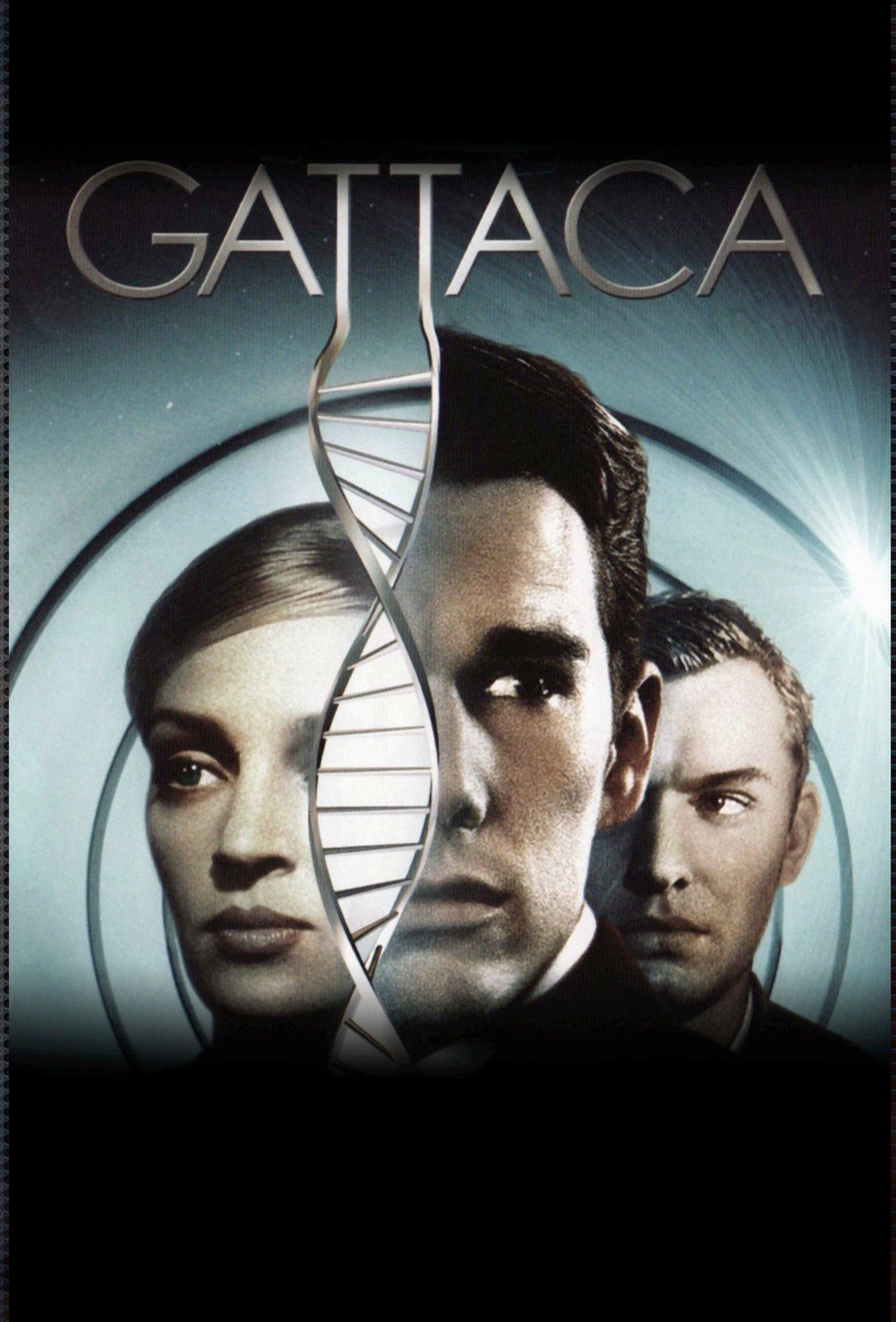
BIO390
Introduction to Bioinformatics

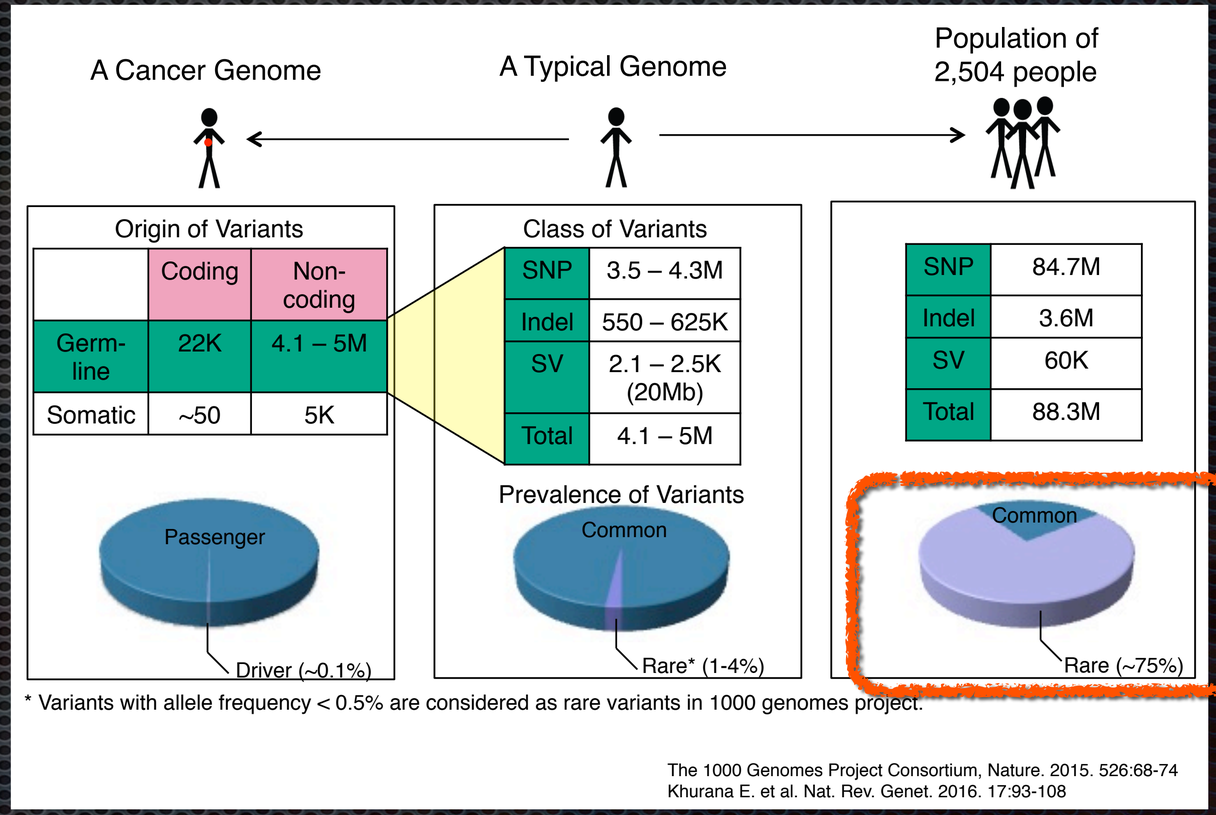
Genome Data & Privacy



Michael Baudis **UZH SIB**
Computational Oncogenomics

