Genomic Data & Privacy

Risks & opportunities

Michael Baudis | UZH BIO392 HS22

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**.

Genome *Beacons* Compromise Security?

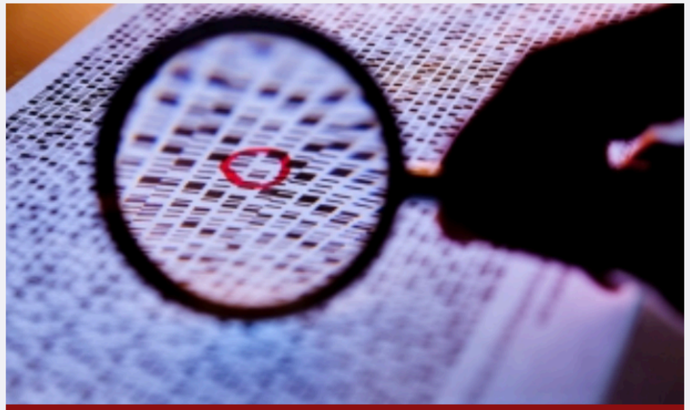
Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals in an anonymized genomic data collection

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.



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

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.

IDENTIFICATION OF INDIVIDUALS FROM MIXED COLLECTIONS USING RARE ALLELES

The American Journal of Human Genetics 97, 631–646, November 5, 2015

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.

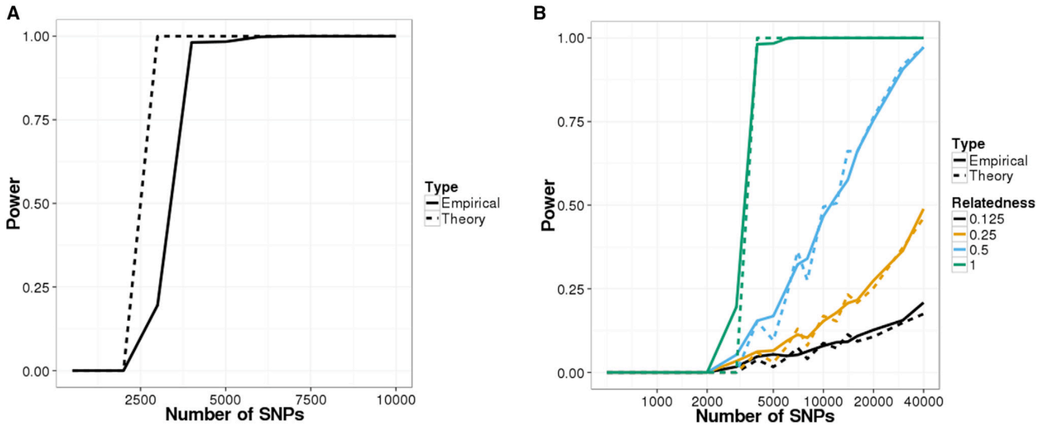


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- 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

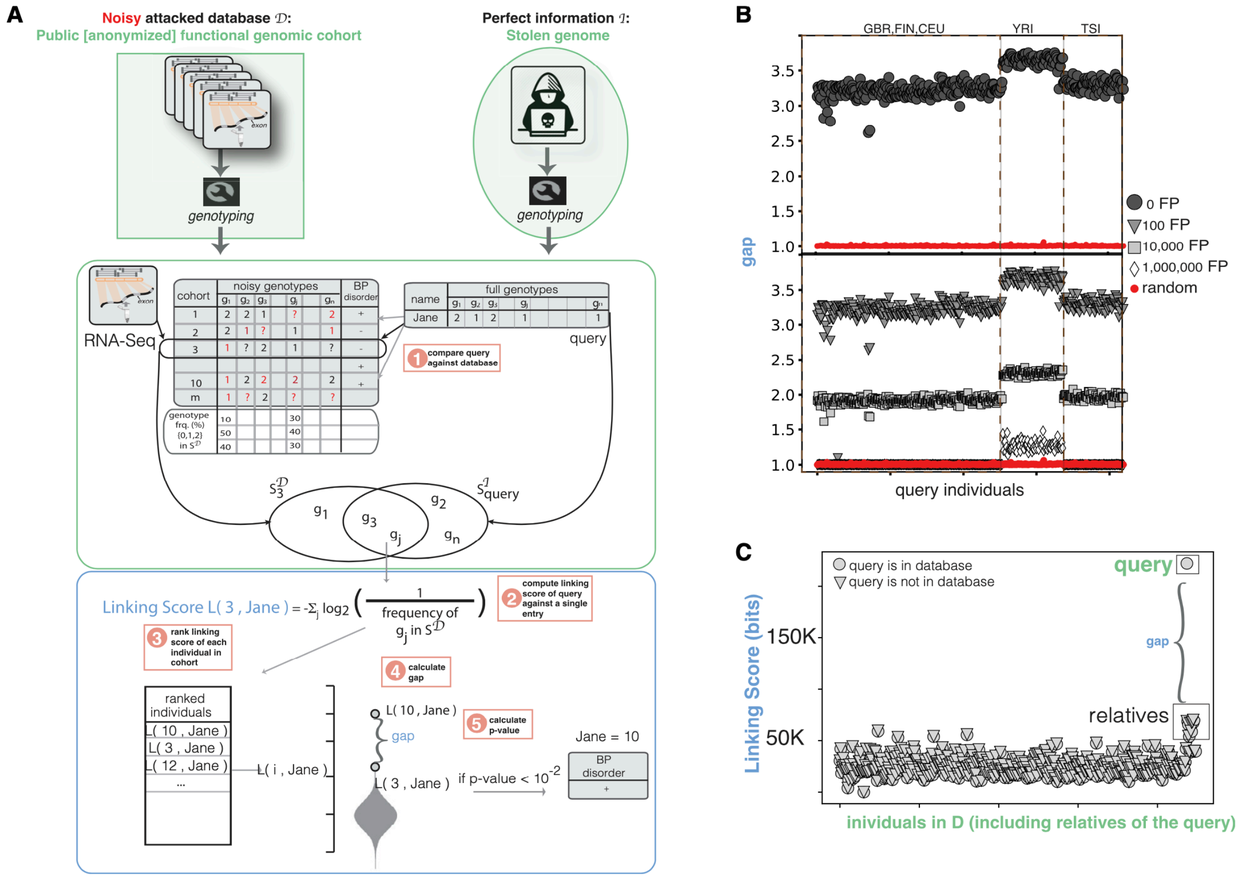


Figure 1. Functional Genomics Data De-anonymization Scheme with Perfect Genomes

(A) Anonymized functional genomics data from a cohort of individuals can be seen as a database *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.

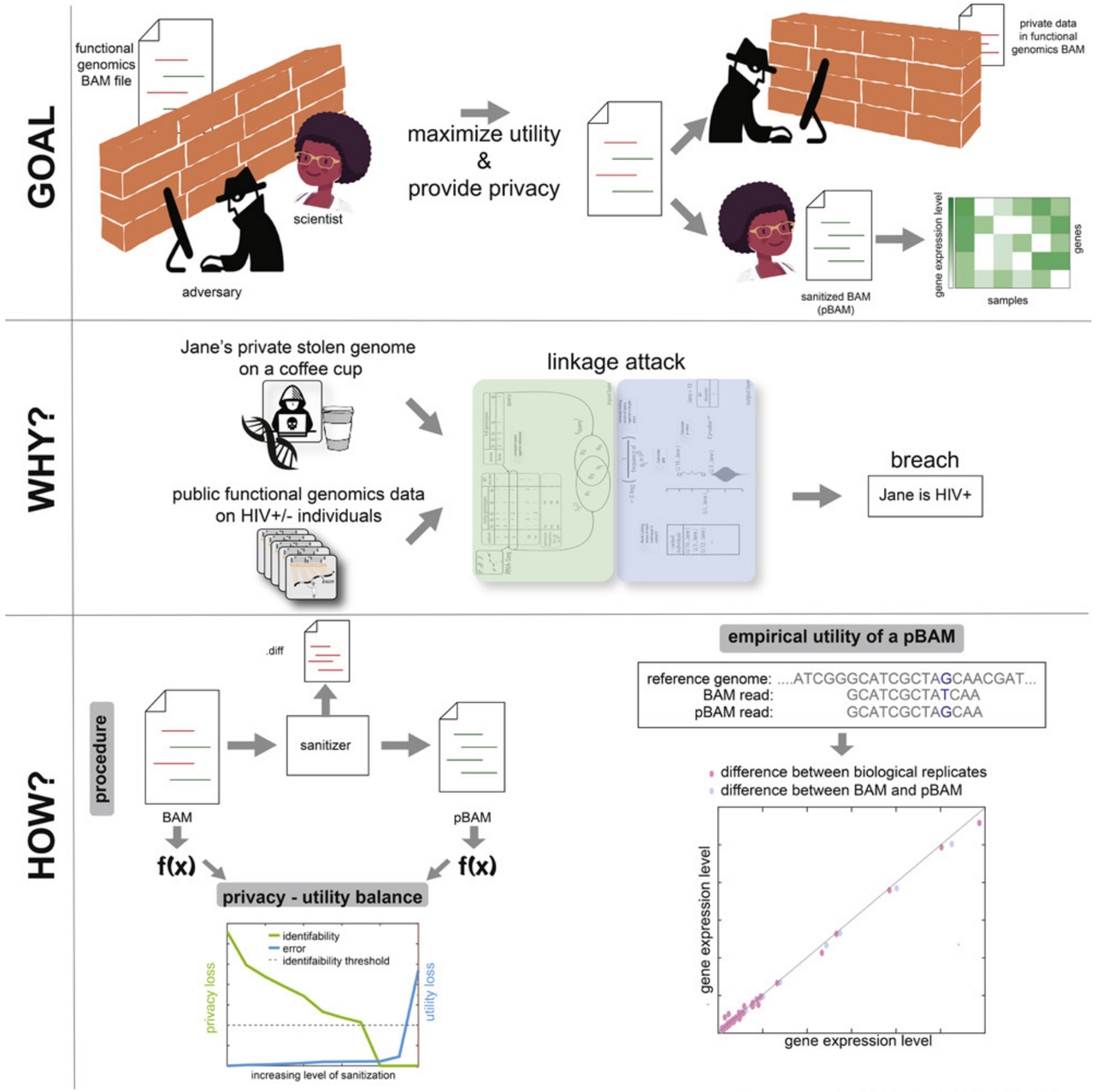
(C) The linking scores for each individual in the functional genomics cohort after the addition of genetically related individuals to the query, with and without the query individual present in the database.

