Privacy in Healthcare and Genomic Privacy

Asst. Prof. Sinem Sav

With gratitude to the biomedical and CS researchers I have the privilege to work with and special thanks to Jean-Pierre Hubaux, David Froelicher, Christian Mouchet, Erman Ayday and Juan Troncoso-Pastoriza who contributed to some of the slides

Attacks on and breaches in medical databases

Anthem Hacking

Anthem Hacking Points to Security Vulnerability of Health Care Industry

By REED ABELSON and MATTHEW GOLDSTEIN FEB. 5, 2015

The New York Times

An Anthem Health Insurance facility in Indianapolis. Hackers gained access to about 80 million company records, and some fear the stolen data will be used for identity theft. Aaron P. Bernstein/Getty Images

- Anthem: one of US largest health insurers
- 60 to 80 million *unencrypted* records stolen in the hack (revealed in February 2015)
- Contain social security numbers, birthdays, addresses, email and employment information and income data for customers and employees, including its own chief executive

US Healthcare “Wall of Shame”

Around 2 declared breaches per week, each affecting 500+ people

https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

Breaches Affecting 500 or More Individuals

As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals. These breaches are now posted in a new, more accessible format that allows users to search and sort the posted breaches. Additionally, this new format includes brief summaries of the breach cases that OCR has investigated and closed, as well as the names of private practice providers who have reported breaches of unsecured protected health information to the Secretary. The following breaches have been reported to the Secretary:

Show Advanced Options

Breach Report Results

Name of Covered Entity	State	Covered Entity Type	Individuals Affected	Breach Submission Date	Type of Breach	Location of Breached Information
Alliance Health Networks, LLC	UT	Healthcare Provider	42372	02/15/2016	Hacking/IT Incident	Network Server
Radiology Regional Center, PA	FL	Healthcare Provider	483063	02/12/2016	Loss	Paper/Films
DataStat, Inc.	MI	Business Associate	552	02/12/2016	Unauthorized Access/Disclosure	Paper/Films
Washington State Health Care Authority (HCA)	WA	Health Plan	91187	02/09/2016	Unauthorized Access/Disclosure	Email
SEIM JOHNSON, LLP	NE	Business Associate	30972	02/08/2016	Theft	Laptop

18 April 2023

Turkey: KVKK announces Beytıp Sağlık Hizmetleri data breach

Investigations | Health and Pharmaceutical | Breach Notification | Incident and Breach

The Personal Data Protection Authority ('KVKK') announced, on 13 April 2023, a data breach that occurred within Beytıp Sağlık Hizmetleri Ltd. Şti., a medical center. In particular, the KVKK highlighted that Beytıp Sağlık Hizmetleri notified the same of a data breach in accordance with Article 12(5) of the Law on Protection of Personal Data No. 6698. Moreover, the KVKK noted that, on 18 February 2023, it was discovered that some computers accessing the program containing medical records of patients and their relatives within Beytıp Sağlık Hizmetleri were unable to be accessed, that unidentified individuals gained unauthorised access to the network and information system connected to those computers, and that the access to the program in question was encrypted.

Furthermore, the KVKK clarified that the personal data affected by the breach includes identity, communication, employment, legal transactions, customer transactions, operational security, risk management, financial, marketing, visual and audio recordings, race and ethnic origin, and health information. In addition, the KVKK provided that, although the number of indi-

Medical Data Breach in Switzerland

Le Matin

MONDE SUISSE SPORTS FAITS DIVERS PEOPLE LOISIRS SOCIÉTÉ HIGH-TE

Images

Bilans de santé en balade sur le net

GAFFE — Des données médicales ultraconfidentielles de patients romands ont été librement accessibles durant des jours sur Internet. Le groupe Synlab déplore une erreur humaine.

Par Raphaël Pomey . Mis à jour le 08.04.2015
5 Commentaires

Tests du papillomavirus, dépistages du sida ou d'une hépatite ont circulé pendant des jours sur le Web.
Image: Jason Butcher / Corbis / Montage

- 8300 files of medical analyses freely available online, for days
- With full patient identifications
- Including HIV and other tests
- Sheer carelessness (no attack)

(April 2015)

Fitness-Tracking by Health Insurers

NEWS > BUSINESS > HEALTH CARE

Companies making fitness-tracking deals with workers for cheaper insurance

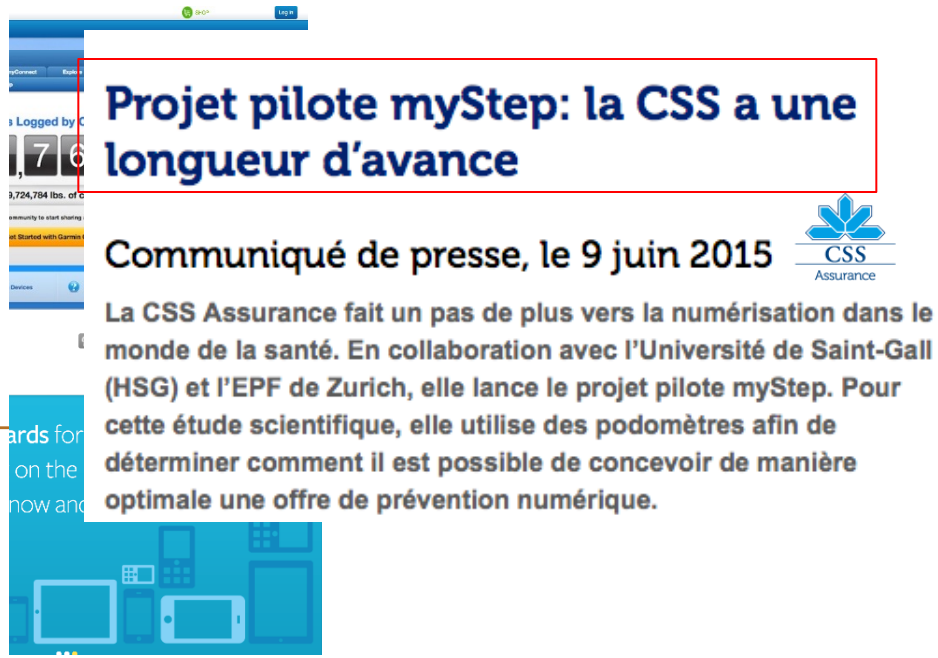
BLOOMBERG

08/21/2014 1:24 PM | Updated: 08/21/2014 1:24 PM

Story Comments

To fight rising medical costs, oil company BP last year offered Cory Slagle – a 260-pound former football lineman – an unusual way to trim $1,200 from his annual insurance bill.

One option was to wear a fitness-tracking bracelet from Fitbit Inc. to earn points toward cheaper health insurance.



Projet pilote myStep: la CSS a une longueur d'avance

Communiqué de presse, le 9 juin 2015

CSS Assurance

La CSS Assurance fait un pas de plus vers la numérisation dans le monde de la santé. En collaboration avec l'Université de Saint-Gall (HSG) et l'EPF de Zurich, elle lance le projet pilote myStep. Pour cette étude scientifique, elle utilise des podomètres afin de déterminer comment il est possible de concevoir de manière optimale une offre de prévention numérique.

Woman finds boyfriend cheating on her after Fitbit watch sends calorie spike alert, her TikTok goes viral

UK's Nadia Essex got to know about her cheating boyfriend through her Fitbit smartwatch.

Listen to Story

Share

News

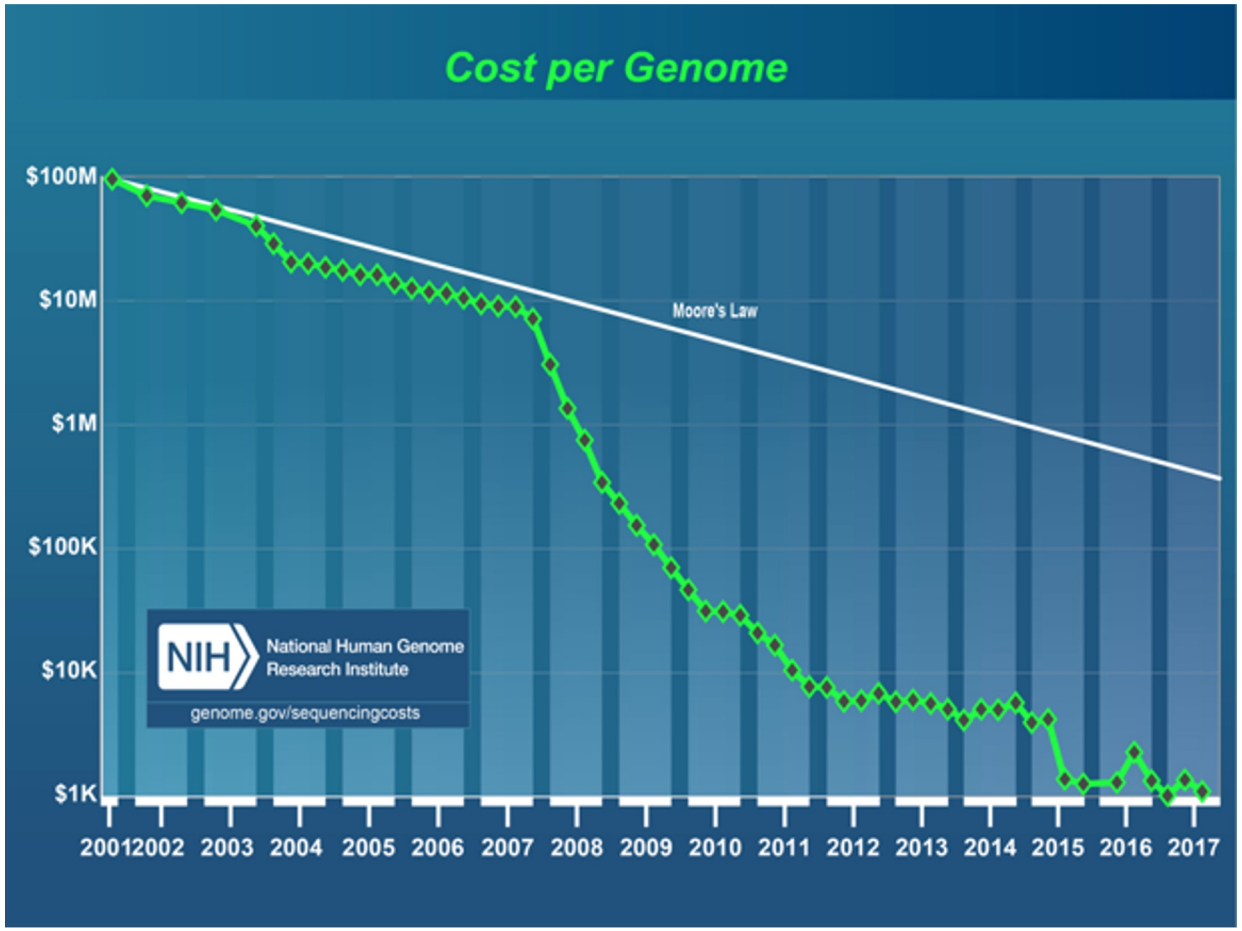
Husband discovers wife has been cheating on her smartwatch

"I wanted to end this marriage of 13 years right there and then."

Introduction to genomics

The Genomic Avalanche Is Coming…

Personalized Health



The massive digitalization of clinical and genomic information is providing unprecedented opportunities for improvements in diagnosis, preventive medicine and targeted therapies

Medical Use of Genetics

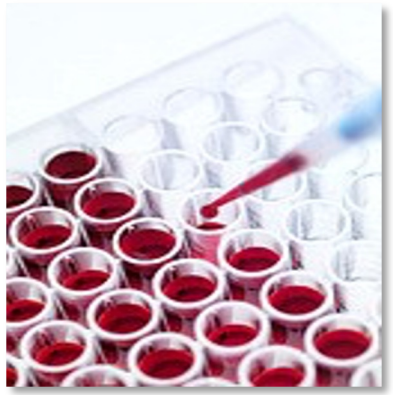
- Genetic disease risk tests help early diagnosis of serious diseases
- Pharmacogenomics - personalized medicine


IEEE
SPECTRUM
FOR THE TECHNOLOGY INSIDER | 0615
A CODER'S GIFT TO HIS SON
A brilliant hack by a diabetes dad
P. 21
WEARABLES WORN BY PRO ATHLETES
The tech that gives top stars an edge
P. 44
THE DOC IN YOUR POCKET
Inside the Tricorder XPrize competition
P. 48
WHY THE EBOLA MODELS FAILED
Big gaps in info couldn't be filled
P. 62
HACKING THE HUMAN OS
How Big Data Will Transform Medicine and Health
SPECIAL REPORT
IEEE

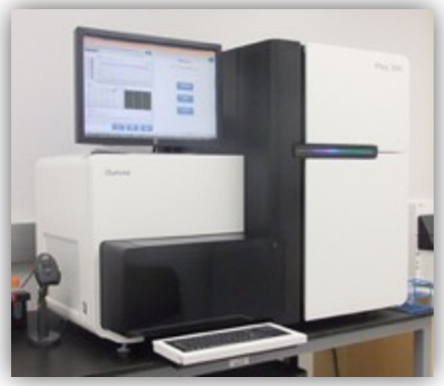

MIT Technology Review
VOL. 118 NO. 3 MAY/JUNE 2015 $6.99
Feature p. 48
HP Tries to Reinvent the Computer
Business Report p. 63
Persuasion
Review p. 72
The Problem with Fake Meat
WE CAN NOW ENGINEER THE HUMAN RACE
p26

- 1990 – 2003
- Goals
  - sequencing and identifying all three billion chemical units in the human genetic instruction set,
  - Finding the genetic roots of disease
  - Developing treatments
- Mostly US/UK effort

From Blood Sample to Genome Analysis



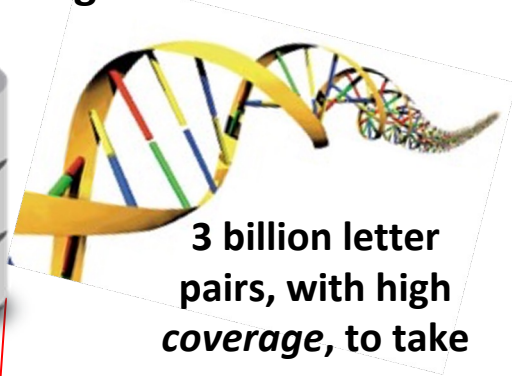
Sample
s



Sequencing machine

Raw data
(short reads)
FastQ files

Alignment



3 billion letter pairs, with high *coverage*, to take into account:

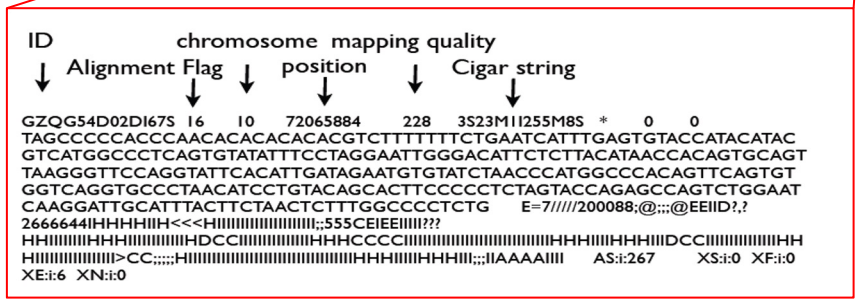
- Sequencing errors
- Possible mutations

SAM/BAM file
(aligned reads)

Variant call

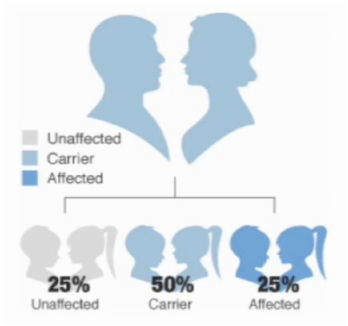
VCF Files

Delta with respect to the *reference genome*



Significance and Popularity of Genomic Data

Counsyl



23andMe

Name	Confidence	Your Risk	Avg. Risk
Atrial Fibrillation	★★★★	33.9%	27.2%
Prostate Cancer ♂	★★★★	29.3%	17.8%
Alzheimer's Disease	★★★★	14.2%	7.2%
Age-related Macular Degeneration	★★★★	11.1%	6.5%
Colorectal Cancer	★★★★	7.8%	5.6%
Chronic Kidney Disease	★★★★	4.2%	3.4%
Restless Legs Syndrome	★★★★	2.5%	2.0%
Parkinson's Disease	★★★★	2.2%	1.6%

Microsoft

Google Genomics

GenomeCanada
GLOBAL CHALLENGES GENOMIC SOLUTIONS

amazon

Global Alliance for Genomics & Health

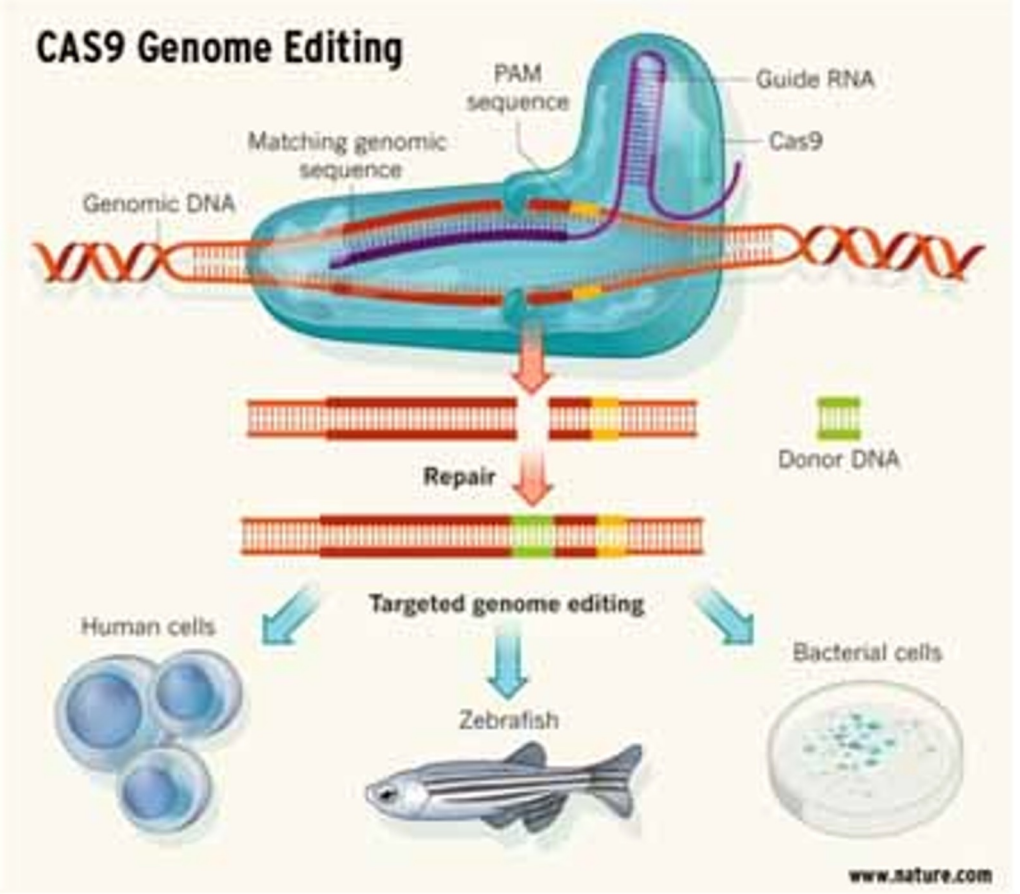
COGR
Canadian Open Genetics Repository

1000 Genomes
A Deep Catalog of Human Genetic Variation

openSNP

Genome Editing (CRISPR-CAS9)

- Potential to alter the **human** genome
- Strong potential for treatment of (human) genetic diseases
- Moratorium pronounced in December 2015 for edition of inheritable parts of the human genome
- Moratorium not fully respected…



CRISPR: Clustered regularly interspaced short palindromic repeats
CAS9 is a protein

MIT Technology Review 02.22

China's CRISPR twins might have had their brains inadvertently enhanced

CRISPR: Gene-editing technique

- Deletion of gene CCR5 was intended to make babies HIV-resistant
- This deletion could increase cognition capabilities

Rewriting Life

The Chinese scientist who claims he made CRISPR babies is under investigation

He Jiankui says he created twin girls whose genes were edited to make them resistant to HIV. Was that ethical? Or even legal?

by Antonio Regalado November 26, 2018

A **Chinese researcher who claims to have created the first** gene-edited babies, He Jiankui of the Southern University of Science and Technology (SUST), in Shenzhen, is now facing investigation over whether the experiment broke Chinese laws or regulations.

On Sunday, MIT Technology Review was first to disclose a secretive project in China to produce children whose genomes had been modified to make them resistant to HIV.

