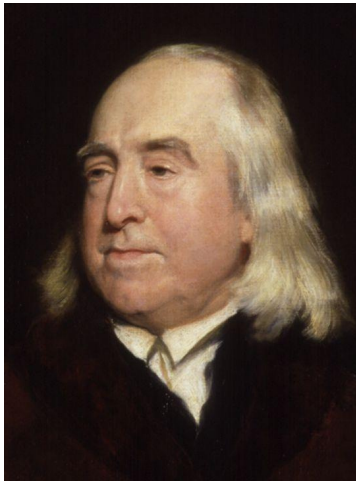


The Chills and Thrills of Whole Genome Sequencing

Emiliano De Cristofaro
https://emilianodc.com

Thanks to: E. Ayday, P. Baldi, R. Baronio, G. Danezis, S. Faber, P. Gasti, J-P. Hubaux, G. Tsudik

TL;DR

Progress in Genomics...

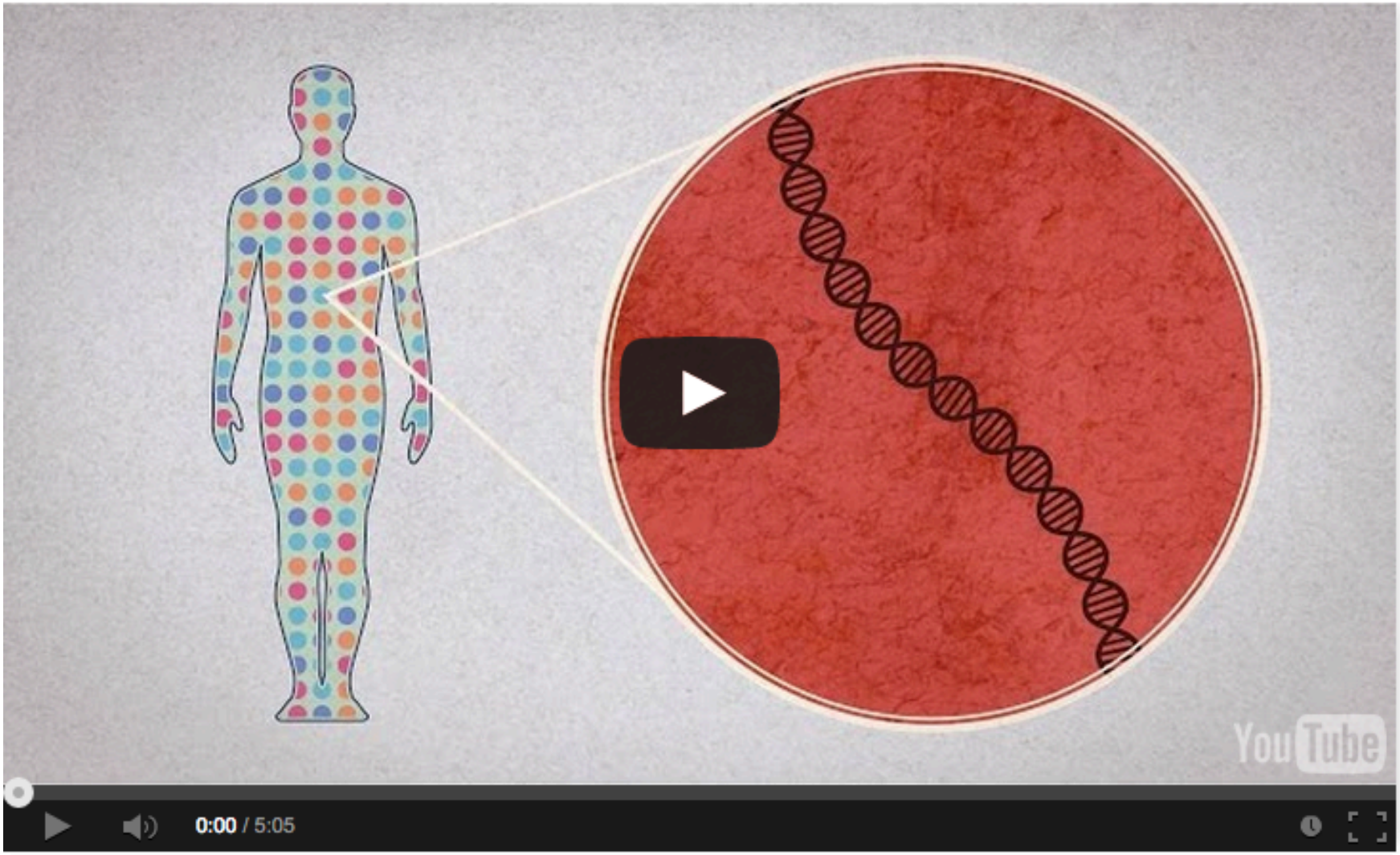
- Enables advances in medicine and healthcare
- Gives rise to a market for genetic testing
- Prompts a greater good vs privacy tension