Gürsoy et al., 2020, Cell *183*, 905–917
November 12, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cell.2020.09.036

Information Leakage from Functional Genomics Data

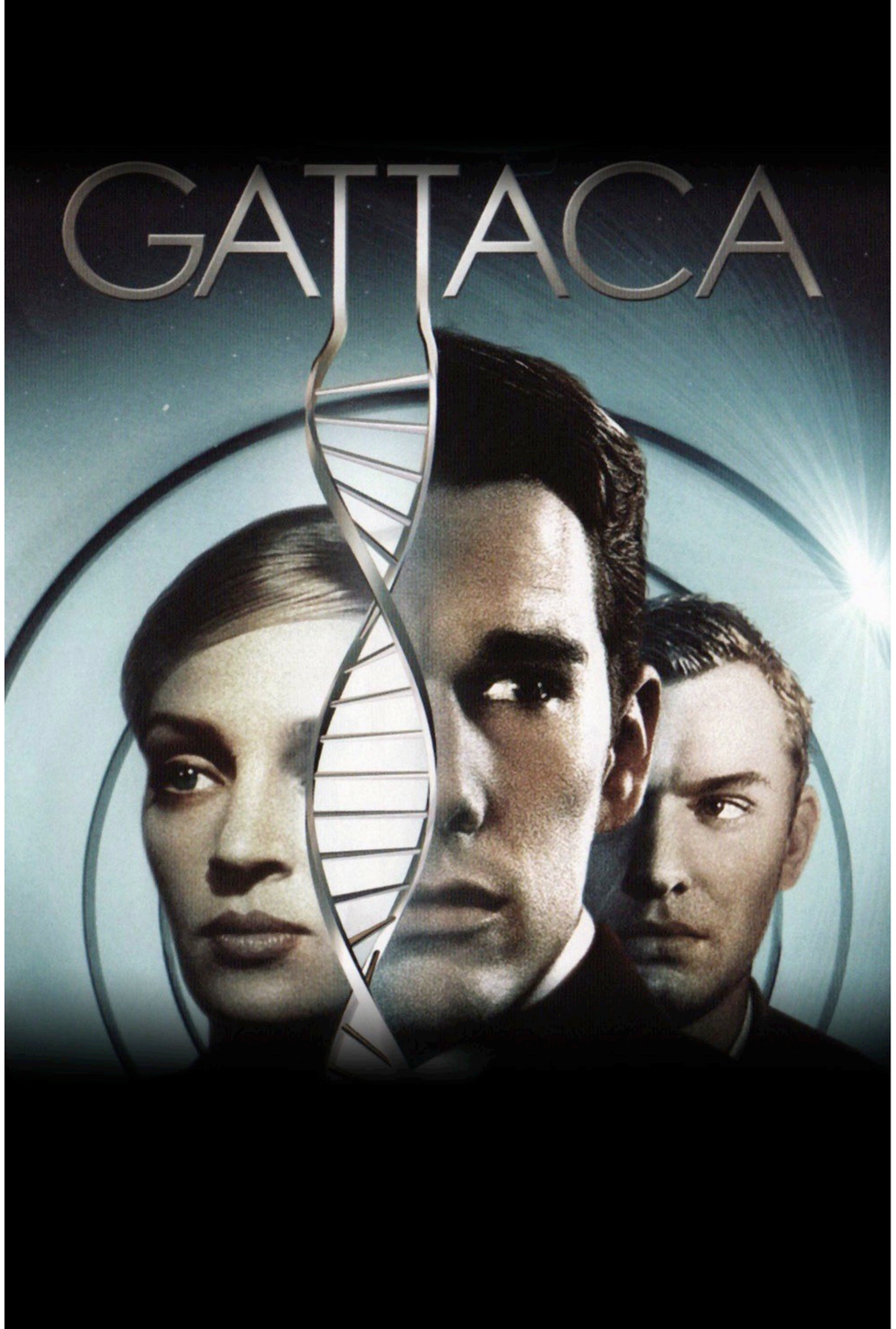
"Sanitize"...

- "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



Gürsoy et al., 2020, Cell *183*, 905–917
November 12, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cell.2020.09.036

Genomes
Privacy
Society

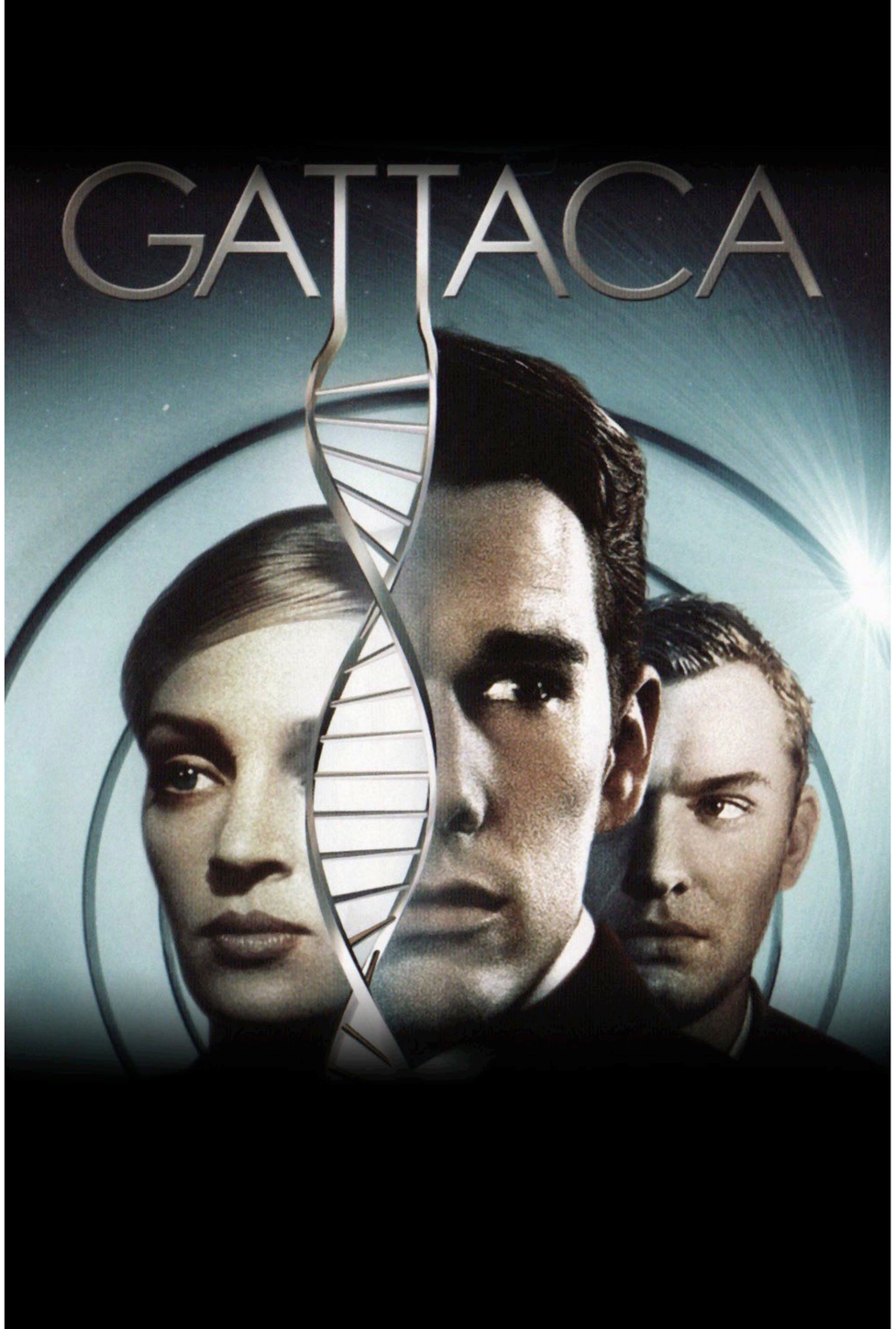


Gattaca (1997)