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)

GA4GH API promotes sharing

GENOMICS

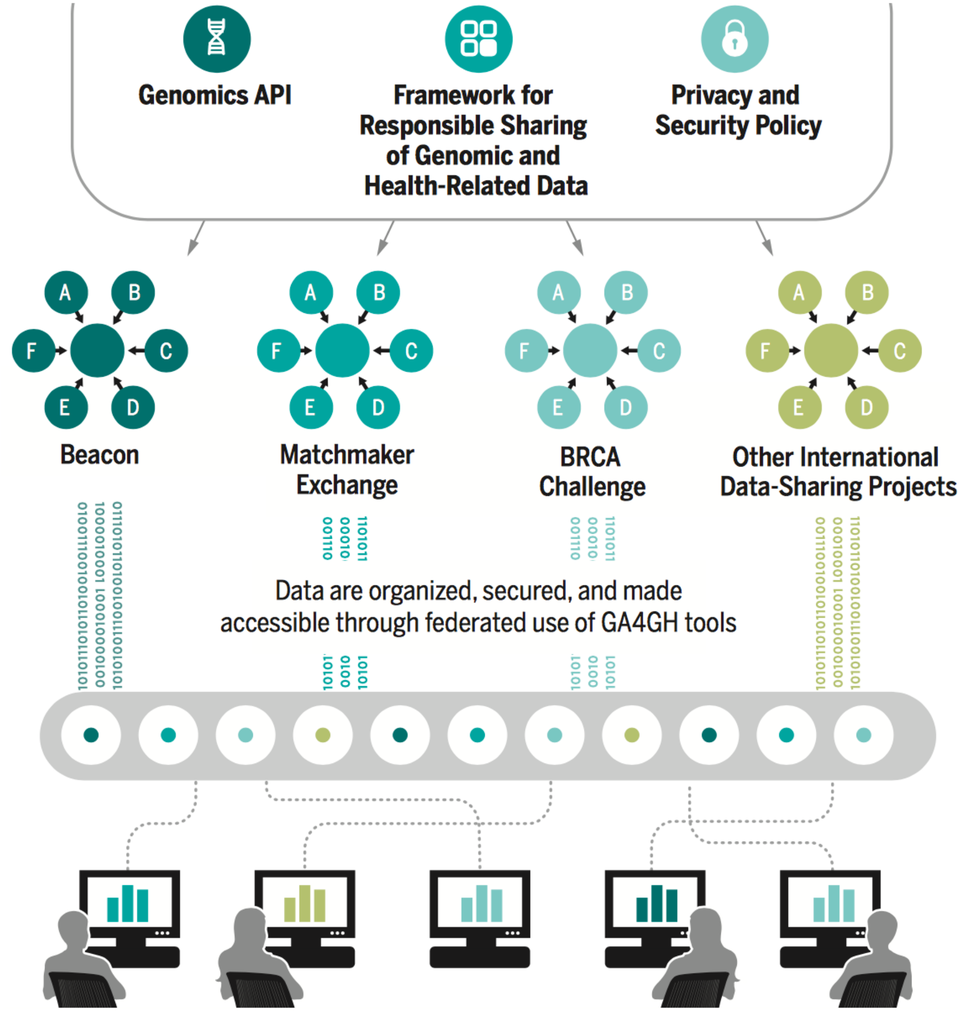
A federated ecosystem for sharing genomic, clinical data

