Genome Privacy: A Computer Scientist's Perspective

Emiliano De Cristofaro
http://emilianodc.com

Genomics: A CS Perspective

ex1.sam

ex1.sam 〉 No Selection

```
1  @HD VN:1.0 SO:coordinate
2  @SQ SN:seq1 LN:5000
3  @SQ SN:seq2 LN:5000
4  @CO Example of SAM/BAM file format.
5  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
6  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
7  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
8  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
9  EAS56_59:8:38:671:758   137 seq1    9   99  35M *   0   0   GCTCATTGTAAATGTGTGGTTTAACTCGTCCATGG
       <<<<<<<<<<<<<<;<;7<<<<<<<<7<<;:<5% MF:i:18 Aq:i:72 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
10 EAS56_61:6:18:467:281   73  seq1    13  99  35M *   0   0   ATTGTAAATGTGTGGTTTAACTCGTCCCTGGCCCA
       <<<<<<<<;<<<8<<<<<;8:;6/686&;(16666 MF:i:18 Aq:i:39 NM:i:1  UQ:i:5  H0:i:0  H1:i:1
11 EAS114_28:5:296:340:699 137 seq1    13  99  36M *   0   0   ATTGTAAATGTGTGGTTTAACTCGTCCATGGCCCAG
       <<<<<;<<<;<;<<<<<<<<<<8<8<3<8;<;<0;    MF:i:18 Aq:i:73 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
12 B7_597:6:194:894:408    73  seq1    15  99  35M *   0   0   TGTAAATGTGTGGTTTAACTCGTCCATTGCCCAGC
       <<<<<<<<<7<<;<<<<;<<<7;;<<<*,;;572< MF:i:18 Aq:i:43 NM:i:1  UQ:i:9  H0:i:0  H1:i:1
13 EAS188_4:8:12:628:973   89  seq1    18  75  35M *   0   0   AAATGTGTGGTTTAACTCGTCCATGGCCCAGCATT
       ==;=:;:;:;:====;=;===:=======;==;=== MF:i:64 Aq:i:0  NM:i:0  UQ:i:0  H0:i:1  H1:i:0
14 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
15 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
16 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
17 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
18 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
19 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
20 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
```

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

Ultimate identifier

- Hard to anonymize / de-identify
- Once leaked, we cannot "revoke" it

Extremely sensitive information

- Ethnic heritage, predisposition to diseases
- Leaking one's genome ≈ leaking relatives' genome*

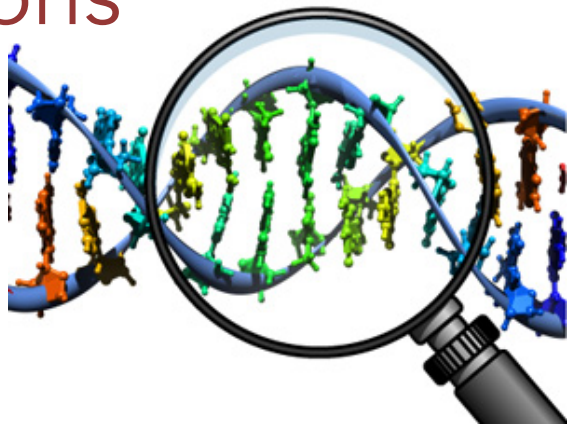
Sensitivity of the genome is (almost) perpetual

- Long after owner's death

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

The rise of a new research community

Studying the privacy implications



Exploring techniques to protect privacy

Studying Privacy

Re-identification of anonymous DNA donors

Infer surnames using (public) information available from popular genealogy sites*

* 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?

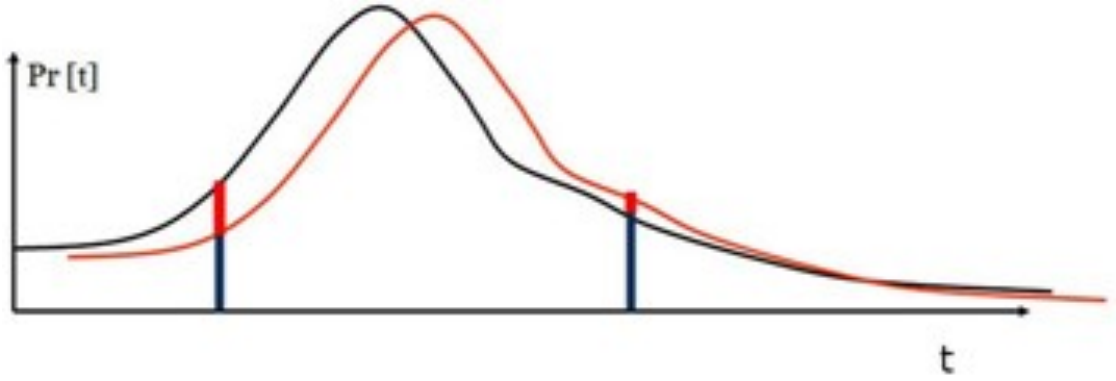
Even statistics from allele frequencies can be used to identify genetic trial participants

