COM 402

Security and Privacy

Data Protection for Personalized Health

Prof. Jean-Pierre Hubaux
Head of LCA1 Lab
Academic Director of the Center for Digital Trust
School of Computer and Communication Sciences
EPFL

With gratitude to the biomedical and CS researchers I have the privilege to work with

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

Source: The New Yorker

AARGAUER ZEITUNG

ABO E-PAPER WEI

START AARGAU SCHWEIZ AUSLAND WIRTSCHAFT SPORT LEBEN

Jetzt aktuell: Vierfachmord Rupperswil • Lawine Crans-Montana • Karl Lagerfeld

GESUNDHEIT

Rückenwind für den E-Patienten: Parlament will obligatorische elektronische Patientendossiers

von Michel Burtscher - az Aargauer Zeitung • 18.2.2019 um 04:15 Uhr

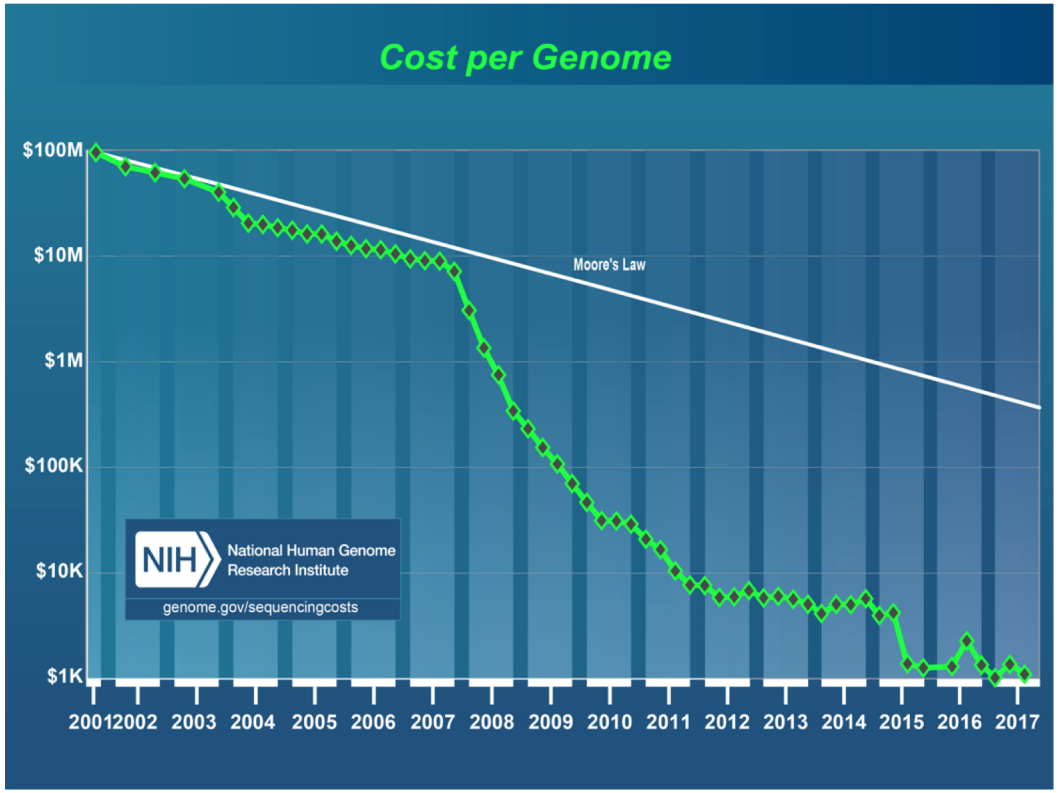
Der gläserne Patient: Nun nimmt die Politik neuen Anlauf für elektronische Gesundheitsdossiers.
Zur Verfügung gestellt

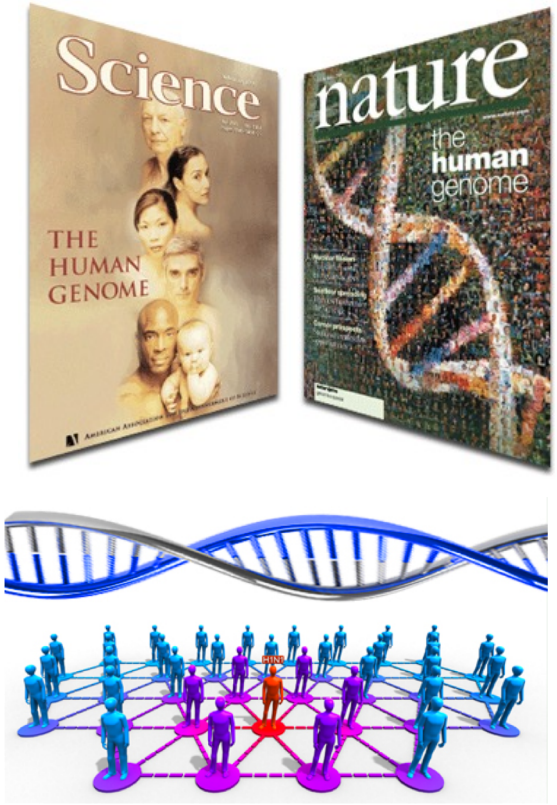
Damit sich das elektronische Patientendossier etablieren kann, müssen möglichst viele Ärzte, Spitäler und Kliniken mitmachen. Doch für Praxisärzte ist die Teilnahme bisher freiwillig. Das könnte sich bald ändern – zumindest für einen Teil von ihnen.

- Electronic Health Records are mandated by law in Switzerland
- Mandatory for hospitals as of 2020
- 2 systems will be deployed (by Swisscom and Swiss Post)

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

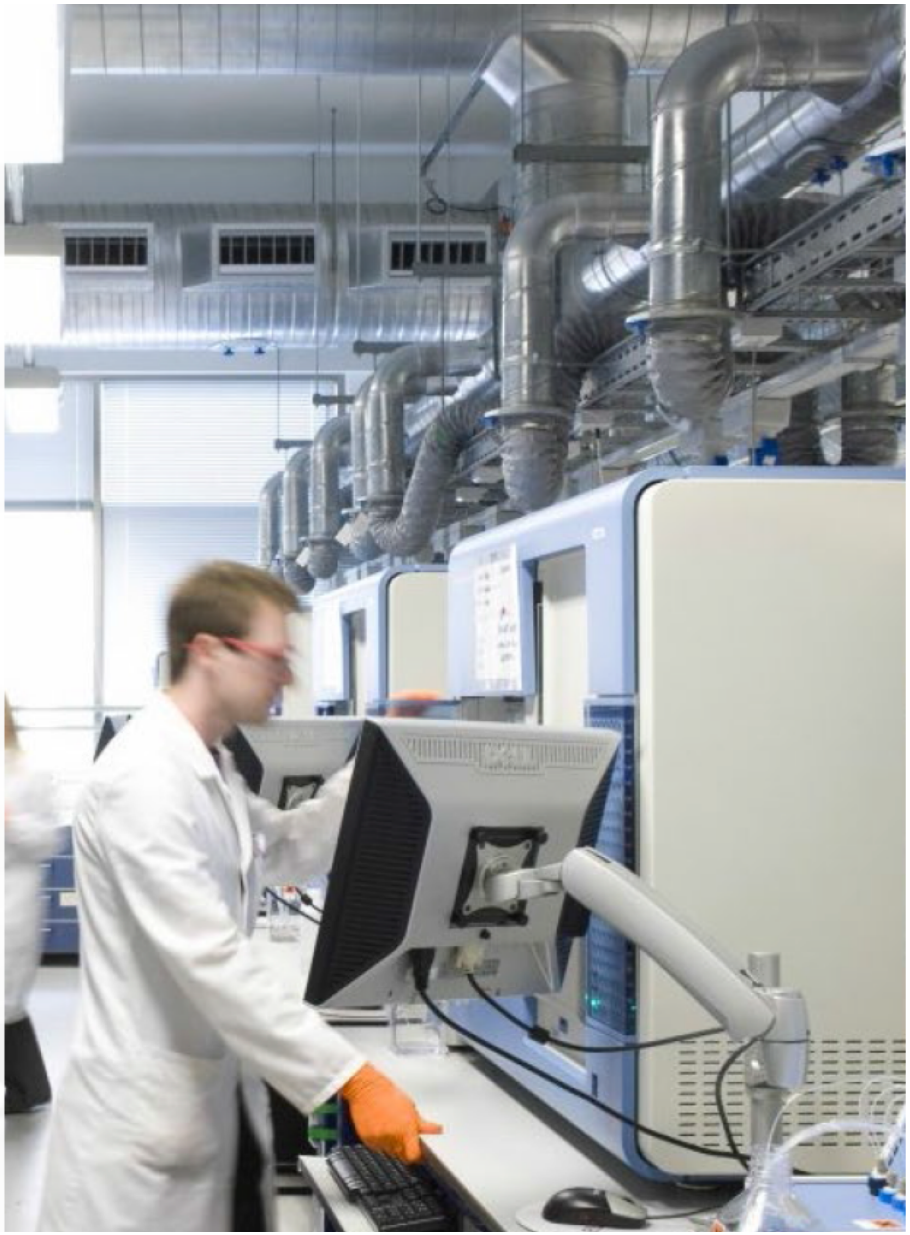
- Genetic disease risk tests help early diagnosis of serious diseases
- Pharmacogenomics ➔ personalized medicine

