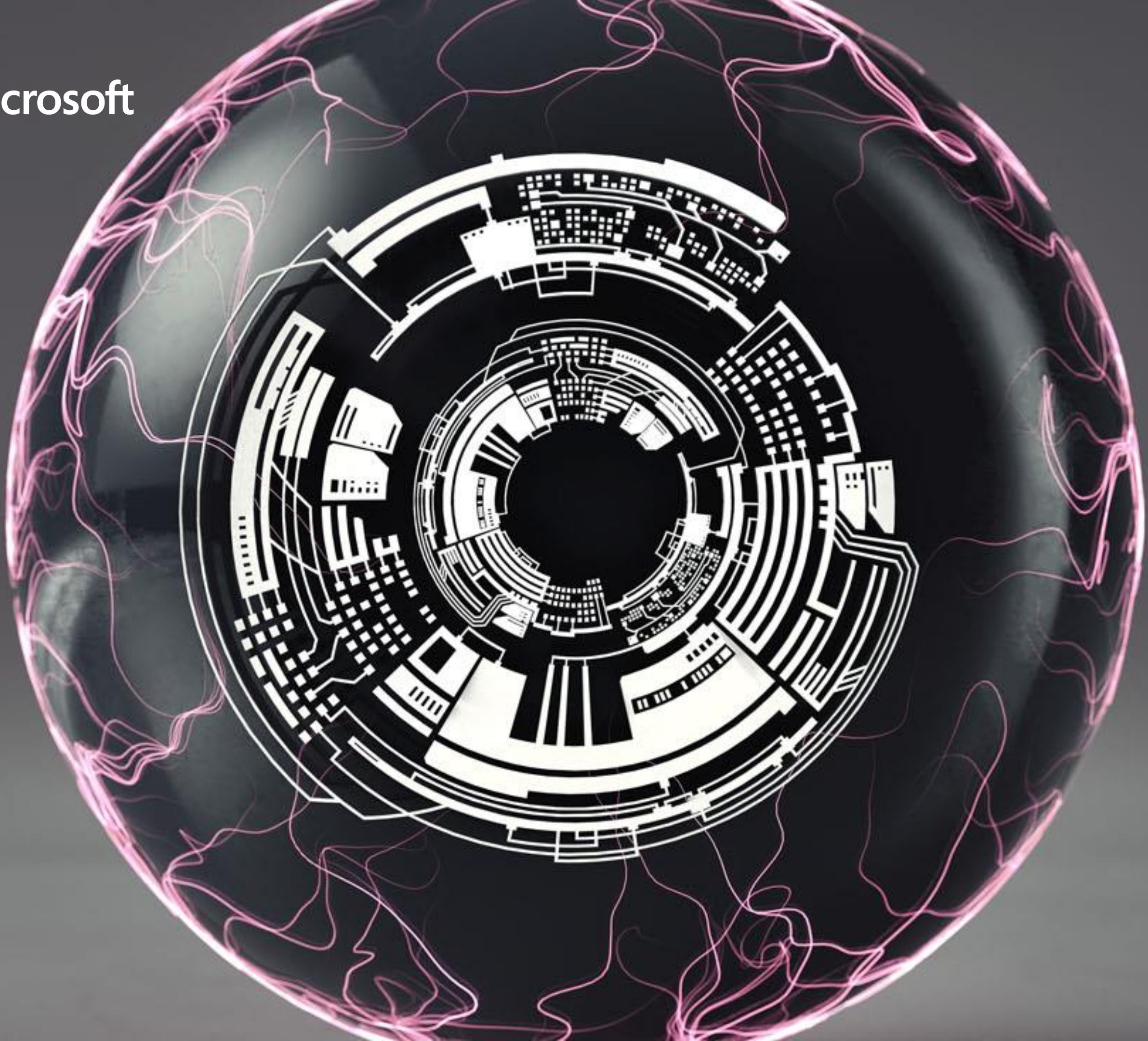
Microsoft

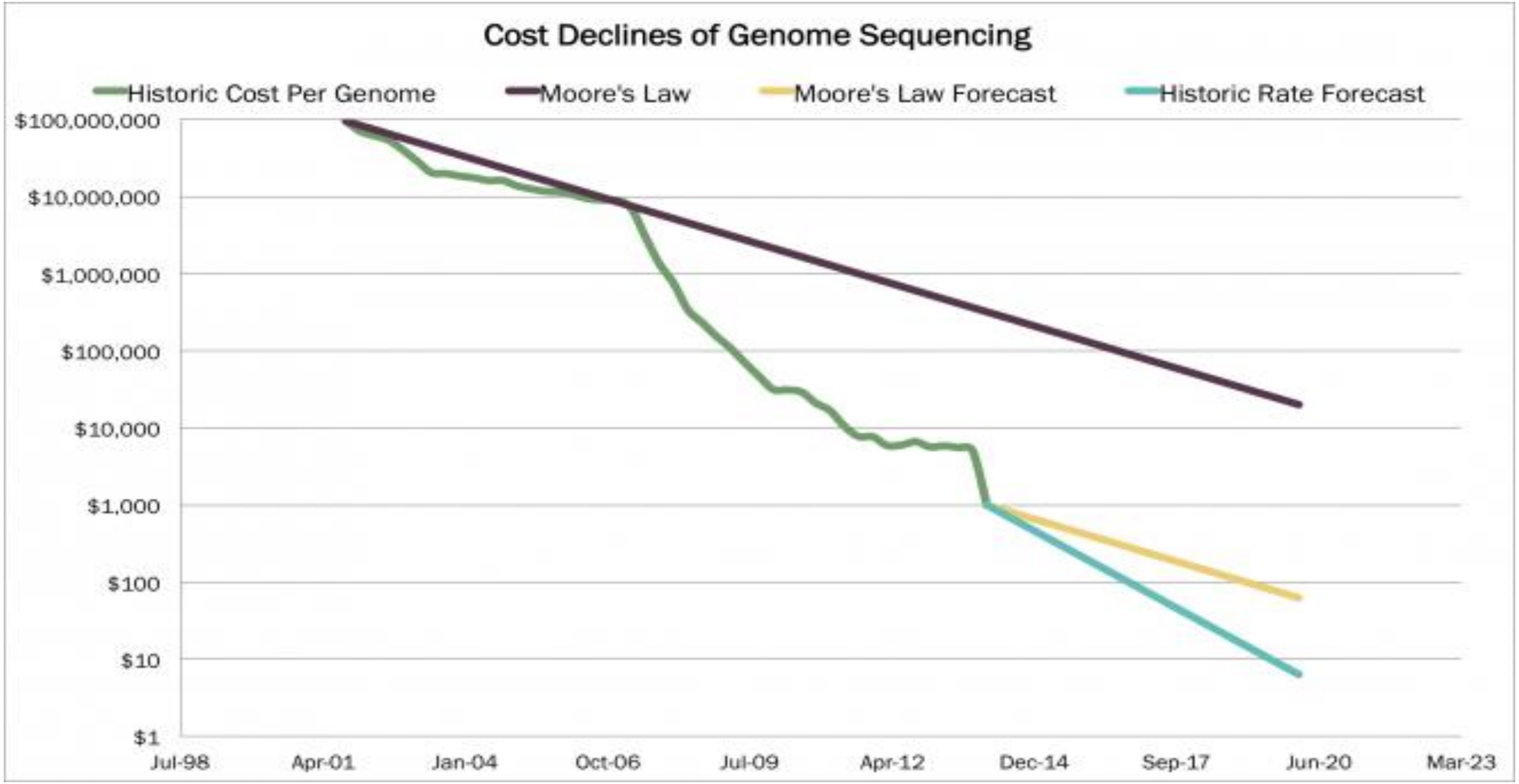


Microsoft Research Faculty Summit 2016

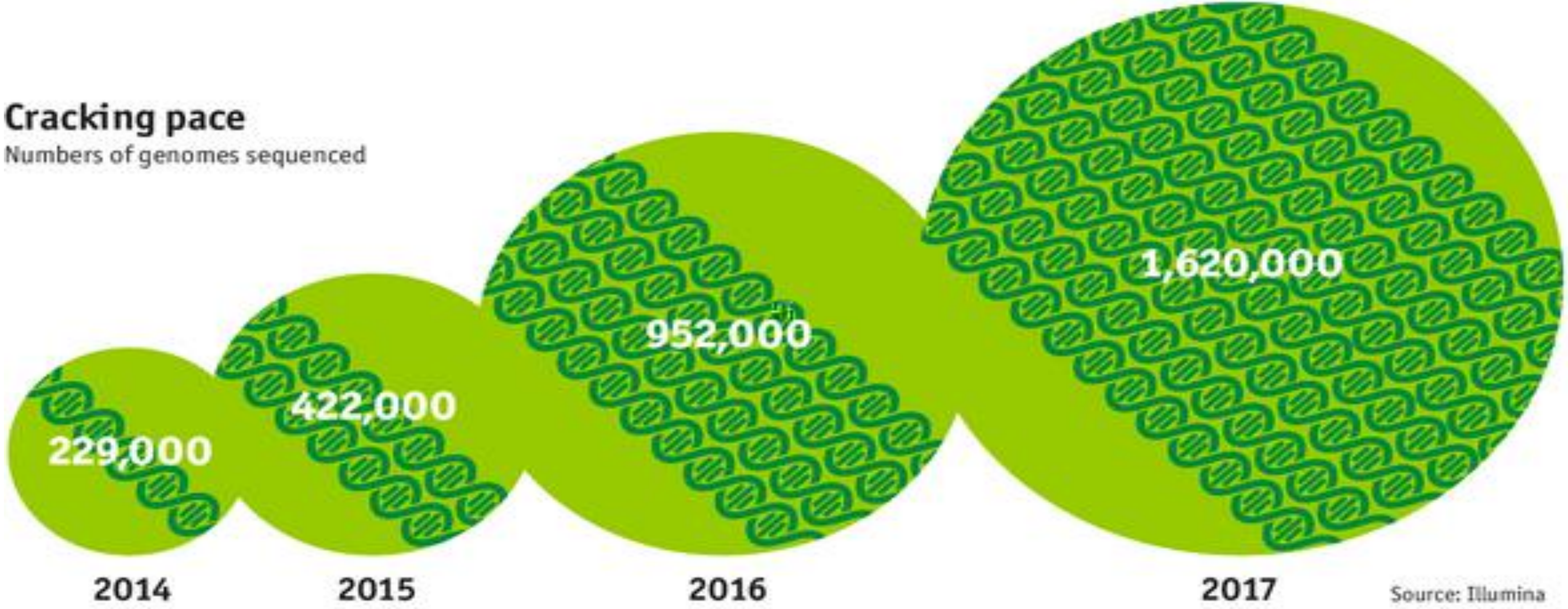
The Genomics Revolution: The Good, The Bad, and The Ugly

(The Privacy Edition)

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



From: James Bannon, ARK