Genomic Data is…

- Extremely sensitive
- Inherently hard to anonymize

In this talk…

- A computer scientist's perspective


0:00 / 5:05

How to read the genome?

Sequencing

Determining the full nucleotide order of an organism's genome

Genotyping

Determining genetic differences using a set of markers

WGS Progress

Some dates

1970s: DNA sequencing starts

1990: The “Human Genome Project” starts

2003: First human genome fully sequenced

2005: Personal Genome Project (PGP) starts

2012: UK announces sequencing of 100K genomes

Some numbers

$3B: Human Genome Project

$250K: Illumina (2008)

$5K: Complete Genomics (2009), Illumina (2011)

$1K: Illumina (2014)

1/05/2011 @ 4:57PM | 30,076 views

The First Child Saved By DNA Sequencing

+ Comment Now + Follow Comments

Genetic Gamble

New Approaches to Fighting Cancer

PART ONE
A Race to Leukemia's Source

PART TWO
Promise and Heartbreak

In Treatment for Leukemia, Glimpses of the Future

LETTER

doi:10.1038/nature13394

Genome sequencing identifies major causes of severe intellectual disability

Christian Gilissen[1]*, Jayne Y. Hehir-Kwa[1]*, Djie Tjwan Thung[1], Maartje van de Vorst[1], Bregje W. M. van Bon[1], Marjolein H. Willemsen[1], Michael Kwint[1], Irene M. Janssen[1], Alexander Hoischen[1], Annette Schenck[1], Richard Leach[2], Robert Klein[2], Rick Tearle[2], Tan Bo[1,3], Rolph Pfundt[1], Helger G. Yntema[1], Bert B. A. de Vries[1], Tjitske Kleefstra[1], Han G. Brunner[1,4]*, Lisenka E. L. M. Vissers[1]* & Joris A. Veltman[1,4]*

The Good News

Affordable WGS facilitates the creation of large datasets for research purposes

- Crucial for hypothesis-driven research

Low-cost WGS will bring genomics to the masses

- Motivated by clinical care and/or personal curiosity, a large number of individuals will have the means to have their (fully) genome sequenced, and possibly store/retain it

In general, genomic tests can be done “in silico”, using specialized computation algorithms


MAY 27, 2013
TIME
THE
ANGELINA
EFFECT
Angelina Jolie's double mastectomy puts genetic testing in the spotlight. What her choice reveals about calculating risk, cost and peace of mind
BY JEFFREY KLUGER & ALICE PARK
time.com
Time

disease risk

Share my health results with family and friends

Show results for

See new and recently updated reports »

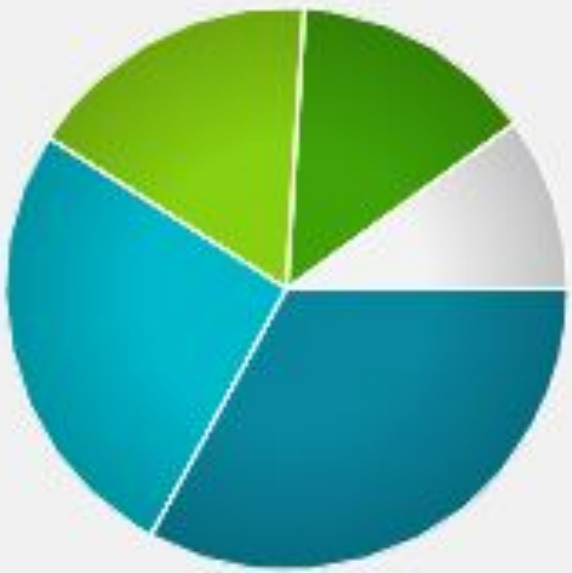
23andWe Discoveries were made possible by 23andMe members who took surveys.