- 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

NEXT GENERATION SEQUENCING

TECHNOLOGY ABOUT TO PROVIDE AFFORDABLE (IN TIME AND COST) AND PRECISE DETERMINATIONS OF GENOME WIDE:

- DNA sequence/variations (DNA-seq)
- Gene subregions' activity (RNA-seq)
- Protein-DNA interaction regions (ChIP-seq)



The Genomic Era

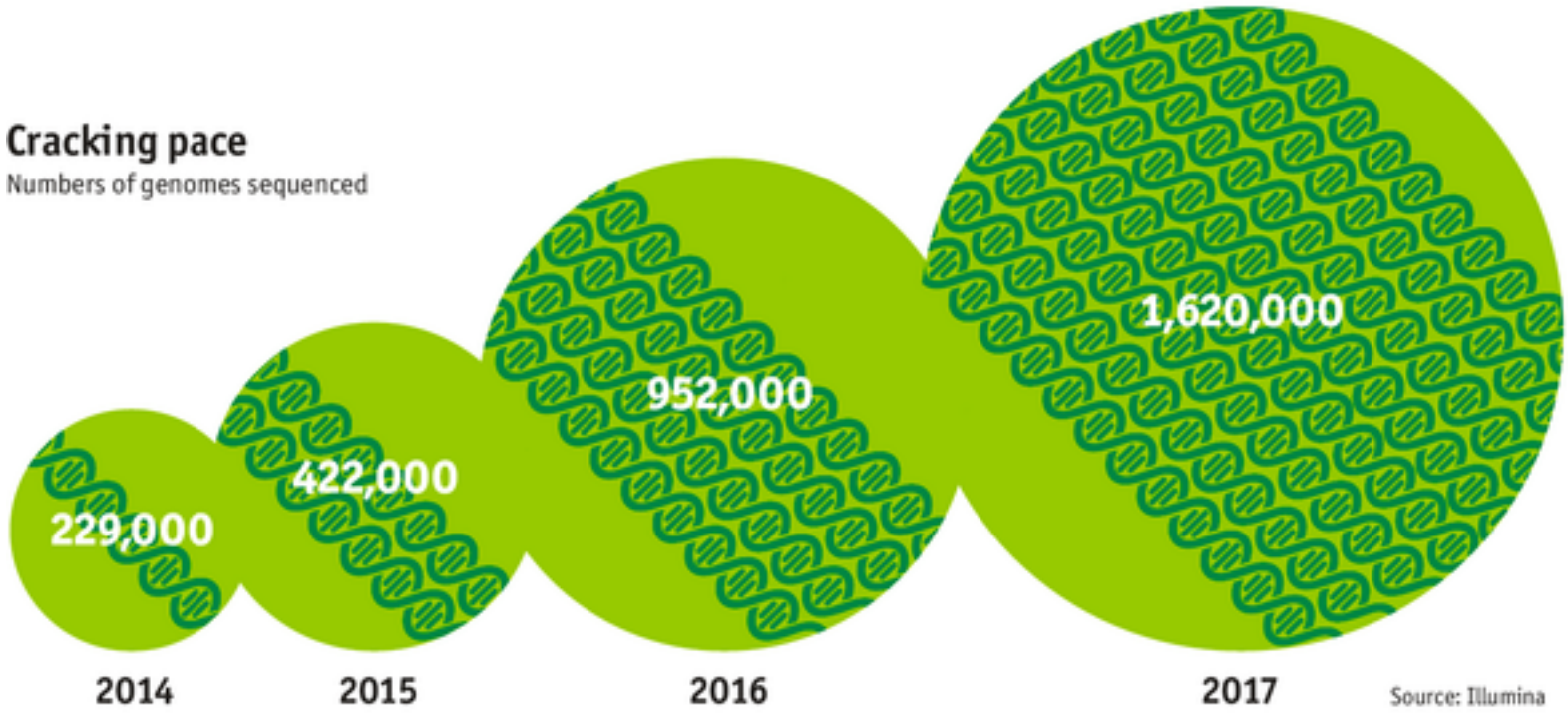
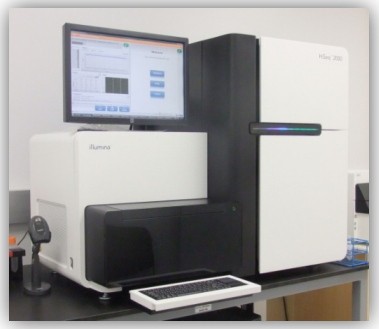


Figure from The Economist

From Blood Sample to Genome Analysis

Samples



Sequencing machine

Raw data (short reads) FastQ files

Alignment

SAM/BAM file (aligned reads)

Variant call

VCF Files

Delta with respect to the *reference genome*

ID chromosome mapping quality
Alignment Flag position Cigar string

```
GZQG54D02DI67S 16  10  72065884  228  3S23M1I255M8S  *  0  0
TAGCCCCCACCCAACACACACACACGTCTTTTTTTCTGAATCATTTGAGTGTACCATACATAC
GTCATGGCCCTCAGTGTATATTTCCTAGGAATTGGGACATTCTCTTACATAACCACAGTGCAGT
TAAGGGTTCCAGGTATTCACATTGATAGAATGTGTATCTAACCCATGGCCCACAGTTCAGTGT
GGTCAGGTGCCCTAACATCCTGTACAGCACTTCCCCCTCTAGTACCAGAGCCAGTCTGGAAT
CAAGGATTGCATTTACTTCTAACTCTTTGGCCCCTCTG      E=7/////200088;@;;;@EEIID?,?
2666644IHHHHIIH<<<HIIIIIIIIIIIIIIIIII;;555CEIEEIIII???
HHIIIIIIIHHHIIIIIIIIIIIHDCCIIIIIIIIIIIIIIIHHHCCCCIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHHHIIIIHHHIIIDCCIIIIIIIIIIIIIIHH
HIIIIIIIIIIIIIII>CC;;;;;;HIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHHHIIIIIHHHII;;;IIAAAAIIII      AS:i:267      XS:i:0  XF:i:0
XE:i:6  XN:i:0
```

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

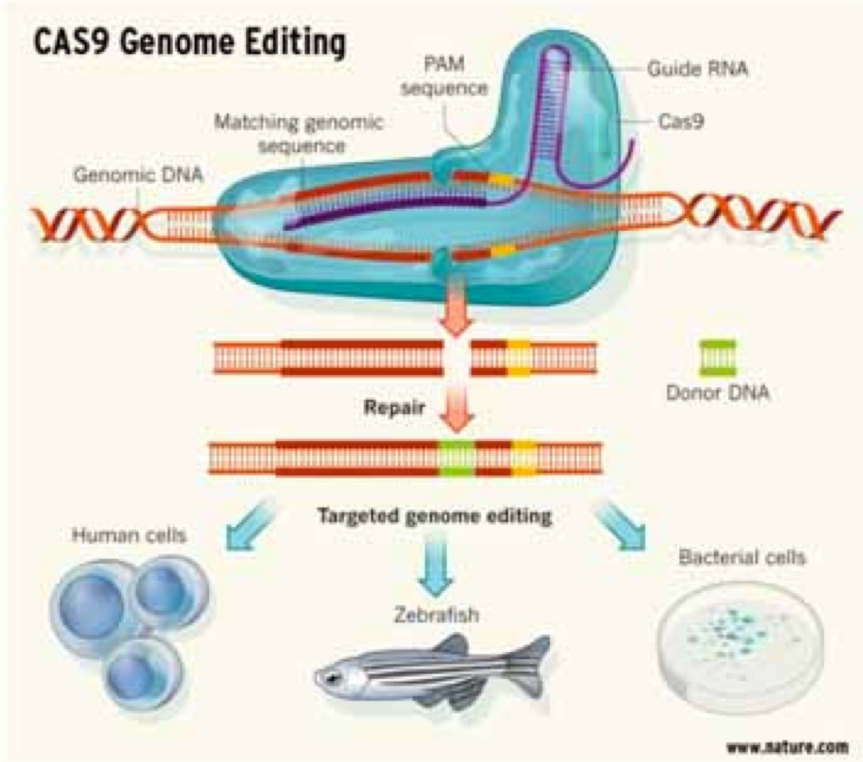
- Sequencing errors
- Possible mutations