Cracking pace

Numbers of genomes sequenced

Source: Illumina

From: The Economist

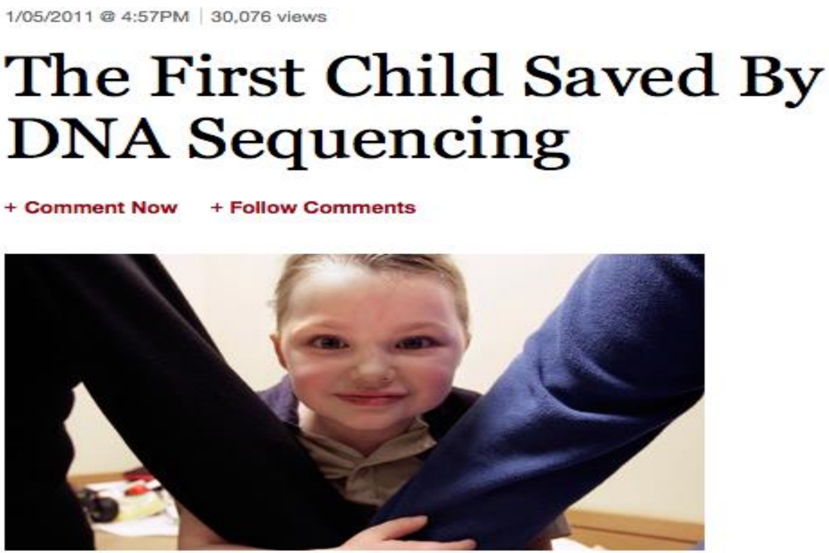
How to read the genome?

Genotyping

Testing for genetic differences using a set of markers

Sequencing

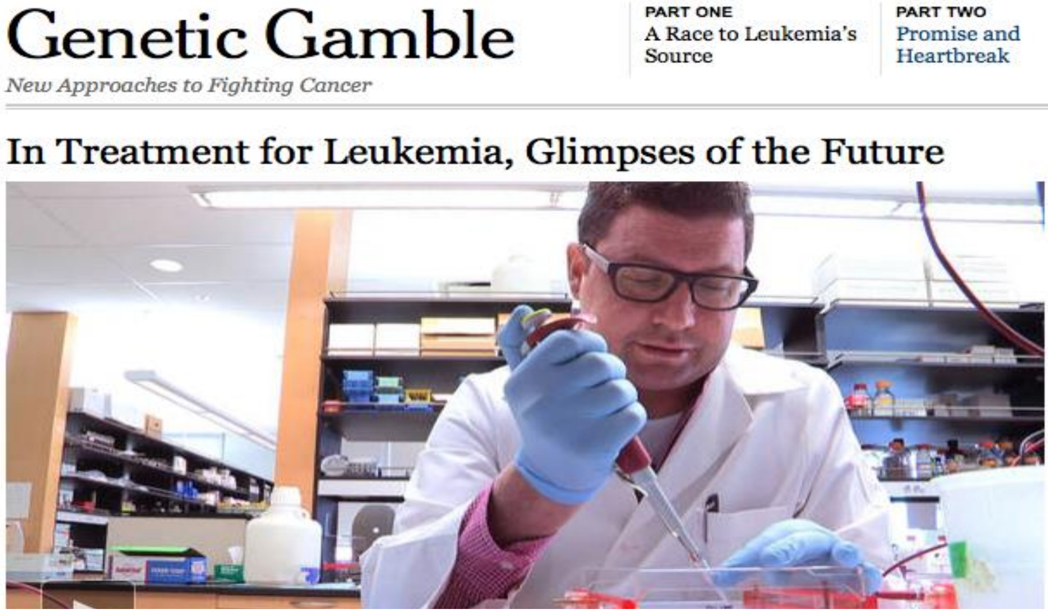
Determining the full nucleotide order of an organism’s genome



