P&S Accelerating Genomics

Lecture 11: Genomic Data Sharing Under Differential Privacy

Dr. Nour Almadhoun Alserr

ETH Zurich

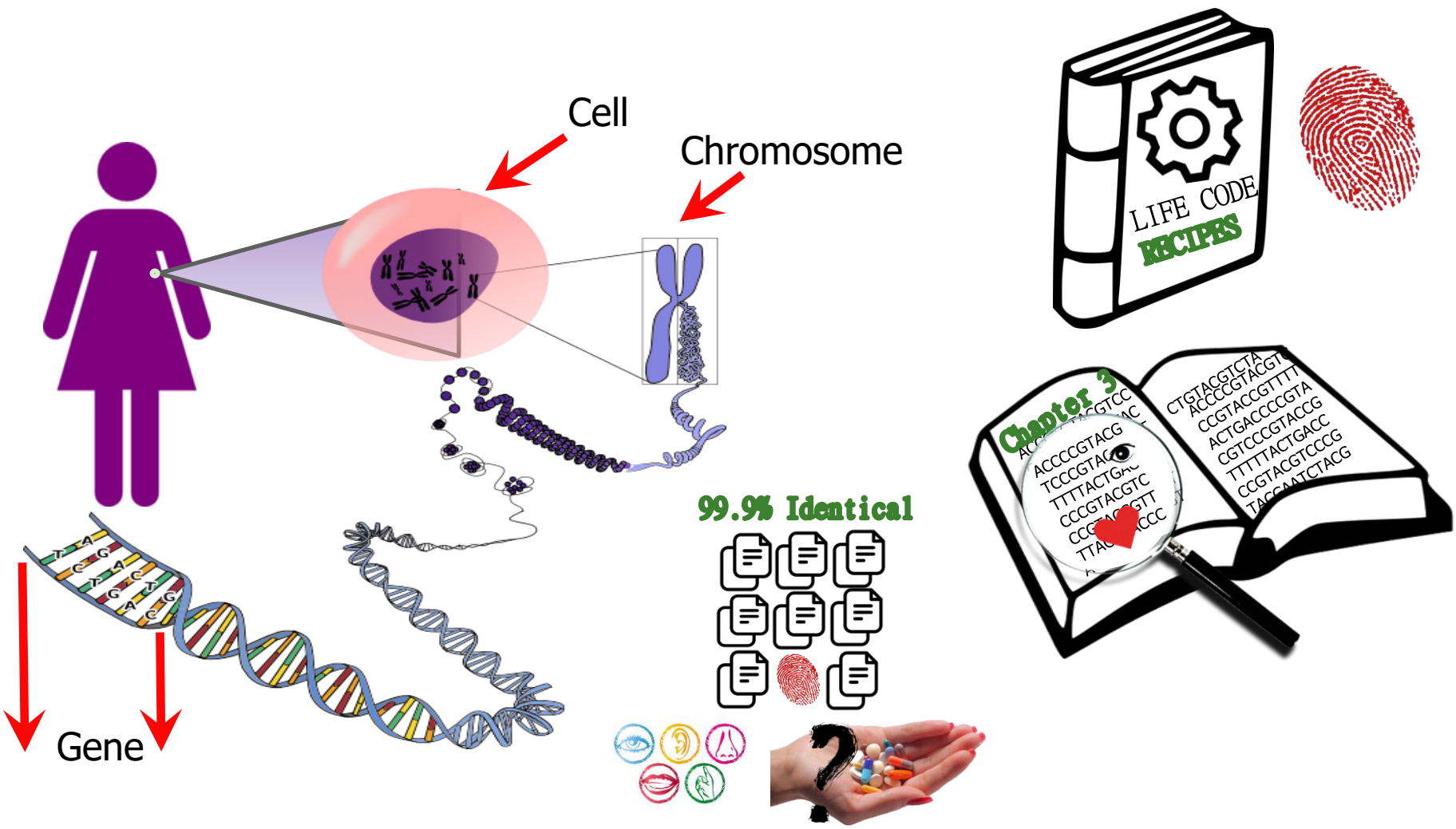
Fall 2022

12 January 2023

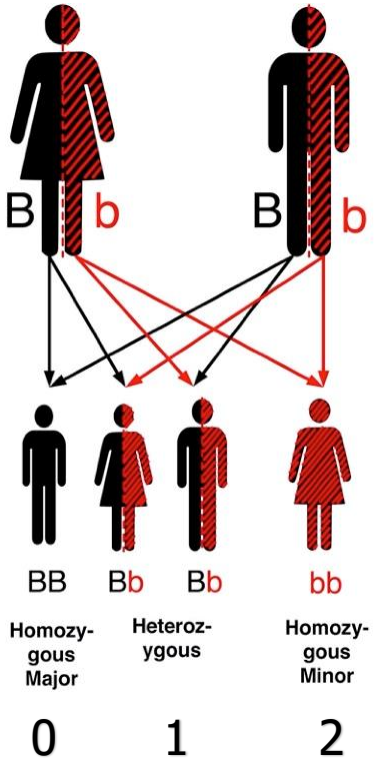
SAFARI

ETH zürich

Genome


Cell
Chromosome
Gene
99.9% Identical
LIFE CODE
RECIPES
Chapter 3

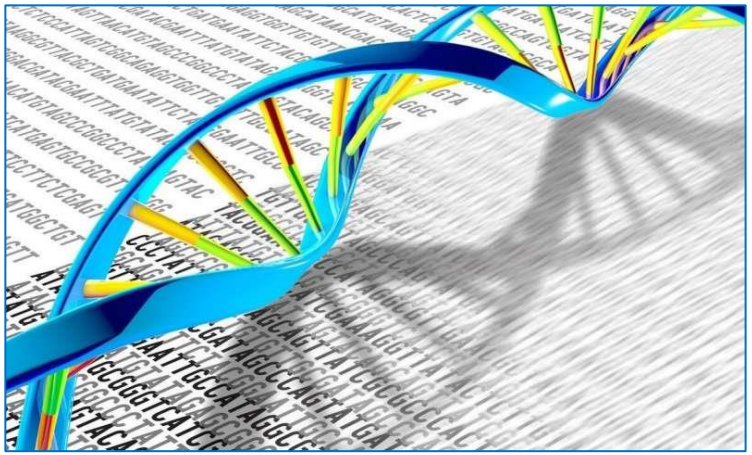
Mendel's Law



Mother \ Father	BB	Bb	bb
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)

Mother \ Child	BB	Bb	bb
BB	(0.5,0.5,0)	(0,0.5,0.5)	N/A
Bb	(0.5,0.5,0)	(0.33,0.33,0.33)	(0,0.5,0.5)
bb	N/A	(0.5,0.5,0)	(0,0.5,0.5)

The Genomic Era



1 Zetta-Bases/year (10^{21}) capacity

105 Million Sequenced Human genome

Stephens, Zachary D., et al. PLoS biology (2015)

The Genomic Era

Home › Genomics › 100,000 Genomes Project

100,000 Genomes Project

The 100,000 Genomes Project is cementing the NHS's position as one of the most advanced healthcare systems in the world, and is providing the foundation for a new era of personalised medicine, and this in turn will contribute towards delivering high quality care for all, now and for future generations.

