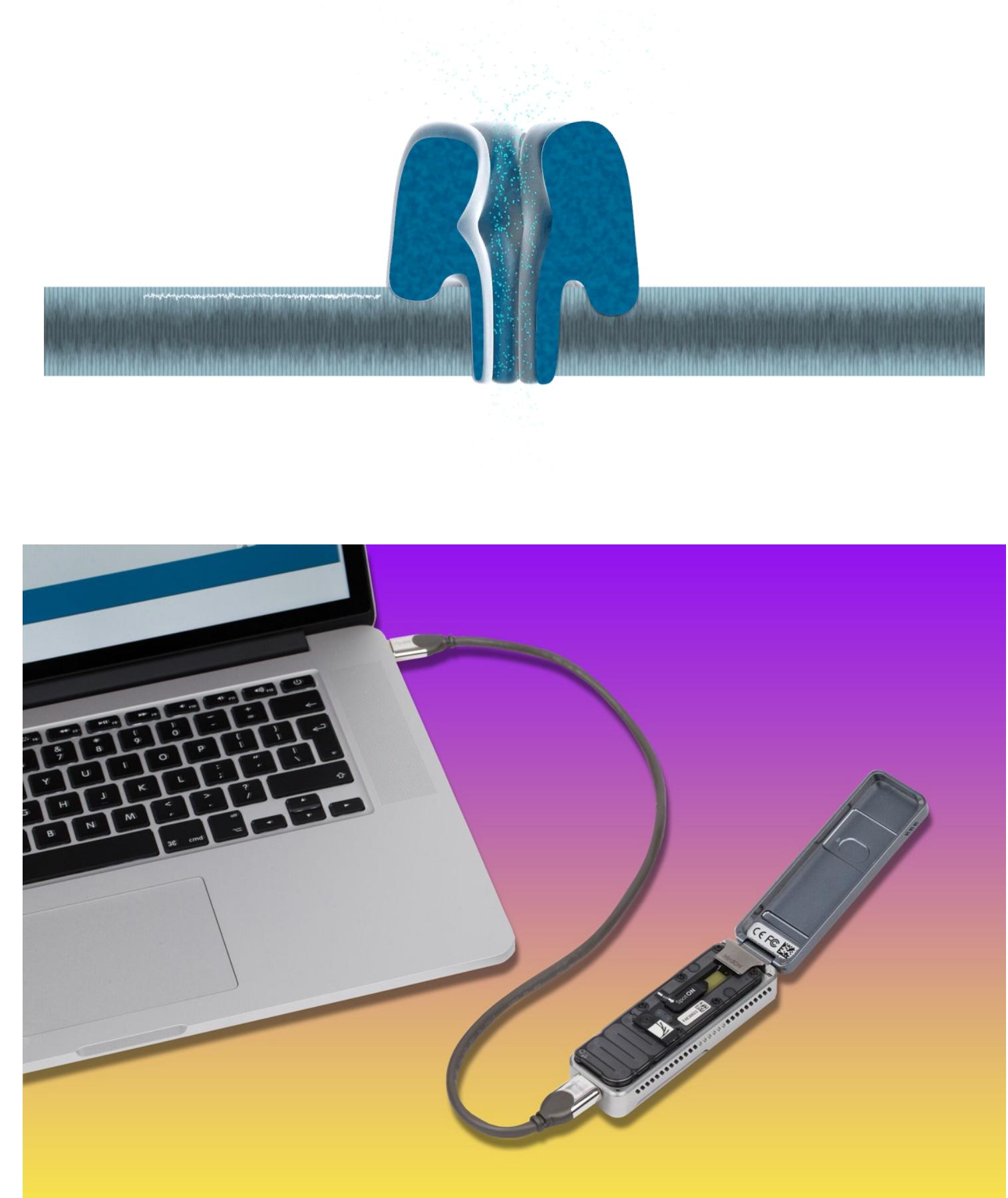
Clinic

Clinics process many samples, either for analysis or, for example, organ donations. These samples are DNA fingerprinted to prevent sample mix-up mistakes. Our method can be implemented in the clinic for rapid sanity-check of all incoming samples.



Cell line identification

Cross contamination of cell lines in science is a major problem. It results in unrepeatable data, and clinical trials based on inaccurate findings. This problem costs billions of dollars per year. We envision labs can adopt our identification method to ensure the purity of the cell line, and detect contamination.



The MinION (Oxford Nanopore)

Source: Sophie Zaaijer

<https://medium.com/neodotlife/nanopore-6443c81d76d3>

DEMOCRATIZING DNA FINGERPRINTING

Sophie Zaaijer, Assaf Gordon, Robert Piccone, Daniel Speyer, Yaniv Erlich, 2016
ddf.teamerlich.org



DNA sequencing for identification/fingerprinting soon “commodity” technology (in contrast with technological/data challenges in “precision medicine”)

MinION by Oxford Nanopore Technologies



The MinION is the smallest DNA sequencer currently around. It's the size of a Mars bar, and can be simply plugged into a laptop with a USB3.0 port.

For more information about the MinION please click:
[Oxford Nanopore Technologies](#)

Bento Lab



The Bento lab is a miniature lab with a centrifuge, thermocycler and a electrophoresis compartment.

For more information about the Bento-lab please click:
[Bento Lab](#)



Data can be loaded into the person ID pipeline
matches inferred between 3-30 minutes

Typical Data Scopes in Genomics (Research) Collections

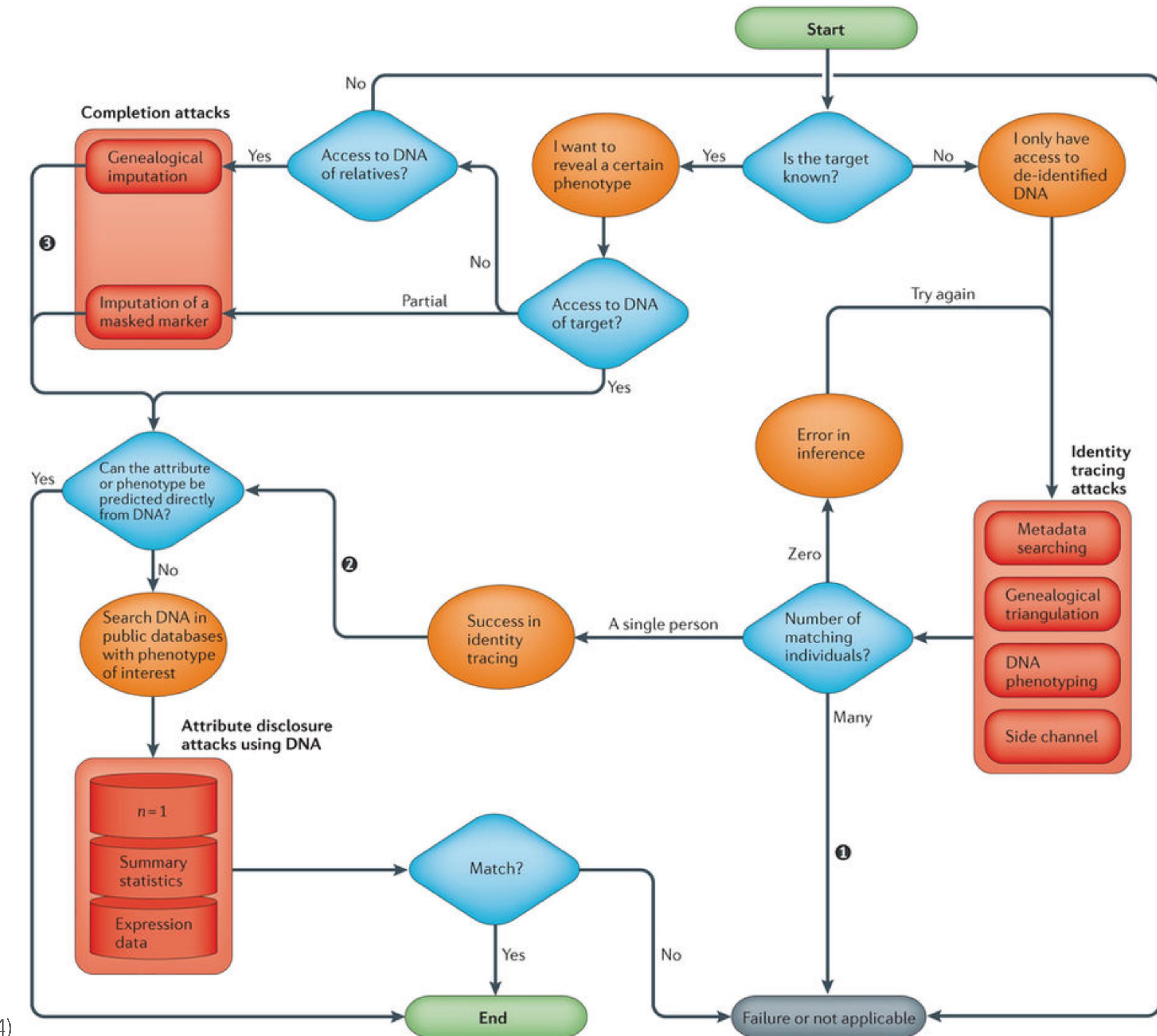
Biomedical and procedural "Meta"data types

- Diagnostic classification
 - mapping text-based cancer diagnoses to standard classification systems
- Provenance data
 - store identifier-based pointers
 - geographic attribution (individual, biosample, experiment)
- Clinical information
 - **core set** of typical cancer study values:
 - ➡ stage, grade, followup time, survival status, genomic sex, age at diagnosis
 - balance between annotation effort and expected usability

Routes for breaching and protecting genetic privacy

The map contrasts different scenarios, such as identifying de-identified genetic data sets, revealing an attribute from genetic data and unmasking of data. It also shows the interdependencies between the techniques and suggests potential routes to exploit further information after the completion of one attack. There are several simplifying assumptions (black circles).

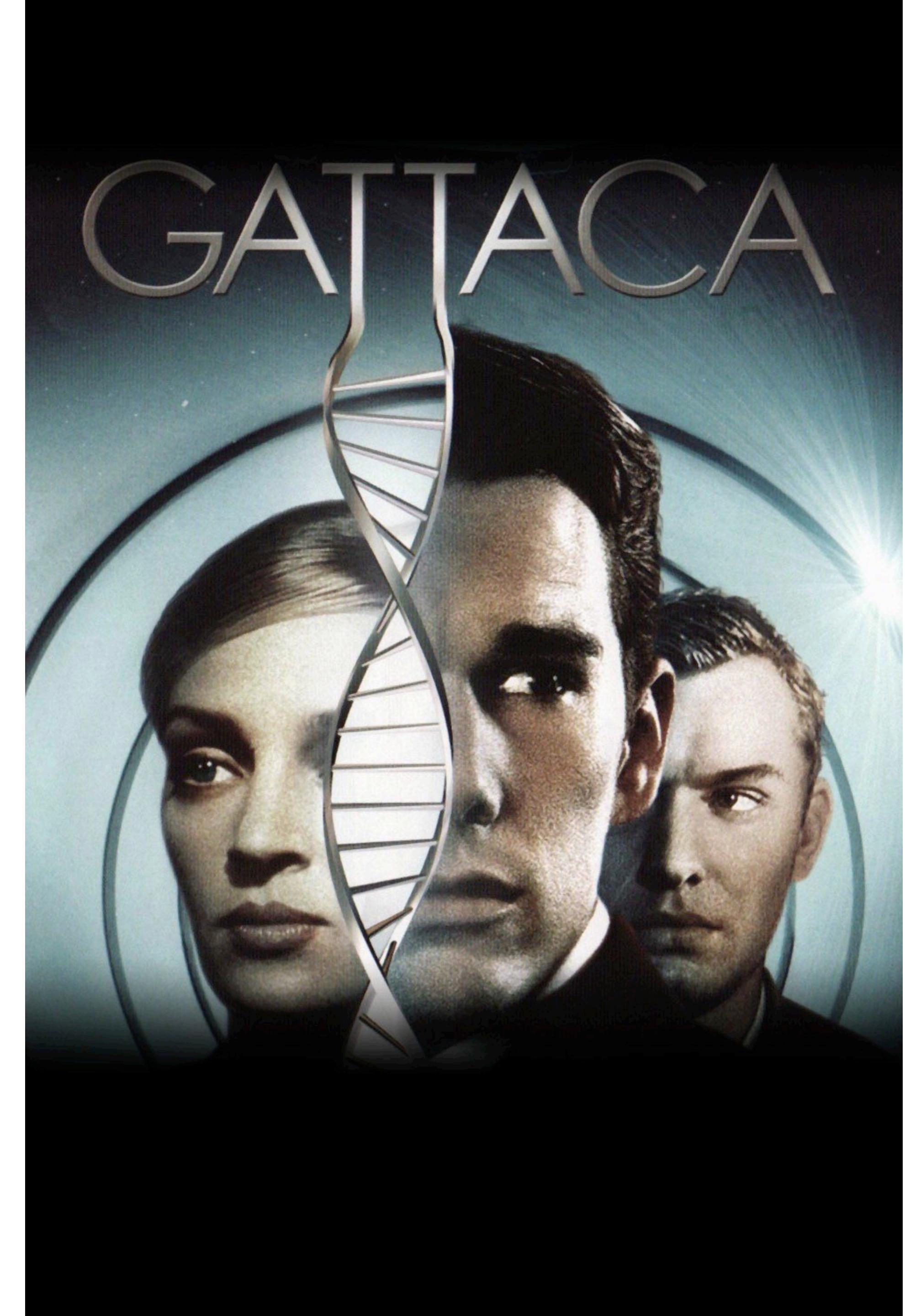
In certain scenarios (such as insurance decisions), uncertainty about the target's identity within a small group of people could still be considered a success (assumption 1). For certain privacy harms (such as surveillance), identity tracing can be considered a success and the end point of the process (assumption 2). The complete DNA sequence is not always necessary (assumption 3).



