Health Information Technology

CS463/ECE424

University of Illinois

Outline

Privacy attacks on genomic data

Privacy protecting genomic research

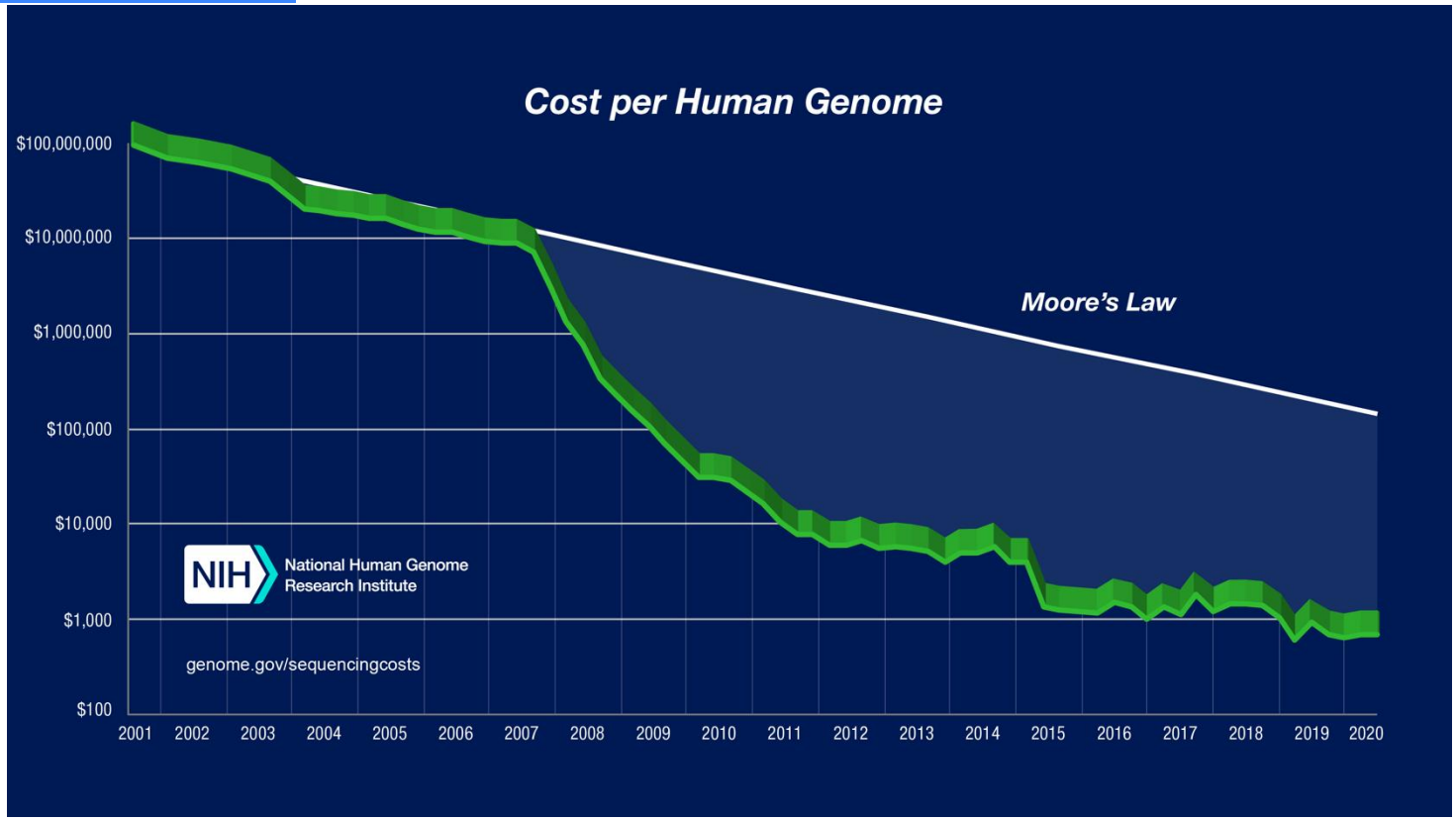
Genome

- Contains all of the biological information needed to build and maintain a “living example” of an organism

- Encoded in **DNA**, one polymer of **nucleotides**
  - A,G,C,T

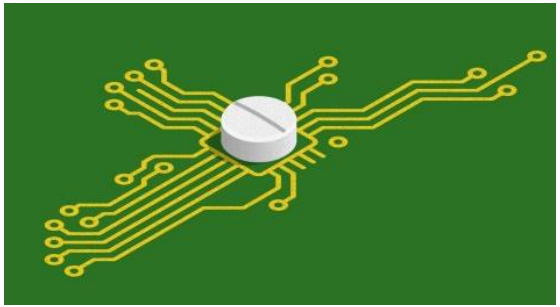
- Human Genome:
  - Approximately **3 billion nucleotides**
  - Stored in **23 chromosome pairs** (plus mtDNA)

Cost Per Genome



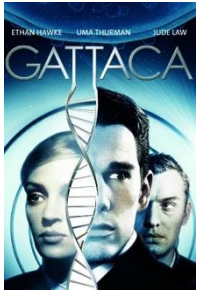
New Frontiers

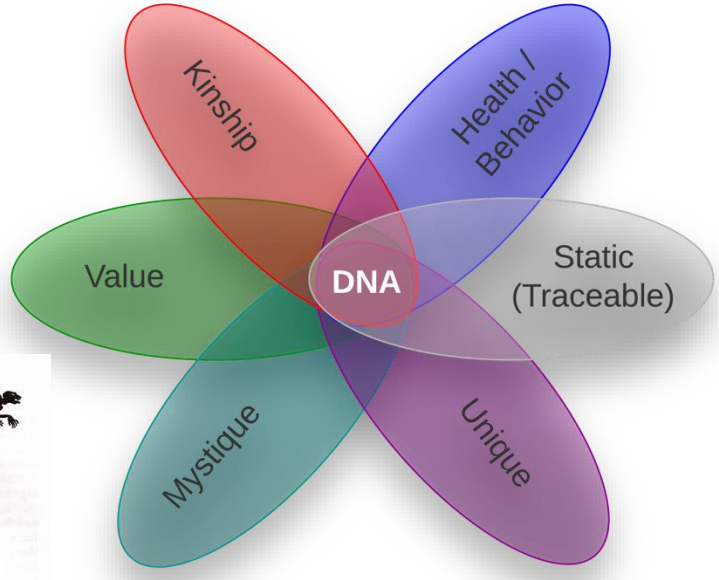
- Better understanding of human genome
- Many individuals have access to key parts of their genomes
  - Precision medicine enabled
- Testing possible not only in-vitro but also via sumulation



Genetic Exceptionalism

How Special is Genomic Data?





McGuire, Amy L., Rebecca Fisher, Paul Cusenza, Kathy Hudson, Mark A. Rothstein, Deven McGraw, Stephen Matteson, John Glaser, and Douglas E. Henley. "Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider." *Genetics in Medicine* 10, no. 7 (2008): 495-499.