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

With information from https://www.imdb.com/title/tt0119177/

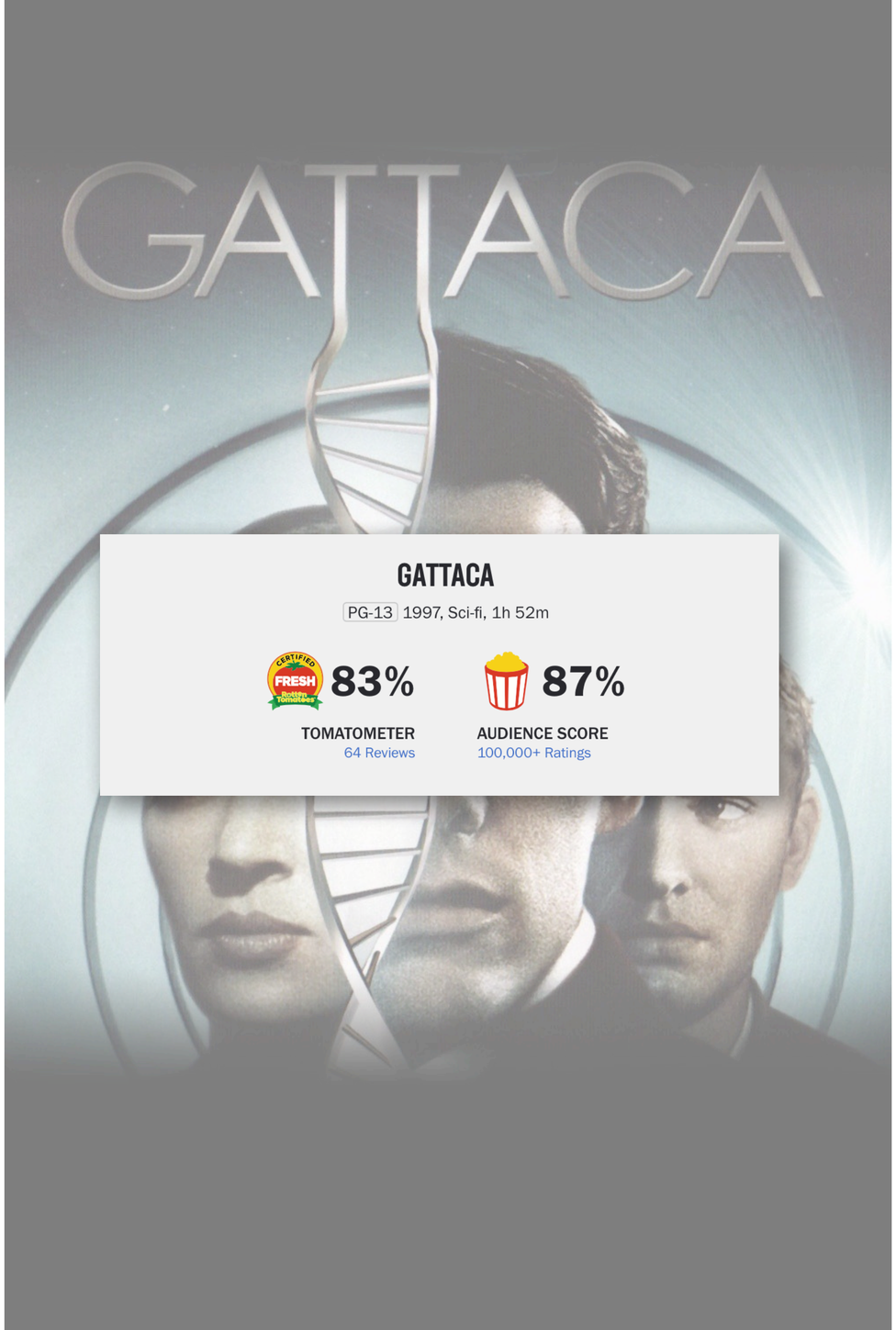


Gattaca (1997)

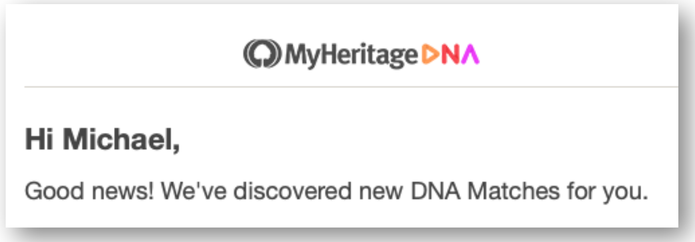
A genetically inferior man assumes the identity of a superior one in order to pursue his lifelong dream of space travel.

- 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

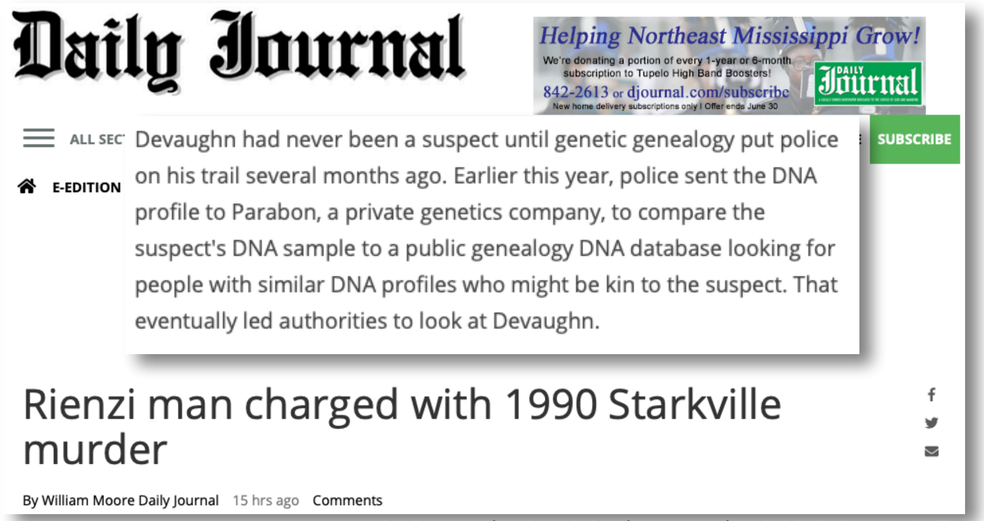
With information from https://www.imdb.com/title/tt0119177/



Long-Range Familial Searches



- Commercial, "Direct to Customer" DNA analyses are provided through independend 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

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. yyyyy — with ancestry records to connect Mr. xxxxx 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. xxxxx, 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. yyyyyy's boot, the affidavit said."

https://www.nytimes.com/2020/11/09/us/terrence-miller-suicide.html?referringSource=articleShare

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."

https://www.nytimes.com/2019/07/01/us/dna-genetic-genealogy-trial.html?referringSource=articleShare

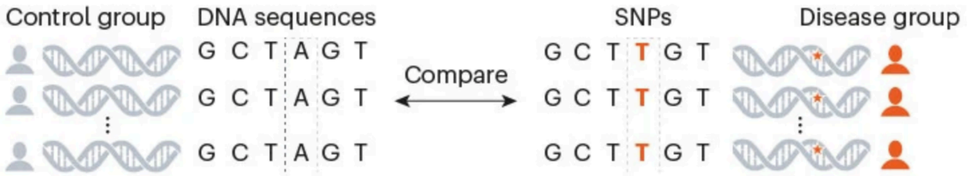
Polygenic Risk Scores for Embryo Selection

SCORING EMBRYOS

Some companies are selling genetic tests for embryos generated by *in vitro* fertilization that they say can identify an embryo with the lowest risk of developing common diseases, such as diabetes and some cancers. The method relies on sequencing the embryo's genome and comparing it with existing data on genetic risk factors in adult populations.