IEEE
SPECTRUM
FOR THE TECHNOLOGY INSIDER 08.15
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. 46
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

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

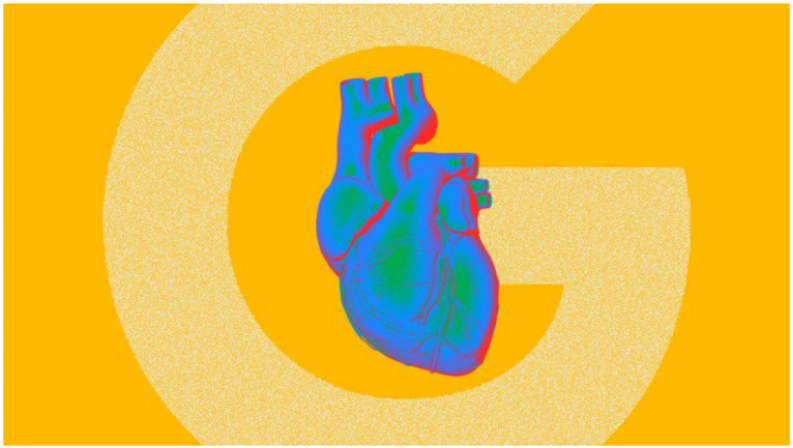
- Deletion of gene CCR5 was intended to make babies HIV-resistant
- This deletion could increase cognition capabilities

CRISPR: Gene-editing technique

MIT Technology Review

Biotechnology May 6, 2019

Google backs a bid to use CRISPR to prevent heart disease

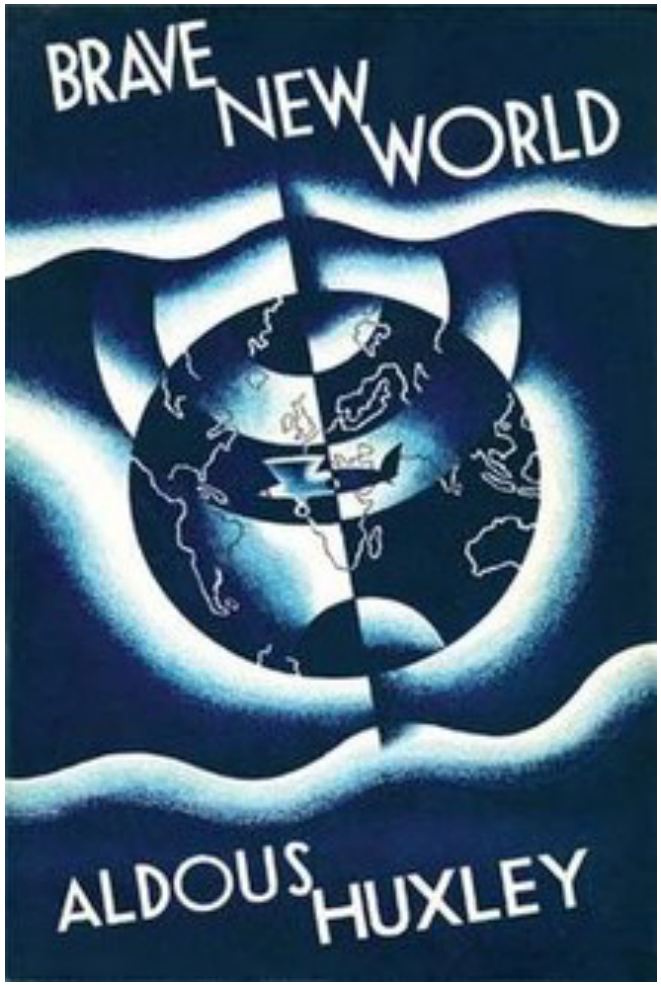


Ever wonder why some fortunate people eat chips, don't exercise, and still don't get clogged arteries? It could be because they've got lucky genes.

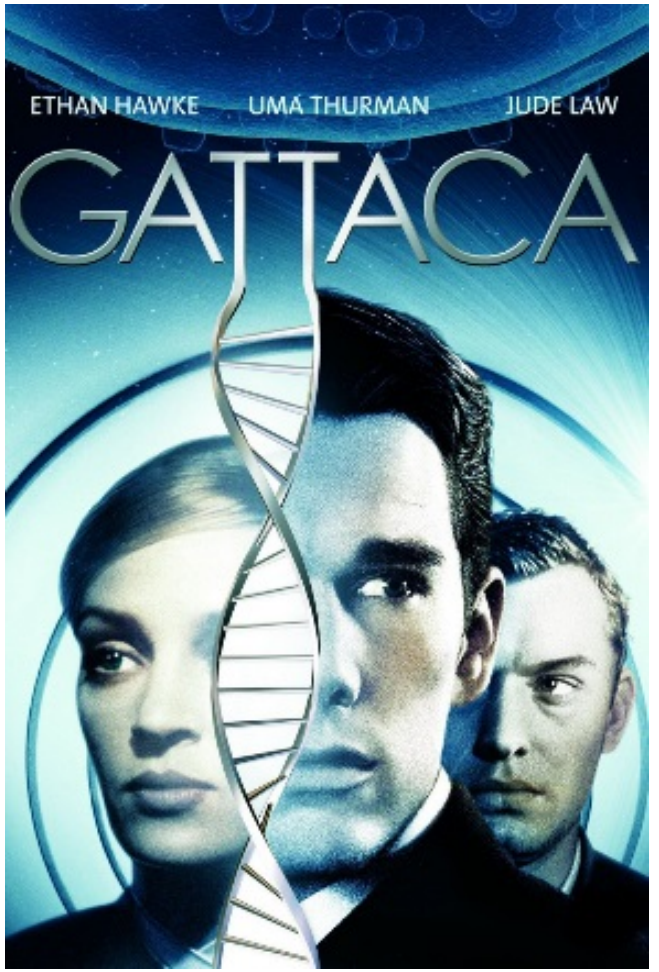
Now Alphabet (Google's parent company) is bankrolling a startup company that plans to use gene editing to spread fortunate DNA variations with "one-time" injections of the gene-editing tool CRISPR.

Heart doctors involved say the DNA-tweaking injections could "confer lifelong protection" against heart disease.

Big idea: The startup, Verve Therapeutics, said it had raised $58.5 million from investors including Alphabet's venture fund, GV. What makes Verve different? Most gene-therapy companies have gone after rare diseases like hemophilia. But Verve thinks editing people's DNA could instead help solve the most common cause of death.



Novel, 1931



Movie, 1997

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)

Ransomware Attack against German Hospitals

20. März 2016, 10:05 Uhr Klinikum Neuss

Wenn Cyberkriminelle ein Krankenhaus lahmlegen

Das Lukaskrankenhaus der Städtischen Kliniken in Neuss wurde Opfer von Cyberkriminellen. (Foto: dpa)

“WannaCry” Ransomware Virus (May 2017)

Do state institutions have the resources to fight hackers?