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

The First Child Saved By DNA Sequencing

+ Comment Now   + Follow Comments



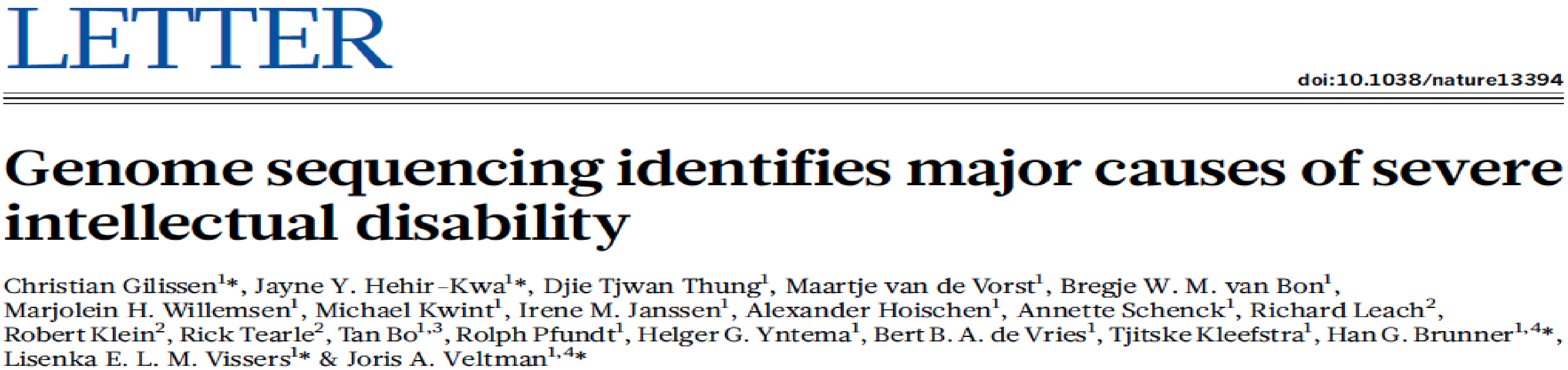
Genetic Gamble

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

TIME

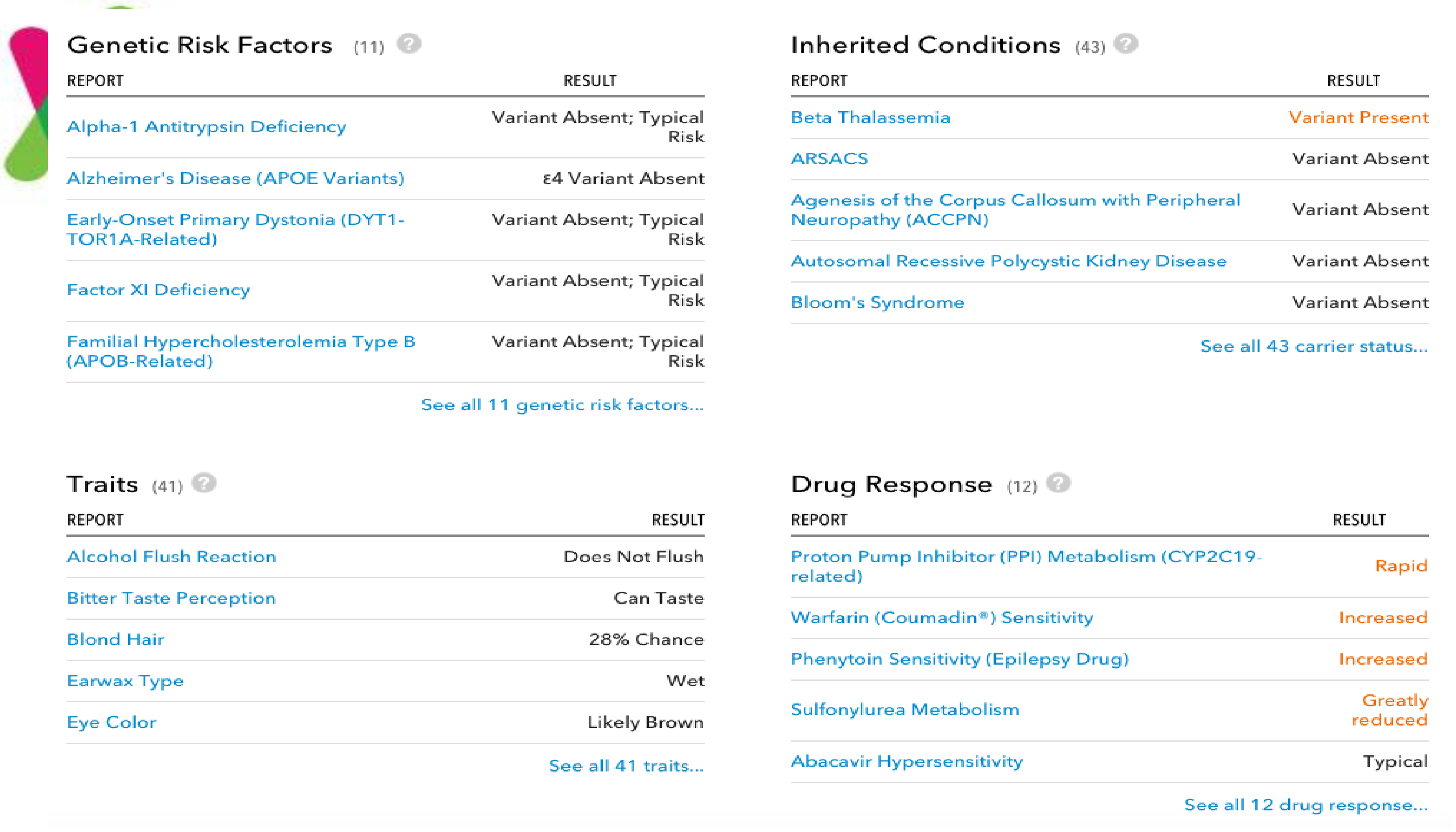
THE ANGELINA EFFECT

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



Genetic Risk Factors (11)

REPORT	RESULT
Alpha-1 Antitrypsin Deficiency	Variant Absent; Typical Risk
Alzheimer's Disease (APOE Variants)	ε4 Variant Absent
Early-Onset Primary Dystonia (DYT1-TOR1A-Related)	Variant Absent; Typical Risk
Factor XI Deficiency	Variant Absent; Typical Risk
Familial Hypercholesterolemia Type B (APOB-Related)	Variant Absent; Typical Risk

See all 11 genetic risk factors...