Elevated Risk

Name	Confidence	Your Risk	Avg. Risk	Compared to Average
Type 2 Diabetes	★★★★	23.6%	18.2%	1.30x higher risk
Age-related Macular Degeneration	★★★★	16.1%	7.0%	2.30x higher risk
Exfoliation Glaucoma	★★★★	2.9%	1.0%	2.89x higher risk
Bipolar Disorder	★★★★	0.2%	0.1%	1.44x higher risk
Ankylosing Spondylitis	★★★			↑
Gallstones	★★★			↑
High Blood Pressure (Hypertension)	★★★			↑
Primary Biliary Cirrhosis	★★★			↑
Stomach Cancer	★★★			↑
Thyroid Cancer	★★★			↑
Cleft Lip and Cleft Palate	★★			↑
Essential Tremor	★★			↑

AncestryDNA Results

Your genetic ethnicity reveals where your ancestors lived hundreds—perhaps even thousands—of years ago.



Scandinavian	33%
British Isles	26%
Eastern European	17%
Central European	14%
Uncertain	10%

See Full Results

Genome: A CS Perspective

Genomics: A CS Perspective

Once sequenced… a genome becomes an (annotated) file

- Needs to be stored somewhere
- Can be queried/searched/tested/etc

But… not all data are created equal!

Security Researcher's Perspective

Genome = the ultimate identifier

- Hard to anonymize / de-identify

Treasure trove of sensitive information

- Ethnic heritage, predisposition to diseases

Sensitivity is perpetual

- Cannot be "revoked"
- Leaking one's genome ≈ leaking relatives' genome

The Greater Good vs Privacy?

Genomic advances dependent on data sharing

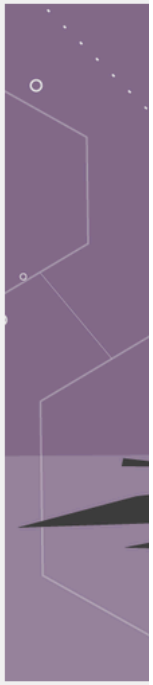
Sharing is an important asset for research in genomics

Privacy and discrimination fears are top concerns

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 SHARE



TECH 4/25/2013 @ 3:47PM | 17,111 views

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study

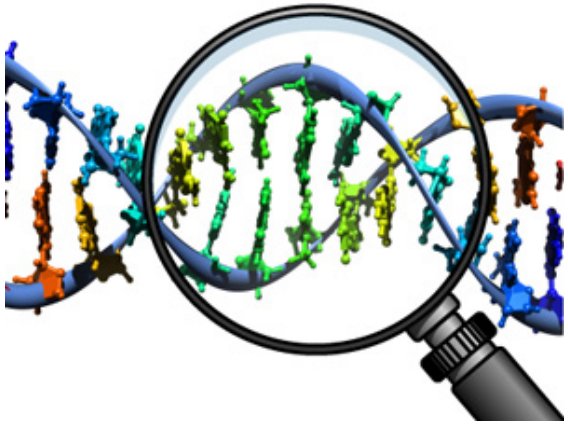
+ Comment Now + Follow Comments

A Harvard professor has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.

From the onset, the Personal Genome Project,

The rise of a new research community

Studying privacy issues



Exploring techniques to protect privacy

Kin Privacy

Quantifying how much privacy do relatives lose when one genome is leaked?

M. Humbert et al., *"Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy."* Proceedings of ACM CCS, 2013

Re-Identification

TECH 4/25/2013 @ 3:47PM | 17,111 views

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study

+ Comment Now + Follow Comments

A Harvard professor has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.