Public sector has lessons to learn as hospital trusts and GPs struggle to recover from ransomware attack

A ransomware attack bought computers to a standstill across the world on Friday. Photograph: Ritchie B. Tongo/EPA

The Guardian,
14 May 2017

US Healthcare Official “Wall of Shame”

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

Around 5 declared breaches per week, each affecting 500+ people

Under Investigation | Archive | Help for Consumers

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. The following breaches have been reported to the Secretary:

Cases Currently Under Investigation

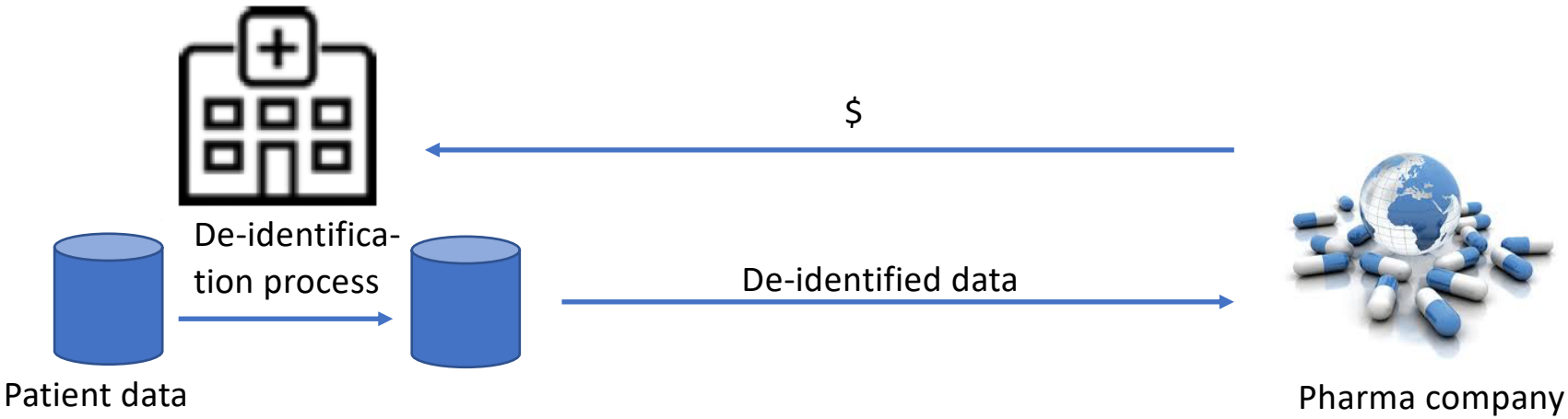
This page lists all breaches reported within the last 24 months that are currently under investigation by the Office for Civil Rights.

Show Advanced Options

Breach Report Results

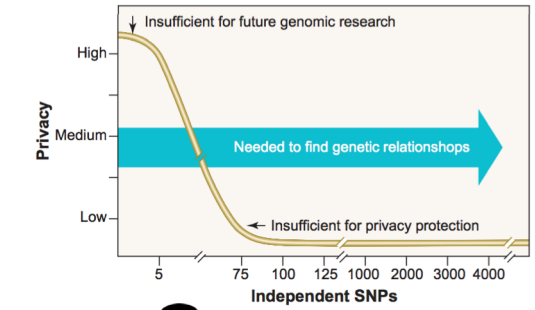
Expand All	Name of Covered Entity	State	Covered Entity Type	Individuals Affected	Breach Submission Date	Type of Breach	Location of Breached Information
	Ohio Living	OH	Healthcare Provider	6510	09/07/2018	Hacking/IT Incident	Email
	Rockdale Blackhawk, LLC d/b/a Little River Healthcare	TX	Healthcare Provider	1494	09/07/2018	Unauthorized Access/Disclosure	Electronic Medical Record, Other
	J.A. Stokes Ltd.	NV	Healthcare Provider	3200	09/05/2018	Hacking/IT Incident	Desktop Computer, Electronic Medical Record, Network Server
	Reliable Respiratory	MA	Healthcare Provider	21311	09/01/2018	Hacking/IT Incident	Email
	Port City Operating Company doing business as St. Joseph's Medical Center	CA	Healthcare Provider	4984	08/31/2018	Loss	Other Portable Electronic Device
	Carpenters Benefit Funds of Philadelphia	PA	Health Plan	20015	08/31/2018	Hacking/IT Incident	Email

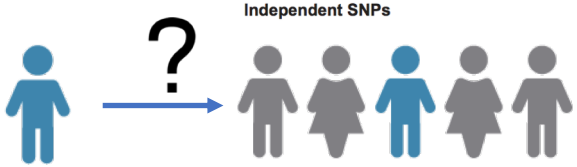
Hospital / pharma interface



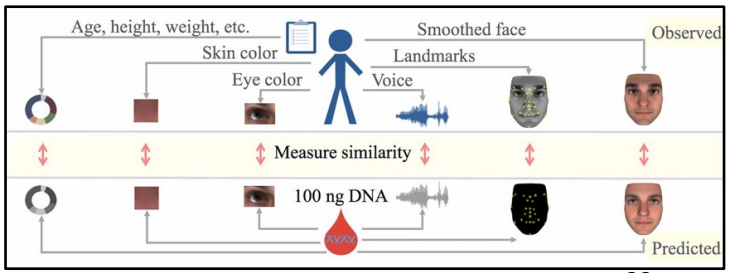
De-identification of genomic data is impossible

- ***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…***



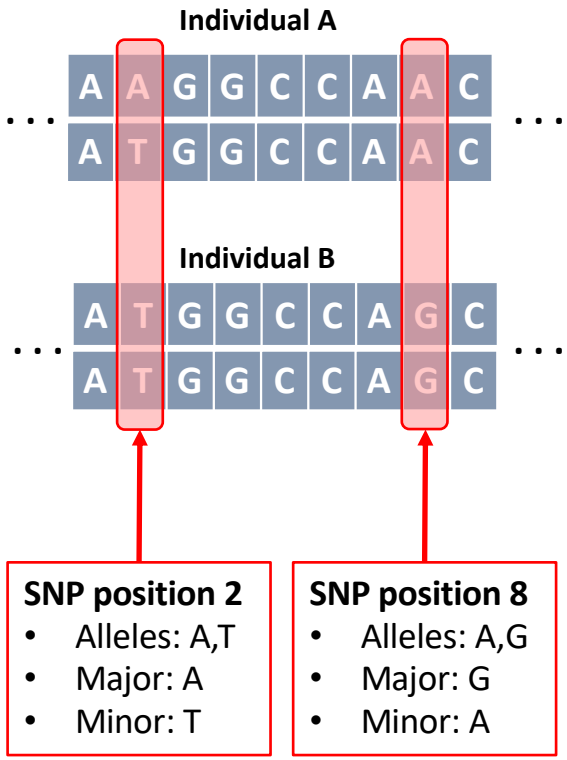


ysearch



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

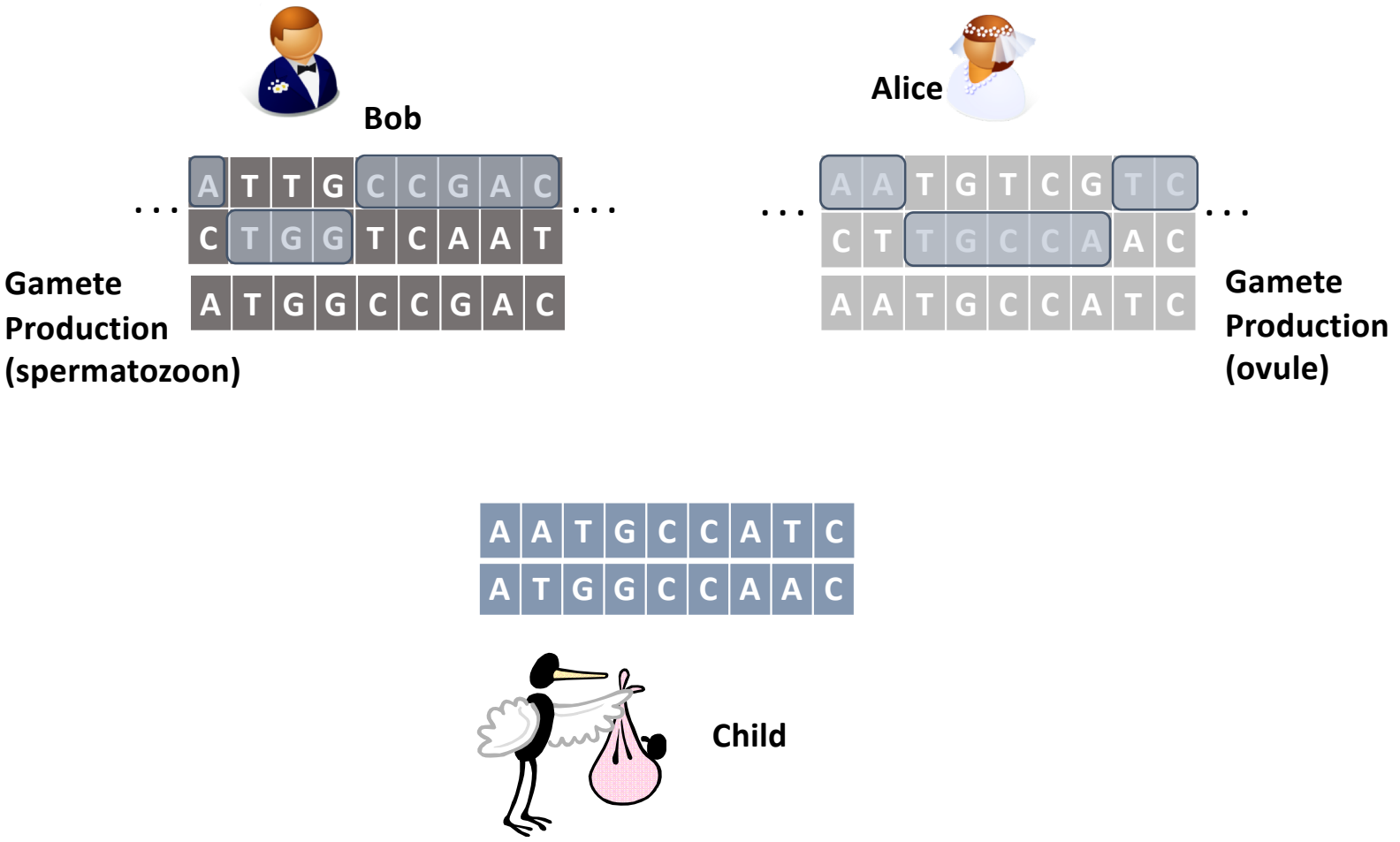
After having extensively authenticated each other,