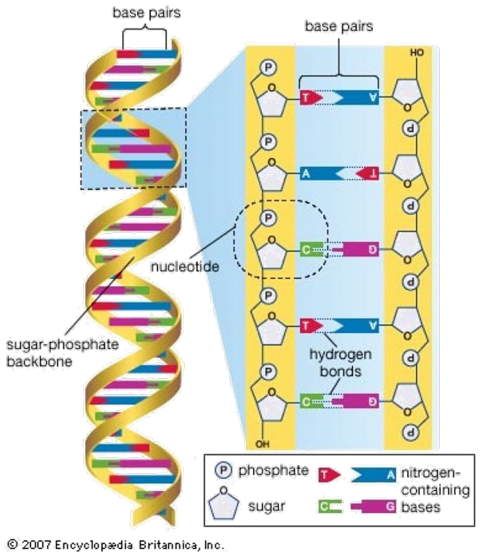
Genomics101

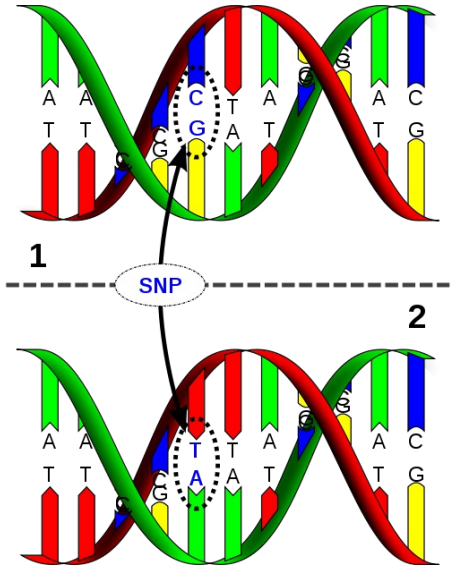
"...and if anyone here suspects that the algorithm that put these two together might be flawed, speak now..."

Source: The New Yorker

Genomics 101 - DNA and SNP

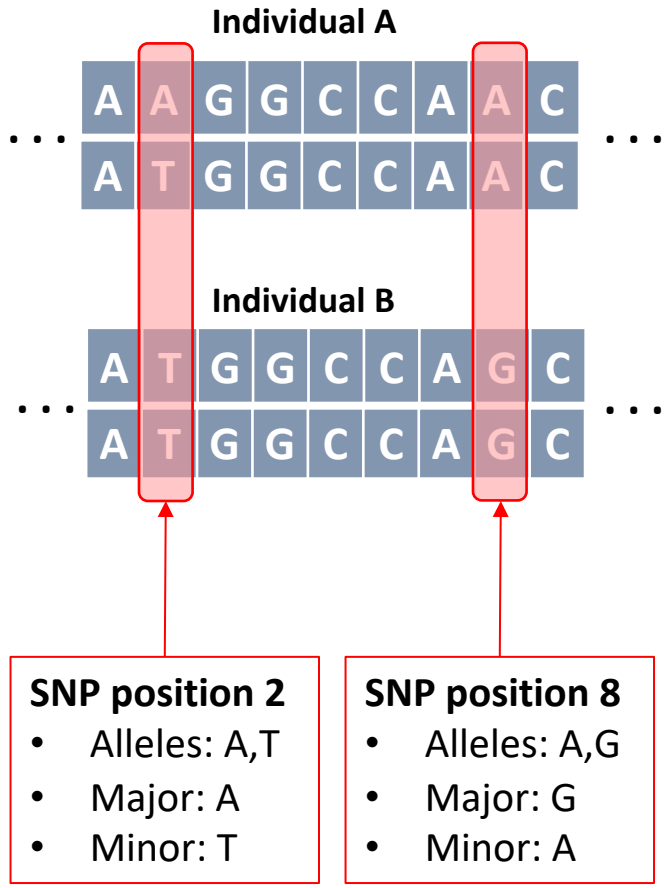
- The human genome consists of approximately 3 billion letters
  - 99.9% is identical between any two individuals
  - Remaining: human genetic variation
- Single Nucleotide Polymorphism (SNP): Most common human genetic variation.
  - A single nucleotide (A, C, G, or T) differs between members of the same species or paired chromosomes of an individual
  - Disease risk can be computed by analyzing particular SNPs
    - Angelina Jolie BRCA1 Mutation
    - 23andMe genetic disease risk tests





Most common genetic variation: Single Nucleotide Polymorphism (SNP)

- Occurs when, at a specific position, at least a single nucleotide (A,C,G, or T) differs between members of the same species in more than 1% of the population
- Potential nucleotides for a SNP are called **alleles**
- 2 different alleles can be observed for each SNP:
  - **Major** allele (M)
  - M**inor** allele (m)
- Every genome carries 2 alleles at each SNP position A SNP can be either:
  - **Homozygous minor** [m,m]
  - **Heterozygous** [m,M] or [M,m]
  - **Homozygous major** [M,M]



Alice and Bob: The Long Awaited Happy End

After having extensively authenticated each other,

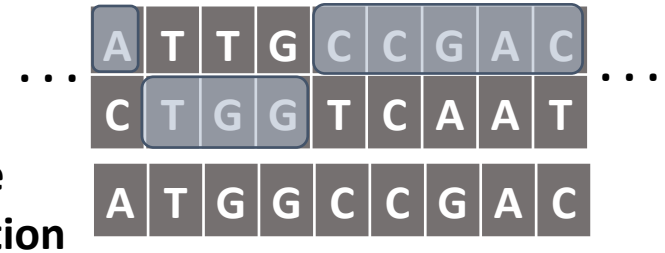
after having exchanged thousands of highly private messages,

after having established numerous secure channels between each other,

after years of intense but platonic relationship, finally, finally…❤️

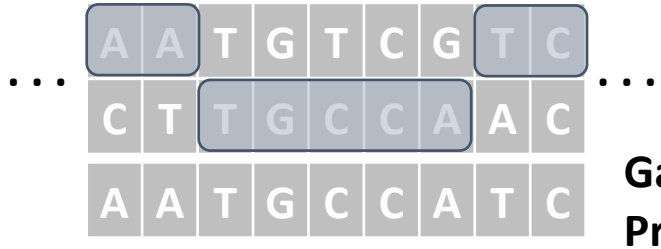
… Alice and Bob got closer to each other

Bob

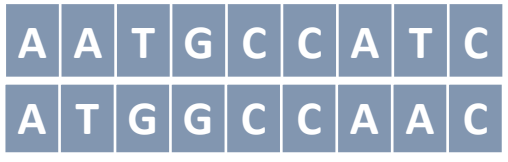


Gamete Production (spermatozoon)

Alice

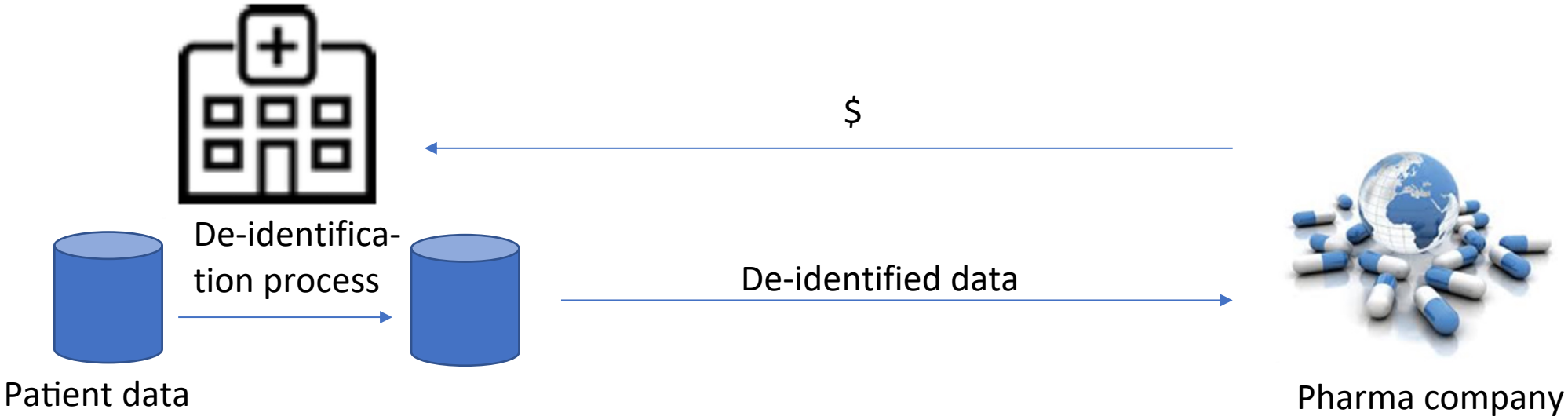


Gamete Production (ovule)



Child

Hospital / pharma interface

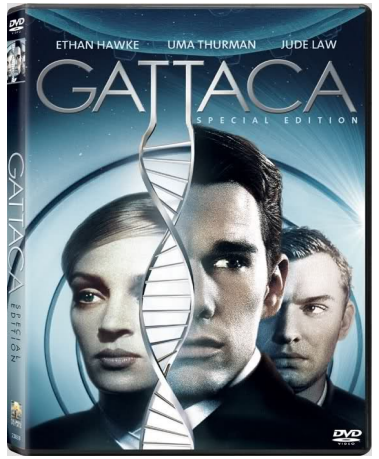


Why Protect Genomic Data?

- Genome carries sensitive information about
  - genetic condition and predispositions to specific diseases
  - physical appearance

- Leakage of genomic information could cause *genetic discrimination*
  - Denial of access to health insurance, mortgage, education, and employmen

- Anonymization is ineffective
- Genome carries information about kinship
- Genomic data is non-revokable



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]*

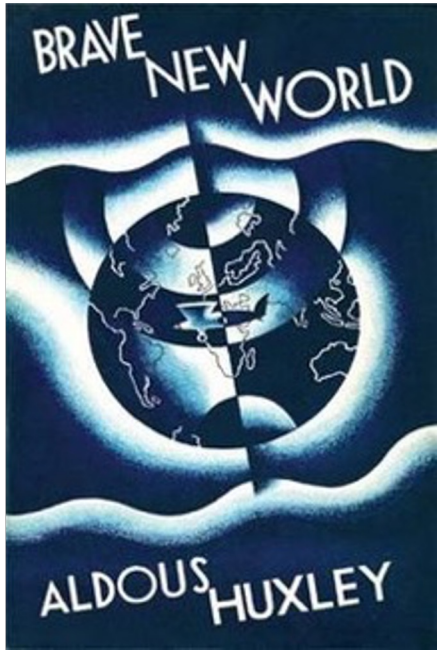
Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a
can be used to tr
relies on free, pu
identification for
with high probab

Privacy in the Genomic Era

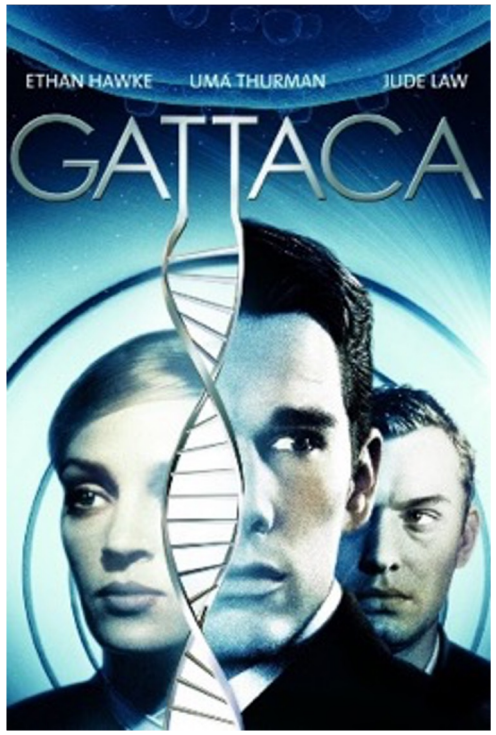
MUHAMMAD NAVEED, University of Illinois at Urbana-Champaign
ERMAN AYDAY, Bilkent University
ELLEN W. CLAYTON, Vanderbilt University
JACQUES FELLAY, Ecole Polytechnique Federale de Lausanne
CARL A. GUNTER, University of Illinois at Urbana-Champaign
JEAN-PIERRE HUBAUX, Ecole Polytechnique Federale de Lausanne
BRADLEY A. MALIN, Vanderbilt University
XIAOFENG WANG, Indiana University at Bloomington

Genome sequencing technology has advanced at a rapid pace and it is n
detailed genotypes inexpensively. The collection and analysis of such da
various applications, including personalized medical services. While the be

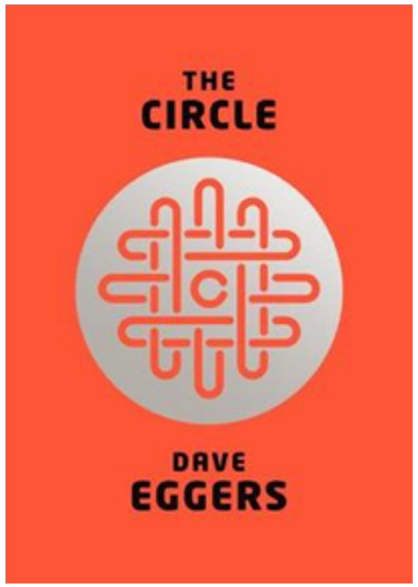
Related Fiction



Novel, 1931



Movie, 1997

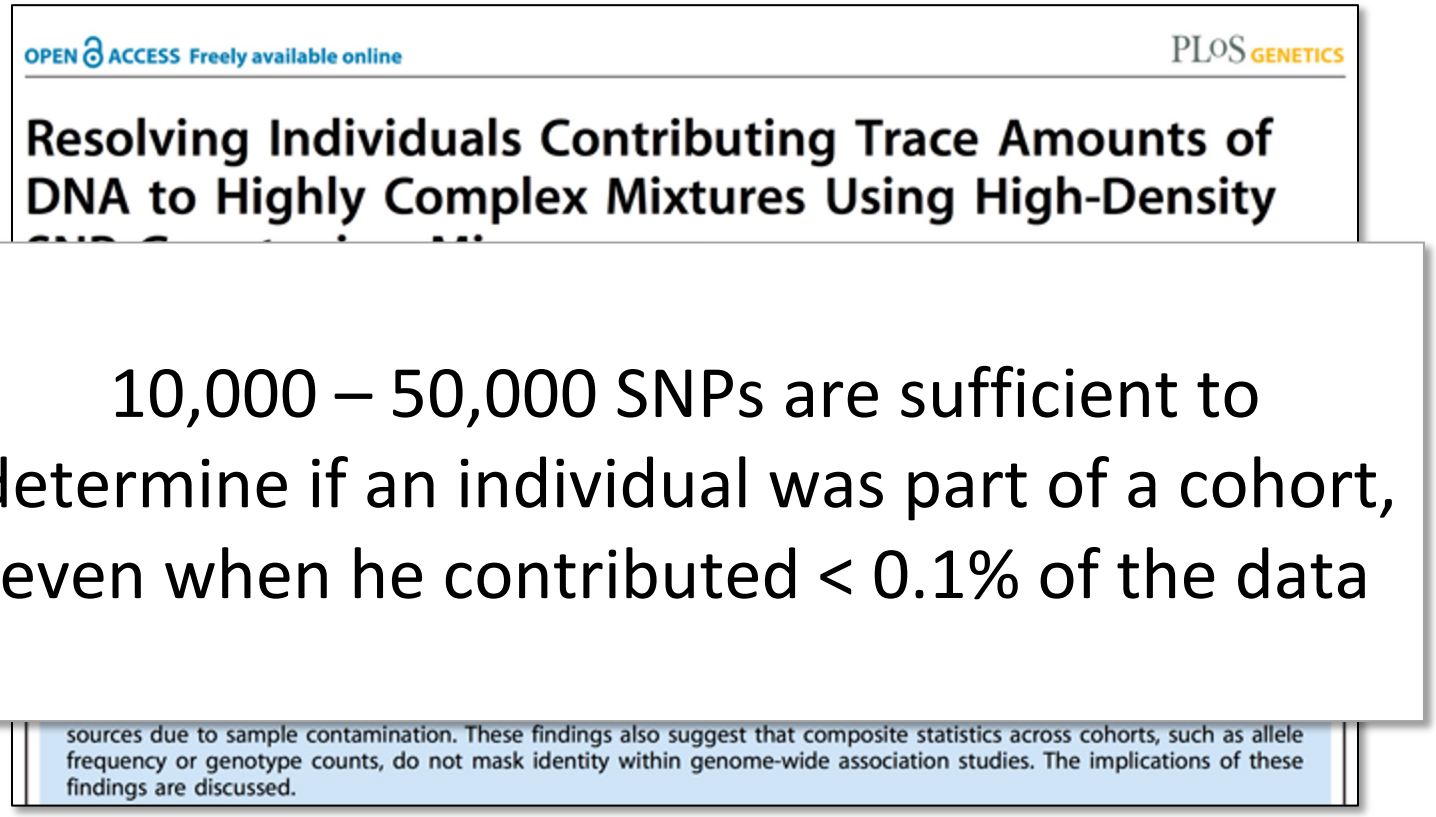


Novel, 2013

Movie, 2017

Attacks against genomic privacy

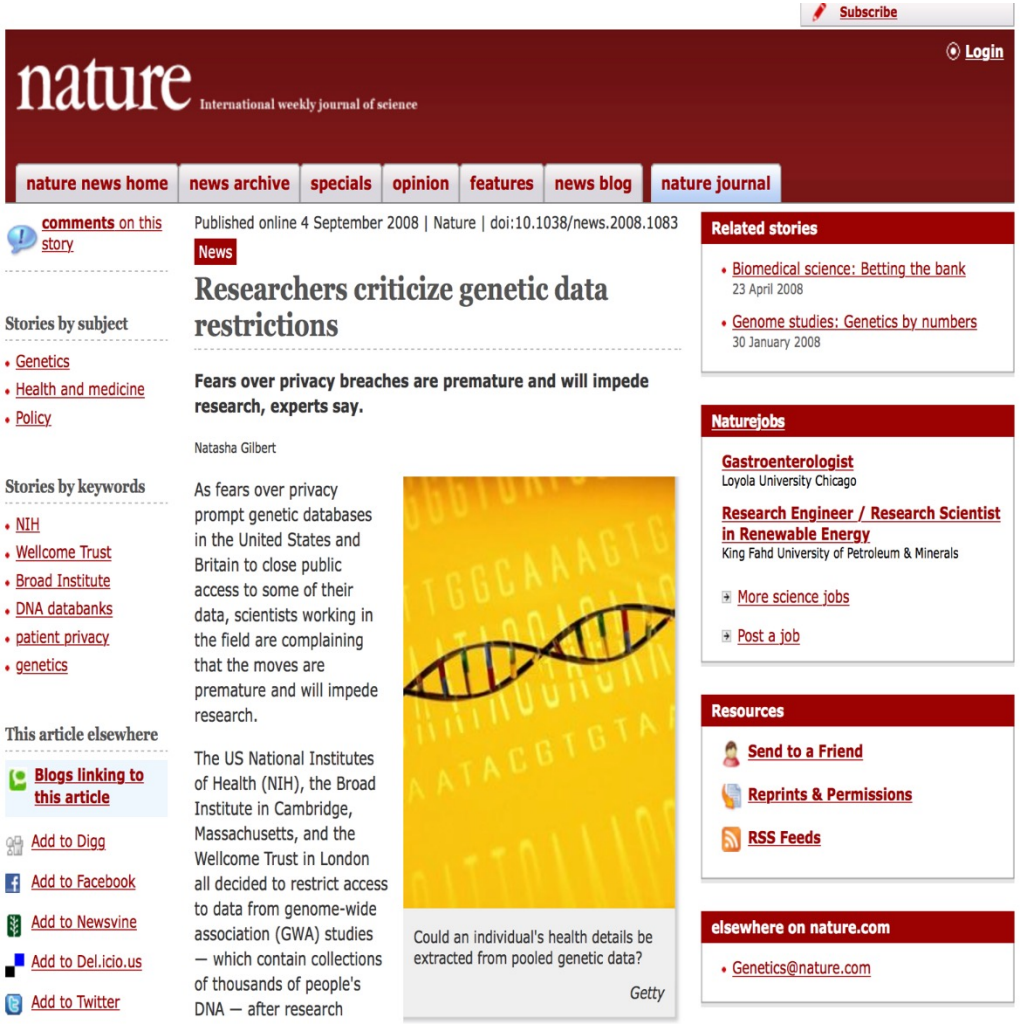
Major Concern: Re-identification Attacks against Genomic Databases



Many other subsequent studies extended the range of vulnerabilities for summary statistics:

[Jacobs et al. *Nature Genet.* '09], [Vissecher and Hill *PLoS Genet.* '09], [Sankararaman et al. *Nature Genet.* '09], [Wang et al. *CCS*'09], [Clayton *Biostatistics* '10], [Im et al. *Am. J. Hum. Genet.* '12], …

Major Concern: Re-identification Attacks against Genomic Databases



Subscribe

Login

nature International weekly journal of science

nature news home | news archive | specials | opinion | features | news blog | nature journal

comments on this story

Published online 4 September 2008 | Nature | doi:10.1038/news.2008.1083

News

Researchers criticize genetic data restrictions

Fears over privacy breaches are premature and will impede research, experts say.

Natasha Gilbert

As fears over privacy prompt genetic databases in the United States and Britain to close public access to some of their data, scientists working in the field are complaining that the moves are premature and will impede research.

The US National Institutes of Health (NIH), the Broad Institute in Cambridge, Massachusetts, and the Wellcome Trust in London all decided to restrict access to data from genome-wide association (GWA) studies — which contain collections of thousands of people's DNA — after research

Could an individual's health details be extracted from pooled genetic data?