Evans, James P., and Wylie Burke. "Genetic exceptionalism. Too much of a good thing?." *Genetics in Medicine* 10, no. 7 (2008): 500-501.

Privacy Concerns

- Genomic data carry sensitive information that may reveal
  - Identity
  - Predisposition to diseases
  - Facial features …
- Disclosure may propagate the privacy risks to blood relatives.
- Data are irrevocable once they are disseminated
- New privacy threats may emerge over time with new discoveries of human genetics and the advance of attack methods.
  - Aggregated results removed from the public domain hosted by NIH.

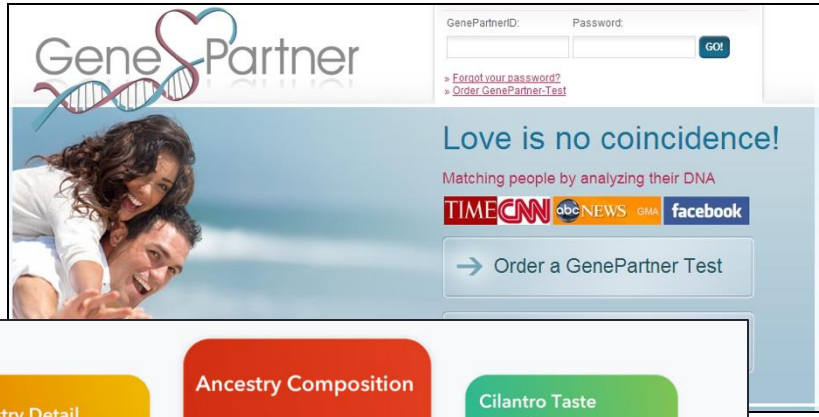
Genomic Privacy Attack

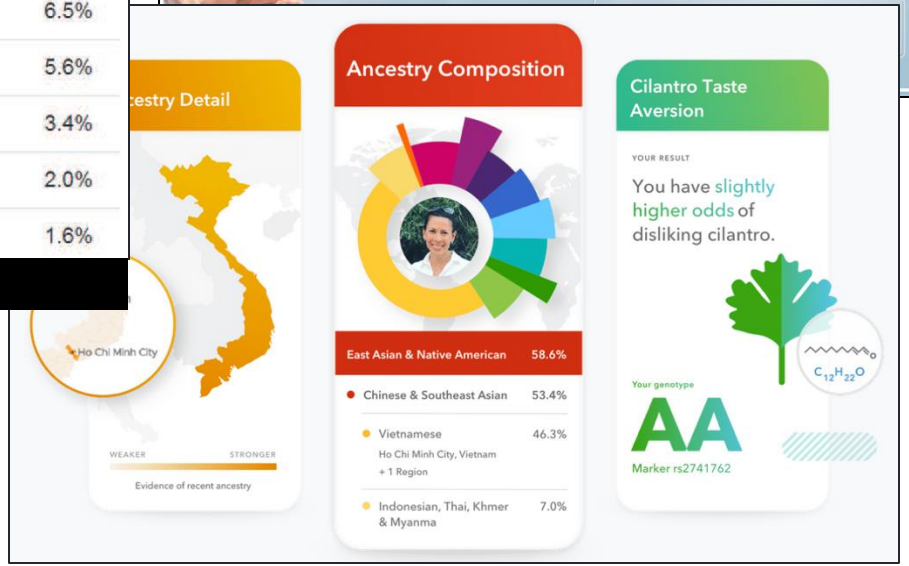
Quantification of Kin Genomic Privacy, CCS 2013

Sharing Genomic Data Online?

Name	Confidence	Your Risk	Avg. Risk
Atrial Fibrillation	★★★★	33.9%	27.2%
Prostate Cancer ♂	★★★★	29.3%	17.8%
Alzheimer's Disease	★★★★	14.2%	7.2%
Age-related Macular Degeneration	★★★★	11.1%	6.5%
Colorectal Cancer	★★★★	7.8%	5.6%
Chronic Kidney Disease	★★★★	4.2%	3.4%
Restless Legs Syndrome	★★★★	2.5%	2.0%
Parkinson's Disease	★★★★	2.2%	1.6%

23andme.com.