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

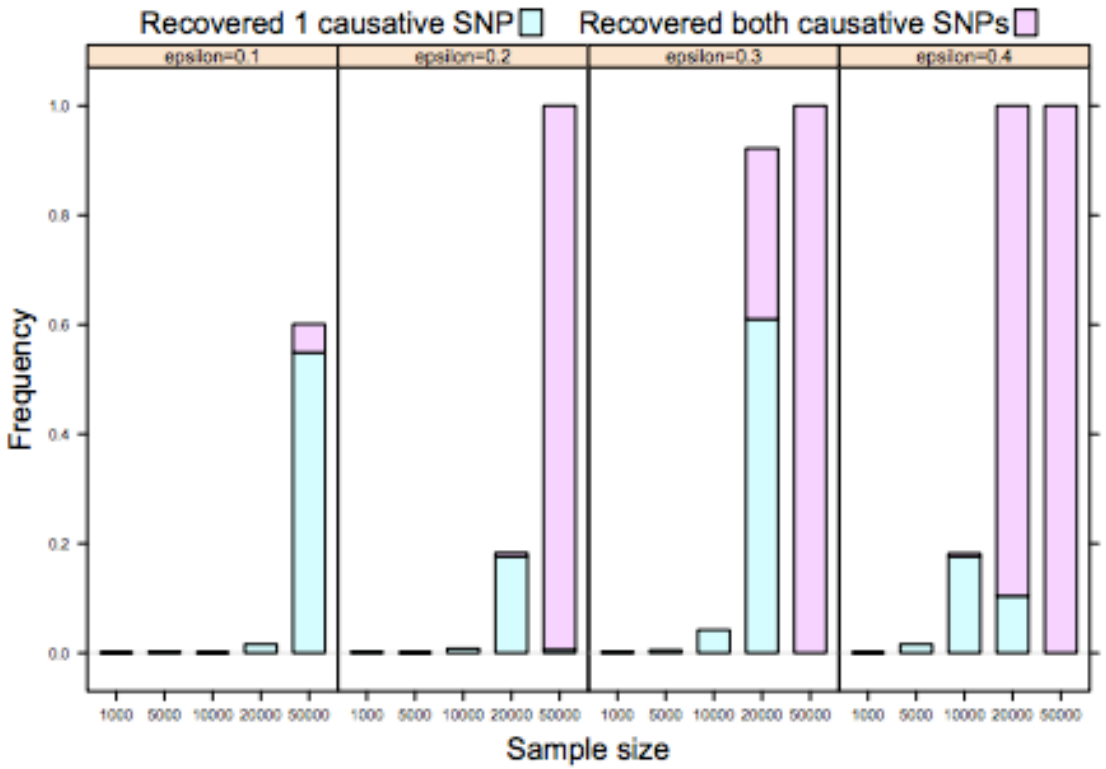
Differential Privacy



Maximizing the accuracy of queries from statistical databases

Minimizing the chances of identifying its records

Differential Privacy



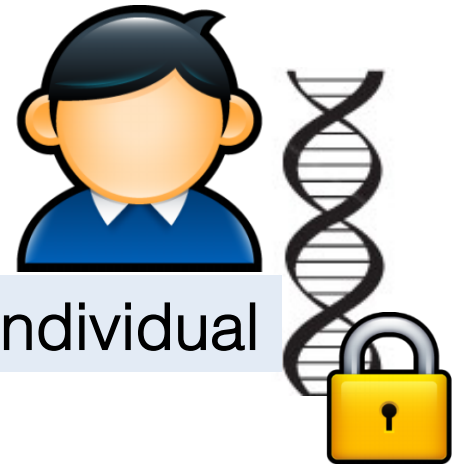
Supporting Genome Wide Association Studies (GWAS)

