BIO392
Bioinformatics of Genome Variations

Ontologies | Privacy

Michael Baudis UZH SIB
Computational Oncogenomics

University of Zurich[UZH]

BIO320: Course Resources

University of Zurich[UZH]

- Course repository & website on Github
  - links to articles and information resources
  - downloads

https://baudisgroup.github.io/BIO392-Github/
https://github.com/baudisgroup/BIO392-Github

BIO392-Github

View the Project on GitHub
baudisgroup/BIO392-Github

This project is maintained by
baudisgroup

Hosted on GitHub Pages — Theme by orderedlist

BIO392

This is a repository for materials related to the BIO392 *Bioinformatics of Genome Variation* introductory course at the University of Zürich, Autumn 2018.

The home page of this repository can be found at https://baudisgroup.github.io/BIO392-Github/.

Abstract

One of the fastest growing areas of bioinformatics is in the analysis, warehousing and representation of genomic and protein sequence variants, particularly with view on the use of molecular data in personalised health and biomedical applications in general. This course will engage participants to explore common data formats, online resources and analysis techniques, with a focus on human genome variation data.

Links

- BIO392 in the UZH OLAT system
- BIO392 in the UZH directory

Literature and Resources

- articles and other reading (files)
- article links
- resource links (browsers and online repositories)

Schedule

The course schedule consists of 1 afternoon (Tuesday) and 3 “full” days (Wednesday - Friday) in the first two weeks, with 1.5 days in the last week.

The “full days” run from 09:00 - 17:00, with a break from 12:00 - 13:30.

Conceptually, mornings are predominantly related to introductions, presentations and discussion of the previous days, and are afternoons reserved for independent work on the examples and tasks.

Bioinformatics: **Ontologies**

- ontologies in information sciences describe concrete and abstract **objects**, there precisely defined **hierarchies** and **relationships**
- ontologies in bioinformatics support the move from a descriptive towards an **analytical science** in describing biological data and relations among it

"The widest use of ontologies within biology is for conceptual annotation – a representation of stored knowledge more computationally amenable than natural language."*

- Gene ontology (GO)
- NCIt Neoplasm Core
- Uberon anatomical structures
- Experimental Factor Ontology (EFO)
- Disease Ontology (DO)

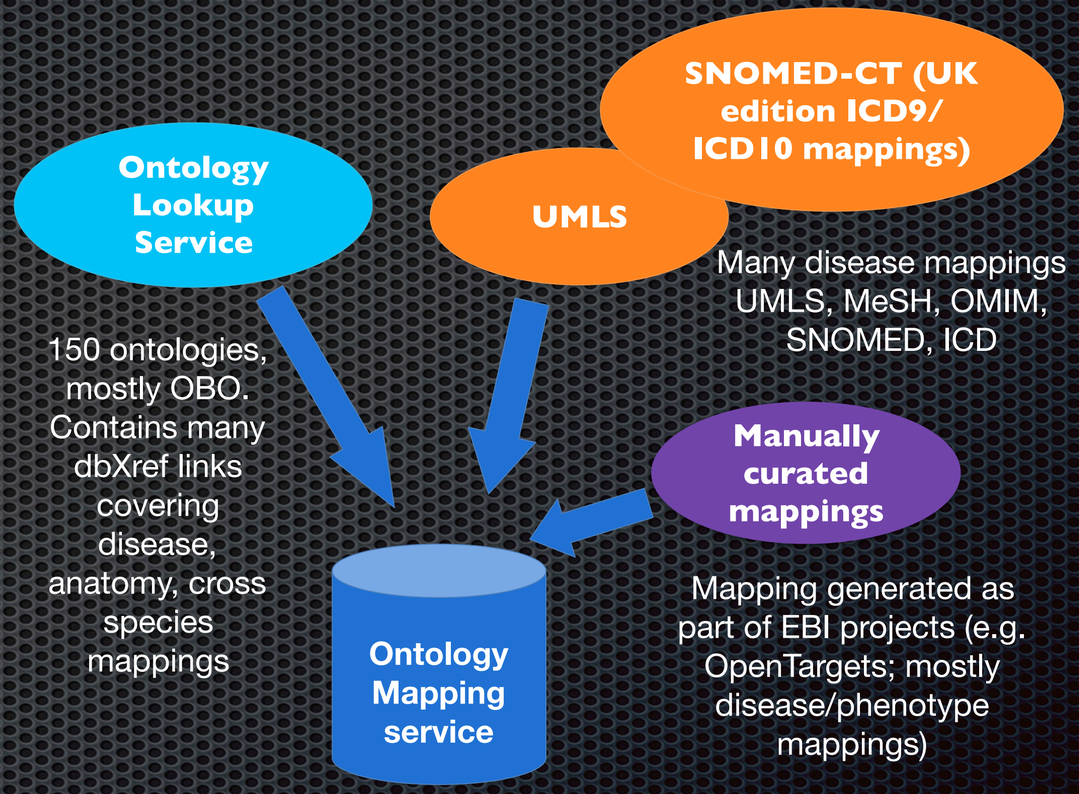
```
id: GO:0000118
name: histone deacetylase complex
namespace: cellular_component
def: "A protein complex that possesses histone deacetylase activity." [GOC:mah]
comment: Note that this term represents a location, not a function; the activity possessed by this complex is mentioned in the
definition for the purpose of describing and distinguishing the complex. The function of this complex is represented by the
molecular function term 'histone
synonym: "HDAC complex" EXACT [
is_a: GO:0044451 ! nucleoplasm
is_a: GO:1902494 ! catalytic co
```

- ⊟Neoplasm by Morphology
  - ⊞ Epithelial Neoplasm C3709
  - ⊞ Germ Cell Tumor C3708
  - ⊞ Giant Cell Neoplasm C7069
  - ⊞ Hematopoietic and Lymphoid Cell Neoplasm C27134
  - ⊟ Melanocytic Neoplasm C7058
    - ⊞ Benign Melanocytic Skin Nevus C7571
    - • Dysplastic Nevus C3694
    - ⊟ Melanoma C3224
      - • Amelanotic Melanoma C3802
      - ⊞ Cutaneous Melanoma C3510
      - • Epithelioid Cell Melanoma C4236
      - • Mixed Epithelioid and Spindle Cell Melanoma C66756
      - • Non-Cutaneous Melanoma C8711
      - • Spindle Cell Melanoma C4237
  - ⊞ Meningothelial Cell Neoplasm C6971

Ontologies in bioinformatics. R Stevens, C Wroe, P Lord and C Goble, Department of Computer Science, University of Manchester. Retrieved from http://homepages.cs.ncl.ac.uk/phillip.lord/download/publications/handbook.pdf

Making Ontologies Work for GA4GH Implementation Studies

- biomedical "metadata" in different resources frequently follows incompatible classification systems
- medical coding systems are driven by different paradigms compared to biological ontologies (e.g. for cross-species comparisons)
- frequently used classifications (ICD, Snomed…) are either not “ontologised” or cannot be referenced in open resources