Getty

Stories by subject
- Genetics
- Health and medicine
- Policy

Stories by keywords
- NIH
- Wellcome Trust
- Broad Institute
- DNA databanks
- patient privacy
- genetics

This article elsewhere

Blogs linking to this article

Add to Digg

Add to Facebook

Add to Newsvine

Add to Del.icio.us

Add to Twitter

Related stories
- Biomedical science: Betting the bank 23 April 2008
- Genome studies: Genetics by numbers 30 January 2008

Naturejobs

Gastroenterologist
Loyola University Chicago

Research Engineer / Research Scientist in Renewable Energy
King Fahd University of Petroleum & Minerals

More science jobs

Post a job

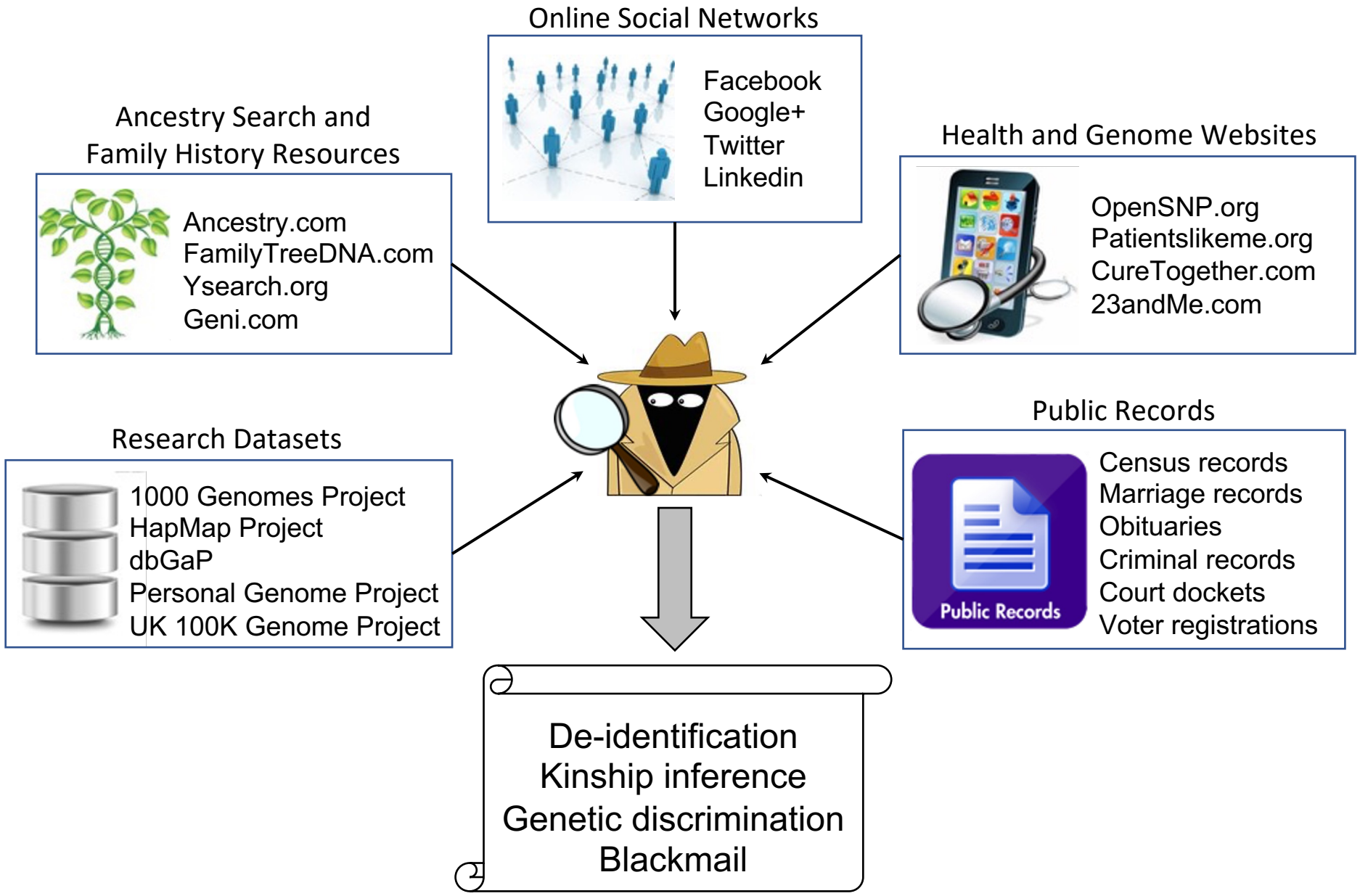
Resources

Send to a Friend

Reprints & Permissions

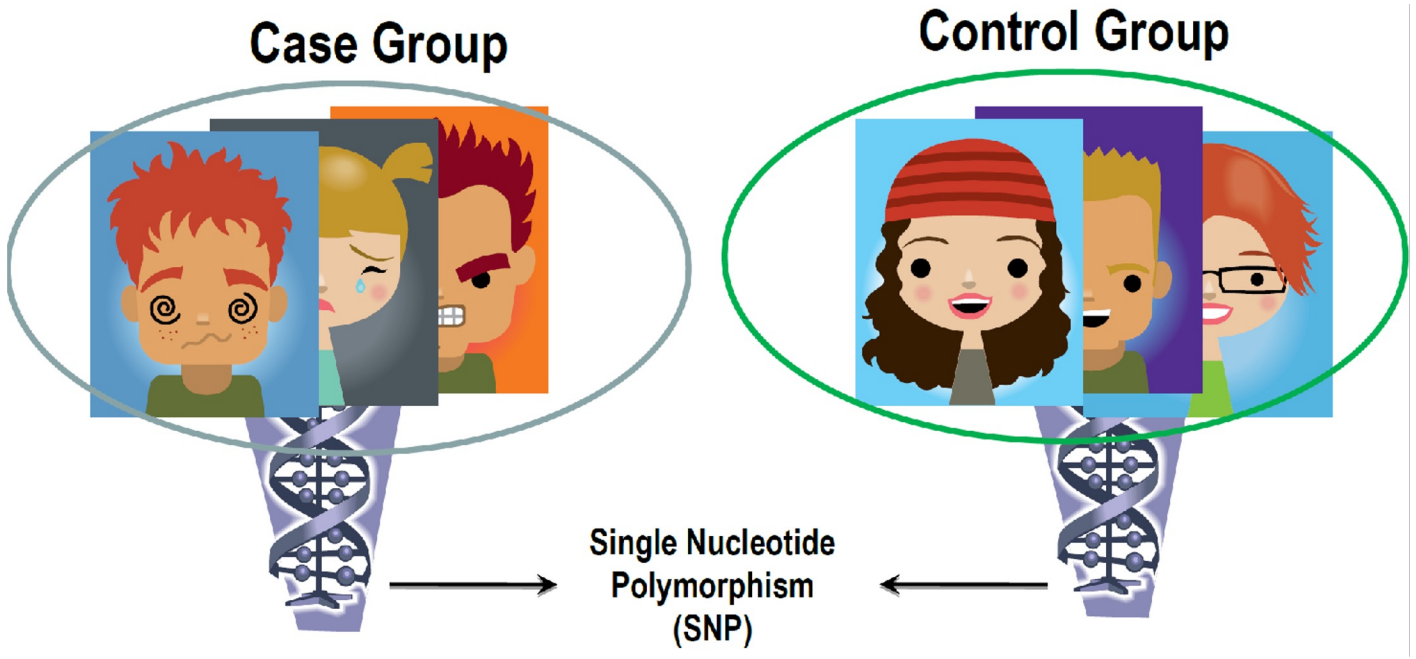
RSS Feeds

elsewhere on nature.com
- Genetics@nature.com


Online Social Networks
Facebook
Google+
Twitter
Linkedin
Ancestry Search and
Family History Resources
Ancestry.com
FamilyTreeDNA.com
Ysearch.org
Geni.com
Health and Genome Websites
OpenSNP.org
Patientslikeme.org
CureTogether.com
23andMe.com
Research Datasets
1000 Genomes Project
HapMap Project
dbGaP
Personal Genome Project
UK 100K Genome Project
Public Records
Census records
Marriage records
Obituaries
Criminal records
Court dockets
Voter registrations
Public Records
De-identification
Kinship inference
Genetic discrimination
Blackmail

Homer's Attack

- Adversary has access to a known participant's genome
- Goal: determine if the target individual is in the case group
- Uses simple correlation in the genome (linkage disequilibrium)
- Attack later improved by Wang et al.



N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics, 4, Aug. 2008.

Homer's Attack

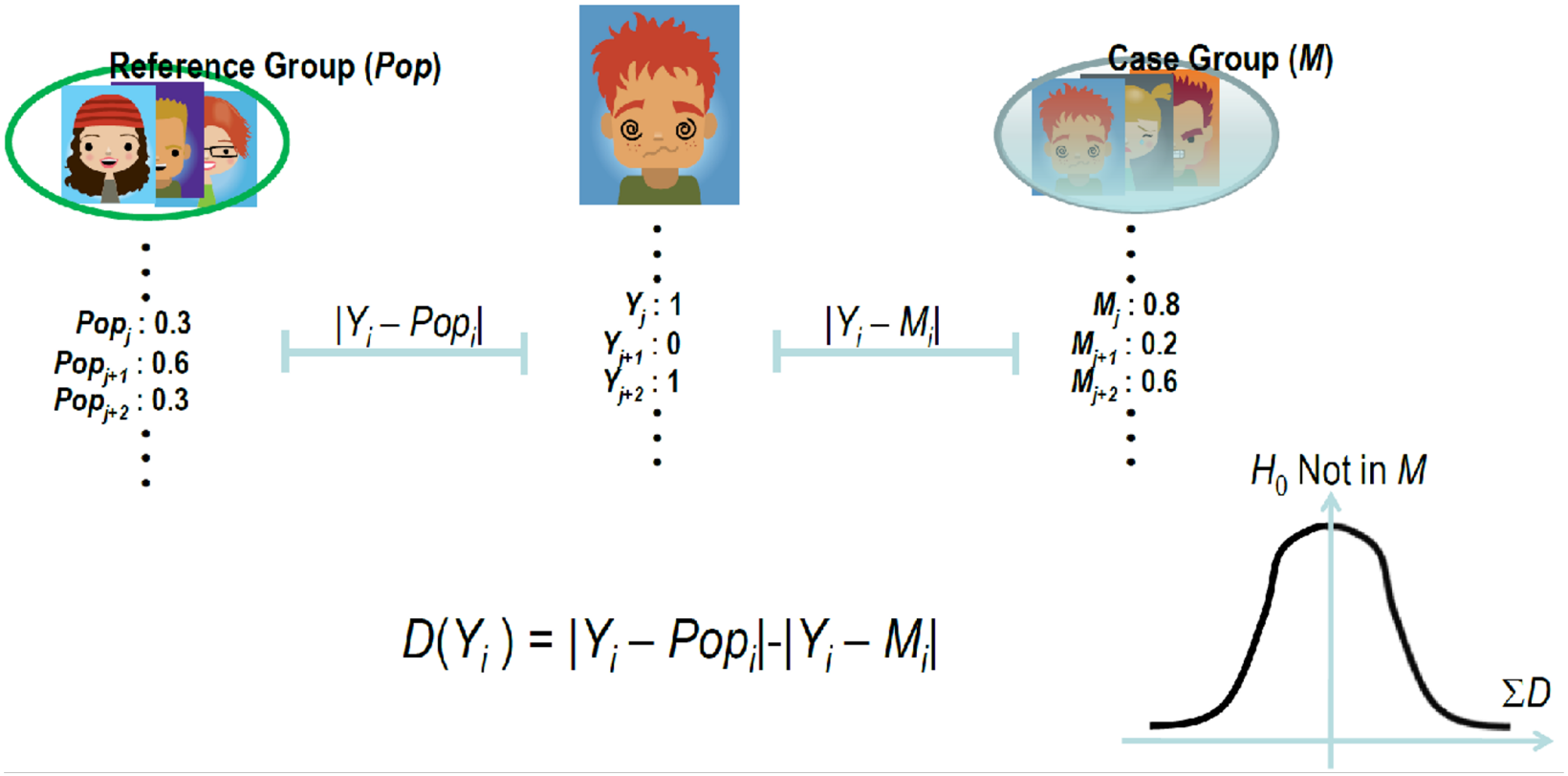


Figure from: Wang et al. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome-Wide Association Study

N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics, 4, Aug. 2008.

Homer's attack in a nutshell

The attacker knows:

- The genome of the victim (her set of variants)
- The size of the mixture he's attacking
- Population allele frequencies

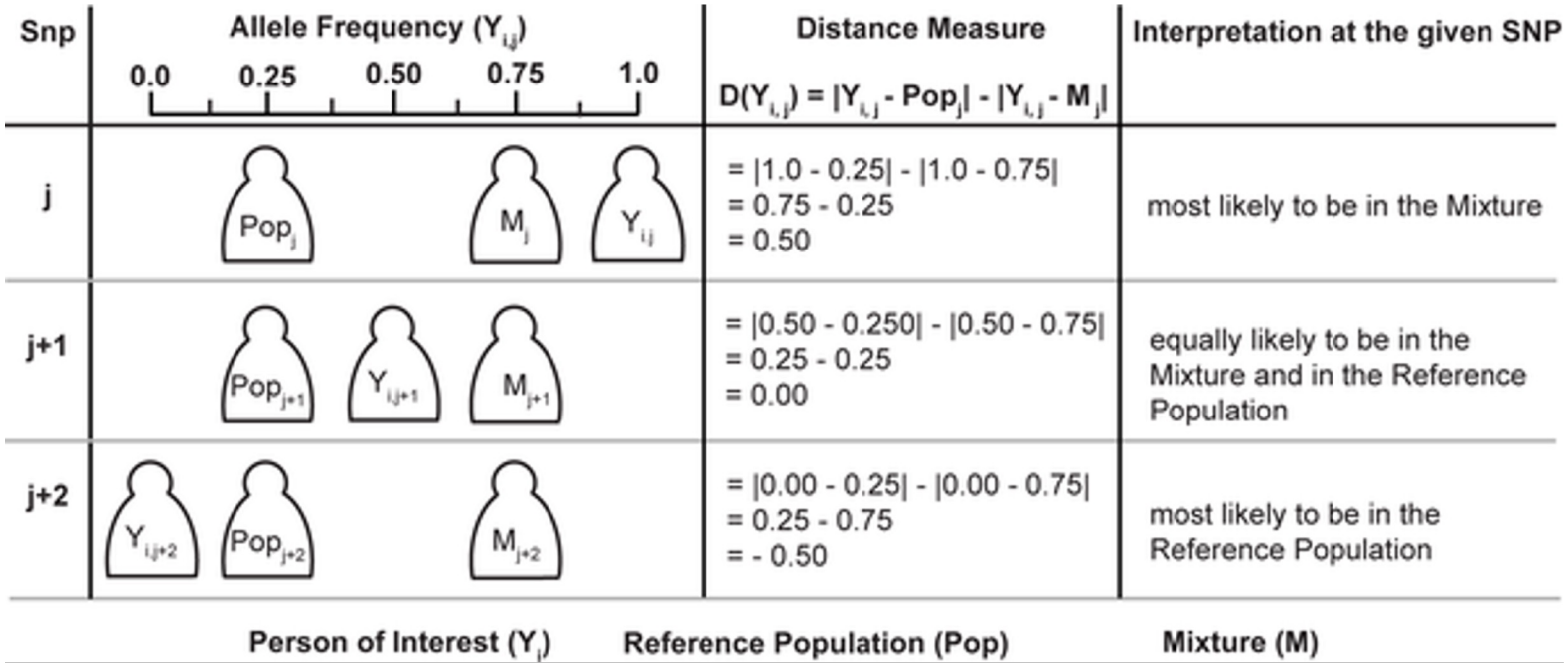
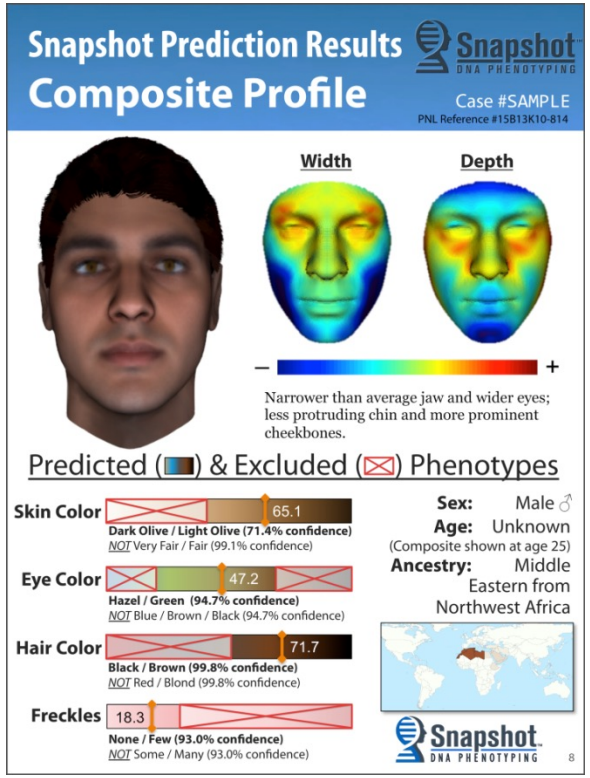
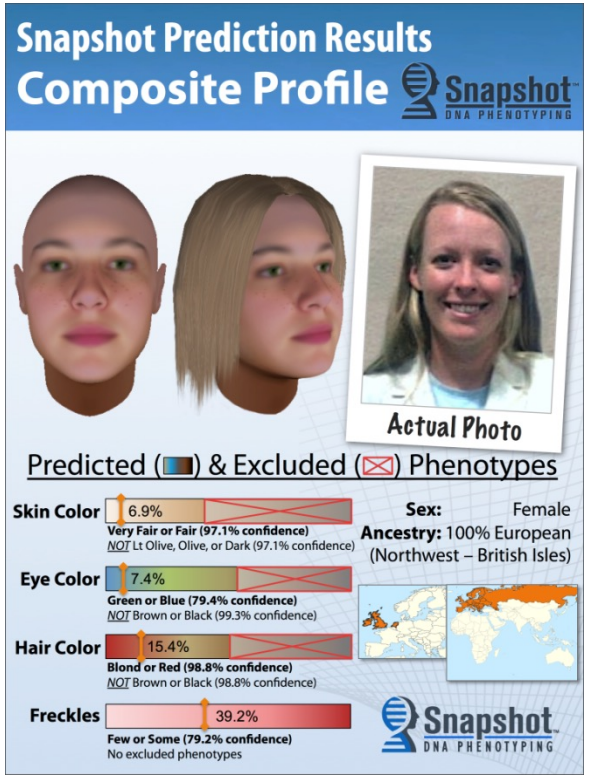


Figure taken from: Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genet 4(8): e1000167. doi:10.1371/journal.pgen.1000167

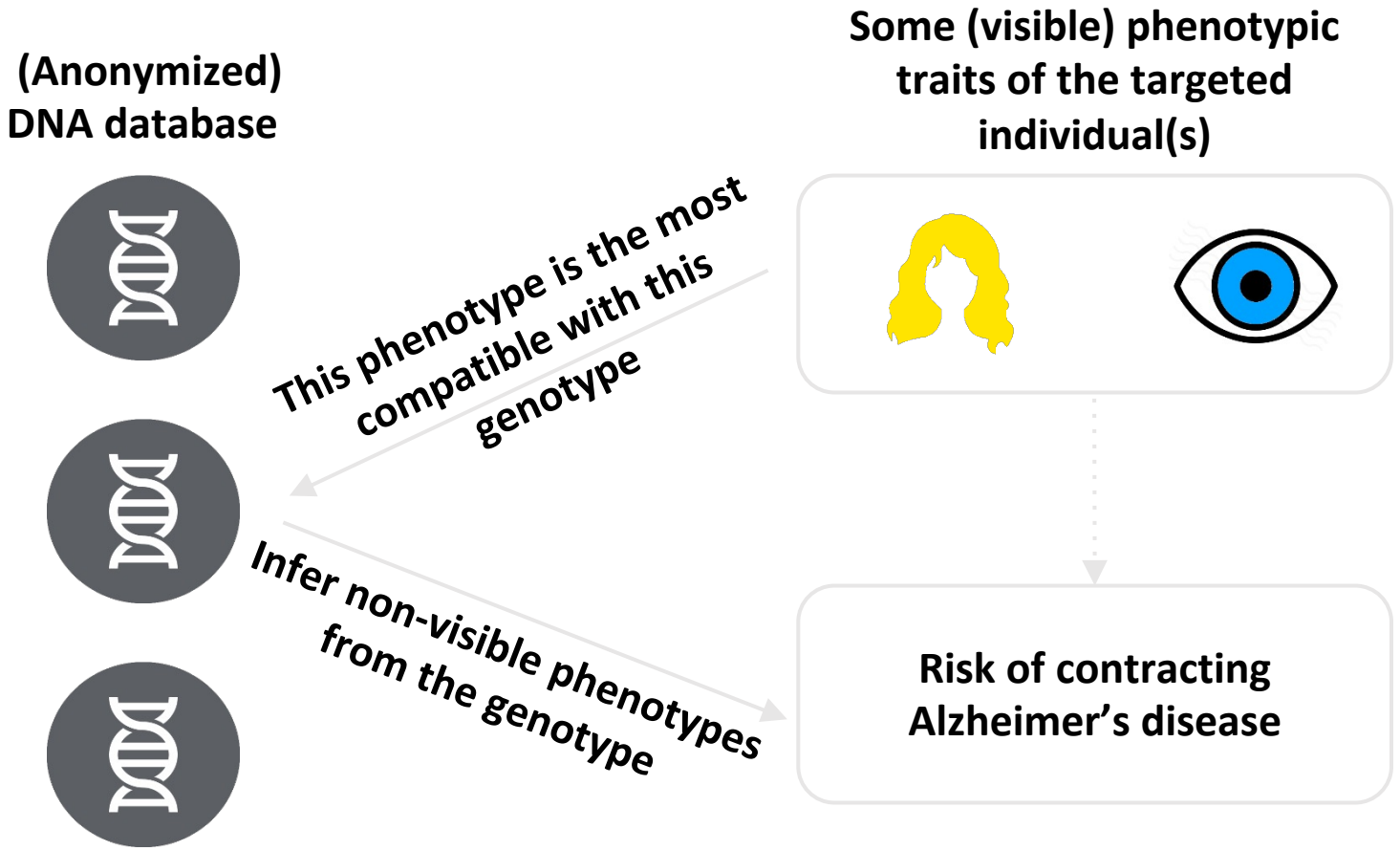
Genomic-Phenotypic Relations

- Genomic data can be used to infer physical traits
- Phenotypic information can be used to infer genomic data





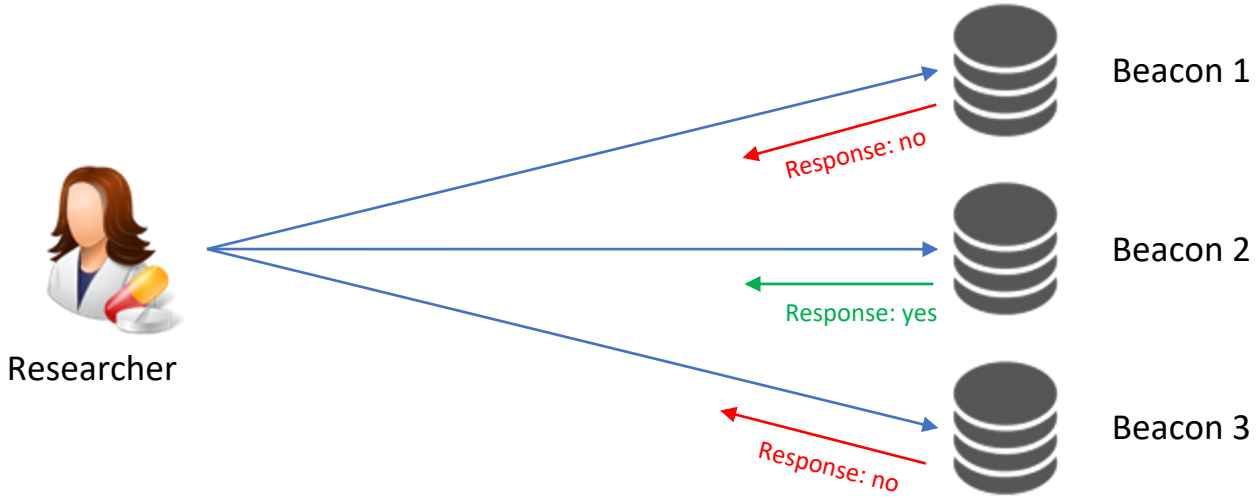
Threat



Given a database with anonymized DNA, can we find some specific individuals from some of their visible phenotypic traits?