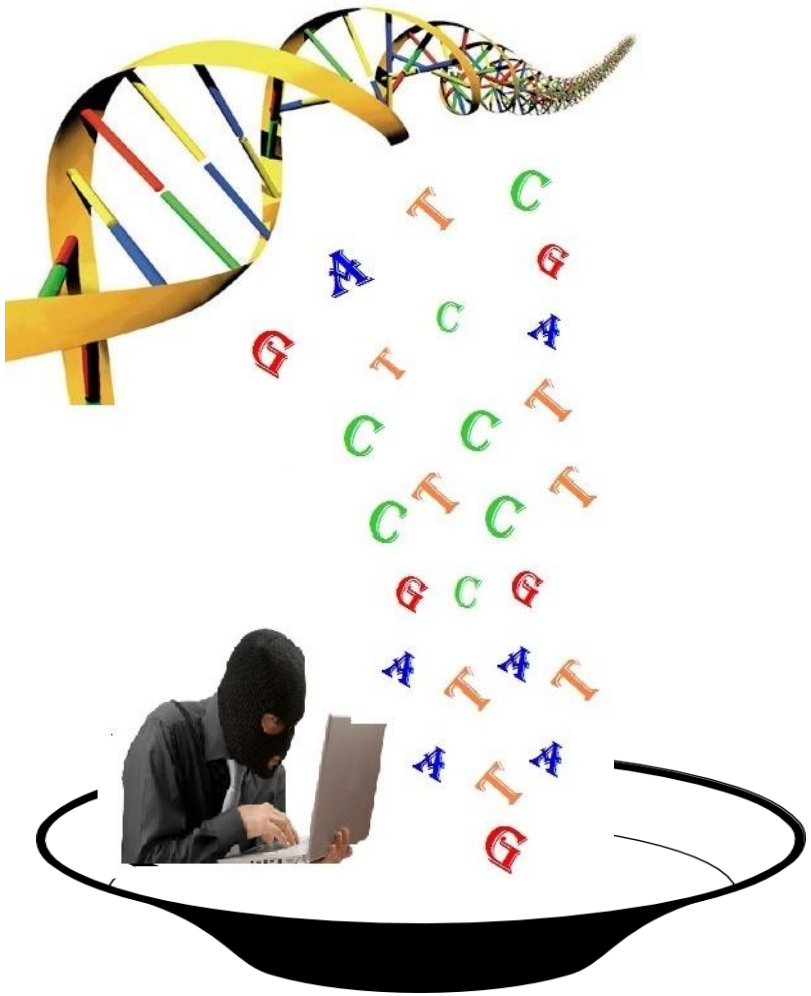
The 100,000 Genomes Project aims to bring the benefits of personalised medicine to the NHS. To make sure patients benefit from innovations in genomics, the Government has committed to sequencing 100,000 whole human genomes, from 70,000 patients, by the end of 2018.

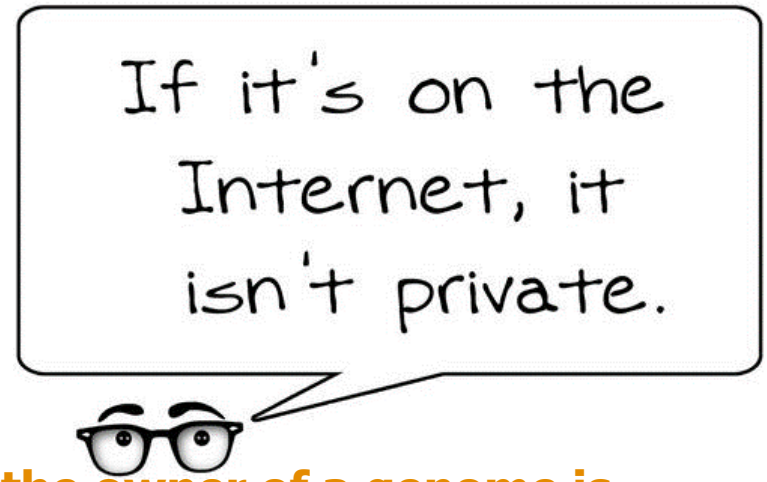
European '1+ Million Genomes' Initiative

The Signatories of the declaration of cooperation "Towards access to at least 1 million sequenced genomes in the EU by 2022" are setting up a collaboration mechanism with the potential to improve disease prevention, allow for more personalised treatments and provide a sufficient scale for new clinically impactful research.

69–92% of the respondents in these studies had positive attitudes towards genomics research and donating their DNA samples.

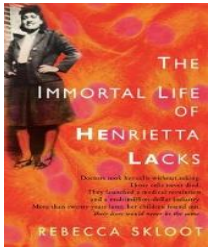
Privacy Risks





If the owner of a genome is identified:

- He/she will face the risk of discrimination by employers or insurance companies.
- DNA sequences are highly correlated to the relatives' sequences, so relative's privacy will be at risk (Henrietta Lacks).



Genome-Wide Association Study (GWAS)

Detecting genetic variants associated with phenotypes using two groups of people.

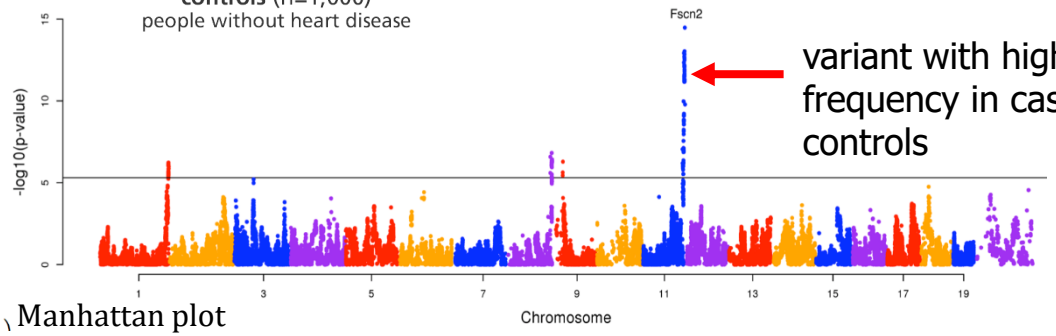
cases (n=1,000)
people with heart disease

controls (n=1,000)
people without heart disease

Table 1. GWAS genotype distribution for a 2 × 3 contingency table (left) and a 2 × 2 contingency table (right).