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

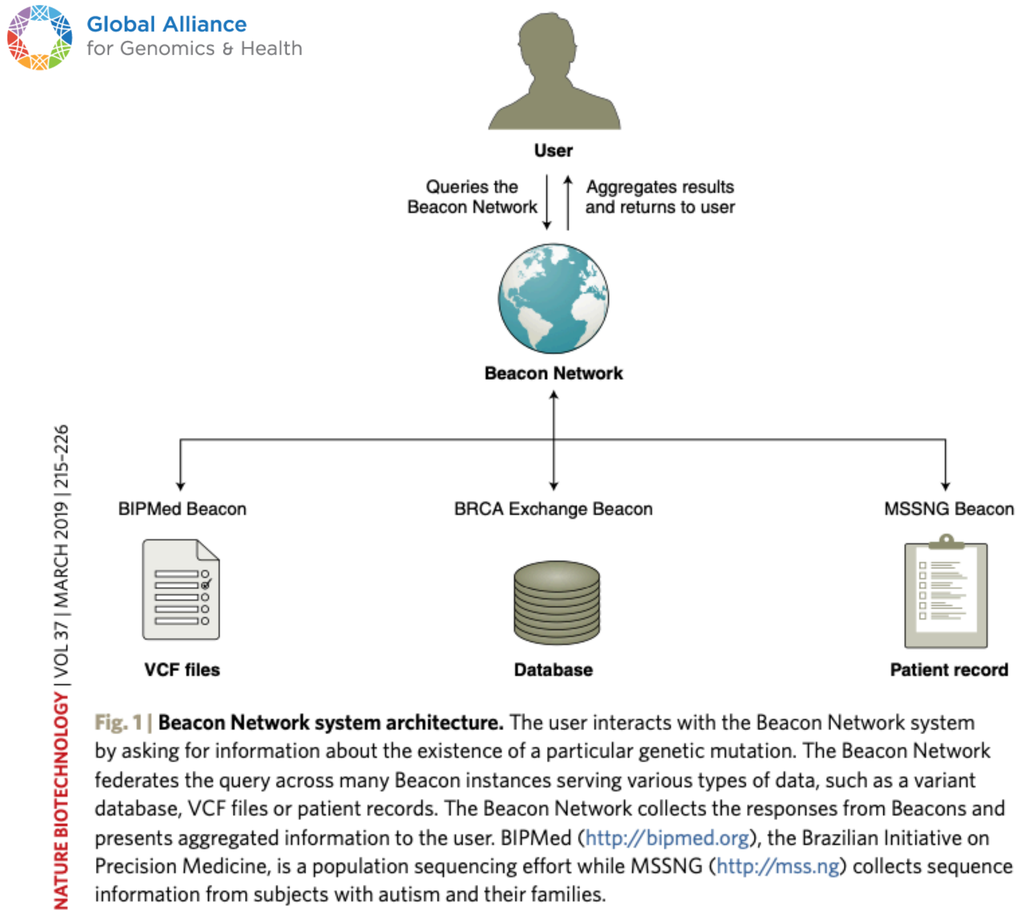
A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