Harvard Professor Latanya Sweeney

From the onset, the Personal Genome Project,

Melissa Gymrek et al. *"Identifying Personal Genomes by Surname Inference."* Science Vol. 339, No. 6117, 2013

Studying Privacy

OK… anonymization doesn't really work. What about aggregation?

Even statistics from allele frequencies can be used to identify genetic trial participants

Rui Wang et al. "Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study." Proceedings of ACM CCS, 2009

Routes for breaching privacy

Y. Erlich and A. Narayanan. "Routes for Breaching and Protecting Genetic Privacy." Nature Review Genetics, Vol. 15, No. 6, 2014

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

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

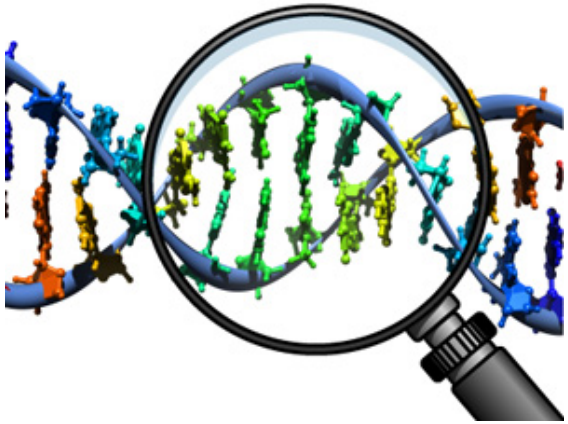
TWEET SHARE +

Most Read

1. Read the Iranian foreign minister passive aggressive response to Ton
2. Where the world's migrants go, in
3. Why there's a roaring controver Hillary Clinton's "homebrewed"
4. A new theory for why the bees are v
5.