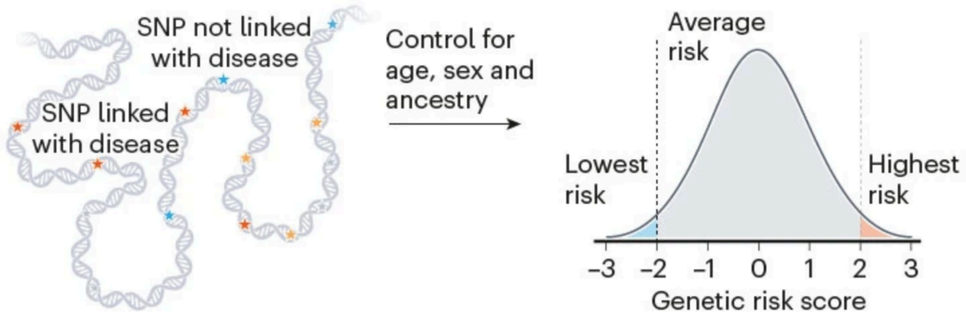
Genetics and risk

Researchers have discovered links between individual DNA letter changes, called single nucleotide polymorphisms (SNPs), and the risk of having a disease. To do this, they compare genome and health data from thousands of people with or without various diseases.



They look across the genome for SNPs that correlate with disease risk.

Combining information on all SNPs yields a prediction for how likely someone is to develop a disease.

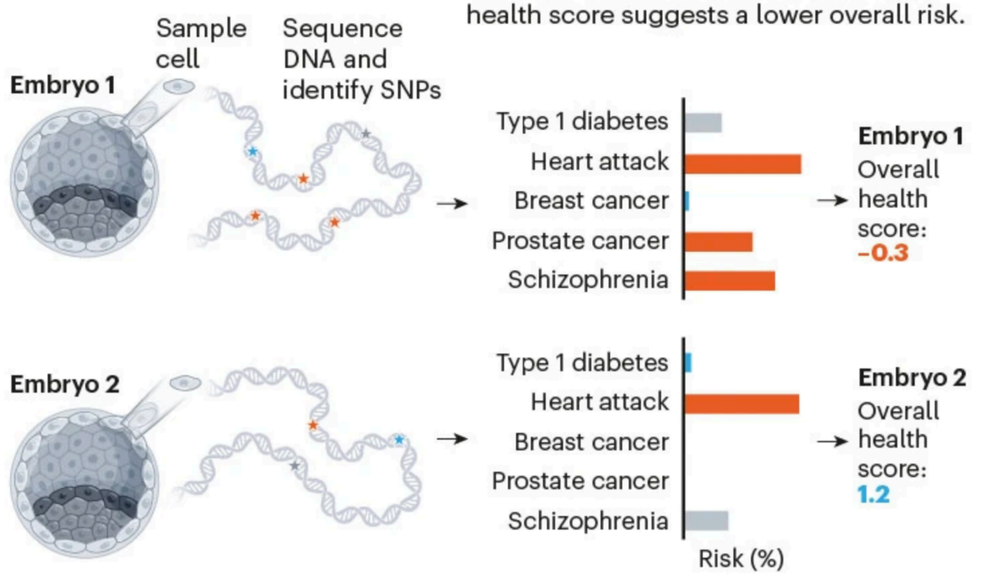


Embryo sampling

To apply this information to an embryo, clinicians take a few cells from embryos that are about 5 days old. They extract and sequence the DNA and look for SNPs.

Calculate scores

One company, Genomic Prediction, uses this method to help clients to identify embryos with a low risk of developing 12 disorders, including diabetes and some cancers. Each embryo's overall score is calculated by combining the risk of each disease and weighting them by their effect on life expectancy. A higher health score suggests a lower overall risk.



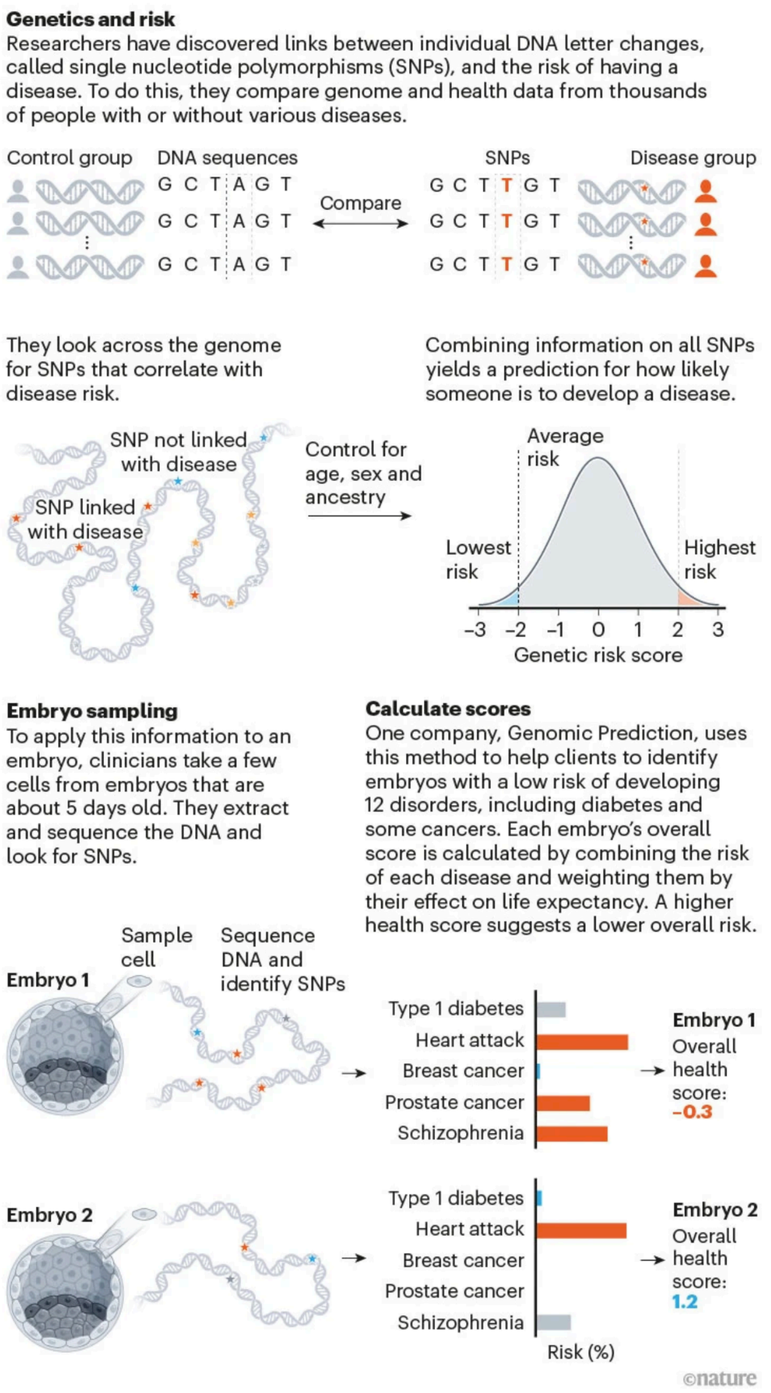
Source for 'Calculate scores': Genomic Prediction

Polygenic Risk Scores & IVF

"Designer Babies"?

- combination of polygenic risk score (PRS) assessment, *in vitro* fertilization (IVF) and pre-implantation diagnostics (PID / PGT)
- while justified for high-penetrance single gene diseases many problems are associated with PRS/IVF/PID:
  - lack of robust models - no easy transferable to individual case from population scale (ethnicity etc.)
  - "nature vs. nurture", unknown trade-offs, small effect size
  - possible antagonistic pleiotropy
  - health risks (mother!) of IVF have to be balanced against benefit
  - stigmatization, economic stratification lock-in if pervasive

> "She has her mother's eyes," begins the advertisement, "but will she also inherit her breast cancer diagnosis?"*



Source for 'Calculate scores': Genomic Prediction

* Genomic Prediction advert

https://www.nature.com/articles/d41586-022-02961-9

But genotyping 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

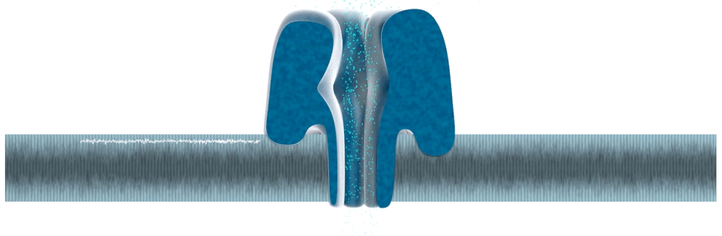
Identification of abandoned meterial using DNA fingerprinting is a common practice. The main challange currently being: time. Our method allows rapid sample preparation at the crime scene (see movie). We envision that the method can be adopted in the field for rapid checks, after a mass disaster, and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

Clinic

Clinics procces 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 implemnted in the clinic for rapid sanitiy-check of all incoming samples.