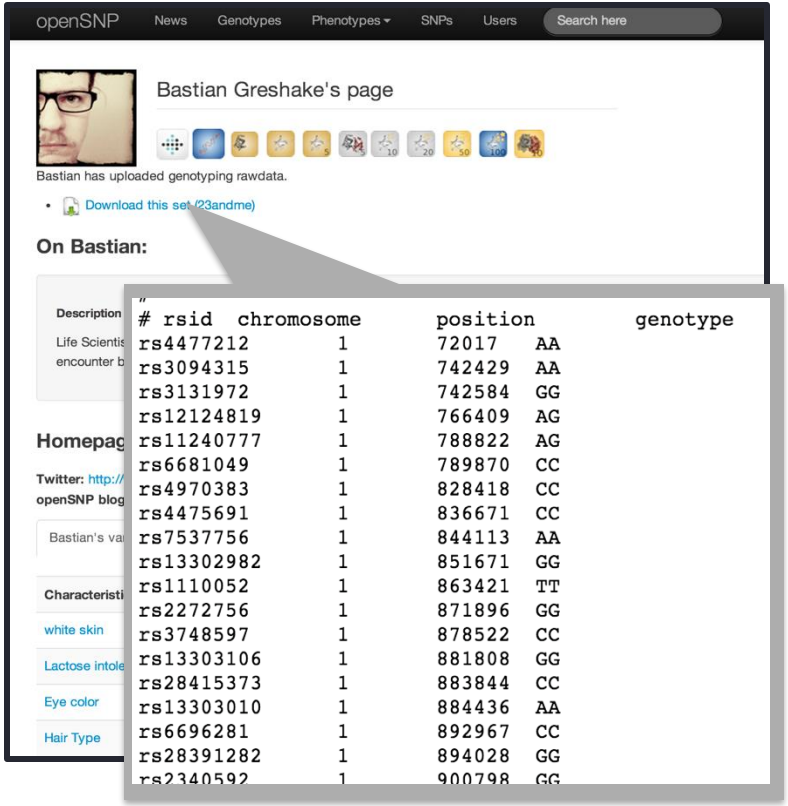


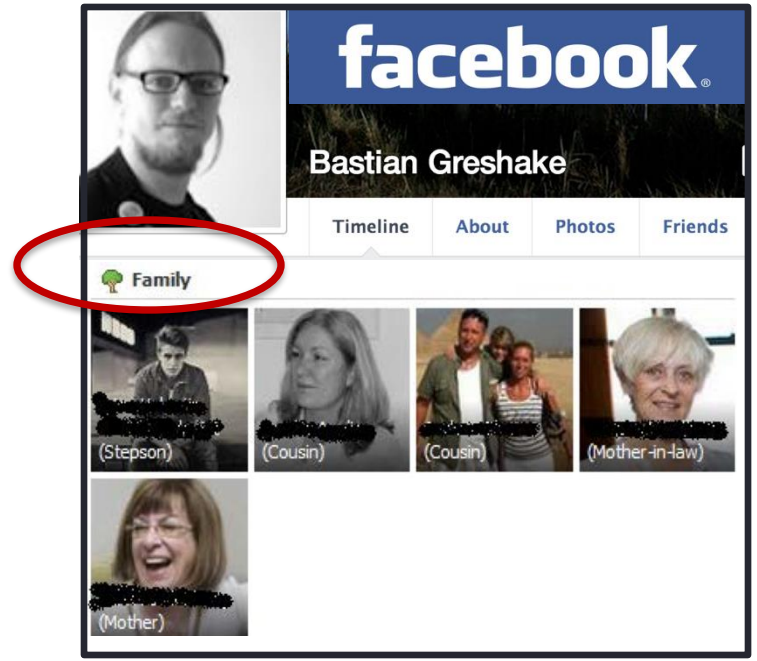


Privacy Concerns

- Kin genomic privacy
  - Correlated genomic info between family members
  - Partial leakage of one member → threaten the whole family
- Example threat: blackmail, denial of insurance, discrimination …
- How much is the individual's genomic privacy threatened by their relatives revealing their genomes?

Kin Genomic Privacy

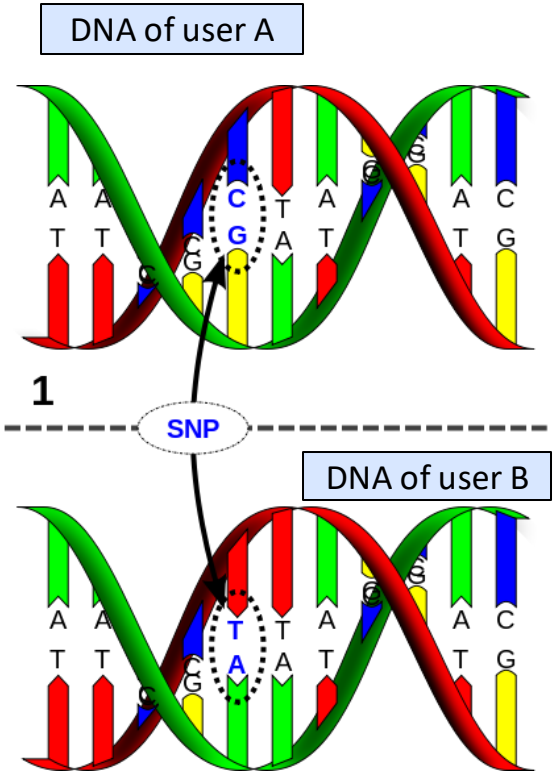




SNP

- Human DNA sequence
  - Identical at 99.5% of the positions

- SNP (Single Nucleotide Polymorphism)
  - Positions where a nucleotide is different between people
  - Define physical characteristics, indicator of diseases
  - 50 million SNP positions



Rules of Reproduction

B: Major allele
b: Minor allele

SNP_i

		B		
		b		

Father

Disease!

SNP_i

		b		
		b		

Mother

SNP_i

		B		
		b		

P(Bb)=1/2

		b		
		b		

P(bb)=1/2

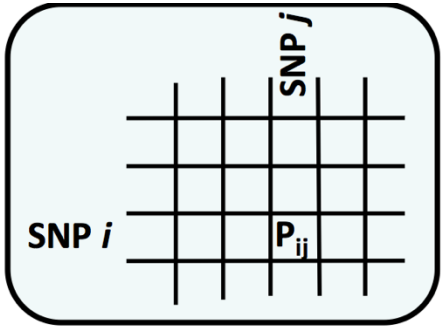
P(BB)=0

(P(BB), P(Bb), P(bb))

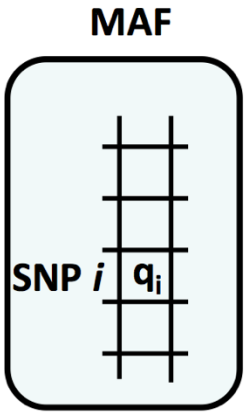
		Father (*F*)		
		BB	Bb	bb
Mother (*M*)	BB	(1,0,0)	(0.5,0.5,0)	(0,1,0)
	Bb	(0.5,0.5,0)	(0.25,0.5,0.25)	(0,0.5,0.5)
	bb	(0,1,0)	(0,0.5,0.5)	(0,0,1)

Linkage Disequilibrium (LD)

- SNPs are NOT entirely independent
  - A given SNP value can be inferred from other SNPs
  - Results of existing genetic research, publicly known



- MAF (minor allele frequencies)
  - Also public knowledge, from medical research

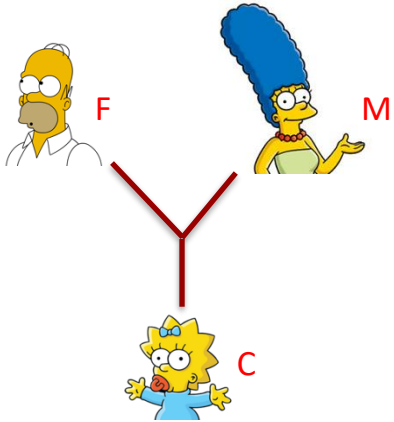


Attack Model

- Using your relatives (partial) SNP values to infer yours
- Attacker knows
  - Family member relationships: Social network sites
  - Partial SNPs of a subset of family members: Genome sharing sites
  - Linkage Disequilibrium *(LD) :* Public available knowledge
  - Minor allele frequencies (MAF): Public available knowledge

- Attacker's goal
  - Infer all family members' unknown SNPs

Inference Algorithm: A Toy Exampl



- **n** family members, each has **m** SNPs
  - n = 3, father, mother, child
  - m = 3, three SNPs
  - Each SNP: three possible values: (BB, Bb, bb), denoted as (0, 1, 2)

	SNP_1	SNP_2	SNP_3
F	?	0	1
M	2	1	?
C	?	?	?