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

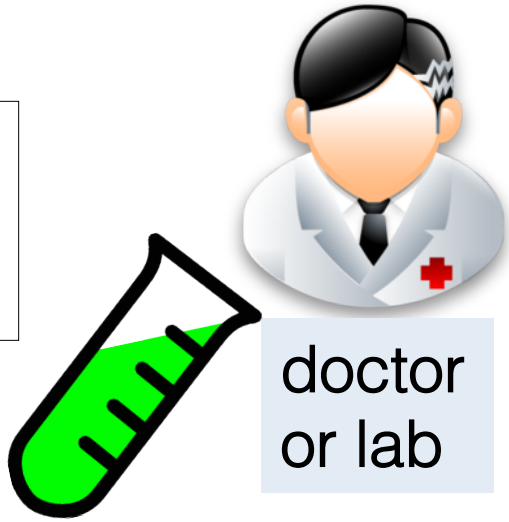
Privacy-Friendly Personal Genomics



individual

genome

- Private Set Intersection (PSI)
- Authorized PSI
- Cardinality-Only PSI
- […]



doctor or lab

test specifics

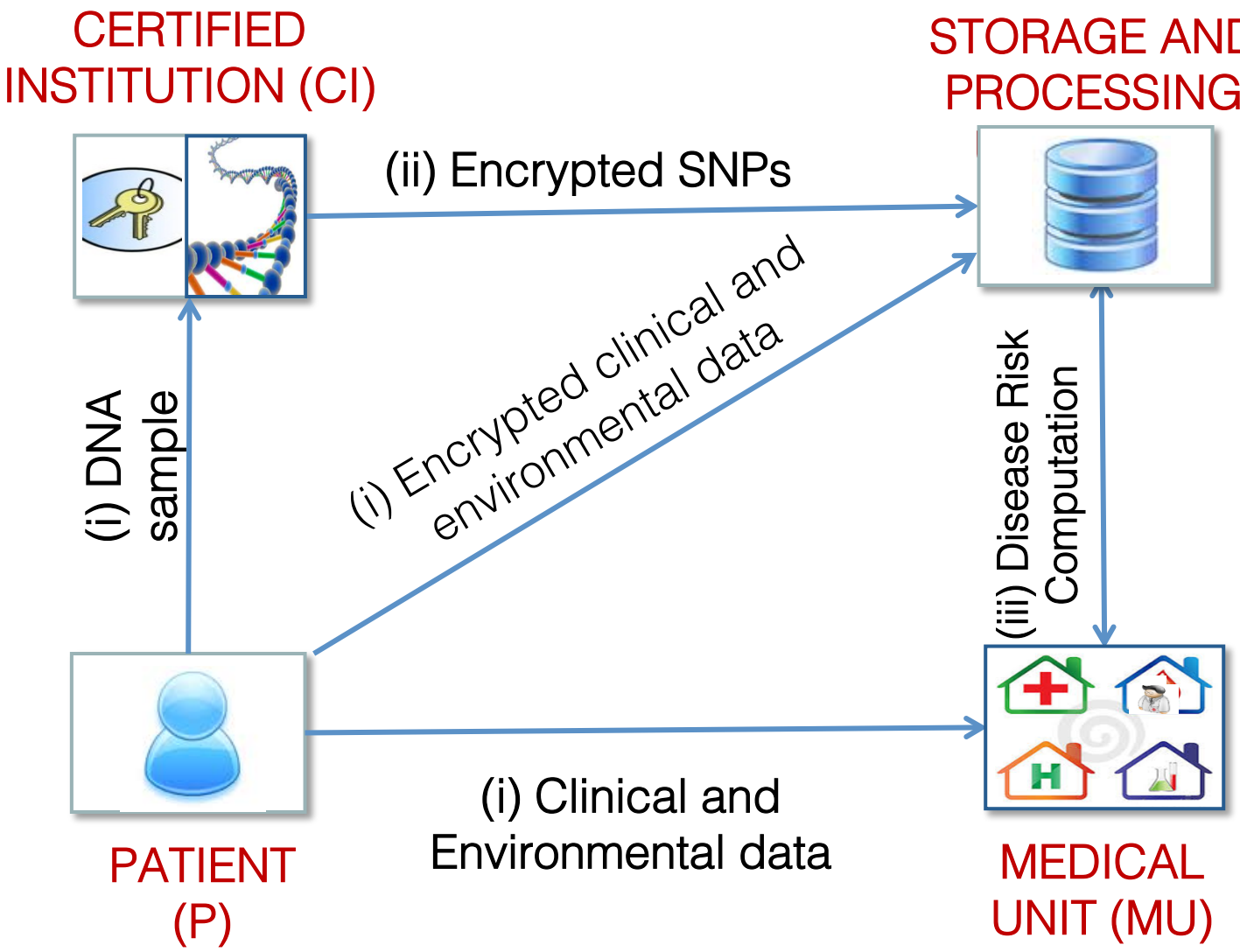
Secure Function Evaluation

- 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)

Open Problems

Guarantee security past 30 years?

User's involvement? Usability?

Computation overhead?

Data representation assumptions?

2nd International Workshop on Genome Privacy and Security

(GenoPri 2015)

May 21, San Jose, CA

http://genopri.org

Submission deadline: January 25

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

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!