after having exchanged thousands of highly private messages,

after having established numerous secure channels between each other,

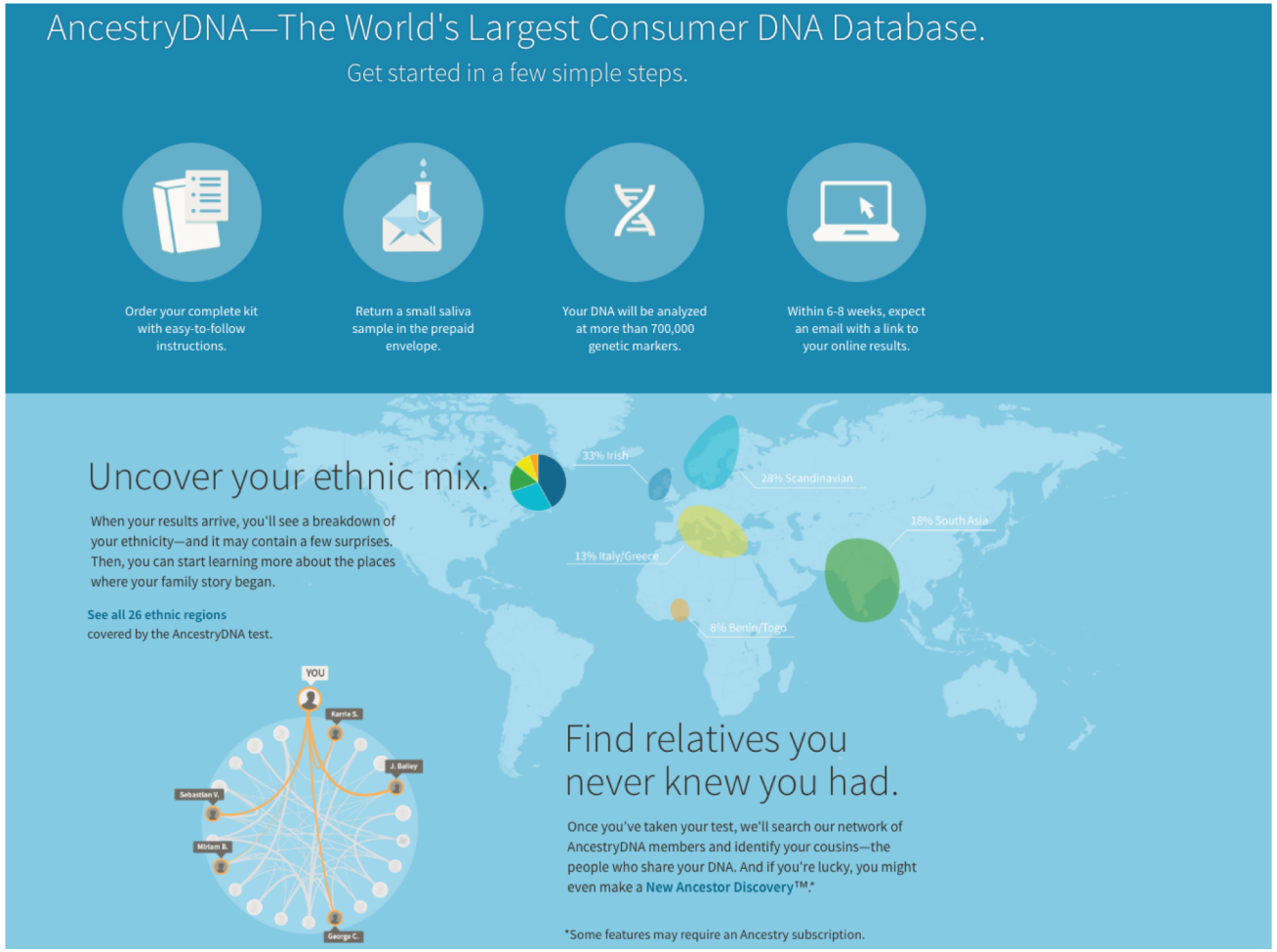
after years of intense but platonic relationship, finally, finally…❤️

… Alice and Bob got closer to each other



Direct to Consumer Genomics (1/2)

- Ancestry.com (millions of customers)



Direct to Consumer Genomics (2/2)

- 23andMe.com
  (millions customers)

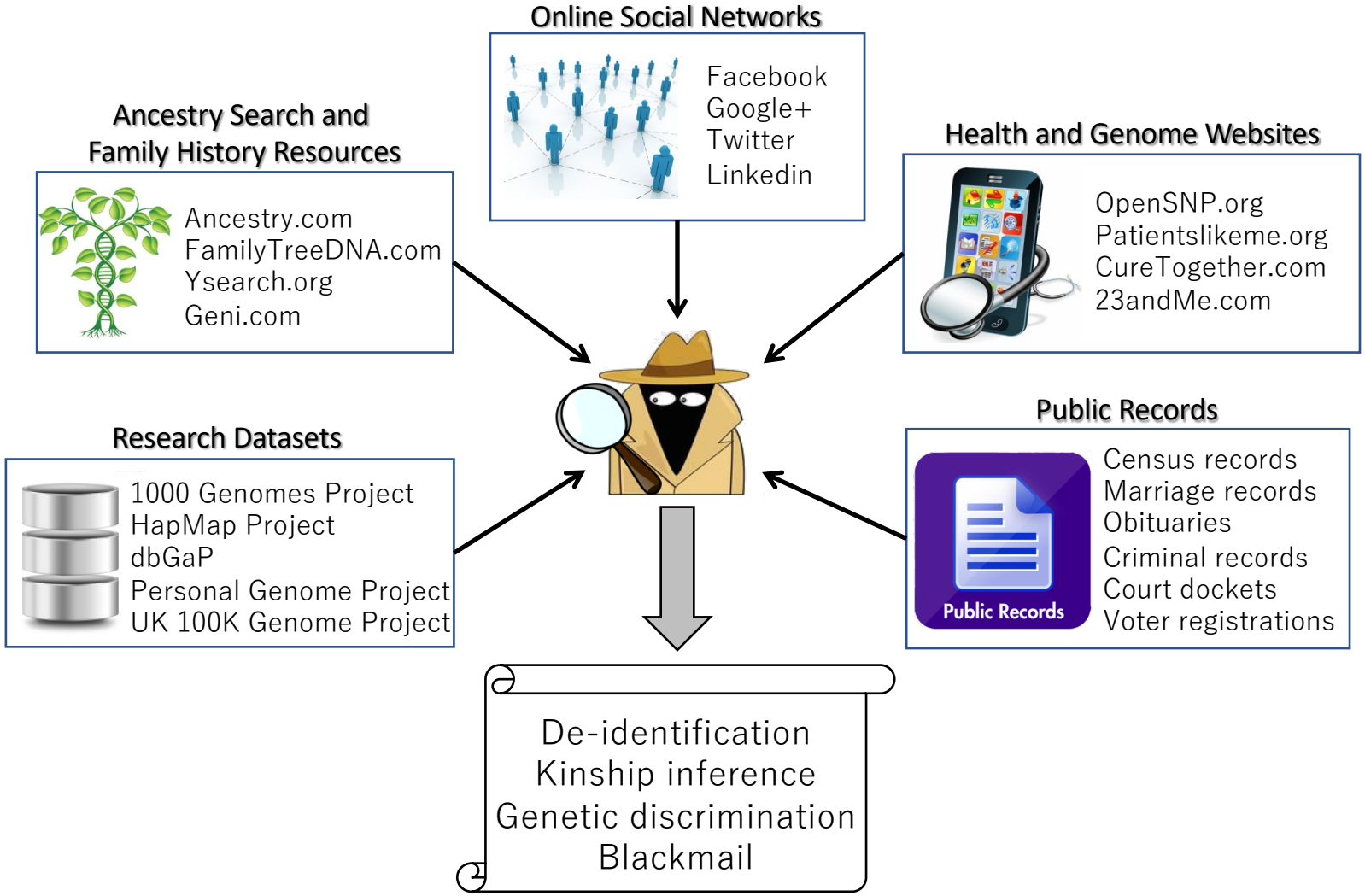
Name	Confidence	Your Risk	Avg. Risk
Atrial Fibrillation	★★★★	33.9%	27.2%
Prostate Cancer ♂	★★★★	29.3%	17.8%
Alzheimer's Disease	★★★★	14.2%	7.2%
Age-related Macular Degeneration	★★★★	11.1%	6.5%
Colorectal Cancer	★★★★	7.8%	5.6%
Chronic Kidney Disease	★★★★	4.2%	3.4%
Restless Legs Syndrome	★★★★	2.5%	2.0%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