Cell line identification

Cross contamination of cell lines in science is a major problem. It results in unreproducible data, and clinical trails 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

MinION by Oxford Nanopore Technologies

The MinION is the smallest DNA sequencer currently around. Its 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:

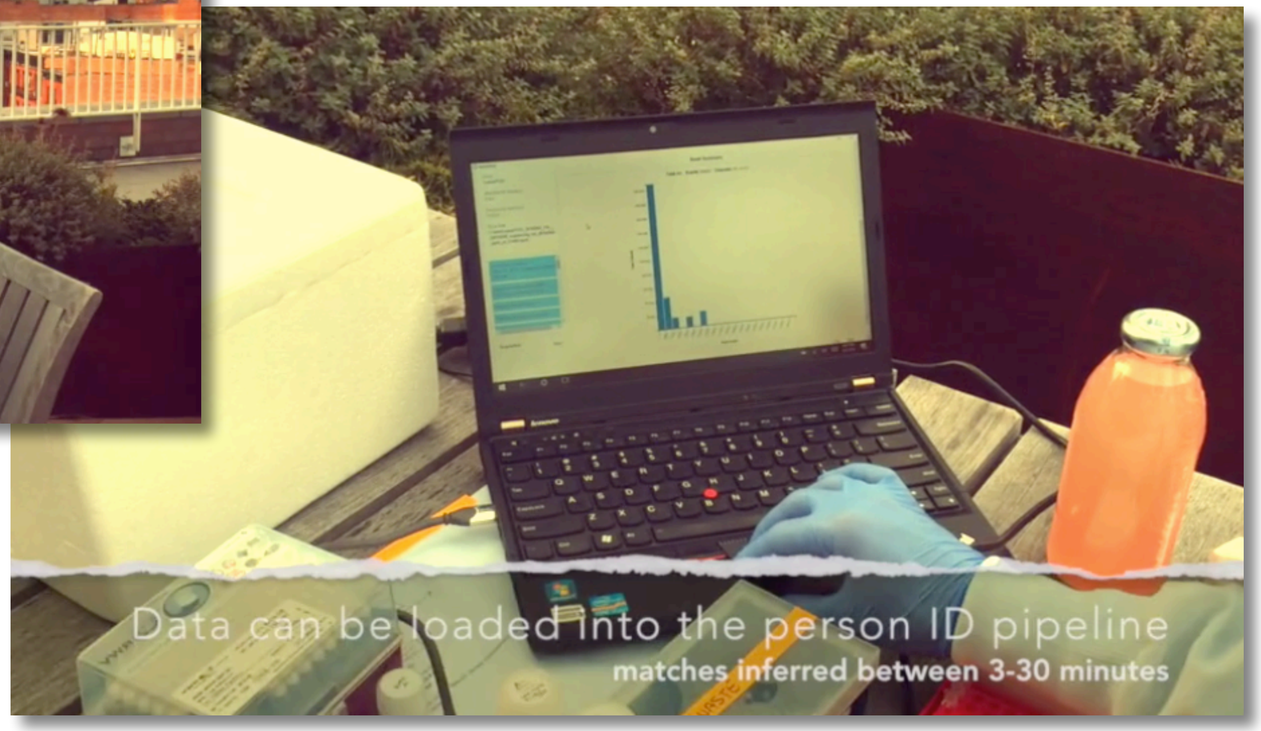
Oxford Nanopore Technologies

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



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

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 nail 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

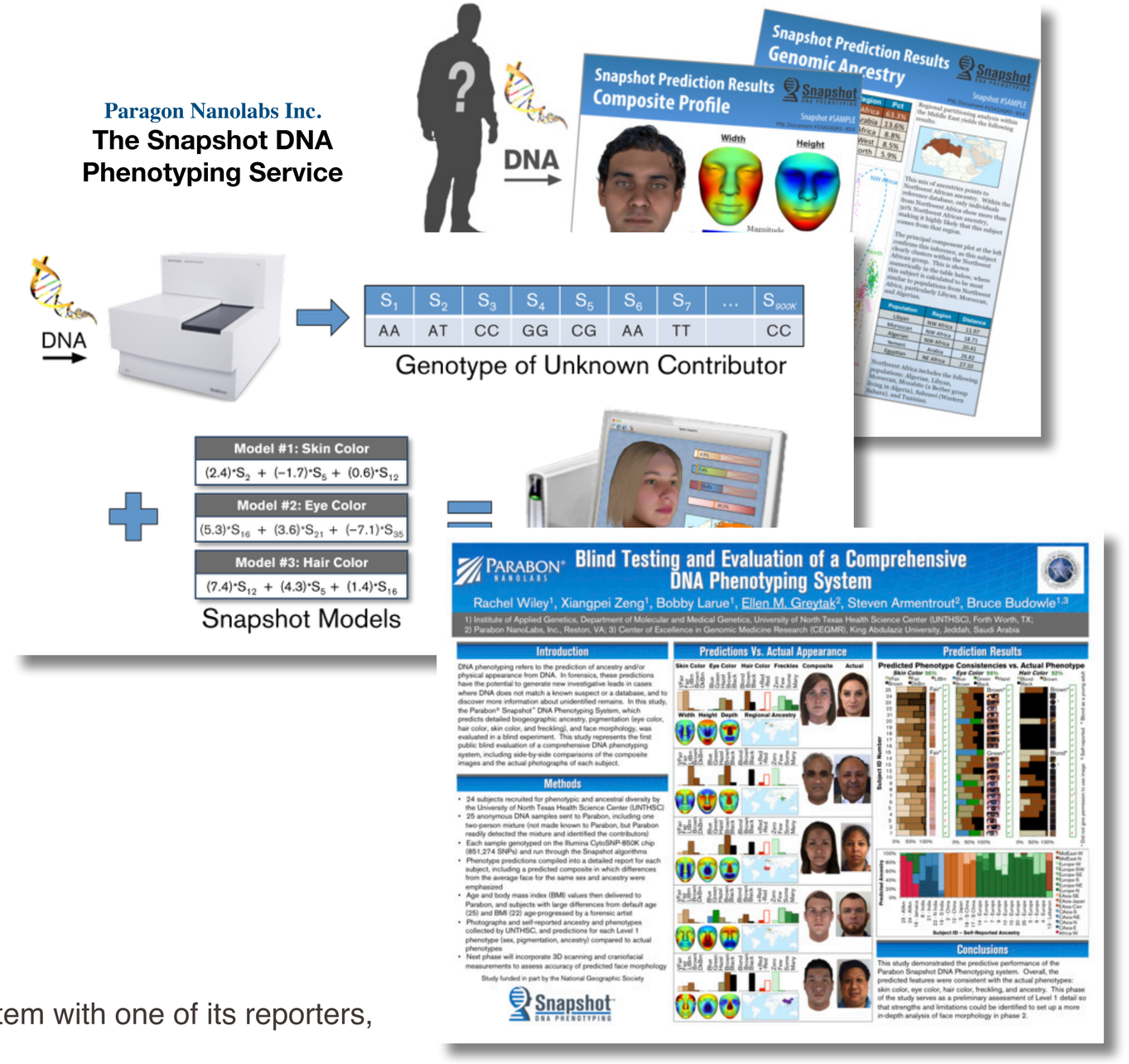
Fig. 1. Individual examples of HIrisPlex-based eye and hair color DNA prediction. Probability outcomes are provided for eye and hair color categories as obtained from complete HIrisPlex SNP profiles [50] using the enhanced IrisPlex eye color and the enhance HIrisPlex hair color prediction models [25] (http://www.erasmusmc.nl/fmb/resources/Irisplex_HIrisPlex/) 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 HIrisiPlex 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.

Manfred Kayser. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Science International: Genetics 18 (2015) 33–48

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 Parabon system with one of its reporters, it failed badly." (ACLU.org)

https://snapshot.parabon-nanolabs.com/phenotyping

Federal Act on the Use of DNA Profiles in Criminal Proceedings and for Identifying Unidentified or Missing Person, DNA Profiles Act

An Area in Transition...

- Currently: «Genetic Fingerprint»
- Future: Will it be allowed to take a deeper look and how far can genetic data be used to determine the characteristics of an unknown perpetrator (colour of hair and eyes, height, ethnicity, etc.)
- Switzerland: Bundesrat decision on 2020-12-04 to allow phenotyping for law enforcement purposes

Content provided by Julian Mausbach | Oberassistent für Strafrecht und Strafprozessrecht

HGTA : Federal Act on Human Genetic Testing

HGTA new 2021	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

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.)

Content provided by Julian Mausbach | Oberassistent für Strafrecht und Strafprozessrecht

Is Genomic Data Special?