	Genotype			
	0	1	2	Total
Case	S_0	S_1	S_2	S
Control	C_0	C_1	C_2	C
Total	n_0	n_1	n_2	n

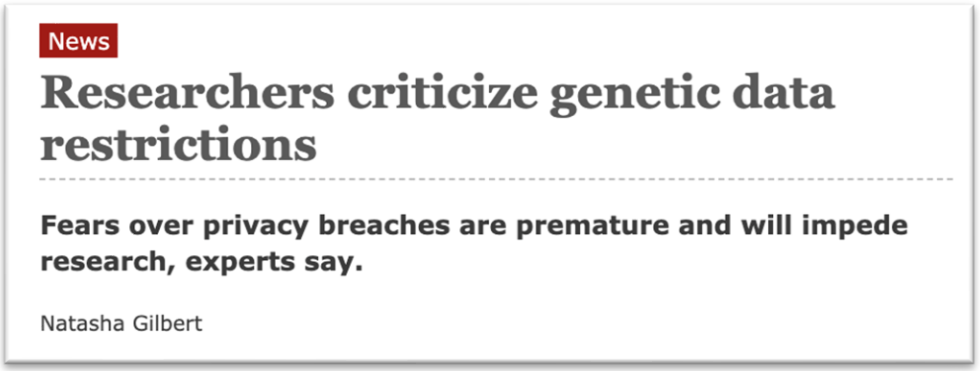
	Genotype		
	0	1	Total
Case	S_0	$S_1 + S_2$	S
Control	C_0	C_1+C_2	C
Total	n_0	n_1+n_2	n



variant with higher frequency in cases than controls

Manhattan plot

Genetic Data Restriction



News

Researchers criticize genetic data restrictions

Fears over privacy breaches are premature and will impede research, experts say.

Natasha Gilbert

- Researchers have assumed that case-control studies are safe to publish aggregate statistics of SNPs. Such belief was challenged when **Homer Attack** happened.

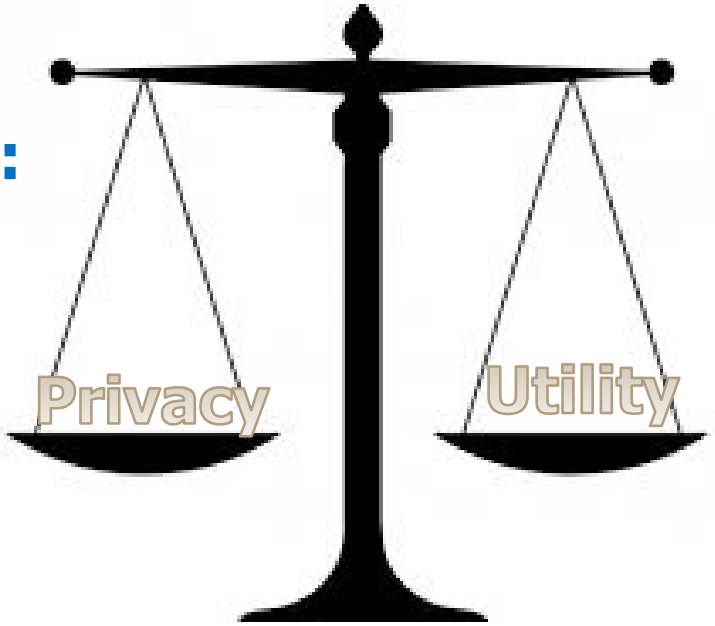
- NIH restricts the access to key results and data of GWAS to only trusted individuals.

Privacy-Utility Tradeoff

- Hiding some important data needs to tradeoff between **privacy** and **utility**.

➢ **Privacy preserving techniques:**

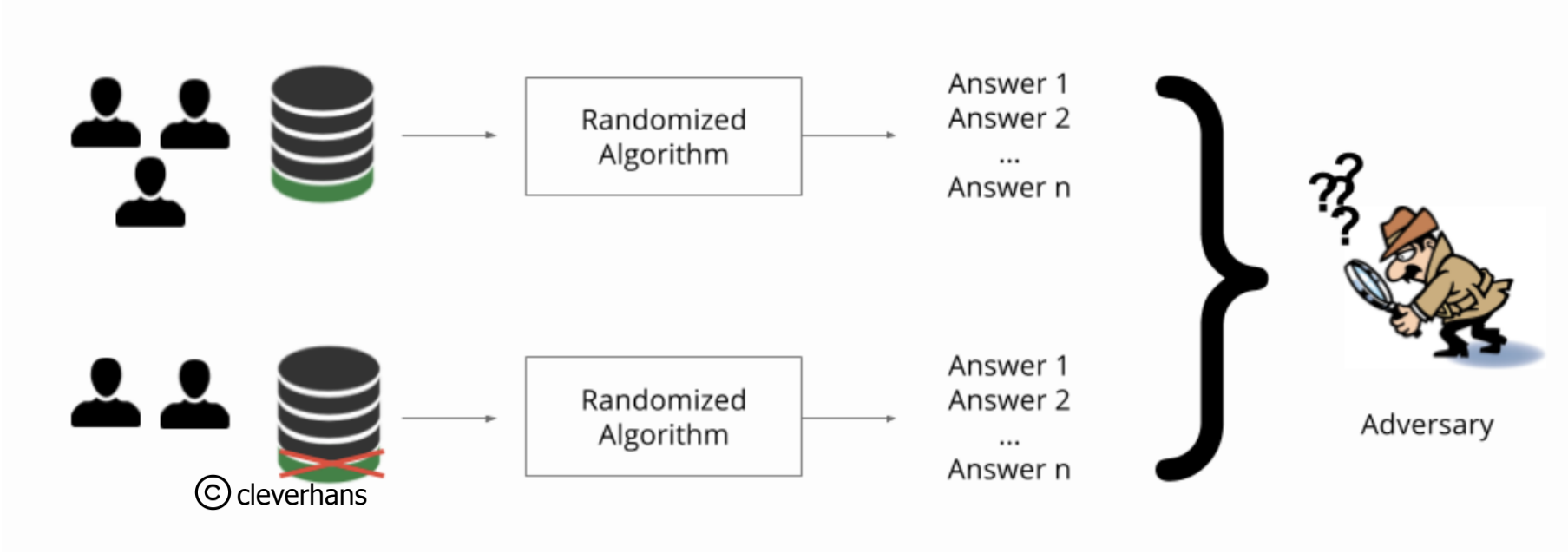
- K-anonymity.
- l-diversity.
- t-closeness.
- **Differential privacy.**
- Crypto-based techniques.