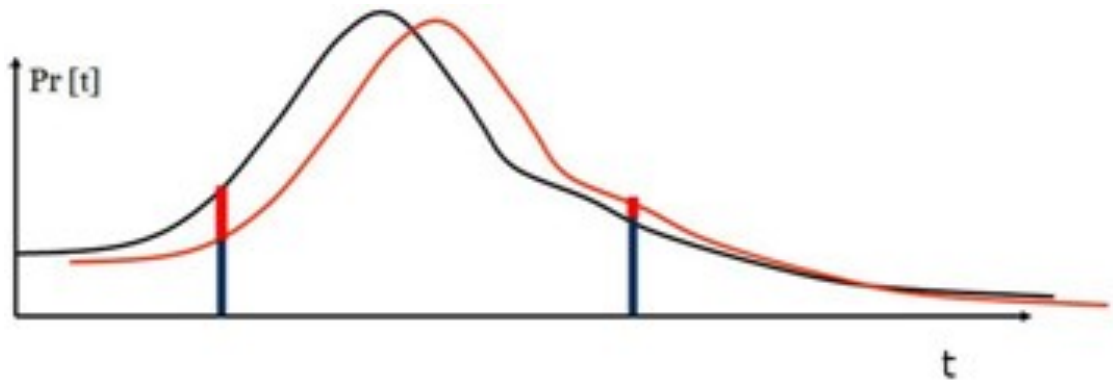
The rise of a new research community

Studying privacy issues



Exploring techniques to protect privacy

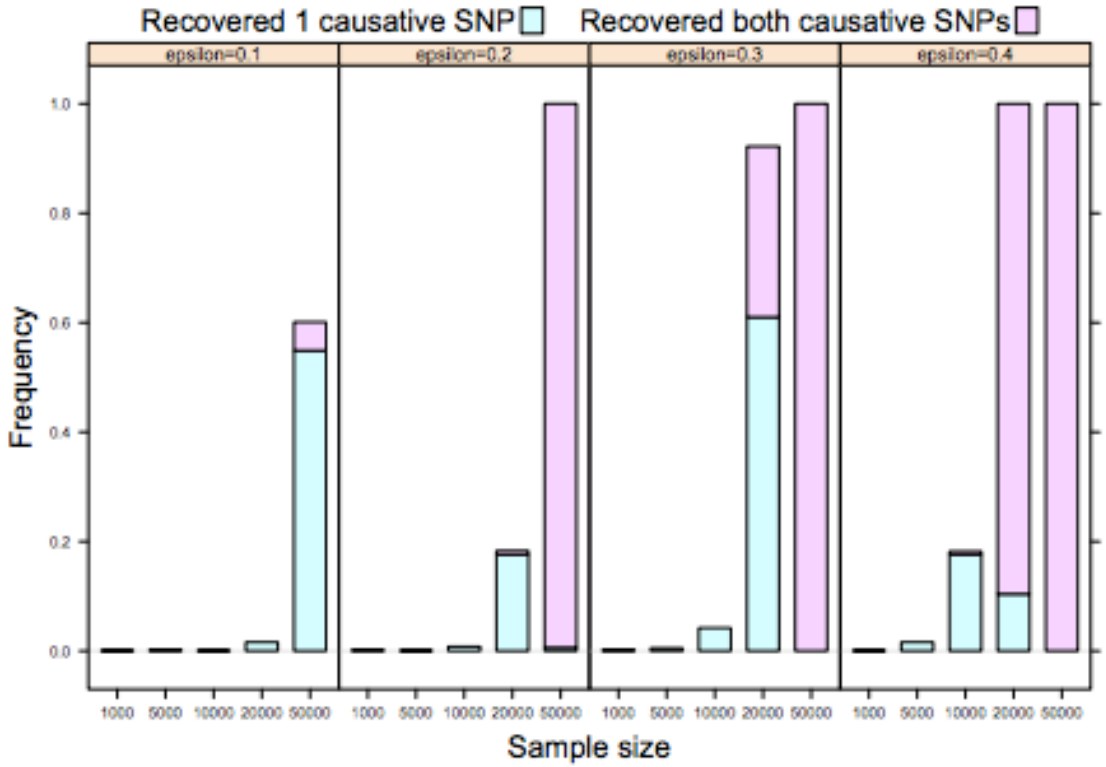
Differential Privacy



Maximizing the accuracy of queries from statistical databases

Minimizing the chances of identifying its records

Differential Privacy



Supporting Genome Wide Association Studies (GWAS)

Computing number and location of SNPs associated to disease

Test significance, correlation, etc. between a SNP and a disease

A. Johnson and V. Shmatikov. *"Privacy-Preserving Data Exploration in Genome-Wide Association Studies."* Proceedings of KDD, 2013