Traits (41)

REPORT	RESULT
Alcohol Flush Reaction	Does Not Flush
Bitter Taste Perception	Can Taste
Blond Hair	28% Chance
Earwax Type	Wet
Eye Color	Likely Brown

See all 41 traits...

Inherited Conditions (43)

REPORT	RESULT
Beta Thalassemia	Variant Present
ARSACS	Variant Absent
Agenesis of the Corpus Callosum with Peripheral Neuropathy (ACCPN)	Variant Absent
Autosomal Recessive Polycystic Kidney Disease	Variant Absent
Bloom's Syndrome	Variant Absent

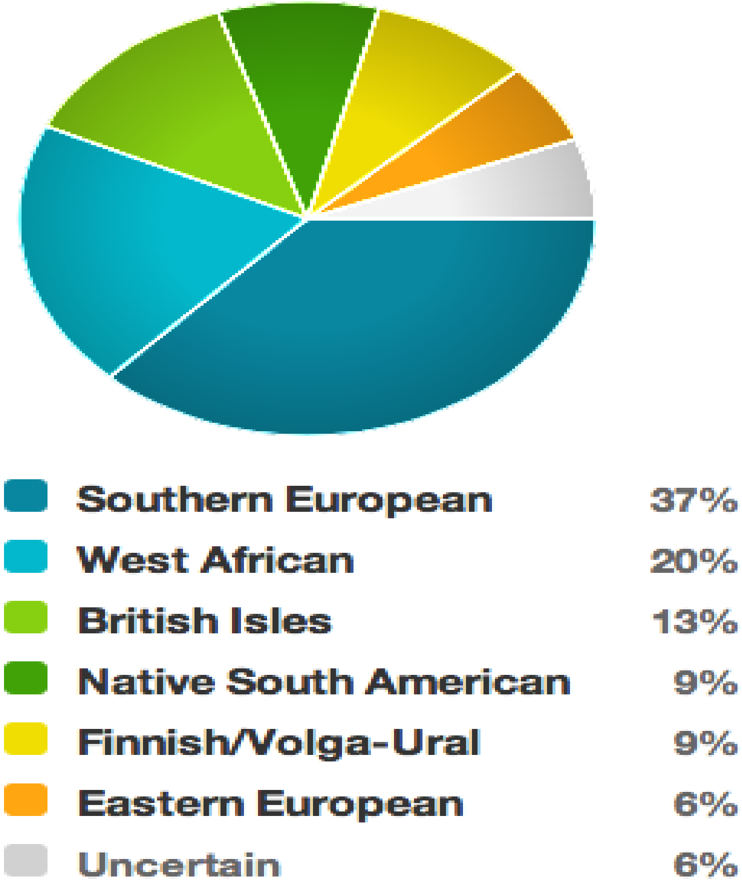
See all 43 carrier status...

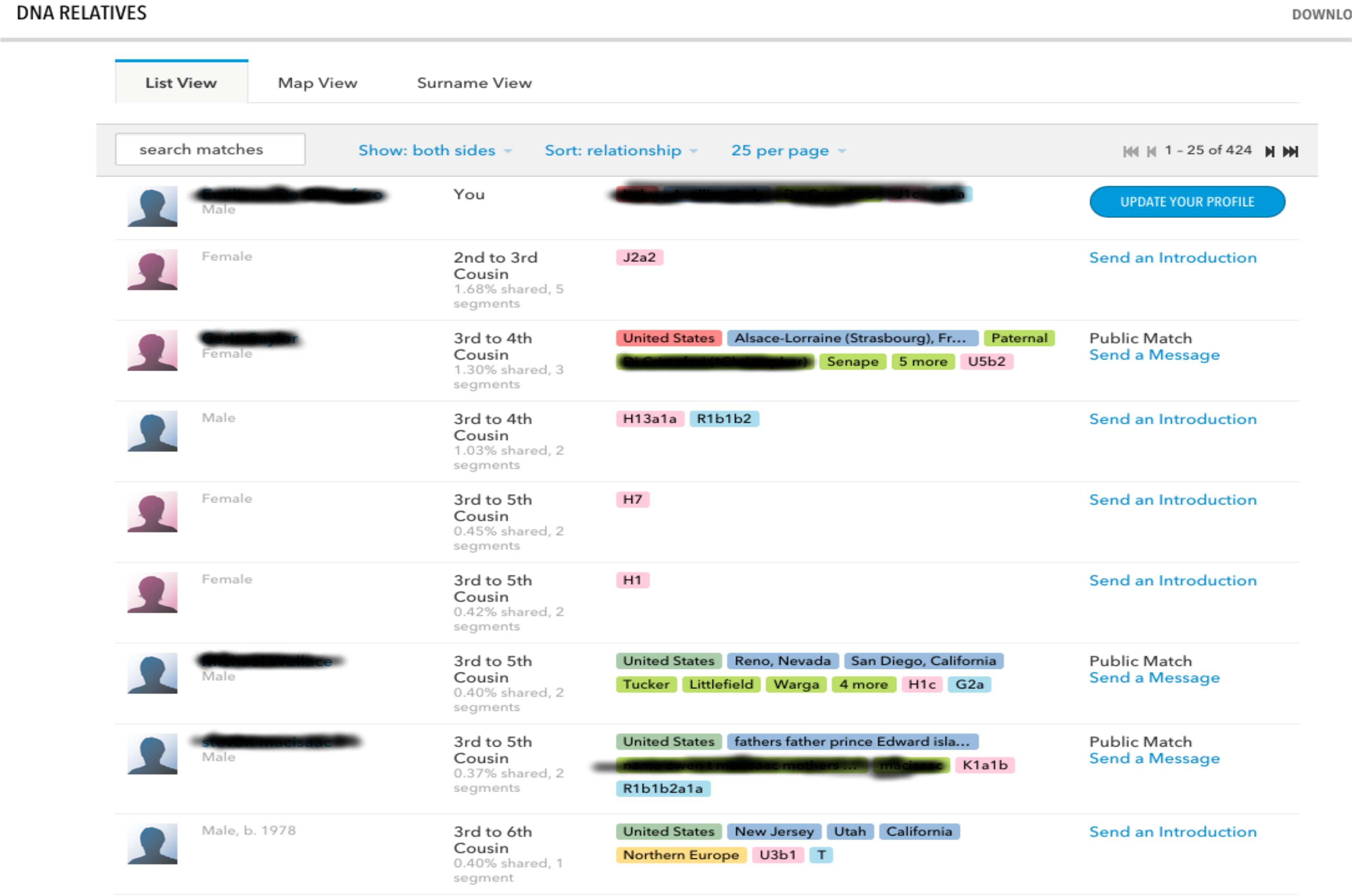
Drug Response (12)

REPORT	RESULT
Proton Pump Inhibitor (PPI) Metabolism (CYP2C19-related)	Rapid
Warfarin (Coumadin®) Sensitivity	Increased
Phenytoin Sensitivity (Epilepsy Drug)	Increased
Sulfonylurea Metabolism	Greatly reduced
Abacavir Hypersensitivity	Typical

See all 12 drug response...