GA4GH Beacon Project

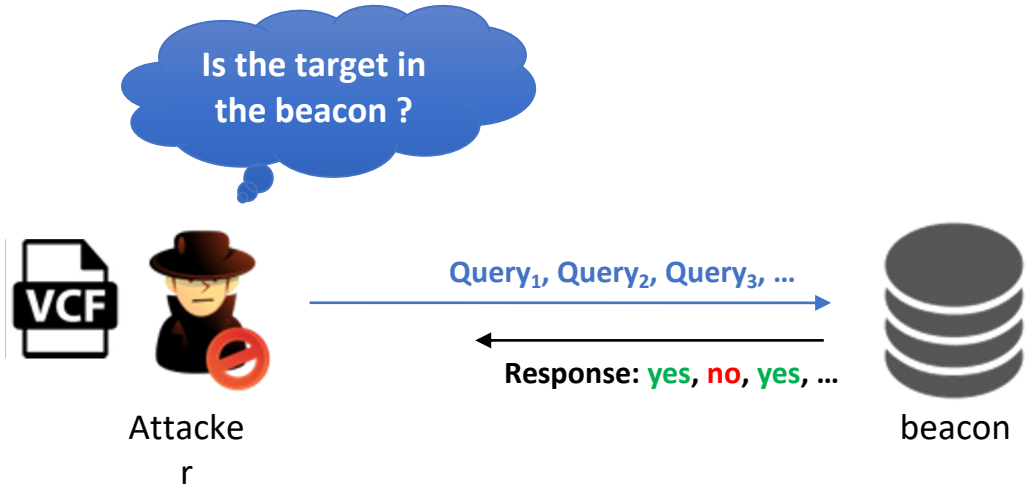
"A Beacon is a genomics discovery tool which allows to aggregate worldwide genomics dataset through a shared query protocol."



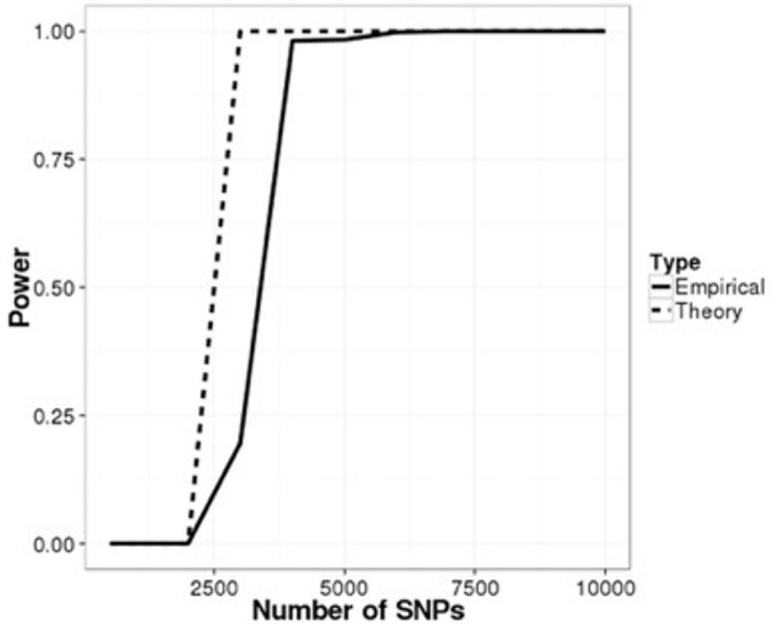
Main features:

- Allows researchers to quickly query multiple database to find the sample they need
- Encourages cross-borders collaboration among researchers
- Only provides minimal responses back in order to mitigate privacy concerns

Beacon used as an oracle: the SB attack



Shringarpure SS, Bustamante CD. Privacy risks from genomic data-sharing beacons. The American Journal of Human Genetics. 2015 Nov 5;97(5):631-46.



- The attack relies on the assumption that the adversary knows:
  - **The set of variants (VCF file) of the target individual**
  - The size of the beacon
- The attack is based on a likelihood ratio test where the adversary repeatedly queries the beacon in order to re-identify the individual
- The attack can be **extremely dangerous** if the beacon is associated with a sensitive phenotype (e.g., cancer)

“Optimal” re-identification attack with real allele frequencies

- **Smarter adversary** who makes use of publicly available information
- **Idea**: rare alleles have higher re-identification power

➔ The attacker queries rarest alleles first $f_1 \leq f_2 \leq \cdots \leq f_M$ ➔

Raisaro JL, et al. “*Addressing Beacon Re-Identification Attacks: Quantification and Mitigation of Privacy Risks*”. JAMIA, February 2017

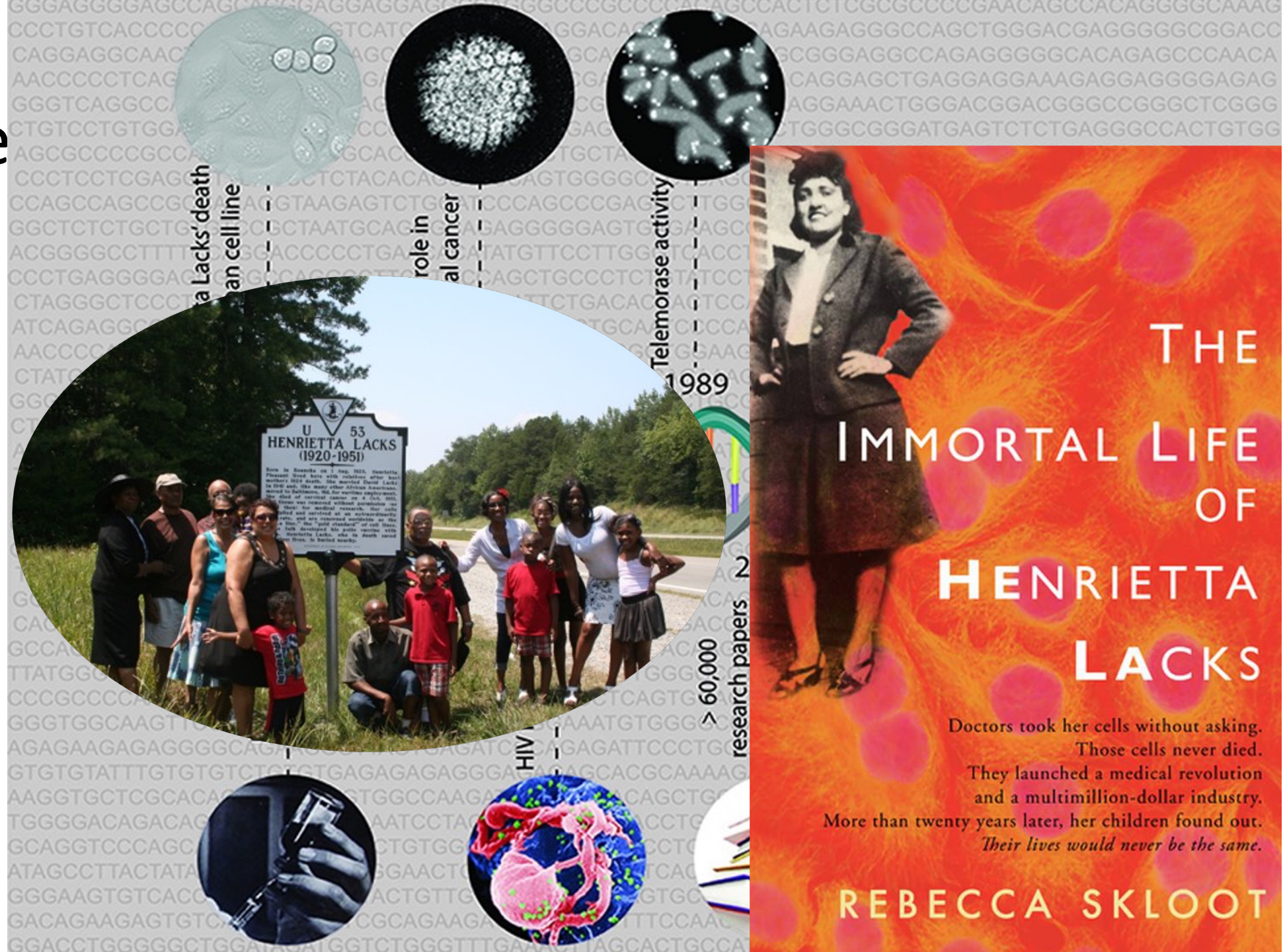
“Even more Optimal” re-identification attack

- von Thenen et al. (2019) exploits the correlations between SNPs:
  - They probabilistically infer the answers of the beacon for a SNP based on answers for other correlated SNPs to reduce the number of queries posed
  - Even if the victim **hides** their rare SNPs from the beacon, the method infers whether the victim has the SNP using answers for correlated SNPs

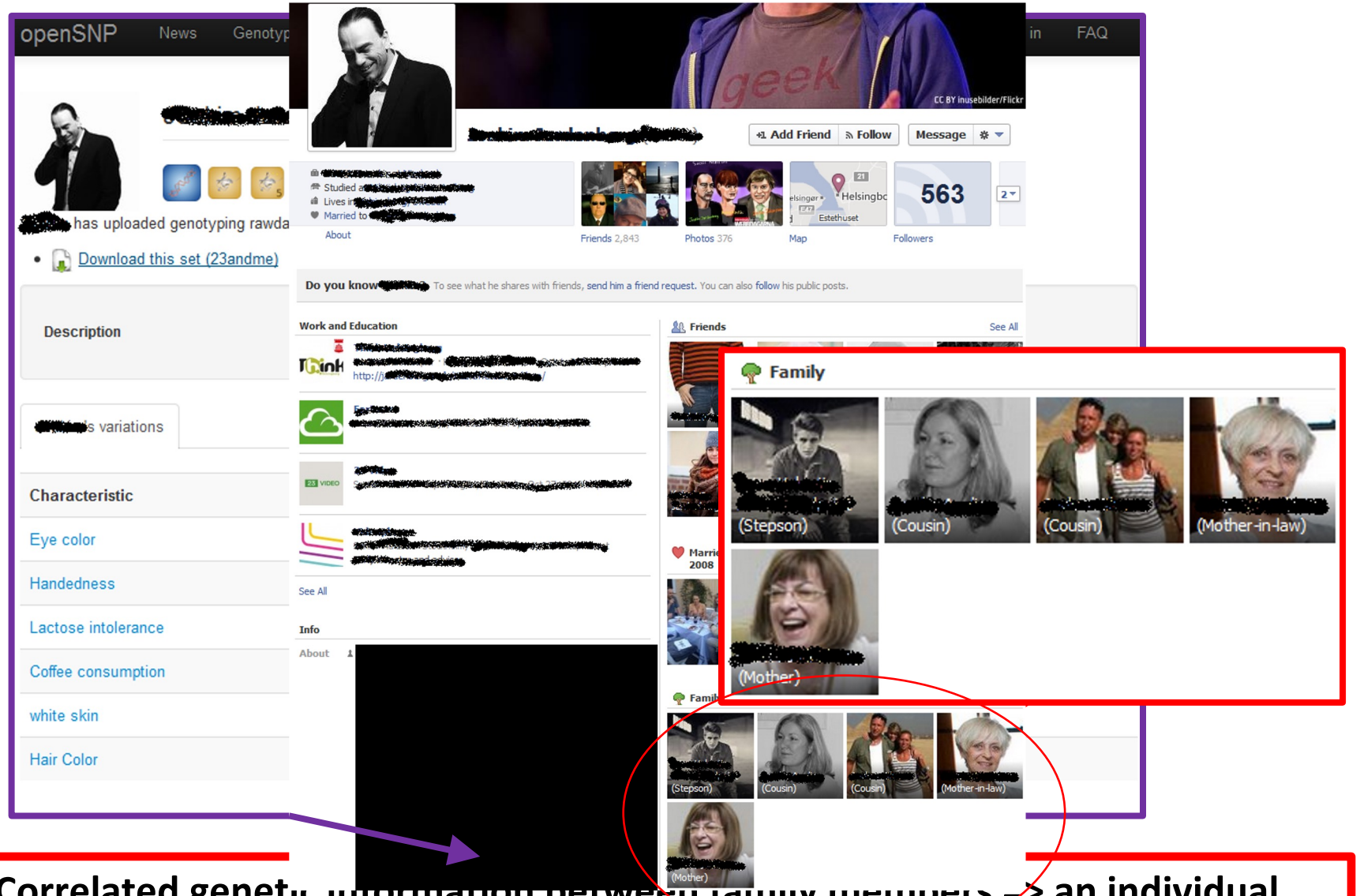
von Thenen N et al. "*Re-identification of individuals in genomic data-sharing beacons via allele inference.*" Bioinformatics 35.3 (2019): 365-371.

Kin Genomic Privacy

He

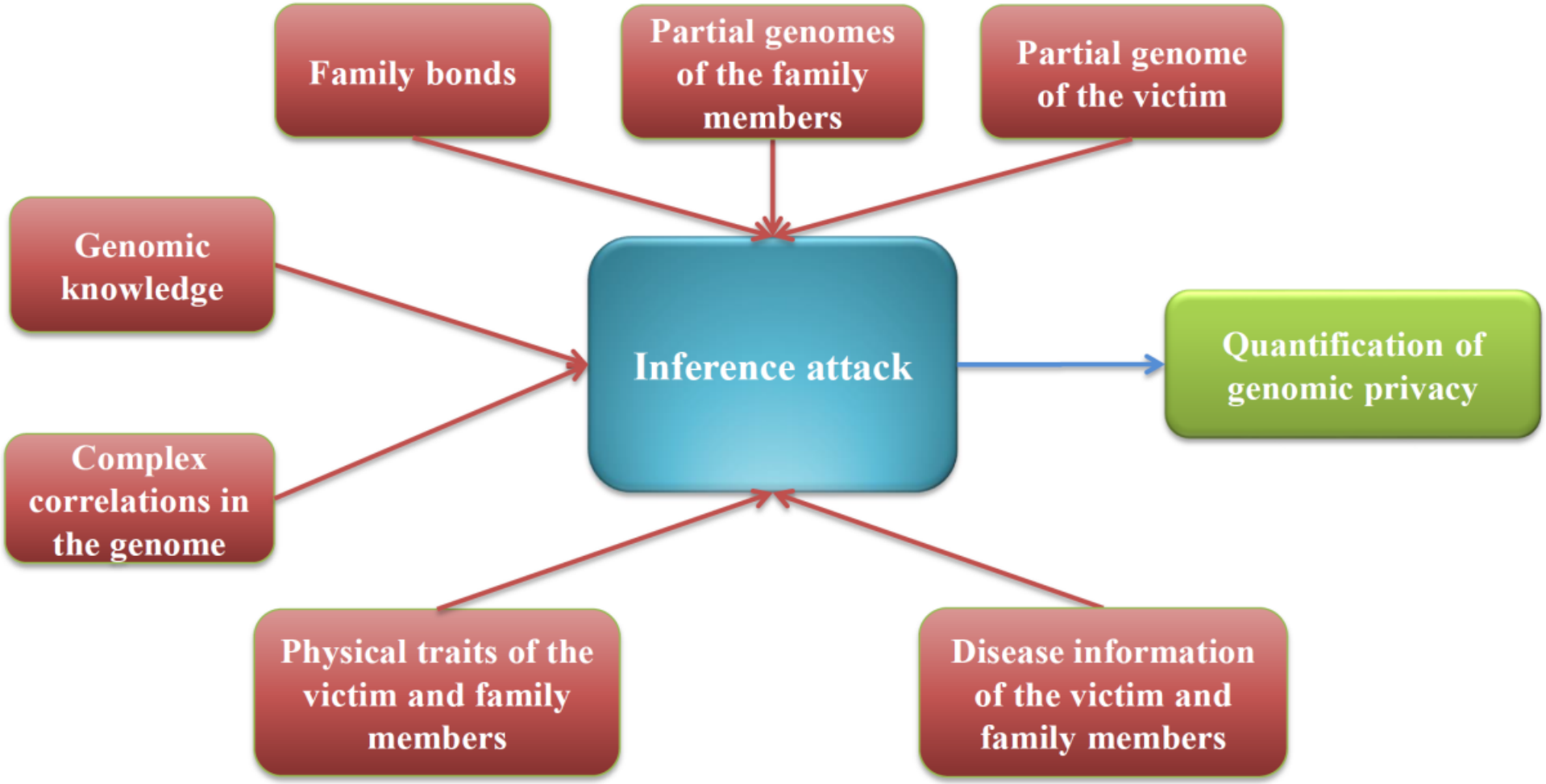

a Lacks' death
an cell line
role in
al cancer
Telemorase activity
1989
2
> 60,000
research papers
HIV
U 53
HENRIETTA LACKS
(1920-1951)
THE IMMORTAL LIFE OF HENRIETTA LACKS
Doctors took her cells without asking.
Those cells never died.
They launched a medical revolution
and a multimillion-dollar industry.
More than twenty years later, her children found out.
Their lives would never be the same.
REBECCA SKLOOT