Privacy-Friendly Personal Genomics

Privacy-preserving Genomic Tests

Privacy as control

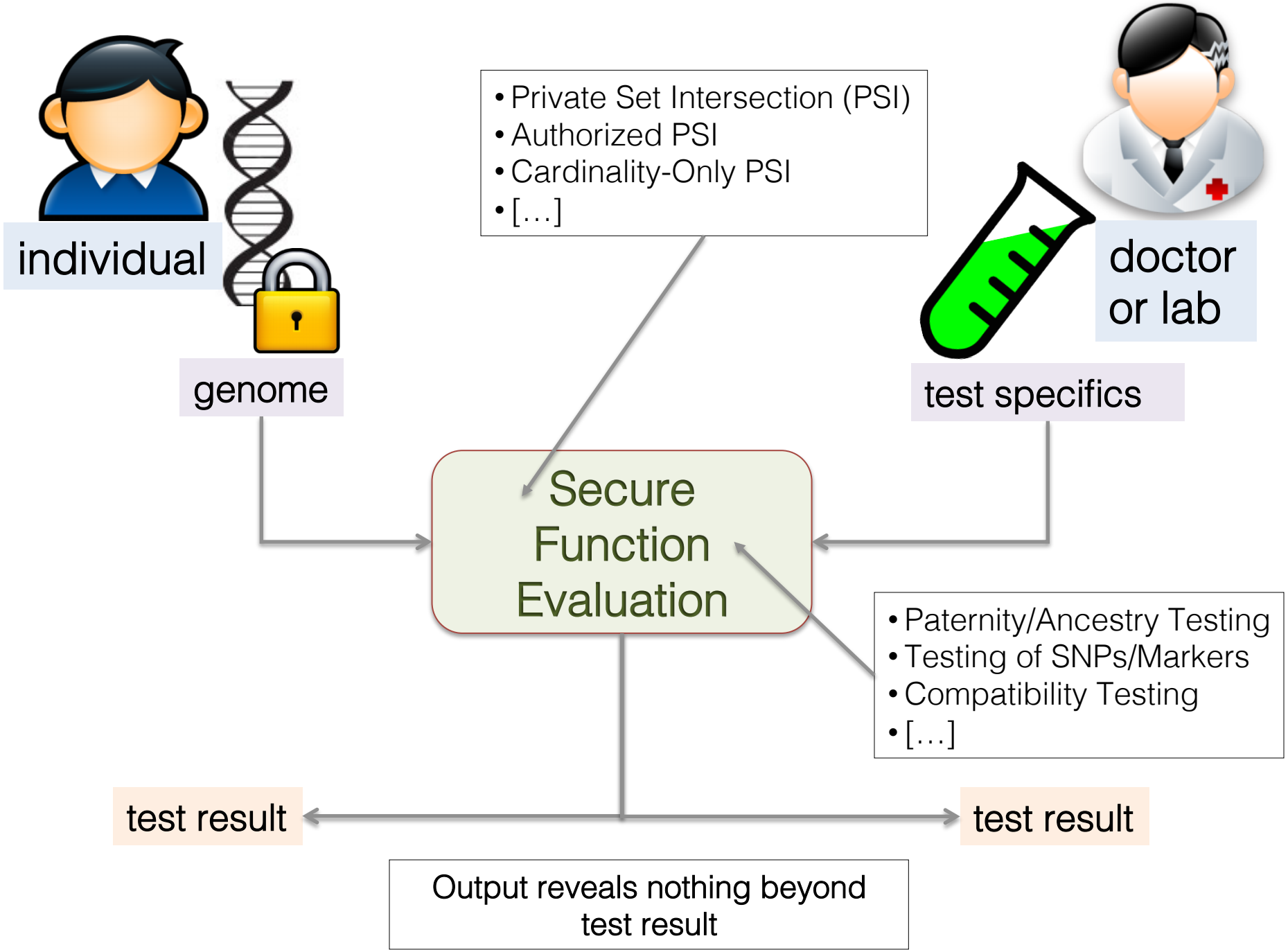
User's genome never disclosed (unless encrypted)