P(x=0)=0
P(x=1)=1
P(x=2)=0

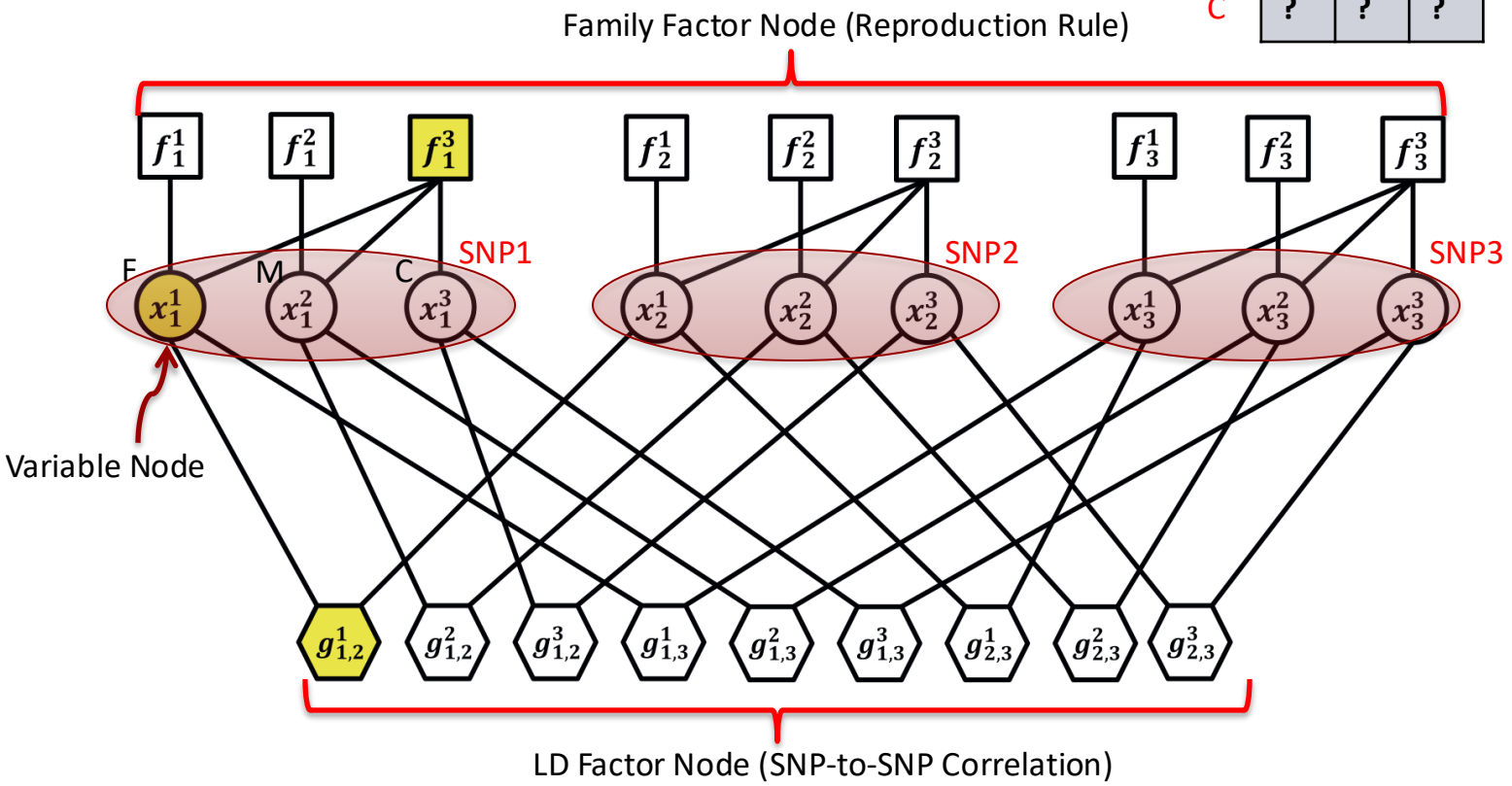
P(x=0)=?
P(x=1)=?
P(x=2)=?

marginal probability distribution $P(x_j^i | X_K)$

- Step1: Construct a factor graph
- Step2: Belief propagation: update node value until converge

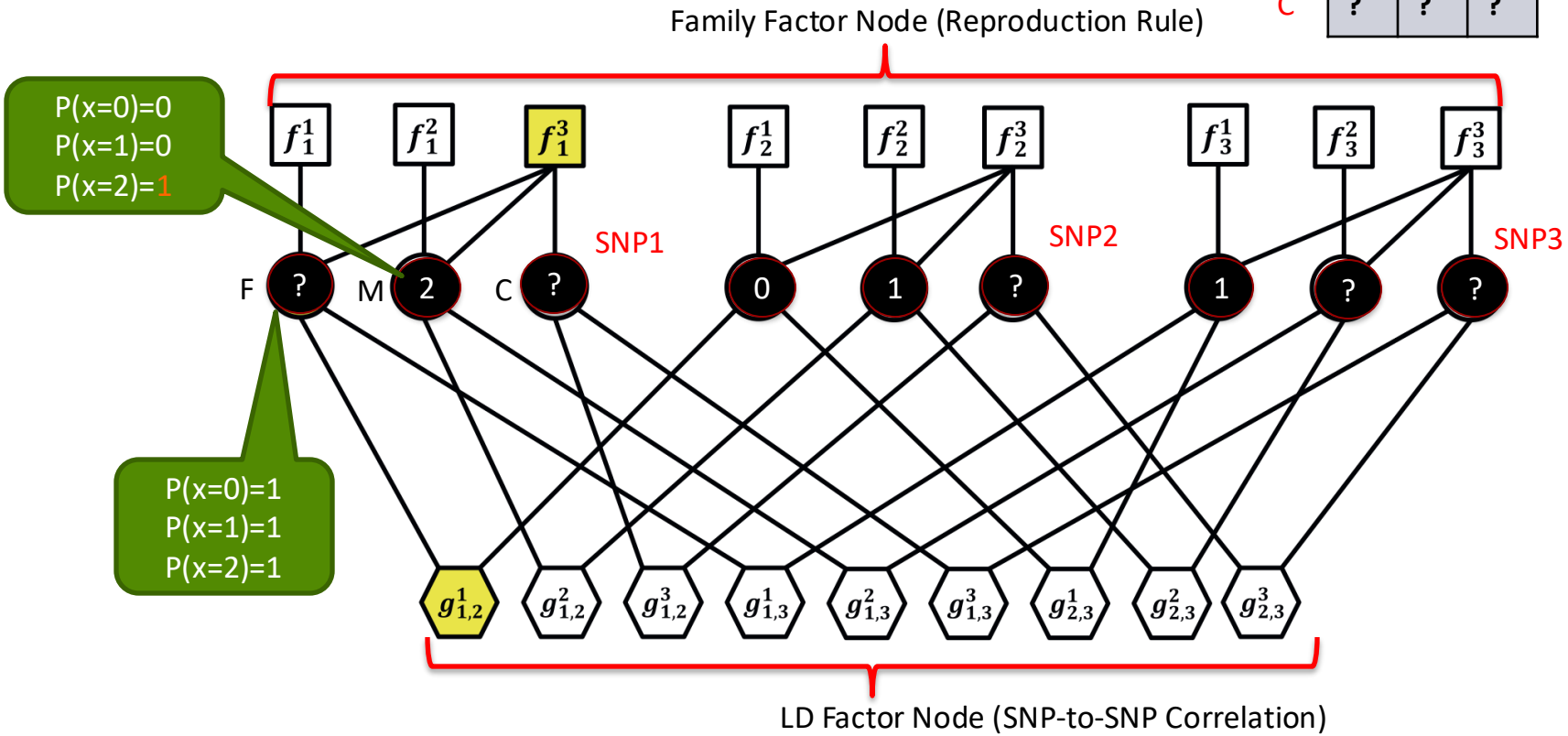
Step1: Factor Graph

	SNP_1	SNP_2	SNP_3
F	?	0	1
M	2	1	?
C	?	?	?