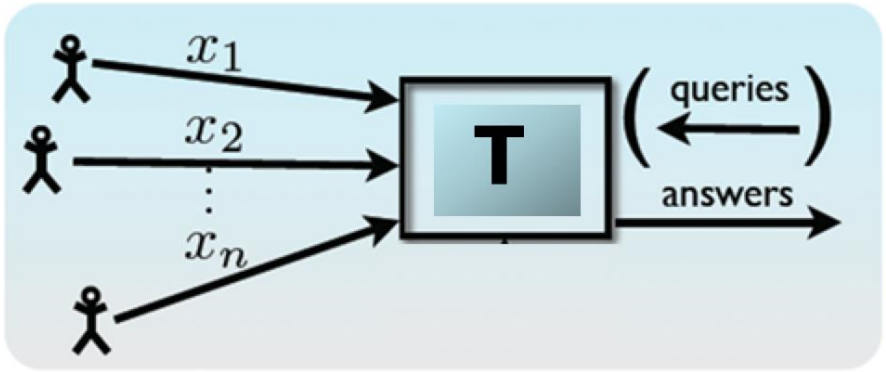
Privacy-Preserving

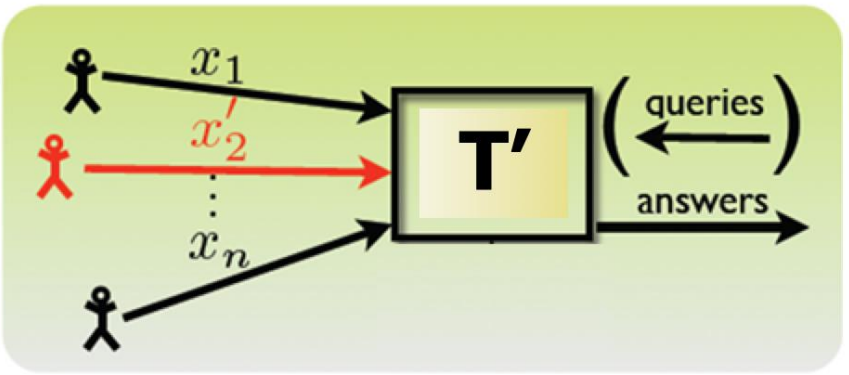
Computing ⟷ **Sharing**

Differential Privacy

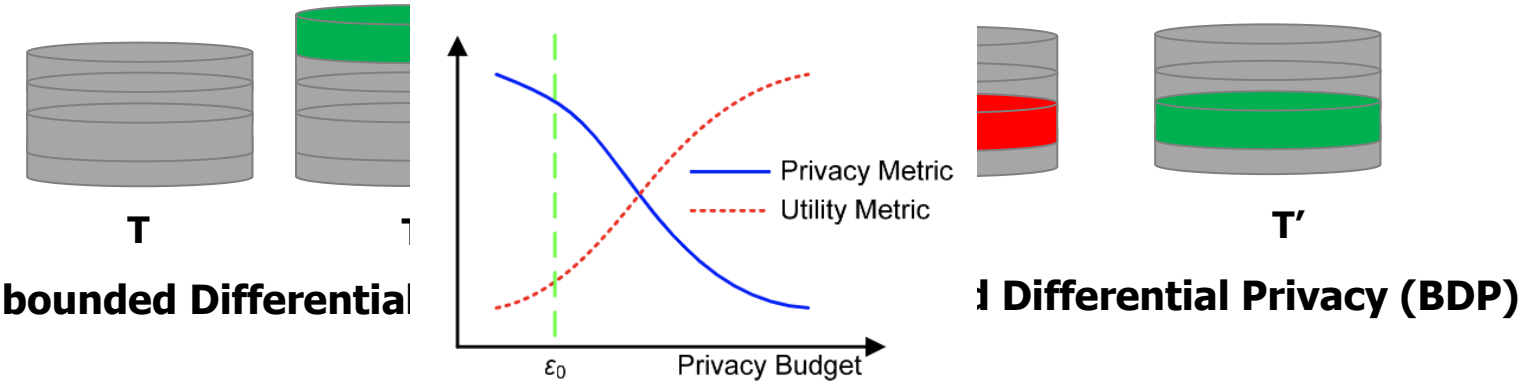


Differential Privacy





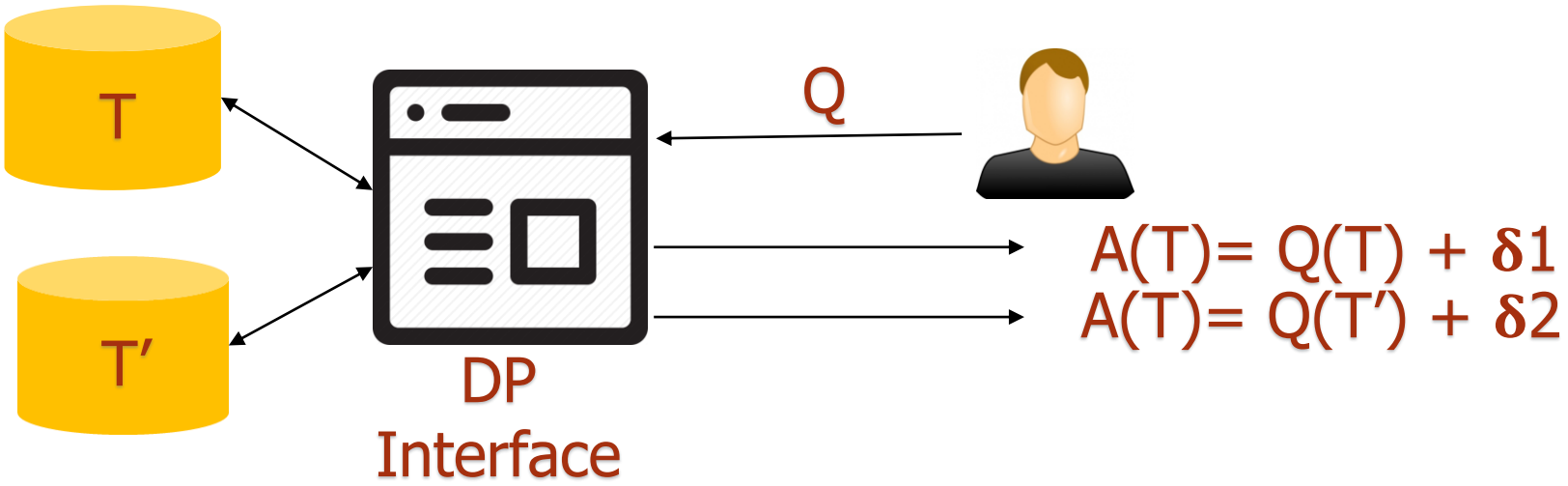
$$\Pr[A(T) \in O] \leq e^{\varepsilon} \Pr[A(T') \in O]$$