Vox GENETICS

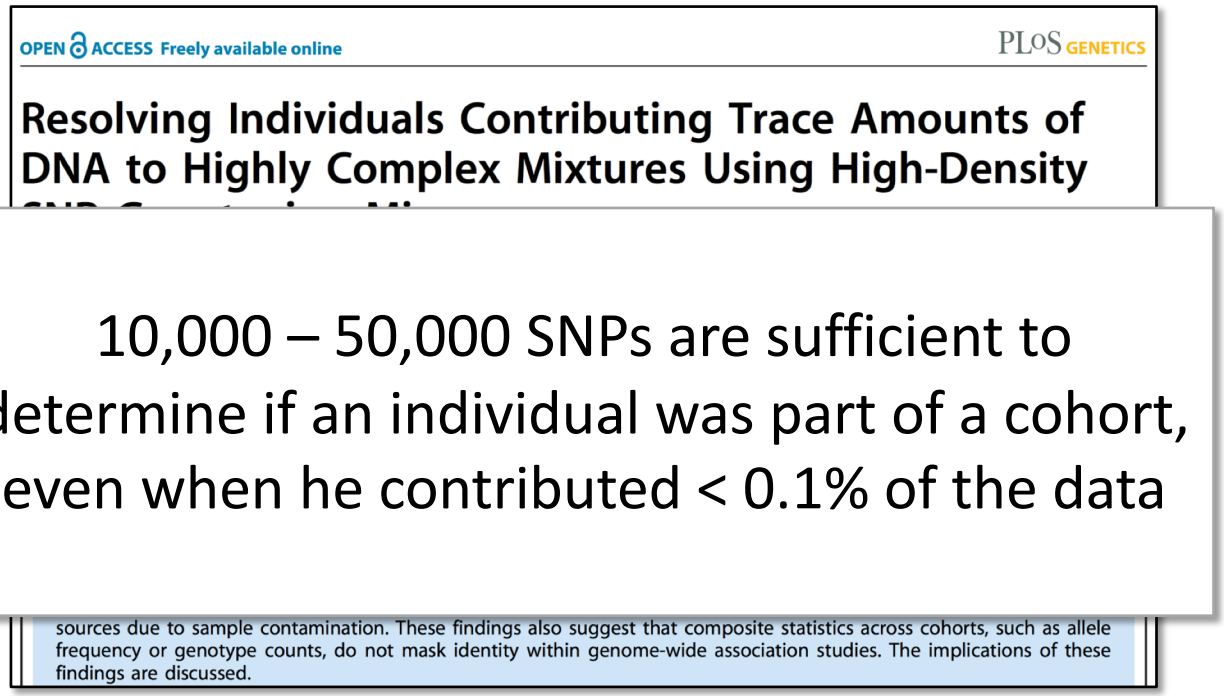
With genetic testing, I gave my parents the gift of divorce

Updated by George Doe on September 9, 2014, 7:50 a.m. ET

TWEET (2,073) SHARE (15K) +


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

Re-identification attacks on genomic data have a long history



OPEN ACCESS Freely available online

PLoS GENETICS

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density

10,000 – 50,000 SNPs are sufficient to determine if an individual was part of a cohort, even when he contributed < 0.1% of the data

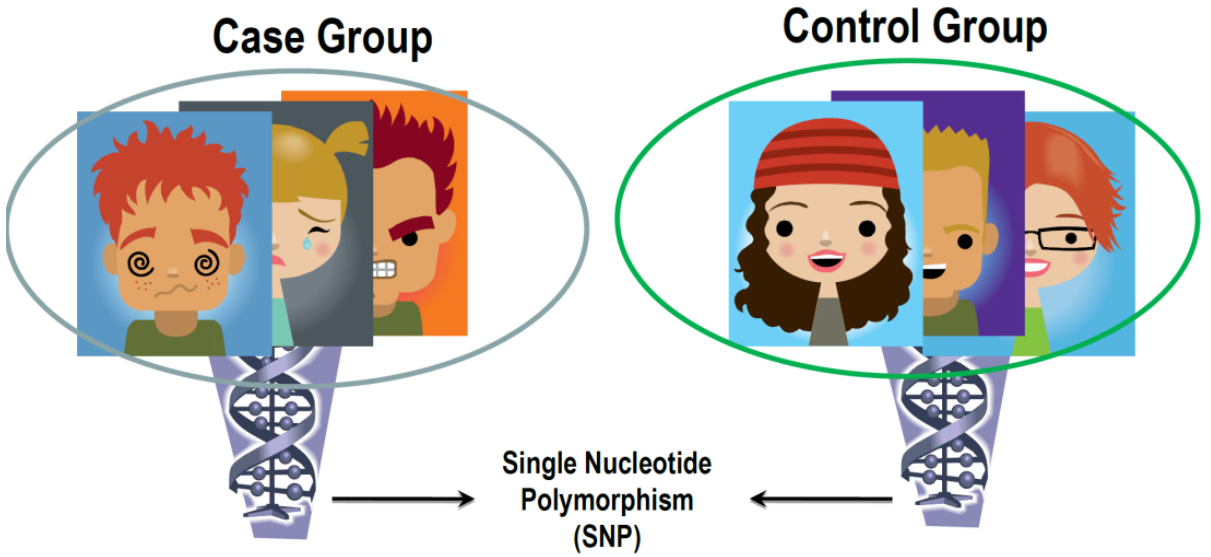
sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