Global Alliance for Genomics & Health

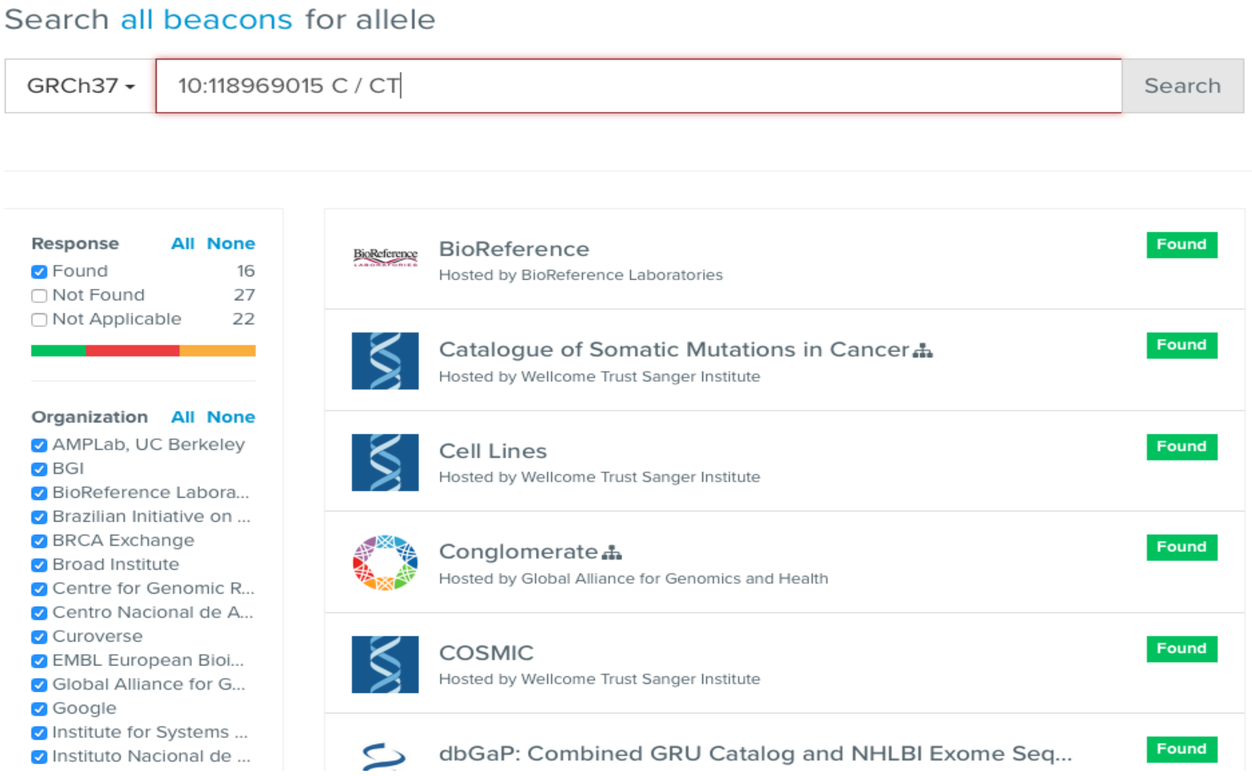
Beacon Project

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



NATURE BIOTECHNOLOGY | VOL 37 | MARCH 2019 | 215–226

Fig. 1 | Beacon Network system architecture. The user interacts with the Beacon Network system by asking for information about the existence of a particular genetic mutation. The Beacon Network federates the query across many Beacon instances serving various types of data, such as a variant database, VCF files or patient records. The Beacon Network collects the responses from Beacons and presents aggregated information to the user. BIPMed (http://bipmed.org), the Brazilian Initiative on Precision Medicine, is a population sequencing effort while MSSNG (http://mss.ng) collects sequence information from subjects with autism and their families.



Genome "Beacons" apply a standardized API, which allows to query the existence of a single genomic variant across many genome resources supporting the protocol. While the practical utility is mostly limited to e.g. excluding candidate variants in rare diseases, the anonymous nature of the responses facilitates the easy uptake and participation of many resources.

Additionally, the widespread participation serves as entry point for extending the usability of the protocol for advanced applications.

Genome ***Beacons*** Compromise Security?

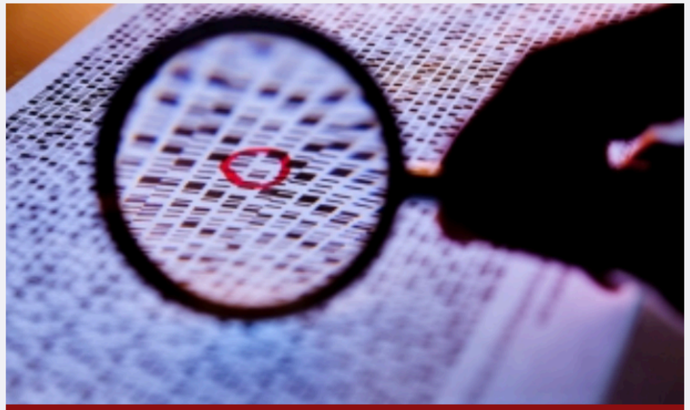
Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals

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.