© Liu, et al. *Applied Sciences* (2018)

Laplace Perturbation Mechanism (LPM)

- **Q(T) + δ** where **δ** is drawn from a Laplace distribution with mean 0 and scale ΔQ/Ɛ
- ΔQ : query global sensitivity



Differential Privacy

Google
Differential privacy library

Google Developer (2019)

Privacy

Differential Privacy Team, Apple (2017)

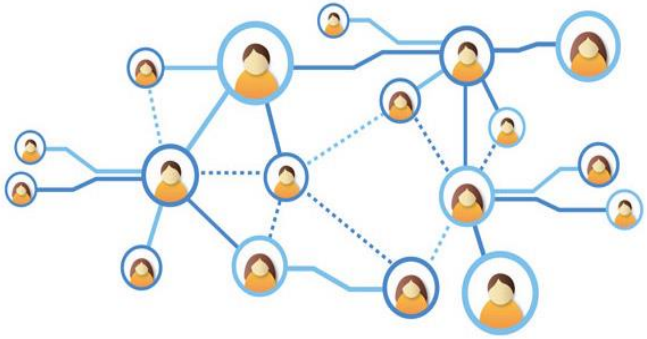
Microsoft Research

Collecting Telemetry Data Privately (2017)

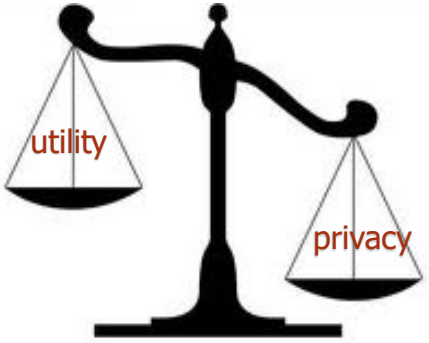
Differentially Private Publication System (2018)

Research Problem

- DP standard mechanism does not consider the dependency between the data tuples in the dataset.

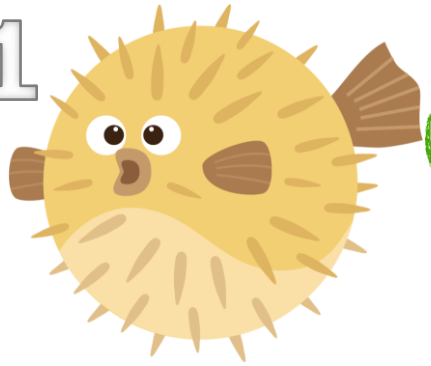


- Current DP-based mechanisms which consider the tuples correlation, provide poor accuracy.



Related Works

1



A generalization of DP

No perturbation algorithm

Pufferfish Framework
(Kifer and Machanavajjhala, 2012)

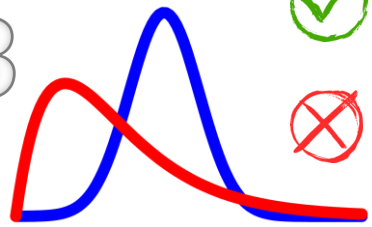
2

Perturbation mechanisms

Deterministic constraints for the adversary

Blowfish Framework
(He et al., 2014)

3



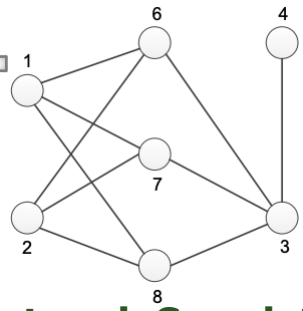
Perturbation mechanisms

Correlation modeled by Gaussian Markov Random Fields

Bayesian DP
(Yang et al., 2015)

Related Works

4



Perturbation mechanisms

Excessive added noise

Network Correlation
(Chen et al., 2014)

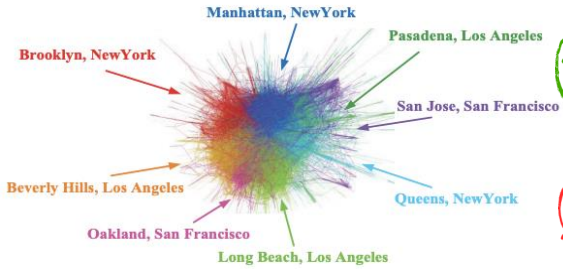
5

Perturbation mechanisms

Correlation modeled by Gaussian Markov Random Fields

Temporal Correlation
(Cao et al., 2017)

6



Perturbation mechanisms

pairwise correlations

Dependent DP
(Liu et al., 2016)

7

Perturbation mechanisms

Less Utility

Dependent DP
(Zhao et al., 2017)

Our Contributions

Attribute Inference Attack

Membership Inference Attack

1 Differentially private **SUM** query results in a static genomic dataset with dependent tuples.
[Bioinformatics'19]

2 Differentially private **MAF** and χ^2 query results in a static genomic dataset with dependent tuples.
[Bioinformatics'20] [ISMB'20]

3 Differentially private **MAF** in a static genomic dataset.
[Bioinformatics'20]
[ISMB'20]

Our Contributions

Countermeasures

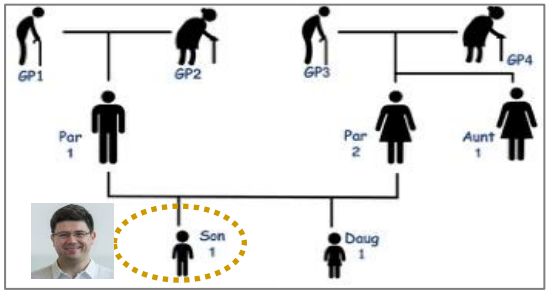
4 Ɛ-differential privacy for sharing genomic datasets with dependent tuples .