Genetic Ethnicity





DNA RELATIVES

DOWNLO

List View | Map View | Surname View

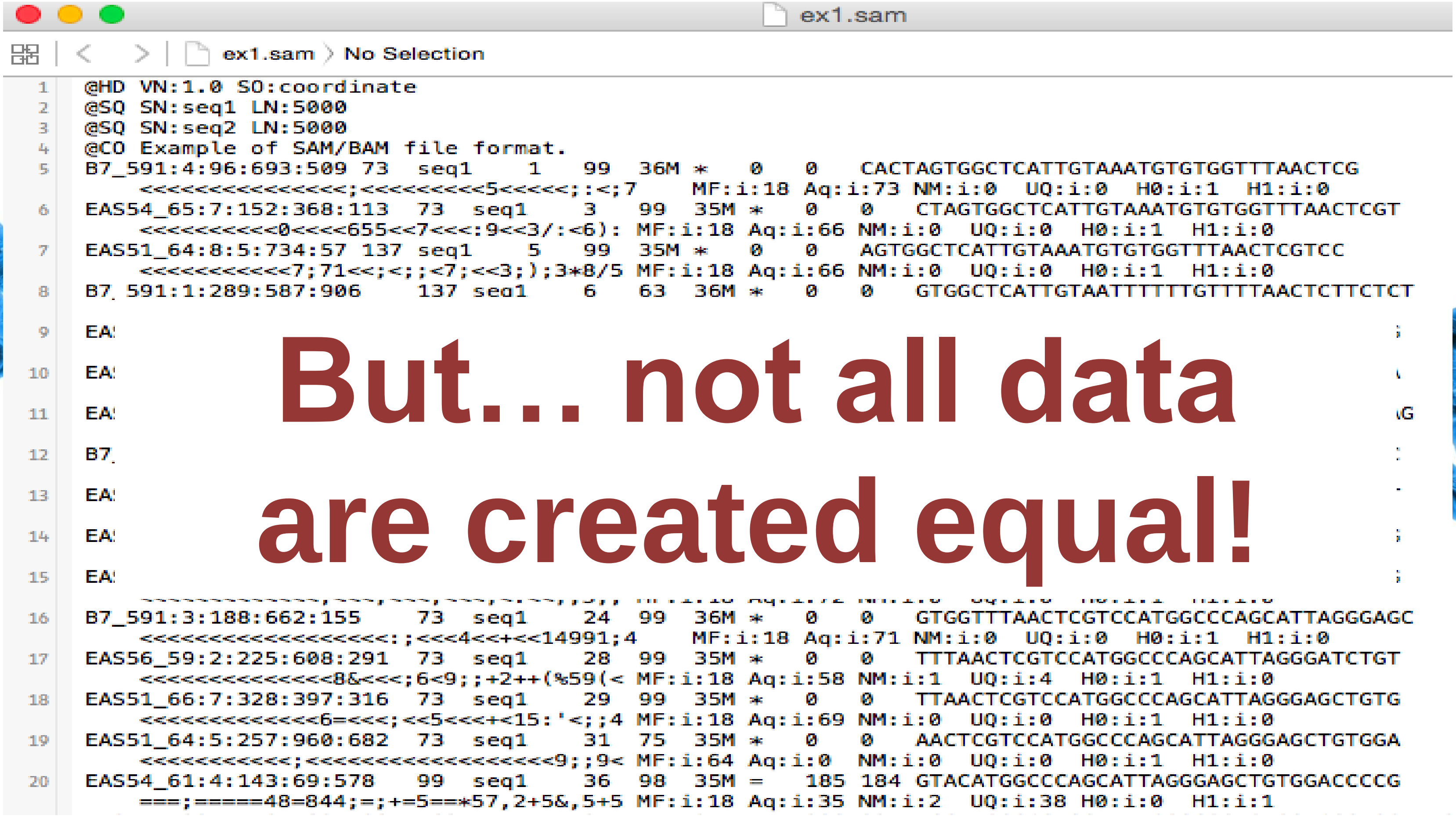
search matches

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

But… not all data are created equal!



```
@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

EA
EA
EA
B7
EA
EA
EA
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
```

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

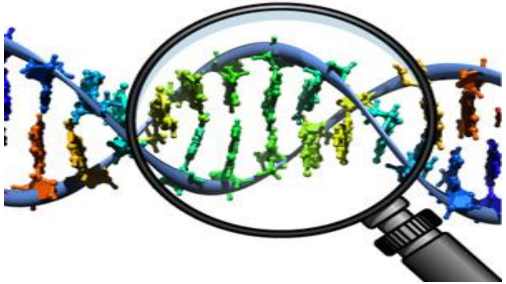
The Greater Good

vs

Privacy?

