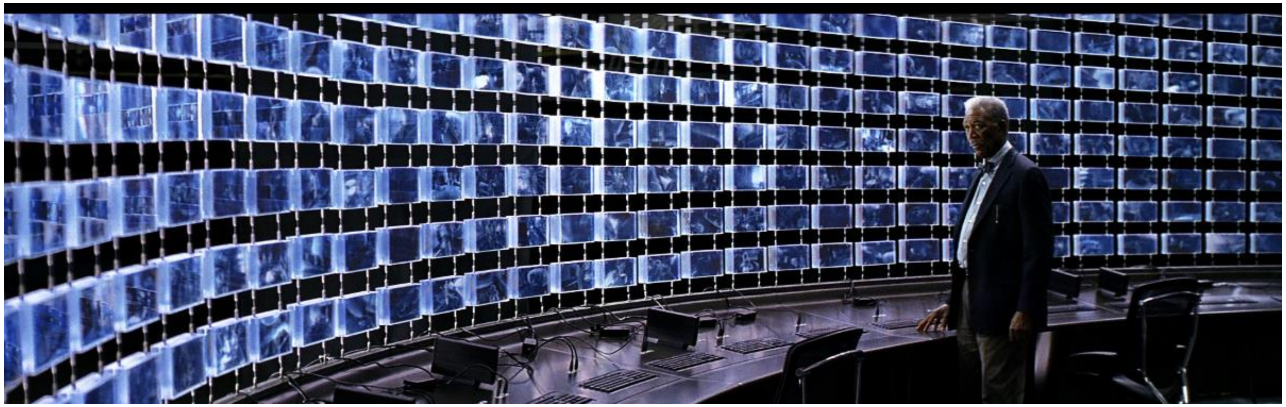
Security and Privacy in the Age of Big Data: The Case of Genomics

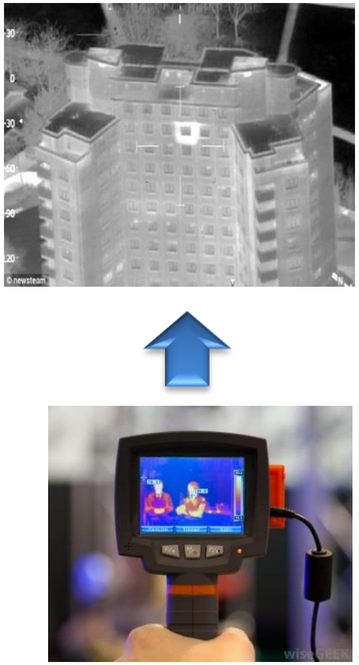
Erman Ayday

Decreasing Privacy - Wholesale Surveillance



Cellular phones

Social networks

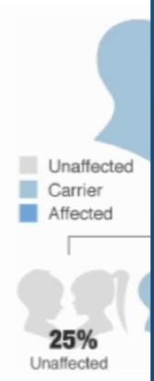


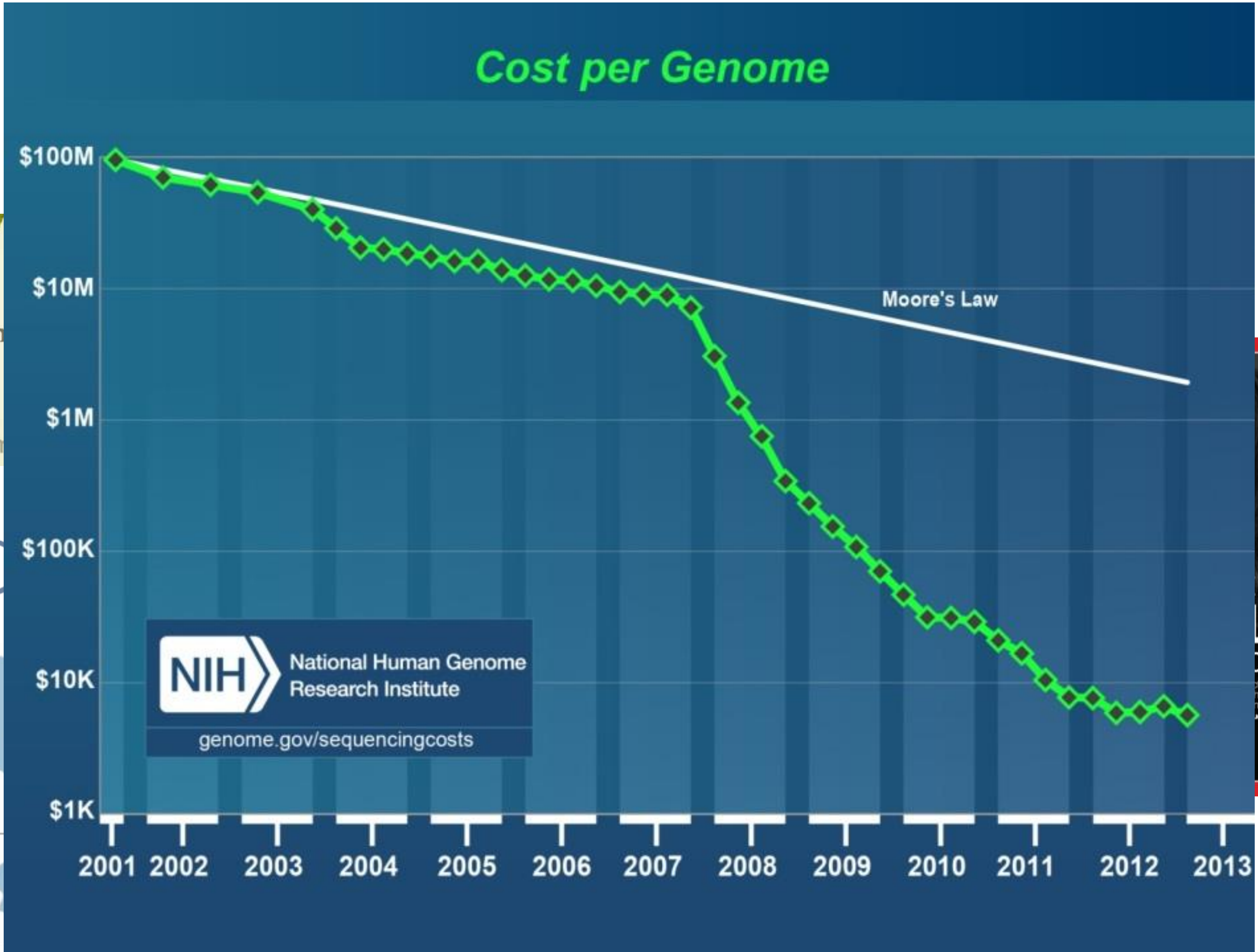
Heat sensors

Traffic cameras

Genomic data

Significance and Popularity of Genomic Data





Why Protect Genomic Data?

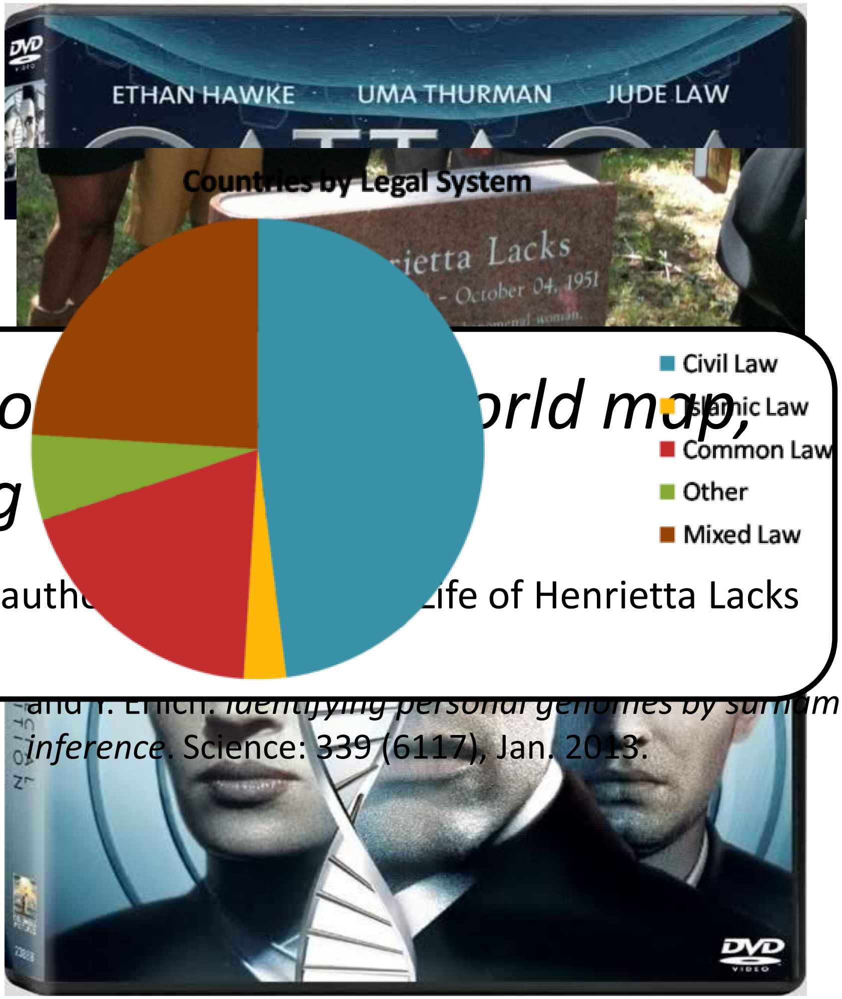
- Genome carries information about a person’s genetic condition and predispositions to specific diseases
  - Leakage of such information could cause *genetic discrimination*
  - Denial of access to health insurance,

“The view we have today of geno *orld map, but Google Street View is coming*

Rebecca Skloot, the auth ife of Henrietta Lacks

Using privacy-sensitive information belonging to a victim retrieved from different sources

- Genomic data is non-revokable
- Law is not universal and hard to enforce



and Y. Erlich. *Identifying personal genomes by surname inference*. Science: 339 (6117), Jan. 2013.

“The Chills and Thrills of Whole Genome Sequencing”

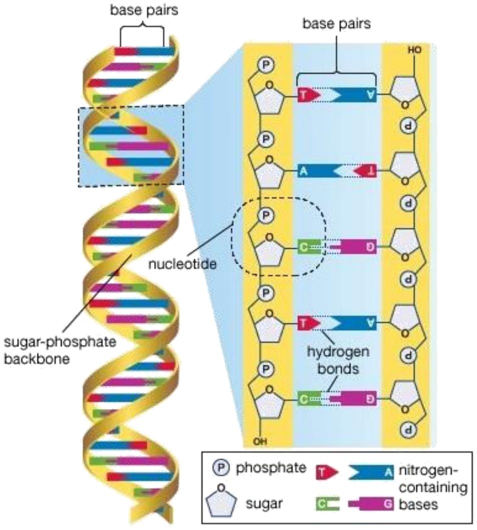
E. Ayday, E. De Cristofaro, J.P. Hubaux, G. Tsudik

Some of Our Contributions

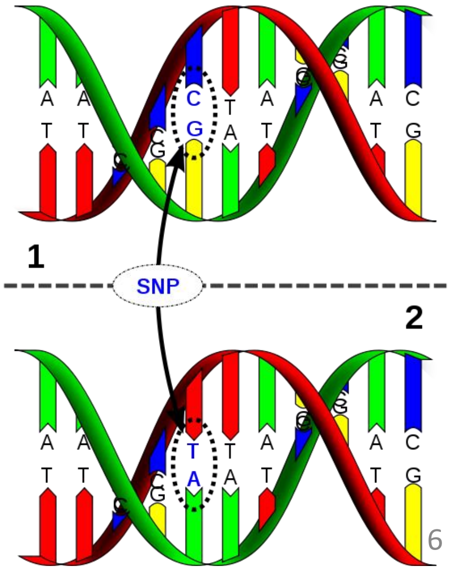
- Inference Attacks and Quantifying Privacy
  - Metrics and methods to infer genomic data
  - Quantifying kin genomic privacy
  - Quantifying genomic privacy in genetic tests
- Protecting Genomic Privacy
  - Computational privacy
    - Applied cryptographic techniques for usable privacy
  - Information theoretical privacy
    - GeneVault via HoneyEncryption
    - Efficient non-cryptographic techniques
- Interdependent Genomic Privacy

Genomics 101 - DNA and SNP

- The human genome consists of approximately 3 billion letters
  - 99.9% is identical between any two individuals
  - Remaining: human genetic variation
- Single Nucleotide Polymorphism (SNP): Most common human genetic variation.
  - A single nucleotide (A, C, G, or T) differs between members of the same species or paired chromosomes of an individual
  - Disease risk can be computed by analyzing particular SNPs
    - Angelina Jolie BRCA1 Mutation
    - 23andMe genetic disease risk tests

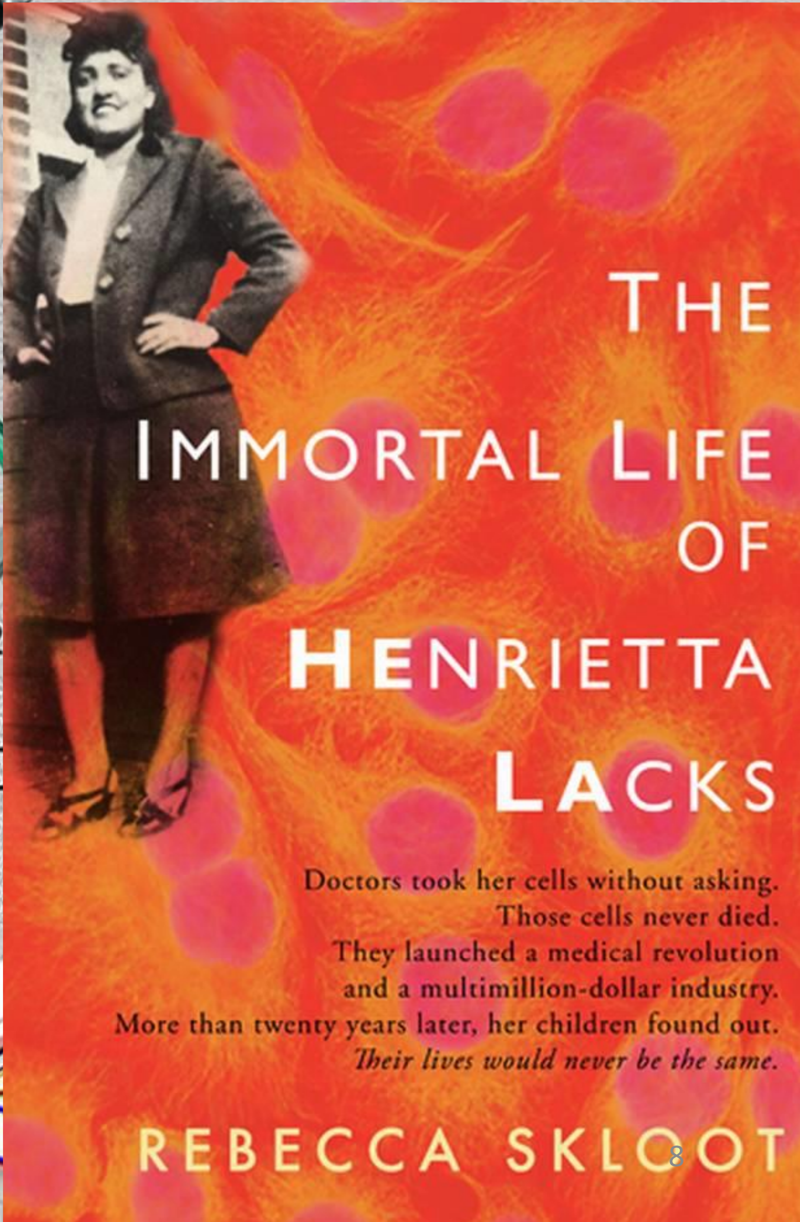


© 2007 Encyclopædia Britannica, Inc.



INFERENCE ATTACKS AND QUANTIFYING GENOMIC PRIVACY

etta Lacks' death
uman cell line
V role in
ical cancer
Telemorase activity
1989
2
> 60,000
research papers
HIV id
Po
U 53
HENRIETTA LACKS
(1920-1951)
Born in Roanoke on 1 Aug. 1920, Henrietta
Pleasant lived here with relatives after her
mother's 1924 death. She married David Lacks
in 1941 and, like many other African Americans,
moved to Baltimore, Md. for wartime employment.
She died of cervical cancer on 4 Oct. 1951.
tissue was removed without permission (as
then) for medical research. Her cells
plied and survived at an extraordinarily
rate, and are renowned worldwide as the
a line," the "gold standard" of cell lines.
Salk developed his polio vaccine with
Henrietta Lacks, who in death saved
lless lives, is buried nearby.

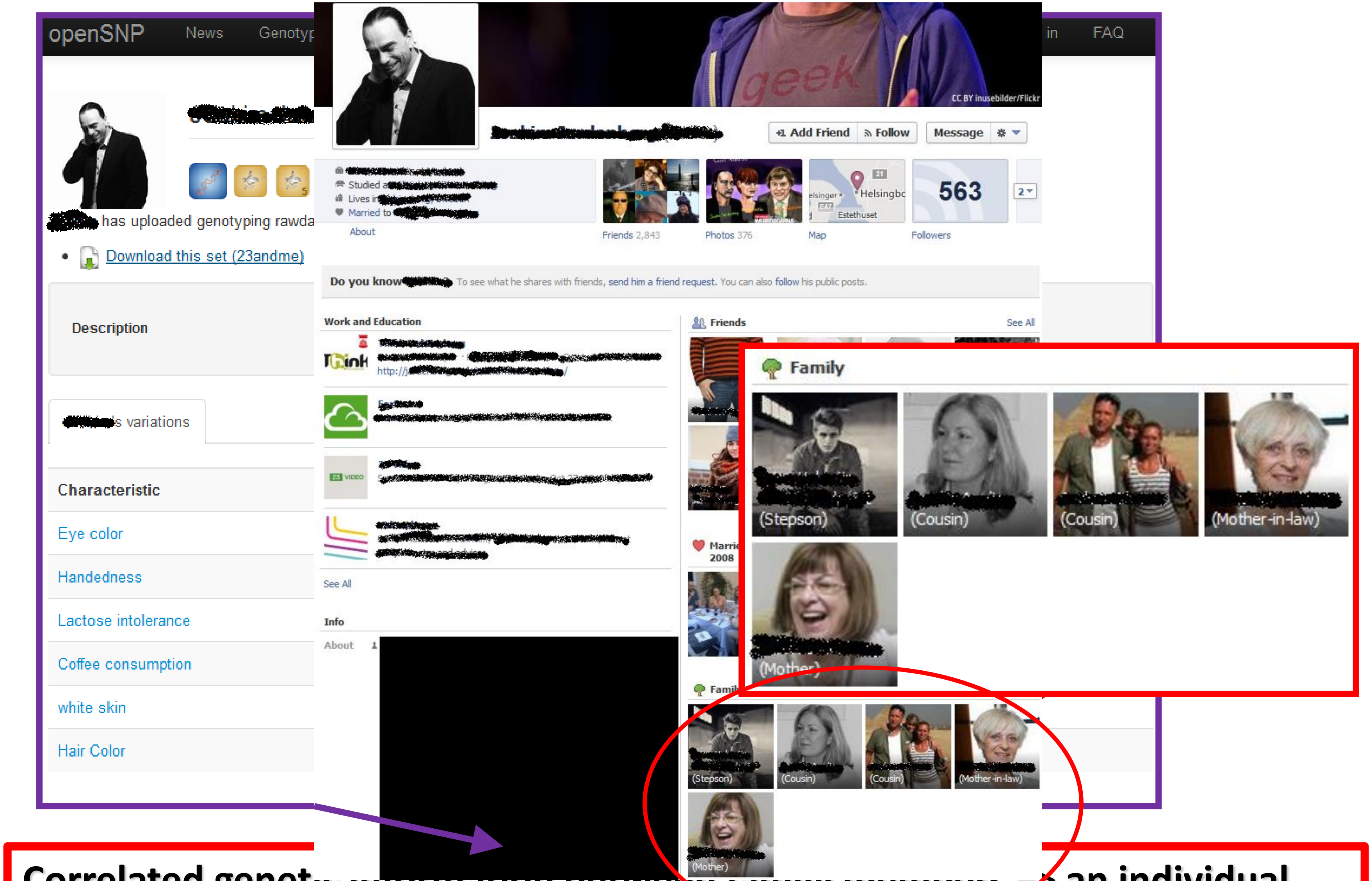


THE IMMORTAL LIFE OF HENRIETTA LACKS

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

Quantifying Kin Genomic Privacy



openSNP News Genotyp in FAQ

[redacted] has uploaded genotyping rawda

- Download this set (23andme)

Description

[redacted]s variations

Characteristic

Eye color

Handedness

Lactose intolerance

Coffee consumption

white skin

Hair Color

CC BY inusebilder/Flickr

[redacted]

Add Friend · Follow · Message

Studied a[redacted]

Lives i[redacted]

Married to [redacted]

About

Friends 2,843 · Photos 376 · Map · 563 Followers

Do you know [redacted]? To see what he shares with friends, send him a friend request. You can also follow his public posts.