“We’re an information economy. They teach you that in school. What they don’t tell you is that it’s impossible to move, to live, to operate at any level without leaving traces, bits, seemingly meaningless fragments of personal information. Fragments that can be retrieved, amplified . . .”

–William Gibson in "Johnny Mnemonic" (1986)

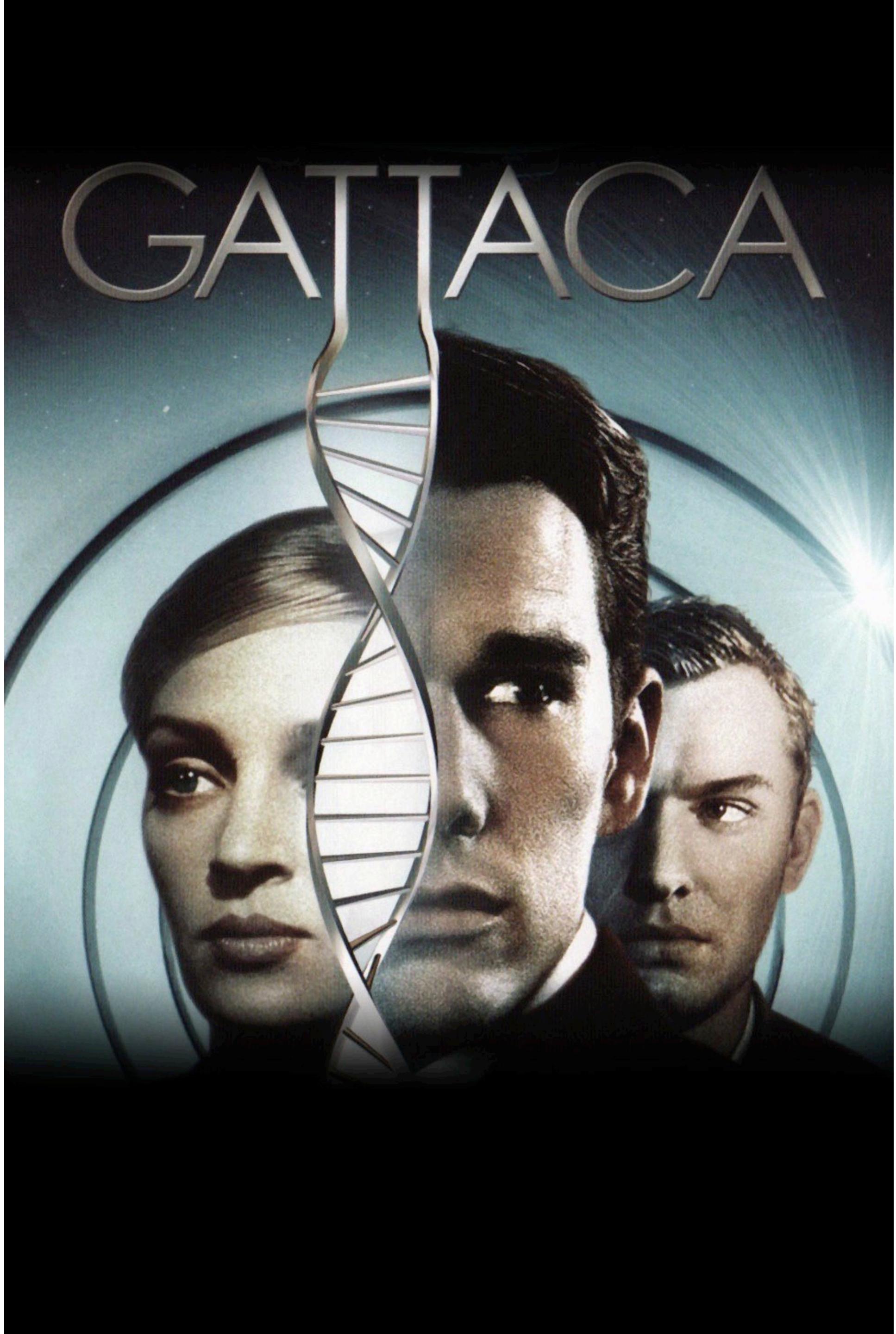
Genomes Privacy Society



Gattaca (1997)

A genetically inferior man assumes the identity of a superior one in order to pursue his lifelong dream of space travel.

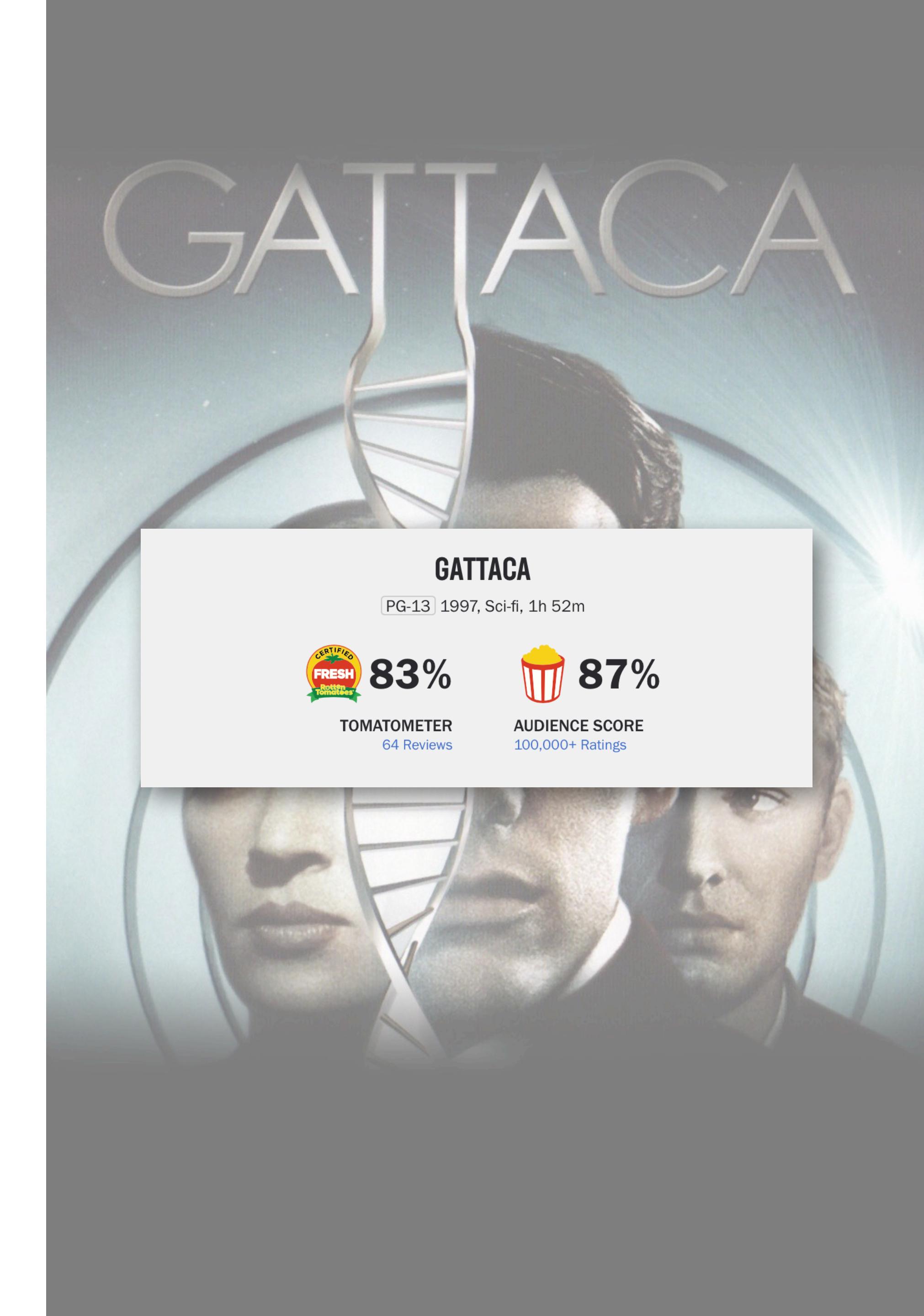
- genetic determinism
 - ▶ main character has been determined to be unsuitable for complex jobs based on genetic analysis
- genetic identification
 - ▶ the use of genetic sampling for personal identification is daily routine



Gattaca (1997)

A genetically inferior man assumes the identity of a superior one in order to pursue his lifelong dream of space travel.

- genetic determinism
 - ▶ main character has been determined to be unsuitable for complex jobs based on genetic analysis
- genetic identification
 - ▶ the use of genetic sampling for personal identification is daily routine



DTC Genomics

Direct-to-Consumer Genomic Testing

- family ancestry or genealogy
 - ▶ >7Mill customers in 2018 at ancestry.com
- DNA-based health & traits information
 - ▶ disease risk
 - ▶ carrier status
 - ▶ lifestyle information
- participation in large cohort studies



23andMe®

ancestry

MyHeritage

Think Before You Spit

You've always known you're unique. Now learn just how unique you are.

Whole Genome Sequencing is the only (and last) DNA screening that you need. This at-home screening gives you clinical-grade results and provides you with ultra-personalized insights that can help you live a healthier life.

[Get Sequenced](#)

Learn more about what you can learn with these reports:



Rare Disease Screen DNA Report: There are [more than 10,000](#) rare diseases, syndromes, conditions, and traits. Outsmart your DNA for yourself today and for your children tomorrow.

Medication and Drug Reaction: Discover through our pharmacogenomics analysis if you need to be more mindful of potential medication side effects or if you may be at risk of addiction to certain medications and illicit drugs.



Complete Genome Analysis: Learn all there is to learn about your DNA, from inherited traits and conditions, disease susceptibility, to ancestry (including mtDNA and Y-DNA analyses).

[Get Check out more DNA Reports](#)

Email Disclaimer Placeholder

[Health](#)[Ancestry](#)[Nutrition](#)[Fitness](#)[Beauty](#)[Lifestyle](#)[Children](#)[Art](#)[Bioinformatics](#)[Test Kits](#)[Featured](#)[Trending](#)[Free](#) Search

Marketplace: DNA Apps & DNA Reports

Health



Wellness and
Longevity
App MD

[\\$120](#)

Medication &
Drug Response
Complete Genome
Science

[\\$59](#)

Genetic
Detoxification
Test
GenelInformed

[\\$69](#)

Inflammation
DNA Wellness
Report
SelfDecode

[\\$49](#)

Rare Disease
Screen
Sequencing.com

[\\$90](#)

Carrier Status
Complete Genome
Science

[\\$29](#)

Cannabis DNA
Health Report
Strain Genie

[\\$29.99](#)

Healthcare Pro
App MD

[\\$140](#)

TBG Total
Wellness
Toolbox Genomics

[\\$119](#)

Disease Risk
Genetic Test
Report
Complete Genome
Science

[\\$59](#)

Genetic
Counseling
DNAVist

[\\$129](#)

Vitamin
Balance DNA
Report
Silverberry Genomix

[\\$4.99](#)

Right to Know?