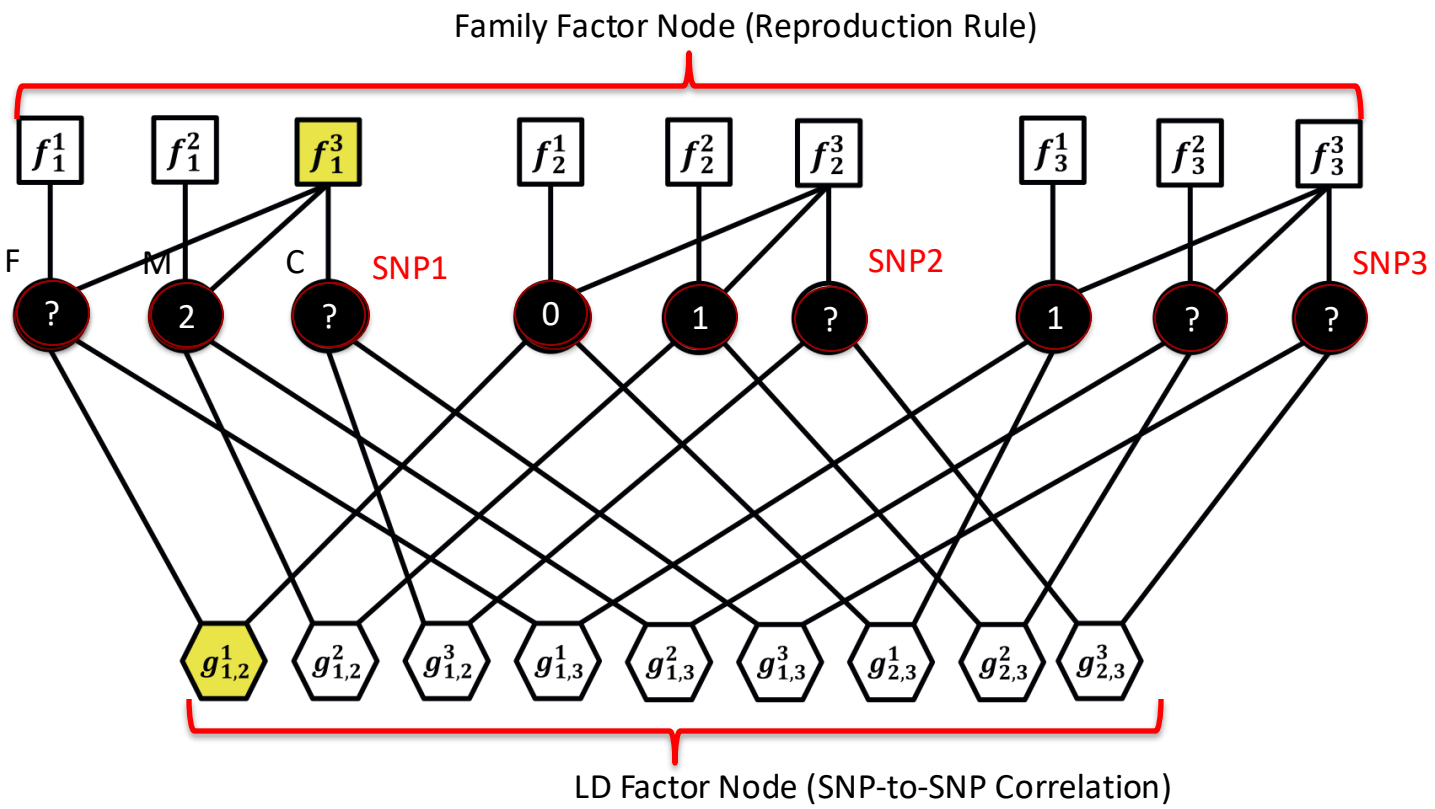


Step2: Belief Propagation

	SNP_1	SNP_2	SNP_3
F	?	0	1
M	2	1	?
C	?	?	?

