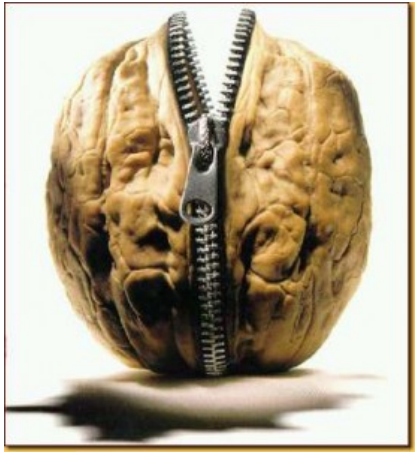
The Genomics Revolution: The Good, The Bad, and The Ugly

(A Privacy Researcher's Perspective)

Emiliano De Cristofaro
University College London
https://emilianodc.com

This Talk In a…



The Good

- Revolution in medicine and healthcare
- Genetic testing for the masses

The Bad

- Collection of highly sensitive data
- Very hard to anonymize / de-identify

The Ugly

- Greater good vs privacy
- Encryption might not be the answer

History

1970s: DNA sequencing starts

1990: The “Human Genome Project” starts

2003: First human genome fully sequenced

2012: UK announces sequencing of 100K genomes

2015: USA announces sequencing of 1M genomes

$$$

$3B: Human Genome Project

$250K: Illumina (2008)

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

$1K: Illumina (2014)

How to read the genome?

Genotyping

Testing for genetic differences using a set of markers

Sequencing

Determining the full nucleotide order of an organism's genome

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]*


MAY 27, 2013
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

health overview

Show results for [redacted]

See new and recently updated reports »

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

Disease Risks (114, 2 locked reports)

Elevated Risks	Your Risk	Average Risk
Psoriasis	22.4%	11.4%
Celiac Disease	0.5%	0.1%
Bipolar Disorder	0.2%	0.1%
Primary Biliary Cirrhosis	0.10%	0.08%
Scleroderma (Limited Cutaneous Type)	0.08%	0.07%

See all 114 risk reports...

Carrier Status (27, 1 locked report)

Hemochromatosis	Variant Present
Alpha-1 Antitrypsin Deficiency	Variant Absent
Bloom's Syndrome	Variant Absent
Canavan Disease	Variant Absent
Congenital Disorder of Glycosylation Type 1a (PMM2-CDG) new	Variant Absent
Cystic Fibrosis	Variant Absent
Familial Dysautonomia	Variant Absent
Factor XI Deficiency	Variant Absent

See all 27 carrier status...

Traits (52)

Alcohol Flush Reaction	Does Not Flush
Bitter Taste Perception	Can Taste
Earwax Type	Wet
Eye Color	Likely Blue
Hair Curl	Slightly Curlier Hair on Average

See all 52 traits...

Drug Response (20)

Warfarin (Coumadin®) Sensitivity	Increased
Abacavir Hypersensitivity	Typical
Alcohol Consumption, Smoking and Risk of Esophageal Cancer	Typical
Clopidogrel (Plavix®) Efficacy	Typical
Fluorouracil Toxicity	Typical

See all 20 drug response...

Genetic Ethnicity

Southern European	37%
West African	20%
British Isles	13%
Native South American	9%
Finnish/Volga-Ural	9%
Eastern European	6%
Uncertain	6%

List View | Map View | Surname View

search matches | Show: both sides | Sort: relationship | 25 per page | 1 - 25 of 424

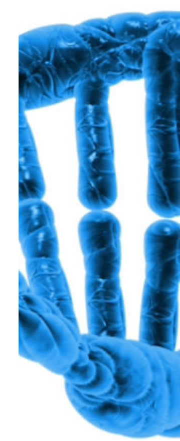
[redacted] Male	You	[redacted]	UPDATE YOUR PROFILE
Female	2nd to 3rd Cousin 1.68% shared, 5 segments	J2a2	Send an Introduction
[redacted] Female	3rd to 4th Cousin 1.30% shared, 3 segments	United States, Alsace-Lorraine (Strasbourg), Fr..., Paternal, [redacted], Senape, 5 more, U5b2	Public Match Send a Message
Male	3rd to 4th Cousin 1.03% shared, 2 segments	H13a1a, R1b1b2	Send an Introduction
Female	3rd to 5th Cousin 0.45% shared, 2 segments	H7	Send an Introduction
Female	3rd to 5th Cousin 0.42% shared, 2 segments	H1	Send an Introduction
[redacted] Male	3rd to 5th Cousin 0.40% shared, 2 segments	United States, Reno, Nevada, San Diego, California, Tucker, Littlefield, Warga, 4 more, H1c, G2a	Public Match Send a Message
[redacted] Male	3rd to 5th Cousin 0.37% shared, 2 segments	United States, fathers father prince Edward isla..., [redacted], K1a1b, R1b1b2a1a	Public Match Send a Message
Male, b. 1978	3rd to 6th Cousin 0.40% shared, 1 segment	United States, New Jersey, Utah, California, Northern Europe, U3b1, T	Send an Introduction