Dealing with "non-actionable" genomic predictions

- diagnostic and direct to consumer genetic tests may provide risk predictions for disease susceptibility
- most will be non-deterministic, non-actionable, and usually be associated with a very low **absolute** risk - but heritable
- understanding such "prognostications" is challenging & potentially fraught with errors - and opens the door to services

VANITY FAIR

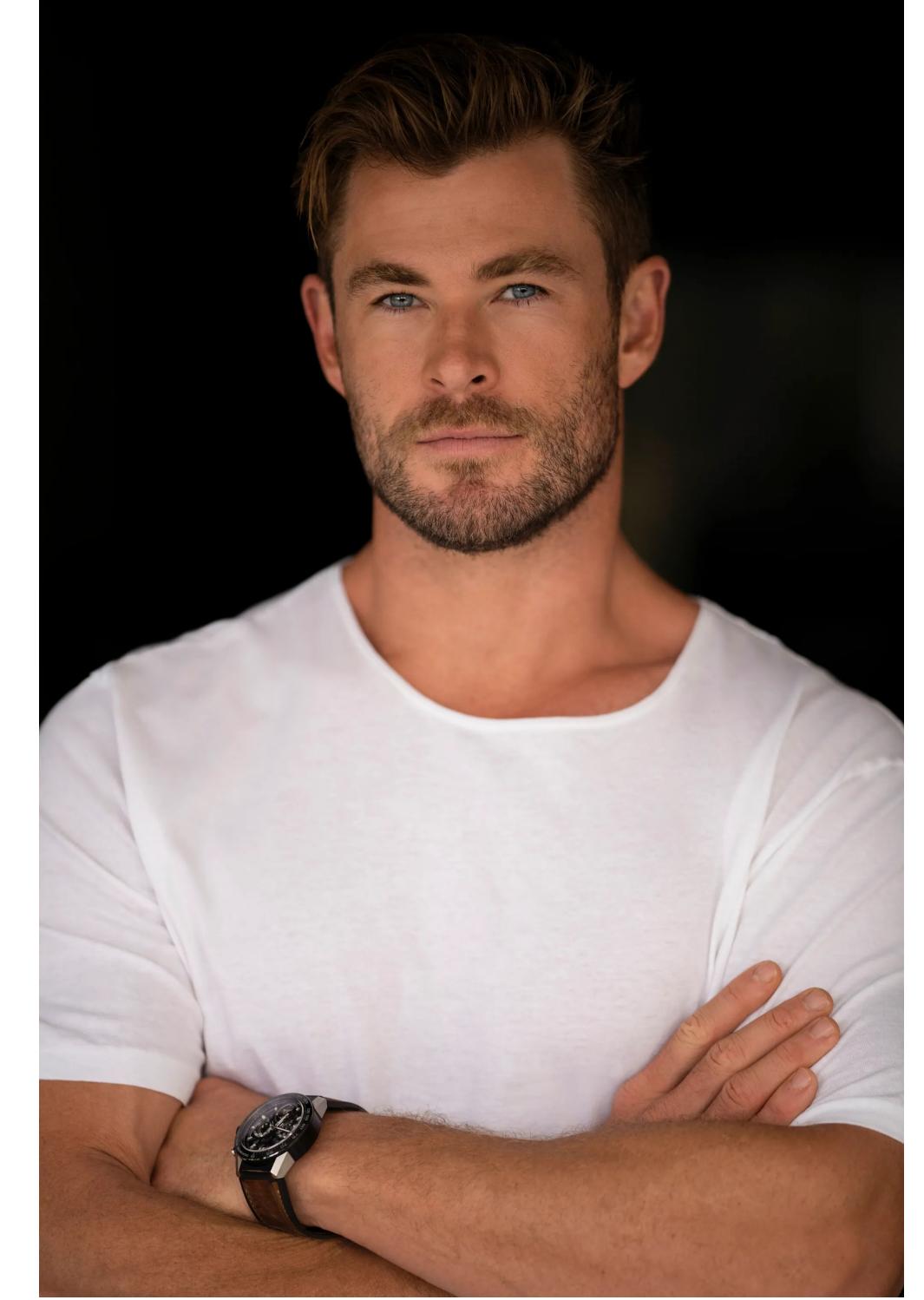
EXCLUSIVE

Chris Hemsworth Changed His Life After an Ominous Health Warning

In an exclusive sit-down with *Vanity Fair*, the actor discusses movies, the future of Thor, his businesses, fatherhood, and how a genetic predisposition for Alzheimer's alters everything.

BY ANTHONY BREZNICAN

NOVEMBER 17, 2022



...His makeup includes two copies of the gene APOE4, one from his mother, the other from his father, which studies have linked to an increased risk of Alzheimer's disease. One in four people carry a single copy of the gene, but only 2 to 3% of the population have both, according to a [2021 study](#) by the National Institutes of Health.

"For me, the positive of it was like, "Right, if I didn't know this [Alzheimer's] information, I wouldn't have made the changes I made." I just wasn't aware of any of it, so now I feel thankful that I have in my arsenal the sort of tools to best prepare myself and prevent things happening in that way."



Hi Michael,

Good news! We've discovered new DNA Matches for you.

- Commercial, "Direct to Customer" DNA analyses are provided through independent sites and such affiliated to genealogy services (MyHeritage, Ancestry.com, 23andMe...)
- Genealogy sites identify individuals with matching haplotype blocks & provide a prediction about degree of genetic relation
- Law enforcement agencies (and who else?!) can send individual SNP profiles (e.g. recovered from evidence many years after a crime) using a *Jane Doe* identity, to identify relatives of the suspect - **long range familial search**

Long-Range Familial Searches

Daily Journal

Helping Northeast Mississippi Grow!

We're donating a portion of every 1-year or 6-month subscription to Tupelo High Band Boosters!
842-2613 or djournal.com/subscribe
New home delivery subscriptions only | Offer ends June 30



SUBSCRIBE

ALL SEC Devaughn had never been a suspect until genetic genealogy put police on his trail several months ago. Earlier this year, police sent the DNA profile to Parabon, a private genetics company, to compare the suspect's DNA sample to a public genealogy DNA database looking for people with similar DNA profiles who might be kin to the suspect. That eventually led authorities to look at Devaughn.

Rienzi man charged with 1990 Starkville murder

By William Moore Daily Journal 15 hrs ago Comments

© Copyright 2018 Daily Journal, 1242 S Green St Tupelo, MS



The New York Times

How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect

Investigators used DNA from crime scenes that had been stored all these years and plugged the genetic profile of the suspected assailant into an online genealogy database. One such service, GEDmatch, said in a statement on Friday that law enforcement officials had used its database to crack the case. Officers found distant relatives of Mr. DeAngelo's and, despite his years of eluding the authorities, traced their DNA to his front door.

The New York Times, April 26, 2018

Attacks Associated With the Golden State Killer



Suspect in 1972 Murder Dies in Suicide Hours Before Conviction

Detectives used genetic genealogy to connect [REDACTED] to the killing of [REDACTED] outside Seattle. He was charged last year.



By Neil Vigdor

Published Nov. 9, 2020 Updated Nov. 11, 2020

The New York Times

"A man who eluded homicide investigators in Washington State for nearly 50 years — until a DNA match on a coffee cup cracked the cold case — died in a suicide on Monday just hours before a jury convicted him of murder, the authorities said. ... Investigators used genetic genealogy, a process that involved crosschecking DNA evidence — taken from a hiking boot worn by Ms. yyyy — with ancestry records to connect Mr. xxxx to the unsolved murder. ..."

In 2008, the samples were sent to the Washington State Patrol Crime Laboratory for DNA testing, but they did not return a match. ...

The breakthrough in the case came in 2018 when investigators, working with Parabon NanoLabs, were able to put together a family tree of possible suspects based on the semen sample found on the heel of the victim's hiking boot. The company uses DNA to help law enforcement agencies find genetic matches.

That's when investigators began their surveillance of Mr. xxxx, whom they followed to a nearby casino and from whom they retrieved a coffee cup that he had thrown in the garbage, the probable cause affidavit said. The DNA sample was an exact match to the semen found on Ms. yyyy's boot, the affidavit said."

Long-Range Familial Searches

Genealogy Sites Have Helped Identify Suspects. Now They've Helped Convict One.

A new forensic technique sailed through its first test in court, leading to a guilty verdict. But beyond the courtroom, a battle over privacy is intensifying.

By Heather Murphy

July 1, 2019

The New York Times

"... Genetic genealogy — in which DNA samples are used to find relatives of suspects, and eventually the suspects themselves — has redefined the cutting edge of forensic science, solving the type of cases that haunt detectives most: the killing of a schoolteacher 27 years ago, an assault on a 71-year-old church organ player, the rape and murder of dozens of California residents by a man who became known as the Golden State Killer.

But until a trial this month in the 1987 murder of a young Canadian couple, it had never been tested in court. Whether genetic genealogy would hold up was one of the few remaining questions for police departments and prosecutors still weighing its use, even as others have rushed to apply it. On Friday, the jury returned a guilty verdict.

"There is no stopping genetic genealogy now," said CeCe Moore, a genetic genealogist whose work led to the arrest in the murder case. "I think it will become a regular, accepted part of law enforcement investigations." ...

A forensic consulting firm, Parabon, offered to generate a **predictive likeness** using DNA. This was **not helpful** either."

Rapid DNA

Legalizing DNA Tests for DNA Indexing

Congress / Bills / H.R. 510 (115th) / Summary

H.R. 510 (115th): Rapid DNA Act of 2017

Overview **Summary** Details Text Study Guide

GovTrack's Summary

[Library of Congress](#)

Rapid DNA is a new technique that can analyze DNA samples in about 90 minutes, instead of days or even weeks as it took previously. A bill that passed the Senate and House last week would expand the use of this technology.

What the bill does

The Rapid DNA Act establishes a system for Rapid DNA's nationwide coordination among law enforcement departments, by connecting it to the FBI's Combined DNA Index System.

Labelled [S. 139](#) in the Senate and [H.R. 510](#) in the House, the legislation was introduced by Sen. Orrin Hatch (R-UT) and Rep. James Sensenbrenner (R-WI5).

Former FBI Director James Comey cited a real-life example of how the technology could be used effectively. “[It will] allow us, in booking stations around the country, if someone’s arrested, to know instantly—or near instantly—whether that person is the rapist who’s been on the loose in a particular community before they’re released on bail and get away or to clear somebody, to show that they’re not the person,” Comey [said in testimony](#).

Rapid DNA was used for the [first time ever in a criminal investigation in 2013](#), to nab burglars who stole more than \$30,000 worth of items from an Air Force Member’s Florida home while they were serving in Afghanistan. Presumably more such cases would be solved and quickly with expanded use of rapid DNA.

What supporters say

Supporters say it will save both time and taxpayer dollars by speeding up the DNA analysis process in a manner that’s no less effective, reducing the backlog of samples waiting to be tested.

“It will enable officers to take advantage of exciting new developments in DNA technology to more quickly solve crimes and exonerate innocent suspects,” Senate lead sponsor Hatch [said in a press release](#). “Under this legislation, rather than having to all send DNA samples to crime labs and wait weeks for results, trained officers will be able to process many samples in less than two hours.”

What opponents say

GovTrack Insider could not locate any members of Congress who expressed public opposition to the legislation, but some members of the public are concerned. The *New Republic* called the rise of rapid DNA [“troubling,”](#) citing the potential for privacy violations and misuses by immigration authorities. They also noted that the FBI already has DNA samples from more than 3.5 percent of Americans, a number likely to grow thanks to a 2015 Supreme Court decision allowing DNA samples to be taken without a warrant.

The Electronic Frontier Foundation expressed doubts about the accuracy of Rapid DNA. “Rapid DNA has only been tested on single-source samples—like a swab taken directly from a person’s inner cheek,” the EFF [writes](#). “And yet, Rapid DNA manufacturers are trying to convince law enforcement agencies to buy these machines to get through their backlog of rape kits and for low-level property crimes—situations where there’s a very good chance the DNA came from multiple people—some of whom may have had no connection to the crime at all.”

Votes and odds of passage

The legislation attracted a bipartisan mix of [12 Senate cosponsors](#), seven Republicans and five Democrats, and [24 House cosponsors](#), 17 Republicans and seven Democrats. It passed both the House and Senate on May 16, by a unanimous consent voice vote in both chambers, meaning no record of individual votes was recorded. It now goes to President Trump’s desk, where he appears likely to sign it.