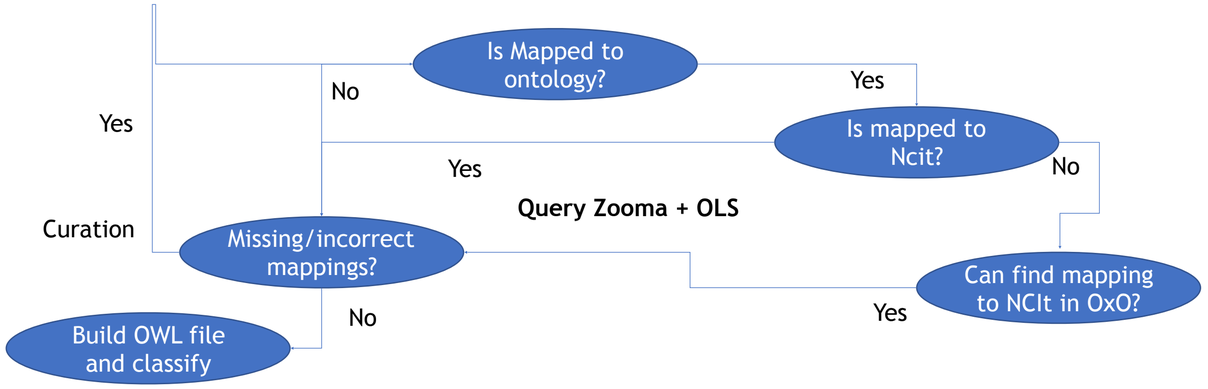
Federated queries across resources need **curated mappings** of classifications/ ontologies

Ontology Mapping Service OXO

Working towards ontologies w/ arrayMap: Mapping >55'000 samples from ICD-O to NCIt neoplasm core

ICDM	ICDMORPHOLOGY	NCItcode	NCItlabel	ICDT	ICDTOPOGRAPHY	NCItcode	NCItlabel	NCIt_mappi	NCIt_mapped_ICDM_T_label
8021/3	Carcinoma anaplastic type			C739	thyroid gland	C12400	thyroid gland	C3878	Thyroid Gland Undifferentiated (Anaplastic) Carcinoma
9451/3	Oligodendroglioma anaplastic	C4326	anaplastic oligodendroglioma	C719	Brain	C12439	brain	C4326	Anaplastic Oligodendroglioma
9051/3	Desmoplastic mesothelioma	C6747		C499	connective and soft tissue			C6747	Desmoplastic Mesothelioma
9732/3	Plasma cell myeloma	C3242	multiple myeloma	C42	hematopoietic and reticuloendothelial systems			C3242	Plasma Cell Myeloma
8070/3	Squamous cell carcinoma	C2926	non-small cell lung carcinoma	C140	pharynx			C102872	Pharyngeal Squamous Cell Carcinoma
8380/3	Endometrioid adenocarcinoma	C3769	endometrioid carcinoma	C54	corpus uteri	C12316		C6287	Endometrial Endometrioid Adenocarcinoma
8070/3	Squamous cell carcinoma	C2926	non-small cell lung carcinoma	C44	skin	C12470	zone of skin	C4819	Skin Squamous Cell Carcinoma
8430/3	Mucoepidermoid carcinoma	C45544	pulmonary mucoepidermoid carcinoma	C089	salivary gland	C12426	saliva-secreting gland	C5953	Minor Salivary Gland Mucoepidermoid Carcinoma
9680/3	Diffuse large B-cell lymphoma	C8851	diffuse large B-cell lymphoma	C42	hematopoietic and reticuloendothelial systems			C8851	Diffuse Large B-Cell Lymphoma
8800/3	Sarcoma	C9118	sarcoma	C559	uterus nos			C9306	Soft Tissue Sarcoma
8441/3	Serous adenocarcinoma	C7550	ovarian serous adenocarcinoma	C570	fallopian tube	C12403	fallopian tube	C40101	Serous Adenocarcinoma
9689/3	splenic marginal zone lymphoma nos			C422	spleen	C12432	spleen	C4663	Splenic Marginal Zone Lymphoma
8077/2	Squamous intraepithelial neoplasia grade III			C53	cervix uteri	C12311		C89476	Grade III Vaginal Intraepithelial Neoplasia
8140/0	Adenoma	C4196	adenoma	C189	large intestine excl. rectum and rectosigmoid junction			C4349	Colon Adenocarcinoma
8272/3	Pituitary carcinoma	C4536	Pituitary carcinoma	C751	pituitary gland	C12399	pituitary gland	C4536	Pituitary Gland Carcinoma
8500/2	Ductal carcinoma in situ	C3641	ductal carcinoma in situ	C50	breast	C12971	breast	C2924	Ductal Breast Carcinoma In Situ
8200/3	Adenoid cystic carcinoma	C2970	adenoid cystic carcinoma	C32	larynx	C12420	larynx	C2970	Adenoid Cystic Carcinoma
9370/3	Chordoma	C2947	Chordoma	C419	bone	C13076	bone tissue	C2947	Chordoma
9717/3	Enteropathy type T-cell lymphoma			C17	small intestine	C12386	small intestine	C4737	Enteropathy-Associated T-Cell Lymphoma
9698/3	Follicular lymphoma grade 3			C42	hematopoietic and reticuloendothelial systems			C3460	Grade 3 Follicular Lymphoma
9863/3	Chronic myeloid leukemia	C3177	chronic myelogenous leukemia	C42	hematopoietic and reticuloendothelial systems			C3174	Chronic Myelogenous Leukemia BCR-ABL1 Positive
8852/3	Liposarcoma myxoid	C3735	myxoid liposarcoma	C499	connective and soft tissue			C27781	Myxoid Liposarcoma
9080/3	Teratoma malignant			C809	unknown	C35882	Hereditary elliptocytosis	C3403	Tetratoma
8530/3	Inflammatory carcinoma	C4872	breast carcinoma	C50	breast	C12971	breast	C4001	Inflammatory Breast Carcinoma
8140/3	Adenocarcinoma	C27745	lung adenocarcinoma	C809	unknown	C35882	Hereditary elliptocytosis	C2852	Adenocarcinoma
						C12762	oropharynx	C2970	Adenoid Cystic Carcinoma
						C12415	kidney	C3158	Leiomyosarcoma
						C12499	internal ear	C2970	Adenoid Cystic Carcinoma
						C12683	bronchus	C2923	Bronchioloalveolar Carcinoma
						C12343	retina	C3224	Melanoma
					ial systems			C8459	Hepatosplenic T-Cell Lymphoma
						C12393	pancreas	C8294	Pancreatic Adenocarcinoma
					ial systems			C3996	Monoclonal Gammopathy of Undetermined Significance
								C4817	Ewing Sarcoma
								C3288	Oligodendroglioma
						C12422	tongue	C4648	Tongue Squamous Cell Carcinoma
					ial systems			C2862	Primary Myelofibrosis
								C4833	Oral Cavity Squamous Cell Carcinoma
						C12390	rectum	C9383	Rectal Adenocarcinoma
								C3158	Leiomyosarcoma
					ial systems			C3898	Extranodal Marginal Zone Lymphoma of Mucosa-Associate
						C12404	female gonad	C4512	Ovarian Mucinous Cystadenoma