Kin Genomic Privacy



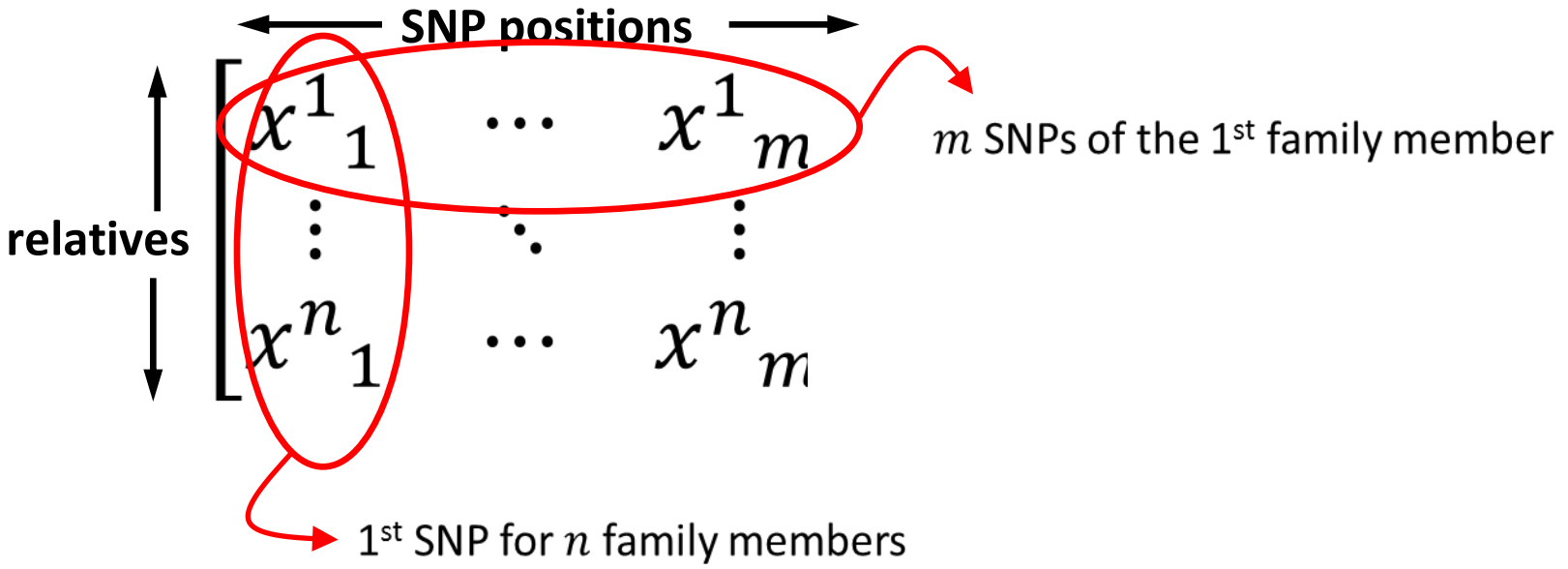
Correlated genetic information between family members –> an individual sharing his/her genome threatens his (known) relatives' genomic privacy

Framework


Family bonds
Partial genomes of the family members
Partial genome of the victim
Genomic knowledge
Inference attack
Quantification of genomic privacy
Complex correlations in the genome
Physical traits of the victim and family members
Disease information of the victim and family members

Parameters

- m : Number of SNPs
- n : Number of family members
- x^i_j: Value of SNP j for individual i
- $x^i_j \in \{0,1,2\}$
- $\mathbb{X}$: $m \times n$ matrix that stores the SNPs of all family members



Reconstruction Attack

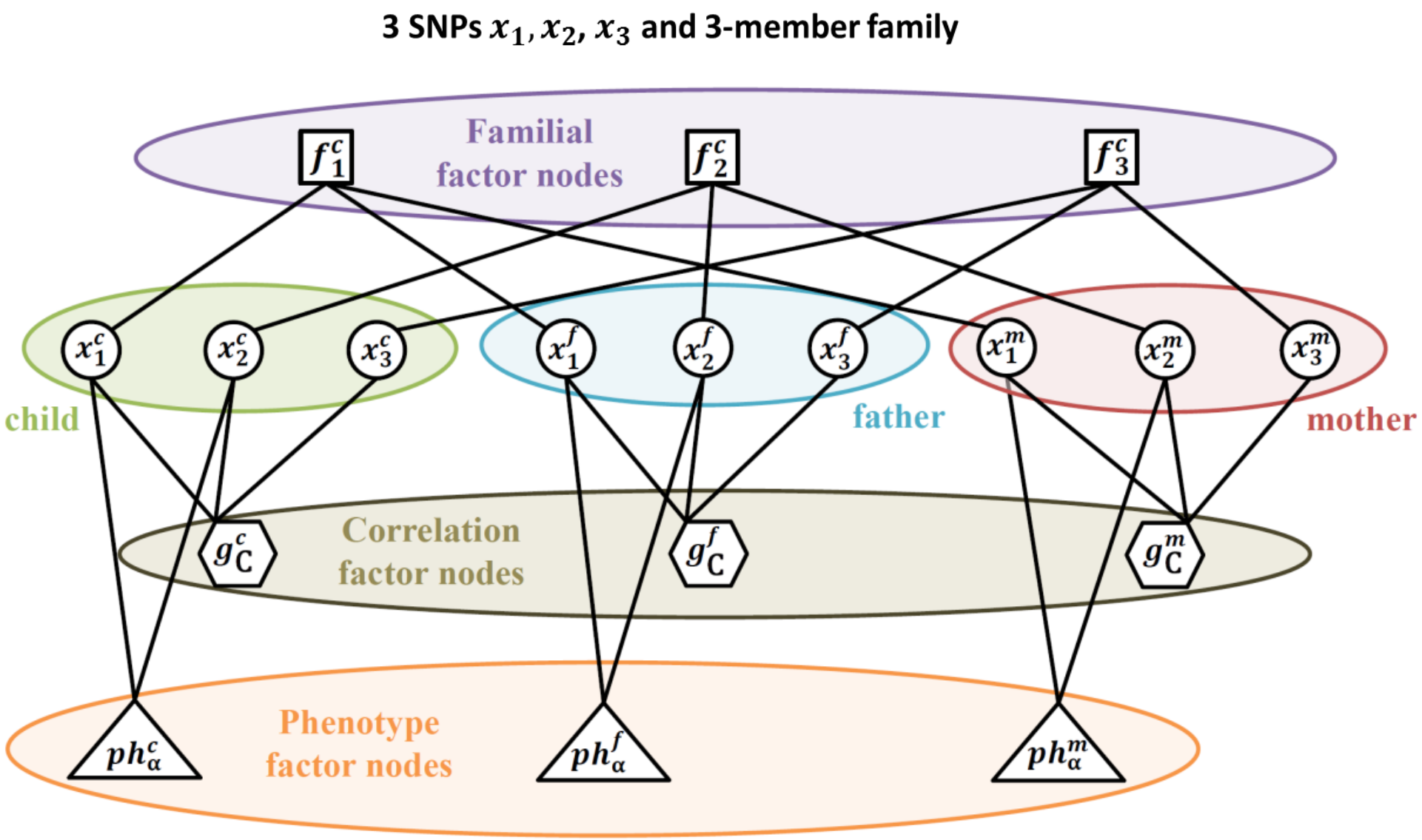
- $\mathbb{X}_U$: Set of unknown SNPs
- $\mathbb{X}_K$: Set of known SNPs
- Attacker's objective: Compute the marginal probabilities of the SNPs in $\mathbb{X}_U$
  - $p(x^i{}_j|\mathbb{X}_K) = \sum_{\mathbb{X}_U \setminus \{x^i{}_j\}} p(\mathbb{X}_U|\mathbb{X}_K, \mathcal{B})$,
    - $p(x^i{}_j|\mathbb{X}_K)$: Marginal probability distribution of SNP j for individual i can be obtained from
    - $p(\mathbb{X}_U|\mathbb{X}_K, \mathcal{B})$: Joint probability distribution function of the variables in $\mathbb{X}_U$ such that:
    - $\mathcal{B} = (\mathcal{F}_R(x^M{}_j, x^F{}_j, x^C{}_j), \mathbb{L}, \mathcal{G}_F, \mathbf{P})$: Background knowledge of the attacker

Mendel's rule | Markov chain or HMM | Family tree | Phenotype

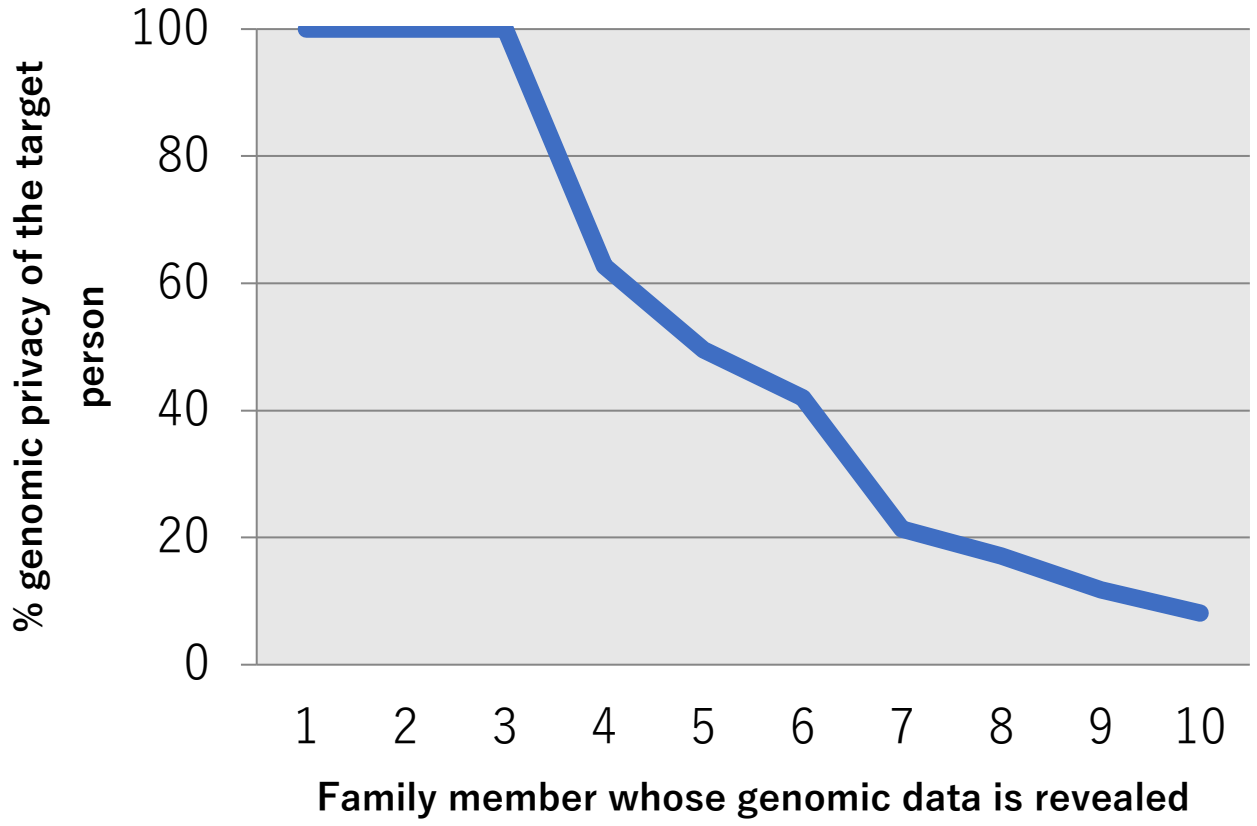
Efficient Inference Algorithm

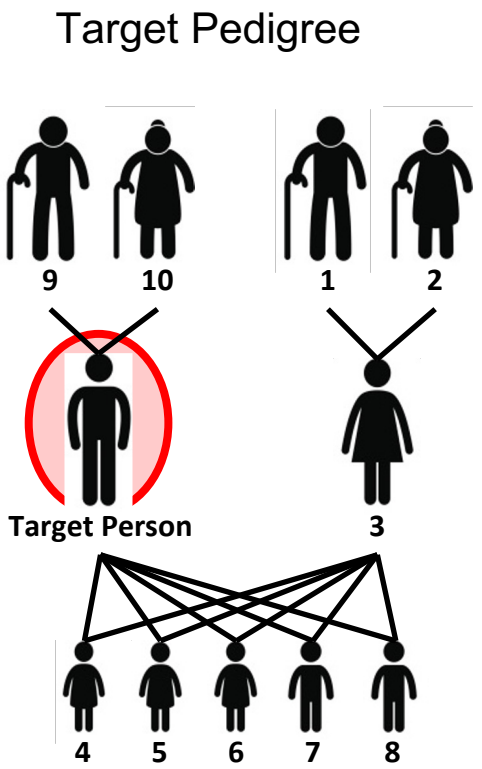
- Naive marginalization has computational complexity $\mathcal{O}(3^{mn})$
  - m is on the order of 10s of millions for human genome
- Run the belief propagation algorithm on a factor graph to reduce the computational complexity
  - Complexity = $\mathcal{O}(mn)$ per iteration

Setup and Message Passing



Quantifying Kin Genomic Privacy

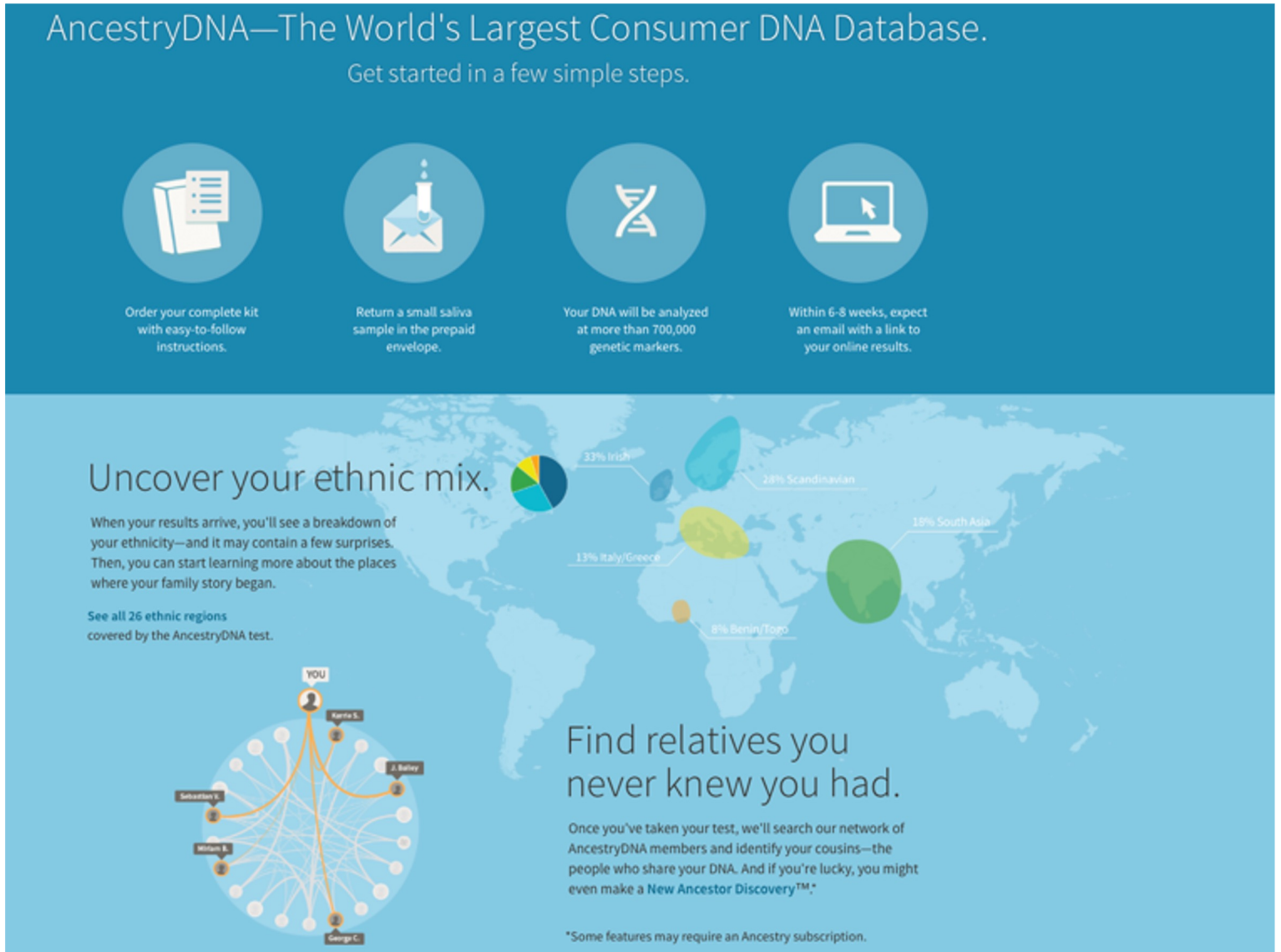




M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Addressing the concerns of the Lacks family: Quantication of kin genomic privacy. CCS, 2013.

Direct to Consumer Genomics (1/2)

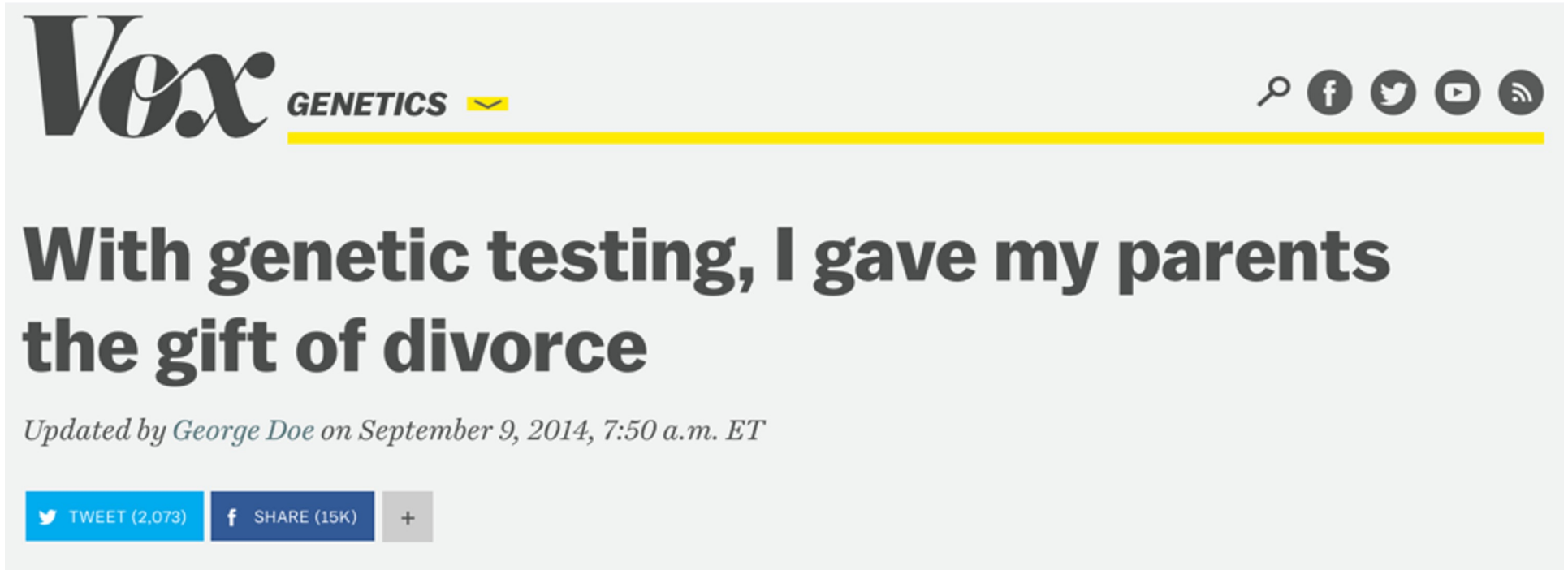
- Ancestry.com (millions of customers)



Direct to Consumer Genomics (2/2)

- 23andMe.com
  (millions customers)

Name	Confidence	Your Risk	Avg. Risk
Atrial Fibrillation	★★★★	33.9%	27.2%
Prostate Cancer ♂	★★★★	29.3%	17.8%
Alzheimer's Disease	★★★★	14.2%	7.2%
Age-related Macular Degeneration	★★★★	11.1%	6.5%
Colorectal Cancer	★★★★	7.8%	5.6%
Chronic Kidney Disease	★★★★	4.2%	3.4%
Restless Legs Syndrome	★★★★	2.5%	2.0%
Parkinson's Disease	★★★★	2.2%	1.6%


Vox
GENETICS
With genetic testing, I gave my parents the gift of divorce
Updated by George Doe on September 9, 2014, 7:50 a.m. ET
TWEET (2,073)
SHARE (15K)
+

Surname Inference Attack

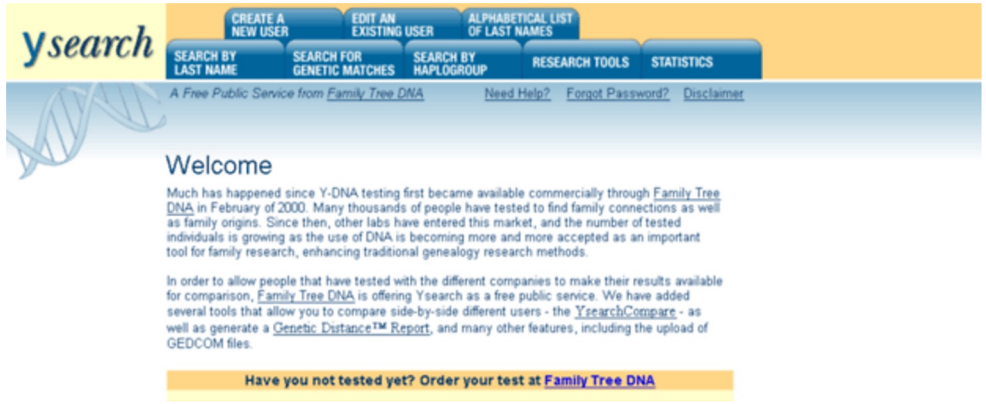
- Goals:
  - Recover the surname of sequence donors from 1000 Genome Project
  - Triangulate the identity of a sequence donor using his surname, age and state
- Using:
  - Surnames are paternally inherited in most human societies
  - Y-chromosome haplotypes in male individuals are directly inherited from the father

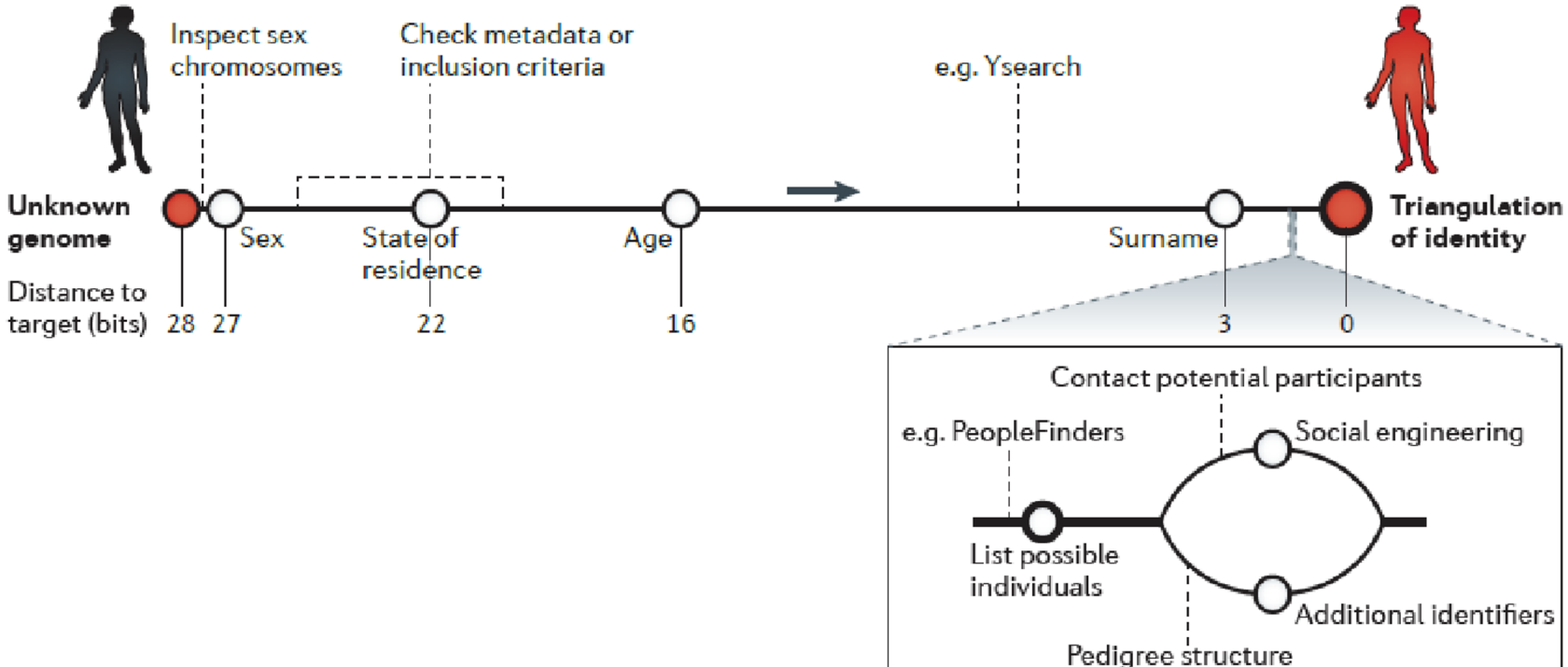
M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. Science: 339 (6117), January 2013.