<https://www.govtrack.us/congress/bills/115/hr510/summary>

Forensic G2P



Fig. 1. Individual examples of HlrisPlex-based eye and hair color DNA prediction. Probability outcomes are provided for eye and hair color categories as obtained from complete HlrisPlex SNP profiles [50] using the enhanced IrisPlex eye color and the enhance HlrisPlex hair color prediction models [25] (http://www.erasmusmc.nl/fmb/resources/Irisplex_HlrisPlex/) for 12 individuals chosen with varying eye and hair colors. Eye and hair photographs are provided to allow visual phenotype inspection and comparison with DNA predicted conclusions. Those probabilities that led to the eye and hair color conclusions are highlighted in grey based on the highest probability rule for eye color and by using the HlrisPlex hair color prediction guide described elsewhere [25,50]. Individual numbering is 1–6 on the left side and 7–12 on the right side. DNA-based prediction conclusions are as follows 1: black hair and brown eyes, 2: dark brown/black hair and brown eyes, 3: dark brown/black hair and blue eyes, 4: brown/dark brown hair and blue eyes, 5: brown/medium brown hair and brown eyes, 6: brown hair and brown eyes (likely with non-brown parts), 7: blond/dark blond hair and blue eyes, 8: blond hair and blue eyes, 9: blond/dark blond hair and blue eyes, 10: red hair and blue eyes, 11: red hair and brown eyes (likely with non-brown parts), and 12: red hair and blue eyes.

Phenotyping from DNA

From DNA to "Wanted" Posters?

- association of genomic variants with phenotypic data collection
- while hair, eye color are easy targets not useful for relevant phenotypic features especially if large environmental component
- huge biases based on input/collection data
- Belgium and Germany do not allow forensic DNA phenotyping
- Switzerland: Bundesrat decision on 2020-12-04 to allow phenotyping for law enforcement purposes



"When the New York Times ran an informal test of the Paragon system with one of its reporters, it failed badly." (ACLU.org)

DNA & Law Enforcement

Legal minefields, hard to avoid?

- "...when police in Edmonton, Canada, released a suspect's image, the **crude graphic** ... came **from the suspect's DNA**."
- "...every time a **family member** sends in their swab, they're sending in your data too..."
- "...**many players** in this growing movement offer to translate our genetic code into phenotypes (that is, observable features like eye color), often with **scant commitment to scientific accuracy**..."
- "...Veering into **pseudoscience**, they are a modern **sales pitch for** the long-discredited **phrenology** of the past. They wrongly treat race as a biological fact, rather than the social construct that it is. And in the process, they open all the **flaws of facial recognition** to new realms..."
- "...we first have to change our focus from preventing DNA collection to **preventing misuse** and managing access..."
- "The answer is simple: **Ban DNA searches** ... beyond the types of one-to-one DNA tests that are subject to judicial oversight..."



Cops Might Already Have Your DNA, Without Your Consent

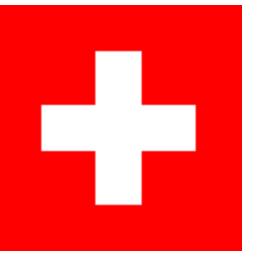
| FAREWELL PRIVACY |

We've entered the era of genetic surveillance and nothing—not even our own cells—is off-limits.

Albert Fox Cahn | Ayesha Rasheed

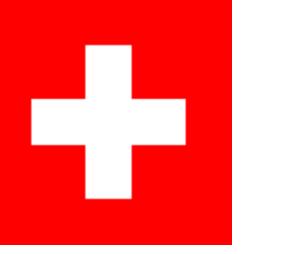
Published Nov. 14, 2022 4:51AM ET

"The unchallenged expansion of DNA collection and law enforcement misuse of the data has also spurred a surge in DNA surveillance startups."



The Swiss Way

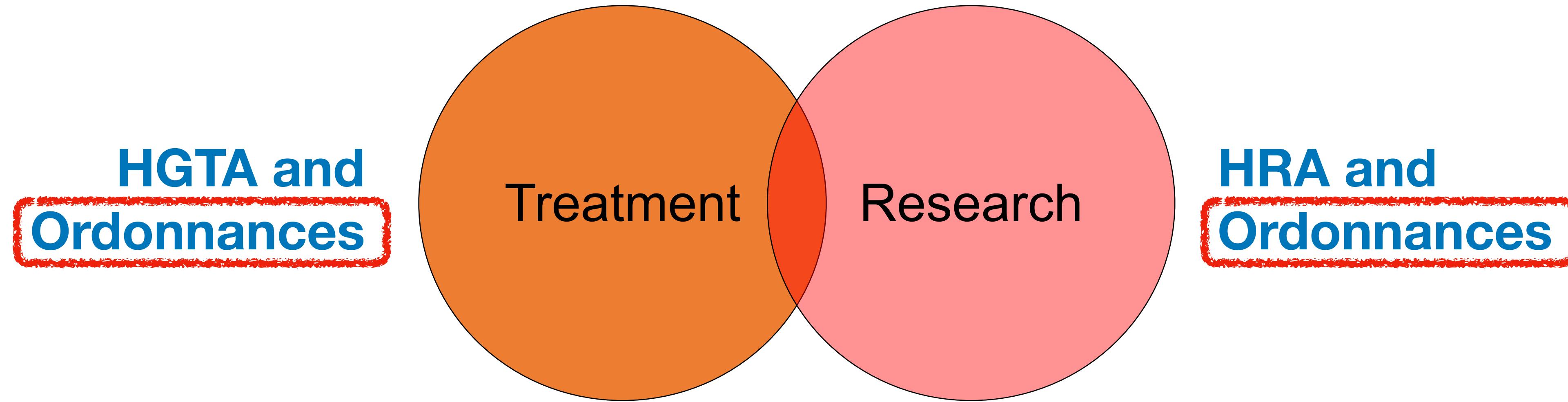
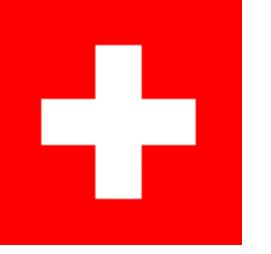
Genomic Data & Privacy Protection - The Swiss View



Relevant areas

- Medical treatment (Federal Act on Human Genetic Testing, HGTA)
- Human Research (Human Research Act, HRA)
- Tests other than for medical purposes (*new* in the HGTA from 2021 on)
- Law enforcement (Federal Act on the Use of DNA Profiles in Criminal Proceedings and for Identifying Unidentified or Missing Person, DNA Profiles Act)
- Data protection (Data Protection Act, DPA)
- ...

Law's View on Modern Medicine

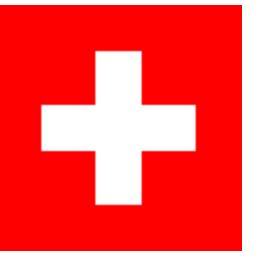


- How do we handle the growing overlap area?
 - unclear; current legislative movement:
HRA will relate more to HGTA in the future

HGTA : Federal
Act on Human
Genetic Testing

HRA: Human
Research Act

2021 Data Protection Act



Art. 5 Definitions

The following definitions apply in this Act:

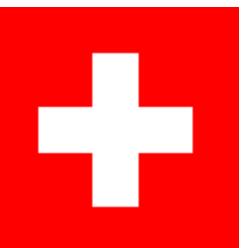
- a. **personal data**: all information relating to an identified or identifiable natural person;
- b. **data subject**: natural person whose personal data is processed;
- c. **sensitive personal data**:
 - 1. data on religious, ideological, political or trade union-related views or activities,
 - 2. data on health, the intimate sphere or the racial or ethnic origin,
 - 3. genetic data,
 - 4. biometric data which unequivocally identifies a natural person,

....

Therewith **Genetic Data is always sensitive data**, and especially Art. 6 Principles of data processing and **High-risk profiling**: profiling which involves a high risk to the personality or fundamental rights of the data subject, as it creates a pairing between data that enables an assessment of essential aspects of the personality of a natural person, needs to be considered deeper.

HGTA : Federal Act on Human Genetic Testing

HGTA new	medical field	outside the medical field	
Investigated characteristics	medical relevant	especially protective values characteristics	other characteristics
General Requirements	Non-discrimination, information and consent, right to information, right not to know, avoidance of surplus information, protection of samples and genetic data, Circulation concerning public advertising, state of science and technology, penal provisions		
Initiation	Physician	Health professional (controlled taking of samples)	Consumer (DTC)
Persons concerned	Persons with and without capacity of judgement, pregnant woman (PND)	ONLY persons with Capacity of judgement	ONLY persons with Capacity of judgement
Communication of surplus information	as a rule according to decision of the person concerned	Not allowed	Not allowed
Laboratory	subject to authorization (cyto and molecular genetic studies)	subject to authorization (cyto and molecular genetic studies)	not subject to authorisation
Employers and Insurance institutions	Studies and Recovery of Results / Data only in regulated exceptional cases	Prohibition to carry out investigations and the Recovery of Results / Data	Prohibition to carry out investigations and the Recovery of Results / Data



Verordnung über genetische Untersuchungen beim Menschen (GUMV)

vom 23. September 2022 (Stand am 1. Dezember 2022)

Dieser Text ist in Kraft

Der Schweizerische Bundesrat,

gestützt auf das Bundesgesetz vom 15. Juni 2018¹ über genetische Untersuchungen beim Menschen (GUMG) und Artikel 8 Absatz 2 des Fortpflanzungsmedizingesetzes vom 18. Dezember 1998² (FMedG), verordnet:

- Art. 3 Schutz von Proben und genetischen Daten

(Art. 6 Bst. c und 10 GUMG)

¹ Wer genetische Daten bearbeitet, muss sicherstellen, dass der Schutz der Daten insbesondere vor unbefugter oder unbeabsichtigter Bekanntgabe, Veränderung, Löschung, Vernichtung oder Erstellung von Kopien sowie vor Verlust gewährleistet ist.

² Der Schutz ist durch angemessene technische und organisatorische Massnahmen zu gewährleisten, insbesondere durch:

- die Beschränkung der Bearbeitung der genetischen Daten auf diejenigen Personen, die die Daten zur Erfüllung ihrer Aufgaben benötigen;
- die Protokollierung aller zur Gewährleistung der Rückverfolgbarkeit massgeblichen Bearbeitungsvorgänge;
- die sichere Übermittlung genetischer Daten;
- die Pseudonymisierung genetischer Daten, wenn sie in ein Land übermittelt werden, dessen Gesetzgebung keinen angemessenen Schutz gewährleistet.

³ Die Massnahmen sind anhand einer Risikoabschätzung und unter Berücksichtigung des Stands der Technik zu bestimmen und zu aktualisieren.

⁴ Werden genetische Daten pseudonymisiert und in ein Land übermittelt, dessen Gesetzgebung keinen angemessenen Schutz gewährleistet, so muss die betroffene Person im Rahmen ihrer Aufklärung darüber informiert werden.

² Für die Erstellung von DNA-Profilen zur Klärung der Abstammung oder zur Identifizierung gilt die Verordnung vom 14. Februar 2007³ über die Erstellung von DNA-Profilen im Zivil- und im Verwaltungsbereich (VDZV).

	Untersuchung	Erforderlicher Titel (x)				
		C	H	I	P	MP
1.	Creutzfeldt-Jakob-Krankheiten, fatale familiäre Insomnie, Gerstmann-Sträussler-Krankheit					x
2.	Familiär defektes Apolipoprotein B-100	x			x	
3.	Familiäre Krebssyndrome; direkte oder indirekte Mutationsanalyse bei Prädispositionen für Karzinome, Sarkome, Lymphome, Leukämien, neurogene, melanozytäre oder embryonale Tumore					x
4.	Genetische Untersuchungen zur Typisierung von Blutgruppen sowie Blut- und Gewebemarken im Rahmen der Abklärung einer Erbkrankheit oder einer Krankheitsveranlagung	x	x	x	x	
5.	Hämochromatose, familiäre; direkte Mutationsanalyse	x	x		x	x
6.	Hämoglobinopathien; direkte oder indirekte Mutationsanalyse bei Thalassämien, Sichelzellanämie		x		x	
7.	Hämostasestörungen; direkte oder indirekte Mutationsanalyse bei Faktor II- und Faktor V-Störung	x	x		x	

- Art. 61 Genetische Untersuchungen von pathologisch verändertem biologischem Material bei Krebserkrankungen

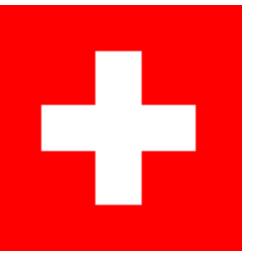
¹ Genetische Untersuchungen von pathologisch verändertem biologischem Material, die bei Krebserkrankungen durchgeführt werden und nicht zur Abklärung von erblichen Eigenschaften des Erbguts dienen, sind vom Geltungsbereich des GUMG ausgenommen, wenn aufgrund der Zusammensetzung des untersuchten biologischen Materials und des gewählten Untersuchungsverfahrens davon auszugehen ist, dass keine Überschussinformationen zu erblichen Eigenschaften entstehen.