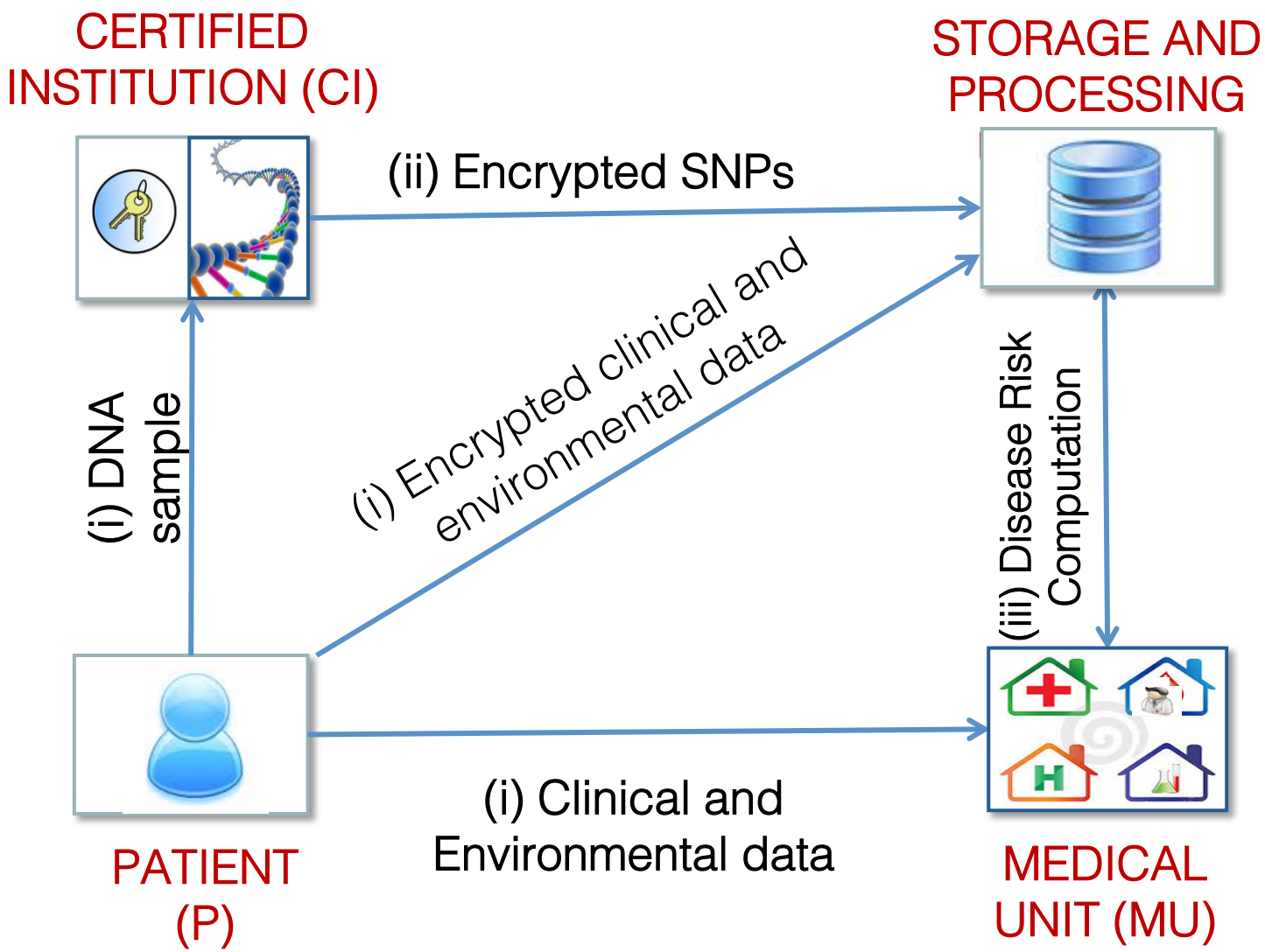
Doctors/labs run genomic tests on encrypted data

Disclose only the minimum **required amount of information**:

- Individuals don't reveal their entire genome, and
- Testing facilities keep test specifics confidential

Two main approaches…


individual
genome
• Private Set Intersection (PSI)
• Authorized PSI
• Cardinality-Only PSI
• […]
doctor or lab
test specifics
Secure Function Evaluation
• Paternity/Ancestry Testing
• Testing of SNPs/Markers
• Compatibility Testing
• […]
test result
test result
Output reveals nothing beyond test result


CERTIFIED INSTITUTION (CI)
STORAGE AND PROCESSING
(ii) Encrypted SNPs
(i) DNA sample
(i) Encrypted clinical and environmental data
(iii) Disease Risk Computation
(i) Clinical and Environmental data
PATIENT (P)
MEDICAL UNIT (MU)

1. Users keep sequenced genomes

Baldi et al. (CCS'11)

- Privacy-preserving version of a few genetic tests, based on private set operations
- Paternity test, Personalized Medicine, Compatibility Tests
- (First work to consider fully sequenced genomes)

De Cristofaro et al. (WPES'12), extends the above

- Framework and prototype deployment on Android
- Adds Ancestry/Genealogy Testing

More by yours truly ☺

2. Using Semi-Trusted Parties

Ayday et al. (WPES'13)

- Data is encrypted and stored at a "Storage Process Unit"
- Disease susceptibility testing

Ayday et al. (DPM'13)

- Encrypting raw genomic data (short reads)
- Allowing medical unit to privately retrieve them

Open Problems

Storage

If storing encrypted genomes at semi-trusted parties, encryption can't guarantee security past 30-50 yrs

If users keep (encrypted) copies of their genome, reliability and availability issues?

Miscellaneous

How much understanding/involvement required from users? Key distribution?

Efficiency overhead incurred by privacy protection layer?

Data representation assumptions:

- Insertions, deletions, sequencing errors, …

Question?

Why do we care about genome privacy???
We all leave biological cells behind…

Hair, saliva, etc., can be collected and sequenced?

But… collecting and sequencing samples is
expensive, illegal, prone to mistakes

Different scale of attacks!