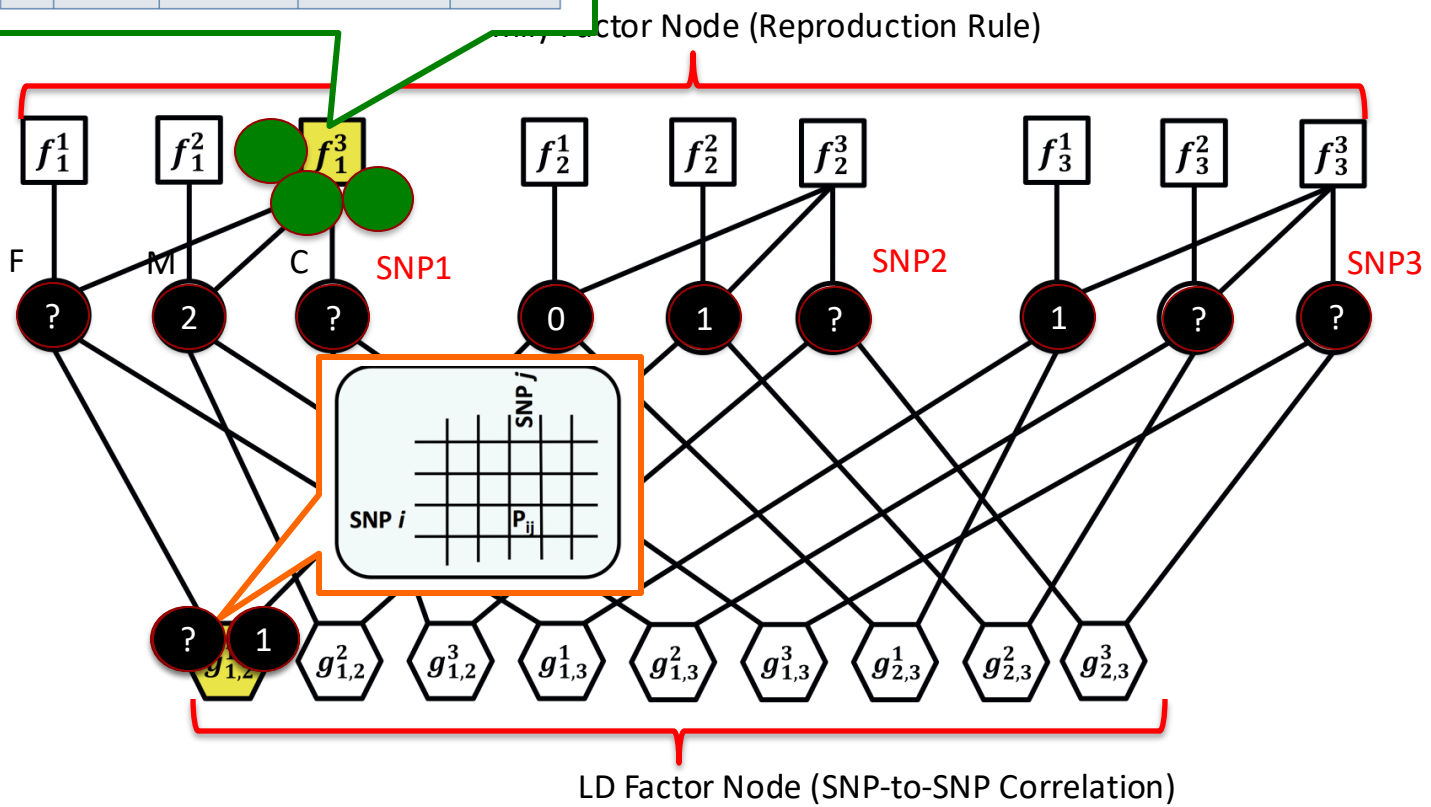


Step2: Belief Propagation



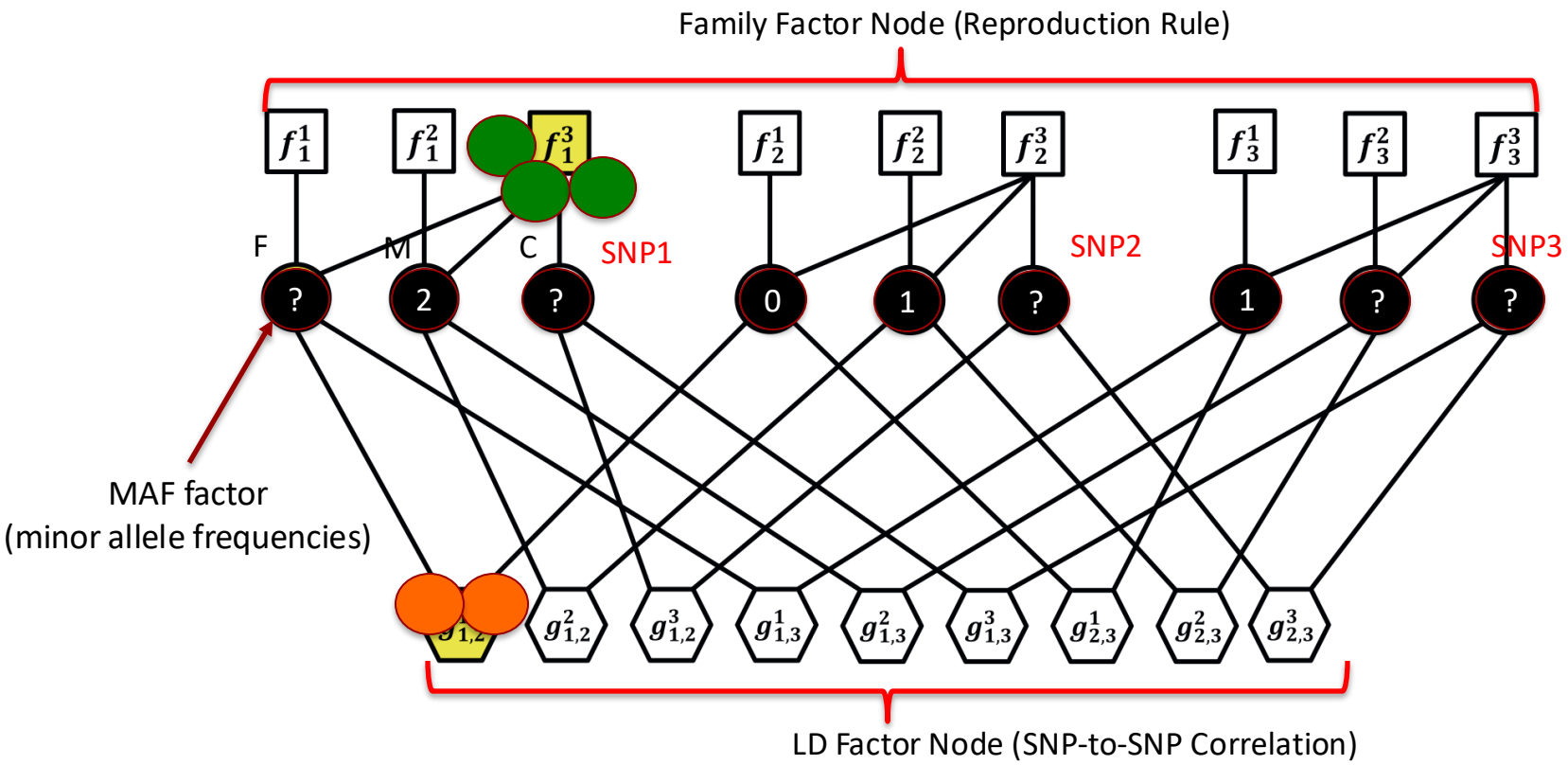
Step [illegible] gation

		Father (*F*)		
		BB	Bb	bb
Mother (*M*)	BB	(1,0,0)	(0.5,0.5,0)	(0,1,0)
	Bb	(0.5,0.5,0)	(0.25,0.5,0.25)	(0,0.5,0.5)
	bb	(0,1,0)	(0,0.5,0.5)	(0,0,1)



Step2: Belief Propagation

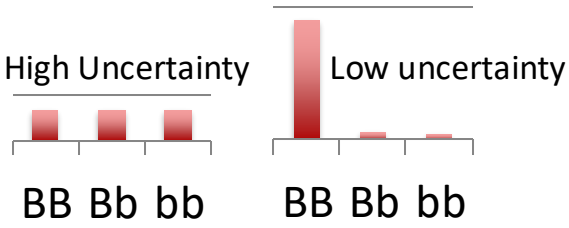
10-15 iterations until converge



Genomic Privacy Metrics

- **Estimation Error (E)**
  - Expected estimation error
  - Needs ground-truth
- **Uncertainty (H)**
  - Entropy of estimated distribution
  - Don't require ground-truth
- **Mutual Information (I)**
  - Mutual dependency between *unknown* SNP and *observed* SNP
  - Privacy decreases with mutual information

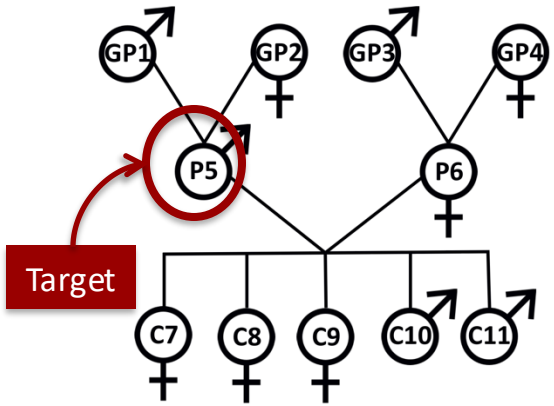
$$E_j^i = \sum_{x_j^i \in \{0,1,2\}} p(x_j^i | \mathbb{X}_{\mathrm{K}}) || x_j^i - \hat{x_j^i} ||$$