² Pathologisch verändertes biologisches Material bei Krebserkrankungen umfasst insbesondere:

- pathologisch oder potenziell pathologisch veränderte Gewebe, Zellen oder Körperflüssigkeiten;
- im Blut vorhandene pathologisch veränderte Zellen oder deren Bestandteile.

³ Ist bei genetischen Untersuchungen von pathologisch verändertem biologischem Material, die bei Krebserkrankungen durchgeführt werden und nicht zur Abklärung von erblichen Eigenschaften des Erbguts dienen, davon auszugehen, dass Überschuss-informationen zu erblichen Eigenschaften entstehen, so gelten die Artikel 3–5, 7–15, 27 und 56–58 GUMG.

Data Ownership



- Within Switzerland, there is no coherent approach on ownership of data as such (but academic discussion is ongoing, if that is needed).
- Restrictions of usage and disclosure of data other than personal data mainly stem from contractual relationships.
- In the field of research this leads mostly to a data ownership by the research institution.

Of course the restrictions of the different acts that are in the field need to be respected (procuring data lawfully, consent for further use, etc.)

Way forward...

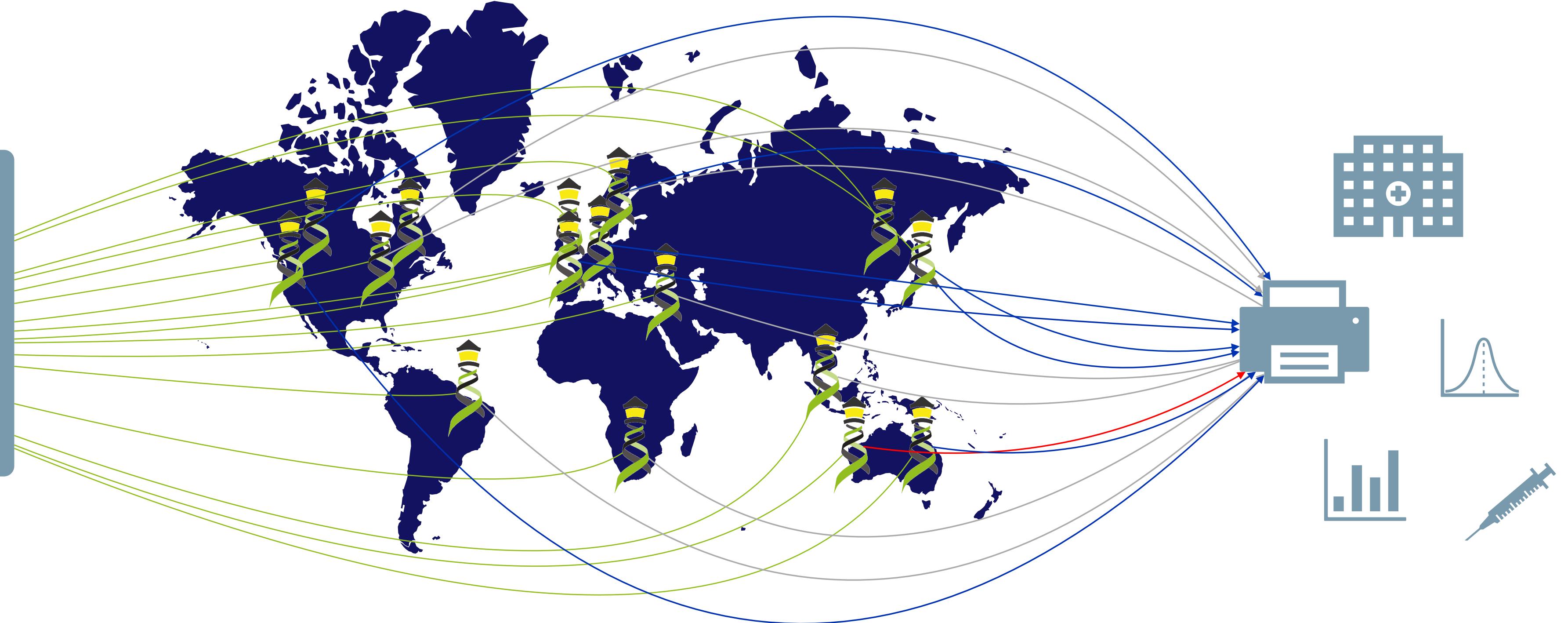
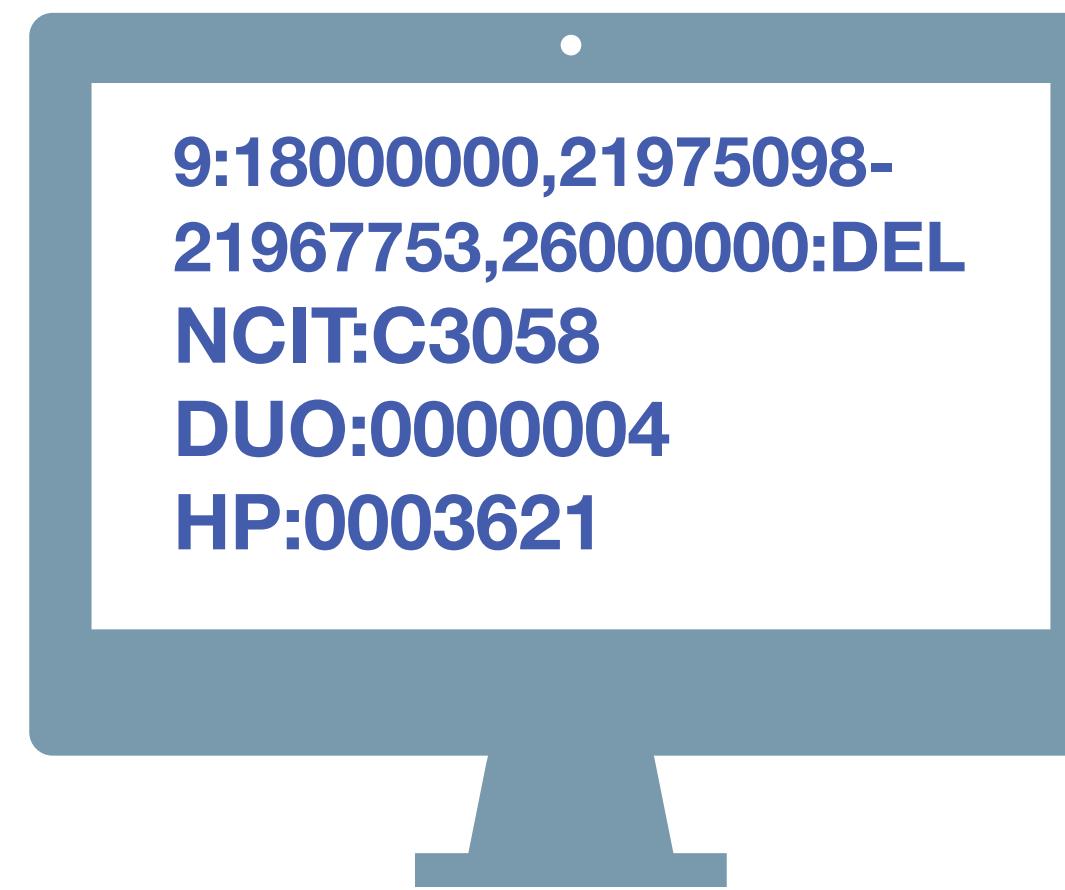
The vision for genomic research: **Federation** of data





Making Beacons Biomedical - Beacon v2

- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
 - ▶ cytogenetic annotations, named variants, variant effects
- Beacon queries as entry for **data delivery**
 - ▶ Beacon v2 permissive to respond with variety of data types
 - Phenopackets, biosample data, cohort information ...
 - ▶ handover to stream and download using htsgt, VCF, EHRs
- Interacting with EHR standards
 - ▶ FHIR translations for queries and handover ...
- Beacons as part of local, secure environments
- Authentication to enable non-aggregate, patient derived datasets
 - ▶ ELIXIR AAI with compatibility to other providers (OAuth...)



Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?



Beacon v2 API

The Beacon API v2 proposal opens the way for the design of a simple but powerful "genomics API".



Making Beacons Biomedical - Beacon v2

- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
 - ▶ cytogenetic annotations, named variants, variant effects
- Beacon queries as entry for **data delivery**
 - ▶ Beacon v2 permissive to respond with variety of data types
 - Phenopackets, biosample data, cohort information ...
 - ▶ handover to stream and download using htsgt, VCF, EHRs
- Interacting with EHR standards
 - ▶ FHIR translations for queries and handover ...
- Beacons as part of local, secure environments
- Authentication to enable non-aggregate, patient derived datasets
 - ▶ ELIXIR AAI with compatibility to other providers (OAuth...)

Definitely breaks the
"Relative Security
by Design"
Concept!

Generalkonsent

BENEFIT

BLOCKCHAIN

HEALTH

PRIVACY

CONSENT

SECURITY

ACCESS

Right to Research

HACKERS

LAWS

**Genetic
Information
Nondiscrimination
Act**

**Health
Insurance
Portability and
Accountability
Act**

SAFETY

CRYPTOGRAPHY

The Right to Scientific Knowledge

In 1948, the General assembly of the United nations adopted the Universal Declaration of Human Rights (UDHr) to guarantee the rights of every individual in the world. Included were twin rights “to share in scientific advancement and its benefits” and “to the protection of the moral and material interests resulting from any scientific...production of which [a person] is the author” (art. 27, United nations 1948).

from Knoppers et al, 2014

A human rights approach to an international code of conduct for genomic and clinical data sharing

Bartha M. Knoppers · Jennifer R. Harris · Isabelle Budin-Ljøsne · Edward S. Dove

Received: 9 December 2013 / Accepted: 16 February 2014 / Published online: 27 February 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract Fostering data sharing is a scientific and ethical imperative. Health gains can be achieved more comprehensively and quickly by combining large, information-rich datasets from across conventionally siloed disciplines and geographic areas. While collaboration for data sharing is increasingly embraced by policymakers and the international biomedical community, we lack a common ethical and legal framework to connect regulators, funders, consortia, and research projects so as to facilitate genomic and clinical data linkage, global science collaboration, and responsible research conduct. Governance tools can be used to responsibly steer the sharing of data for proper stewardship of research discovery, genomics research resources, and their clinical applications. In this article, we propose that an international code of conduct be designed to enable global genomic and clinical data sharing for biomedical research. To give this proposed code universal application and accountability, however, we propose to position it within a human rights framework. This proposition is not without precedent: international treaties have long recognized that everyone has a right to the benefits of scientific

progress and its applications, and a right to the protection of the moral and material interests resulting from scientific productions. It is time to apply these twin rights to internationally collaborative genomic and clinical data sharing.

Introduction

In 1948, the General Assembly of the United Nations adopted the *Universal Declaration of Human Rights* (UDHR) to guarantee the rights of every individual in the world. Included were twin rights “to share in scientific advancement and its benefits” and “to the protection of the moral and material interests resulting from any scientific...production of which [a person] is the author” (Art. 27, United Nations 1948). In the 21st century, where are we in realizing the sharing of scientific advancement and its benefits, and the importance of protecting a scientific producer’s moral and material interests? In this article, we argue that these little-developed twin rights, what we call the right “to benefit from” and “to be recognized for”, have direct application to internationally collaborative genomic and clinical data sharing, and can be activated through an international code of conduct.