Work and Education

[redacted]

See All

Info

About [redacted]

Friends See All

Married 2008

Family

[redacted] (Stepson) · [redacted] (Cousin) · [redacted] (Cousin) · [redacted] (Mother-in-law) · [redacted] (Mother)

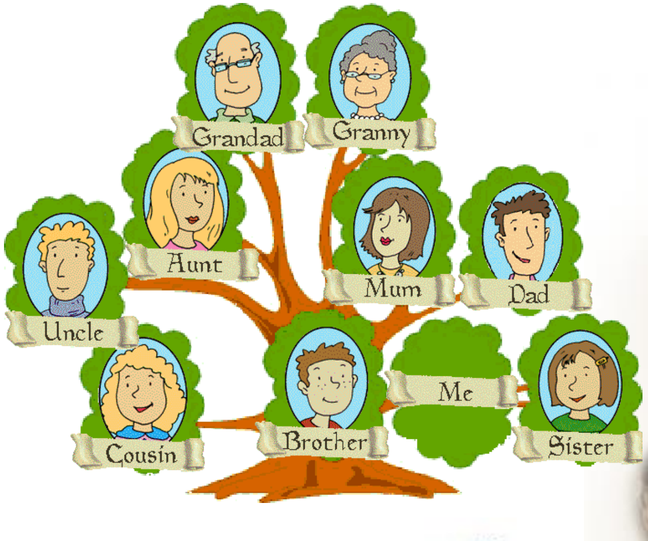
Family

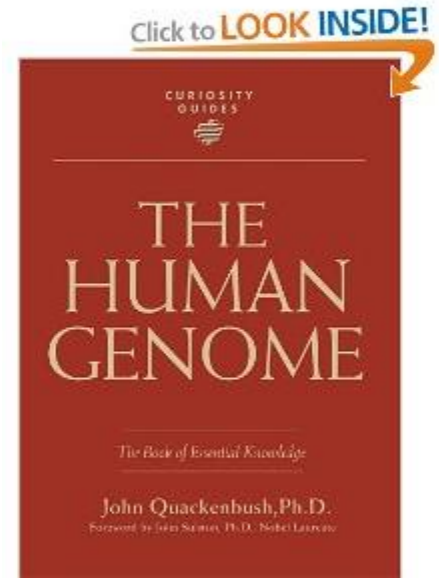
[redacted] (Stepson) · [redacted] (Cousin) · [redacted] (Cousin) · [redacted] (Mother-in-law) · [redacted] (Mother)

Correlated genetic information between family members => an individual sharing his/her genome threatens his (known) relatives' genomic privacy

How much can we infer about one’s genome?

Big Picture




Click to LOOK INSIDE!

CURIOSITY GUIDES

THE HUMAN GENOME

The Book of Essential Knowledge

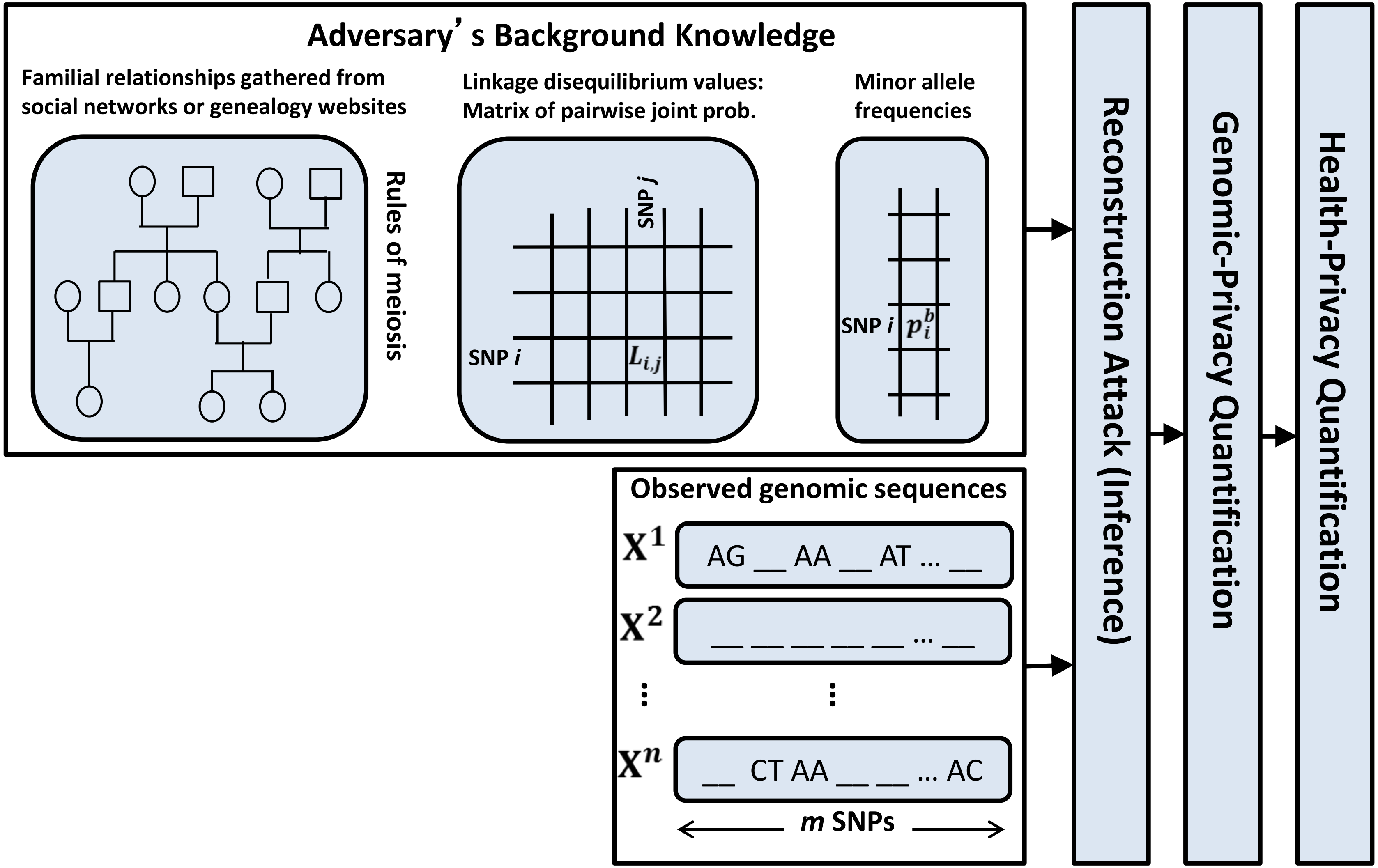
John Quackenbush, Ph.D.

Foreword by John Sulston, Ph.D., Nobel Laureate

- Given:
  - Family tree
  - (Partial) genomes of one or more family members
  - Public genomic knowledge
    - Minor allele frequencies
    - Linkage Disequilibrium
    - Reproduction

➢ (Probabilistically) infer the unrevealed genomes

Quantification and Protection Framework



Adversary's Background Knowledge

Familial relationships gathered from social networks or genealogy websites

Linkage disequilibrium values: Matrix of pairwise joint prob.

Minor allele frequencies

Observed genomic sequences

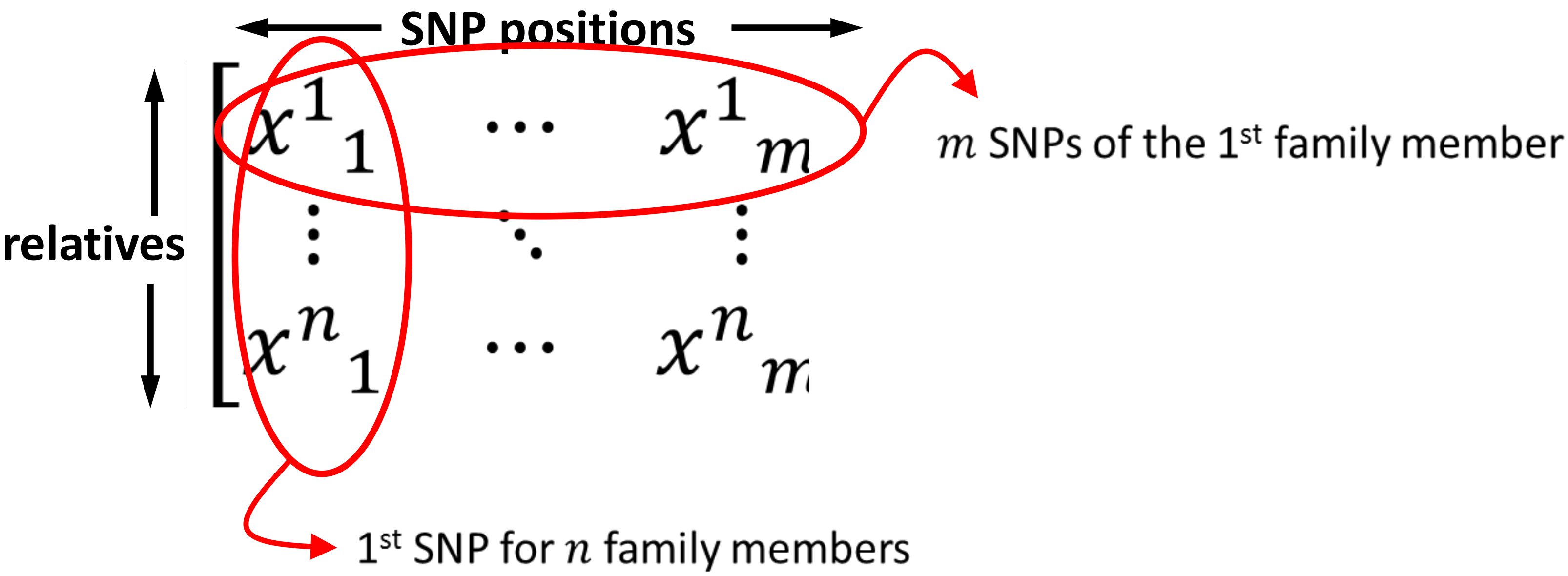
Reconstruction Attack (Inference)

Genomic-Privacy Quantification

Health-Privacy Quantification

Parameters

- m : Number of SNPs
- n : Number of family members
- ${x^i}_j$: Value of SNP j for individual i
- ${x^i}_j \in \{0,1,2\}$
- $\mathbb{X}$: $m \times n$ matrix that stores the SNPs of all family members



SNP positions

relatives

$$\begin{bmatrix} {x^1}_1 & \cdots & {x^1}_m \\ \vdots & \ddots & \vdots \\ {x^n}_1 & \cdots & {x^n}_m \end{bmatrix}$$

m SNPs of the 1st family member

1st SNP for n family members

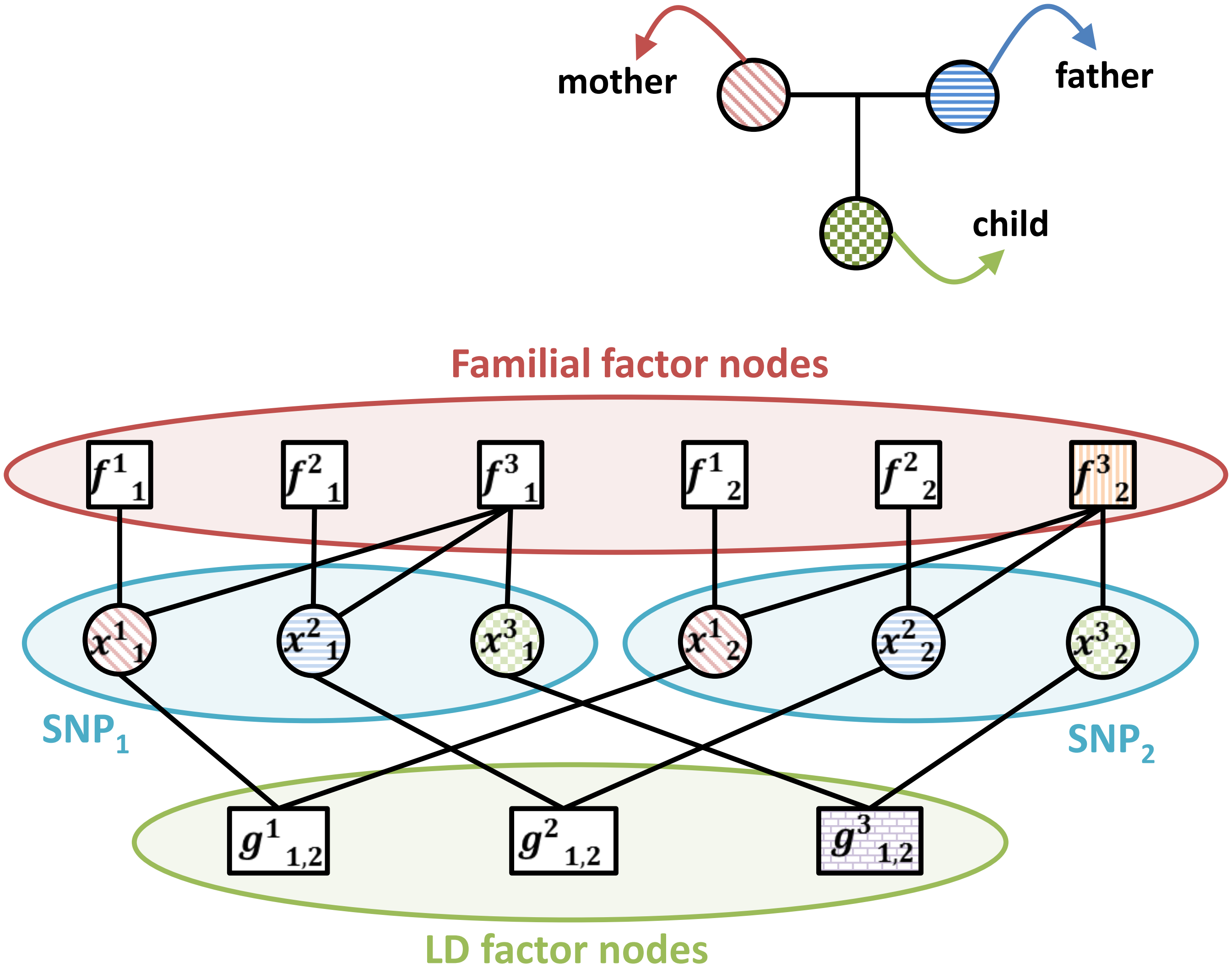
Reconstruction Attack

- $\mathbb{X}_{\mathrm{U}}$: Set of unknown SNPs
- $\mathbb{X}_{\mathrm{K}}$: Set of known SNPs
- Attacker's objective: Compute the marginal probabilities of the SNPs in $\mathbb{X}_{\mathrm{U}}$
  - $p({x^i}_j|\mathbb{X}_{\mathrm{K}}) = \sum_{\mathbb{X}_{\mathrm{U}}\setminus\{{x^i}_j\}} p(\mathbb{X}_{\mathrm{U}}|\mathbb{X}_{\mathrm{K}}, \mathcal{B})$,
    - $p({x^i}_j|\mathbb{X}_{\mathrm{K}})$: Marginal probability distribution of SNP j for individual i can be obtained from
    - $p(\mathbb{X}_{\mathrm{U}}|\mathbb{X}_{\mathrm{K}}, \mathcal{B})$: Joint probability distribution function of the variables in $\mathbb{X}_{\mathrm{U}}$ such that:
    - $\mathcal{B} = (\mathcal{F}_R({x^M}_j, {x^F}_j, {x^C}_j), \mathbb{L}, \mathcal{G}_F, \mathbf{P})$: Background knowledge of the attacker

Efficient Inference Algorithm

- Naive marginalization has computational complexity $\mathcal{O}(3^{mn})$
  - m is on the order of 10s of millions for human genome
- Run the belief propagation algorithm on a factor graph to reduce the computational complexity
  - Technique developed for trust and reputation management (Ph.D. thesis)
  - Factorize the joint probability distribution function into products of simpler local functions
  - Local functions represent conditional dependences between variables
    - LD and reproduction
  - Complexity = $\mathcal{O}(mn)$ per iteration

Factor Graph Representation



▨ mother

▤ father

▦ offspring

Variable nodes

▥ *Familial factor node* for SNP_2 of the offspring

▦ *LD factor node* between SNP_1 and SNP_2 of the offspring

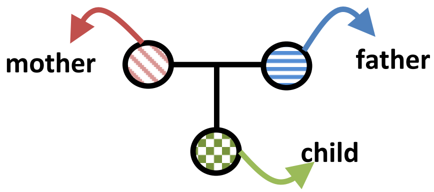
Factorization

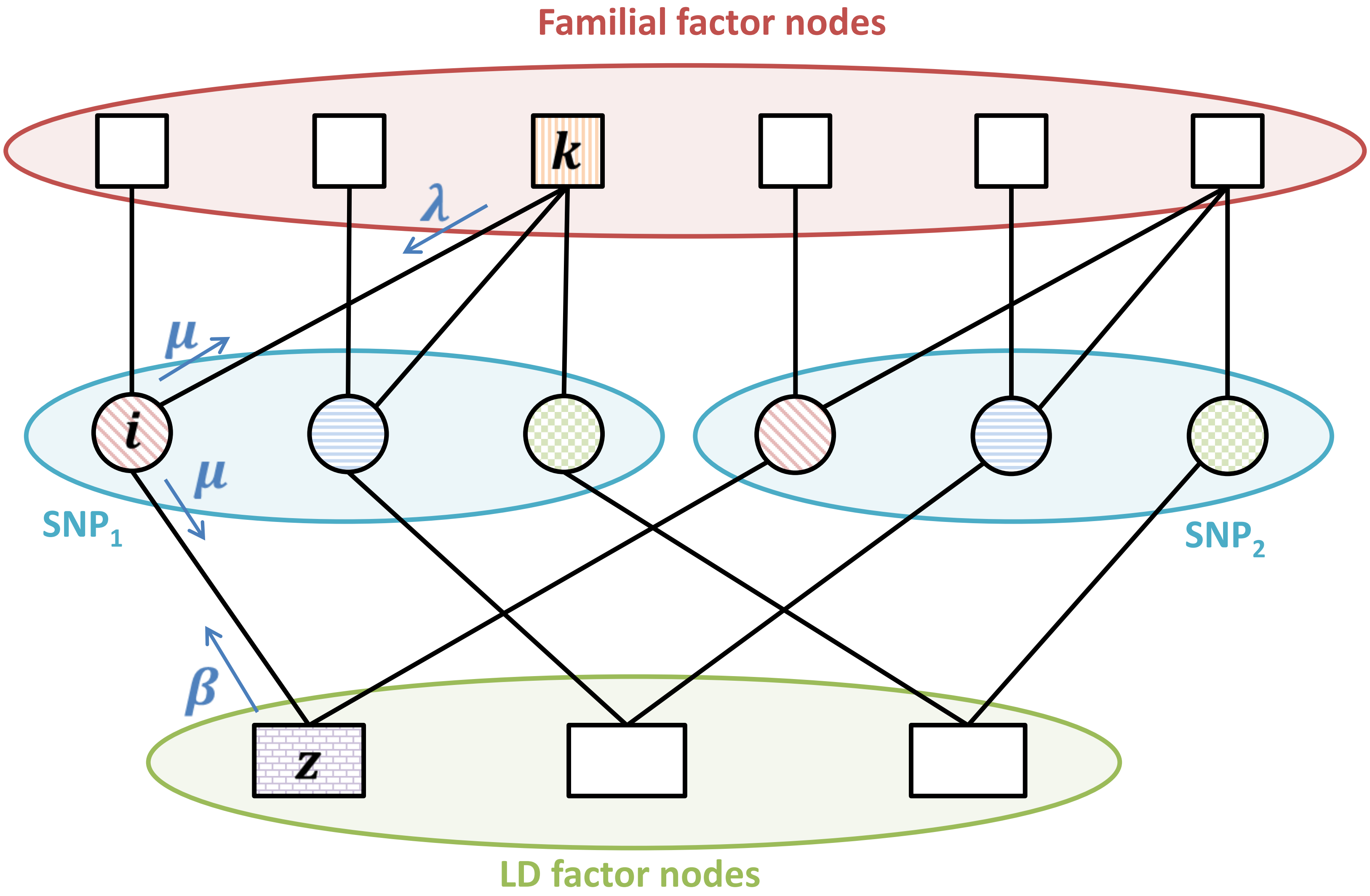
- Factorize the joint probability distribution function into products of simpler local functions
- $p(\mathbb{X}_{\mathrm{U}}|\mathbb{X}_{\mathrm{K}}, \mathcal{B}) =$ $\left[\prod_i \prod_j {f^{i}}_{j}\left({x^{i}}_{j}, \Theta\left({x^{i}}_{j}\right), \mathcal{F}_R\left({x^{M}}_{j}, {x^{F}}_{j}, {x^{C}}_{j}\right), \mathbf{P}\right)\right] \times$ $\left[\prod_i \prod_{(j,m)} {g^{2}}_{1,2}\left({x^{i}}_{j}, {x^{i}}_{m}, \mathbb{L}_{j,m}\right)\right]$