- From 456 pairs of ICD-O terms Morphology and Topography representative of cancer entities in arrayMap
- Develop Python script to take ICD-O Morphology and Topography labels separately QUERY ZOOMA, OxO and OLS to find mapping to NCIt



From 456 pairs of ICD-O
70% ICD-O Morphology - NCIt
65% ICD-O Topography - NCIt

45% ICD-O-3 Pairs mapped to NCIt terms

=> MANUAL CURATION of >50%

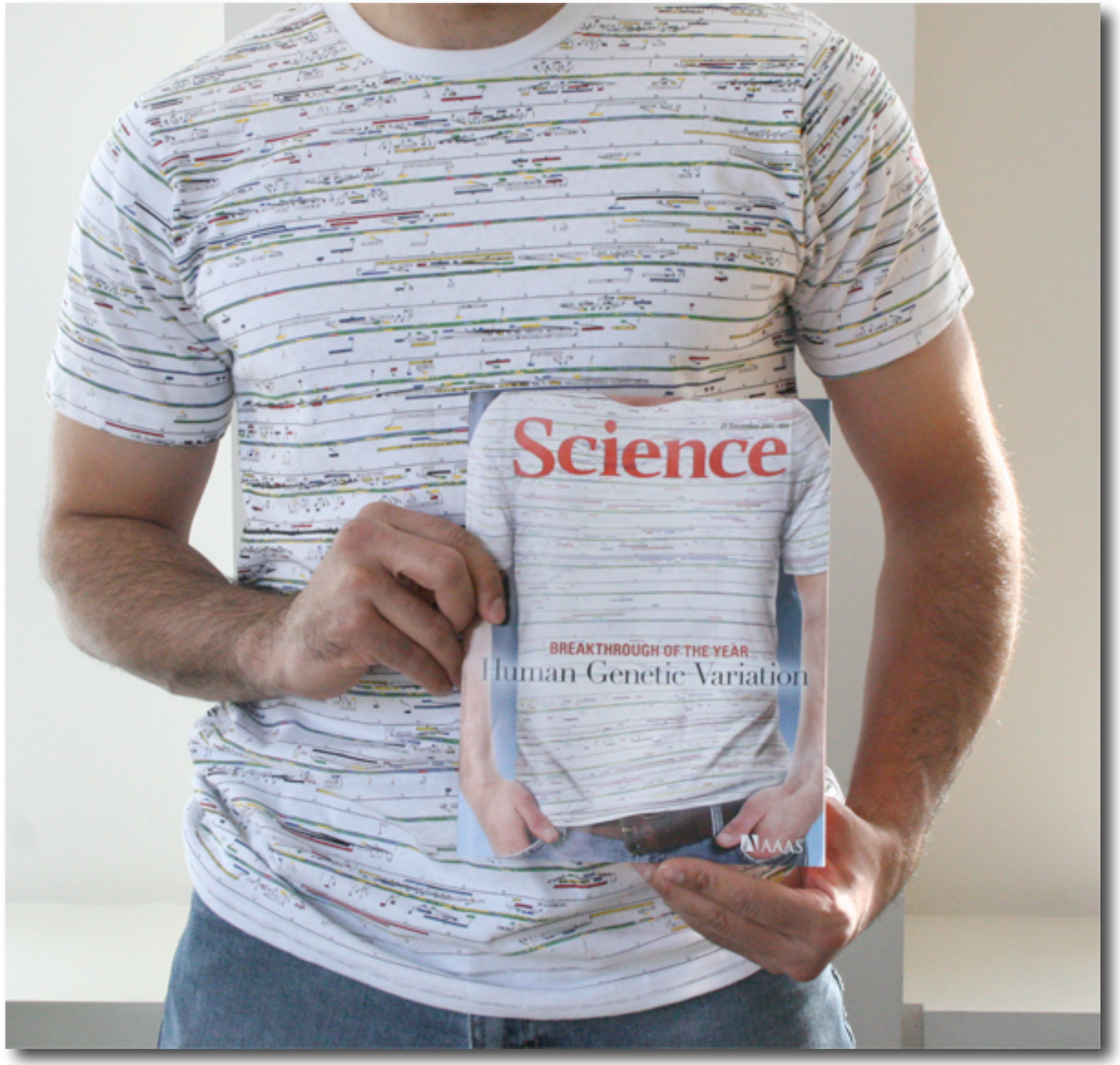
BIO392
Bioinformatics of Genome Variations

Genomes: Core of "Personalized Health" & "Precision Medicine"

Michael Baudis **UZH SIB**
Computational Oncogenomics

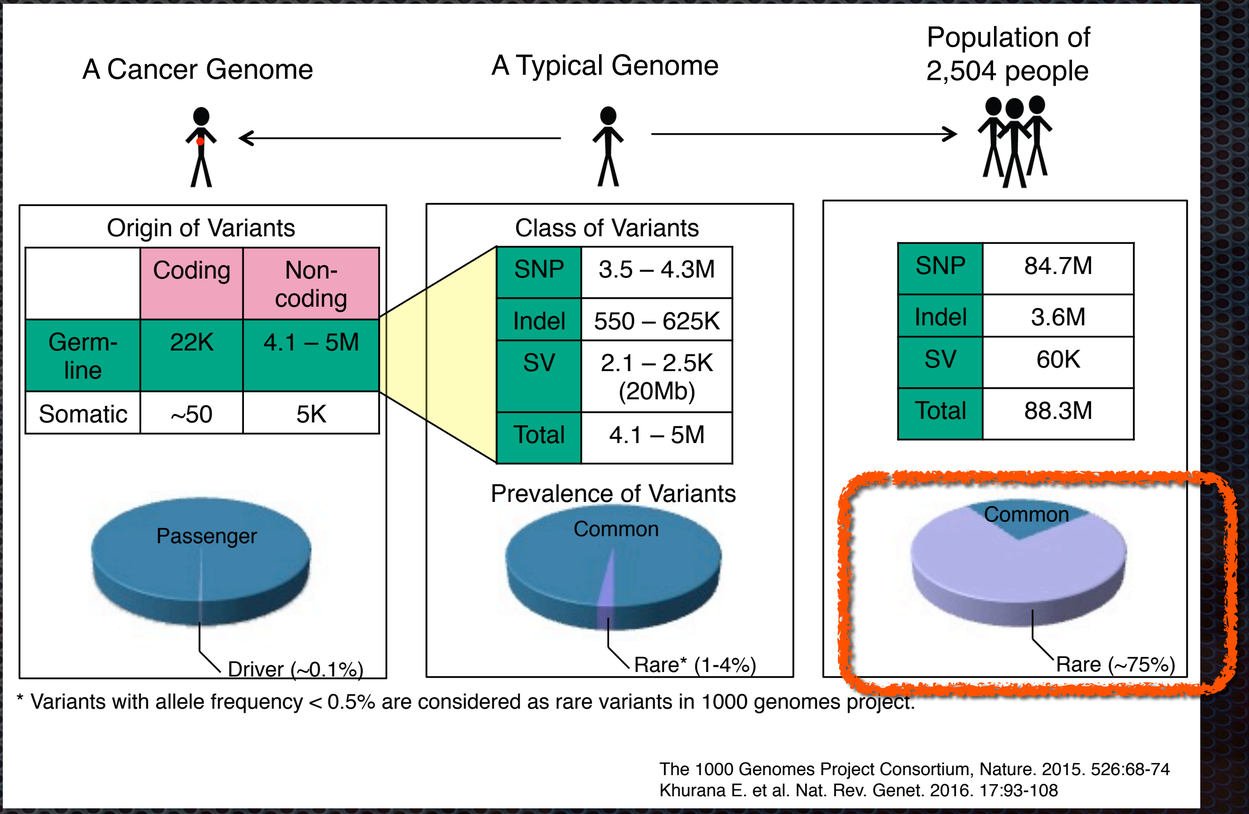
University of Zurich[UZH]

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



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

GA4GH to solve accessibility…

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Search...