ex1.sam

ex1.sam 〉 No Selection

```
@HD VN:1.0 SO:coordinate
@SQ SN:seq1 LN:5000
@SQ SN:seq2 LN:5000
@CO Example of SAM/BAM file format.
B7_591:4:96:693:509 73  seq1    1   99  36M *   0   0   CACTAGTGGCTCATTGTAAATGTGTGGTTTAACTCG
    <<<<<<<<<<<<<<<;<<<<<<<<<5<<<<<;:<;7    MF:i:18 Aq:i:73 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS54_65:7:152:368:113  73  seq1    3   99  35M *   0   0   CTAGTGGCTCATTGTAAATGTGTGGTTTAACTCGT
    <<<<<<<<<<0<<<<655<<7<<<:9<<3/:<6): MF:i:18 Aq:i:66 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS51_64:8:5:734:57 137 seq1    5   99  35M *   0   0   AGTGGCTCATTGTAAATGTGTGGTTTAACTCGTCC
    <<<<<<<<<<<7;71<<;<;;<7;<<3;);3*8/5 MF:i:18 Aq:i:66 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
B7_591:1:289:587:906    137 seq1    6   63  36M *   0   0   GTGGCTCATTGTAATTTTTTGTTTTAACTCTTCTCT
    &----,----)-)-),'--)---',+-,),''*,  MF:i:130    Aq:i:63 NM:i:5  UQ:i:38 H0:i:0  H1:i:0
EAS56_59:[illegible]671:758  137 seq1    9   99  [illegible]5M * [illegible]   GC[illegible]ATTGT[illegible]AATGTGTGGTTTAACTCGTCCATGG
    <<<<[illegible];7<<<<<<<<7<[illegible]18 [illegible]NM:[illegible]i:1[illegible]
EAS56_61:[illegible]73 seq1 [illegible]35M [illegible]0 AT[illegible]GGCCCA
    <<<<<<<<;<<<8<<<<<;8:;6/686&;(16666 MF:i:18 Aq:i:39 NM:i:1  UQ:i:5  H0:i:0  H1:i:1
EAS114_28:5:296:340:699 137 seq1    13  99  36M *   0   0   ATTGTAAATGTGTGGTTTAACTCGTCCATGGCCCAG
    <<<<<;<<<;<;<<<<<<<<<<<8<8<3<8;<;<0;    MF:i:[illegible] Aq:i:73 NM:i:0  UQ:i:[illegible]0:i:1  H1:i:0
B7_597:6:194:894:408 [illegible]   0 [illegible]TAA[illegible]AACTCGTCCATTGCCCAGC
    <<<<<<<<<7<<;<<<<;[illegible]:0  H1:i:1
EAS188_4:8:12:628:973   89  seq1    18  75  35M *   0   0   [illegible]AATGTGTGGTTTAACTCGTCCATGGCCCAGCATT
    ==;=:;:;;:====;=;===:=======;==;=== MF:i:64 Aq:i:0  NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS51_66:7:68:402:50    137 seq1    22  99  35M *   0   0   GTGTGGTTTAACTCGTCCATGGCCCAGCATTTGGG
    <<<<<<<<<<<<<<:<<<9<6;9;;&697;7&<55 MF:i:18 Aq:i:66 NM:i:1  UQ:i:5  H0:i:1  H1:i:0
EAS114_30:6:298:115:564 137 seq1    22  99  35M *   0   0   GTGTGGTTTAACTCGTCCATGGCCCAGCATTAGGG
    <<<<<<<<<<<<<;<<<;<<<;<<<;<:<<;;5;; MF:i:18 Aq:i:72 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
B7_591:3:188:662:155    73  seq1    24  99  36M *   0   0   GTGGTTTAACTCGTCCATGGCCCAGCATTAGGGAGC
    <<<<<<<<<<<<<<<<<<<:;<<<4<<+<<14991;4   MF:i:18 Aq:i:71 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS56_59:2:225:608:291  73  seq1    28  99  35M *   0   0   TTTAACTCGTCCATGGCCCAGCATTAGGGATCTGT
    <<<<<<<<<<<<<<8&<<<;6<9;;+2++(%59(< MF:i:18 Aq:i:58 NM:i:1  UQ:i:4  H0:i:1  H1:i:0
EAS51_66:7:328:397:316  73  seq1    29  99  35M *   0   0   TTAACTCGTCCATGGCCCAGCATTAGGGAGCTGTG
    <<<<<<<<<<<<<6=<<<;<<5<<<+<15:'<;;4 MF:i:18 Aq:i:69 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS51_64:5:257:960:682  73  seq1    31  75  35M *   0   0   AACTCGTCCATGGCCCAGCATTAGGGAGCTGTGGA
    <<<<<<<<<<<;<<<<<<<<<<<<<<<<<<9;;9< MF:i:64 Aq:i:0  NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS54_61:4:143:69:578   99  seq1    36  98  35M =   185 184 GTACATGGCCCAGCATTAGGGAGCTGTGGACCCCG
    ===;=====48=844;=;+=5==*57,2+5&,5+5 MF:i:18 Aq:i:35 NM:i:2  UQ:i:38 H0:i:0  H1:i:1
```