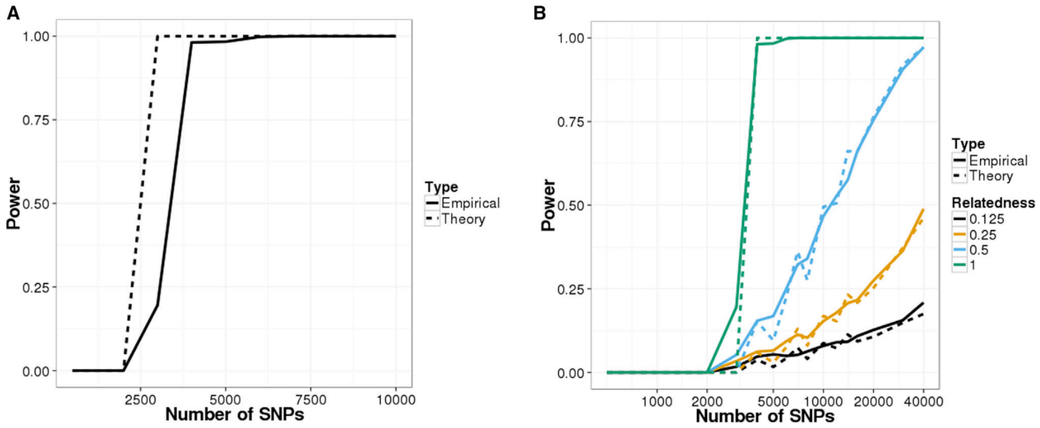
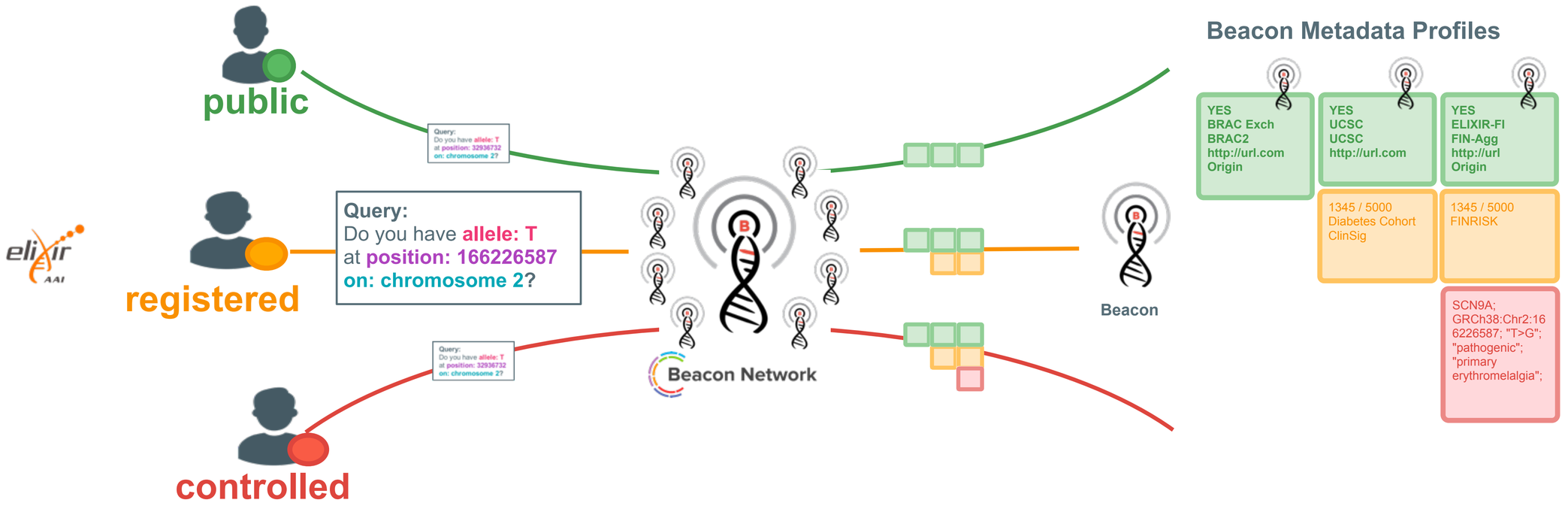


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- 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**

Beyond open Beacons: Integrating permissions and discovery



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

Each Beacon's data type release policy should cover all rhetorical questions

Factual circumstance	Example data	 Public	 Registered	 Controlled
Who?	Specimen/sample ID	No	No	Yes
What?	HGVS allele	Yes, some loci restricted	Yes	Yes
When?	Date of report	No	Retrospective, no Prospective, yes for repetitive queries	Yes
Where?	Variant frequencies in sufficiently described population samples.	No	Yes	Yes
Why?	Not well captured. Best example might be: germline | somatic | tumor	Yes if beacon is restricted to specific mutational class	Yes when beacon supports reporting by mutational class	Yes
In what way?	cis|trans, single | compound haplotype, phenotypes	Yes*	Yes Limited phenotype	Yes
By what means?	Experimental details | instrument data	No	Aggregate	Yes

Source: Stephen Sherry, NCBI (2016)

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

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

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

Long-Range Familial Searches - Rapid Uptake, No Escape?

Table 1. Public cases of long-range familial cases. A "–" indicates data not available.

Case	Announcement	Solved by	Closest match	Comments
Buckskin Girl	9 April 2018	DNA Doe Project	First cousin once removed	
Golden State Killer	24 April 2018	Barbara Rae-Venter	Third cousin	
Lyle Stevik	8 May 2018	DNA Doe Project	Second cousin	Inbreeding complicated the estimation of the match.
William Earl Talbott II	21 May 2018	Parabon	Half–first cousin once removed	Second cousins were identified as well.
Joseph Newton Chandler III	21 June 2018	DNA Doe Project	Second cousin once removed	
Gary Hartman	22 June 2018	Parabon	Half–first cousin	Genealogists were able to overcome a nonpaternity event in the family tree of the suspect.
Raymond "DJ Freez" Rowe	25 June 2018	Parabon	–	
James Otto Earhart	26 June 2018	Parabon	Second cousin	
John D. Miller	15 July 2018	Parabon	–	
Matthew Dusseault and Tyler Grenon	28 July 2018	Parabon	–	
Spencer Glen Monnett	29 July 2018	Parabon	–	This was an active case for a crime that occurred in April 2018.
Darold Wayne Bowden	23 August 2018	Parabon	–	
Michael F. Henslick	29 August 2018	Parabon	–	