[Bioinformatics'19]

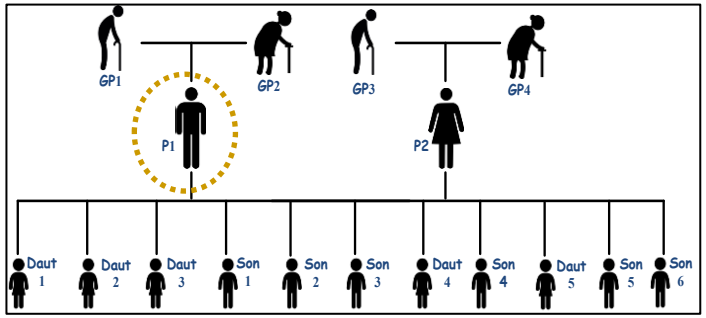
5 Selective hiding mechanism and differential privacy.

[arXiv'21]

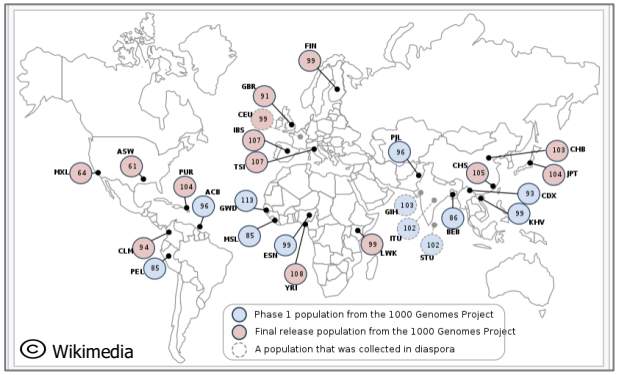
Dataset Description

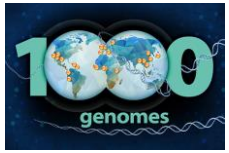


Manuel Corpas Family



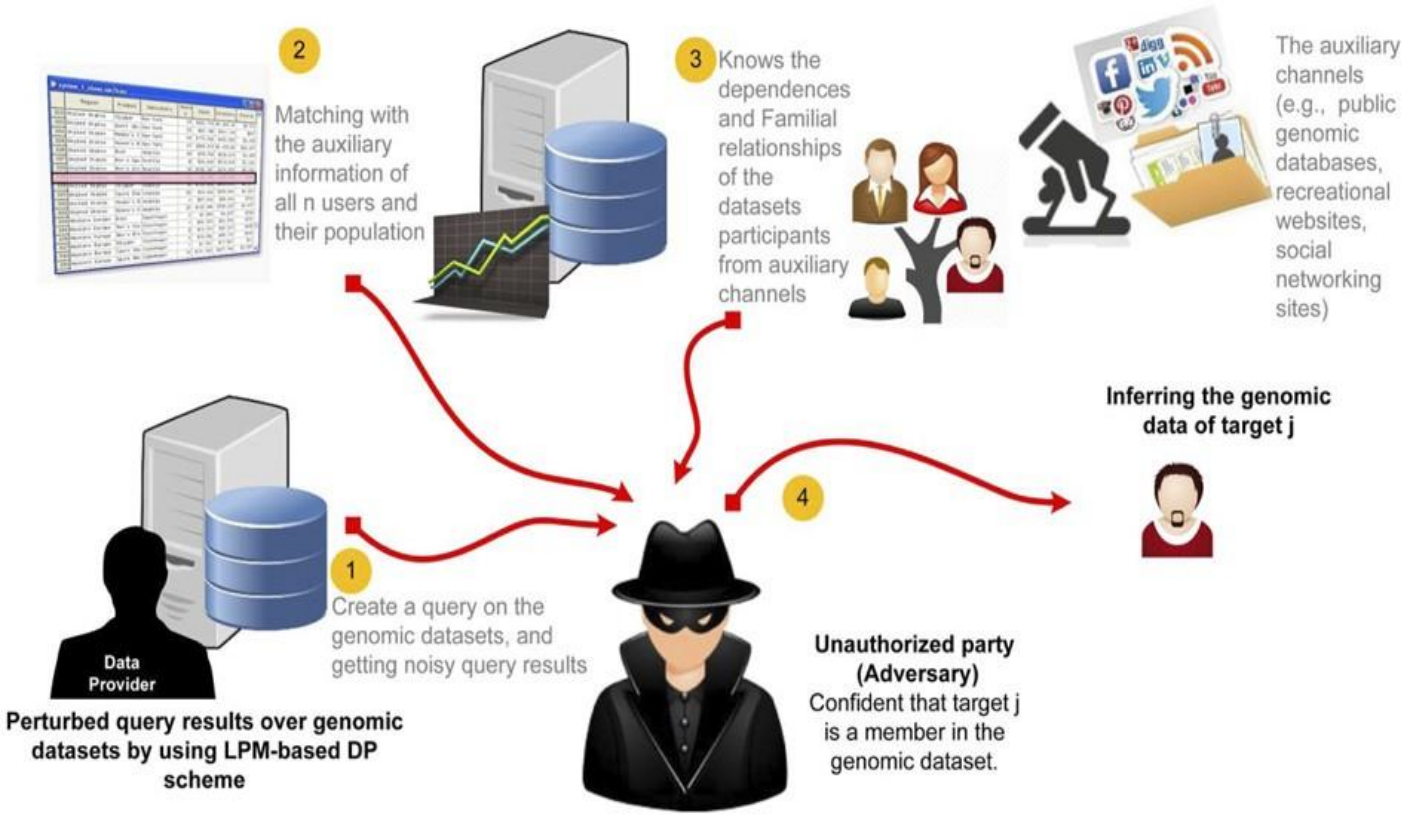