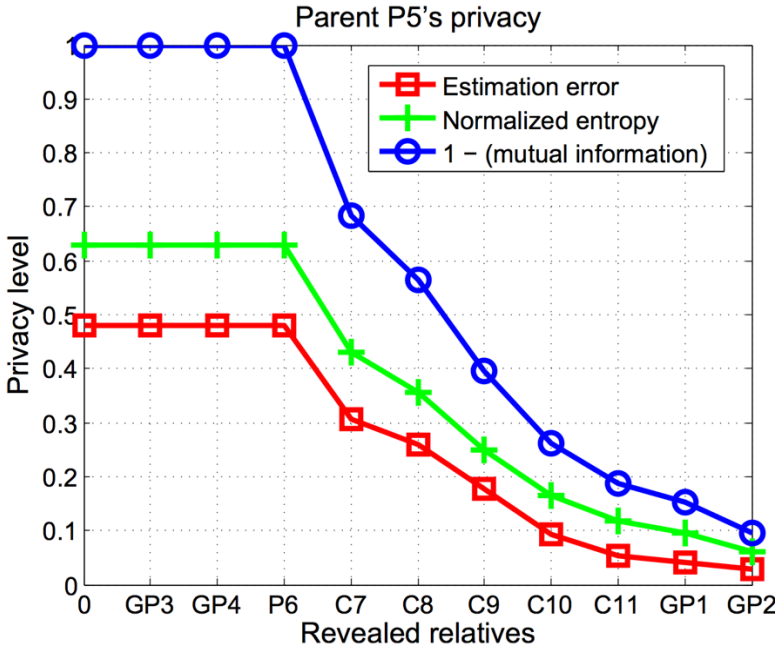


Good Privacy = High Estimation Error, High Uncertainty, and Low Mutual Info

Ground-truth Evaluation

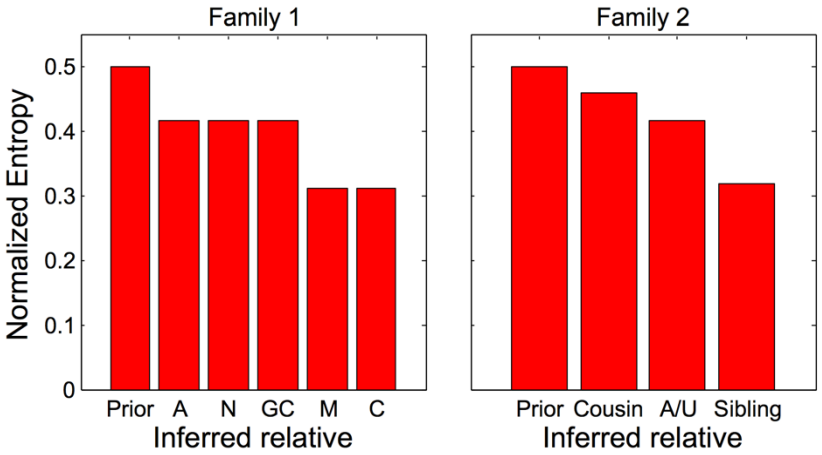
- 1 family (11 people), SNPs (80k), CEPH/Utah Pedigree 1463
- Example: Target P5, gradually reveal relative's SN info to attacker, from distant relatives to close relatives





Attacking People in the Wild

- Attacking two families, focusing on health privacy
  - Genome from OpenSNP, family tree from Facebook
  - Only 1 person in each family revealed genome to attacker
  - No ground-truth, the only usable metric is “entropy”



Privacy Protection Schemes

Privacy-enhancing Technology for Genomic Data

- Homomorphic encryption
- Secure multi-party computation (MPC)
  - Garbled circuits
- Secure two-party computation
  - Private Set Intersection (PSI)
- Differential privacy
  - Adding noise
- Trusted execution environments
  - SGX

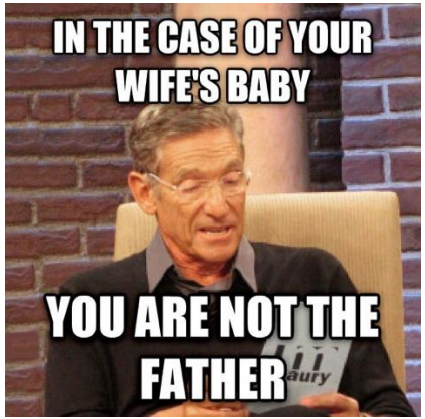
Privacy-Preserving Genetic Paternity Test (1 of 2)

Strawman Approach for Paternity Test

- On average, ~99.5% of any two human genomes are identical
- Parents and children have even more similar genomes
- Compare candidate's genome with that of the alleged child:
  - **Test positive if % of matching nucleotides is > 99.5 + τ**

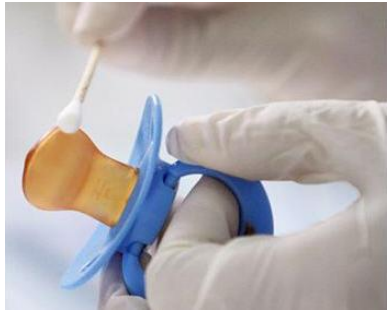
First-Attempt Privacy-Preserving Protocol

- Use an appropriate secure two-party protocol for the comparison
- PROs: High-accuracy and error resilience
- CONs: Performance not promising (3 billion symbols in input)
- Experiments showed computation takes a few days



Privacy-Preserving Genetic Paternity Test (2 of 2)

- Improved Protocol
  - ~99.5% of any two human genomes are identical
  - Why don't we compare *only* the remaining 0.5%?