Sharing genomic and clinical data is critical to achieve precision medicine (National Research Council 2011), that is, more accurate disease classification based on molecular profiles to enable tailored effective treatments, interventions, and models for prevention. Better communication flow across borders and research teams, encompassing data from clinical and population research, enables researchers to connect the diverse types of datasets and expertise needed to elucidate the genomic basis and complexities of disease etiology. Such data integration can make it possible to reveal the genetic basis of cancer, inherited diseases,

B. M. Knoppers (✉) · E. S. Dove
Centre of Genomics and Policy, McGill University, 740 Dr.
Penfield Avenue, Suite 5200, Montreal H3A 0G1, Canada
e-mail: bartha.knoppers@mcgill.ca

E. S. Dove
e-mail: edward.dove@mcgill.ca

J. R. Harris · I. Budin-Ljøsne
Division of Epidemiology, Department of Genes
and Environment, Norwegian Institute of Public Health,
PO Box 4404, Nydalen 0403, Oslo, Norway
e-mail: Jennifer.Harris@fhi.no

I. Budin-Ljøsne
e-mail: Isabelle.Budin.Ljosne@fhi.no

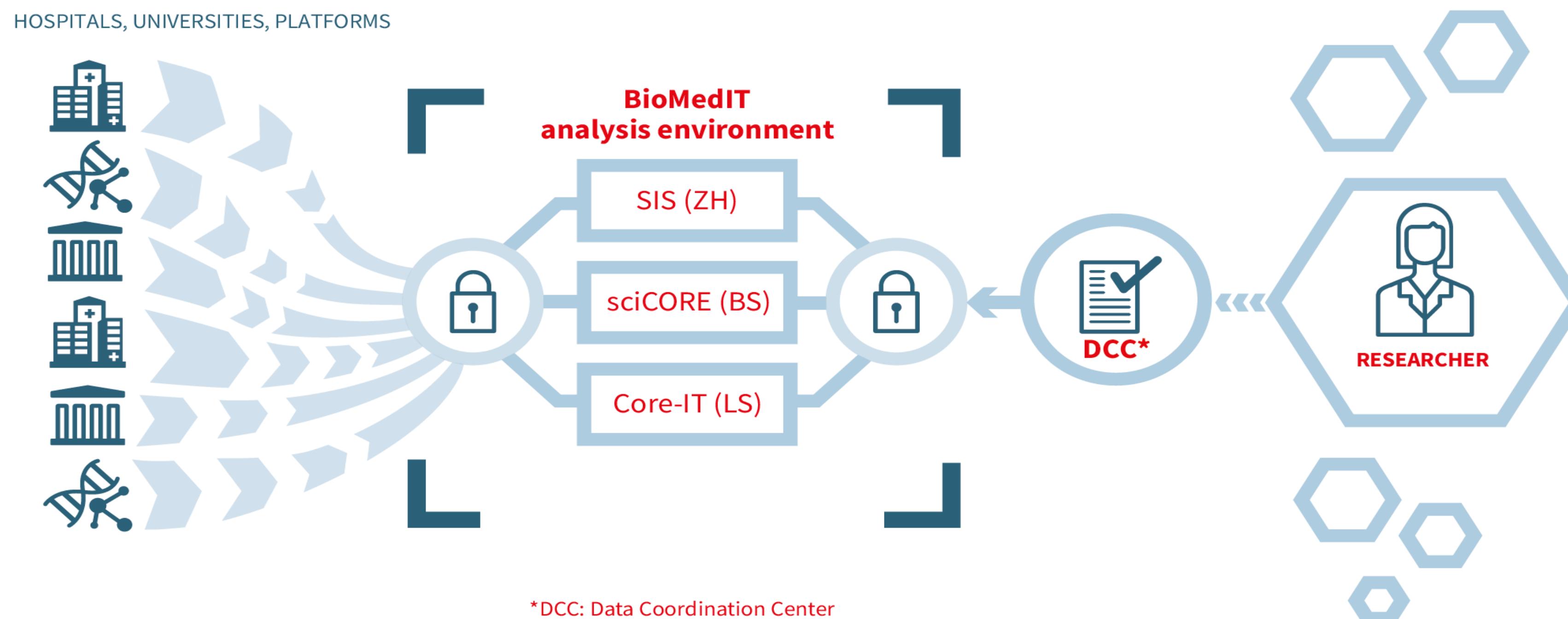
Improving Data Privacy but Empowering Beneficial Use

Intersecting Areas of Development

- Make genomic (and functional) data "obfuscated" for malicious use
 - ▶ e.g. spiking / randomization of variants in "not-disease" loci
- access protection with defined user access using standardized protocols for users' roles and permissions, in contrast to individual per user, per dataset access requests over data access committees (DACs)
 - ▶ digital "differential" consent using e.g. data use ontologies
- intentional and unintentional (!) data providers have to be protected from abuse by legal regulations - though thin line regarding "overzealous" use by law enforcement
- alternative solution for active consent
 - ▶ encrypted wide-area networking solutions with managed access control (e.g. SPHN's BiomedIT) and limited access to anonymized data (e.g. using the Beacon protocol with "handover" scenarios)
 - ▶ (genomic) data ownership by the individual "data donors, together with strong privacy protection by law

The BioMedIT network

BioMedIT provides researchers with access to a secure and protected computing environment for analysis of sensitive data without compromising data privacy





Making Beacons Biomedical - Beacon v2

- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
 - ▶ cytogenetic annotations, named variants, variant effects

- Beacon queries as entry for **data delivery**

- ▶ Beacon v2 permissive to respond with variety of data types
 - Phenopackets, biosample data, cohort information ...

- ▶ handover to stream and download using htsgt, VCF, EHRs

- Interacting with EHR standards

- ▶ FHIR translations for queries and handover ...

- Beacons as part of local, secure environments

- Authentication to enable non-aggregate, patient derived datasets

- ▶ ELIXIR AAI with compatibility to other providers (OAuth...)

Definitely breaks the
"Relative Security
by Design"
Concept!

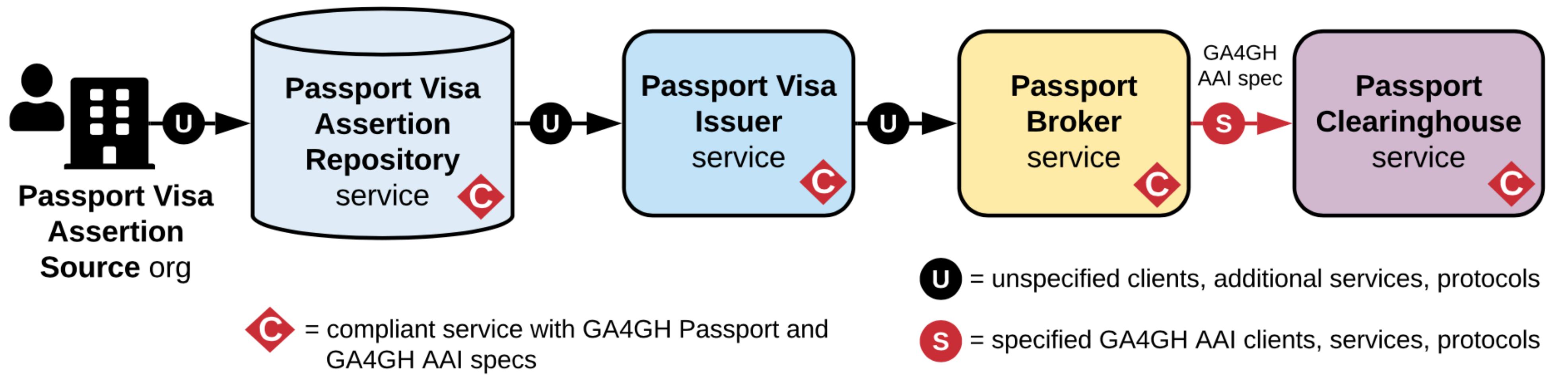
Mitigation by
tailored
implementation and
security practices

GA4GH Passports



Global Alliance
for Genomics & Health

Communicating a user's data access authorizations



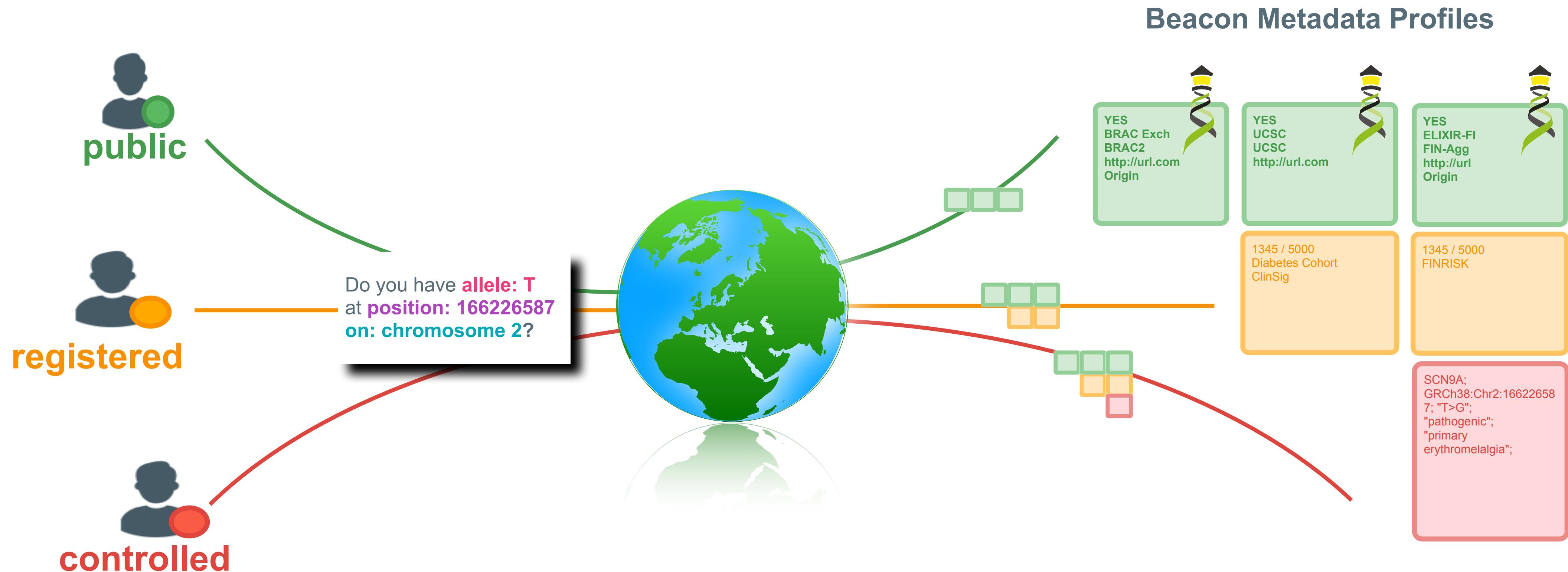
- format to communicate a user's data access authorizations based on either their role (e.g. researcher), affiliation, or access status
- works together with the GA4GH Authentication and Authorization Infrastructure (AAI) OpenID Connect Profile to streamline researchers' data access over federated data access protocols
- both standards approved in Dec 2019 with early implementation by Google Cloud services and ELIXIR