A New Research Community

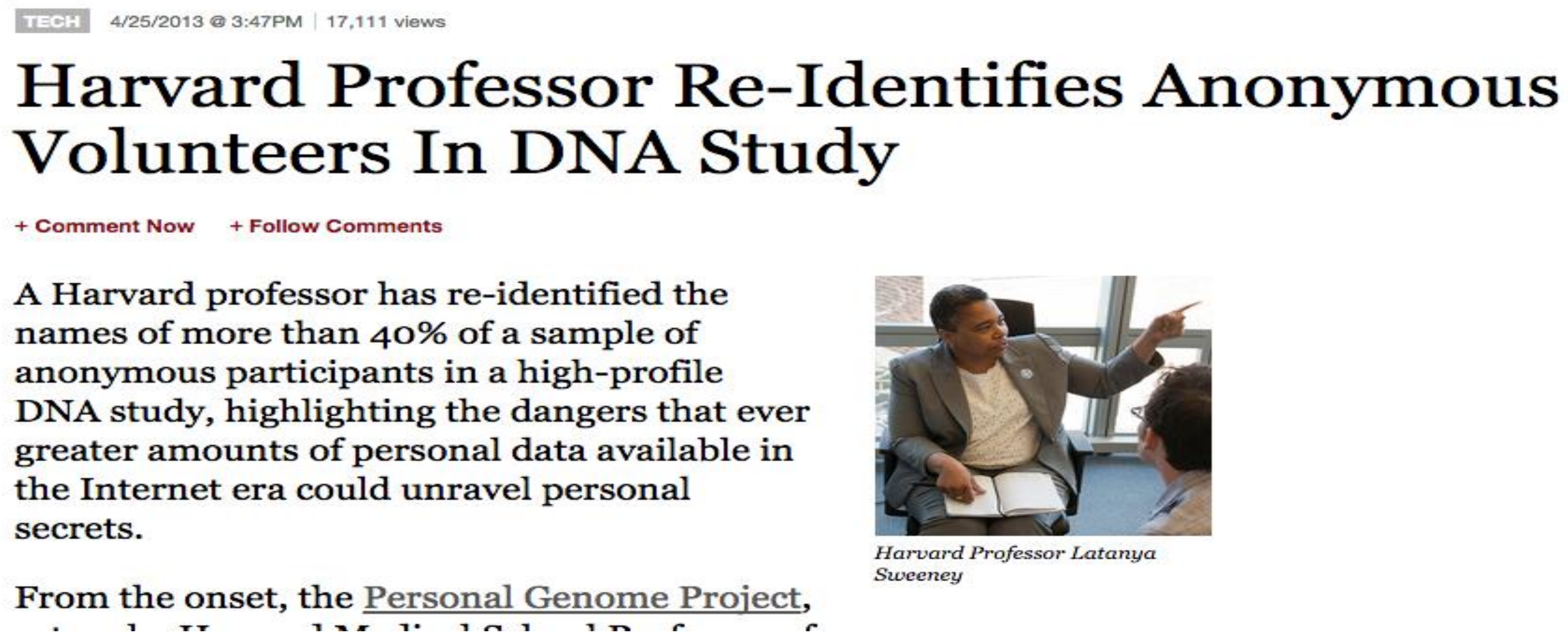
Studying privacy issues



Crypto tools to protect privacy

http://genomeprivacy.org

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.

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

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

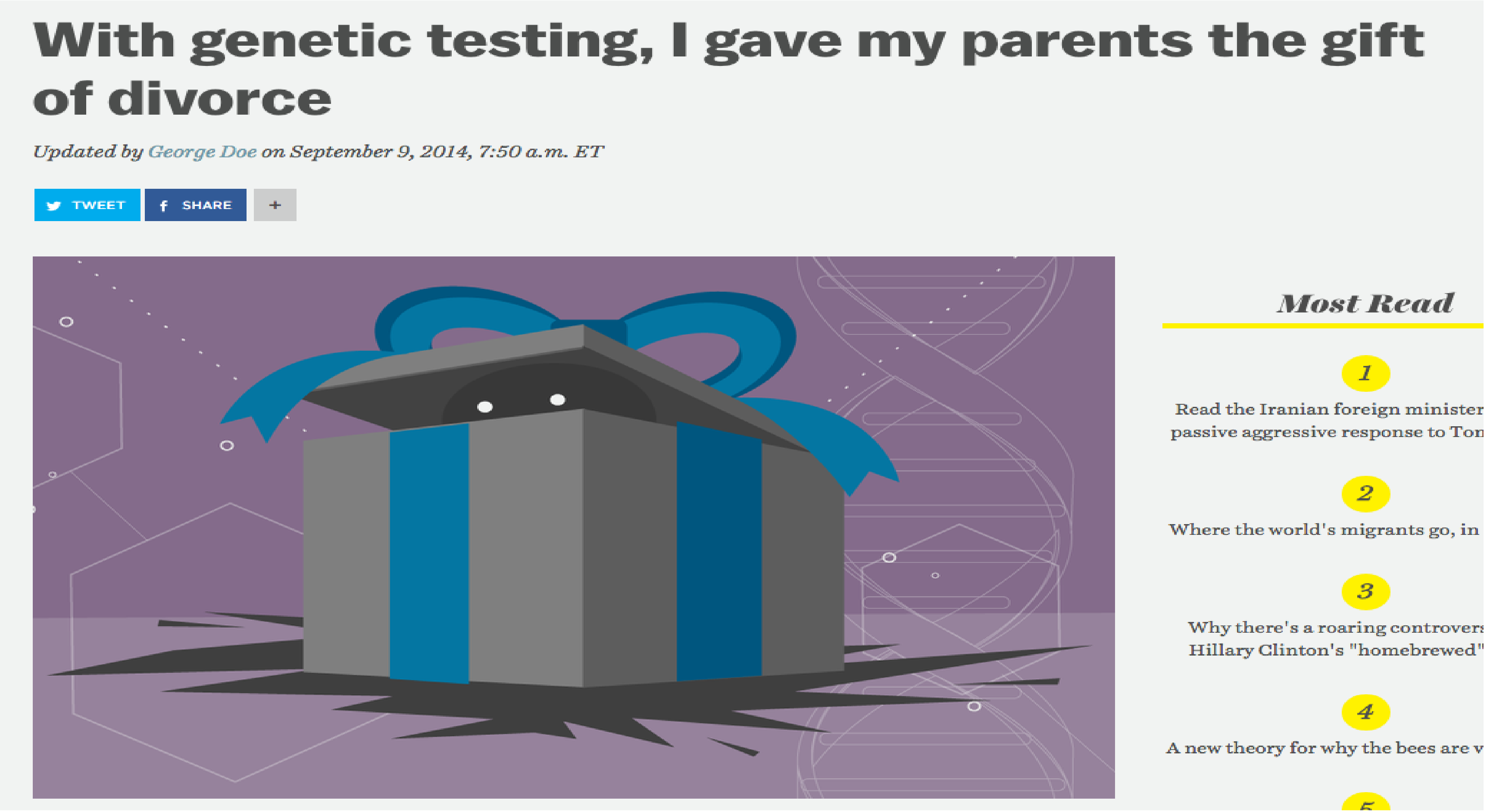
[2] N. Homer et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics,2008

Kin Privacy

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

Also read: Ayday, De Cristofaro, Hubaux, Tsudik. "Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare?"

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

Most Read

1

Read the Iranian foreign minister
passive aggressive response to Tor

2

Where the world's migrants go, in

3

Why there's a roaring controvers
Hillary Clinton's "homebrewed"