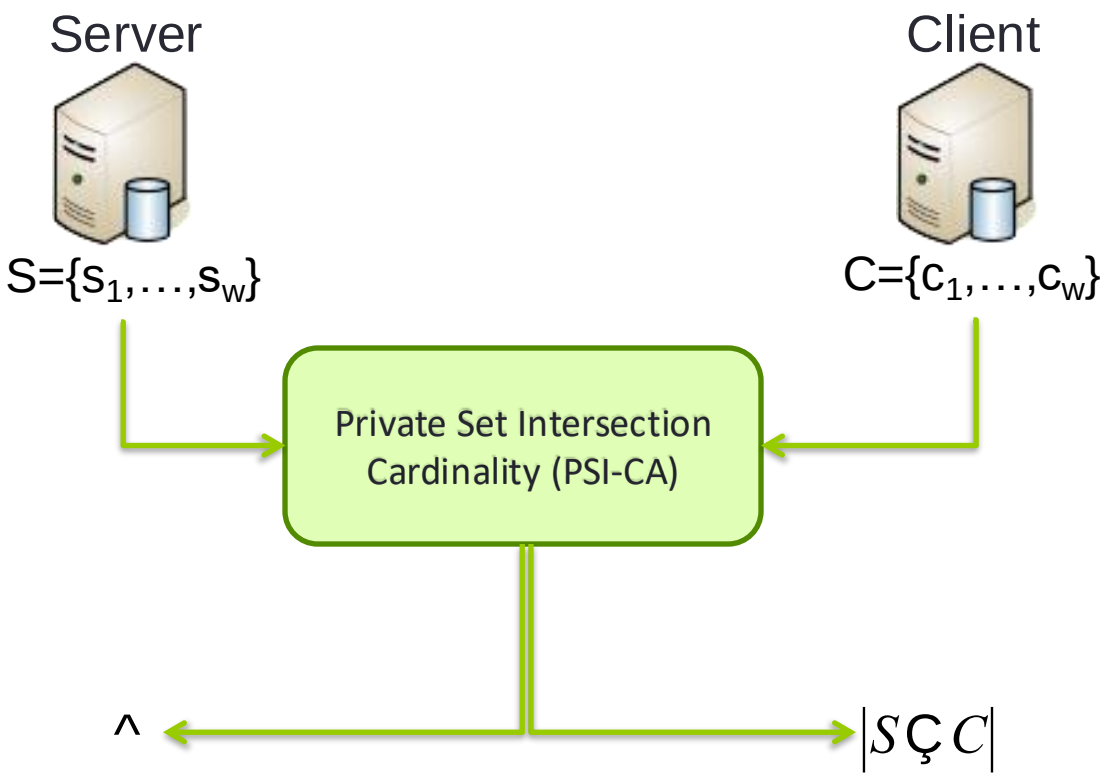


But… We don't know (yet) where *exactly* these 0.5% occur!

Using **Private Set Intersection Cardinality** for privacy-preserving comparison, it takes about 1 hour

Image from dna-testing-for-paternity.com

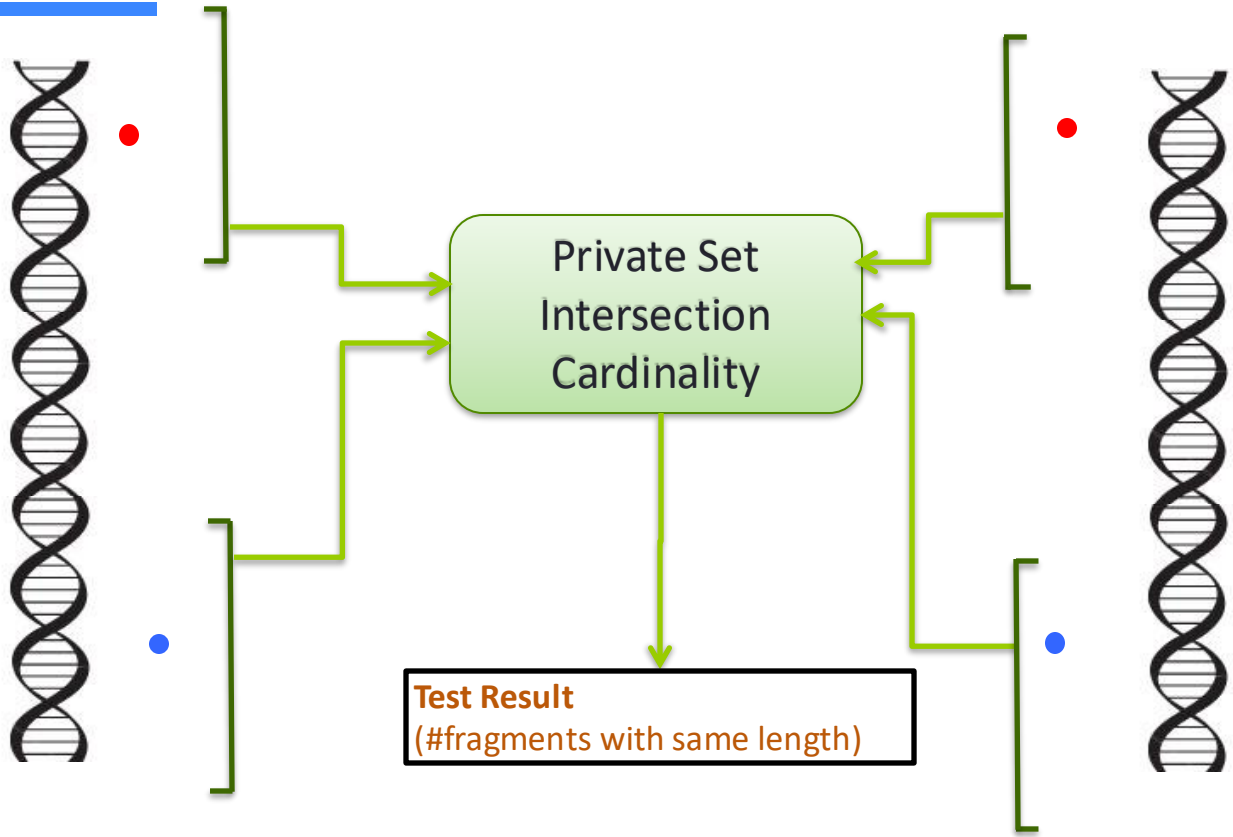
Private Set Intersection Cardinality (PSI-CA)



PPGT Strategy

- **In-vitro emulation – RFLP-based paternity test**
  - Restriction Fragment Length Polymorphism (RFLP) analysis: a difference between samples of homologous DNA molecules from differing locations of restriction enzyme sites
  - DNA sample is cut into fragments by enzymes
    - Fragments separated according to their lengths by *gel electrophoresis*
    - Paternity test is positive if enough fragments have the same length

Privacy-Preserving RFLP-based Paternity Test



PPGT Strategy

- **In-vitro emulation – RFLP-based paternity test**
  - Restriction Fragment Length Polymorphism (RFLP) analysis:
    a difference between samples of homologous DNA molecules from differing locations of restriction enzyme sites
  - DNA sample is cut into fragments by enzymes
    - Fragments separated according to their lengths by *gel electrophoresis*
    - Paternity test is positive if enough fragments have the same length

- **RFLP-based PPGPT – Reduction to PSI-CA**
  - *Participants*: "client" (receives the result), "server" (remains oblivious)
  - *Public input*: t, enzymes $E = \{e_1, \ldots, e_j\}$ markers $M = \{mk_1, \ldots, mk_l\}$
  - *Private input*: digitized genomes

Remarks

- Why compare fragment lengths?
  - Isn't it more accurate to compare actual contents?
  - In practice, RFLP yields "false positives" with very low probability
  - This approach increases resilience to sequencing errors
- Performance Evaluation
  - About **1 min pre-processing** to emulate enzyme digestion process
  - About **10 ms computation** time on Intel Core i5 with 25 fragments
  - Extending to 50 fragments doubles computation time and increases accuracy by orders of magnitudes
  - Communication overhead: only a few **KBs**

Discussion

- What security and privacy issues are raised by DTC (direct-to-consumer) genomics?
- How would you like to see *your* DNA data managed? What about the DNA of your relatives?
- Should it be legal to obtain your DNA without your consent?

References

- *Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. 2013. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security (CCS '13). Association for Computing Machinery, New York, NY, USA, 1141–1152. DOI:https://doi.org/10.1145/2508859.2516707*