Google Cloud

elixir

Empowering Beacon use through Access Levels

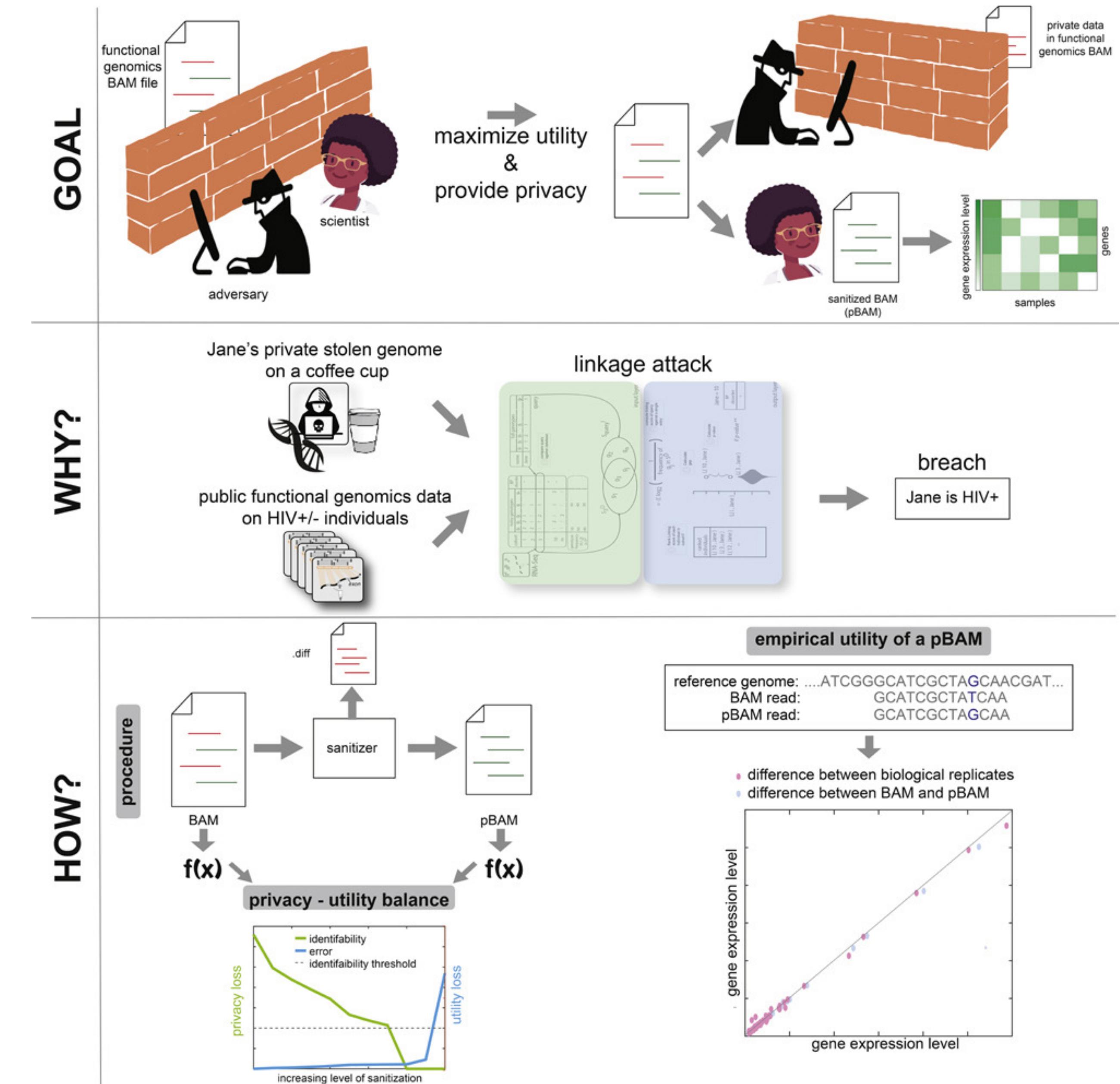
Integrating permissions and discovery



Information Leakage from Functional Genomics Data

"Sanitize"...

- "functional" genomics data can be sanitized by removing features which are not relevant for the specific use cases
- an example could be the randomization of variant alleles in datasets where variant call specificity is of minor concern



Health Related Data & Privacy

Considerations when evaluating risks of data sharing

- Is the genetic condition outwardly visible?
- How severe is it? (serious disease, penetrance, age of onset)
- Is it associated with what could be considered to be stigmatizing health information (e.g., associated with mental health, reproductive care, disability)?
- Is it familial (i.e., potential carrier status/reproductive implications for family/relatives)?
- Does it provide information about the likely geographical location of individuals?
- Does it provide information about ethnicity that may be considered potentially stigmatizing information?

Sharing health-related data: a privacy test?

Stephanie OM Dyke¹, Edward S Dove² and Bartha M Knoppers¹

Greater sharing of potentially sensitive data raises important ethical, legal and social issues (ELSI), which risk hindering and even preventing useful data sharing if not properly addressed. One such important issue is respecting the privacy-related interests of individuals whose data are used in genomic research and clinical care. As part of the Global Alliance for Genomics and Health (GA4GH), we examined the ELSI status of health-related data that are typically considered 'sensitive' in international policy and data protection laws. We propose that 'tiered protection' of such data could be implemented in contexts such as that of the GA4GH Beacon Project to facilitate responsible data sharing. To this end, we discuss a Data Sharing Privacy Test developed to distinguish degrees of sensitivity within categories of data recognised as 'sensitive'. Based on this, we propose guidance for determining the level of protection when sharing genomic and health-related data for the Beacon Project and in other international data sharing initiatives.

npj Genomic Medicine (2016) **1**, 16024; doi:10.1038/npjgenmed.2016.24; published online 17 August 2016

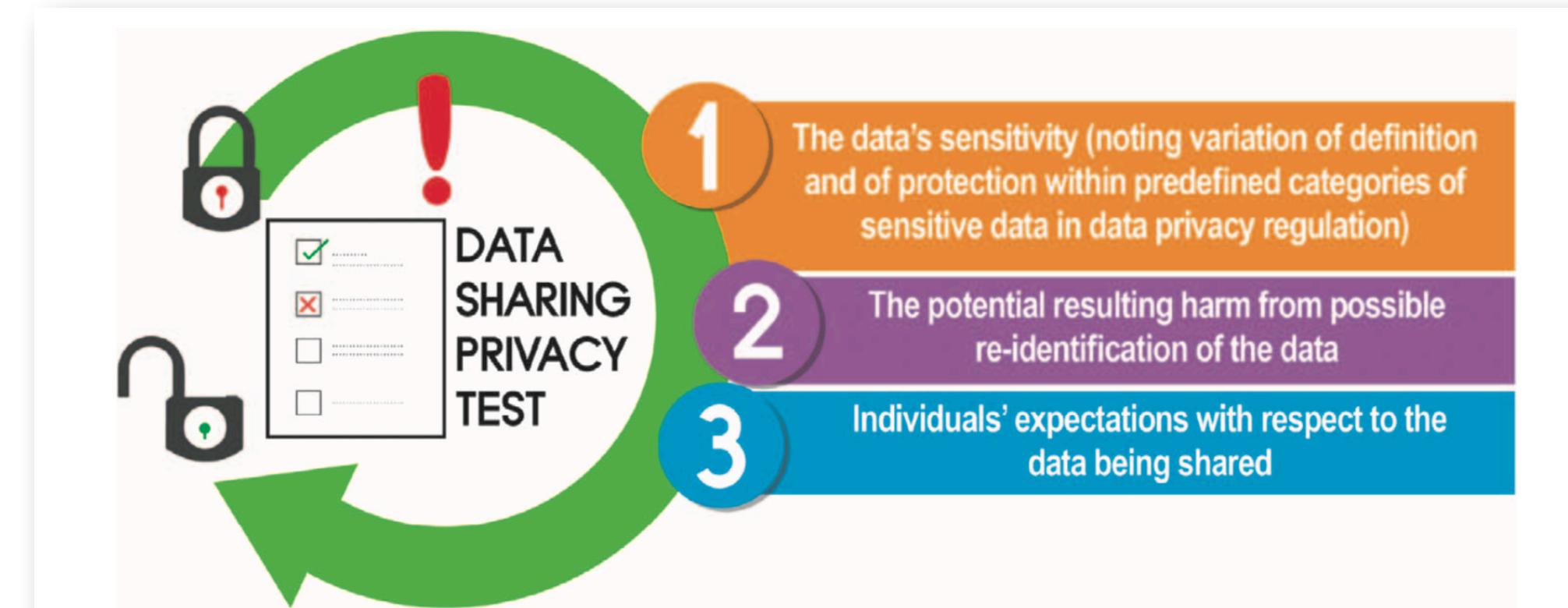


Figure 1. The three steps of a Data Sharing Privacy Test to distinguish degrees of data sensitivity within categories of data recognised as 'sensitive'.

Modernizing Patient Consent

forward looking, transparent and technically feasible regulations for enabling access to research material and data while empowering **patients**

Generalkonsent: Eine einheitliche Vorlage soll schweizweite Forschung erleichtern

Art des Forschungsmaterials	Biologisches Material und genetische Daten	Nicht-genetische Daten
Personenbezug		
Unverschlüsselt (identifizierend)	Information + Einwilligung in jedes einzelne Forschungsprojekt	Information über Weiterverwendung für zukünftige noch unbestimmte Forschungsprojekte + Generalkonsent für Forschungszwecke
Verschlüsselt	Information über Weiterverwendung für zukünftige noch unbestimmte Forschungsprojekte + Generalkonsent für Forschungszwecke	Information über Weiterverwendung für zukünftige noch unbestimmte Forschungsprojekte + Generalkonsent für Forschungszwecke + über Möglichkeit Weiterverwendung abzulehnen > Widerspruchsrecht
Anonymisiert	Genetische Daten: Information über Weiterverwendung für zukünftige noch unbestimmte Forschungszwecke + über Möglichkeit Weiterverwendung abzulehnen > Widerspruchsrecht Proben: Information zur Anonymisierung > Widerspruchsrecht	Ausserhalb des Geltungsbereichs des HFG



Consent Codes: Upholding Standard Data Use Conditions

Stephanie O. M. Dyke^{1*}, Anthony A. Philippakis², Jordi Rambla De Argila^{3,4}, Dina N. Paltoo⁵, Erin S. Luetkemeier⁵, Bartha M. Knoppers¹, Anthony J. Brookes⁶, J. Dylan Spalding⁷, Mark Thompson⁸, Marco Roos⁸, Kym M. Boycott⁹, Michael Brudno^{10,11}, Matthew Hurles¹², Heidi L. Rehm^{2,13}, Andreas Matern¹⁴, Marc Fiume¹⁵, Stephen T. Sherry¹⁶



Consent Codes		
Name	Abbreviation	Description
Primary Categories (I^{IV})		
no restrictions	NRES	No restrictions on data use.
general research use and clinical care	GRU(CC)	For health/medical/biomedical purposes and other biological research, including the study of population origins or ancestry.
health/medical/biomedical research and clinical care	HMB(CC)	Use of the data is limited to health/medical/biomedical purposes, does not include the study of population origins or ancestry.
disease-specific research and clinical care	DS-[XX](CC)	Use of the data must be related to [disease].
population origins/ancestry research	POA	Use of the data is limited to the study of population origins or ancestry.
Secondary Categories (II^{IV}) (can be one or more extra conditions, in addition to I ^{IV} category)		
other research-specific restrictions	RS-[XX]	Use of the data is limited to studies of [research type] (e.g., pediatric research).
research use only	RUO	Use of data is limited to research purposes (e.g., does not include its use in clinical care).
no “general methods” research	NMDS	Use of the data includes methods development research (e.g., development of software or algorithms) ONLY within the bounds of other data use limitations.
genetic studies only	GSO	Use of the data is limited to genetic studies only (i.e., no research using only the phenotype data).
Requirements		
not-for-profit use only	NPU	Use of the data is limited to not-for-profit organizations.
publication required	PUB	Requestor agrees to make results of studies using the data available to the larger scientific community.
collaboration required	COL-[XX]	Requestor must agree to collaboration with the primary study investigator(s).
return data to database/resource	RTN	Requestor must return derived/enriched data to the database/resource.
ethics approval required	IRB	Requestor must provide documentation of local IRB/REC approval.
geographical restrictions	GS-[XX]	Use of the data is limited to within [geographic region].
publication moratorium/embargo	MOR-[XX]	Requestor agrees not to publish results of studies until [date].
time limits on use	TS-[XX]	Use of data is approved for [x months].
user-specific restrictions	US	Use of data is limited to use by approved users.
project-specific restrictions	PS	Use of data is limited to use within an approved project.
institution-specific restrictions	IS	Use of data is limited to use within an approved institution.