ABOUT US | HOW WE WORK | GA4GH TOOLKIT | NEWS & EVENTS | BECOME A MEMBER | CONTACT US

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

GA4GH API promotes sharing

GENOMICS

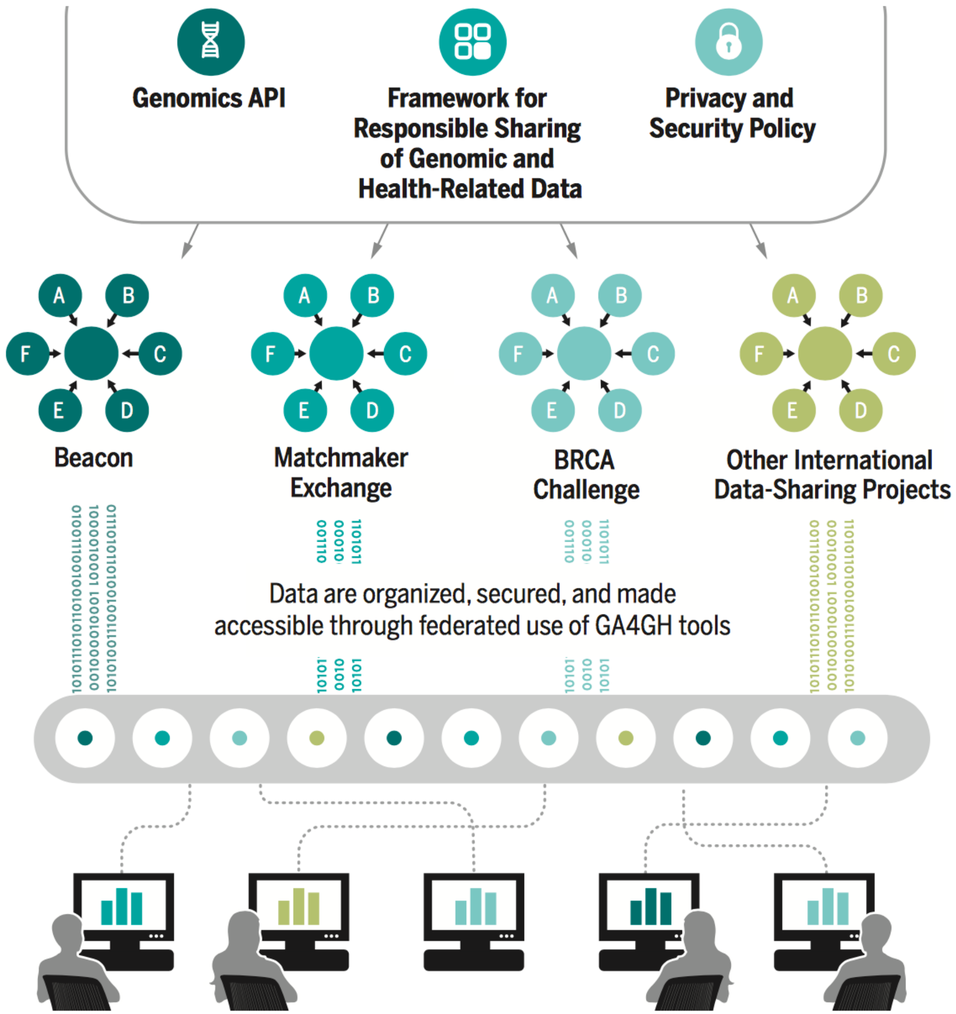
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

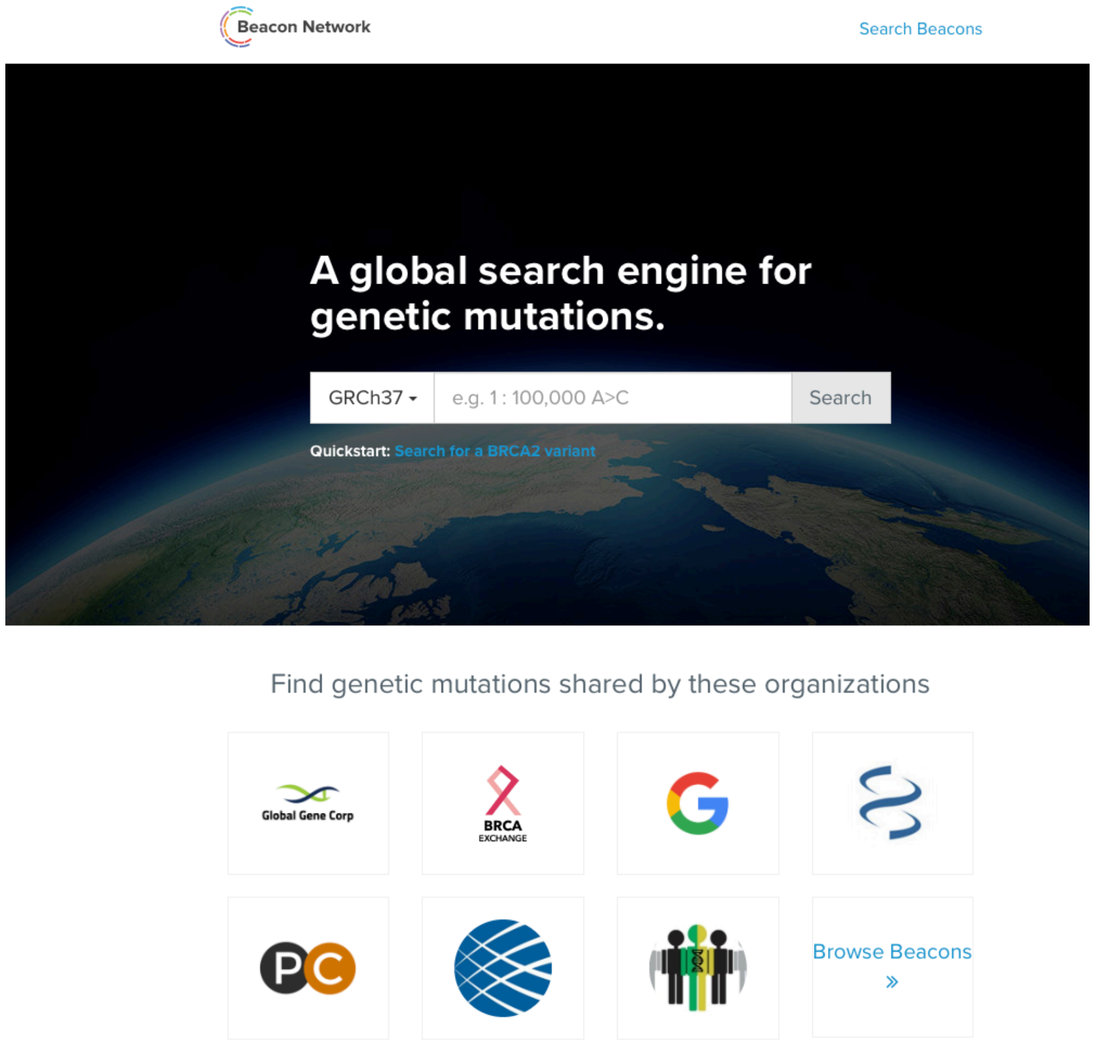
A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