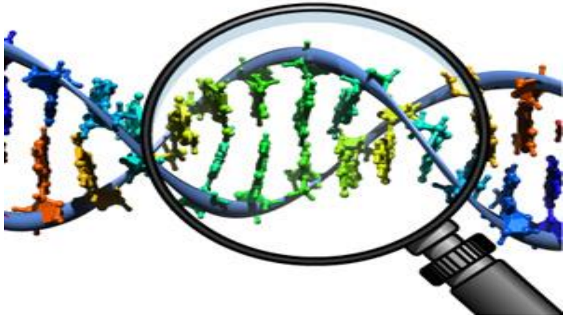
4

A new theory for why the bees are v

5

The rise of a new research community

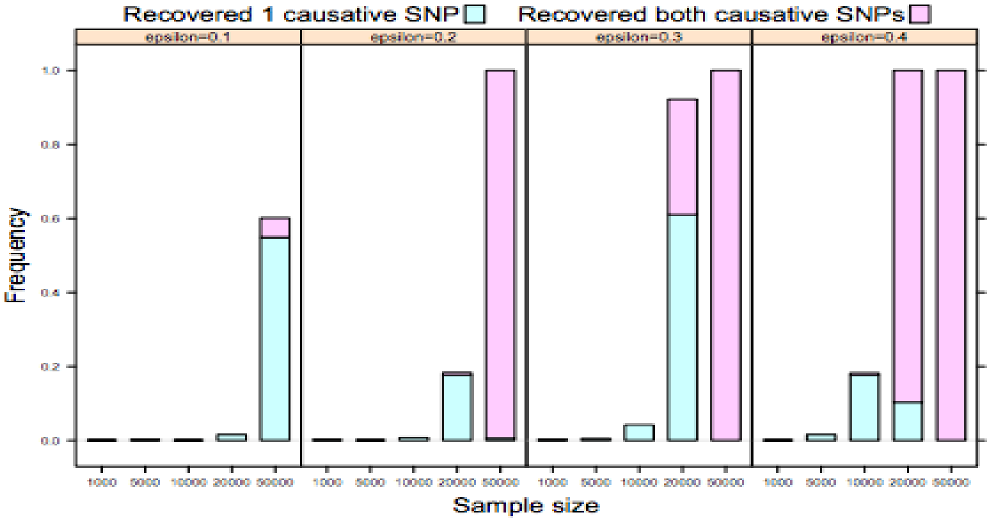
Studying privacy issues



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

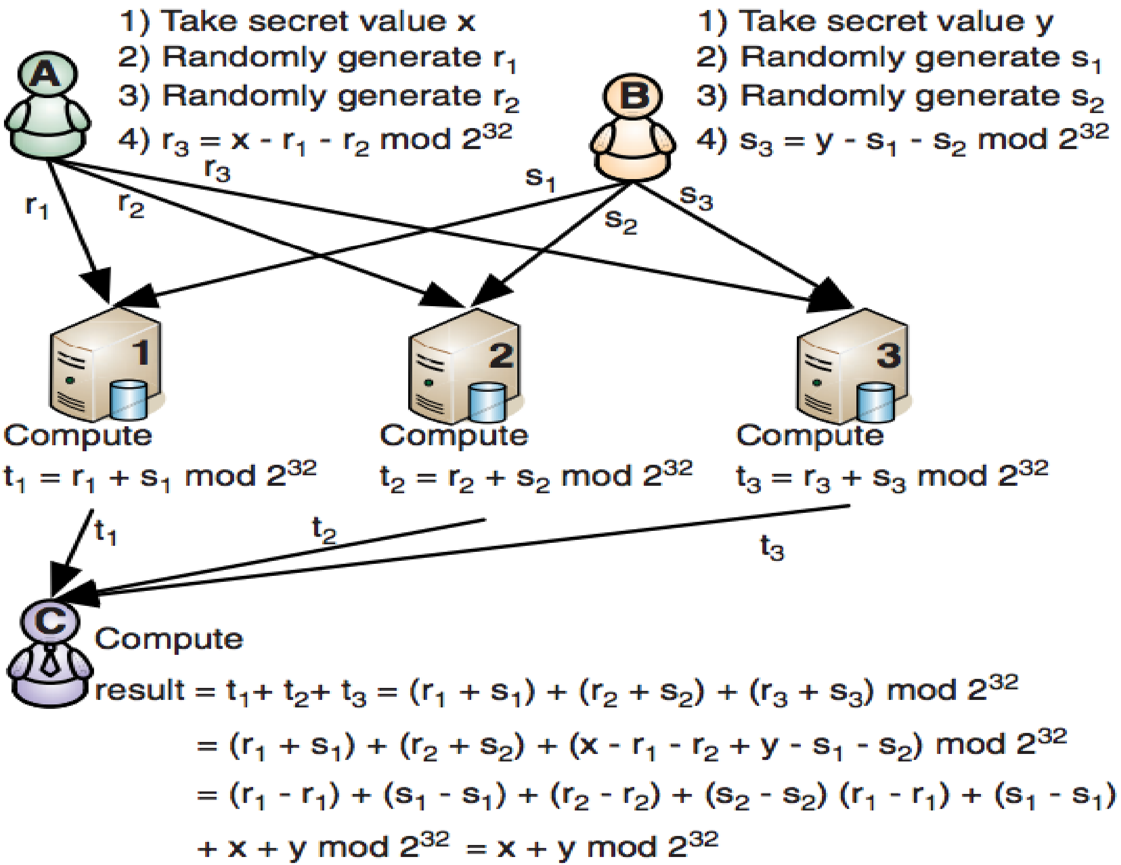
- Perform private computation over encrypted data
- Using partial & fully homomorphic encryption

Examples:

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

K. Lauter, A. Lopez-Alt, M. Naehrig. Private Computation on Encrypted Genomic Data

Computing on Encrypted Genomes



1) Take secret value x
2) Randomly generate r_1
3) Randomly generate r_2
4) $r_3 = x - r_1 - r_2 \bmod 2^{32}$

1) Take secret value y
2) Randomly generate s_1
3) Randomly generate s_2
4) $s_3 = y - s_1 - s_2 \bmod 2^{32}$

Compute
$t_1 = r_1 + s_1 \bmod 2^{32}$

Compute
$t_2 = r_2 + s_2 \bmod 2^{32}$

Compute
$t_3 = r_3 + s_3 \bmod 2^{32}$

Compute

$$\begin{aligned} \text{result} &= t_1 + t_2 + t_3 = (r_1 + s_1) + (r_2 + s_2) + (r_3 + s_3) \bmod 2^{32} \\ &= (r_1 + s_1) + (r_2 + s_2) + (x - r_1 - r_2 + y - s_1 - s_2) \bmod 2^{32} \\ &= (r_1 - r_1) + (s_1 - s_1) + (r_2 - r_2) + (s_2 - s_2)\ (r_1 - r_1) + (s_1 - s_1) \\ &+ x + y \bmod 2^{32} = x + y \bmod 2^{32} \end{aligned}$$