Surname Inference Attack

- Surname inference
  - Profile short tandem repeats (STR) on the Y-chromosome
    - STR: A small repeating sequence in DNA.
  - Query recreational genetic genealogy databases
  - Obtain a list of possible surnames for the sequence in question
- Identity Triangulation
  - Combine surnames with age and state
  - Triangulate the identity of the target (using US census database)

Surname Inference Attack





How an Unlikely Family History Website Transformed Cold Case Investigations

Fifteen murder and sexual assault cases have been solved since April with a single genealogy website. This is how GEDmatch went from a casual side project to a revolutionary tool.

Curtis Rogers enjoys helping people solve family history puzzles. He inadvertently created a da
of Americans of Northern European ancestry. Scott McIntyre for The New York Times

Los Angeles Times

The untold story of how the Golden State Killer was found: A covert operation and private DNA

By Heather Murphy

Oct. 15, 2018

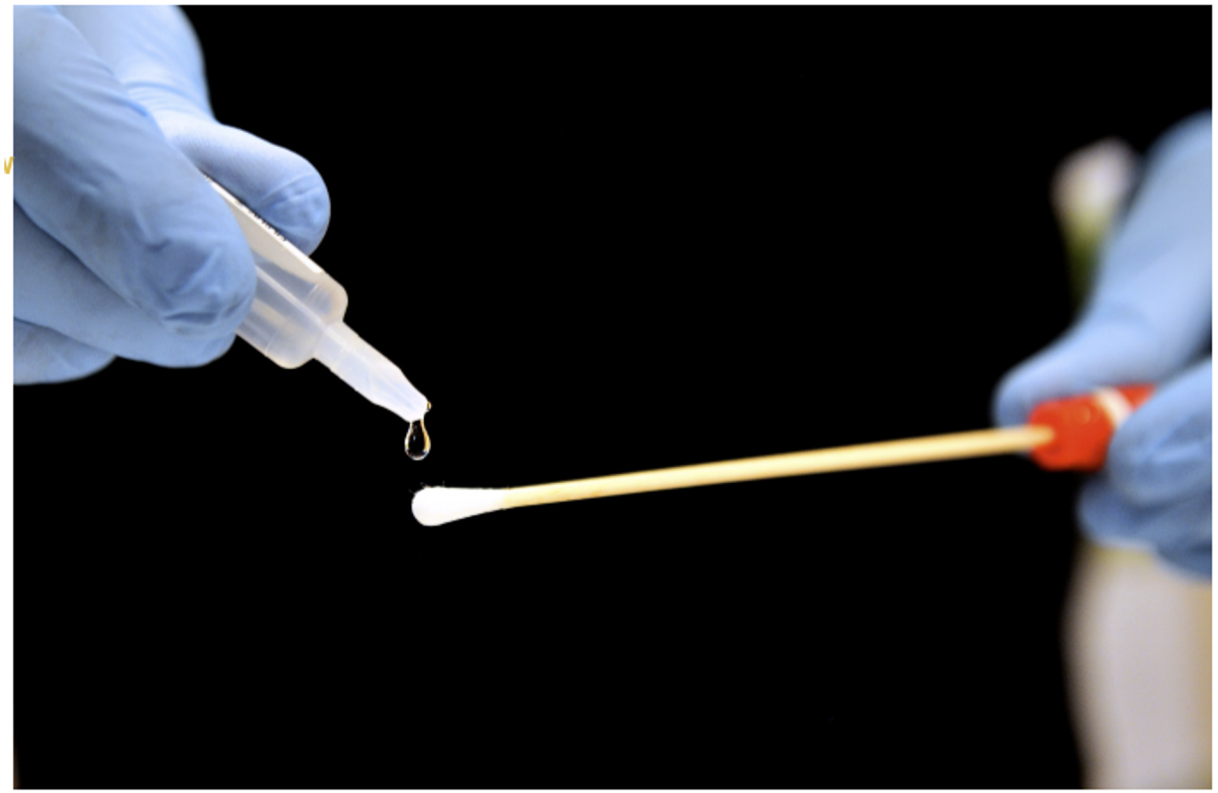
95

LAKE WORTH, FLA. — On Halloween night in 1996, a man in a skeleton mask knocked on the door of a house in Martinez, Calif., handcuffed the

A popular genealogy website just helped solve a serial killer cold case in Oregon

Comment

Taylor Hatmaker @tayhatmaker / 4 months ago



On Thursday, detectives in Portland, Ore. announced that a long-cold local murder case finally came to a resolution, 40 years after the fact.

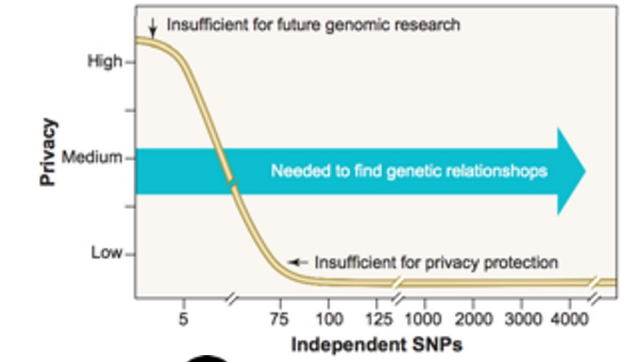
In 1979, 20-year-old Anna Marie Hlavka was found dead in the Portland apartment she shared with her fiance and sister. According to police, she was strangled to death and sexually assaulted. Police followed a number of leads and kept tabs on the case for decades without a breakthrough.

Last May, detectives with Portland's Cold Case Homicide Detail dug back into the case using the methodology made famous when investigators last year tracked down the man believed to be the Golden State Killer.

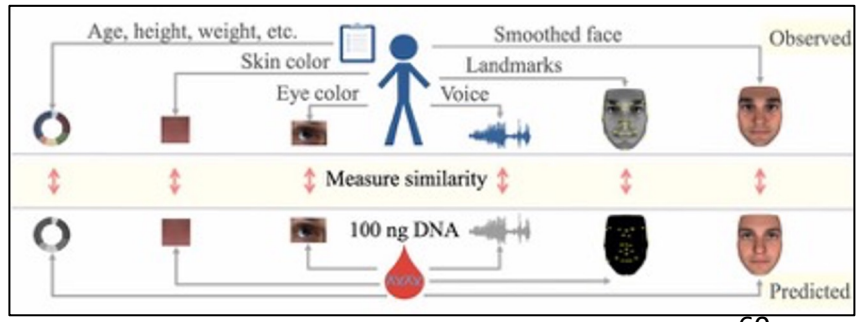
TechCrunch
January 31, 2019

De-identification of genomic data is impossible (?)

- ***Lin et al. 2004 Science***: 75 or more SNPs (Single Nucleotide Polymorphisms) are sufficient to identify a single person
- ***Homer et al. 2008 PLOS Genetics:*** aggregated genomic data (i.e., allele frequencies) can be used for re-identifying an individual in a case group with a certain disease
- ***Gymrek et al. 2013 Science:*** surnames can be recovered from personal genomes, linking "anonymous" genomes and public genetic genealogy databases
- ***Lipper et al. 2017 PNAS:*** Anonymous genomes can also be identified by inferring physical traits and demographic information
- ***Many more to come…***



ysearch

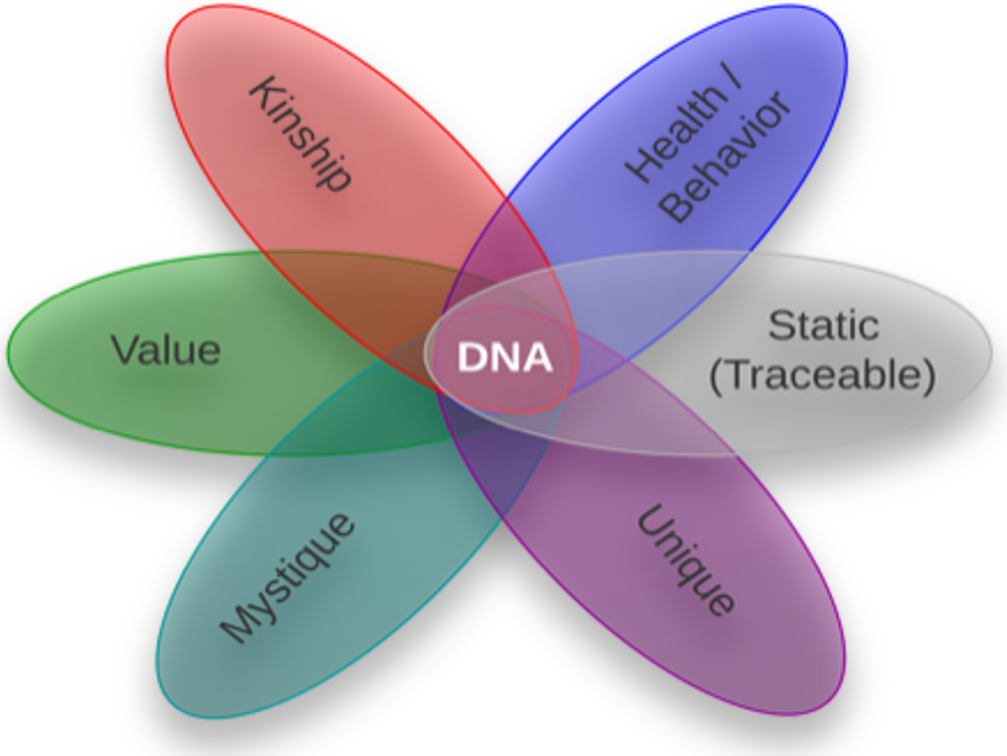


Standard de-identification and anonymization techniques (data suppression, swapping, obfuscation) are ineffective with genomic data

What If Genomic Data Are Leaked?

Genomic data pose special privacy problems:

- They are inherently identifying
- They can't be changed (as opposed to passwords)
- They have unique statistical regularities
- They contain sensitive and personal information (genetic diseases or propensity to develop certain conditions)
- Their leakage can expose individuals to genetic discrimination
- Relatives can also be affected

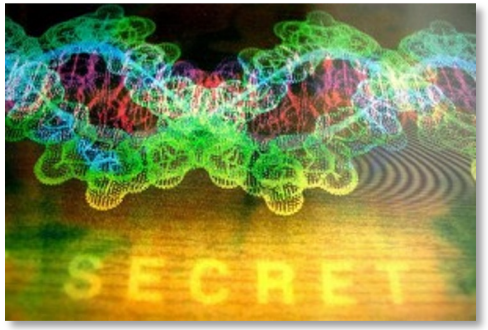


[Naveed et al.'15]

Protecting genomic privacy

Genome Privacy and Security: a Grand Challenge for Mankind

- Required **duration** of protection >> **1 century**
- (Current) **data size**: around **300 GBytes** / person
- Need sometimes to carry out computations on **millions** (if not more) of patient records
- **Noisy** data
- **Correlations**
  - within a single genome (“linkage disequilibrium”)
  - across genomes (kinship, ethnicity)



- **Several “semi-trusted” stakeholders**: sequencing facilities (including Direct-to-Consumer companies), hospitals, genetic analysis labs, private doctors,…
- **Diversity of applications** (and thus of requirements): healthcare, medical research, forensics, ancestry

Canonical Misconception about Genome Privacy and Security

Genome privacy is hopeless, because all of us leave biological cells (hair, skin, droplets of saliva,…) wherever we go

- Those cells can be collected and used for DNA sequencing
- Hence trying to secure genomes is a lost battle
- What is wrong with this reasoning?
- Collecting human biological samples and sequencing them is expensive, illegal, prone to mistakes, and non-scalable! (even if sequencing techniques keep improving)
- The medical community (research and healthcare) **should not be** the (indirect) accomplice of massive leaks of sensitive data

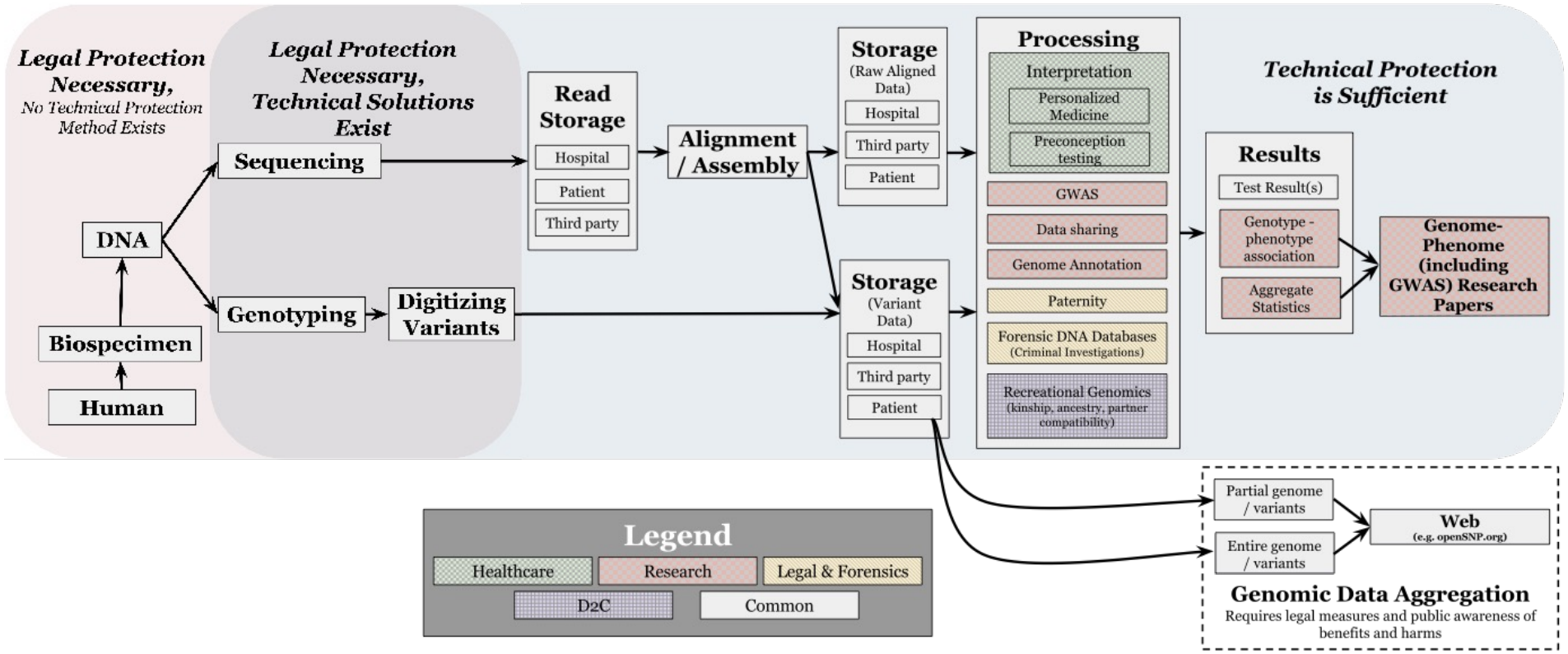
Security / Privacy Requirements for Personalized Health

- Pragmatic approach, **gradual** introduction of new protection tools
- Different **sensitivity levels** of the data
- Different **access rights**
- Exploit **existing** data (electronic health records) and tools
- Be **future-proof** (no short-sighted “bricolage”)
- Awareness and enforcement of **patient consent**

Key Protection Strategies

- Trusted parties
  - Legislation, regulation, and data usage agreements
  - Access control and encryption
- Aggregated or obfuscated release
  - Summary statistics
  - Differential privacy
- Semi-trusted parties using privacy-enhancing technologies (PETs)
  - Homomorphic encryption E(M*N) = E(M)*E(N)
  - Multi-Party Computation (MPC) garbled circuits
  - Functional Encryption

Privacy Protection Landscape



M. Naveed, E. Ayday, E.W. Clayton, J. Fellay, C. Gunter, J.-P. Hubaux, B. Malin, and X. Wang. Privacy in the genomic era. ACM Computing Surveys, 48, 1, Article 6, August 2015.