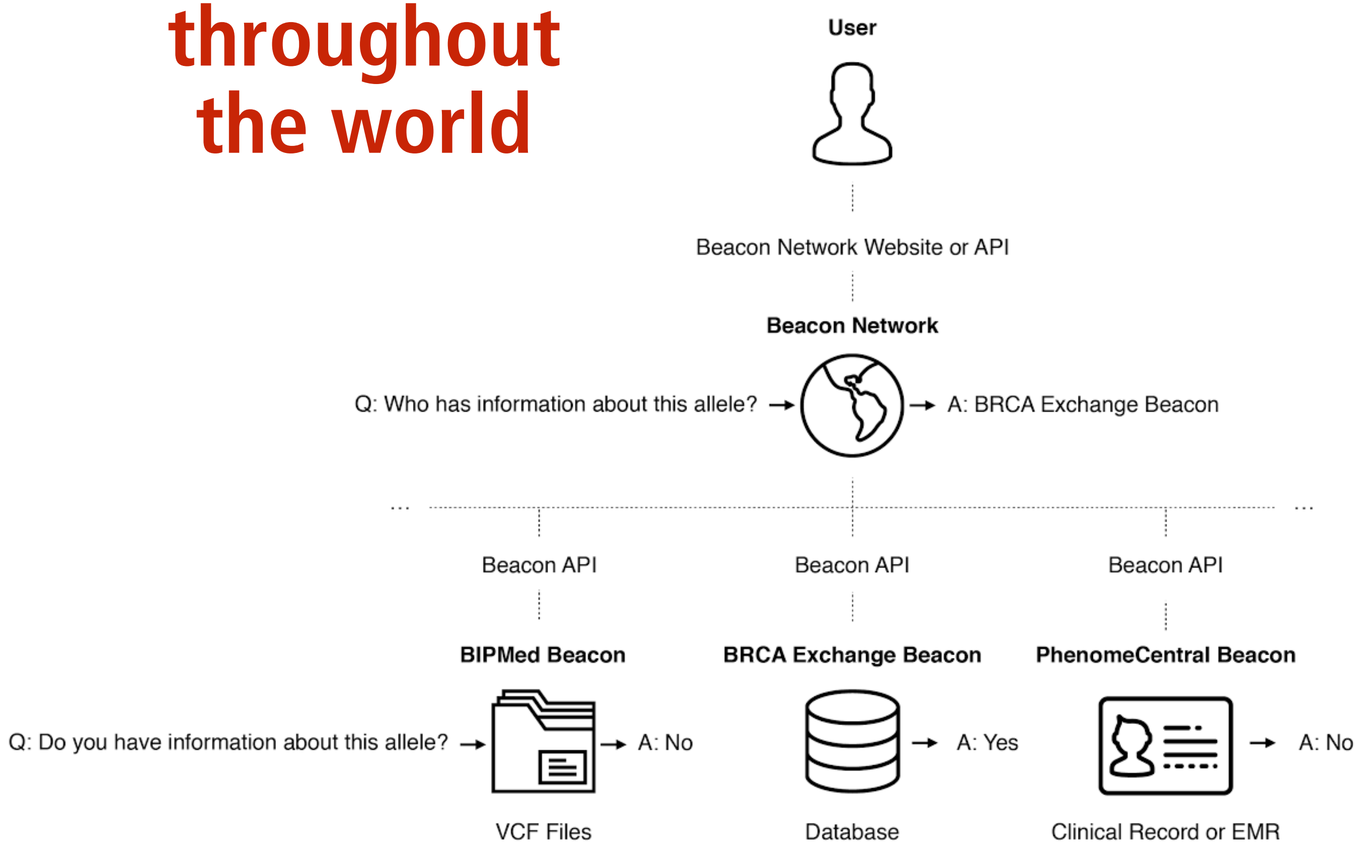
Global Alliance for Genomics & Health

Beacon Project

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

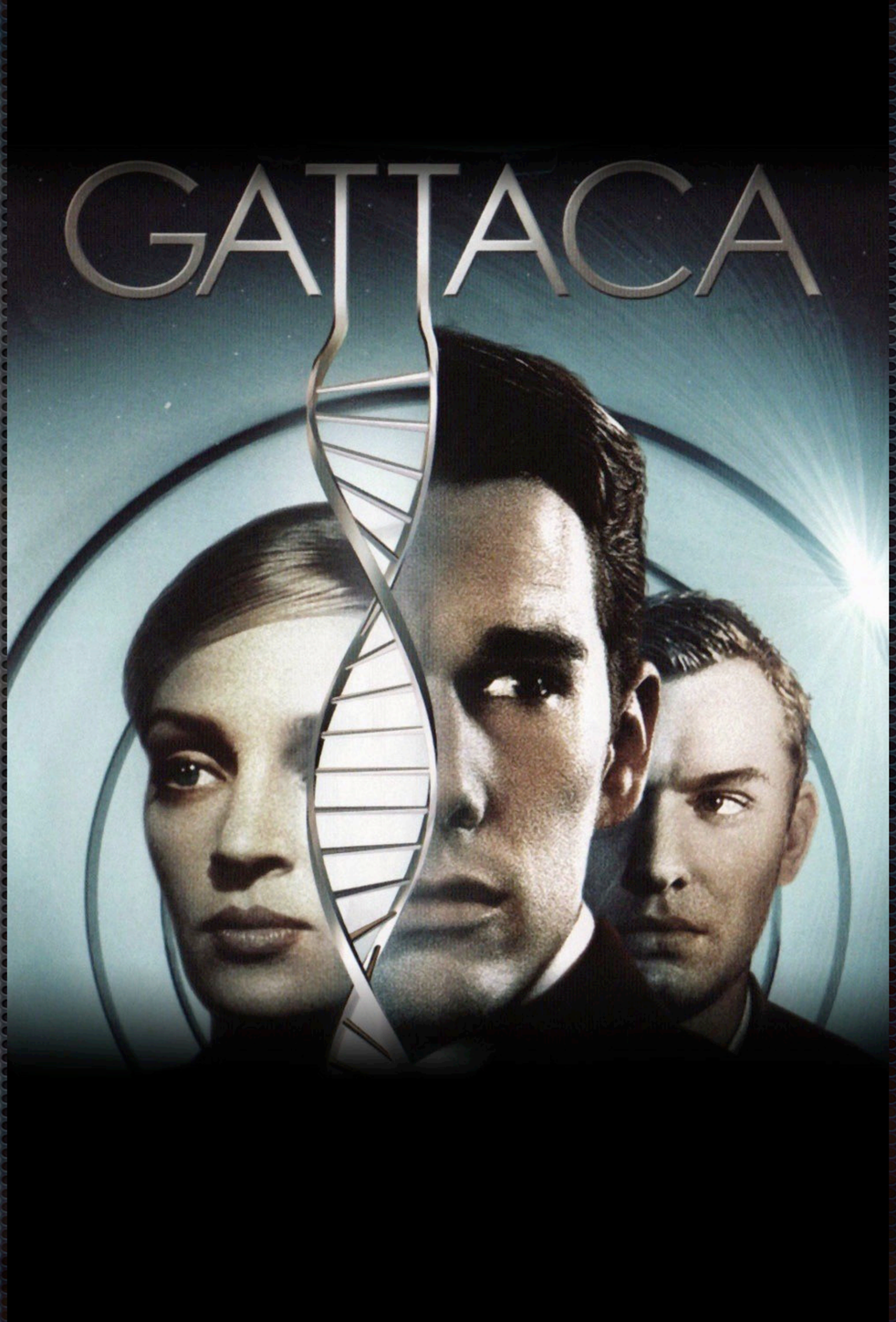


> 50 Beacons throughout the world



Global Alliance for Genomics & Health

Genomes & Privacy



Generalkonsent

BENEFIT

BLOCKCHAIN

HEALTH

PRIVACY

SECURITY

CONSENT

ACCESS

Right to Research

HACKERS

LAWS

Genetic
Information
Nondiscrimination
Act

Health
Insurance
Portability and
Accountability
Act

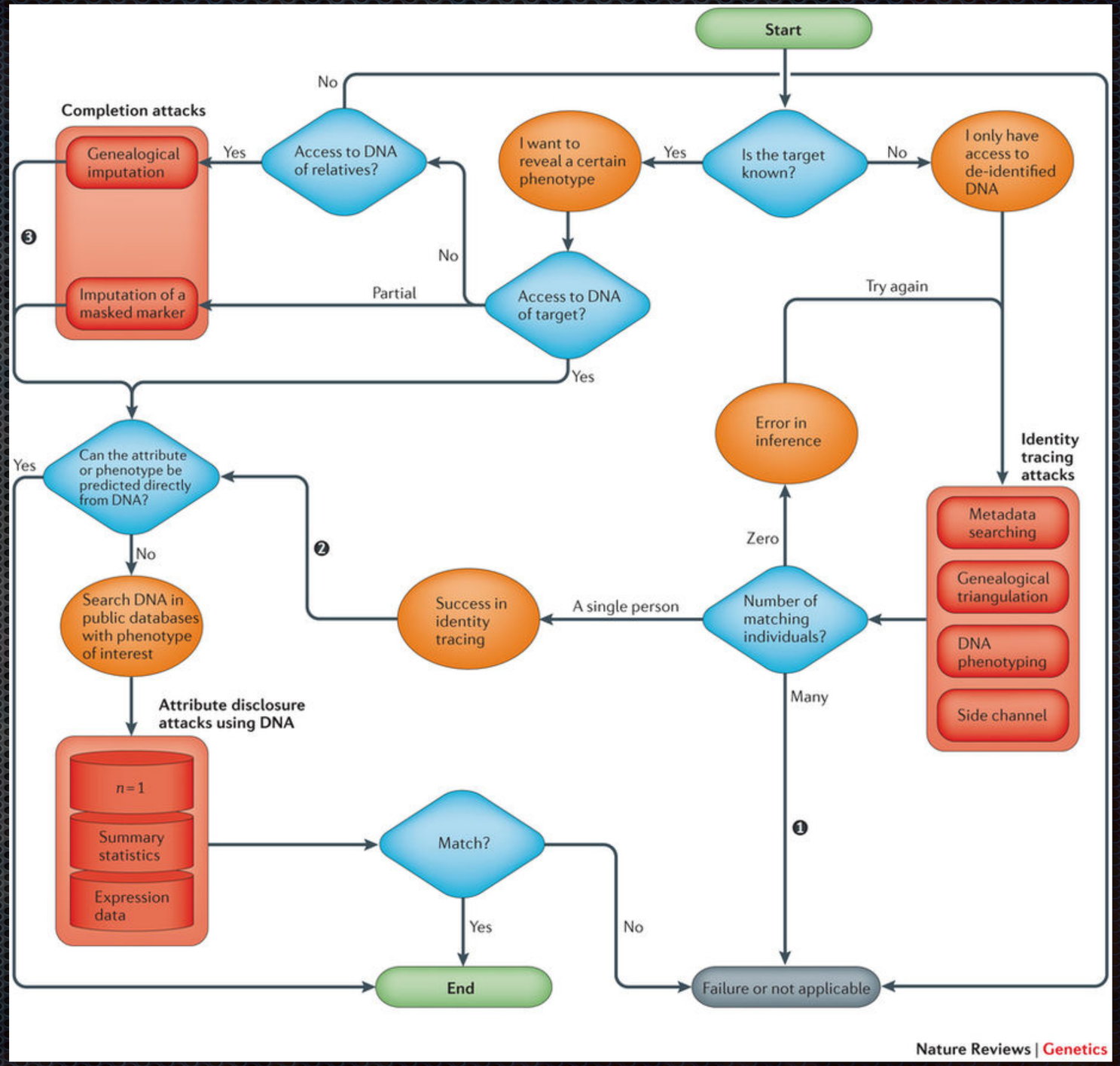
SAFETY

CRYPTOGRAPHY

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)



IDENTIFICATION OF INDIVIDUALS BASED ON "GENOMIC FINGERPRINTS"

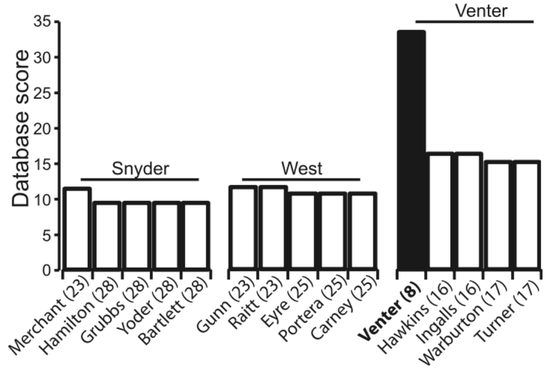
18 JANUARY 2013 VOL 339 SCIENCE www.sciencemag.org

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Fig. 2. The top five records retrieved after searching Ysearch with the Y-STR haplotypes of Michael Snyder, John West, and Craig Venter. The expected number of generations to the MRCA is given in parentheses for each record. Searching with Craig Venter returned a "Venter" record (closed bar) as the top match.