SOM Dyke, et al. Consent Codes: Upholding Standard Data Use Conditions. *PLoS Genetics* 12(1): e1005772.
<http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005772>

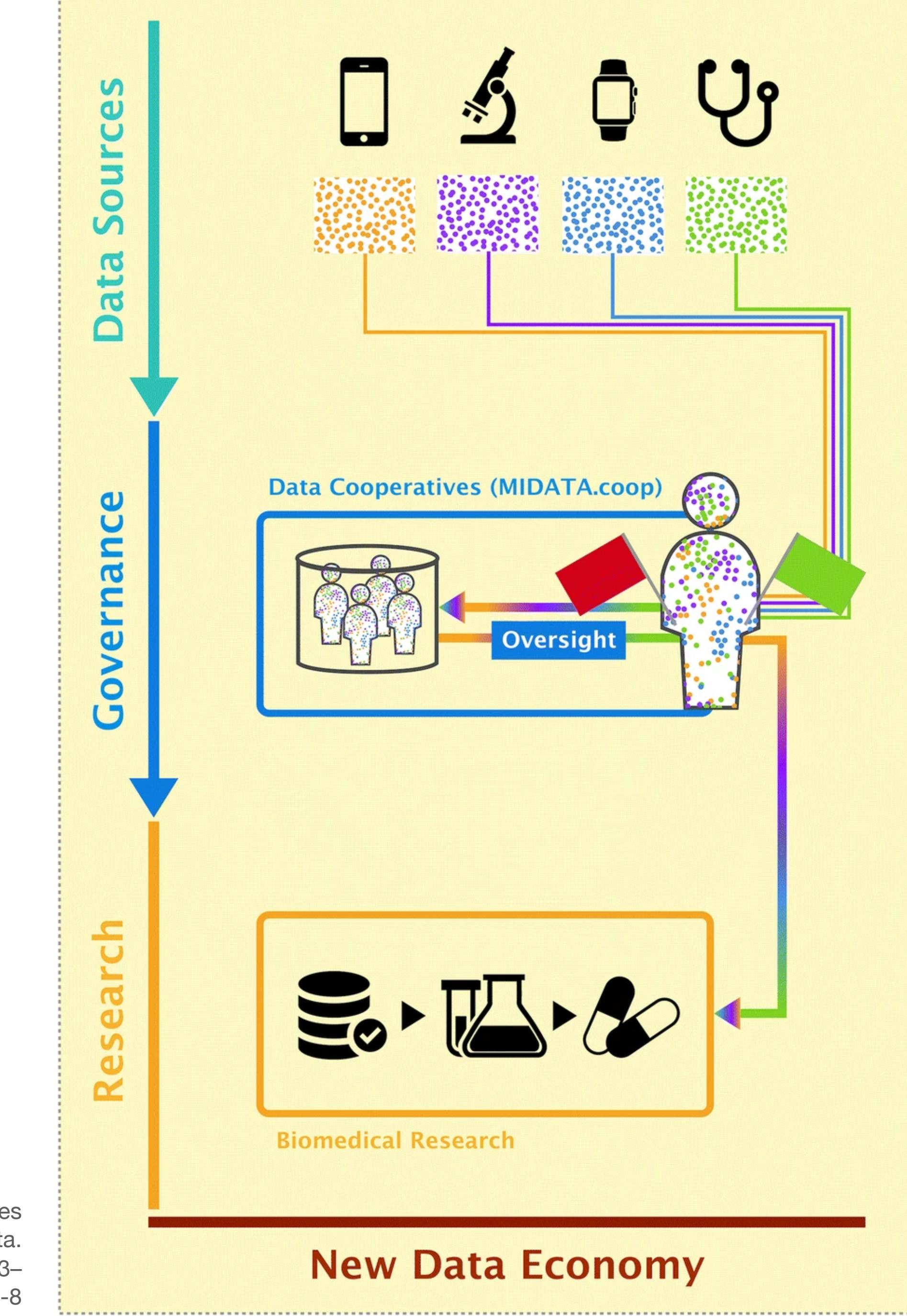
Contact: Dr. Stephanie Dyke (stephanie.dyke@mcgill.ca)

Switzerland: Definition of a unified "Generalkonsent", to provide a single framework to manage permissions for access to patient derived material and related data

Power to the People?!

Individuals as Owners & Managers of their Data

- (genomic) data ownership by the individual "data donors"
- supported by technological frameworks for data management and arbitration
- one vision here are "data cooperatives"
- need strong support from policy makers and financial sustainability support



Genomic Data & Privacy - Key Areas

- **Re-identification**
 - ▶ identification of an individual based on sets of genomic variants they (or close relatives) carry - so one needs some genome data first
 - ▶ information to be gained is circumstantial (e.g. their genome is in a particular disease related dataset)
 - ▶ currently only risk with some practical use (e.g. **long-range familial attacks**)
- **Genotype-to-Phenotype (G2P) attacks**
 - ▶ determination of some disease risk or phenotypic features from a genome itself
 - ▶ needs access to genome data which is illegal in many jurisdictions (but technically more & more feasible)
 - ▶ real-world use cases are limited but abuse through wrong perception of utility
- **Genomic Determinism**
 - ▶ assignment of individual abilities and personal development trajectories from genomic profiling
 - ▶ topic of (some good, most bad) SciFi
 - ▶ but: **Wehret den Anfängen!**

Genomic Data & Privacy - Some Take-Home Messages

- Many clinical and research applications in genomics **need vast numbers of genomes** to evaluate e.g. genotype-phenotype relationships
- Such data cannot simply be provided by a few reference data curation resources - and those again rely on multitudes of original data resources > **federated data access + data curation**
- Genomic data is considered to potentially expose unwilling individuals through **re-identification**/de-anonymization but also through direct information (genotype -> phenotype/disease)
- Legislative bodies and law enforcement have varying and *curious* approaches to "genomic privacy", with a mix of de-legalizing genomic data generation (e.g. in Switzerland) or strictly limiting its use while also using "eminent domain" to co-opt such data for criminal persecution in a possibly extending set of use cases

Share YOUR Genome data?

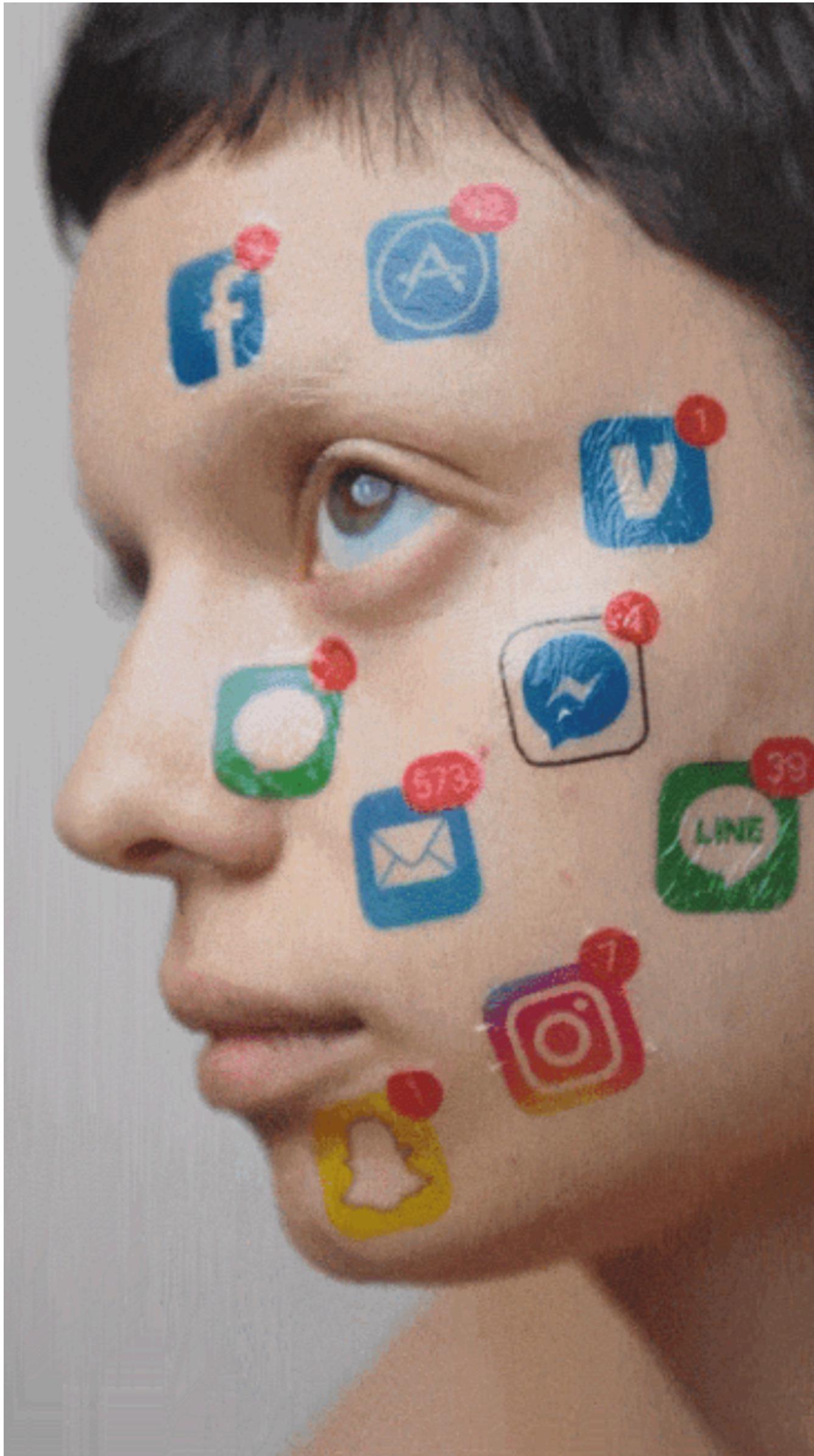
- The Beacon concept - balanced approach for accessing genome variant data from internationally distributed resources
- However: Genome data has the inherent “risk” of being identified and linked to a person

Solutions from Technology or Society? Discourse!

Welcome to openSNP

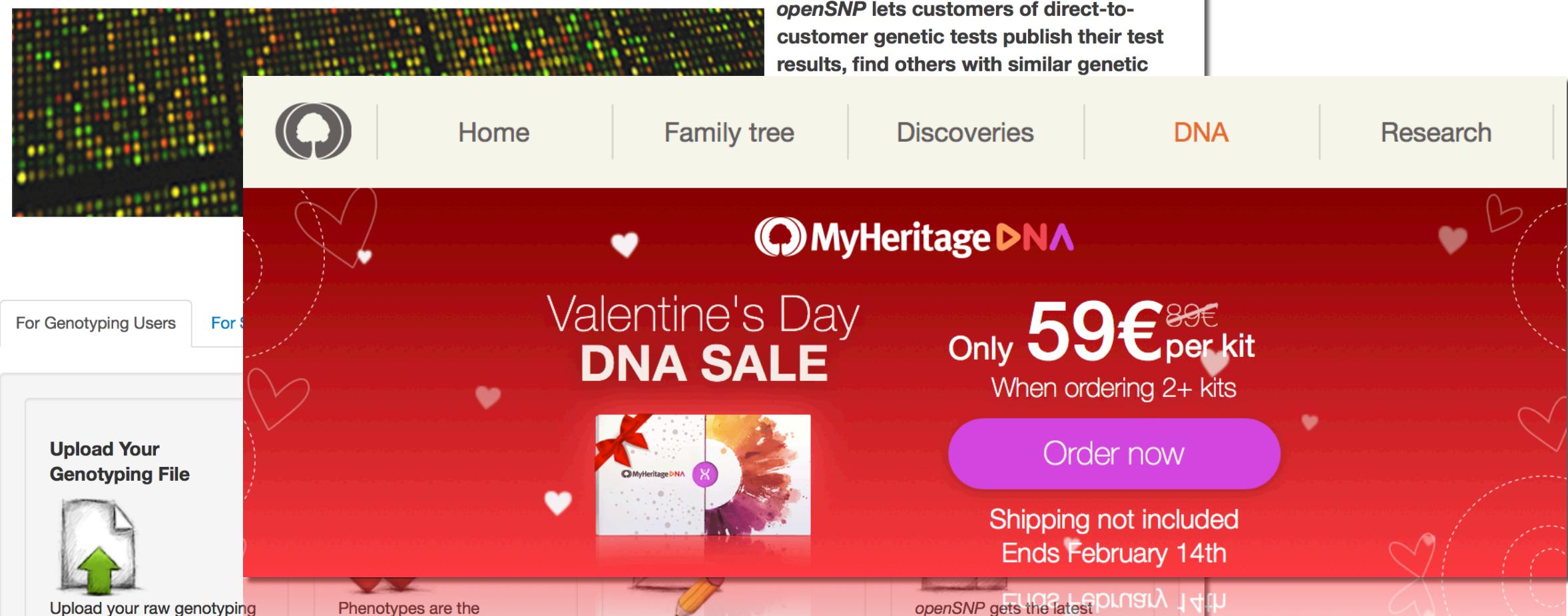
The screenshot shows the openSNP website homepage. At the top, there is a banner for "MyHeritage DNA" with a "Valentine's Day DNA SALE" offer. The banner includes a "Upload Your Genotyping File" button and a "For Genotyping Users" section. The main navigation menu includes Home, Family tree, Discoveries, DNA (selected), and Research. A sidebar on the right contains a snippet of text about openSNP's mission to publish test results and find others with similar genetic backgrounds.

The screenshot shows two side-by-side website snippets. On the left is the 23andMe website, featuring a "Welcome to you" message and a "saliva collection kit". On the right is the AncestryDNA website, which includes a headline stating "THE AVERAGE BRITISH PERSON'S DNA IS ONLY 36% BRITISH". It also features "GROW YOUR TREE" and "Find your ancestors in our database" buttons, along with an "ancestryDNA" logo.



John Yuyi, NYT 2018-02-09

Welcome to *openSNP*



BIO390 HS22

Exam planning

- On site exam!
- 2022-12-20
- time: 08:15-09:45
- multiple (single + multiple) choice w/ one or two open questions
- no material, phones etc.
- student ID for entrance



info.baudisgroup.org