Typical Data Scopes in Genomics (Research) Collections

Biomedical and procedural "Meta"data types

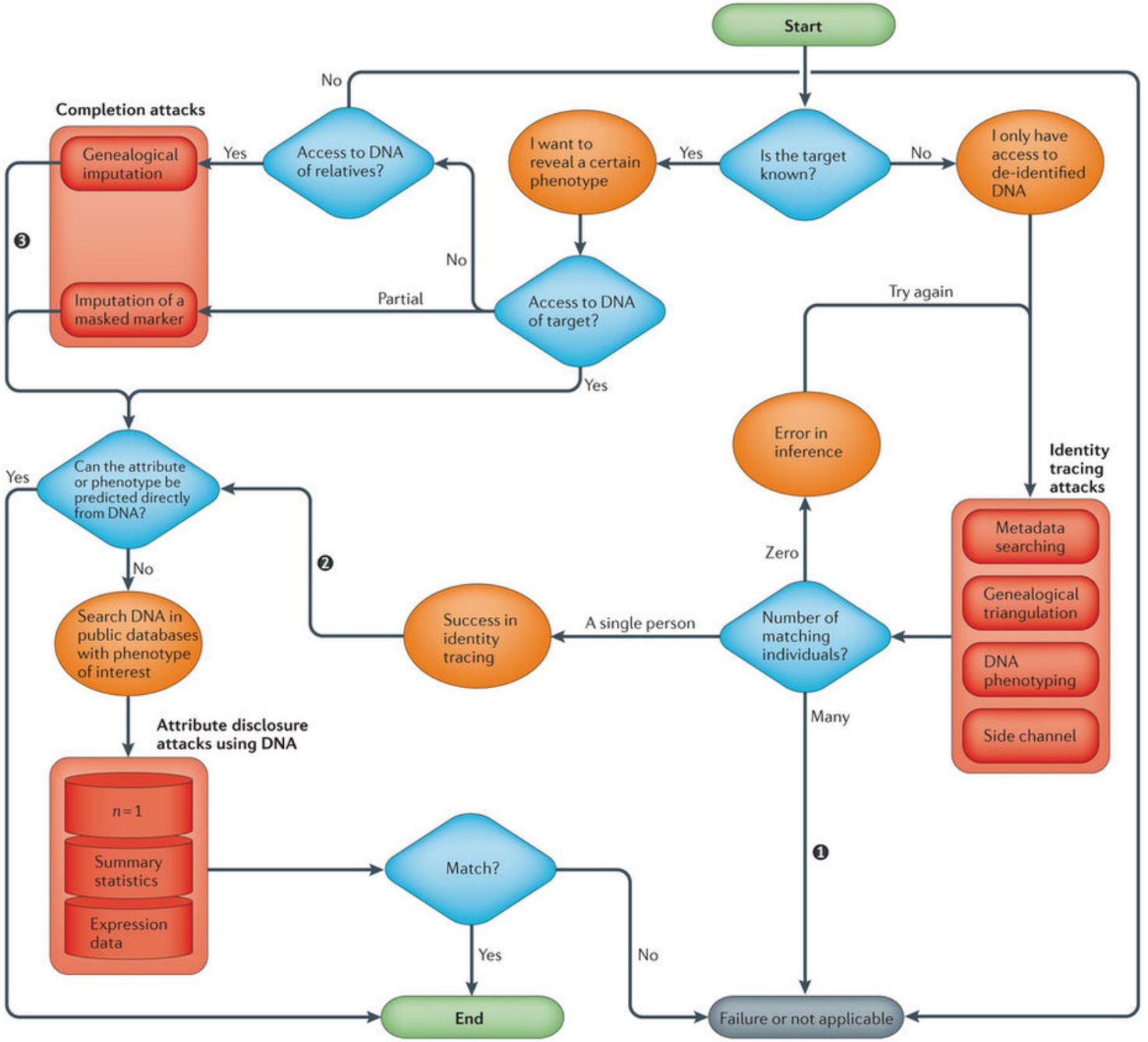
- Diagnostic classification
  - mapping text-based cancer diagnoses to standard classification systems
- Provenance data
  - store identifier-based pointers
  - geographic attribution (individual, biosample, experiment)
- Clinical information
  - **core set** of typical cancer study values:
    - ➡ stage,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

progenetix

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).

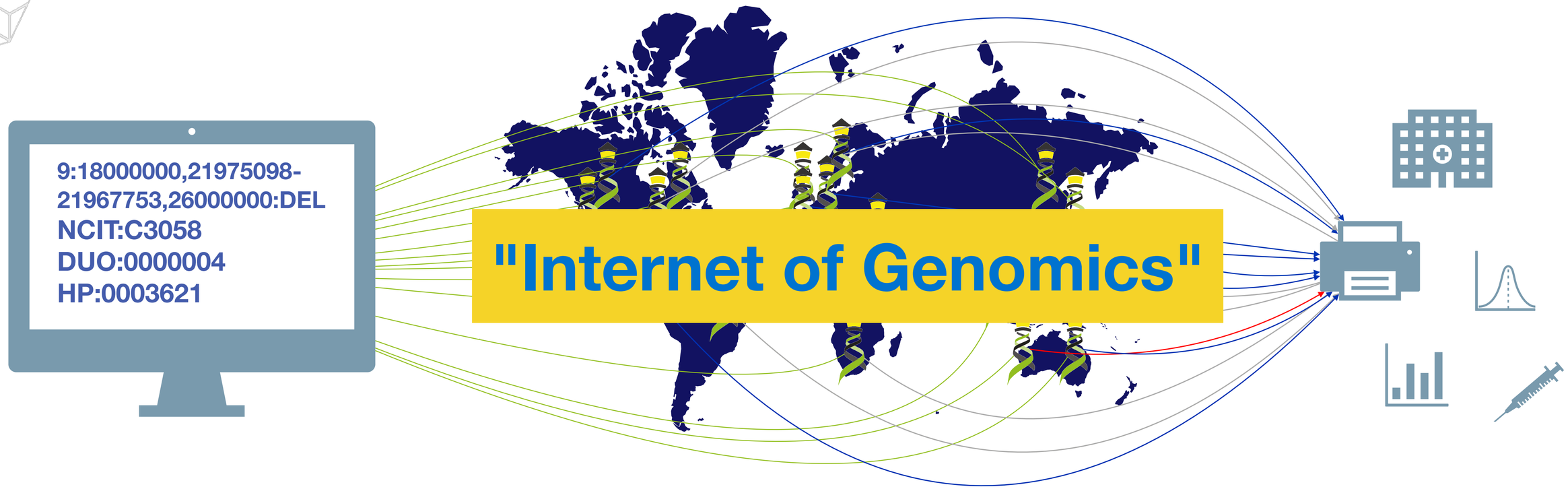
In certain scenarios (such as insurance decisions), uncertainty about the target's identity within a small group of people could still be considered a success (assumption 1). For certain privacy harms (such as surveillance), identity tracing can be considered a success and the end point of the process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



Yaniv Erlich & Arvind Narayanan. Nature Reviews Genetics 15, 409–421 (2014)

“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)



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 htsget, 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...)

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 htsget, 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!

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 htsget, 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...)

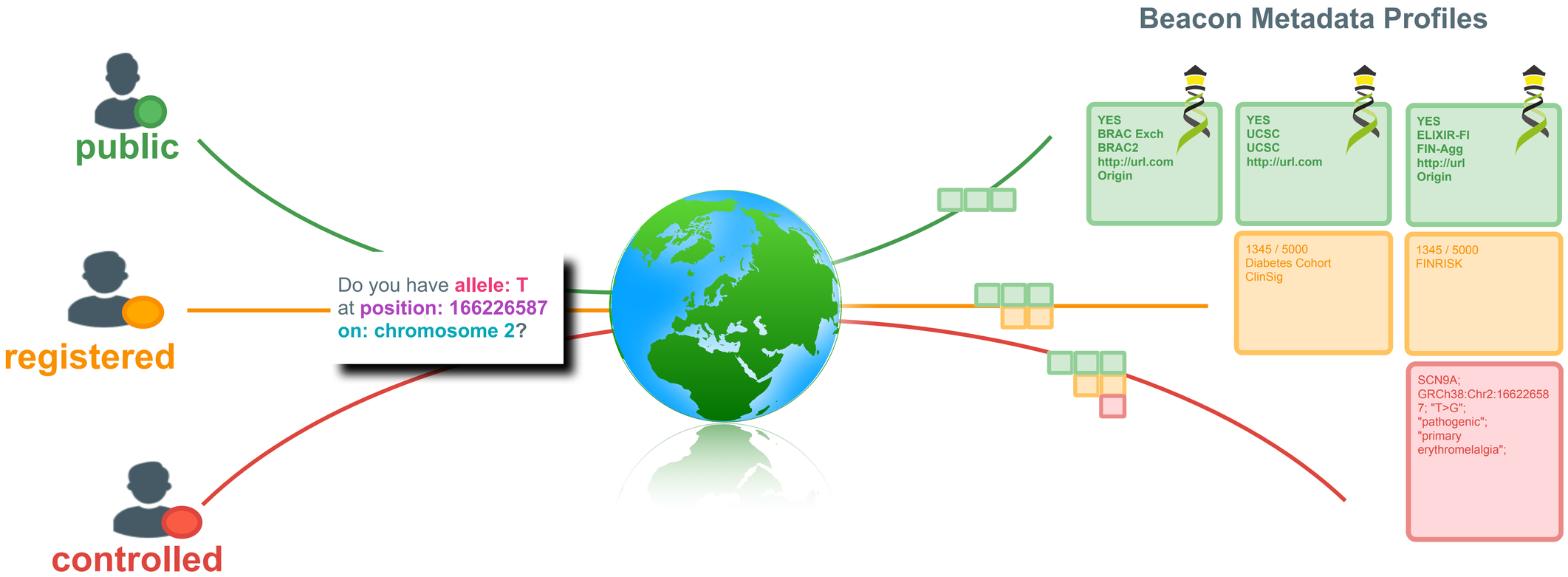
Definitely breaks the "Relative Security by Design" Concept!