But... not all data are created equal!



Privacy Researcher's Perspective

Treasure trove of sensitive information

Ethnic heritage, predisposition to diseases

Genome = the ultimate identifier

Hard to anonymize / de-identify

Sensitivity is perpetual

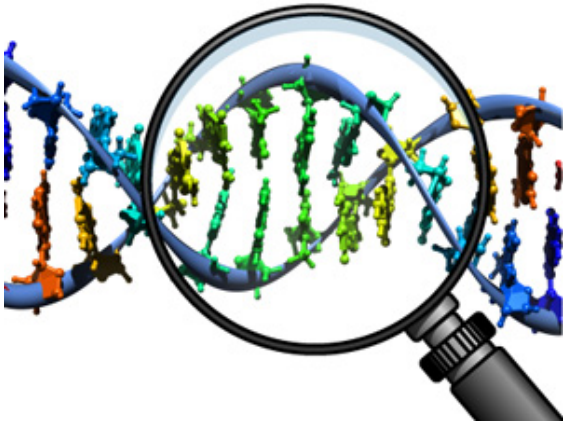
Cannot be "revoked"

Leaking one's genome ≈ leaking relatives' genome

The Greater Good
vs
Privacy?

The rise of a new research community

Studying privacy issues



Exploring techniques to protect privacy

De-Anonymization

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,

Harvard Professor Latanya Sweeney

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

Aggregation

Re-identification of aggregated data

Statistics from allele frequencies can be used to identify genetic trial participants [1]

Presence of an individual in a group can be determined by using allele frequencies and his DNA profile [2]

[1] R. Wang et al. “Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study.” CCS, 2009

[2] N. Homer et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics,2008

Kin Privacy

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

Also read: "Routes for breaching genetic privacy"
Y. Erlich and A. Narayanan,
Nature Review Genetics
Vol. 15, No. 6, 2014

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

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 controvers
Hillary Clinton's "homebrewed"

4. A new theory for why the bees are v

5.

Human Aspects of Genome Privacy

Dynamic Consent:

- Patients electronically control consent through time and receive information about the uses of their data
- Jane Kaye's work at Oxford

Understanding "fears" and "reactions", including:

- Survivor's guilt
- Freedom to withdraw is crucial but poorly understood
- Insurance carriers and big corporations most distrusted

Ethnographic Studies in WGS

Semi-structured interviews with 16 participants

Assessing perception of genetic tests, attitude toward WGS programs, as well as perception of privacy/ethical issues

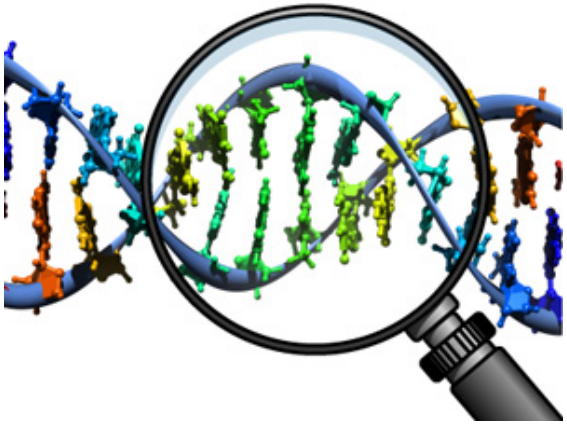
(Some) Preliminary results

1. Preferred method is through doctors not companies (trust)
2. Labor/healthcare discrimination top concerns
3. Differences in correlation with income and education