Representing the familial relationships

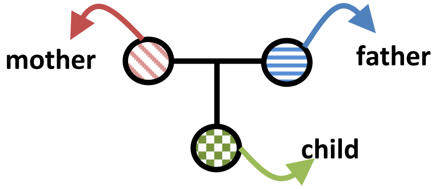
Representing the correlations (LD) between the SNPs

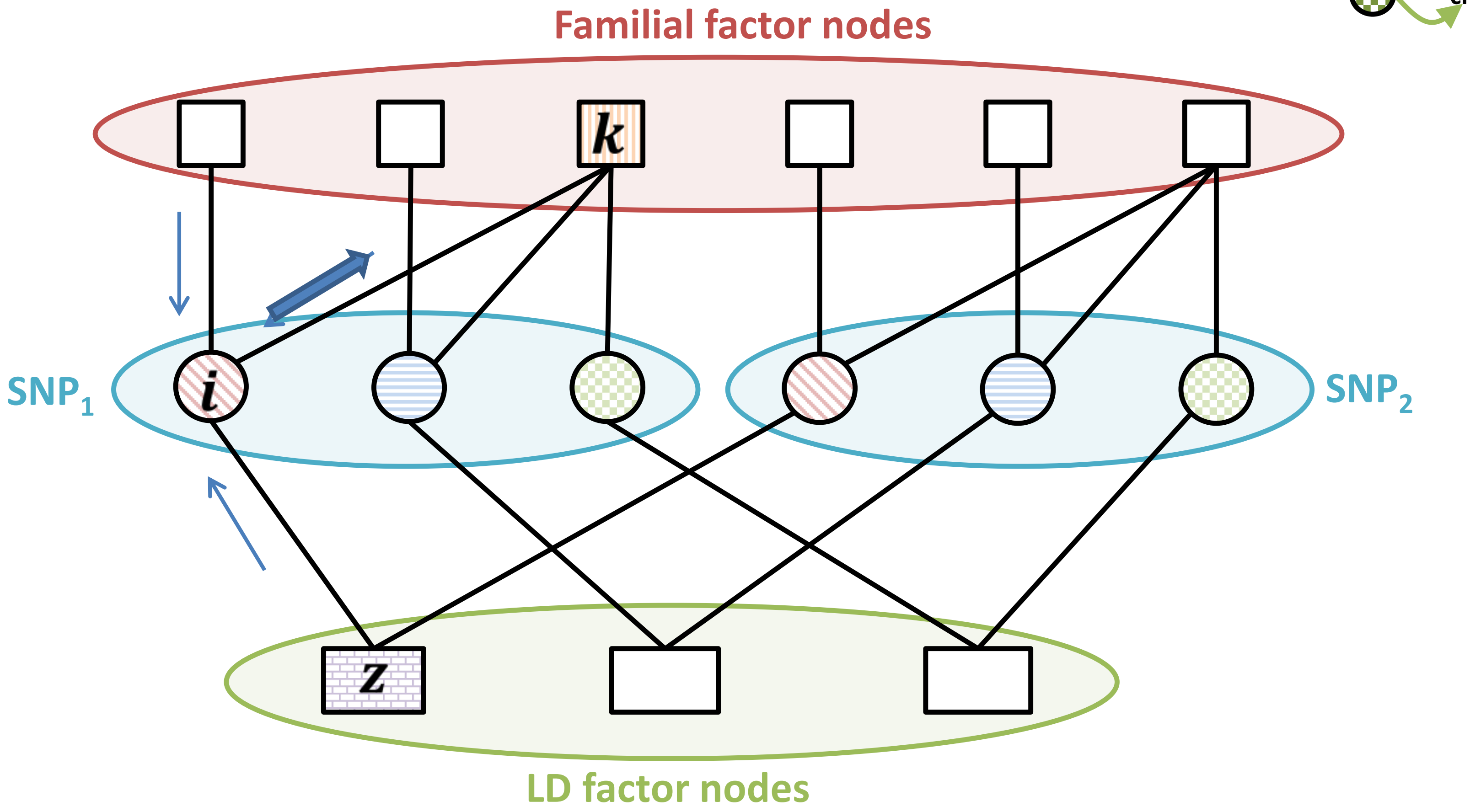
Message Passing





First Round

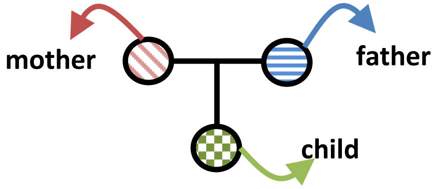


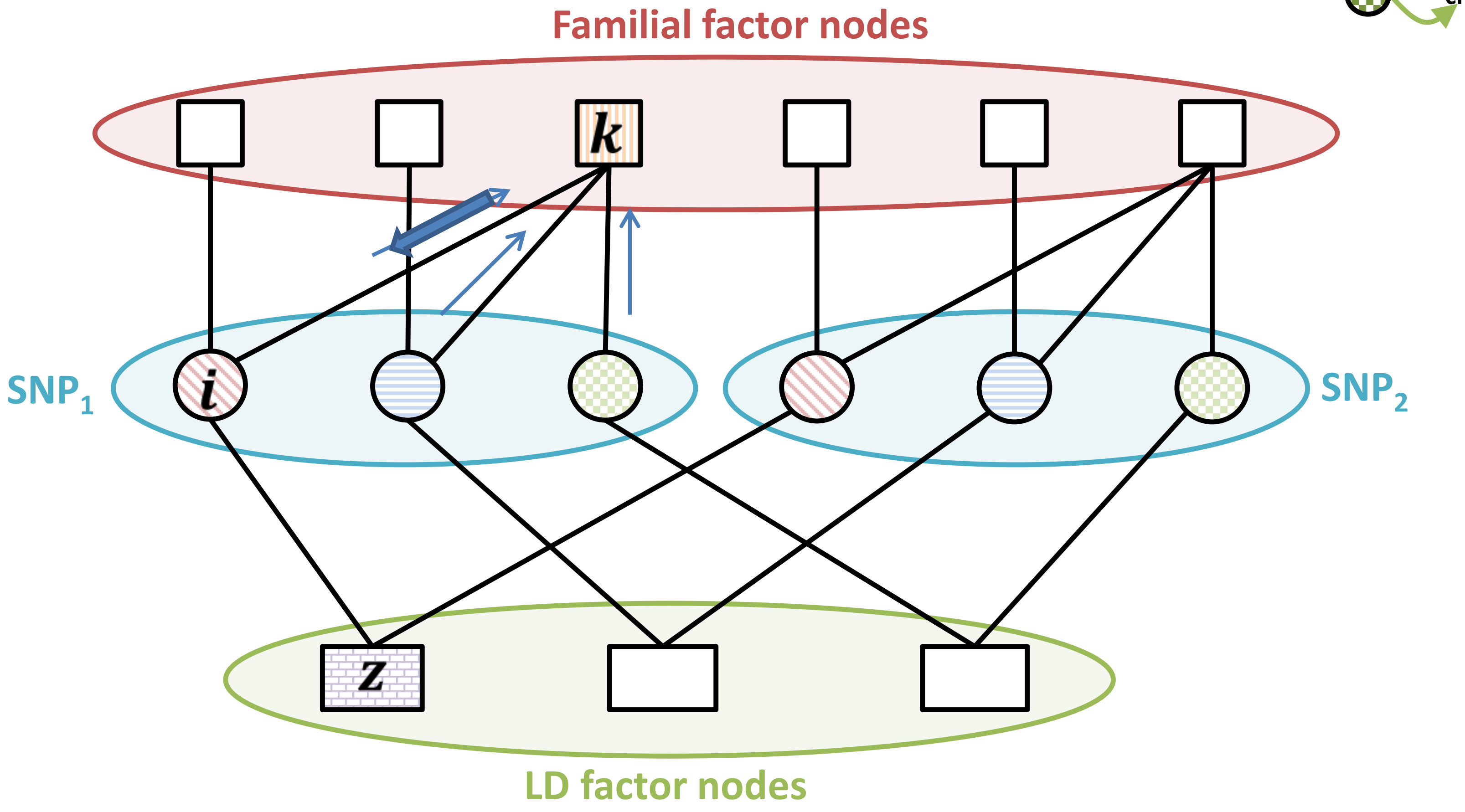


$$\mu^{(v)}{}_{i\to k}\left(x^{1(v)}{}_{1}\right) \propto \prod_{w\in(\sim k)} \lambda^{(v-1)}{}_{w\to i}\left(x^{1}{}_{1}\right) \times \prod_{y=z} \beta^{(v-1)}{}_{y\to i}\left(x^{1}{}_{1}\right)$$

→ Denotes $\Pr(x^{i}{}_{j} = \ell)$

Second Round

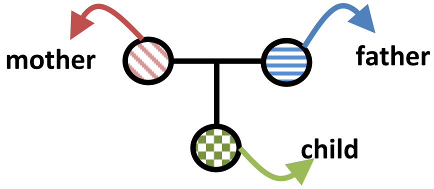




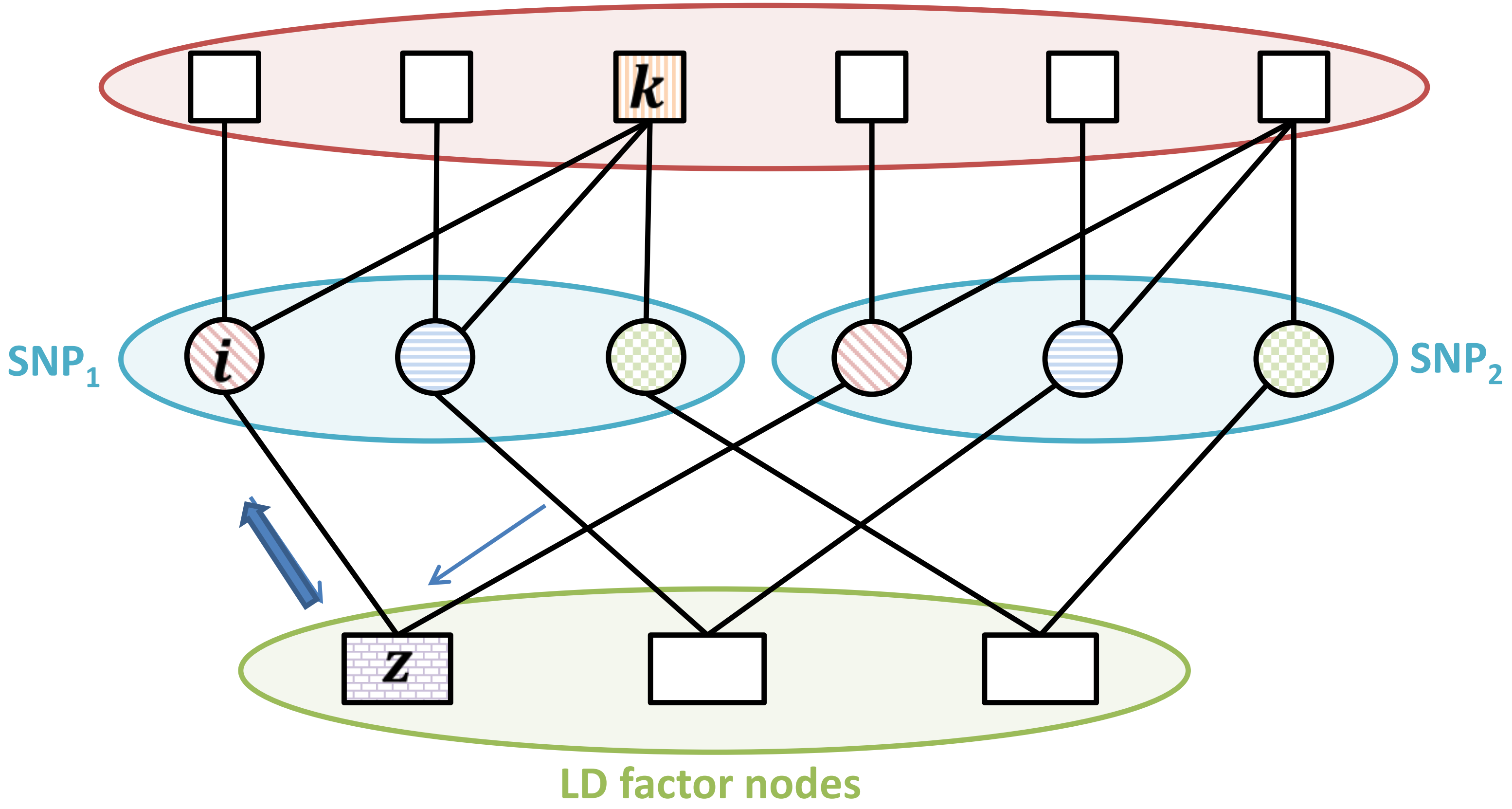
$$\lambda^{(\nu)}{}_{k\to i}(x^1{}_1) \propto \sum_{\Theta(x^1{}_1)} f^3{}_1\left(x^1{}_1, \Theta(x^1{}_1), \mathcal{F}_R\left(x^M{}_j, x^F{}_j, x^C{}_j\right), \mathbf{P}\right) \prod_{y\in\Theta(x^1{}_1)} \mu^{(\nu)}{}_{y\to k}(x^1{}_1)$$

→ Denotes $\Pr(x^i{}_j = \ell)$ given $\Theta\left(x^i{}_j\right), \mathcal{F}_R\left(x^M{}_j, x^F{}_j, x^C{}_j\right), \mathbf{P}$

Third Round



Familial factor nodes



SNP$_1$

SNP$_2$

LD factor nodes

$$\beta^{(\nu)}{}_{z\to i}(x^1{}_1) \propto \sum_{y} g^1{}_{1,2}(x^1{}_1, y, \mathbb{L}_{j,m}) \prod_{y} \mu^{(\nu)}{}_{y\to k}(x^1{}_1)$$

→ Denotes $\Pr(x^i{}_j = \ell)$ given the LD relationships

Convergence and Quantification

- Keep iterating
- At the end of each iteration:
  - Check the inferred marginal distributions of the SNPs in $\mathbb{X}_{\mathrm{U}}$
    - The marginal probability of each variable in $\mathbb{X}_{\mathrm{U}}$ is given by multiplying all the incoming messages at each variable node
- Stop iterations when the values stop changing
- Use the inferred values for quantification of genomic privacy
- Quantify w.r.t:
  - Attacker’s incorrectness
    - Using estimation error metric
  - Attacker’s uncertainty
    - Using Entropy-based metrics

Privacy Metrics

$x^j{}_{i,t}$: actual value

X_k : observed SNPs

- Adversary's incorrectness

Estimation error at SNP *i* for individual $j = \sum_{x^j{}_i} \Pr(x^j{}_i | X_k)\, d(x^j{}_i, x^j{}_{i,t})$

- Adversary's uncertainty [1]

Normalized entropy at SNP *i* for individual $j = \frac{1}{\log(3)} \sum_{x^j{}_i} -\Pr(x^j{}_i | X_k) \log \Pr(x^j{}_i | X_k)$

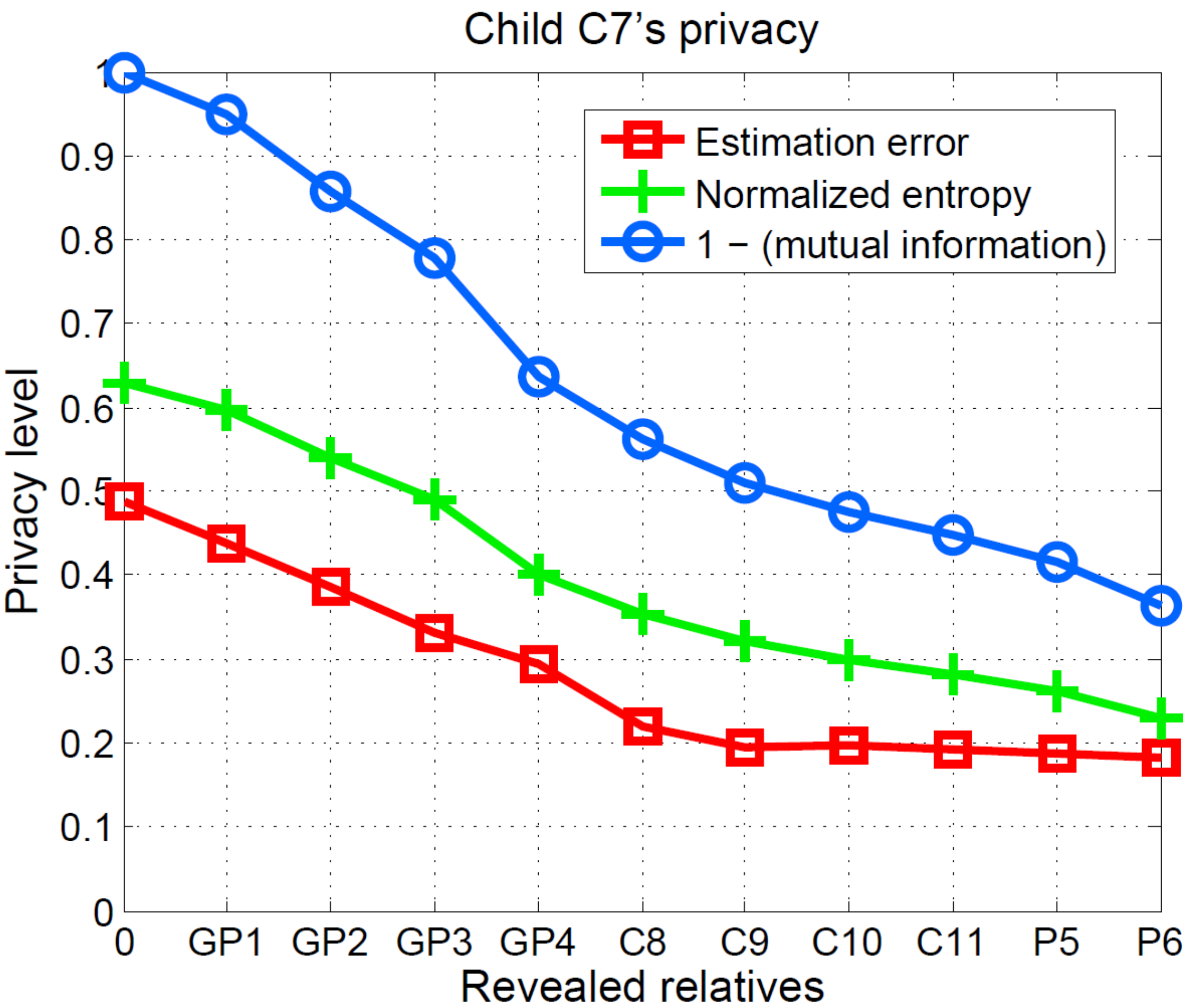
- Mutual information-based metric [2]

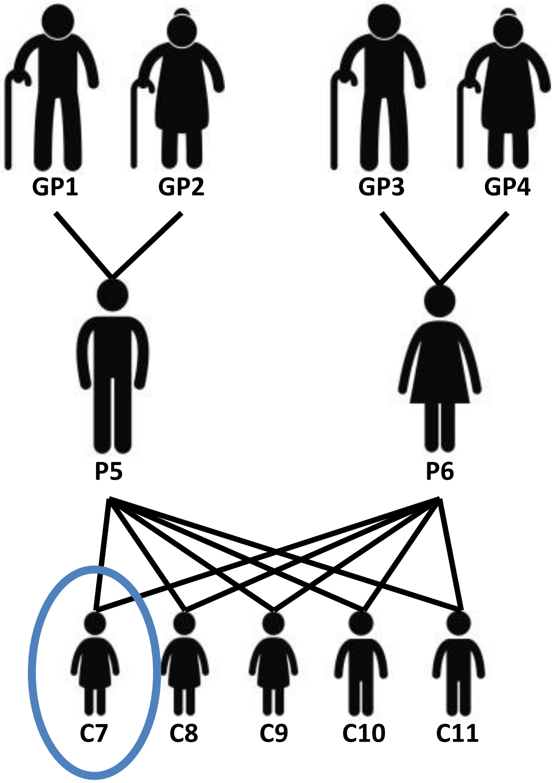
1 – (normalized) mutual information at SNP *i* for individual $j = 1 - \frac{H(x^j{}_i) - H(x^j{}_i | X_k)}{H(x^j{}_i)} = \frac{H(x^j{}_i | X_k)}{H(x^j{}_i)}$