Mitigation by tailored implementation and security practices

Global Alliance for Genomics & Health

Empowering Beacon use through Access Levels

Integrating permissions and discovery

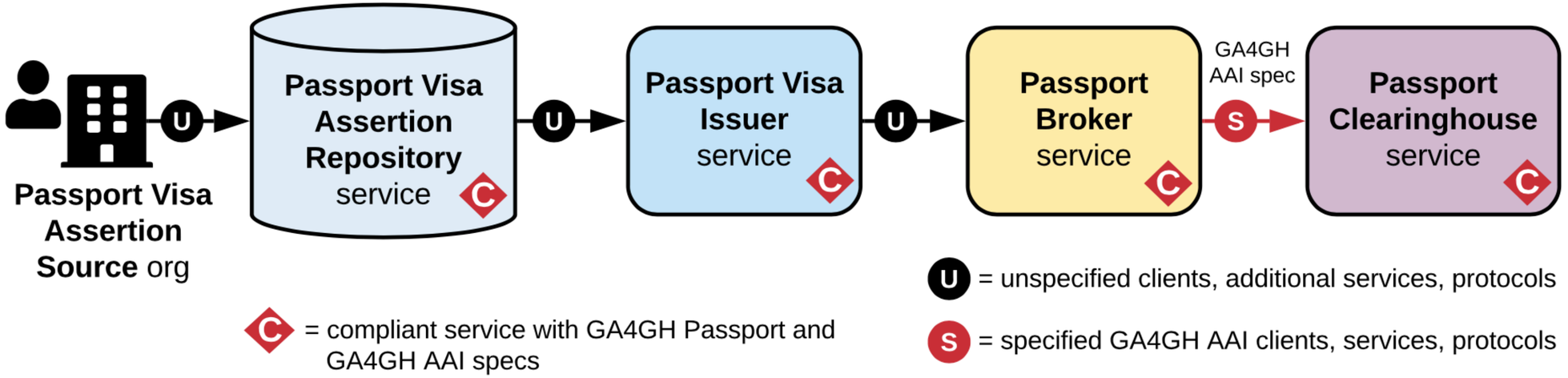


https://www.youtube.com/watch?v=LyfmvAs7LtQ&feature=youtu.be

GA4GH Passports

Global Alliance for Genomics & Health

Communicating a user's data access authorizations



www.ga4gh.org/ga4gh-passports/

- 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

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

Generalkonsent

BENEFIT

BLOCKCHAIN

HEALTH

PRIVACY

SECURITY

CONSENT

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*

Hum Genet (2014) 133:895–903
DOI 10.1007/s00439-014-1432-6

ORIGINAL INVESTIGATION

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

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.

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

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,

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 Dyke[1], Edward S Dove[2] and Bartha M Knoppers[1]

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

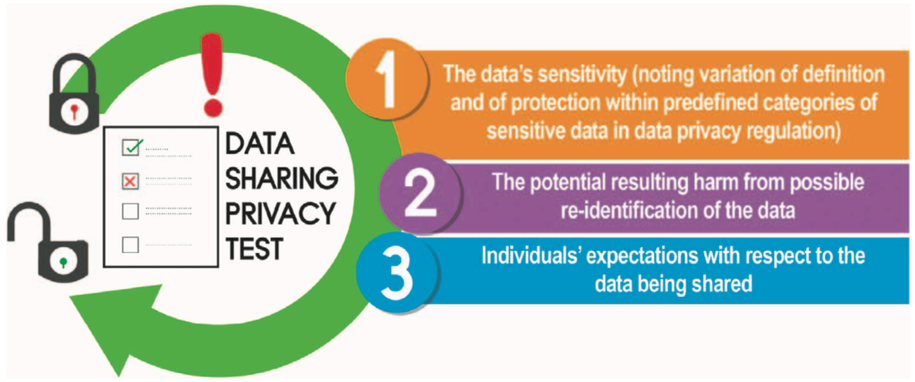
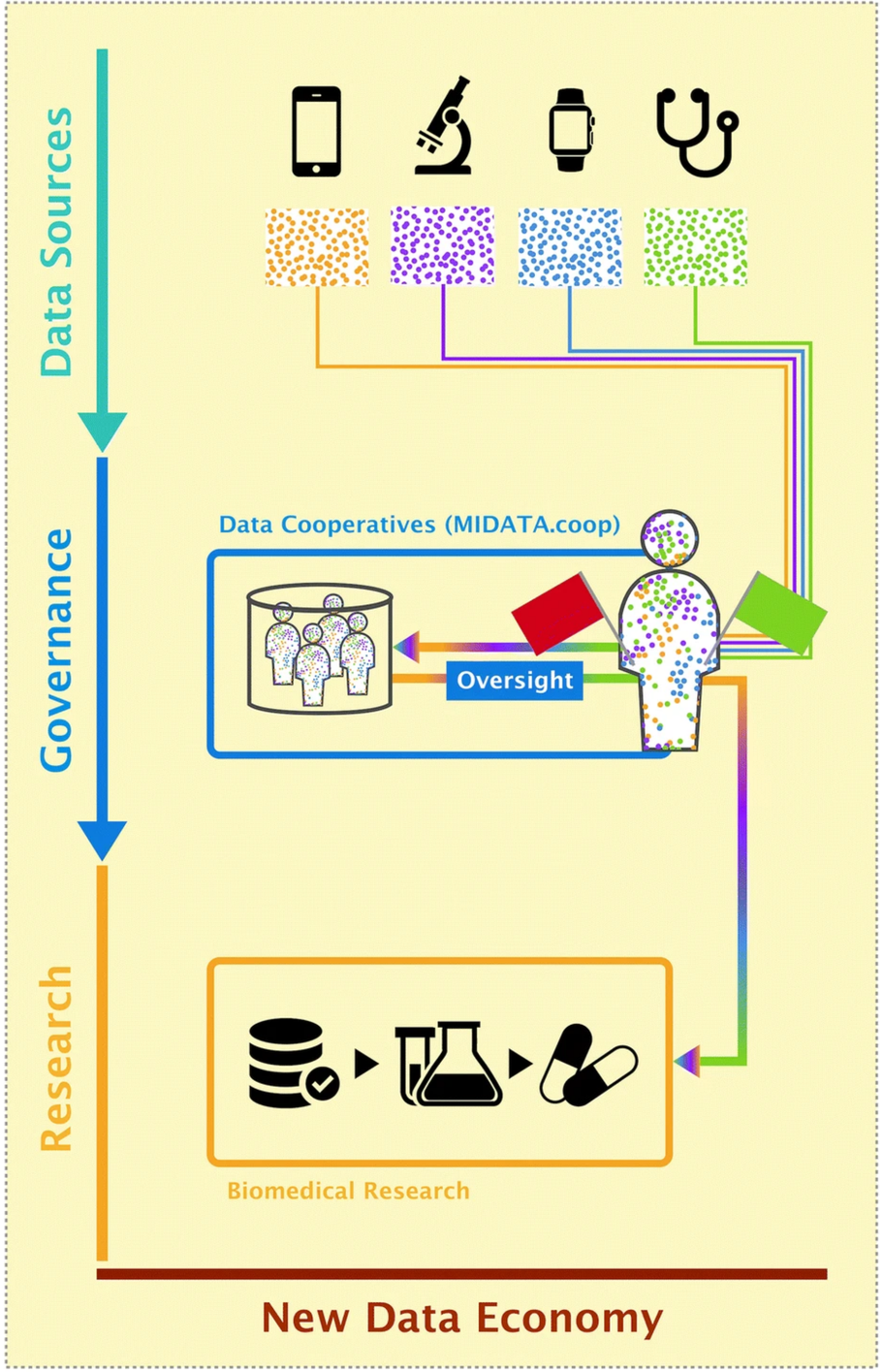


Figure 1. The three steps of a Data Sharing Privacy Test to distinguish degrees of data sensitivity within categories of data recognised as 'sensitive'.