E. De Cristofaro. "*Users' Attitudes, Perception, and Concerns in the Era of Whole Genome Sequencing.*" (USEC 2014)

The rise of a new research community

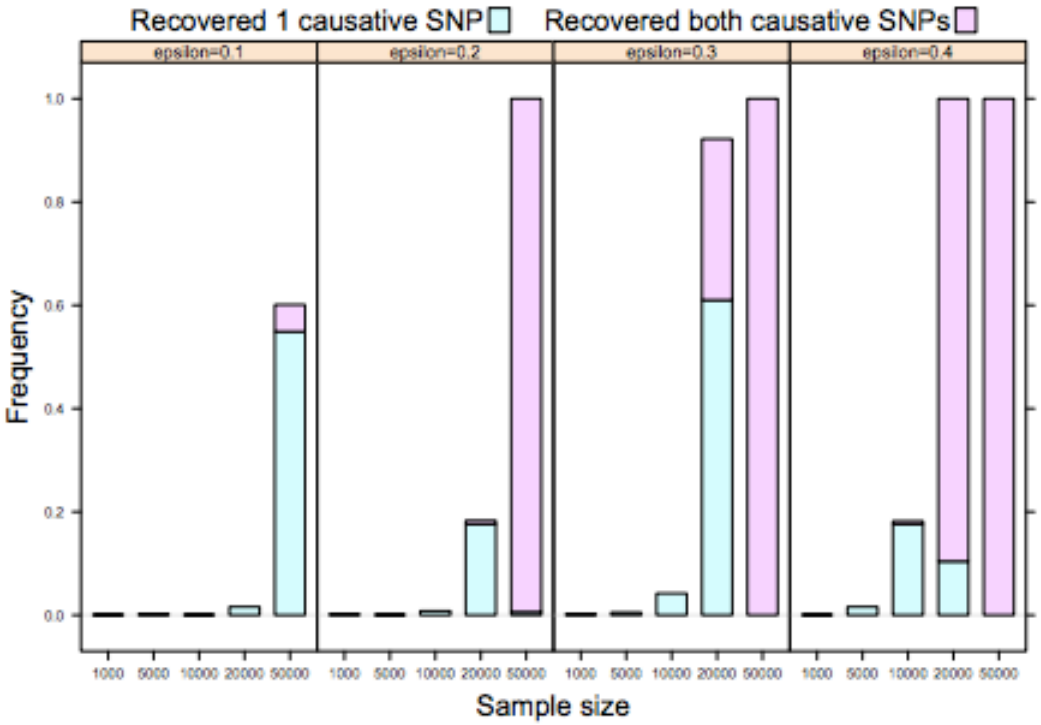
Studying privacy issues



Exploring techniques to protect privacy

Differential Privacy

Genome Wide Association Studies (GWAS)