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], …

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

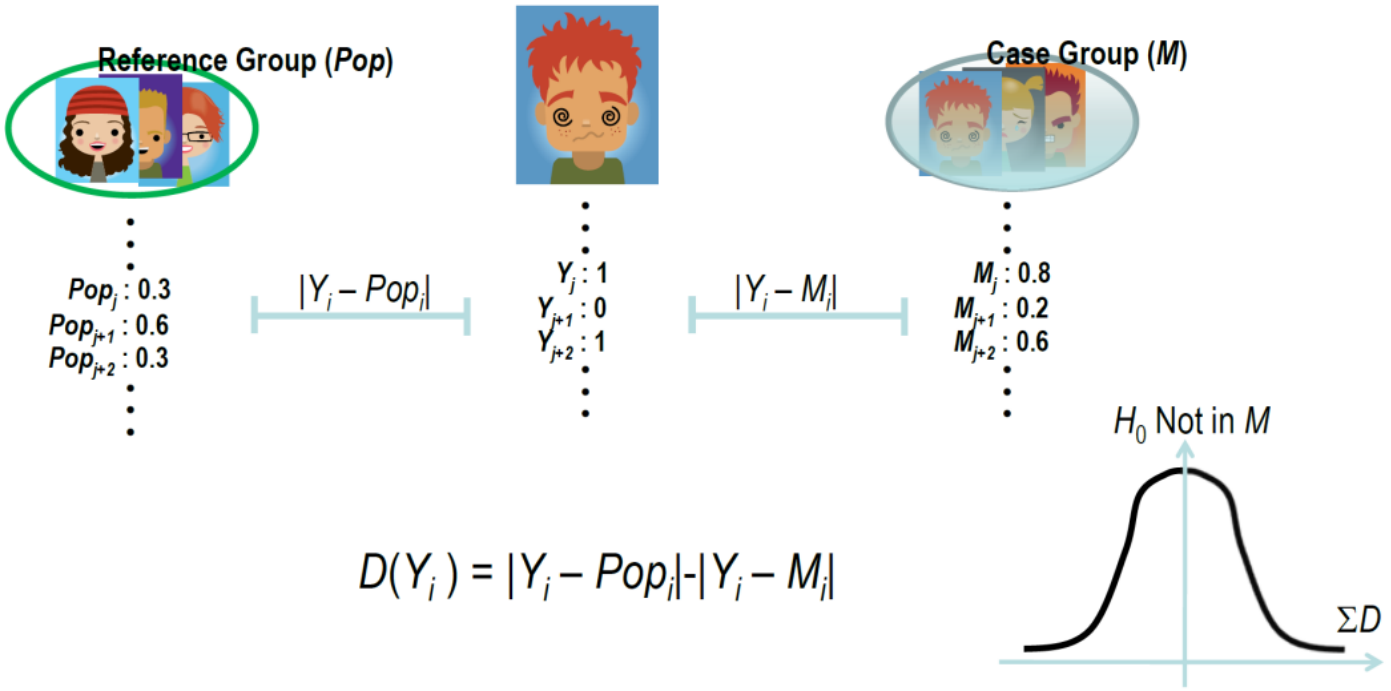


Figure from: Wang et al. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome-Wide Association Study

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

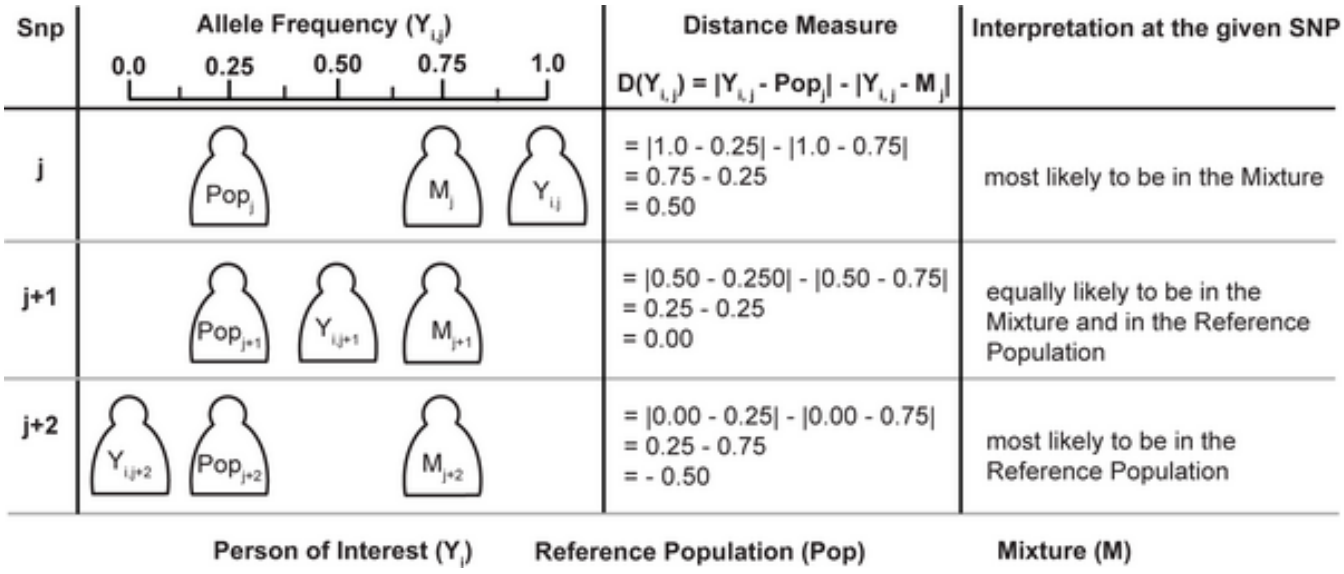
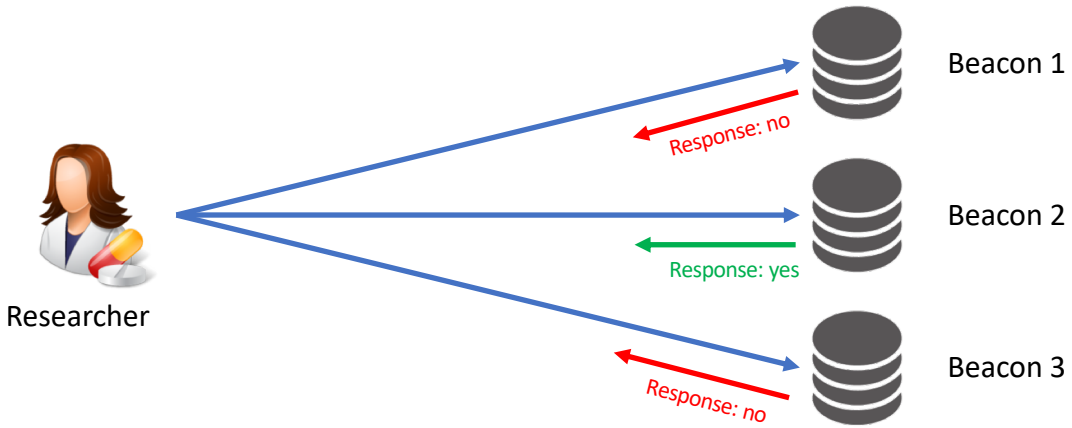


Figure taken from: Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genet 4(8): e1000167. doi:10.1371/journal.pgen.1000167

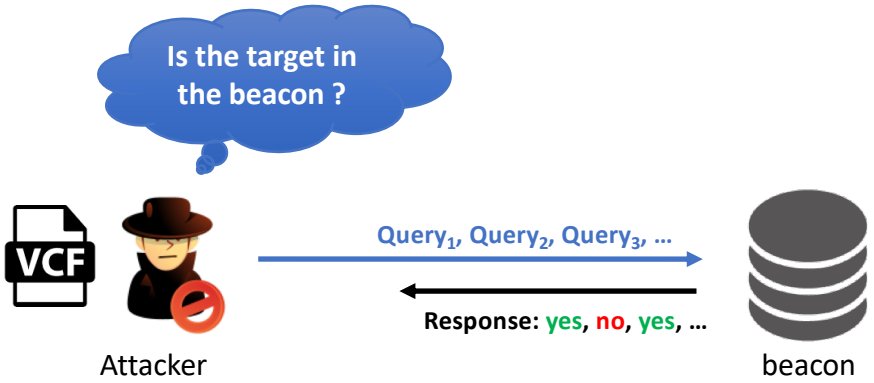
GA4GH Beacon Project

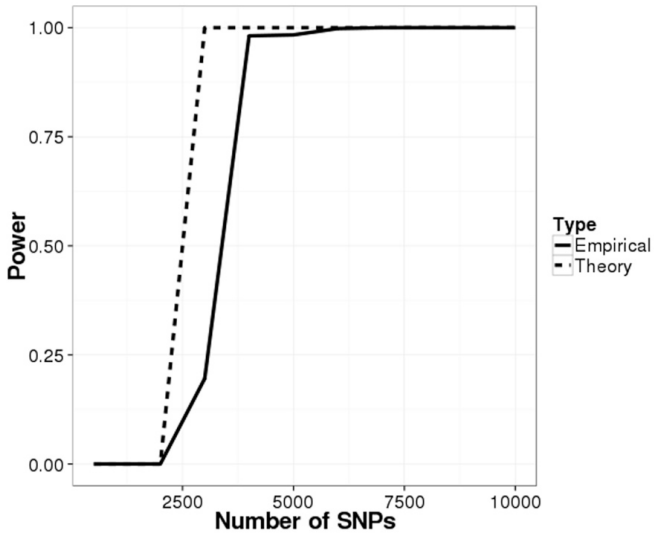


Main features:

- 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)

SB Attack: Likelihood Ratio Test

- H0: the target individual is not in the beacon
- H1: the target individual is in the beacon

- $R = \{x_1, \dots, x_n\}$ is the set of beacon responses
- δ probability of sequencing errors
- D_N^i denote the probability that none of the N other genomes in the beacon have an alternate allele at position i

$$L_{H_0}(R) = \sum_{i=1}^{n} x_i \log\left(1 - D_N^i\right) + (1 - x_i)\log\left(D_N^i\right),$$

$$L_{H_1}(R) = \sum_{i=1}^{n} x_i \log\left(1 - \delta D_{N-1}^i\right) + (1 - x_i)\log\left(\delta D_{N-1}^i\right).$$

$$\Lambda = L_{H_0}(R) - L_{H_1}(R)$$

“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 $f_1 \leq f_2 \leq \cdots \leq f_M$ ➔

$$D_N^i = \Pr(\text{none of the other N genomes have an alternate allele at position i})$$