[1] Serjantov, A. and Danezis, G., Towards an information theoretic metric for anonymity, PET 2003
[2] Agrawal, D. and Aggarwal C.C., On the design and quantification of privacy preserving data mining algorithms, PODS 2001

Evaluation - 80k SNPs, w\o LD

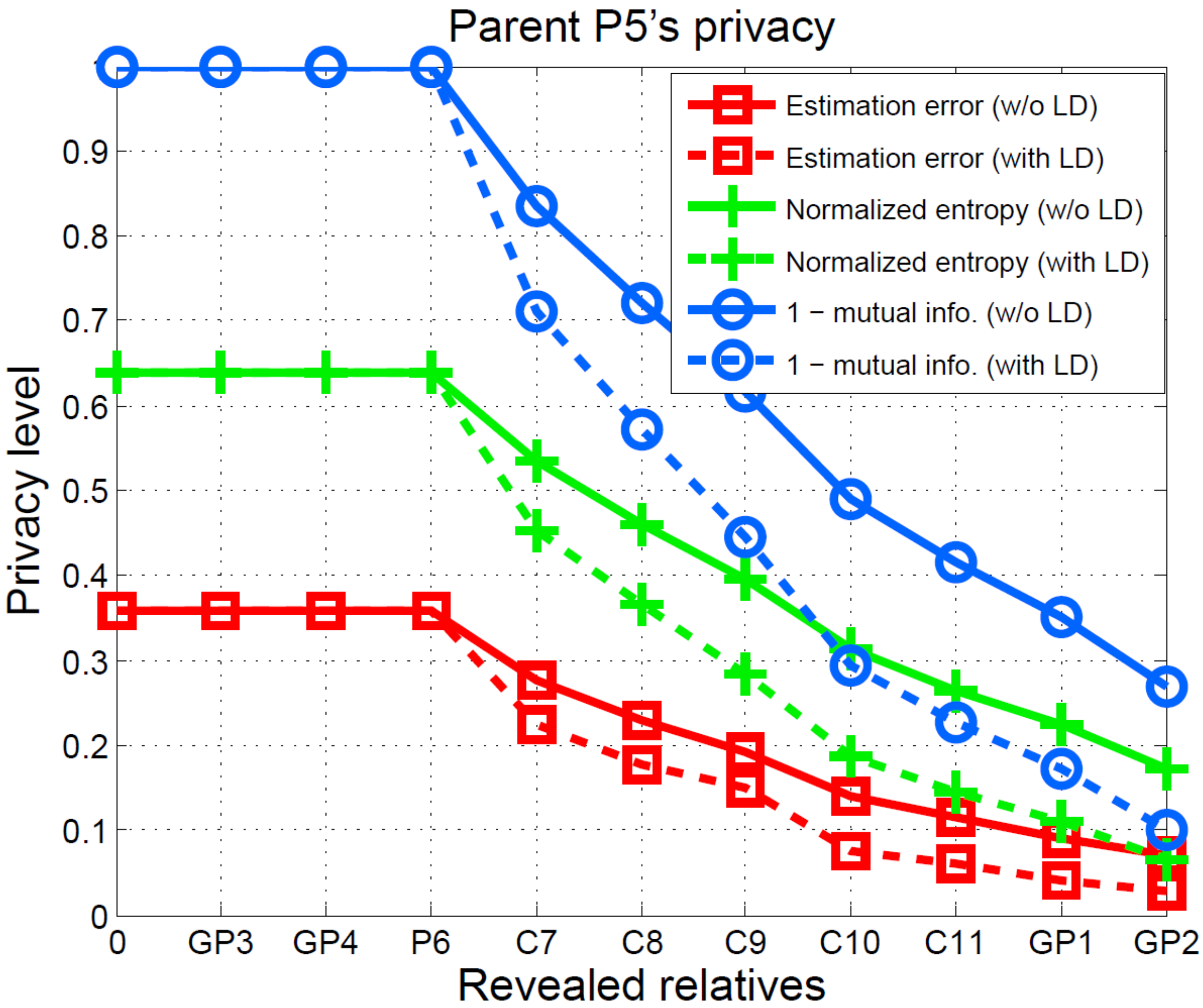
Evolution of the genomic privacy of child C7 by gradually revealing the SNPs of other family members (starting with the most distant family members)

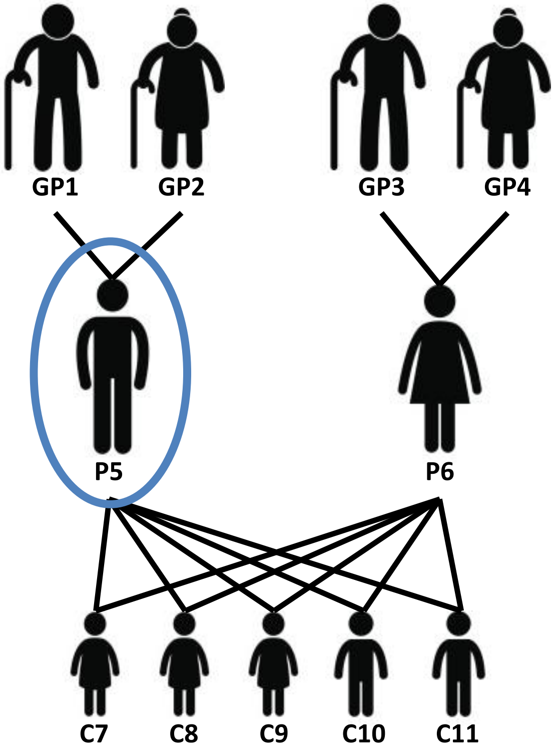




Evaluation

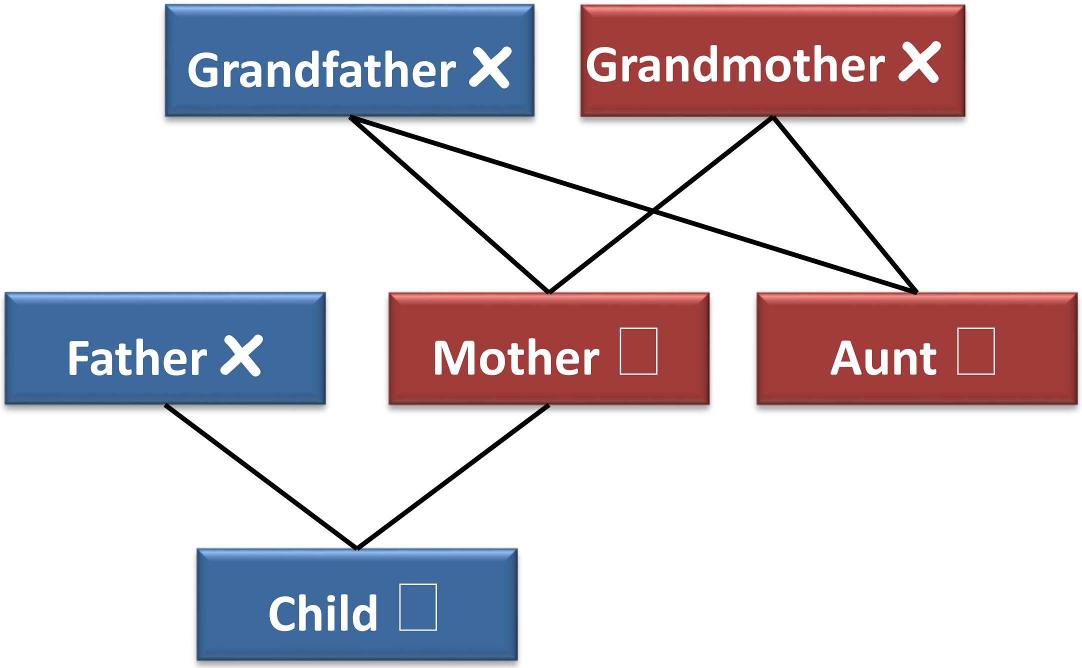
Evolution of the genomic privacy of parent P5 by gradually revealing 50 SNPs (out of 100) of other family members (starting with the most distant family members)





Threat in Online Social Networks

- De-anonymized 149 individuals from OpenSNP
  - Using other publicly available resources
  - (Partially) sharing their genomes (about 1M SNPs each)
- Found the family tree of 47
  - Using the family information on Facebook, 23andMe, Geneology.org, etc.
  - 3 de-anonymized individuals belong to the same family
- Computed health privacy for Alzheimer’s disease



Grandfather ✗

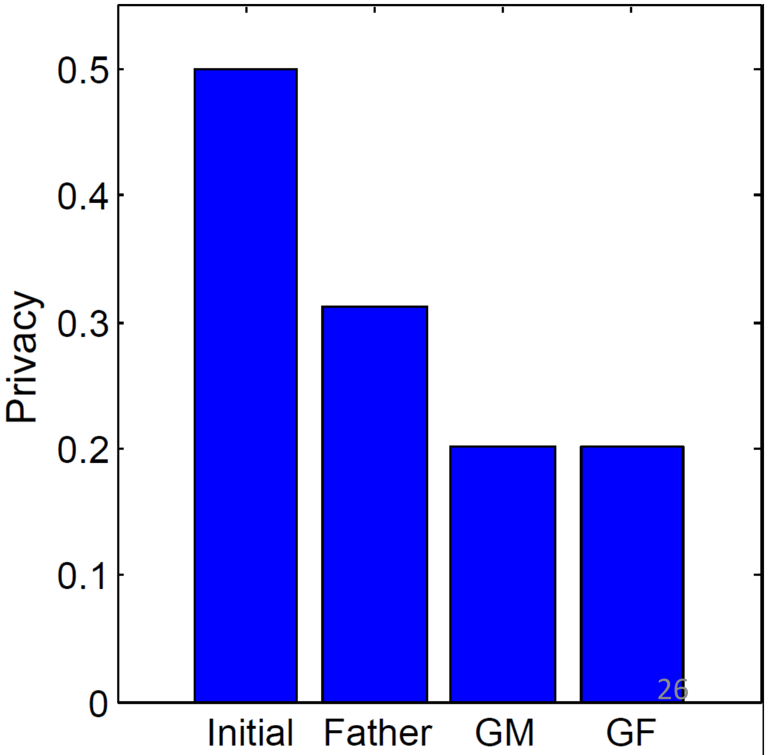
Grandmother ✗

Father ✗

Mother □

Aunt □

Child □



Discussion

- Genomes of relatives are highly correlated and some family members might be opposed to *genetic exhibitionism*
- Making thousands of human genomes publicly available is crucial for genomic researchers

> *"If we are going to solve cancer, it is going to take a movement of tens of thousands, or hundreds of thousands, of patients willing to contribute information from their cancer genomes towards a common good"*
>
> Eric S. Lander, the founding director of the Broad Institute

- Trade-off between privacy and utility
- Design optimal genomic-privacy preserving mechanisms

PROTECTING GENOMIC PRIVACY

Protecting Genomic Privacy - Our Solutions

- Computational Privacy
  - Privacy-preserving personalized medicine
  - Privacy-preserving management of raw genomic data (BAM files)
  - Privacy-preserving genomic research
    - Ancestry inference
    - Genome-wide association studies
  - Data sharing and finding similar patients using functional encryption
  - Real-life implementations with CHUV, Sophia Genetics, and Swiss HIV Cohort
- Information Theoretical Privacy
  - Optimization-based techniques
    - Privacy vs. utility
  - GeneVault via HoneyEncryption

Operation Mincemeat

- Successful British disinformation plan during World War II
- Operation Mincemeat saved an estimated 40,000 Allied lives
- It also gave rise to a movie… The Man Who Never Was

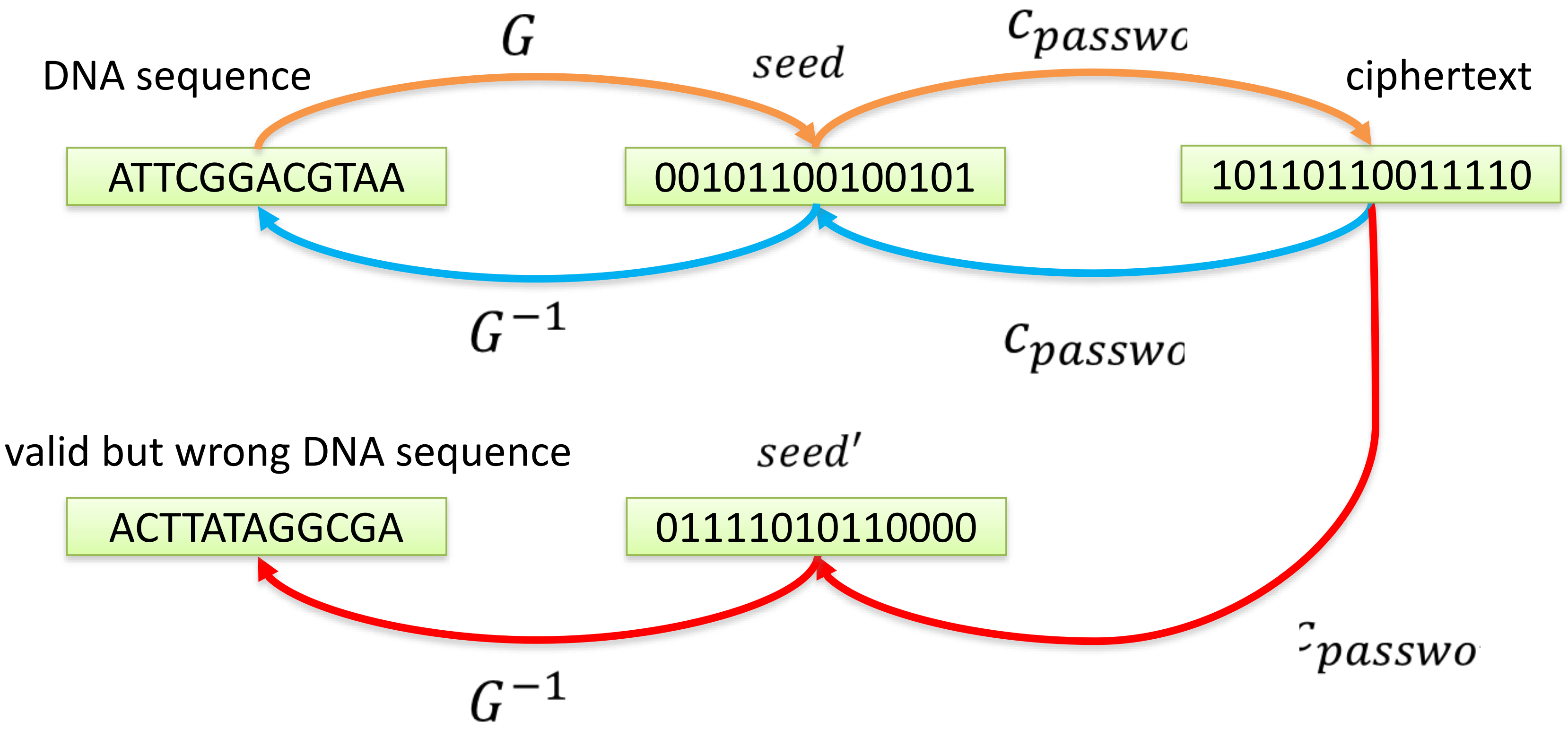
THE STRANGEST MILITARY HOAX OF WORLD WAR II!
THE MAN WHO NEVER WAS
CINEMASCOPE PICTURE
COLOR by DE LUXE
STARRING
CLIFTON WEBB · GLORIA GRAHAME
with ROBERT FLEMYNG · JOSEPHINE GRIFFIN · STEPHEN BOYD
PRODUCED BY ANDRE HAKIM · DIRECTED BY RONALD NEAME · SCREEN PLAY BY NIGEL BALCHIN · FROM THE NOVEL BY EWEN MONTAGU · RELEASED BY 20th CENTURY-FOX

Decoys

- Decoys, fake objects that look real, are a time-honored counterintelligence tools
- In computer security, we have “honey objects”:
  - Honeypots
  - Honeytokens, honey accounts
  - Decoy documents

- **Key question:** How can we apply honey objects to the most pressing computer security / privacy problems?
  - Password breaches in the cloud (Juels et al.)
  - Breaches in genome databases

GeneVault



DNA sequence

ATTCGGACGTAA

G

$seed$

001011001001 01

c_{passwo}

ciphertext

101101100111 10

G^{-1}

c_{passwo}

valid but wrong DNA sequence

ACTTATAGGCGA

$seed'$

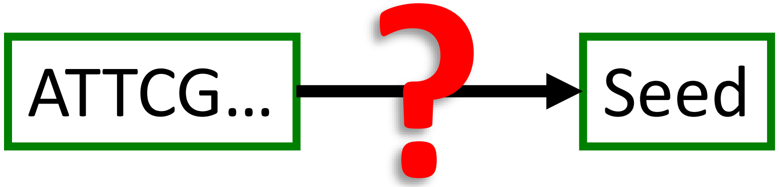
011110101100 00

G^{-1}

c_{passwo}

GeneVault – Main Challenge

- How to build such a generator G that can simulate the distribution of genome sequences?
  - Naïve way: enumerate all genome sequences and compute their probabilities based on allele frequencies and linkage disequilibrium (LD)
  - Works, but impractical
  - Is there a more intelligent way to do so?