Computing number/location of SNPs associated to disease

Significance/correlation between a SNP and a disease

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

Computing on Encrypted Genomes

Genomic datasets often used for association studies

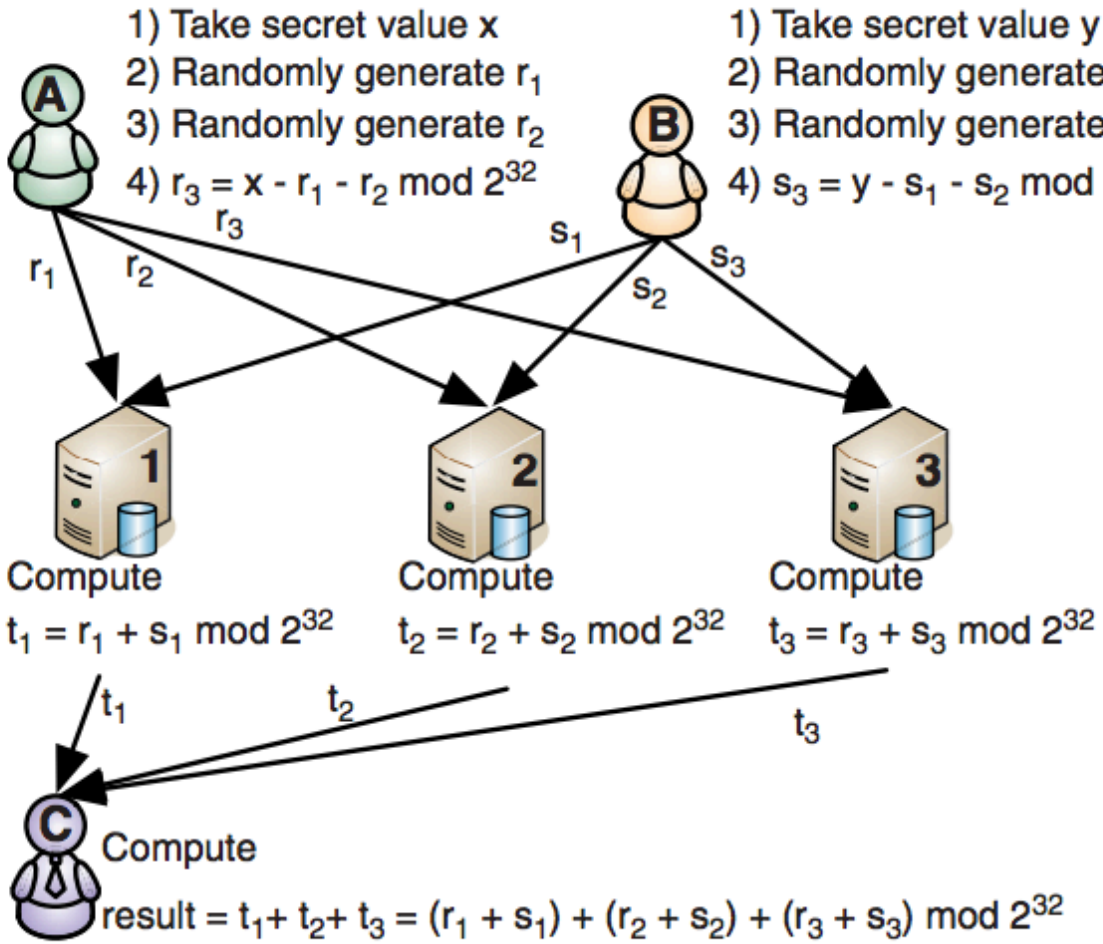
Encrypt data & outsource to the cloud

- Perform private computation over encrypted data
- Using partial & fully homomorphic encryption

Examples:

- Pearson Goodness-of-Fit test, linkage disequilibrium
- Estimation Maximization, Cochran-Armitage TT, etc.

Computing on Encrypted Genomes



L. Kamm, D. Bogdanov, S. Laur, J. Vilo.
A new way to protect privacy in large- scale genome-wide association studies. Bioinformatics 29 (7): 886-893, 2013.