Technologies for Privacy and Security Protection

Traditional Encryption

- Protects data at rest and in transit
- Cannot protect computation

Homomorphic Encryption

- Protects computation in untrusted environments
- Limited versatility vs efficiency

Secure Multiparty Computation

- Protects computation in distributed environments
- High communication overhead

Trusted Execution Environments

- Protects computation with Hardware Trusted Element
- Requires trust in the manufacturer, vulnerable to side-channels

Differential Privacy

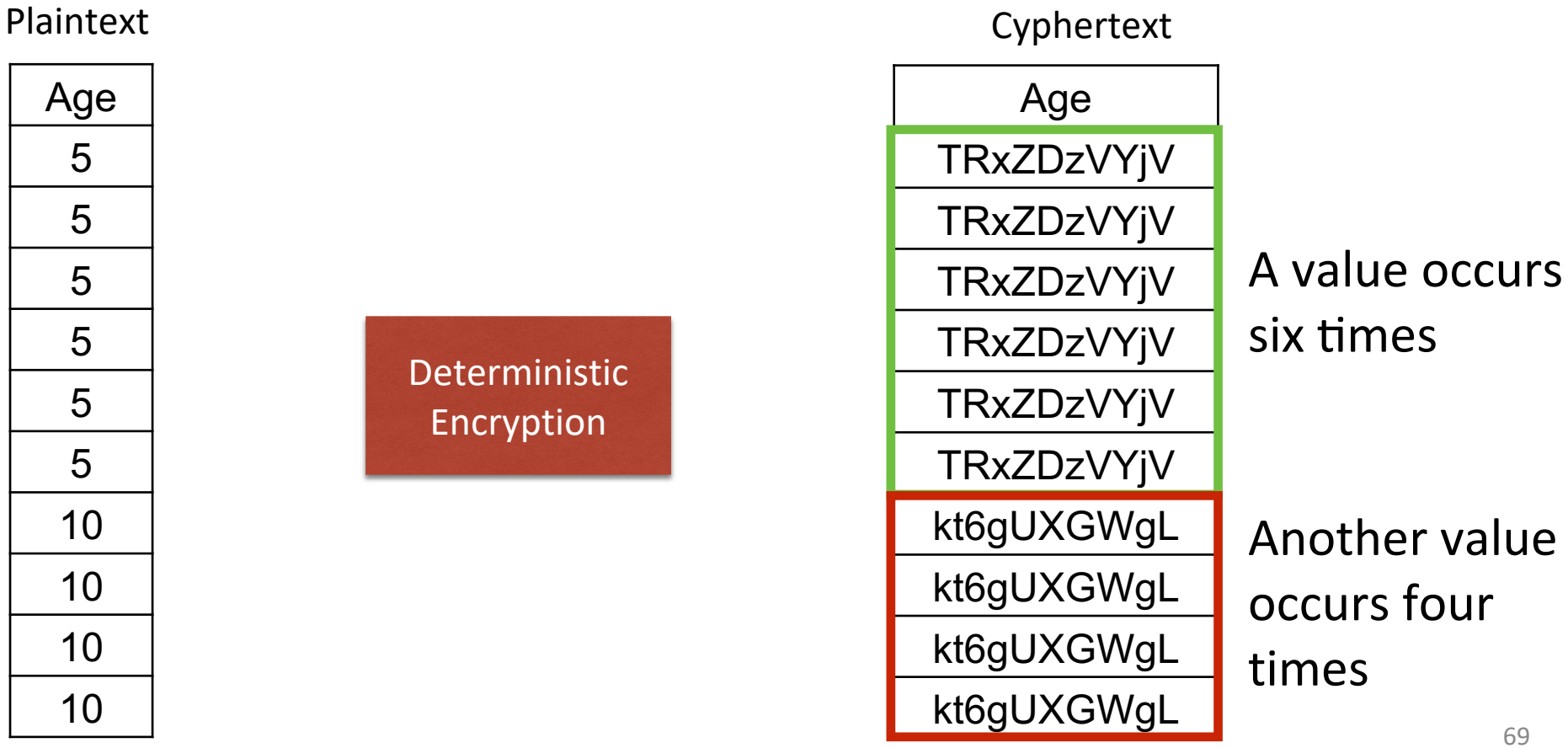
- Protects released data from inferences
- Degrades data utility (privacy-utility tradeoff)

Distributed Ledger Technologies (Blockchains)

- Strong accountability and traceability in distributed environments
- Usually no data privacy

Deterministic vs probabilistic encryption

- Deterministic encryption
  - Preserves and **leaks** equality of the the plaintext
  - More general property-preserving schemes can focus on other properties (e.g., order, format,…)



Deterministic vs probabilistic encryption

- Probabilistic encryption
  - Random salt added to each encryption to achieve semantic security (plaintext properties are not disclosed, and equality of plaintexts cannot be determined with non-negligible advantage)
  - Problem: ciphertexts cannot be compared

Plaintext

Age
5
5
5
5
5
5
10
10
10
10

Probabilistic Encryption

Ciphertext

Age
LKGM8EUnGd
kt6gUXGWgL
TRxZDzVYjV
IgDwwF64cl
ht5UIk8Evw
kbfRQ2nRAy
R8cBg6KRrw
en0yWX5iWA
mcq3uYOAQA
EE3sF0XfTn

Multi-site Studies – Where to Store the Data?

a. Keep them at each site

- Useful especially if the cloud is untrusted
- Better control of the data

b. In the cloud

- Take advantage of the well-known strengths of the cloud

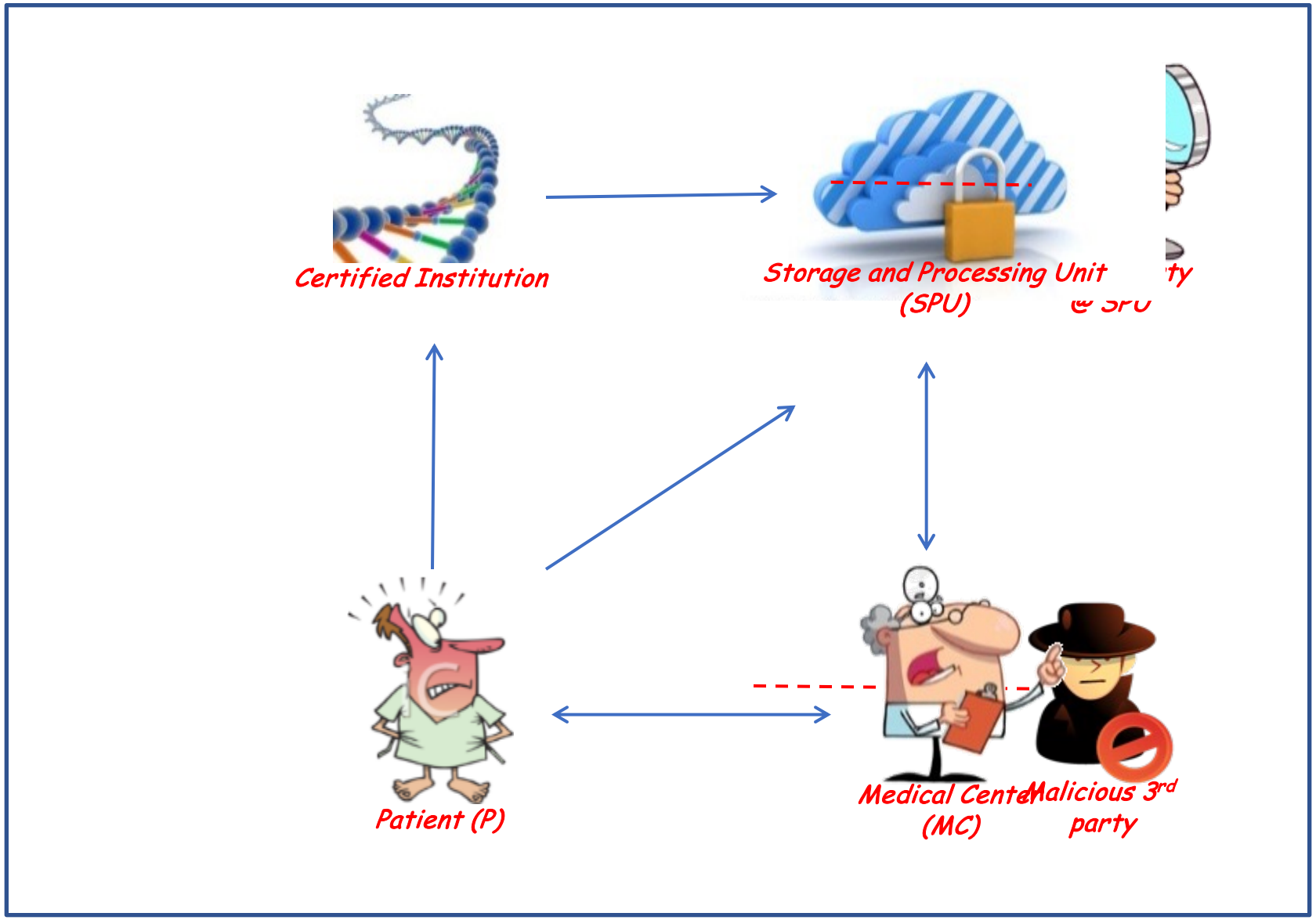
Secure Sharing of Health Data

Rationale

- Lesson of COVID-19: health data management needs vigorous improvement
  - Insufficient, ineffective data sharing
  - Poor data quality
  - Time-consuming ethical approval procedures for research protocols
- With appropriate **tools** and **processes**, it can and should be fixed
- These tools should be based on **mathematically-proven** protection techniques
- This will give data management the **robustness** and **efficiency** it deserves

Privacy-preserving processing using homomorphic encryption

Privacy-Preserving Personalized Medicine

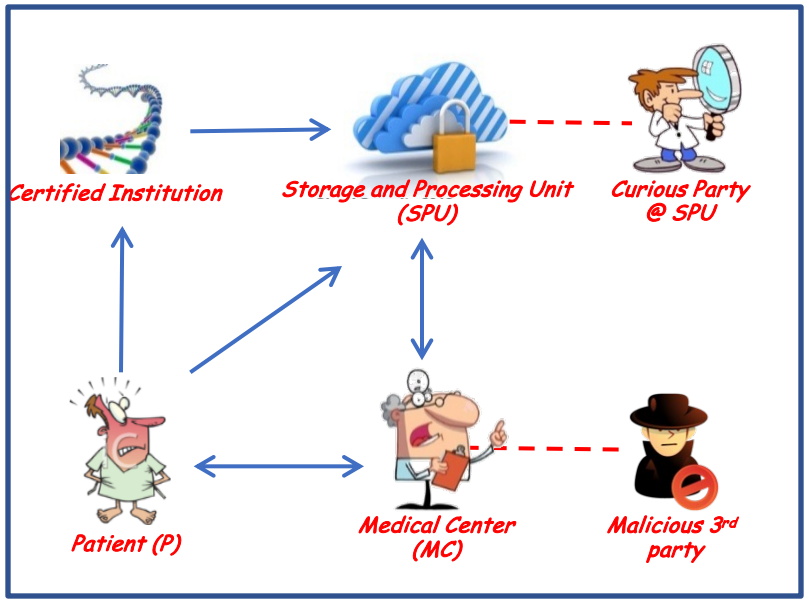


Setting and Goals

- Setting: A medical center (MC) want to conduct a *genetic disease susceptibility test* on a patient (P)
- Protect the privacy of users' genomic data
  - Protect data, including from insiders (e.g., curious sysadmins)
- Protect the privacy of medical center's confidential data
- Allow specialists to access only to the genomic data they need (or they are authorized for)
- Keep the access time to a single patient's genomic data to a few seconds

Threat Model

- The certified institution (CI) is a trusted entity.
  - Indispensable to do the sequencing
- An attacker at the MC
  - A careless or disgruntled employee at the MC or a hacker who breaks into the MC
  - Aims to obtain private genomic information about a patient (for which it is not authorized)
- A curious party at the SPU
  - Existence of a curious party or a disgruntled employee at the SPU
- No collusion between the MC and the SPU
- Access control based on patient's consent



Cryptographic Tools

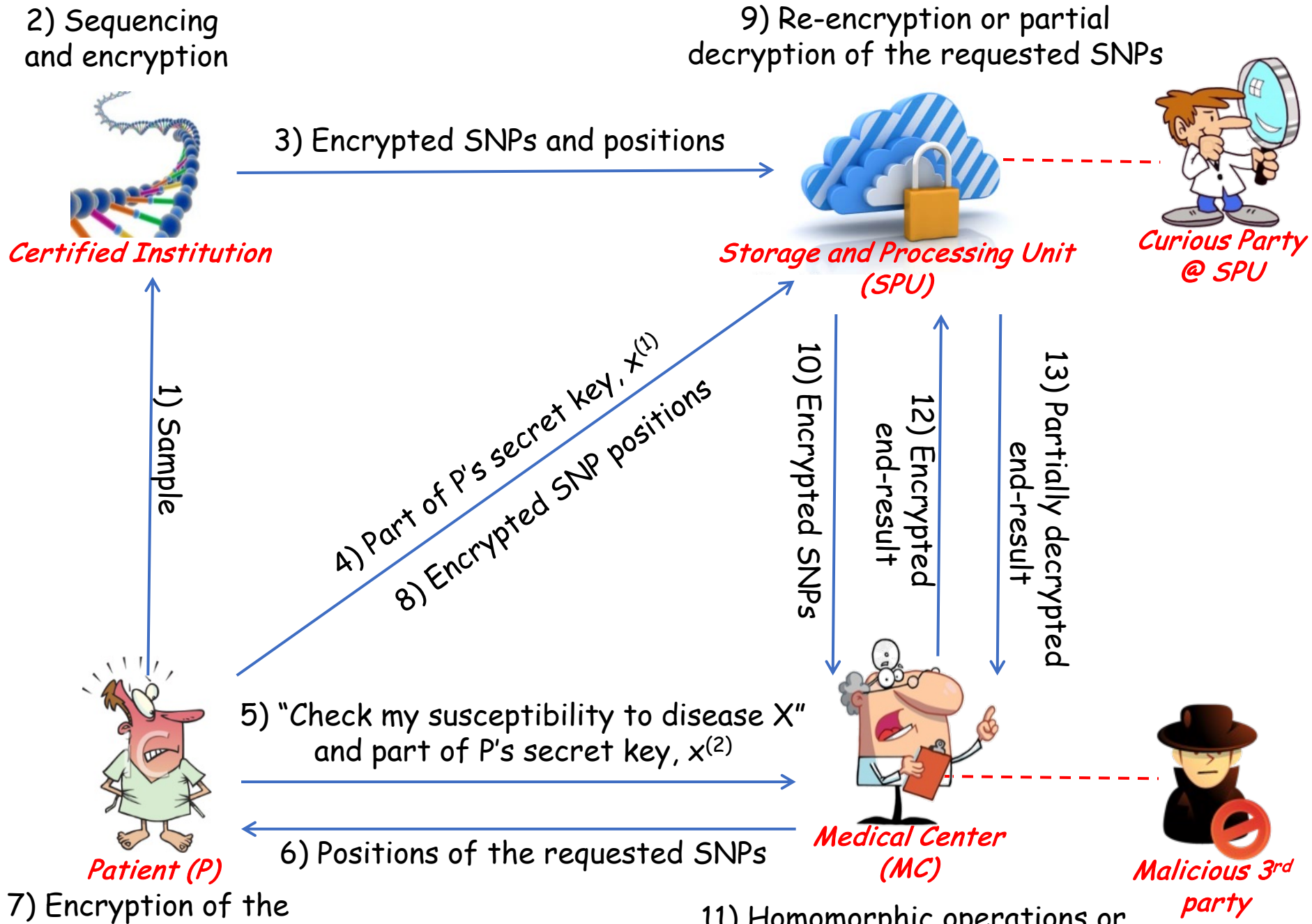
- Modified Paillier Cryptosystem
  - Bresson et. al 2003.
  - Homomorphic addition

$$D(E(m_1, r_1, g^{x_p}) \cdot E(m_2, r_2, g^{x_p})) = D(T_1^1 \cdot T_1^2, T_2^1 \cdot T_2^2 \mod n^2) = m_1 + m_2 \mod n$$

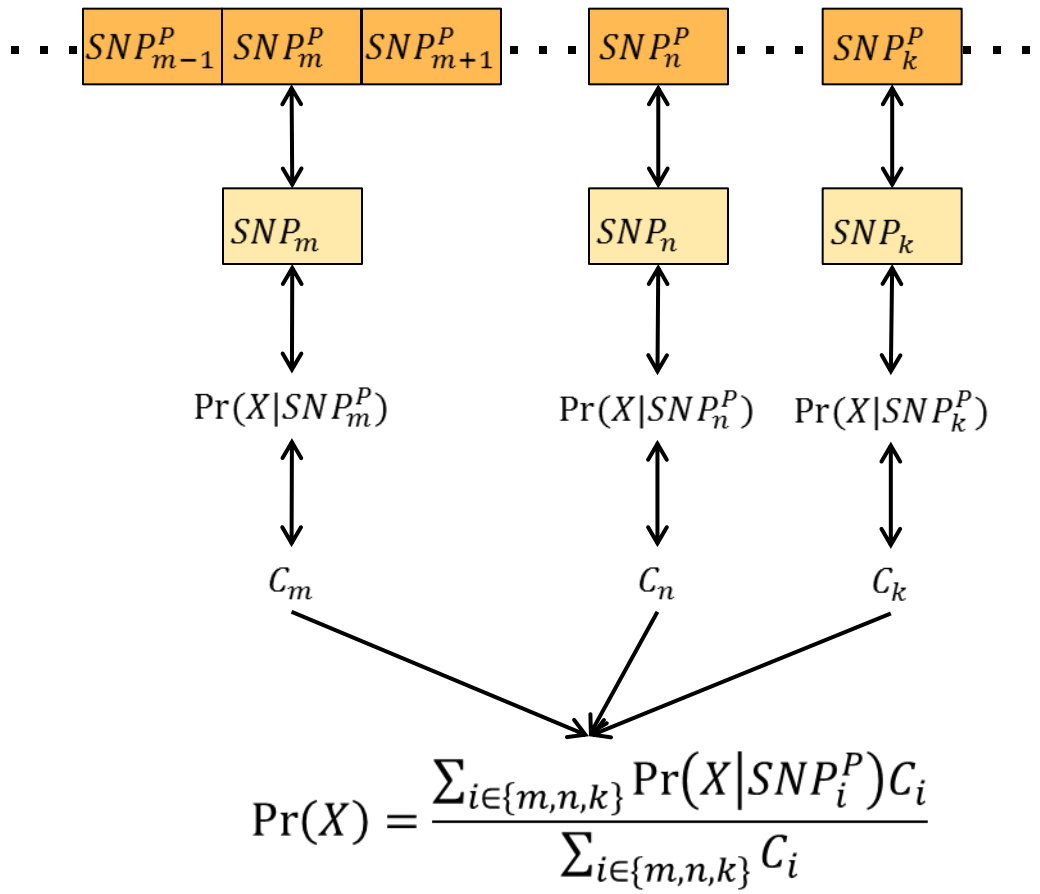
  - Multiplication with a constant

$$D(E(m_1, r_1, g^{x_p})^k) = D((T_1^1)^k, (T_2^1)^k \mod n^2) = km_1 \mod n.$$

  - Proxy re-encryption
    - Divide the weak secret into two shares
    - Distribute the shares to two parties
- Secure multiparty computation (SMC)


2) Sequencing and encryption
9) Re-encryption or partial decryption of the requested SNPs
3) Encrypted SNPs and positions
Certified Institution
Storage and Processing Unit (SPU)
Curious Party @ SPU
1) Sample
4) Part of P's secret key, $x^{(1)}$
8) Encrypted SNP positions
10) Encrypted SNPs
12) Encrypted end-result
13) Partially decrypted end-result
5) "Check my susceptibility to disease X" and part of P's secret key, $x^{(2)}$
6) Positions of the requested SNPs
Patient (P)
Medical Center (MC)
Malicious 3rd party
7) Encryption of the requested positions
11) Homomorphic operations or recovery of relevant SNPs

Computing Disease Susceptibility



- All operations are conducted in ciphertext using homomorphic encryption

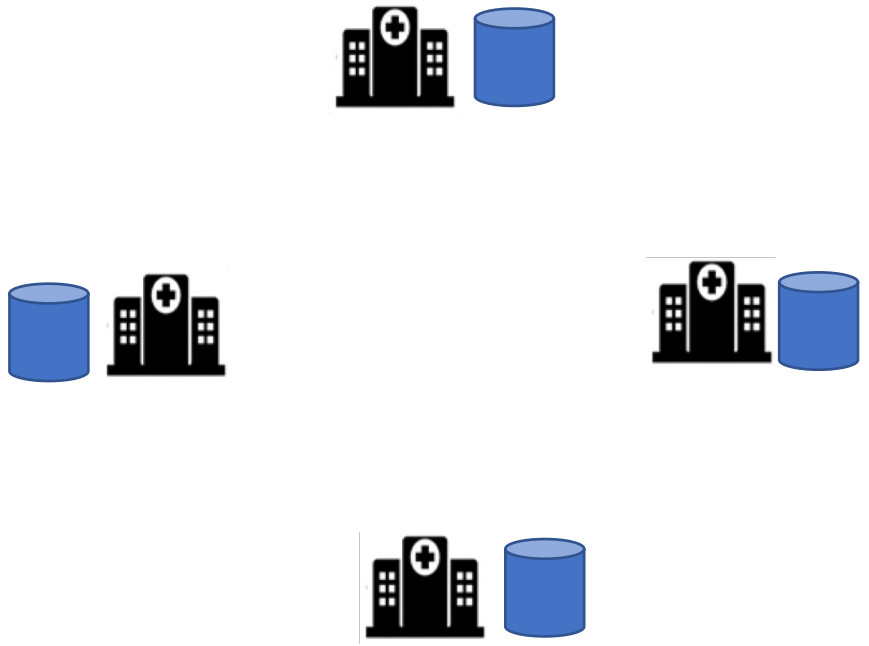
Remarks

- Patient-related steps can be handled via the patient's smart card or mobile device
- Individual contributions of the genetic variant markers remain secret at the MC
  - Homomorphic operations are conducted at the MC
- Solution is possible without the proxy re-encryption by letting the patient decrypt the end-result
  - Secret key of the patient remains only at the patient
- **Does this solve everything?**

Recap: the multi-site studies

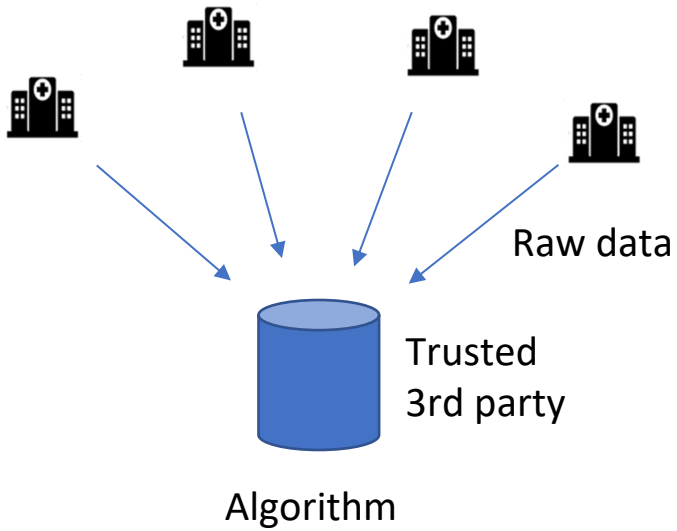
Multi-site studies

- Benefit: increase the amount of available data and thus the statistical significance of findings
- Challenges
  - Need to interface **several** ethics committees and IT services
  - **Diverse** legislations (especially if international)
  - **Heterogeneity** of data semantics and of data quality (partially due to reluctance to data transfer)
  - **Reluctance** to share data (control, publication advantage)
- Widespread practice: de-identification of the data
- Usually:
  - Research usage of de-identified data patient consent needed
  - Not needed for anonymous data
- Overall, very little awareness of cryptography and differential privacy in biomed