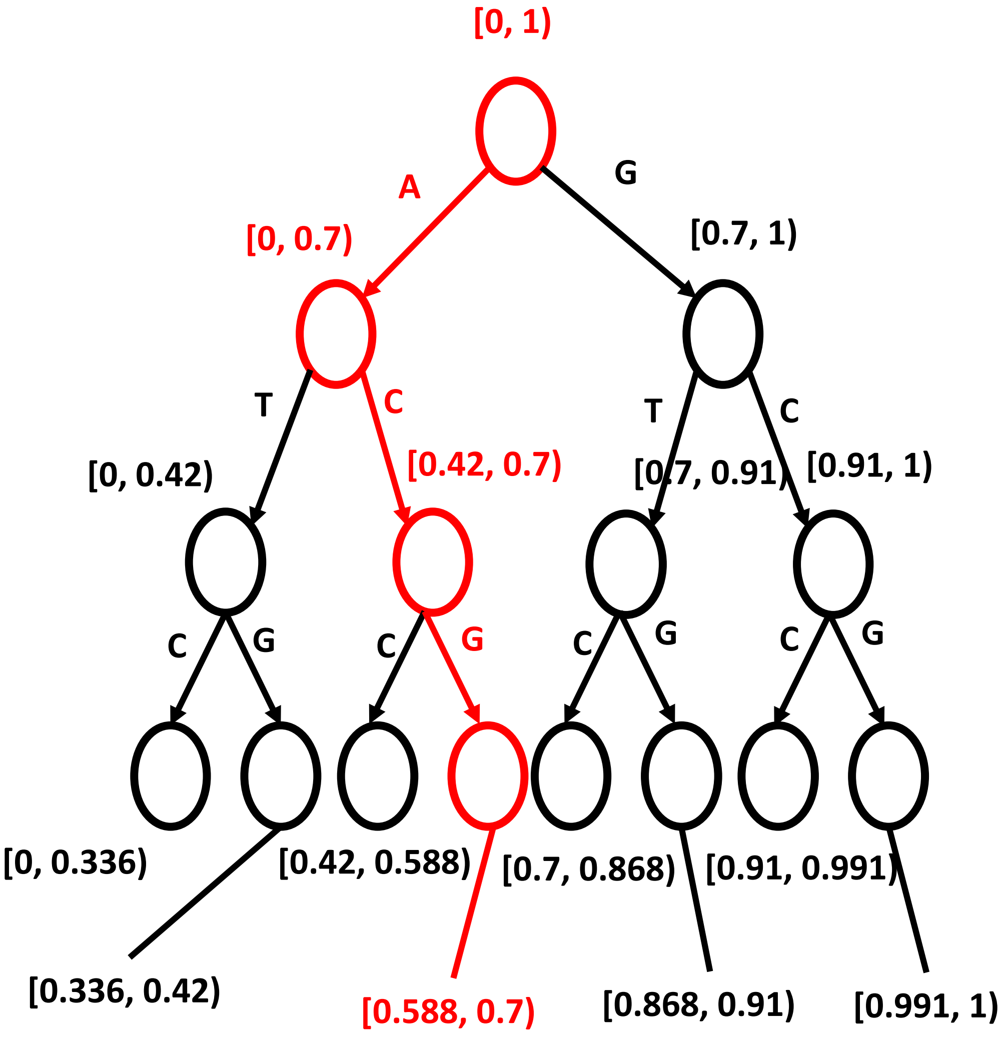
GeneVault - Example

- Transform sequence ACG into a seed:
- Randomly pick a seed $0.6 \in [0.588, 0.7)$
- 8 bits to encode one seed:

$$\lfloor 0.6 \times 256 \rfloor = 153$$
$$= 10011001_2$$

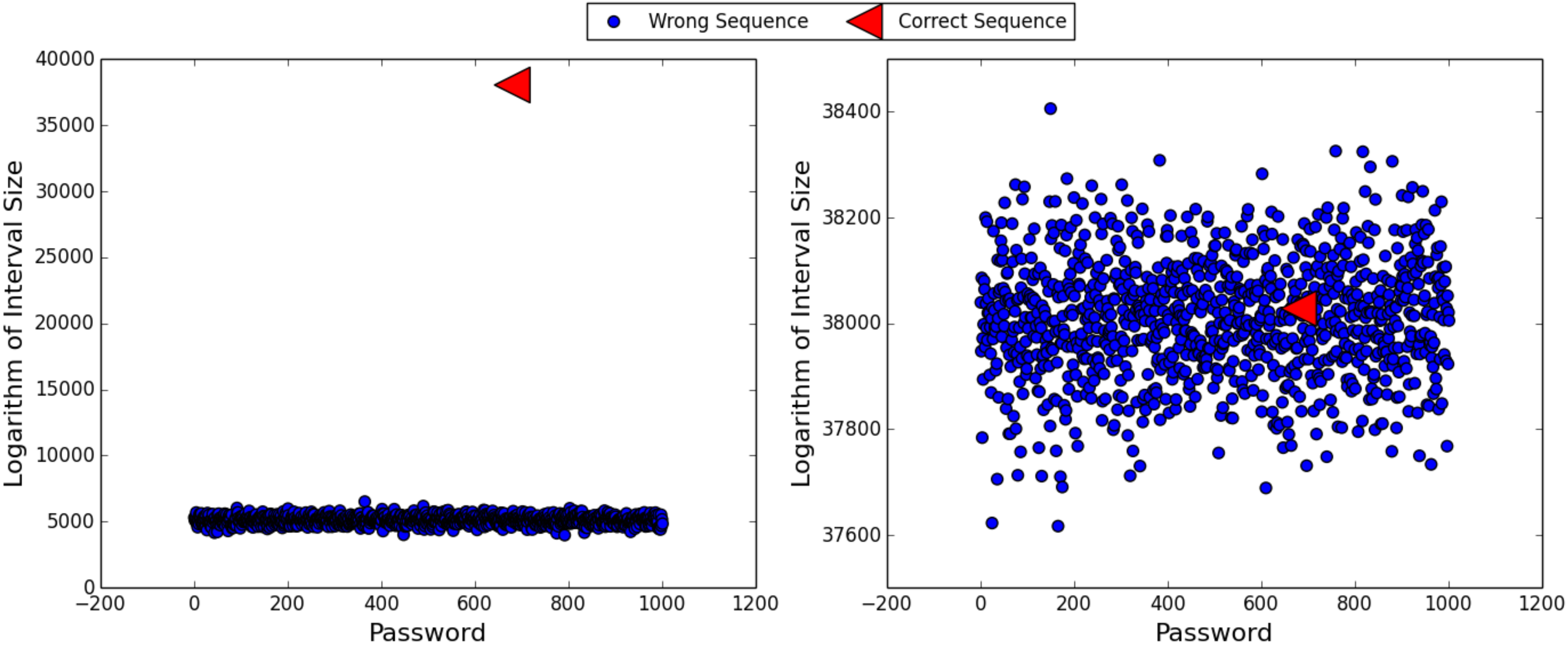
- Password “hzc”
=> Generate Key:
Gen(“hzc”) = 01000110
- Ciphertext:

$$
\begin{aligned}
& 10011001 \quad \text{(seed)} \\
\oplus\ & 01000110 \quad \text{(Key)} \\
= \ & 11011111
\end{aligned}
$$



GeneVault – Security Evaluation

- Probability of a decrypted sequence



Traditional Encryption

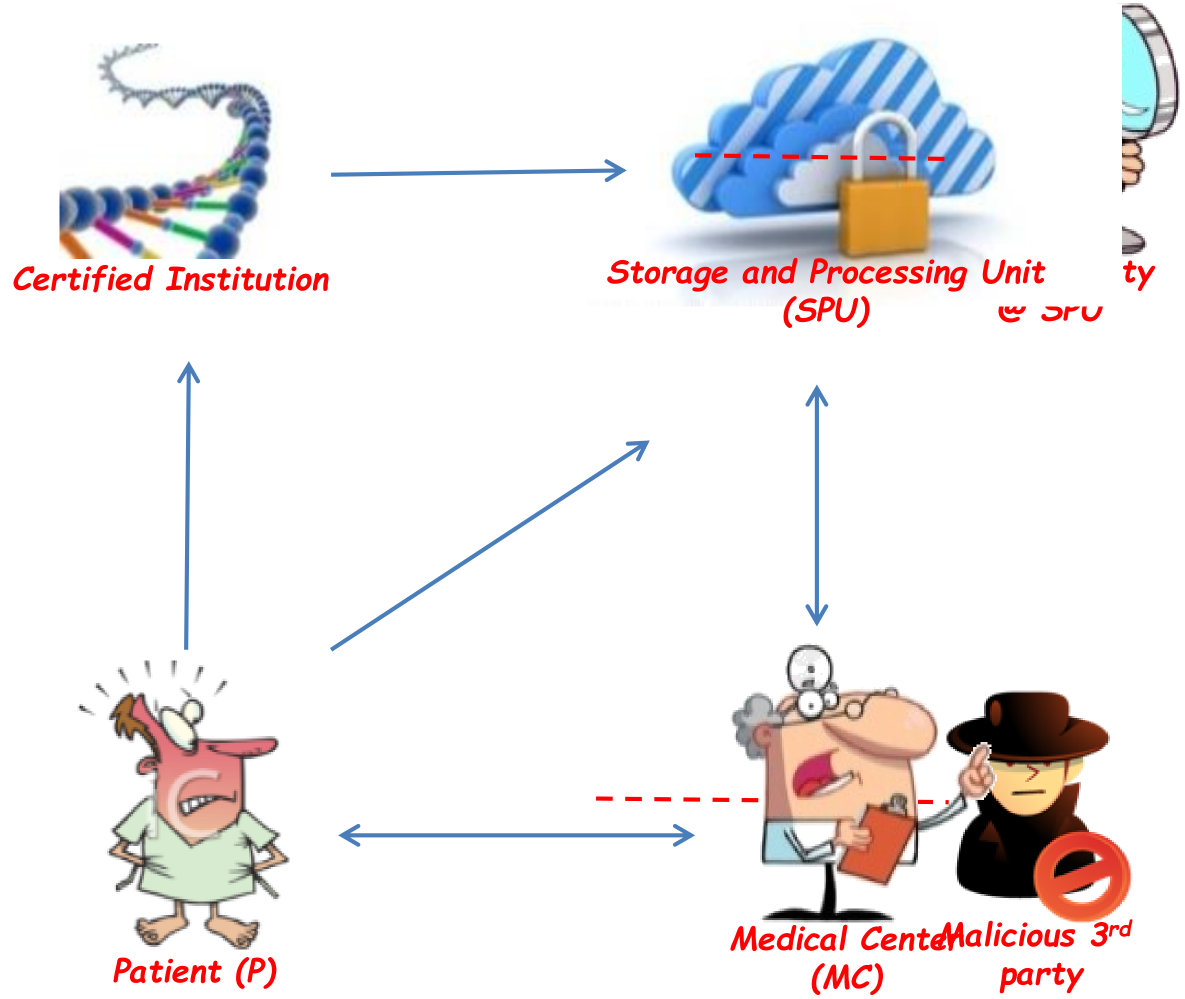
GeneVault

GeneVault – Still to Come

- Partial Retrieval
- Typo
  - When the user incidentally types a wrong password, he will get a plausible sequence
  - Might cause problems if he doesn’t realize it and uses it for medical purposes
- Adversary’ s background knowledge
  - Physical traits, phenotypes (eye color, hair color, etc.)
  - Kinship
  - Can eliminate some (incorrect) keys if the decrypted sequence doesn’t indicate those phenotypes
- Operations on the data

MORE ON PROTECTING GENOMIC PRIVACY

Privacy-Preserving Personalized Medicine



Certified Institution

Storage and Processing Unit (SPU)

ty

@ SPU

Patient (P)

Medical Center (MC)

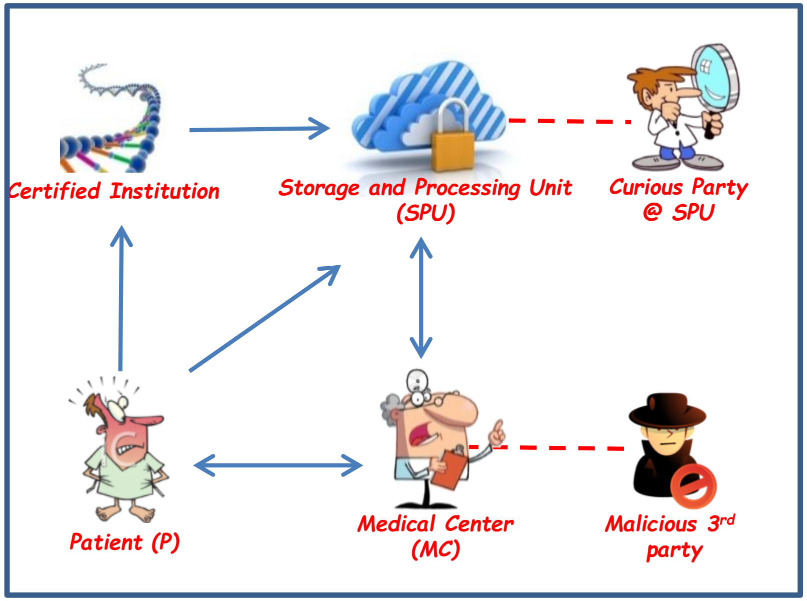
Malicious 3rd party

Setting and Goals

- Setting: A medical center (MC) want to conduct a *genetic disease susceptibility test* on a patient (P)
- Protect the privacy of users' genomic data
  - Protect data, including from insiders (e.g., curious sysadmins)
- Protect the privacy of medical center's confidential data
- Allow specialists to access only to the genomic data they need (or they are authorized for)
- Keep the access time to a single patient's genomic data to a few seconds

Threat Model

- The certified institution (CI) is a trusted entity.
  - Indispensable to do the sequencing
- An attacker at the MC
  - A careless or disgruntled employee at the MC or a hacker who breaks into the MC
  - Aims to obtain private genomic information about a patient (for which it is not authorized)
- A curious party at the SPU
  - Existence of a curious party or a disgruntled employee at the SPU
- Both MC and SPU follows the protocols properly
- No collusion between the MC and the SPU
- Access control based on patient's consent



Cryptographic Tools

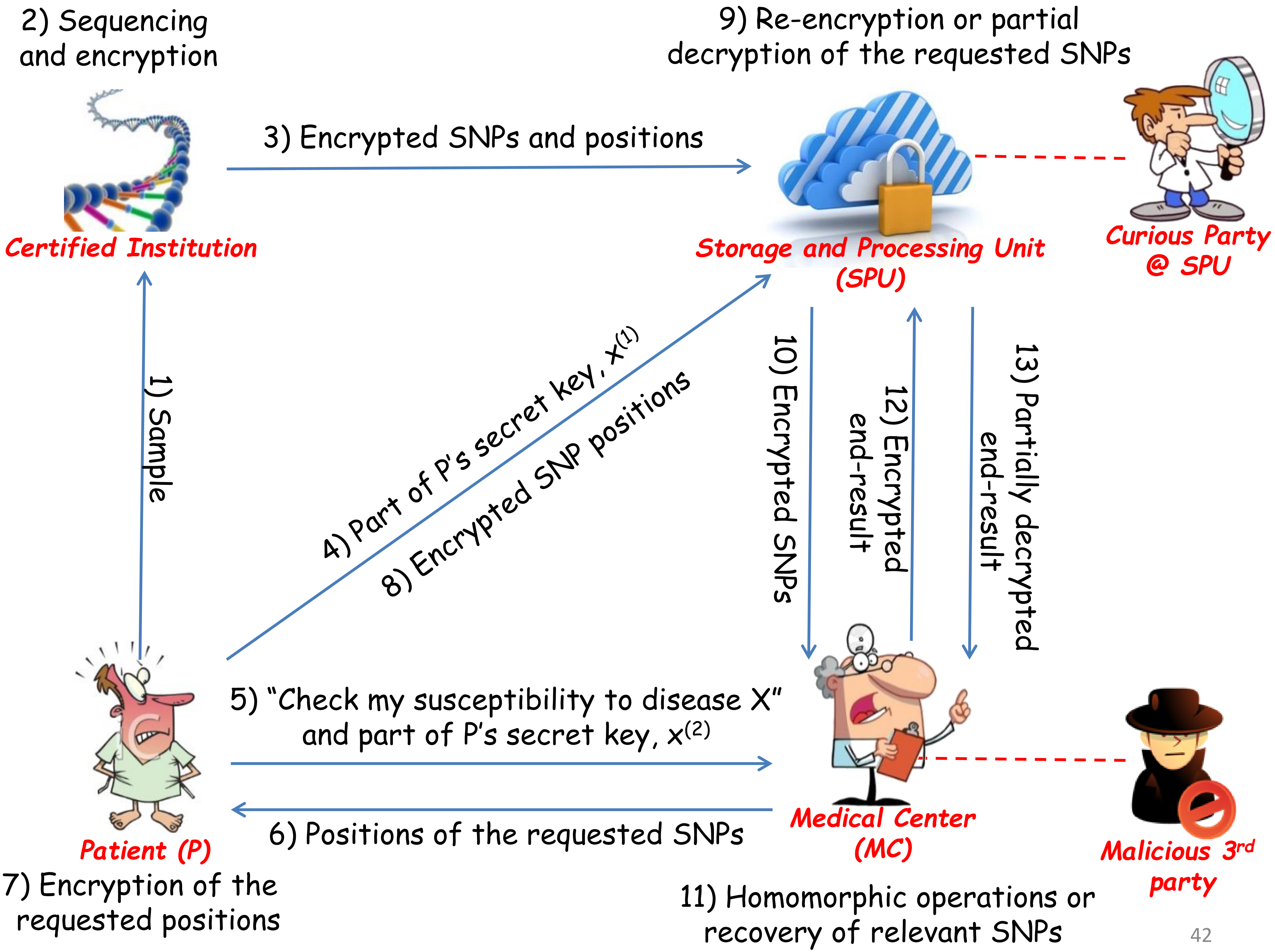
- Modified Paillier Cryptosystem
  - Bresson et. al 2003.
  - Homomorphic addition

$$D(E(m_1, r_1, g^{x_p}) \cdot E(m_2, r_2, g^{x_p})) = D(T_1^1 \cdot T_1^2, T_2^1 \cdot T_2^2 \mod n^2) = m_1 + m_2 \mod n$$

  - Multiplication with a constant

$$D(E(m_1, r_1, g^{x_p})^k) = D((T_1^1)^k, (T_2^1)^k \mod n^2) = km_1 \mod n.$$

  - Proxy re-encryption
    - Divide the weak secret into two shares
    - Distribute the shares to two parties
- Secure multiparty computation (SMC)



2) Sequencing and encryption

9) Re-encryption or partial decryption of the requested SNPs

3) Encrypted SNPs and positions

Certified Institution

Storage and Processing Unit (SPU)

Curious Party @ SPU

1) Sample

4) Part of P's secret key, $x^{(1)}$

8) Encrypted SNP positions

10) Encrypted SNPs

12) Encrypted end-result

13) Partially decrypted end-result

5) "Check my susceptibility to disease X" and part of P's secret key, $x^{(2)}$

6) Positions of the requested SNPs

Patient (P)

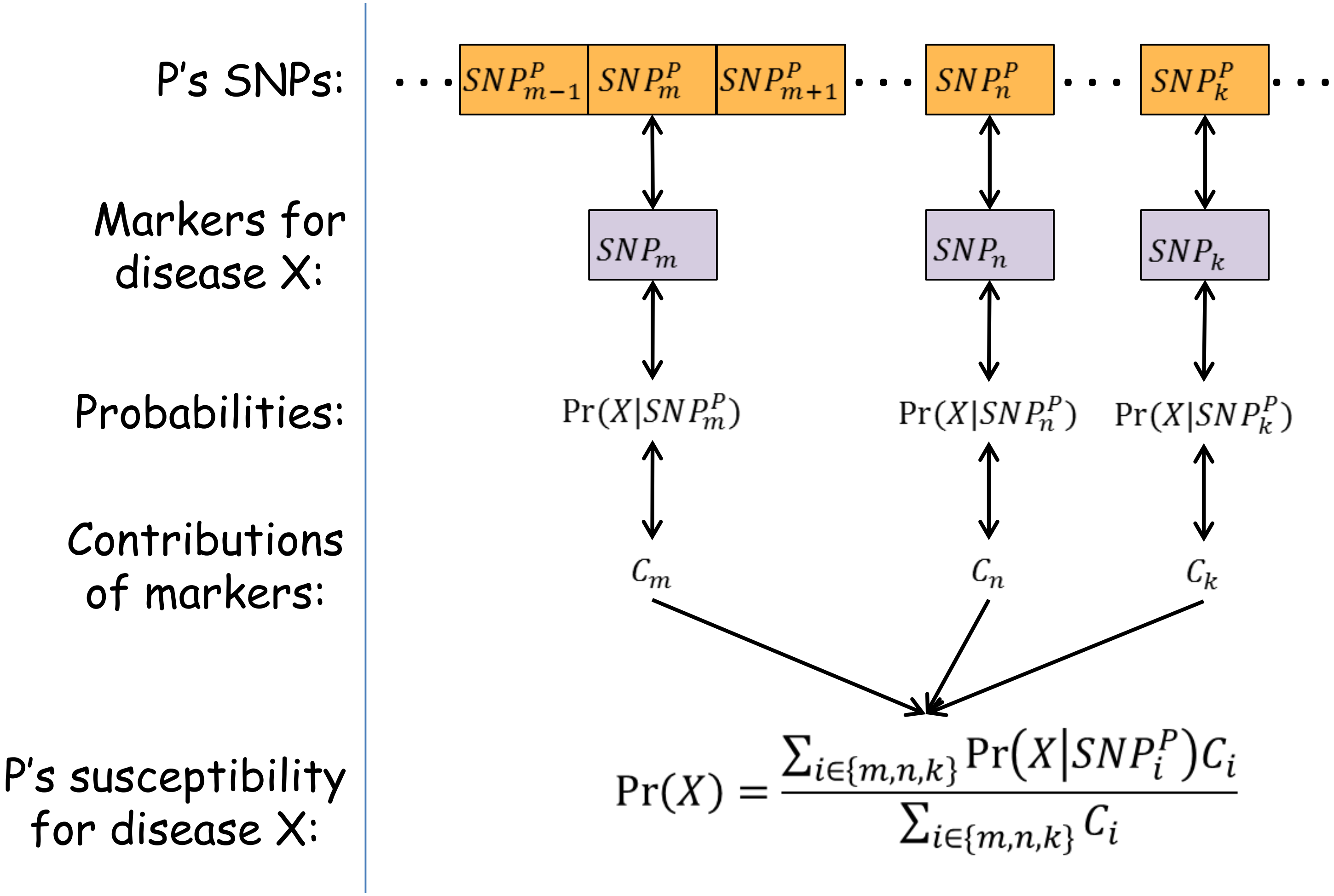
Medical Center (MC)

Malicious 3rd party

7) Encryption of the requested positions

11) Homomorphic operations or recovery of relevant SNPs

Computing Disease Susceptibility



P's SNPs: $\cdots SNP^P_{m-1}\ SNP^P_m\ SNP^P_{m+1} \cdots SNP^P_n \cdots SNP^P_k \cdots$