CEPH/Utah Family



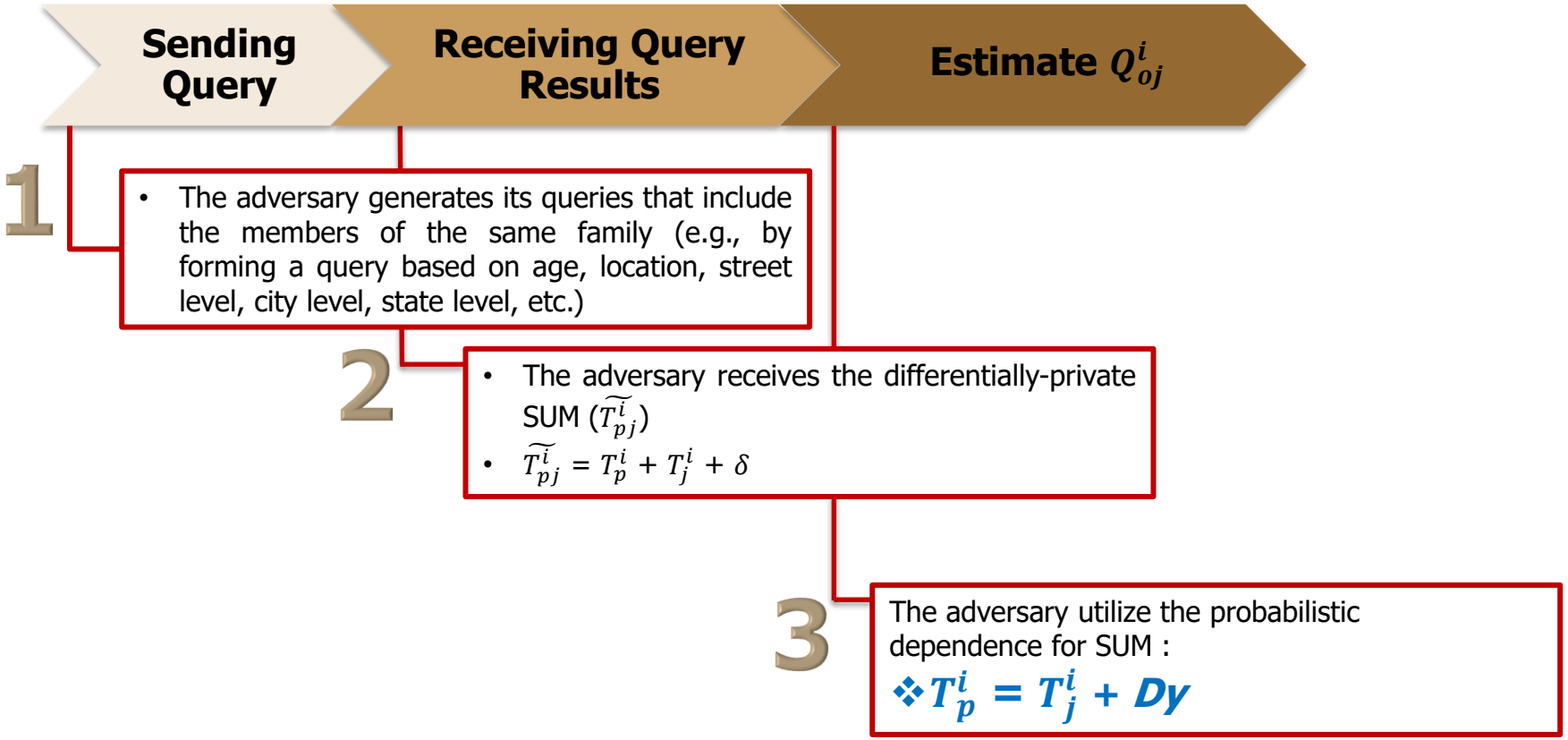


n= 2514 or n= 2508, m =100

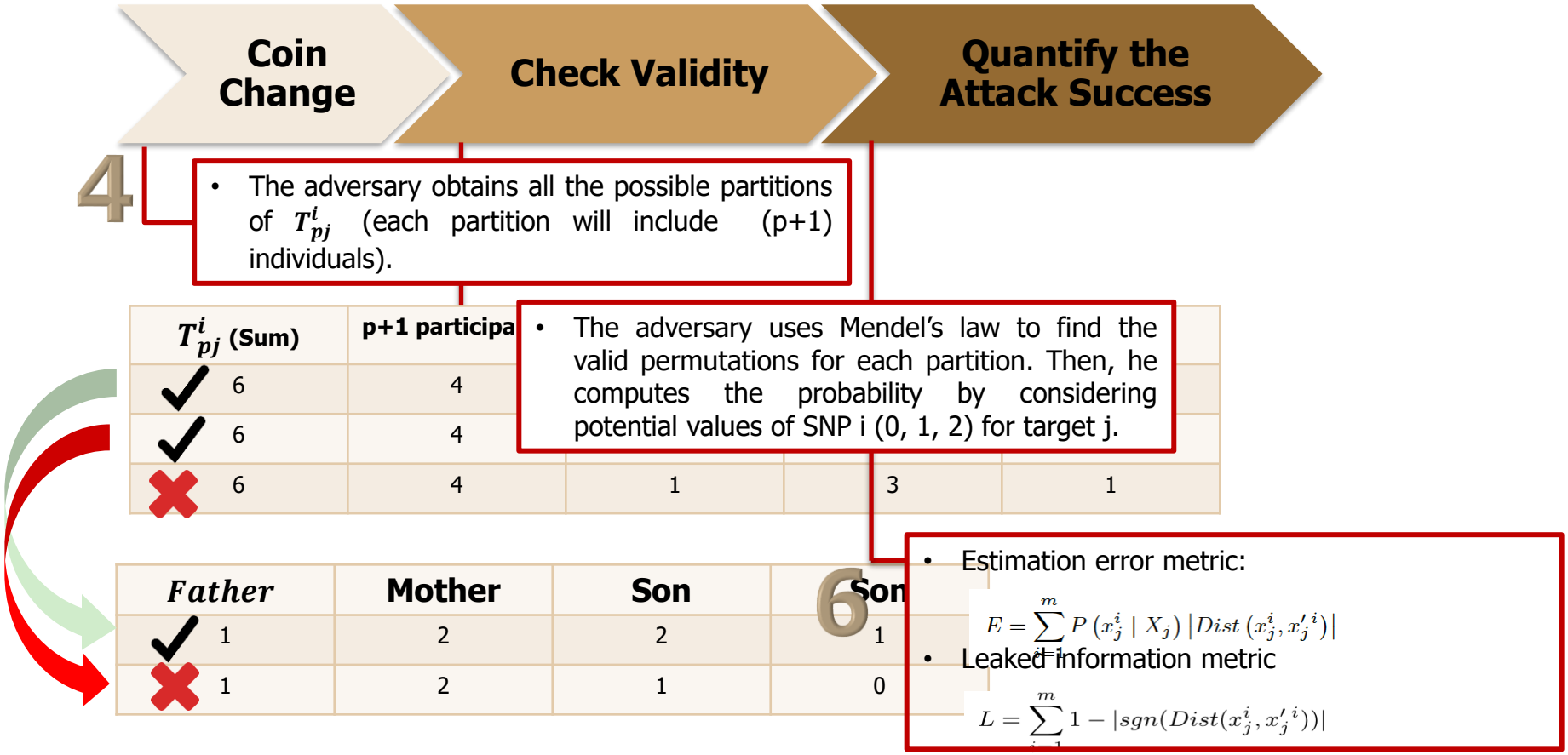
DP Inference Attacks



Attribute Inference Attack



Attribute Inference Attack



Key Results

The adversary can infer the actual value of the targeted SNPs by up to **50%**.