Multi-site studies – Current approaches

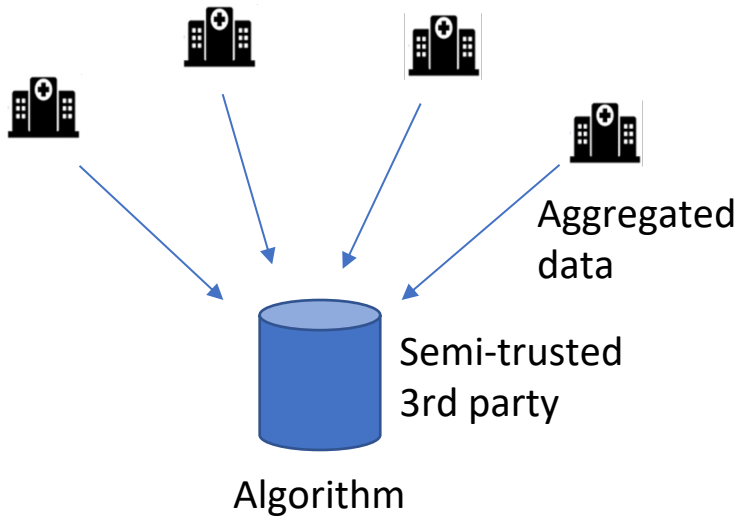
(a) Fully centralized



- Transfer raw, de-identified data to a central database
- Do all computations there
- Data protection: security of the central database
- Individual sites lose control over their data

All of Us
EGA
Genomics England

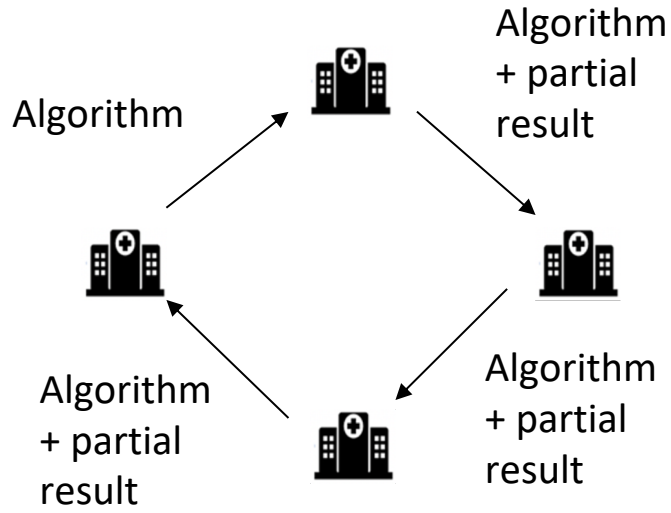
(b) Meta-analysis



- For each study, aggregated data provided by each site
- Still need to trust the central server

https://covidclinical.net/

(c) Decentralized

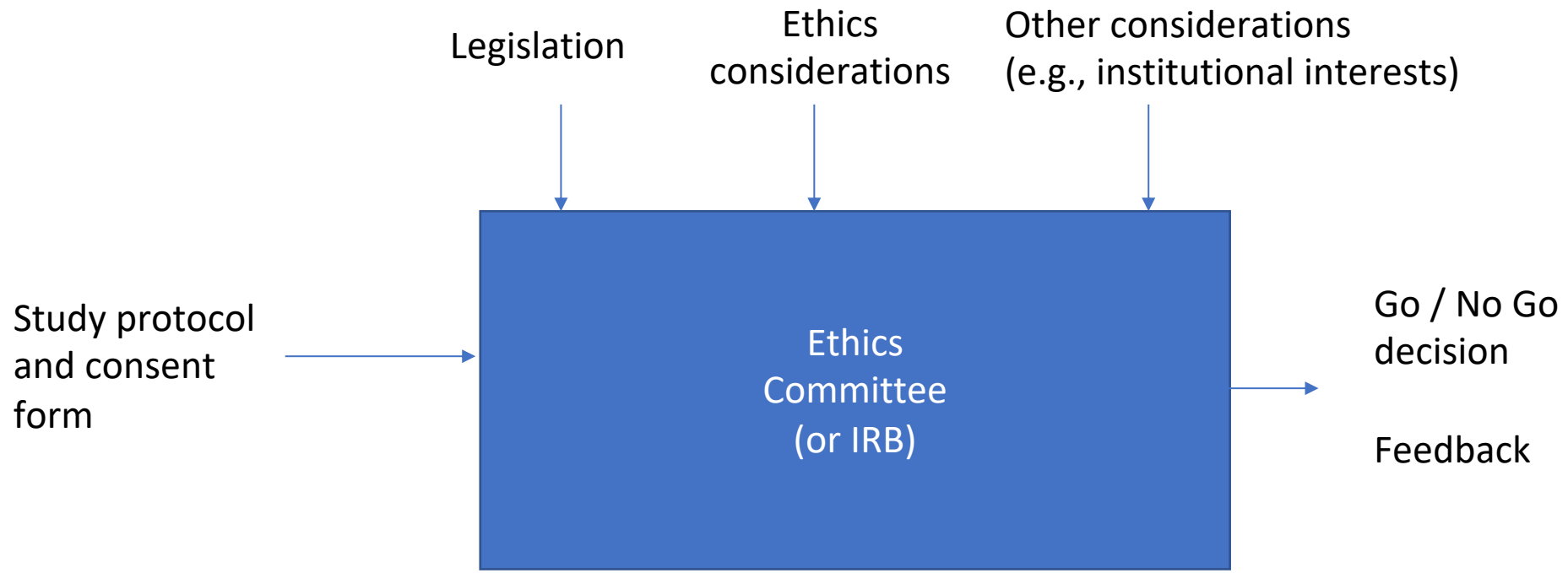


- "Send the algorithm to the data"

http://www.datashield.ac.uk

Personalized Health Train (PHT)

Ethics Committees



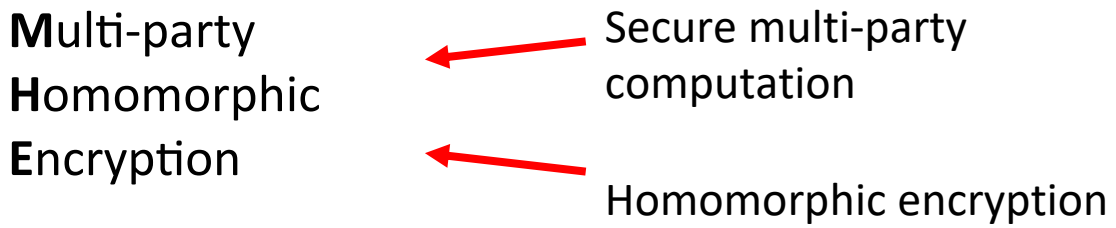
- Very slow, manpower-hungry and tedious process to check the proposed data-protection measures
- Need to obtain informed consent; diversity of consent forms
- Ethics committees make an on-paper *a priori* evaluation, with little control on what happens afterwards
- Risk of "race to the bottom": the researchers that obtain permissions to see more data will extract more value → competitive advantage

Technology Disruption: Multi-site Studies on Health Data

Current situation

- Questionable data protection guarantees
- Reliance on the manual work of ethics committees
- Slow, ad hoc procedures

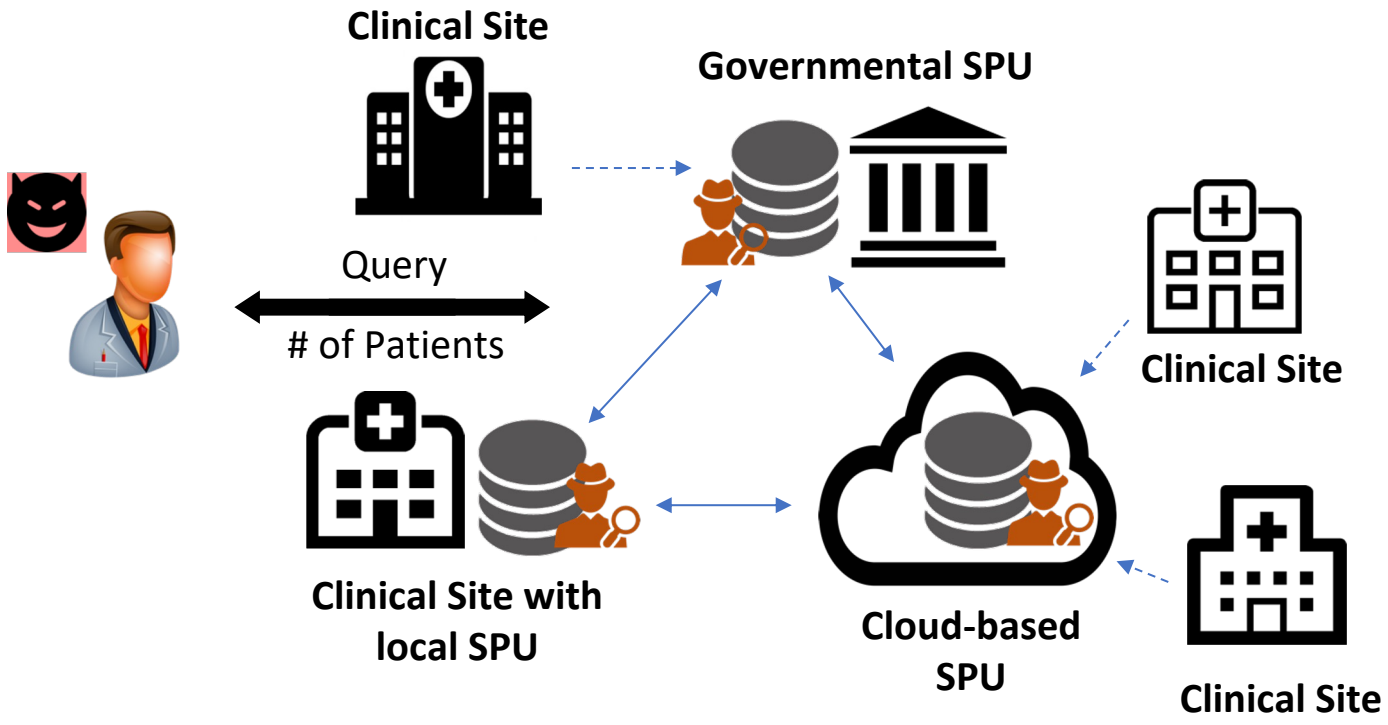
The future: use the right crypto



- Streamlining the approval processes
- Mathematically-proven data protection guarantees
- Really decentralized trust → federated learning

Amalio Telenti- Commoditization Of Medical Data: Advanced Solutions For Modern Risks. Forbes Council Post, 25 March 2020
https://www.forbes.com/sites/forbestechcouncil/2020/03/25/commoditization-of-medical-data-advanced-solutions-for-modern-risks/#5a61741f3286

System and threat models



SPU: Storage and Processing Unit

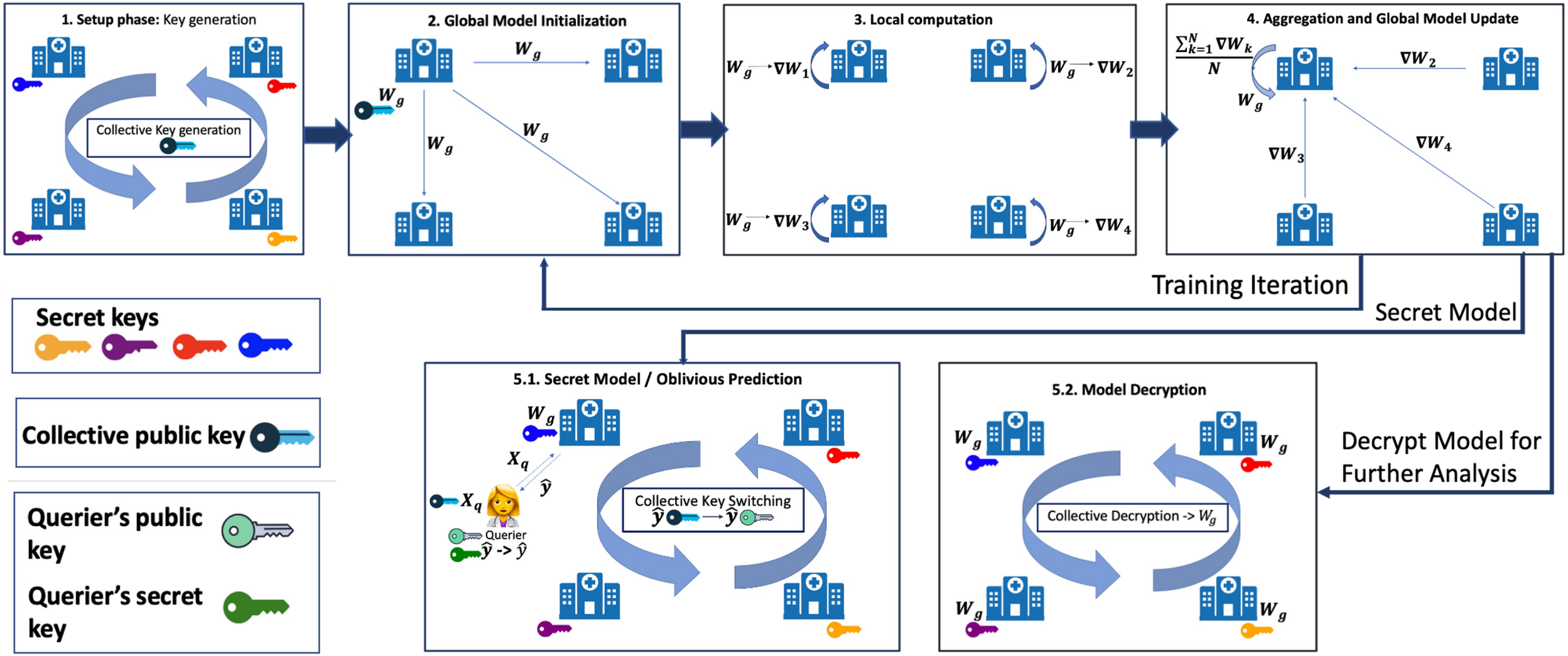
Honest-but-curious adversary:

- honestly follows the protocol
- tries to infer sensitive data from the different steps of the protocol

Malicious-but-covert adversary:

- can tamper with the protocol
- tries to infer sensitive data from the query end-result

Solution Overview



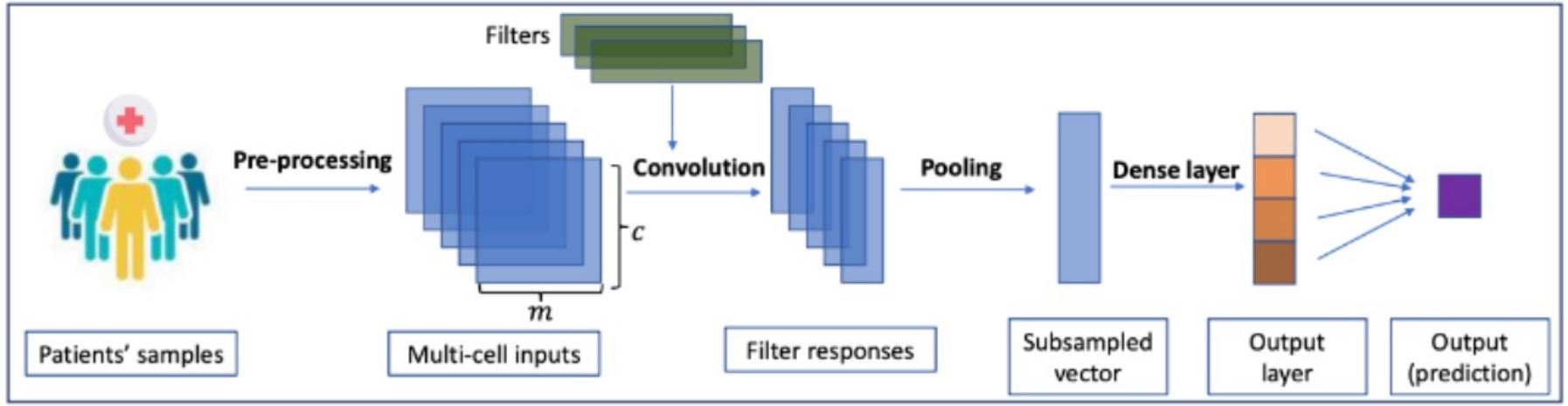
[1] POSEIDON: Privacy-Preserving Federated Neural Network Learning, S.Sav, A. Pyrgelis, J.R. Troncoso-Pastoriza, D. Froelicher, J.P. Bossuat, J.S. Sousa and JP Hubaux, NDSS 2021

Privacy-Preserving Federated Neural Network Learning for Disease-Associated Cell Classification

PriCell

Medical Application: To evaluate this system within the framework of single-cell analysis -> train a convolutional neural network (CellCnn[1]) for the disease classification task.

- In a **privacy-preserving** and **federated** setting
- **Privacy-Preserving Federated Neural Network Learning for Disease-Associated Cell Classification**, S. Sav, J.-P. Bossuat, J. R. Troncoso-Pastoriza, M. Claassen and J.-P. Hubaux., Patterns, 3(5), 2022.



[1] Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via representation learning. Nature Communications, 2017.

Let’s exercise our privacy brain

From the PETs we discussed so far, what kind of technologies can be employed to enable privacy in this field?

- How does the threat model is affected by the solution?
- What is the system model?
- What are the limitations?

Events on Genome Privacy and Security

- **Dagstuhl** seminars on genome privacy and security 2013, 2015
- **Conference on Genome and Patient Privacy (GaPP)**
  - March 2016, Stanford School of Medicine
- **GenoPri**: International Workshop on Genome Privacy and Security

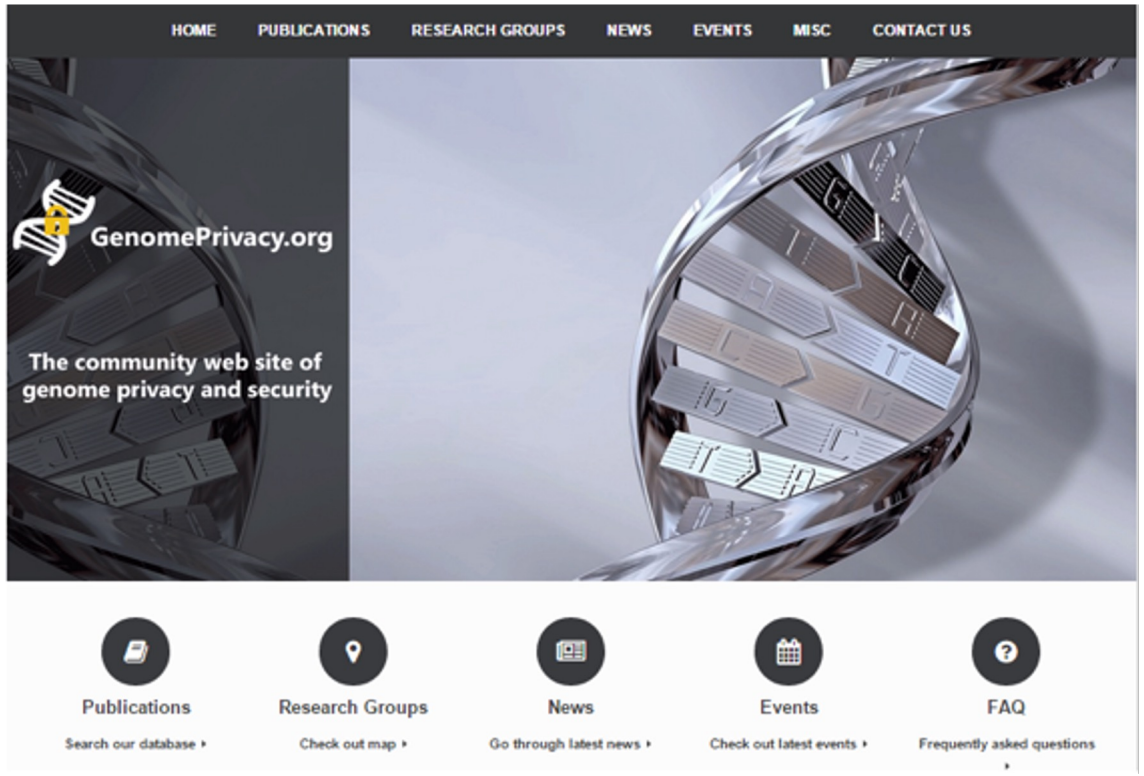
- **iDash**: integrating Data for Analysis, Anonymization and sHar (already in previous years)

- Inst. For Pure and Applied Mathematics (IPAM, UCLA)

Algorithmic Challenges in Protecting Privacy for Biomed Data

- Lots of material online

“genomeprivacy.org”



Community website

- Searchable list of publications on genome privacy and security
- News from major media (from Science, Nature, GenomeWeb, etc.)
- Research groups and companies involved
- Tutorial and tools
- Events (past & future)

Overall Conclusion

- Protecting health data is one of the most formidable challenges for cybersecurity
- With the advent of personalized health and thus genomics:
  - risk is increasing
  - conventional medical data protection techniques based on de-identification do not work anymore
- We have presented technical solutions to address the problem
- They are based on SMC, homomorphic encryption, blockchains and differential privacy
- Legislation will also play an important role