Markers for disease X: SNP_m SNP_n SNP_k

Probabilities: $\Pr(X|SNP^P_m)$ $\Pr(X|SNP^P_n)$ $\Pr(X|SNP^P_k)$

Contributions of markers: C_m C_n C_k

P's susceptibility for disease X:

$$\Pr(X) = \frac{\sum_{i\in\{m,n,k\}} \Pr\left(X\middle|SNP^P_i\right) C_i}{\sum_{i\in\{m,n,k\}} C_i}$$

- All operations are conducted in ciphertext using homomorphic encryption

Remarks

- Patient-related steps can be handled via the patient's smart card or mobile device
- Individual contributions of the genetic variant markers remain secret at the MC
  - Homomorphic operations are conducted at the MC
- Solution is possible without the proxy re-encryption by letting the patient decrypt the end-result
  - Secret key of the patient remains only at the patient
  - Useful when the collusion between the SPU and MC is possible
- **Does this solve everything?**

Quantification of Genomic Privacy

- Privacy is quantified from MC's view-point
- Two types of genetic tests:
  - Test 1: MC obtains a subset of SNPs of P
    - For complex diseases that homomorphic operations fail
    - Privacy loss due to the exposition of a subset of SNPs
  - Test 2: MC obtains the end-result of a genetic test
    - Test is conducted at the MC using homomorphic operations
    - Privacy loss due to the exposition of the end-result

Quantification of Genomic Privacy

- What the MC knows?
  - Markers (SNPs) and their contributions to the diseases (for Test 2)
  - Contributions of two alleles (of a SNP) to a disease
  - Linkage Disequilibrium (LD) values between the SNPs
    - LD occurs when SNPs at the two SNP positions are not independent of each other
- Goal:
  - Compute the decrease in privacy of the patient given his revealed SNPs or the end-result of a genetic test
  - Used asymmetric entropy for the quantification
  - Maximize the genomic privacy of the patient via obfuscation methods or policies

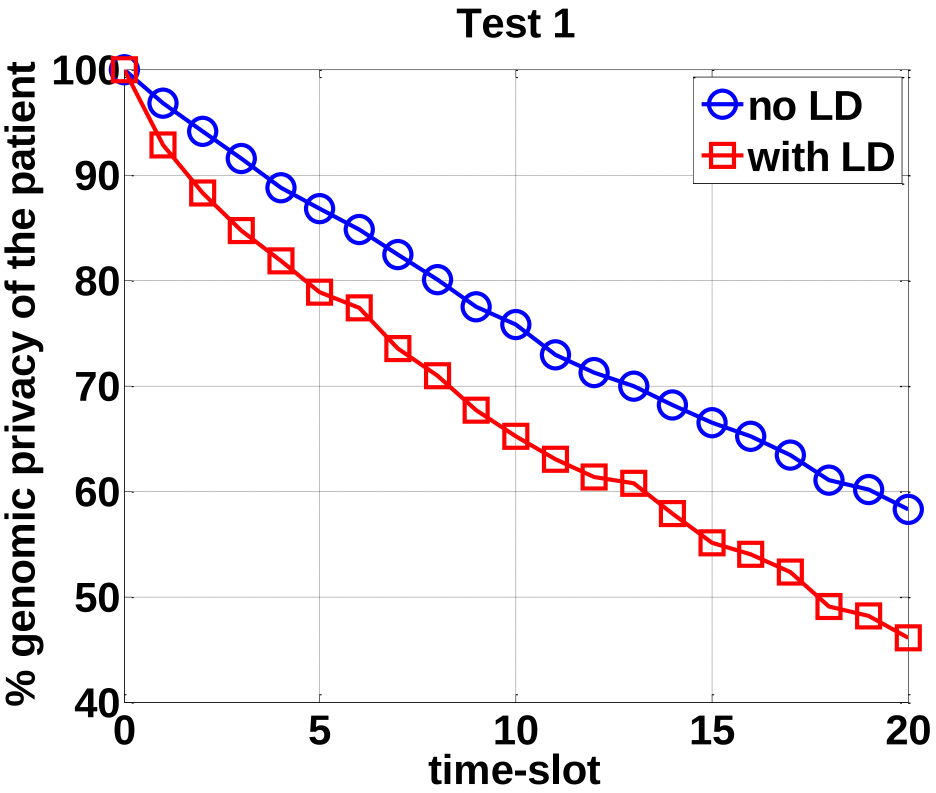
Methodology

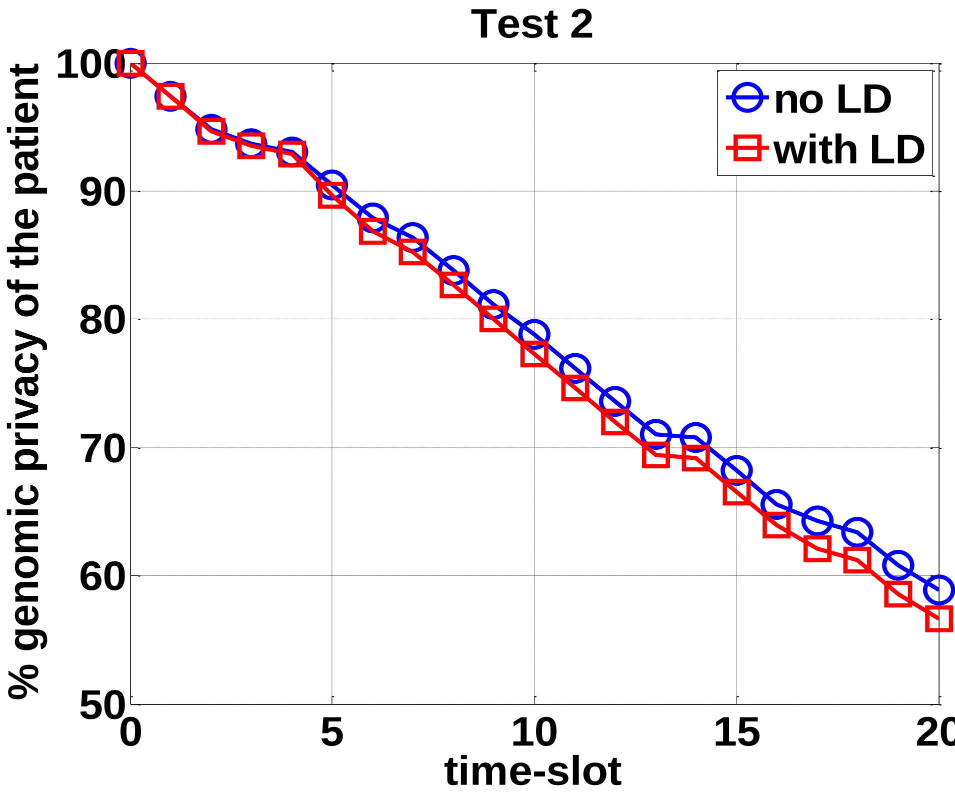
- At each time slot, randomly conduct a test
  - Either Test 1 or Test 2
- Test 1:
  - Min number of markers revealed: 10
  - Max number of markers revealed: 15
  - Update the inferred values of non-revealed SNPs using LD
- Test 2:
  - Randomly chose a disease to test
  - Compute the end-result
    - Weighted averaging (to compute the disease susceptibility)
  - Compute the potential end-results using public information
  - Update the inferred values of the non-revealed SNPs using the end-result of the test

Parameters

- Real human DNA profile from 1000 Genome Project
- Consider a particular subset of SNPs
  - 500 SNPs
- Susceptibility to 40 diseases are determined using these SNPs
- Each disease is associated with at least 1 and at most 15 SNPs
- 12 SNPs are markers of more than one disease
- Real LD values between these SNPs

Decrease in Genomic Privacy

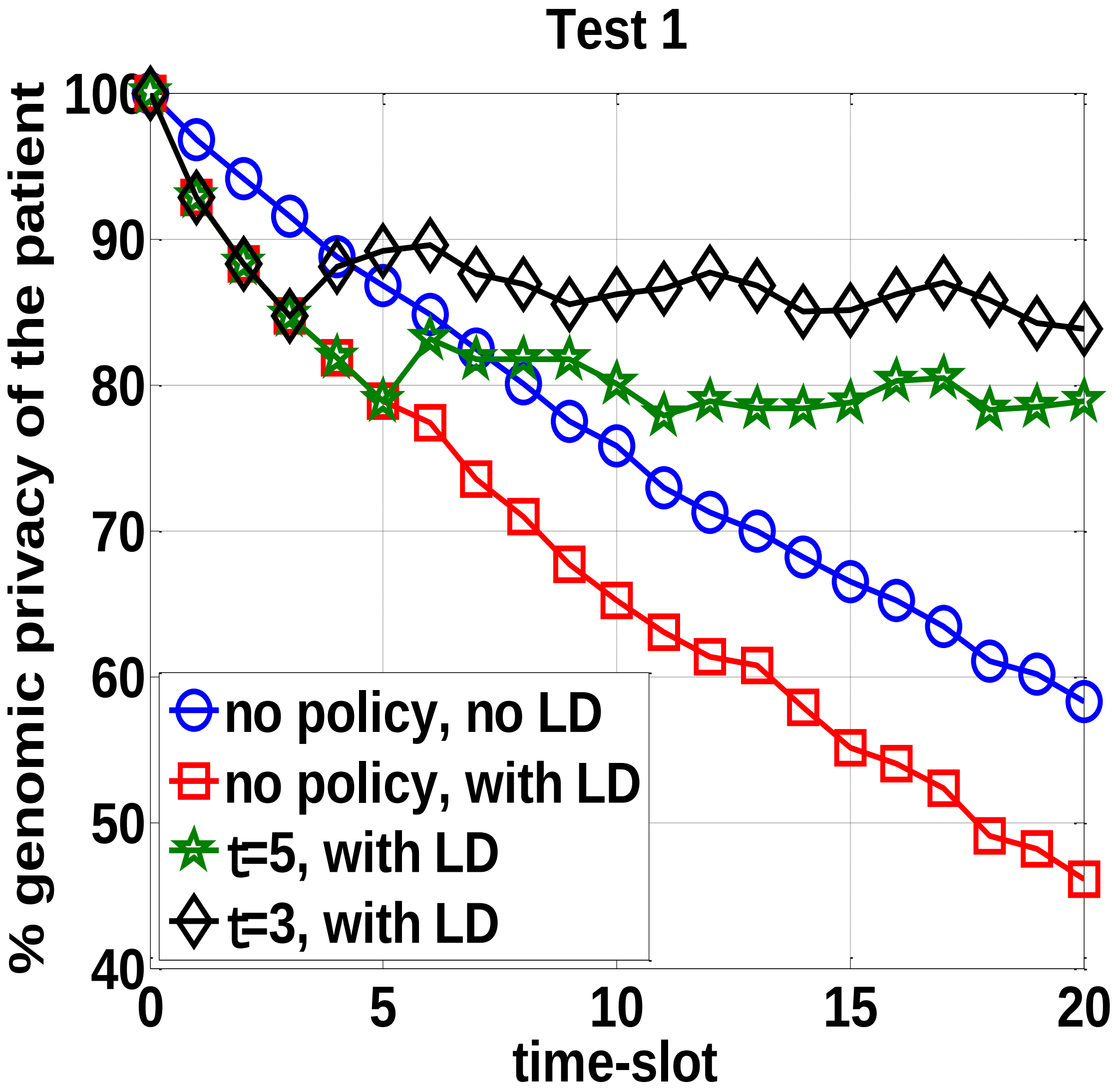




- Need to introduce techniques to keep the genomic privacy above a certain level
  - For Test 1: Define policies to delete the revealed SNPs from MC’s database
  - For Test 2: Use obfuscation methods on the end-result of the genetic test

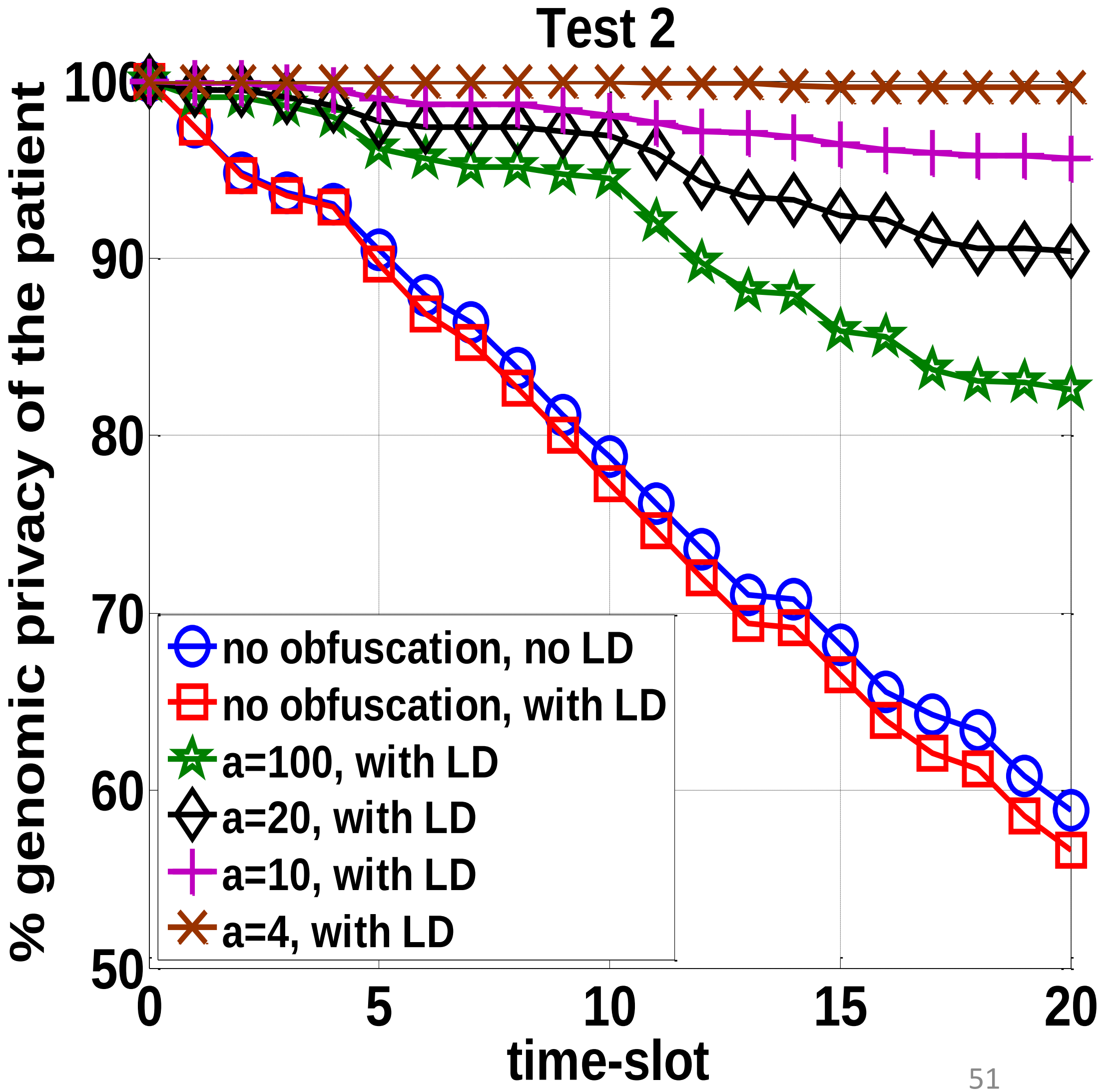
Policies for Test 1

- Delete the revealed SNPs from the MC after t time-slots
  - A set of SNPs in Σ are revealed as a result of Test 1 at time t_0
  - The SNPs in Σ are used to infer other SNPs (via LD) between (t_0, t_0+t)



Obfuscation for Test 2

- Provide the end-result as a range
  - Range can be determined via secure 2PC between the SPU and the MC
- E.g., divide the result range into a=4 ranges:
  - [0,0.25)
  - [0.25,0.5)
  - [0.5,0.75)
  - [0.75,1]

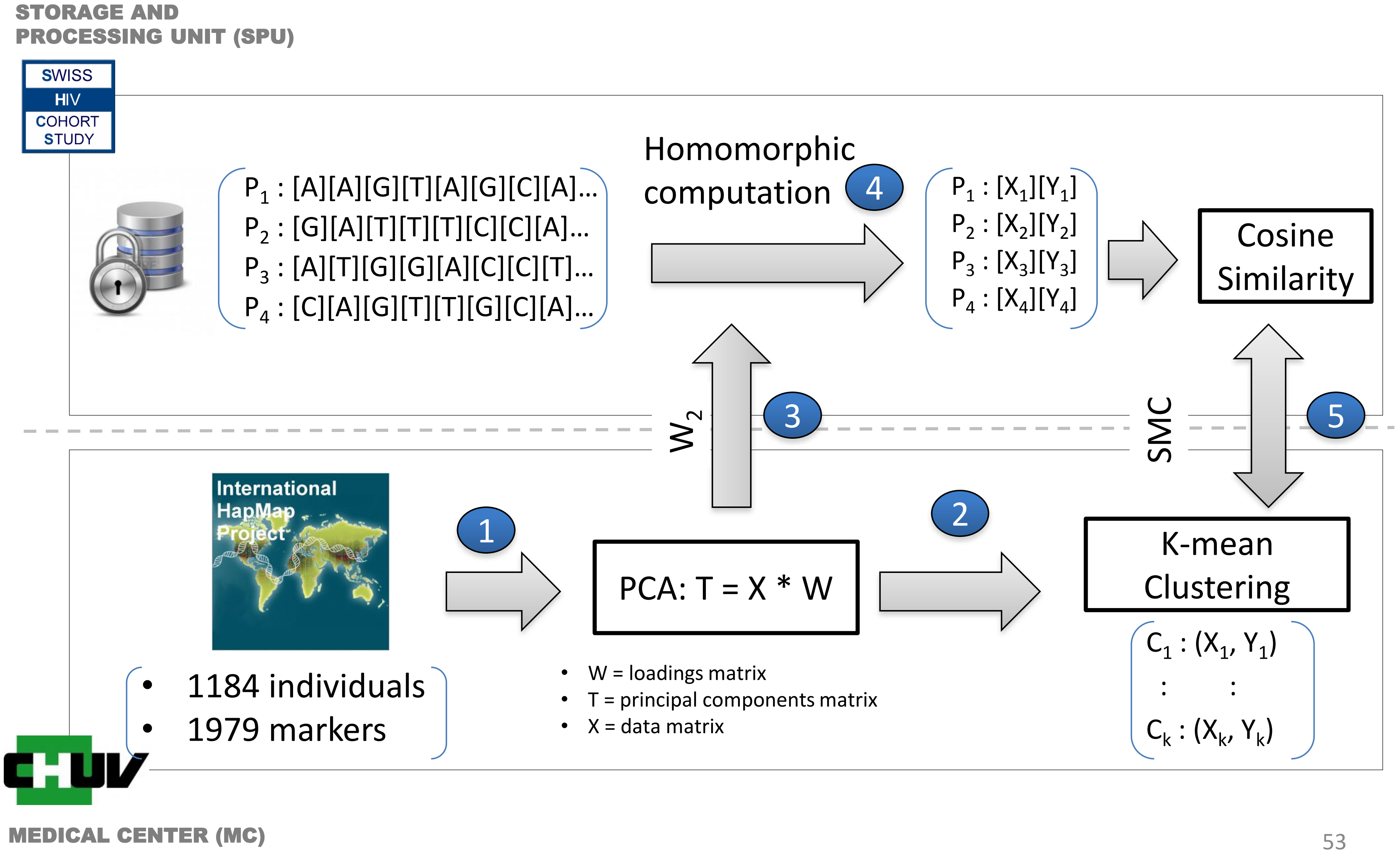


Implementation and Complexity

- Intel Core i7-2620M CPU with 2.70 GHz
- Windows 7
- MySQL 5.5 database
- Java programming language

	@CI	@SPU			@MC	
	Paillier Encryption	Proxy Re-encryption	Re-encryption under the Same Public Key	Storage	Homomorphic Operations	Paillier Decryption
Key Size=2K	0.049 ms./SNP	30 ms.	0.182 ms./SNP)	2.1 GB/patient	43 ms. (10 SNPs)	2 ms.
Key Size=4K	0.168 ms./SNP	42 ms.	0.658 ms./SNP	4.1 GB/patient	173 ms. (10 SNPs)	13 ms.