- Genomic data of many types can be used to re-identify individuals in data collections

Genome ***Beacons*** Compromise Security?

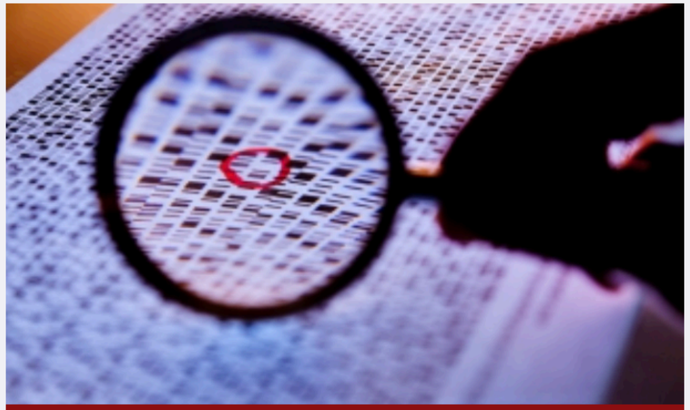
Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals

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.

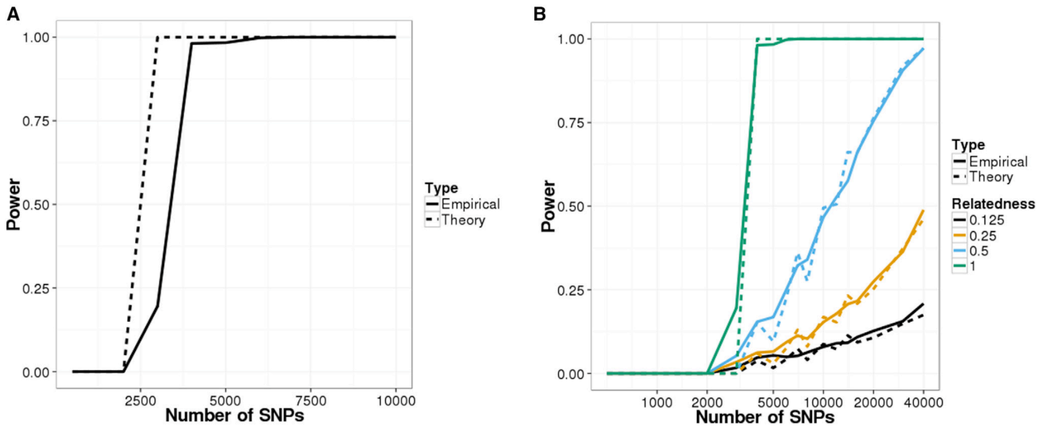


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- 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

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.

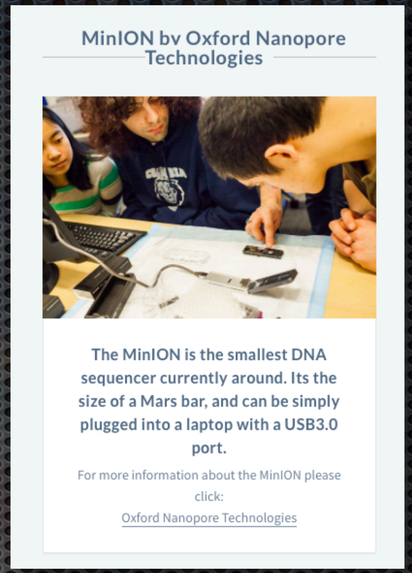
Cell line identification

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.

“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”)

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 scienti c advancement and its bene ts” and “to the protection of the moral and material interests resulting from any scien- ti c...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
© 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.

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,

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

Personenbezug \ Art des Forschungs-materials	Biologisches Material und genetische Daten	Nicht-genetische Daten
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 Forschungs-projekte + Generalkonsent für Forschungs-zwecke	Information über Weiterverwendung für zukünftige noch unbestimmte Forschungs-projekte + über Möglichkeit Weiterverwen-dung 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

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

Consent Codes: Upholding Standard Data Use Conditions

Stephanie O. M. Dyke[1]*, Anthony A. Philippakis[2], Jordi Rambla De Argila[3,4], Dina N. Paltoo[5], Erin S. Luetkemeier[5], Bartha M. Knoppers[1], Anthony J. Brookes[6], J. Dylan Spalding[7], Mark Thompson[8], Marco Roos[8], Kym M. Boycott[9], Michael Brudno[10,11], Matthew Hurles[12], Heidi L. Rehm[2,13], Andreas Matern[14], Marc Fiume[15], Stephen T. Sherry[16]

Global Alliance for Genomics & Health

Consent Codes		
Name	**Abbreviation**	**Description**
Primary Categories (I^{Ty})		
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^{Ty}) (can be one or more extra conditions, in addition to I^{Ty} 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)