Private *Personal* Genomic Tests

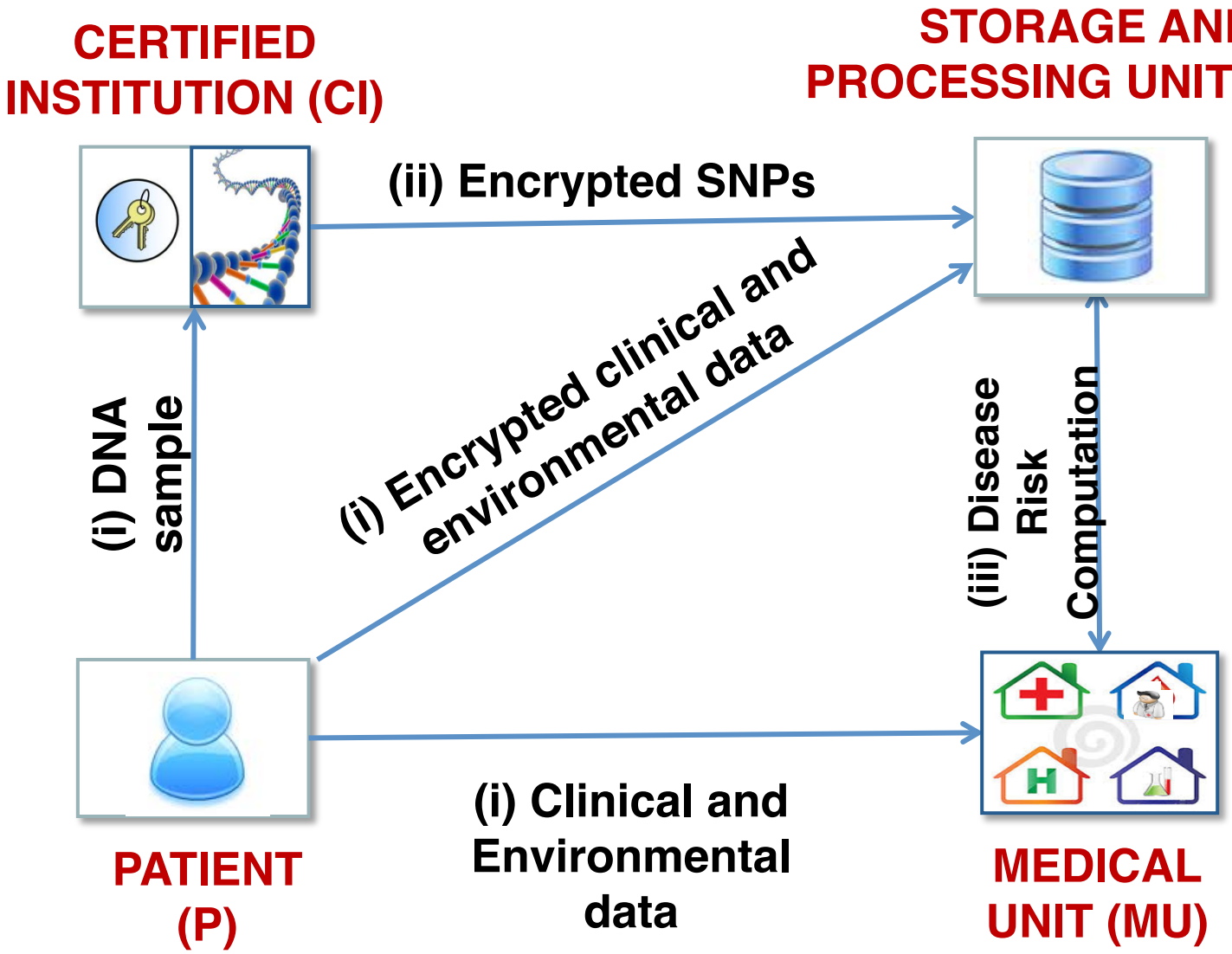
Individuals retain control of their sequenced genome

Allow doctors/labs to run genetics tests, but:

1. Genome never disclosed, only test output is
2. Pharmas can keep test specifics confidential

… two main approaches …

1. Using Semi-Trusted Parties



1. 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)

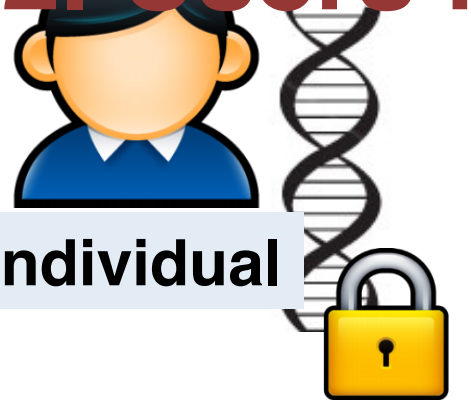
Encrypting raw genomic data (short reads)

Allowing medical unit to privately retrieve them

Danezis and De Cristofaro (WPES'14)

Regression for disease susceptibility

2. Users keep sequenced genomes



individual

genome

- Private Set Intersection (PSI)
- Authorized PSI
- Private Pattern Matching
- Homomorphic Encryption
- Garbled Circtuis
- […]

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

2. 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

Open Problems

Where do we store genomes?

- Encryption can't guarantee security past 30-50 yrs
- Reliability and availability issues?

Cryptography

- Efficiency overhead
- Data representation assumptions
- How much understanding required from users?

Why do we even care about genome privacy?

We all leave biological cells behind…

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

Compare this "attack" to re-identifying millions of DNA donors or hacking into 23andme…

The former: expensive, prone to mistakes, only works against a handful of targeted victims

The latter: very "scalable"

Epilogue

Whole Genome Sequencing

- A revolution in healthcare
- Raises worrisome privacy/ethical concerns

The Genomic Privacy research community

- Understanding the privacy issues
- Privacy-preserving testing on whole genomes
- Possible using efficient crypto protocols & cross-discipline collaboration

A number of open research issues…

For more info:

http://genomeprivacy.org

Also:

E. Ayday, E. De Cristofaro, J.P. Hubaux, G. Tsudik.
"Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare?"
IEEE Computer Magazine

Thank you!

Special thanks to

E. Ayday, P. Baldi, R. Baronio, G. Danezis, S. Faber,
P. Gasti, J-P. Hubaux, B. Malin, G. Tsudik