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

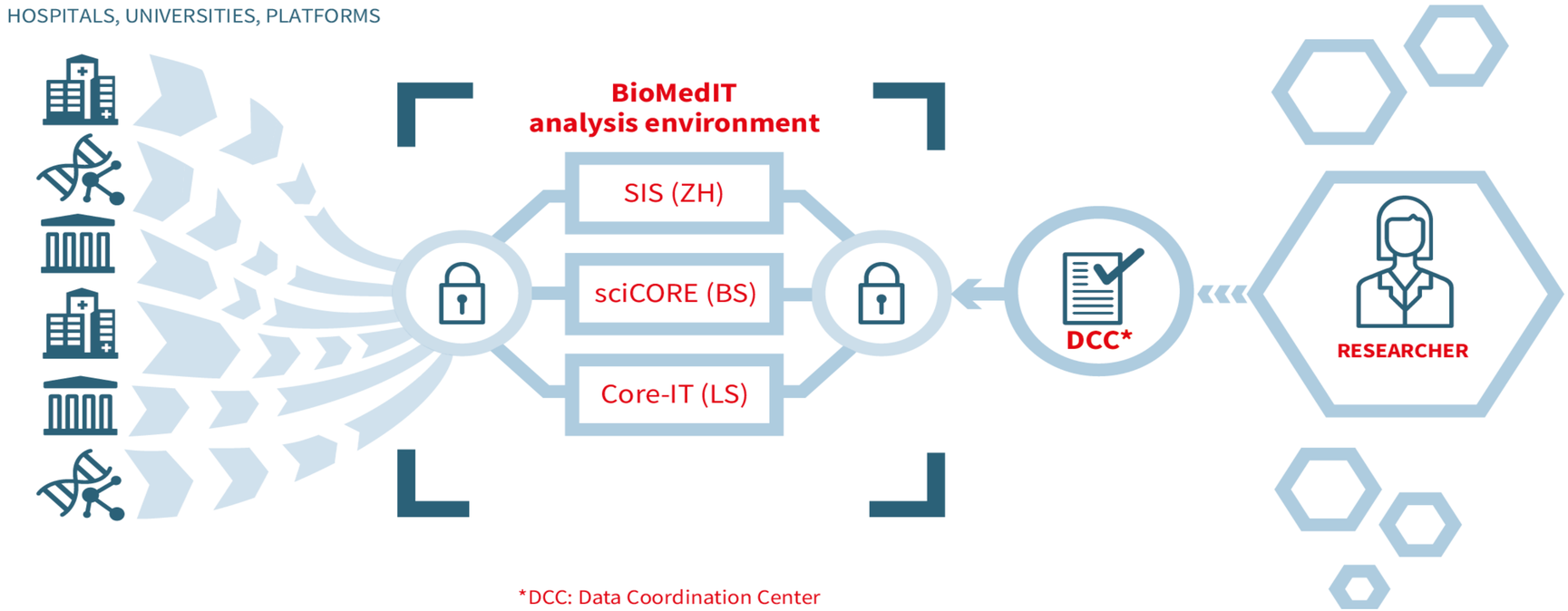


Citizens aggregate data from different sources and make them available for research through data cooperatives. Cooperatives offer oversight mechanisms to filter data access requests and tools for the democratic governance of the data.
Blasimme, A., Vayena, E. & Hafen, E. **Democratizing Health Research Through Data Cooperatives**. *Philos. Technol.* 31, 473–479 (2018). https://doi.org/10.1007/s13347-018-0320-8

graphic credit Manuel Schneider

The BioMedIT network

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



*DCC: Data Coordination Center

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**!

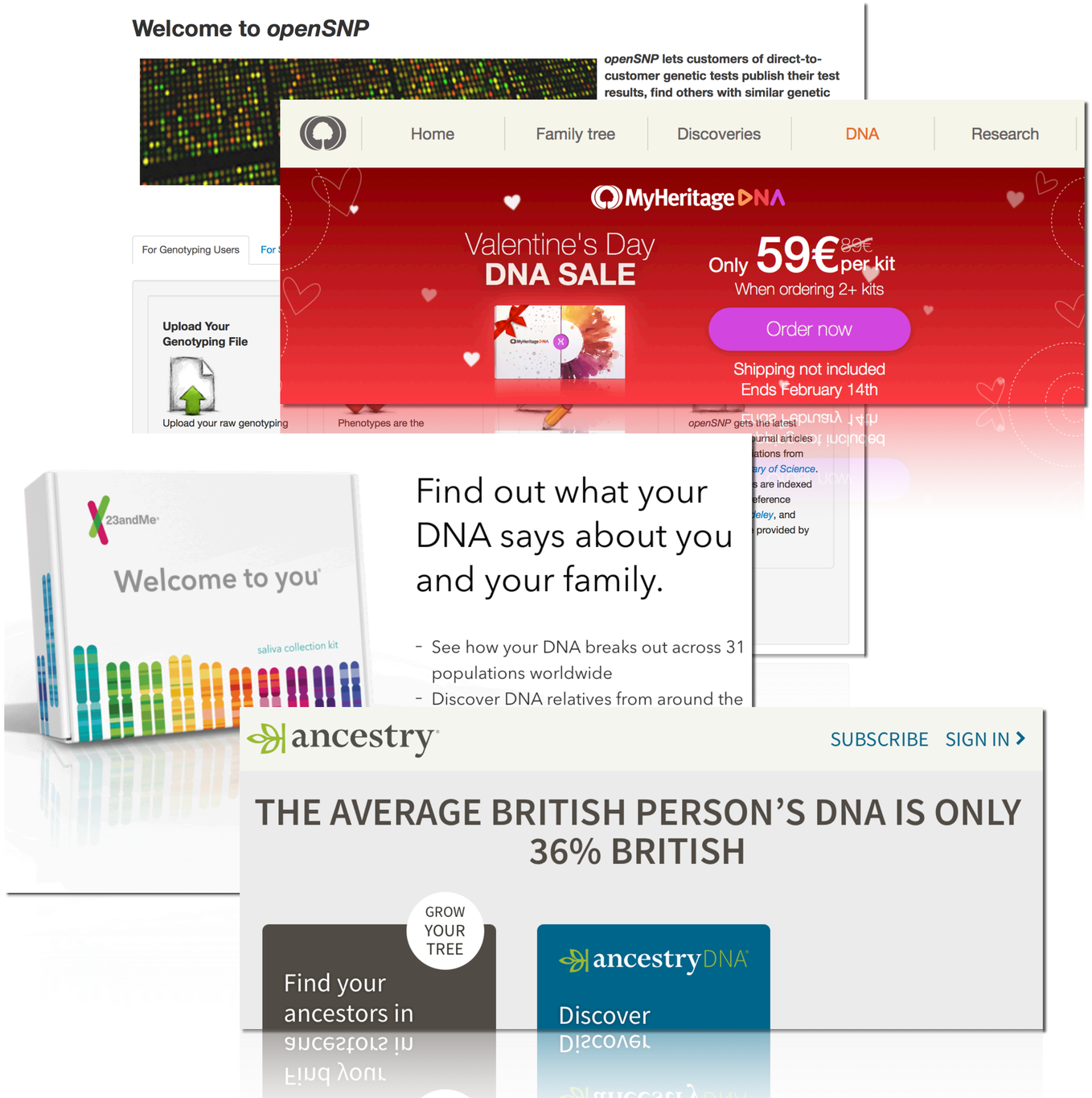
Genomic Data & Privacy - Some Take-Home Messages

- 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!



John Yuyi, NYT 2018-02-09

Welcome to *openSNP*

openSNP **lets customers of direct-to-customer genetic tests publish their test results, find others with similar genetic**

For Genotyping Users

Upload Your Genotyping File

Upload your raw genotyping

Phenotypes are the

openSNP gets the latest

Home | Family tree | Discoveries | DNA | Research

MyHeritage DNA

Valentine's Day
DNA SALE

Only **59€** ~~89€~~ per kit
When ordering 2+ kits

Order now

Shipping not included
Ends February 14th

23andMe

Welcome to you

saliva collection kit

Find out what your DNA says about you and your family.

- See how your DNA breaks out across 31 populations worldwide
- Discover DNA relatives from around the

ancestry

SUBSCRIBE SIGN IN

THE AVERAGE BRITISH PERSON'S DNA IS ONLY 36% BRITISH

GROW YOUR TREE

Find your ancestors in

ancestryDNA

Discover

BIO392 HS22

Exam

- 2022-10-12
- time: 09:30-10:30
- multiple (single + multiple) choice w/ one or two open questions