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

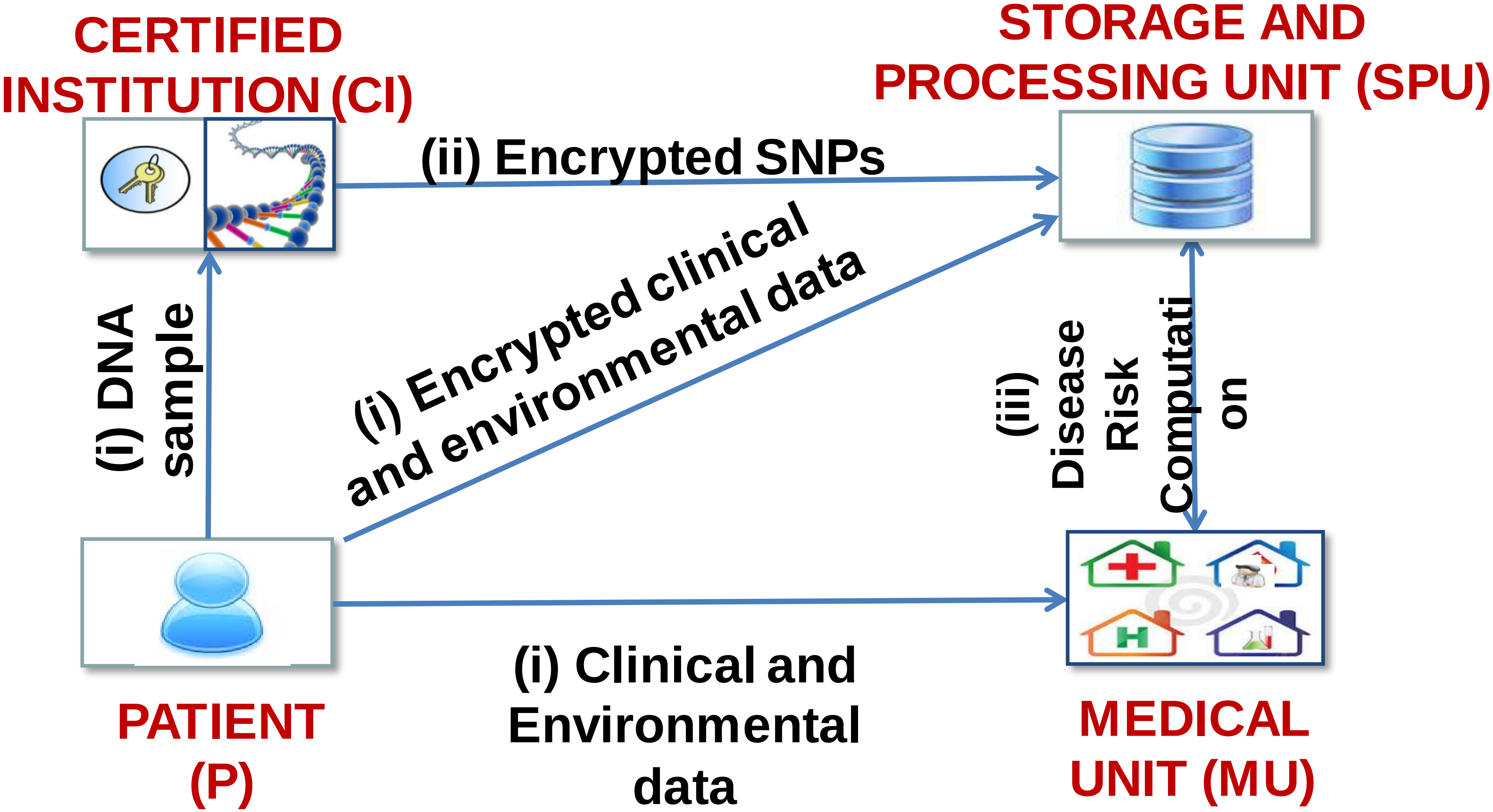
Individuals retain control of their sequenced genome

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



CERTIFIED INSTITUTION (CI)

(ii) Encrypted SNPs

STORAGE AND PROCESSING UNIT (SPU)

(i) DNA sample

(i) Encrypted clinical and environmental data

(iii) Disease Risk Computati
on

PATIENT (P)

(i) Clinical and Environmental data

MEDICAL UNIT (MU)

1. Using Semi-Trusted Parties

Ayday et al. (WPES’13)

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

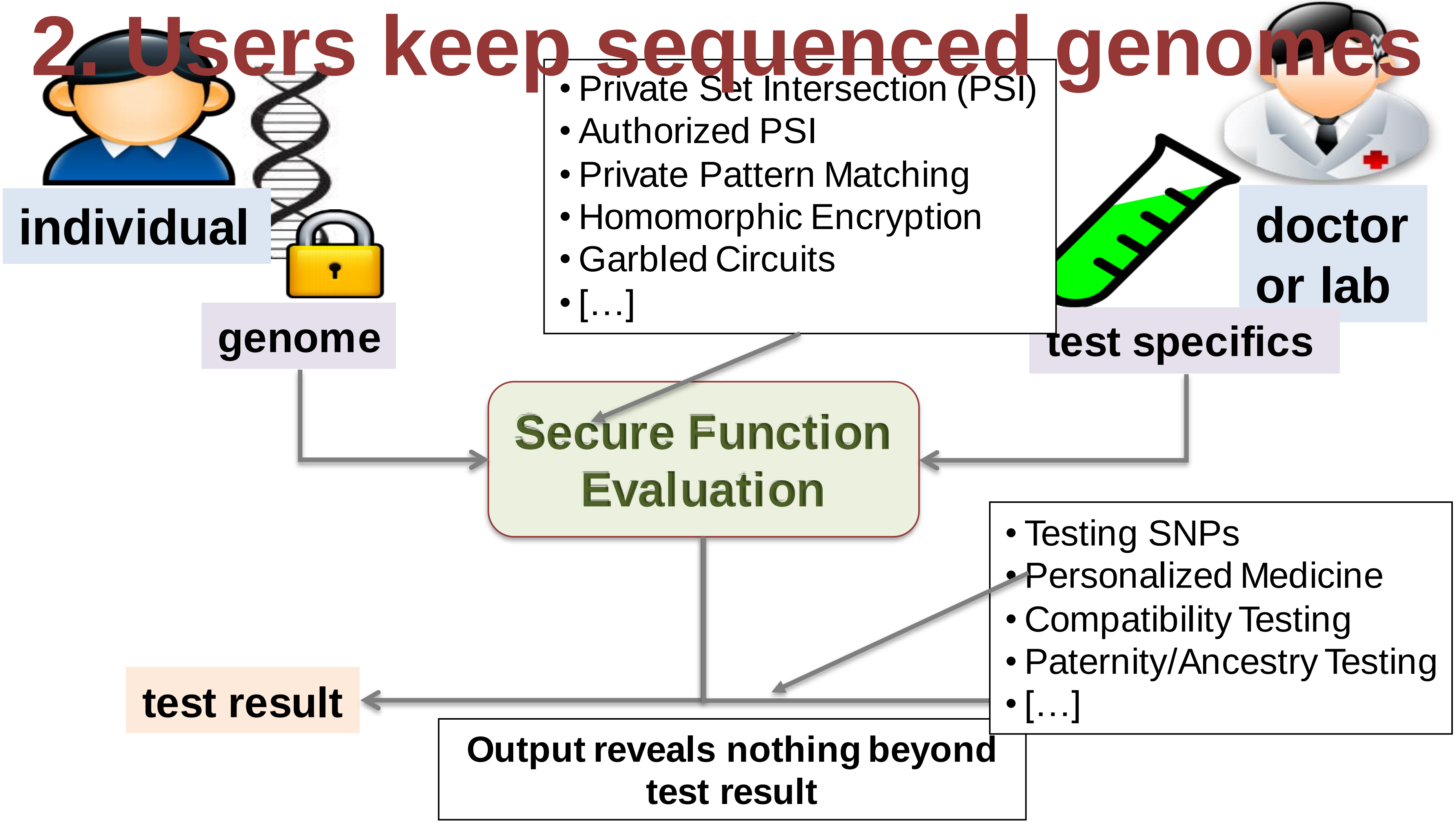
Ayday et al. (DPM’13)

- 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 Circuits
- […]

doctor or lab

test specifics

Secure Function Evaluation

- Testing SNPs
- Personalized Medicine
- Compatibility Testing
- Paternity/Ancestry Testing
- […]

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
- Dealing with sequencing errors
- How much understanding required from users?

parc®
A Xerox Company

MARIE CURIE ACTIONS

Google
Google
Faculty Research Awards

Thank you!

Special thanks to

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

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 a DTC’s DB…

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

The latter: cheaper, more *scalable*

Milestones

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)