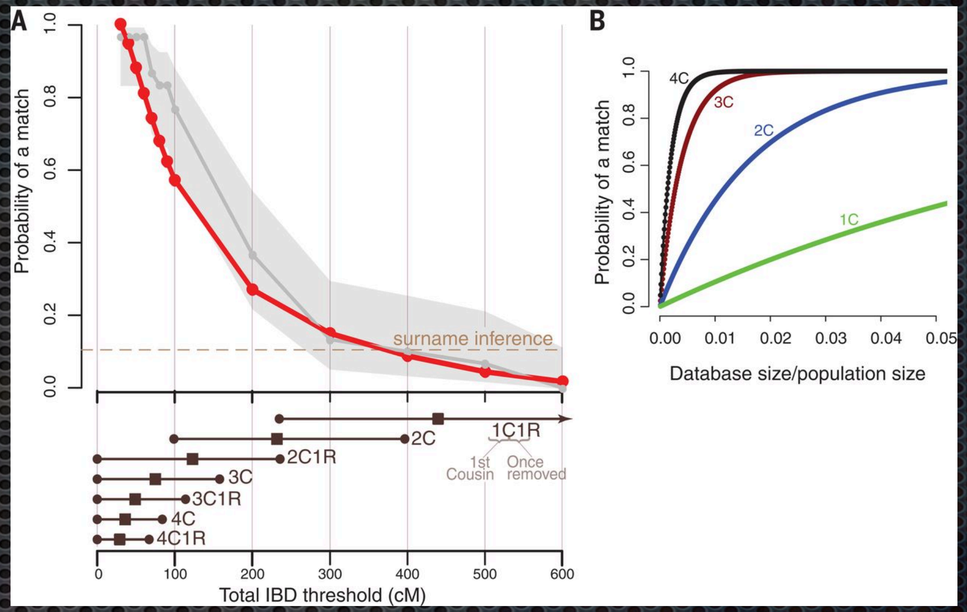


Fig. 1. The performance of long-range familial searches for various database sizes. (**A**) The probability of finding at least one relative for various IBD thresholds (top) with 1.28 million searches of DTC-tested individuals (red) and 30 random GEDmatch searches (gray). Light gray shading indicates the 95% CI for the GEDmatch estimates. The dashed line indicates the probability of a surname inference from Y chromosome data (*17*). The bottom panel shows the 95% CIs (circles) and average total IBD length (squares) for a first cousin once removed (1C1R) to a fourth cousin once removed (4C1R) (*20*). (**B**) A population-genetic theoretical model for the probability of finding relatives up to a certain type of cousinship as a function of the database coverage of the population. 1C to 4C indicate first to fourth cousins.

With the rapid spreading of direct-to-consumer genetic testing, basically everybody will soon have relatives with genome profiles in one of the large genealogy or other genotyping resources. While such by commercial providers restrict access & use of the data, many individuals add their genotyping data to publicly accessible resources.

Erlich et al., Science 362, 690–694 (2018) 9 November 2018

But for re-identification one first needs the DNA of the person to be identified...

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

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:

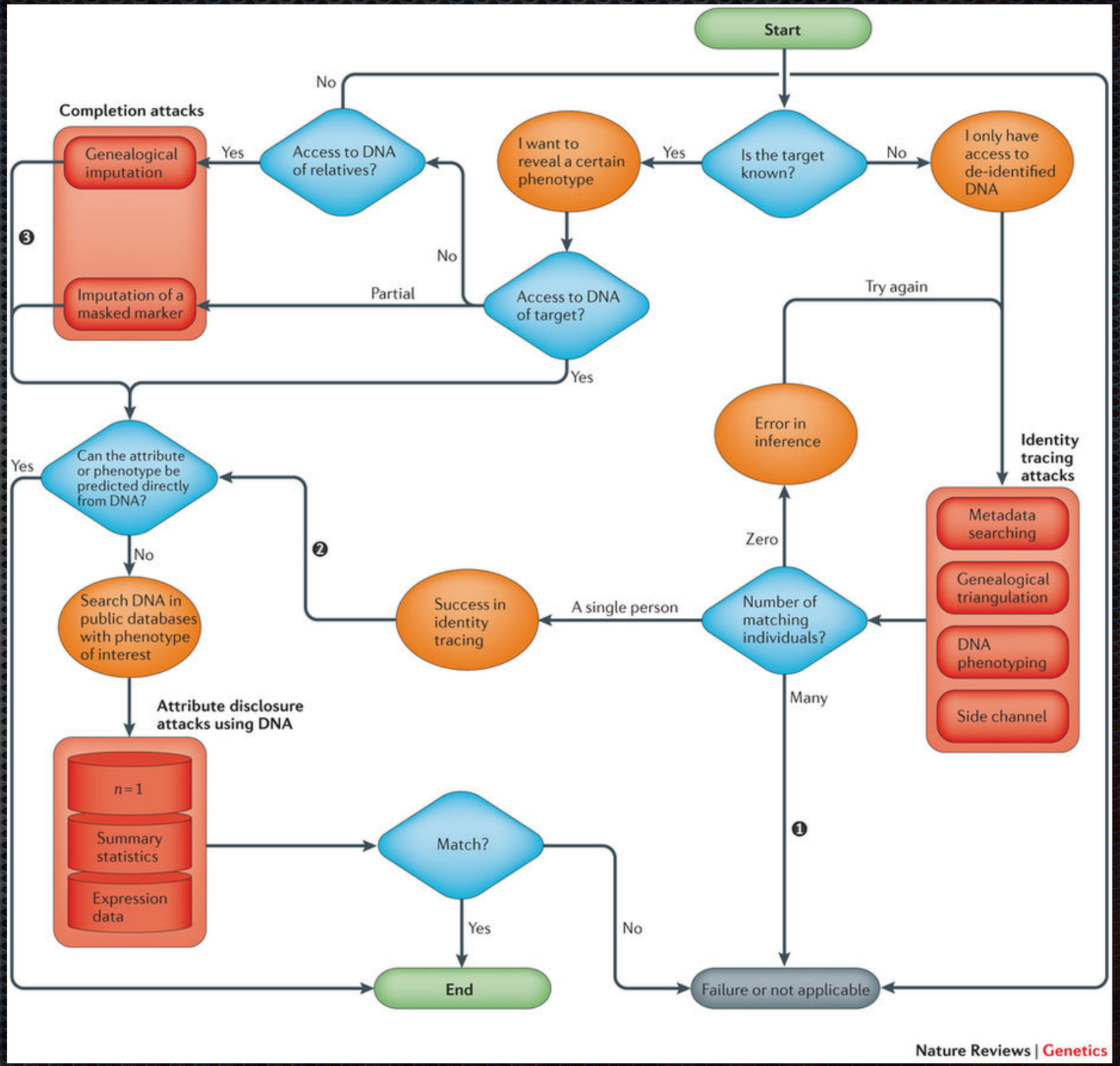
Bento Lab

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

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)



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 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

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

Art des Forschungs-materials / Personenbezug	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^{ry})		
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^{ry}) (can be one or more extra conditions, in addition to I^{ry} 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

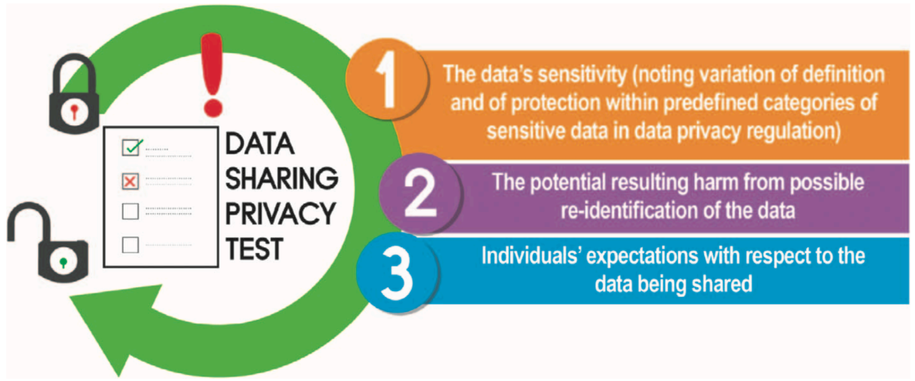
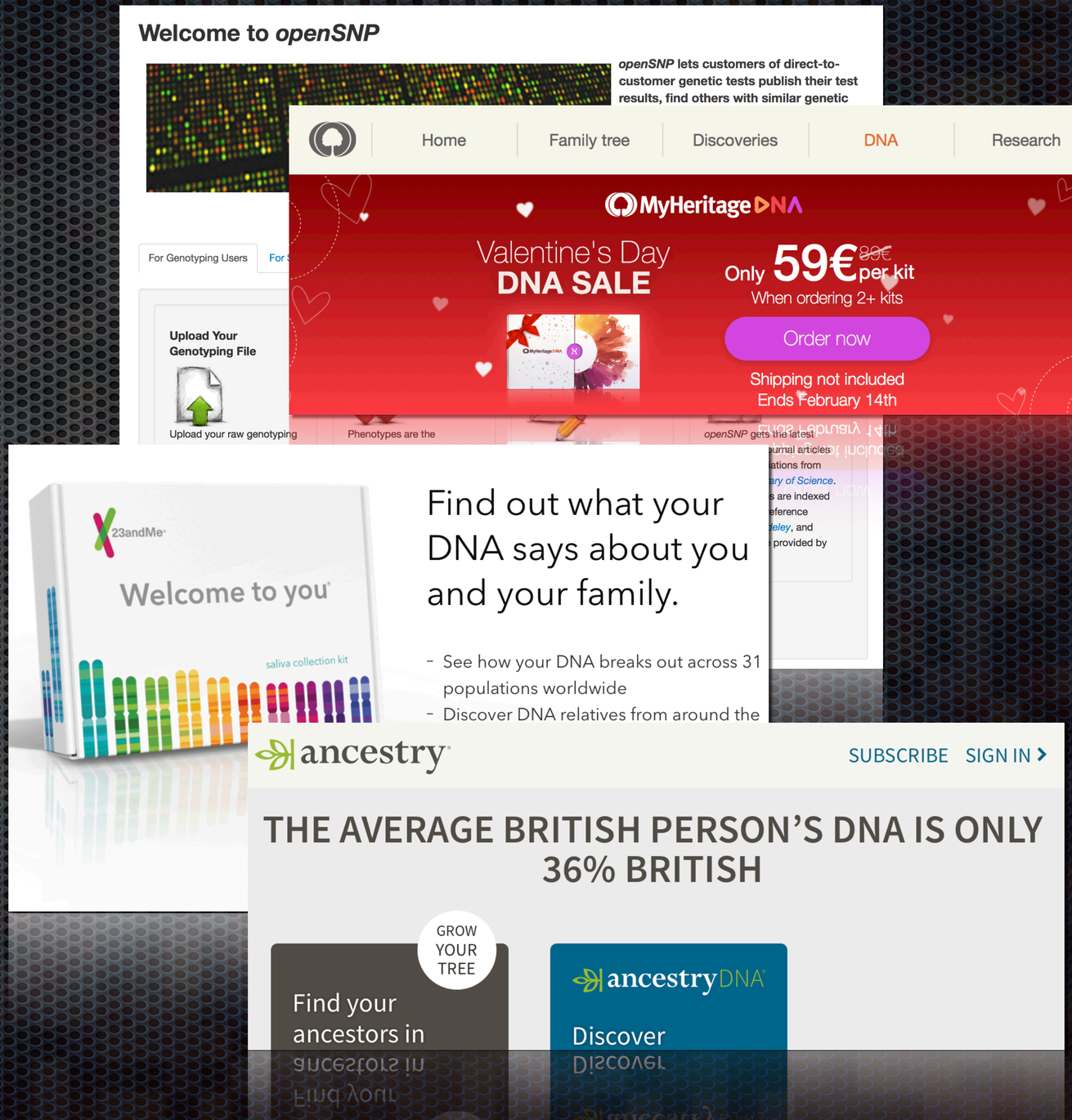