Health Related Data & Privacy

- 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

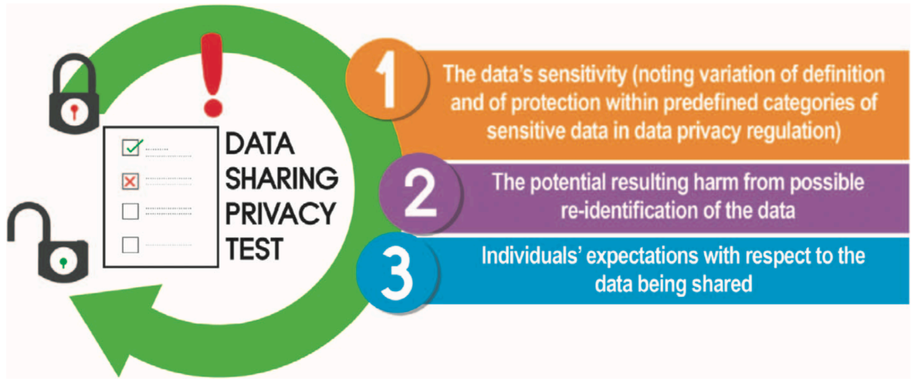
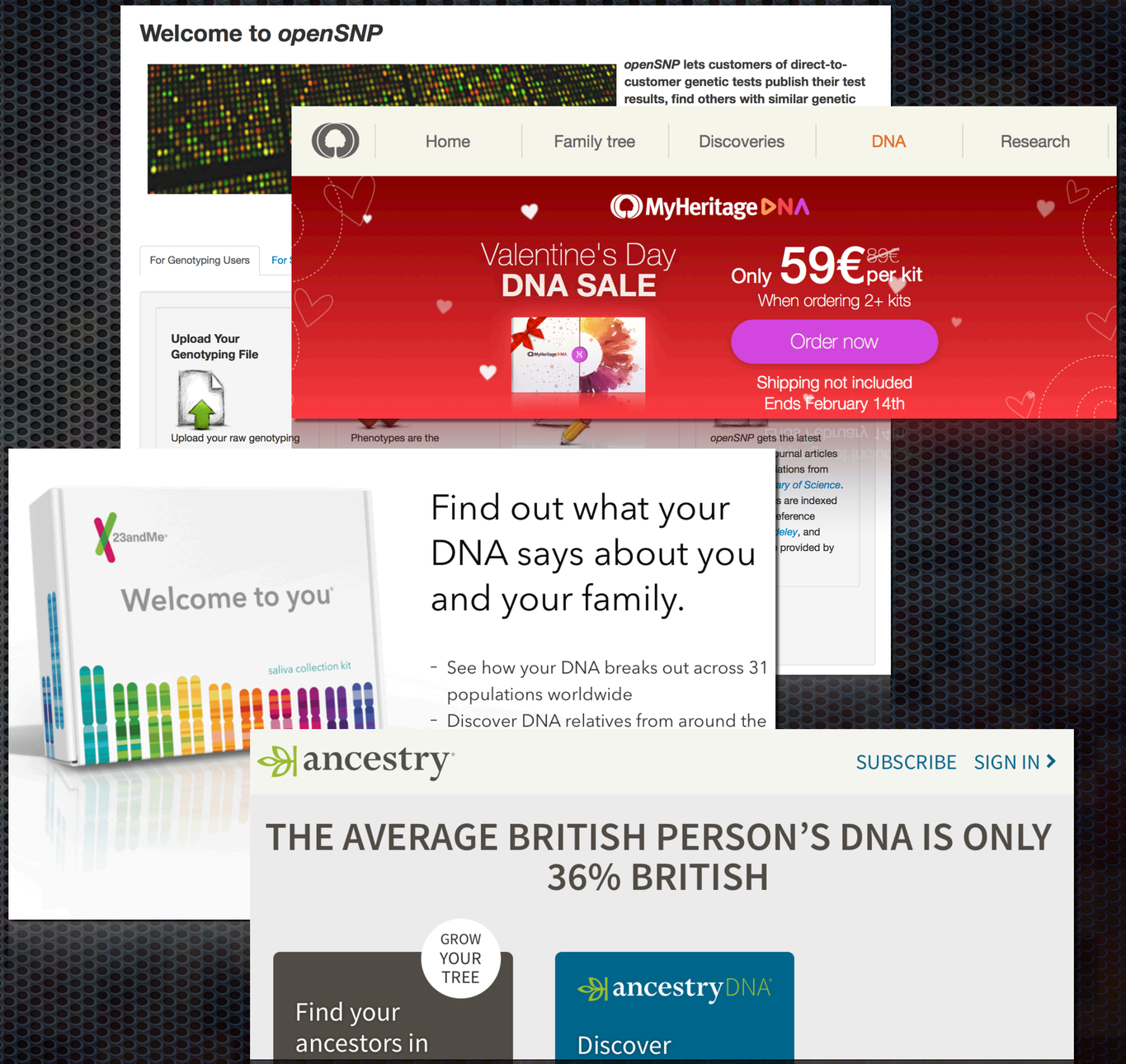


Figure 1. The three steps of a Data Sharing Privacy Test to distinguish degrees of data sensitivity within categories of data recognised as 'sensitive'.

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!



Long-Range Familial Searches

MyHeritage DNA

Hi Michael,

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

- 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**

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

ALL SEC

E-EDITION

SUBSCRIBE

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

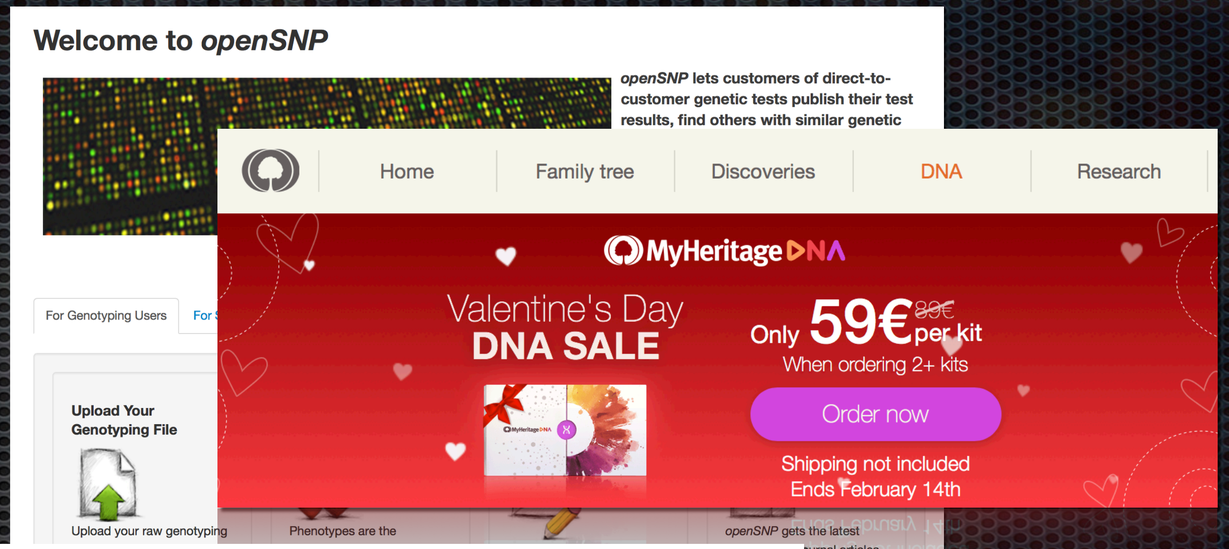
How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect

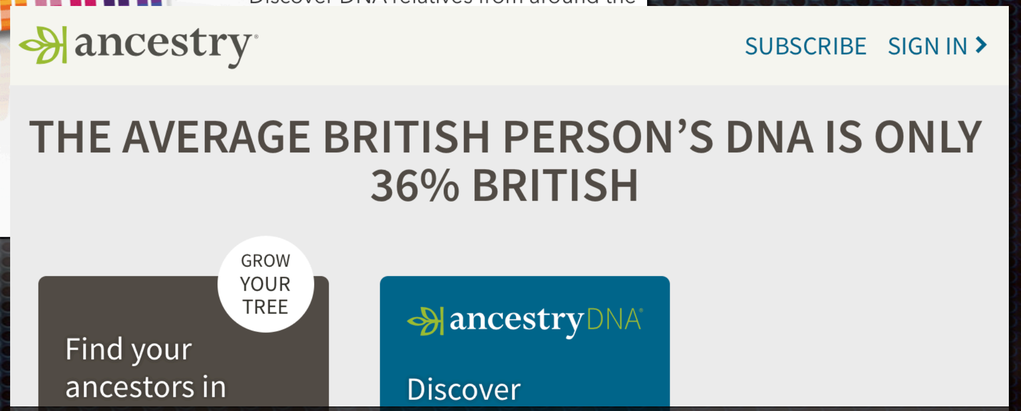
Investigators used DNA from crime scenes that had been stored all these years and plugged the genetic profile of the suspected assailant into an online genealogy database. One such service, GEDmatch, said in a statement on Friday that law enforcement officials had used its database to crack the case. 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

John Yuyi, NYT 2018-02-09





University of Zurich[UZH]

SIB

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

arraymap.org

progenetix.org

info.baudisgroup.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis

Global Alliance
for Genomics & Health

elixir