Privacy-Preserving Ancestry Inference



STORAGE AND PROCESSING UNIT (SPU)

SWISS HIV COHORT STUDY

P_1 : [A][A][G][T][A][G][C][A]…
P_2 : [G][A][T][T][T][C][C][A]…
P_3 : [A][T][G][G][A][C][C][T]…
P_4 : [C][A][G][T][T][G][C][A]…

Homomorphic computation 4

P_1 : $[X_1][Y_1]$
P_2 : $[X_2][Y_2]$
P_3 : $[X_3][Y_3]$
P_4 : $[X_4][Y_4]$

Cosine Similarity

W_2 3

SMC 5

International HapMap Project

- 1184 individuals
- 1979 markers

1

PCA: T = X * W

- W = loadings matrix
- T = principal components matrix
- X = data matrix

2

K-mean Clustering

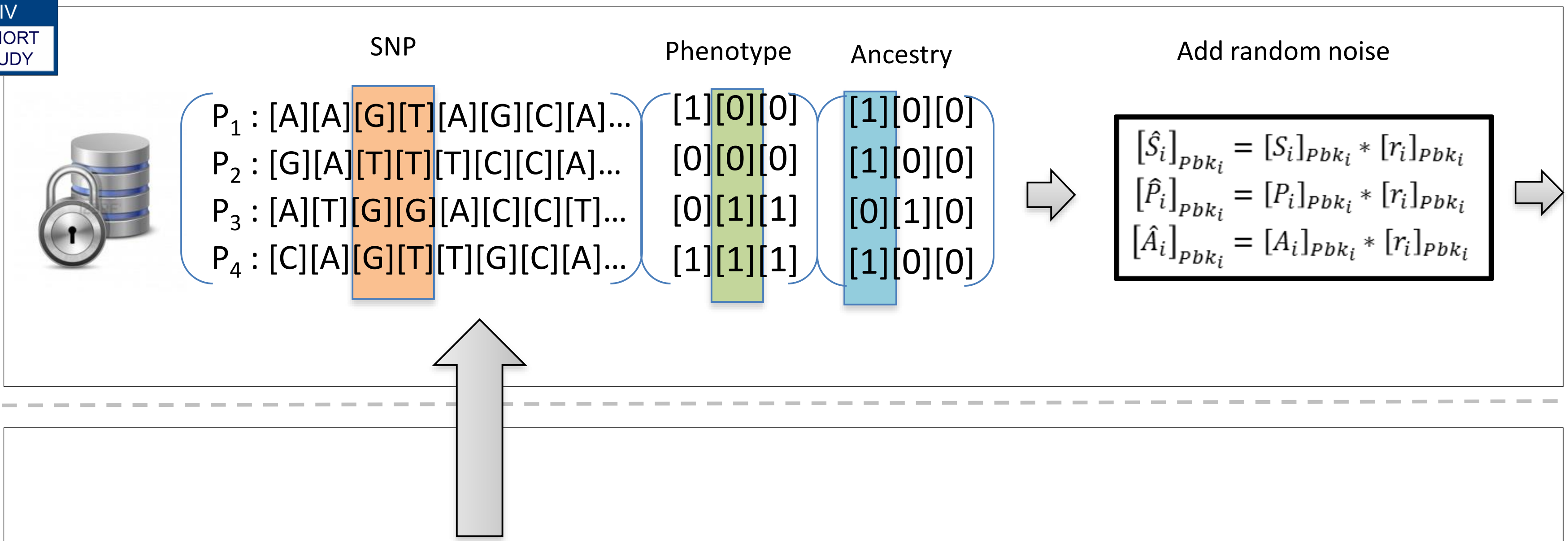
C_1 : (X_1, Y_1)
: :
C_k : (X_k, Y_k)

CHUV

MEDICAL CENTER (MC)

Privacy-Preserving GWAS

STORAGE AND PROCESSING UNIT (SPU)



SNP

P_1 : [A][A][G][T][A][G][C][A]…
P_2 : [G][A][T][T][T][C][C][A]…
P_3 : [A][T][G][G][A][C][C][T]…
P_4 : [C][A][G][T][T][G][C][A]…

Phenotype

[1][0][0]
[0][0][0]
[0][1][1]
[1][1][1]

Ancestry

[1][0][0]
[1][0][0]
[0][1][0]
[1][0][0]

Add random noise

$$[\hat{S}_i]_{Pbk_i} = [S_i]_{Pbk_i} * [r_i]_{Pbk_i}$$
$$[\hat{P}_i]_{Pbk_i} = [P_i]_{Pbk_i} * [r_i]_{Pbk_i}$$
$$[\hat{A}_i]_{Pbk_i} = [A_i]_{Pbk_i} * [r_i]_{Pbk_i}$$

GWAS: SNP = S_2; Phenotype = P_2; Ancestry = A_1

Notation:
- $SNP_i = S_i$
- $Phenotype_i = P_i$
- $Ancestry_i = A_i$
- [ ] = encrypted value

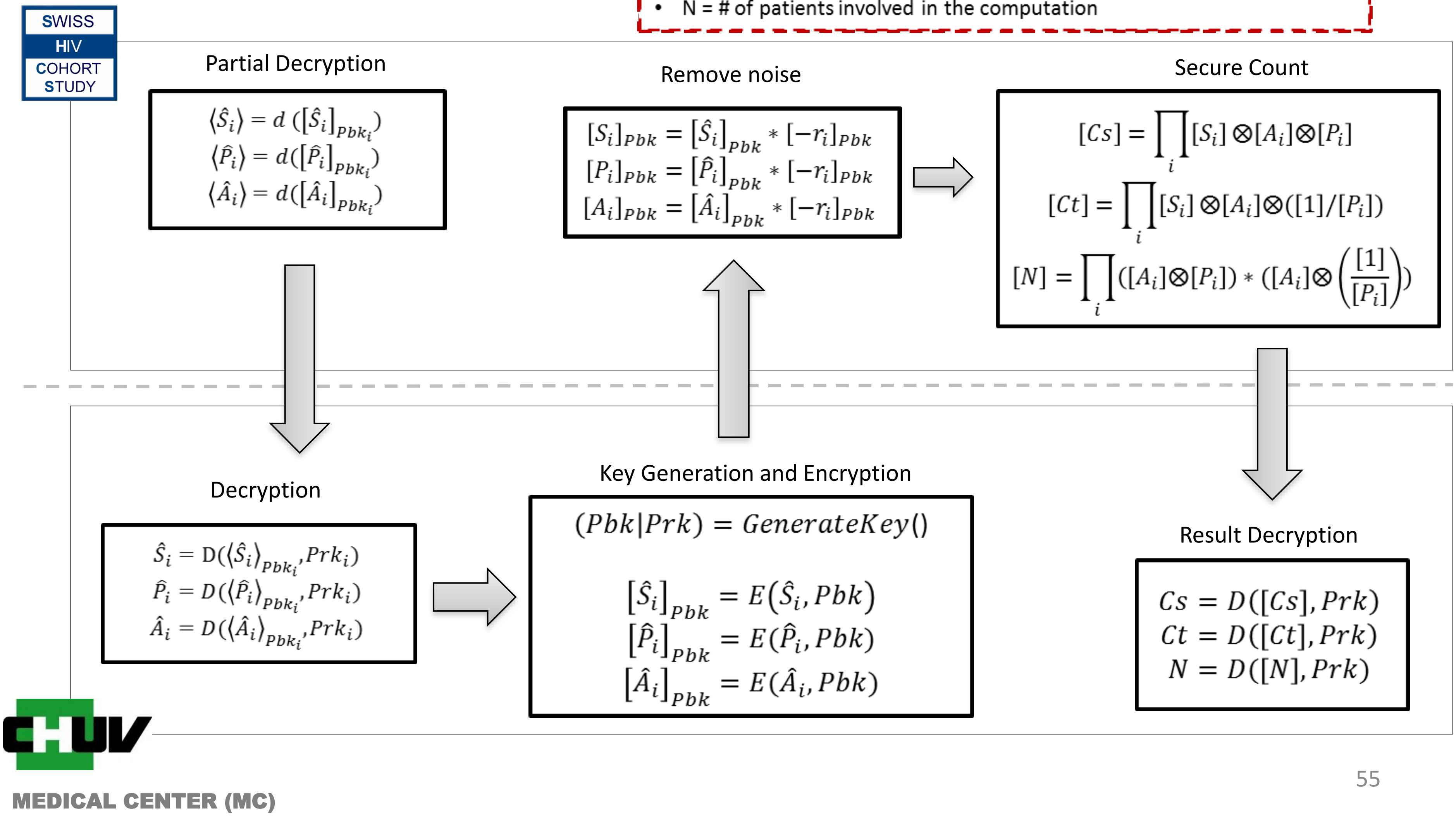
MEDICAL CENTER (MC)

Privacy-Preserving GWAS (cont.)

Notation:

- ⊗ = secure multiplication protocol
- Cs = # of observed alleles in cases group
- Ct = # of observed alleles in control group
- N = # of patients involved in the computation

STORAGE AND PROCESSING UNIT (SPU)



Partial Decryption

$$\langle \hat{S}_i \rangle = d\left([\hat{S}_i]_{Pbk_i}\right)$$
$$\langle \hat{P}_i \rangle = d\left([\hat{P}_i]_{Pbk_i}\right)$$
$$\langle \hat{A}_i \rangle = d\left([\hat{A}_i]_{Pbk_i}\right)$$

Remove noise

$$[S_i]_{Pbk} = [\hat{S}_i]_{Pbk} * [-r_i]_{Pbk}$$
$$[P_i]_{Pbk} = [\hat{P}_i]_{Pbk} * [-r_i]_{Pbk}$$
$$[A_i]_{Pbk} = [\hat{A}_i]_{Pbk} * [-r_i]_{Pbk}$$

Secure Count

$$[Cs] = \prod_i [S_i] \otimes [A_i] \otimes [P_i]$$
$$[Ct] = \prod_i [S_i] \otimes [A_i] \otimes ([1]/[P_i])$$
$$[N] = \prod_i ([A_i] \otimes [P_i]) * ([A_i] \otimes \left(\frac{[1]}{[P_i]}\right))$$

Decryption

$$\hat{S}_i = \mathrm{D}(\langle \hat{S}_i \rangle_{Pbk_i}, Prk_i)$$
$$\hat{P}_i = D(\langle \hat{P}_i \rangle_{Pbk_i}, Prk_i)$$
$$\hat{A}_i = D(\langle \hat{A}_i \rangle_{Pbk_i}, Prk_i)$$

Key Generation and Encryption

$$(Pbk|Prk) = GenerateKey()$$

$$[\hat{S}_i]_{Pbk} = E(\hat{S}_i, Pbk)$$
$$[\hat{P}_i]_{Pbk} = E(\hat{P}_i, Pbk)$$
$$[\hat{A}_i]_{Pbk} = E(\hat{A}_i, Pbk)$$

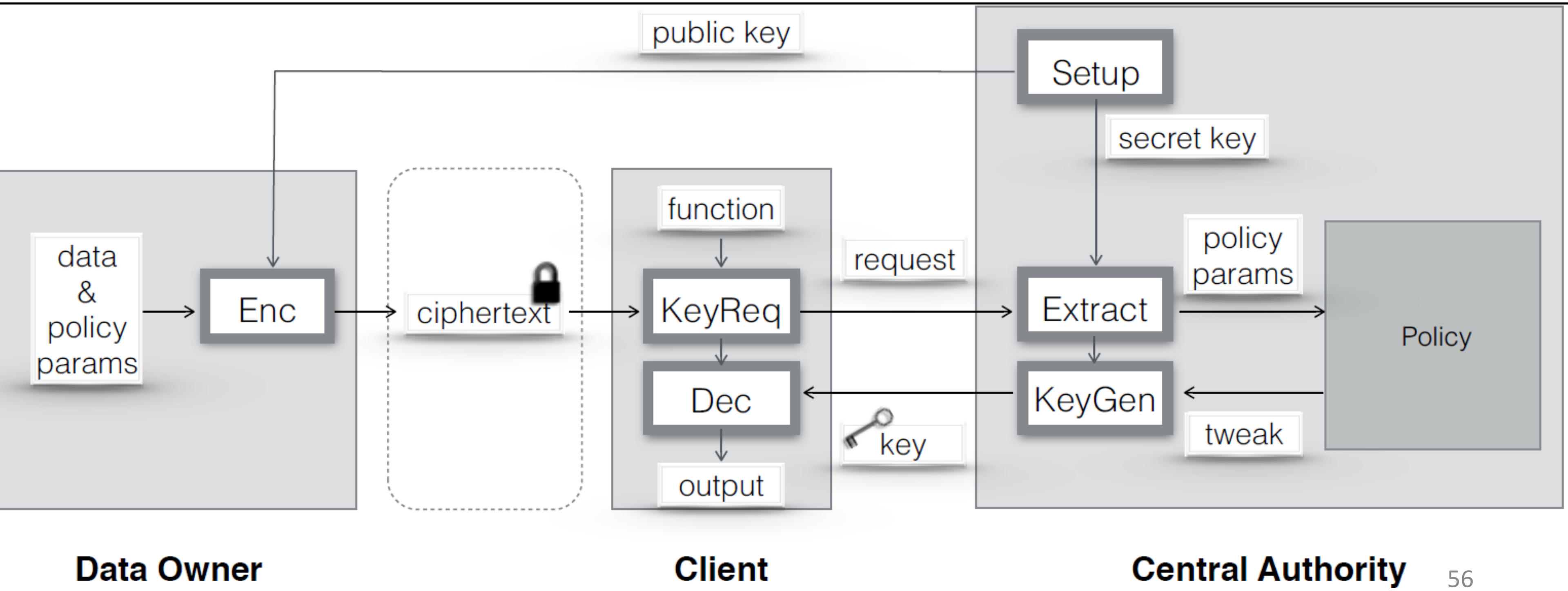
Result Decryption

$$Cs = D([Cs], Prk)$$
$$Ct = D([Ct], Prk)$$
$$N = D([N], Prk)$$

MEDICAL CENTER (MC)

Functional Encryption

- Similarity between genome sequences
  - Genomic data sharing
  - Finding similar patients



OPTIMIZATION

Information Theoretical Privacy - Back to Henrietta Lacks

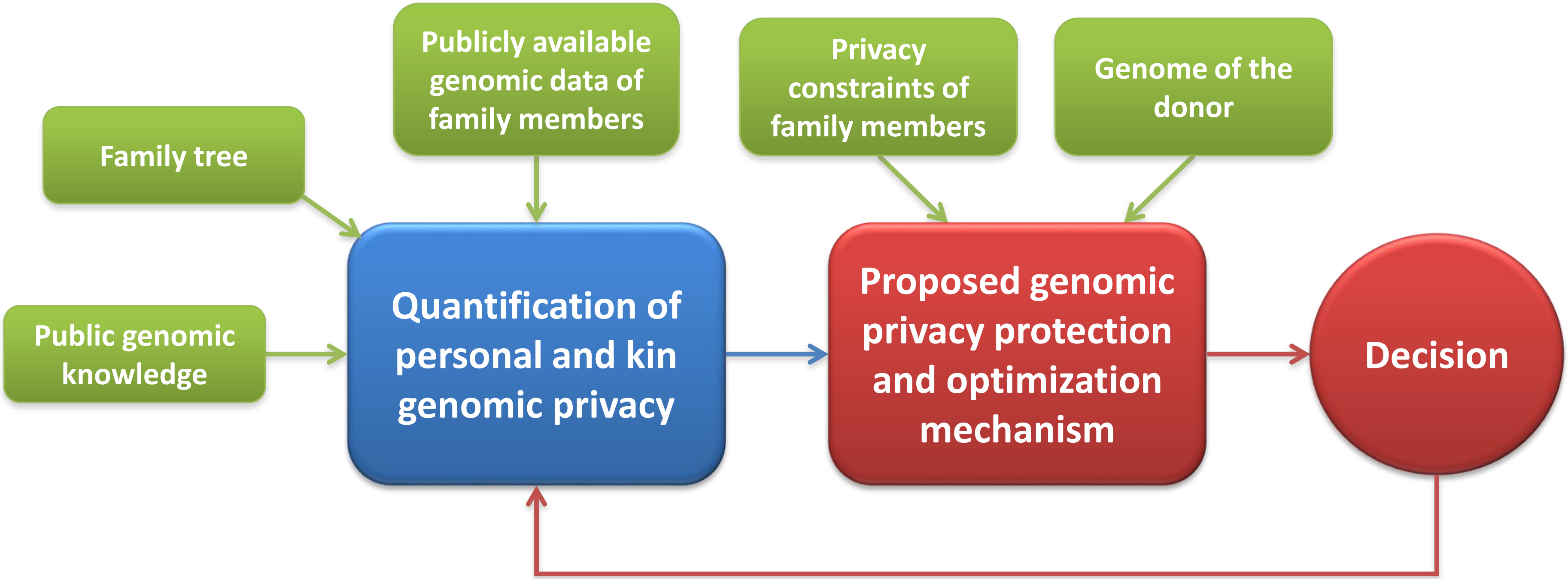
- Agreement between the Lacks Family and NIH
- Gives some control to Lacks Family over how *HeLa Genome* is used
  - Working group in NIH reviewing applications

"There i icy."

ne Mayo Clinic

- It is im of scienti with these i

Protecting Kin Genomic Privacy via Optimization


Publicly available genomic data of family members
Privacy constraints of family members
Genome of the donor
Family tree
Public genomic knowledge
Quantification of personal and kin genomic privacy
Proposed genomic privacy protection and optimization mechanism
Decision

- Decision maker(s): family member(s)
  - One member (*donor*) reveals his genome
  - Other members already (partially) shared their genomic data on the Internet
  - All members have *privacy constraints*
- Decision variables: SNPs to be revealed or not

Goals

- Protect the genomic and health privacy of individuals, considering their personal privacy requirements
  - Each individual has a personal genomic (or health) privacy constraint
  - The donor wants to make sure that both his own privacy constraints and these of his family members' are met after he shares part of his genome
- Make as much genomic data publicly available as possible for genomic research
  - The donor wants to share as much genomic data (e.g., SNPs) as possible
- Potential use:
  - NIH would not need a working group to control the access to the HeLa genome

Optimization Model

- **Assumption:** Independent SNPs

x_1: decision variables = binary vector of length *M*

$$\max_{\boldsymbol{x_1}} \sum_i u_i x_1^i$$

$$\text{subject to} \quad \left.\frac{1}{\sum_i g_1^i} \sum_i g_1^i x_1^i \le \tau_1 \right\}$$

genomic privacy of F1

$$\left.\frac{1}{\sum_{k \in \boldsymbol{S_d}} c_k} \sum_{k \in \boldsymbol{S_d}} c_k x_1^k \le \tau_1^d \,, \forall\, d \in \boldsymbol{D} \right\}$$

health privacy of F1
for any disease *d*

$$\left.\begin{aligned} &1 - \frac{1}{M} \sum_{j=1}^{M} \sum_{v_{r,i}^j} \overbrace{\Pr\left(v_{r,i}^j \middle| x_1^j v_1^j\right)}^{\sim \alpha_i^j x_1^j \text{ => linear constraint}} d\left(v_{r,i}^j, v_i^j\right) \le \tau_i \,, \forall\, i \in \boldsymbol{F}, i \ne 1 \\ &1 - \frac{1}{\sum_k c_k} \sum_k \sum_{v_{r,i}^k} c_k \underbrace{\Pr\left(v_{r,i}^k \middle| x_1^k v_1^k\right)}_{\sim \alpha_i^k x_1^k \text{ => linear constraint}} d\left(v_{r,i}^k, v_i^k\right) \le \tau_i^d \,, \forall\, i \in \boldsymbol{F}, i \ne 1, \forall\, d \in \boldsymbol{D} \end{aligned}\right\}$$

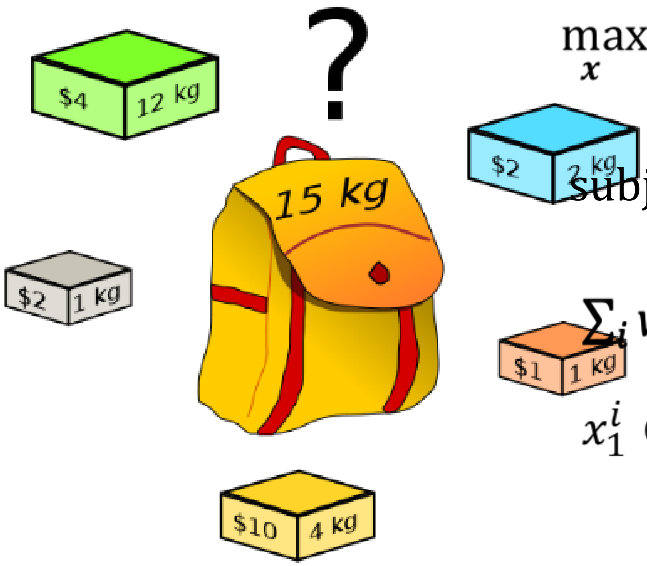
genomic and
health privacy
of other
family
members

$$x_1^j \in \{0,1\}, \forall\, \mathrm{j} \in \{1, \dots, \mathrm{M}\}$$

Solving the Optimization Model

- The Knapsack Problem:

You are given a container with a limited weight capacity, and some items which each have a weight and a value. Choose which items to place in the container such that the weight limit is not exceeded, but the total value of the items is as large as possible.



$$\max_{x} \sum_{i} p_i x_i$$

subject to

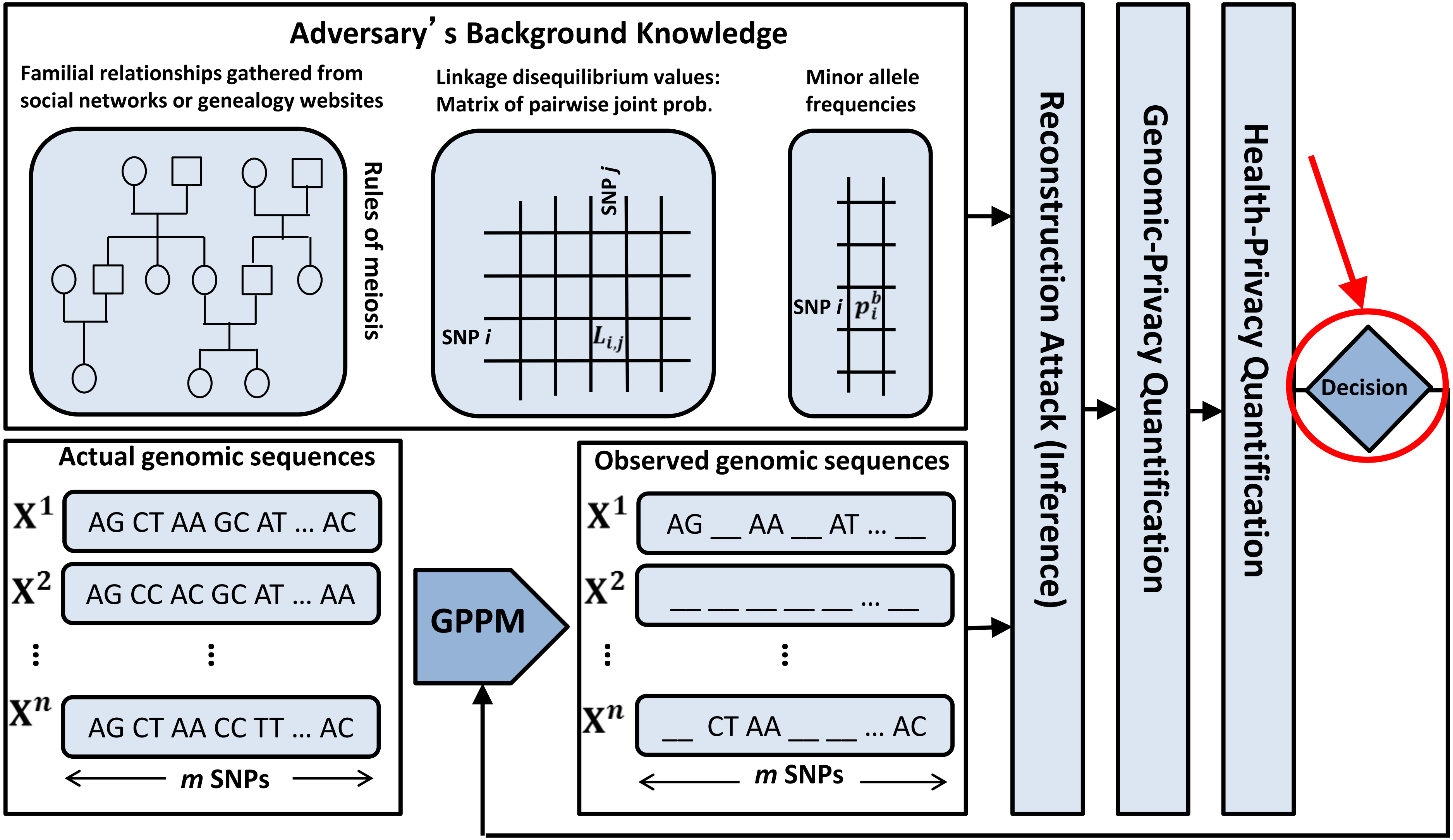
$$\sum_i w_{ij} x_i \leq W_j \ \forall j \ \in \{1, \ldots$$

$$x_1^i \in \{0,1\}, \forall i \in \{1, \ldots, n\}$$

Multidimensional 0-1 Knapsack Problem

- Exact methods
  - Often based on dynamic programming and *branch-and-bound algorithm*
  - Scales linearly with the number of constraints
- Heuristics
  - Competitive alternative to exact methods, especially when the number of constraints is large
  - Can achieve lower time complexity while still providing good (but not necessarily exact) solutions

Back to the Framework

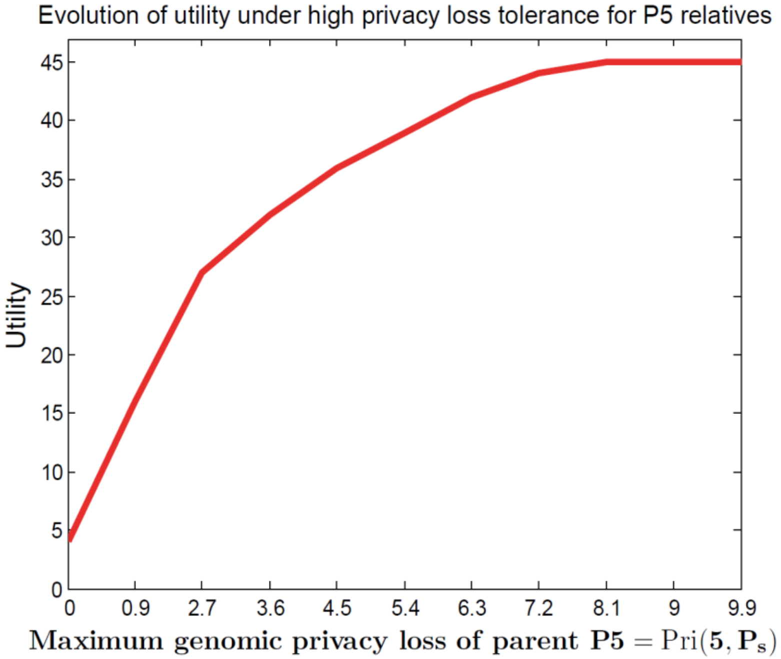


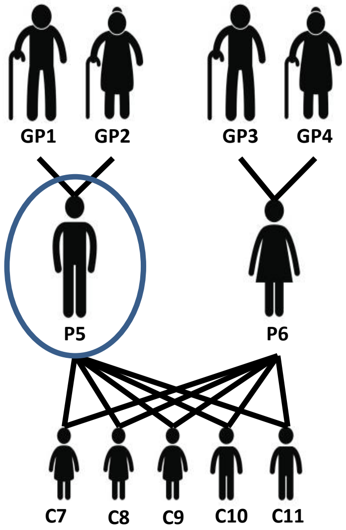
Methodology

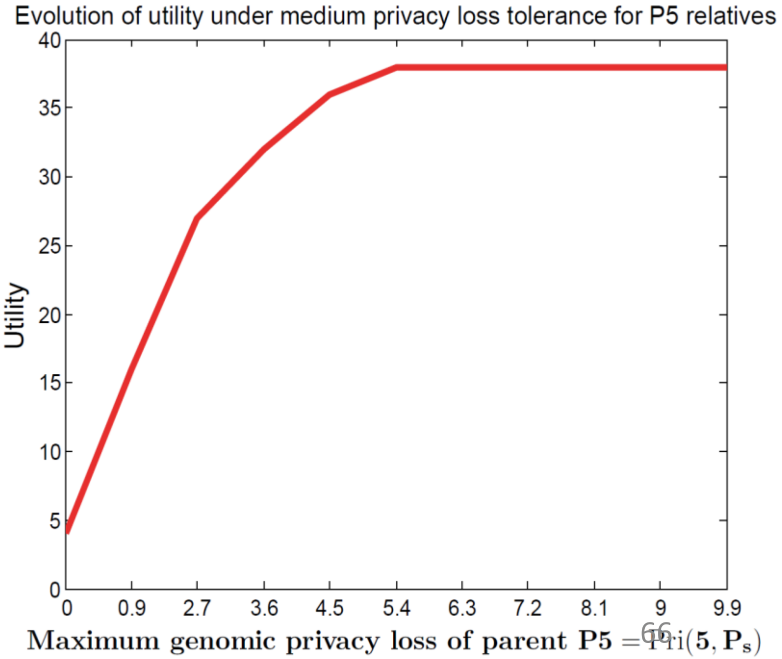
- Optimization using *branch-and-bound algorithm*
  - Independent SNPs (no LD)
    - Familial relationships affect privacy much more than the LD
  - Obtain the first result
    - Set of donor's SNPs that can be publicly revealed
    - Privacy constraints are satisfied
- Iterative Fine-Tuning
  - Using LD
  - Inference Algorithm Quantification
    - Check the privacy constraints again
  - Reveal or hide more SNPs
  - Iterate until privacy constraints are satisfied again

Results

- Focus on 50 SNPs
  - Utility: number of SNPs publicly revealed out of 50
- One genomic privacy constraint for each member
  - Each member is tolerant to *high privacy loss*
  - Each member is tolerant to *medium privacy loss*


Evolution of utility under high privacy loss tolerance for P5 relatives




Evolution of utility under medium privacy loss tolerance for P5 relatives

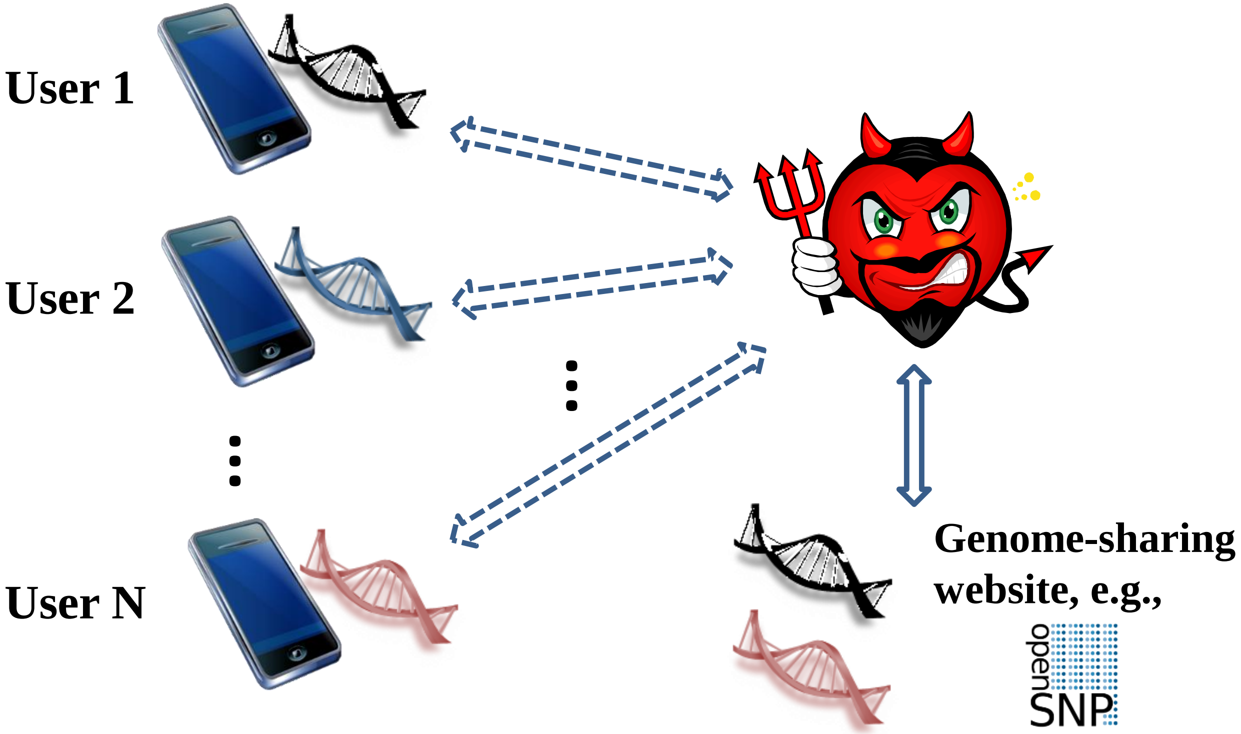
INTERDEPENDENT GENOMIC PRIVACY

Interdependent Privacy Game

- Privacy of family members is inherently interdependent
- If family members are not cooperative (i.e., selfish), then they put other relatives' privacy at risk => externalities
- Similar to «interdependent security (IDS) games» [1]

[1] Laszka A., Felegyhazi M., and Buttyan L., A Survey of Interdependent Security Games, submitted to ACM CSUR, November 2012

System Model



Assumptions:

•Users storing their genomes (SNPs) in a mobile device (smartphone, tablet) for various benefits (cf. [2])

•Some users also publicly sharing their genomic data (or sharing them with untrusted parties?), «genetic exhibitionism»

[2] De Cristofaro et al., GenoDroid: Are Privacy-Preserving Genomic Tests Ready for Prime Time?, WPES' 12

Current Deployments

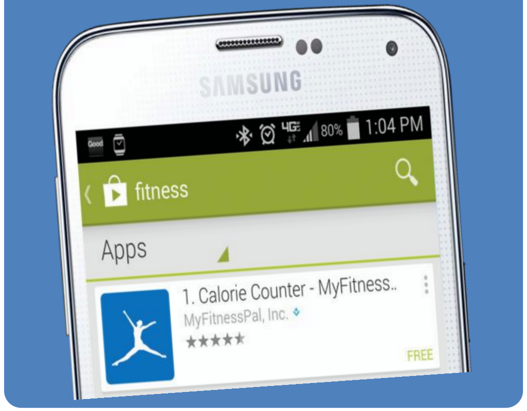
- Swiss HIV Cohort Study:
  - Infrastructure supporting multi-center research project dealing with HIV infected adults
    - Participating clinics of 7 Swiss hospitals
    - Coordination and data center based in Lausanne
      - http://www.shcs.ch/
- Lausanne University Hospital (CHUV)
  - Protection of CHUV biobank 2015
    - Clinical and environmental data
    - Genomic data:
      - 2.5M SNPs / patient
      - 20’ 000 patients
      - http://www.chuv.ch/biobanque
  - Mobile Android App for Doctors: GenoPri
- Sophia Genetics
  - Start-up company, on campus; visualization of genomic data
  - Our contribution: protection of raw genomic data
  - http://www.sophiagenetics.com

Future Research on Security and Privacy for Healthcare Data

- Cryptographic and non-cryptographic solutions
  - Differential privacy, membership privacy
  - Trade-off between privacy and utility
- Inference attacks and mitigations
  - Using genomic and non-genomic data
  - Genotype ↔ Phenotype
- Dynamic access control and database privacy
  - ORAM, PIR for healthcare data
- Protection against different attack models
  - Stronger attacker models
- Economics
  - Incentive of the attacker
- Practical implementations
- Credibility (authenticity) of a genome
- Privacy budget and genomic data sharing

Future Research on Security and Privacy of E-health Platforms

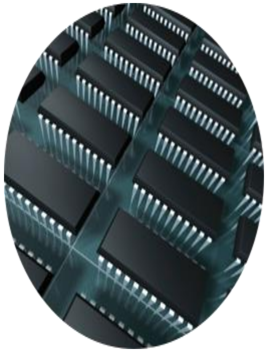
5 YRS AGO



TODAY

NEAR FUTURE

MOBILE



SEMICON & HARDWARE

SENSORS

ALGORITHMS

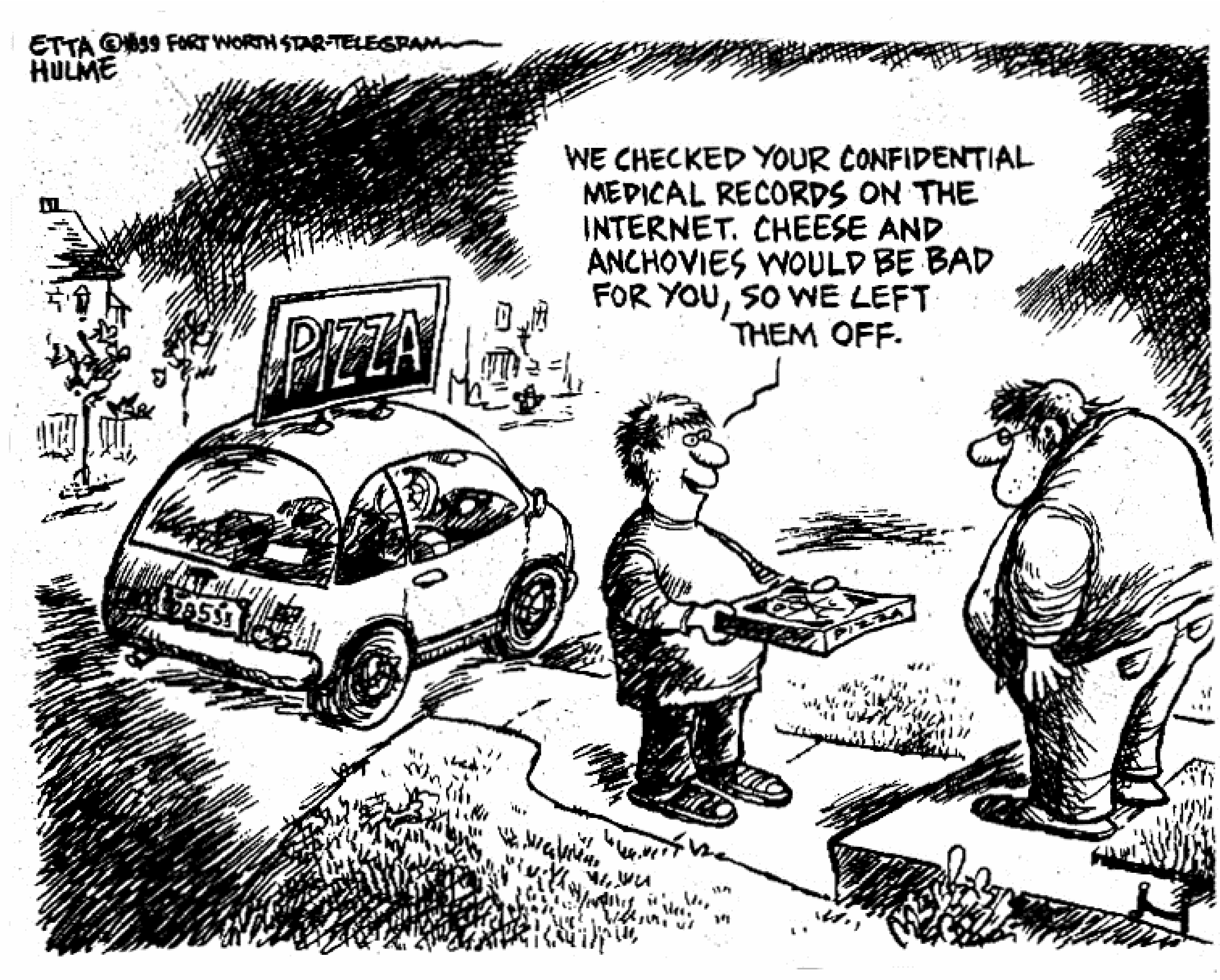
CLOUD & BIG DATA

BEHAVIORAL SCIENCE

Conclusion

- Digital medicine is coming
- It will *forever* change the landscape of privacy protection
- Very few researchers have addressed the topic of genome privacy
  - Much more needs to be done in this field

- Our contributions:
  - Inference attacks and quantification
  - Techniques to protect genomic privacy
  - Real-life deployments (hospitals, biobanks, industry)
  - Workshop on Genome Privacy
    - 2014 with PETS, 2015 with IEEE S&P
  - Dagstuhl Seminar on Genomic Privacy (2013 and 2015)


ETTA ©1999 FORT WORTH STAR-TELEGRAM
HULME
PIZZA
WE CHECKED YOUR CONFIDENTIAL MEDICAL RECORDS ON THE INTERNET. CHEESE AND ANCHOVIES WOULD BE BAD FOR YOU, SO WE LEFT THEM OFF.
PIZZA