Figure 1. The three steps of a Data Sharing Privacy Test to distinguish degrees of data sensitivity within categories of data recognised as 'sensitive'.

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!



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

BIO390: Course Schedule

- 2019-09-17: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2019-09-24: Christian von Mering - Sequence Bioinformatics
- 2019-10-01: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2019-10-08: Mark Robinson - Statistical Bioinformatics
- 2019-10-15: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2019-10-22: Abdullah Kahraman (USZ) - Molecular Interaction Networks
- 2019-10-29: Katja Baerenfaller (SIAF) - Proteomics
- 2019-11-05: Amedeo Caflisch - Molecular Dynamics
- 2019-11-12: Elif Ozkirimli - Protein Structure and Interactions
- 2019-11-19: Christophe Dessimoz (UniL) - Sequence evolution and phylogenetics
- 2019-11-26: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2019-12-03: Andreas Wagner - Biological Networks
- 2019-12-10: Alex Handler Wagner (WUSTL) & Michael Baudis - Human Genome Variation Resources
- **2019-12-17: Exam (Multiple Choice)**

BIO390 Exam

- 8:10 - 9:40
- UZH ID
- Multiple choice questions (single or multiple answers; indicated)

12 **Tidy data is a way of arranging data where each variable is in one column and each observation is in one row. The table below shows some data which has been imported into R as a tibble data.frame. Is this data in tidy format?**

```
## # A tibble: 10 x 5
##     control treatment  time    sex individual
##       <dbl>     <dbl> <chr> <fctr>     <fctr>
##  1 3.711482  4.476728   day female          1
##  2 4.652311  3.680602   day   male          2
##  3 4.478371  4.832210 night     NA       <NA>
##  4 6.273473  5.440702 night   male          2
##  5 6.824521  6.380166 night female          1
##  6 3.488692  5.768513   day female          3
##  7 5.110508  5.480421   day female          4
##  8 4.239204  4.975053   day   male          5
##  9 4.330103  4.090669 night   male          5
## 10 5.274520  3.666011 night   male          6
```

A	No, because there are missing values.	☐
B	No, because the columns “control” and “treatment” are observations of the same variable and for tidy format should be in the same column.	☐
C	No, because the column “time” has type “character” but should be categorial (i.e. with type “factor”).	☐
D	Yes, because “control” and “treatment” should be in separate columns as they are contain the values that will be compared.	☐

HL 1 option

22 **Mark ALL true statements: Data from which of the following method(s) would you use to infer *direct* protein-protein interactions?**

A	AP-MS	☐
B	BioID-MS	☐
C	APEX	☐
D	*in vitro* reconstitution	☐
E	size exclusion chromatography	☐
F	Yeast two hybrid screen	☐

MG multiple options

University of Zurich[UZH]

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB I 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