Original attack [SB15]

$$D_N^i = \mathbb{E}[(1-f_i)^{2N}],$$

Where:
- $f_i \sim \text{beta}(a, b)$

“Optimal” attack [R et al. 16]

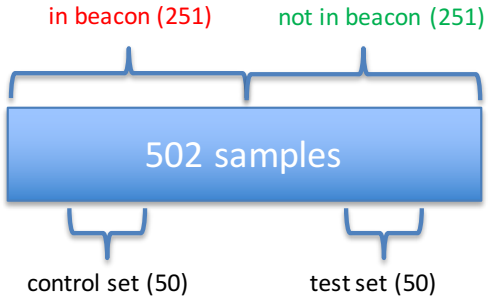
$$D_N^i = (1-f_i)^{2N}$$

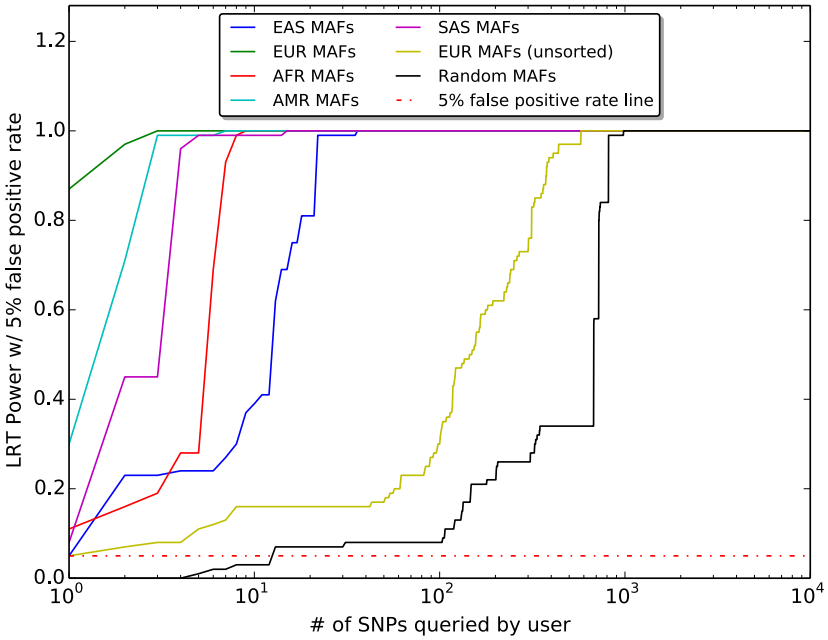
Where:
- f_i is different at every position i

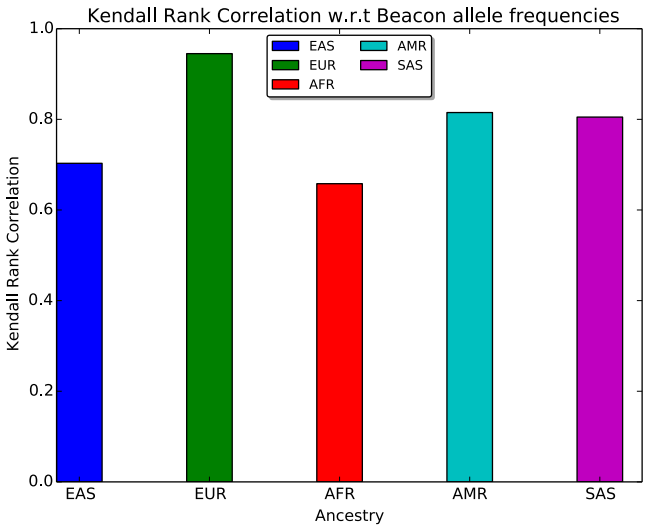
Raisaro JL, et al. “*Addressing Beacon Re-Identification Attacks: Quantification and Mitigation of Privacy Risks*”. JAMIA, February 2017

“Optimal” re-identification attack in single population beacon

- **Experimental setup**:
  - 502 EUR samples from 1000 Genomes Project Phase 3 (chromosome 10)
  - False positive rate of 5%

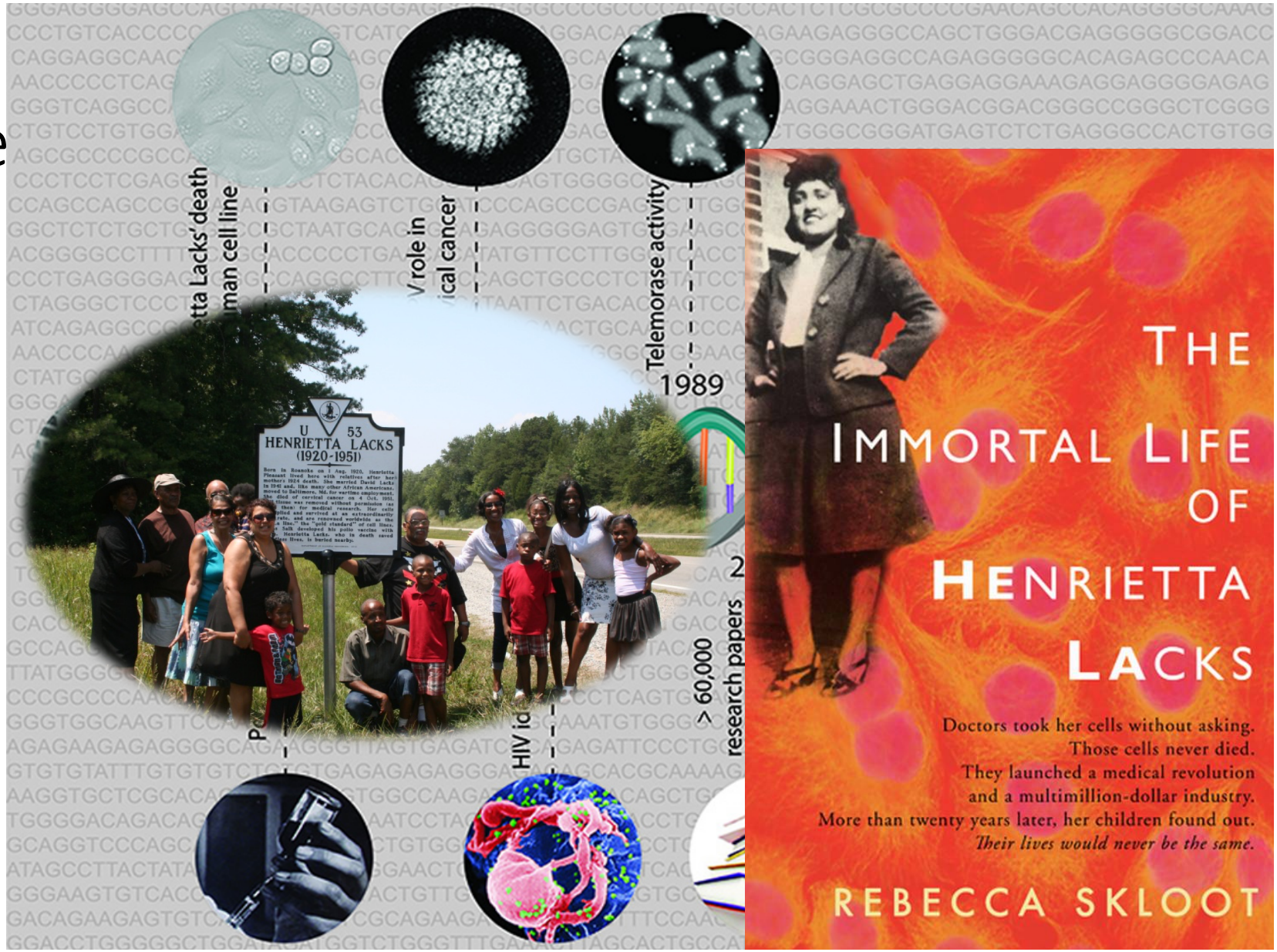






Kin Genomic Privacy

He


tta Lacks' death
man cell line
V role in
ical cancer
Telemorase activity
1989
U 53
HENRIETTA LACKS
(1920-1951)
HIV id
> 60,000
research papers
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