Our proposed mechanism can achieve up to **50%** better privacy guarantees than the traditional DP-based solutions.

DP Inference Attacks

Nour Almadhoun, Erman Ayday, and Ozgur Ulusoy
"Differential privacy under dependent tuples—the case of genomic privacy"
Bioinformatics, 2020
[Source code]

Bioinformatics

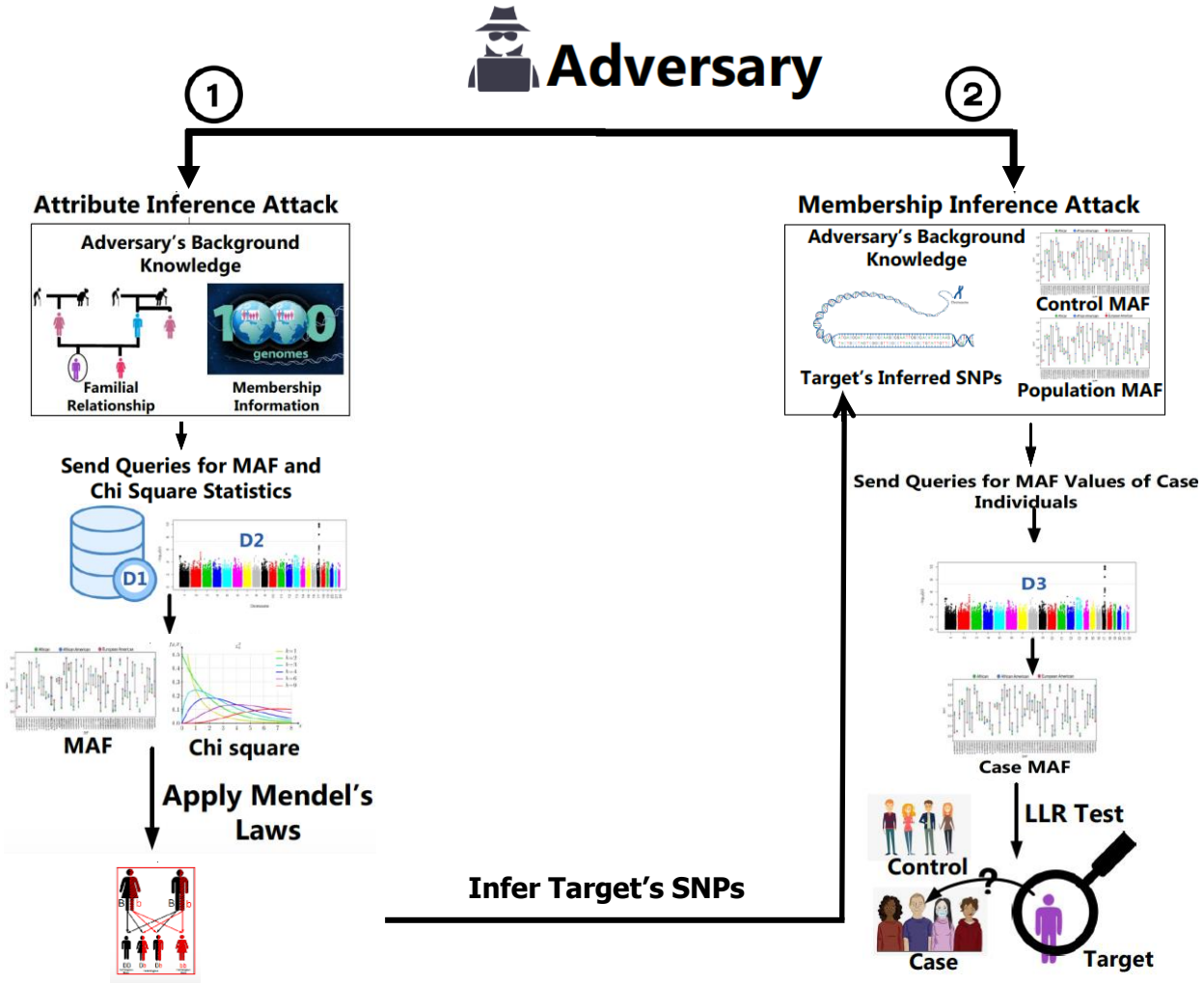
ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Differential privacy under dependent tuples—the case of genomic privacy FREE

Nour Almadhoun, Erman Ayday ✉, Özgür Ulusoy ✉

Bioinformatics, Volume 36, Issue 6, 15 March 2020, Pages 1696–1703,
https://doi.org/10.1093/bioinformatics/btz837
Published: 08 November 2019 **Article history** ▾

Threat Model



Membership Inference Attack

1

The adversary knows:
- the set of MAF values of SNPs for individuals in the control group (MC)
- the set of MAF values of SNPs for a similar population including both the case and control individuals (MP).

2

- The adversary receives the differentially-private MAF value of a SNP i for individuals in the case group $\widetilde{M_S^i} = M_S^i + \delta$.

3

- Null hypothesis: target j is not a part of the case.
- Alternative hypothesis: target j is part of the case group S.

$$LLR = \sum_{i=1}^{m} x_j^i log\frac{M_S}{M_C} + (1 - x_j^i) log\frac{1 - M_S}{1 - M_C}$$

Key Results

An adversary can reveal up to **40% ~ 50%** more sensitive information about the genome of a target (compared to original privacy guarantees of standard DP-based mechanisms).

The inference power of the adversary can be **significantly high** in the membership attack even using inferred (and hence partially incorrect) genomes.

DP Inference Attacks

Nour Almadhoun, Erman Ayday, and Ozgur Ulusoy
"Inference attacks against differentially private query results from genomic datasets including dependent tuples"
Bioinformatics, 2020
[Source code]

Bioinformatics

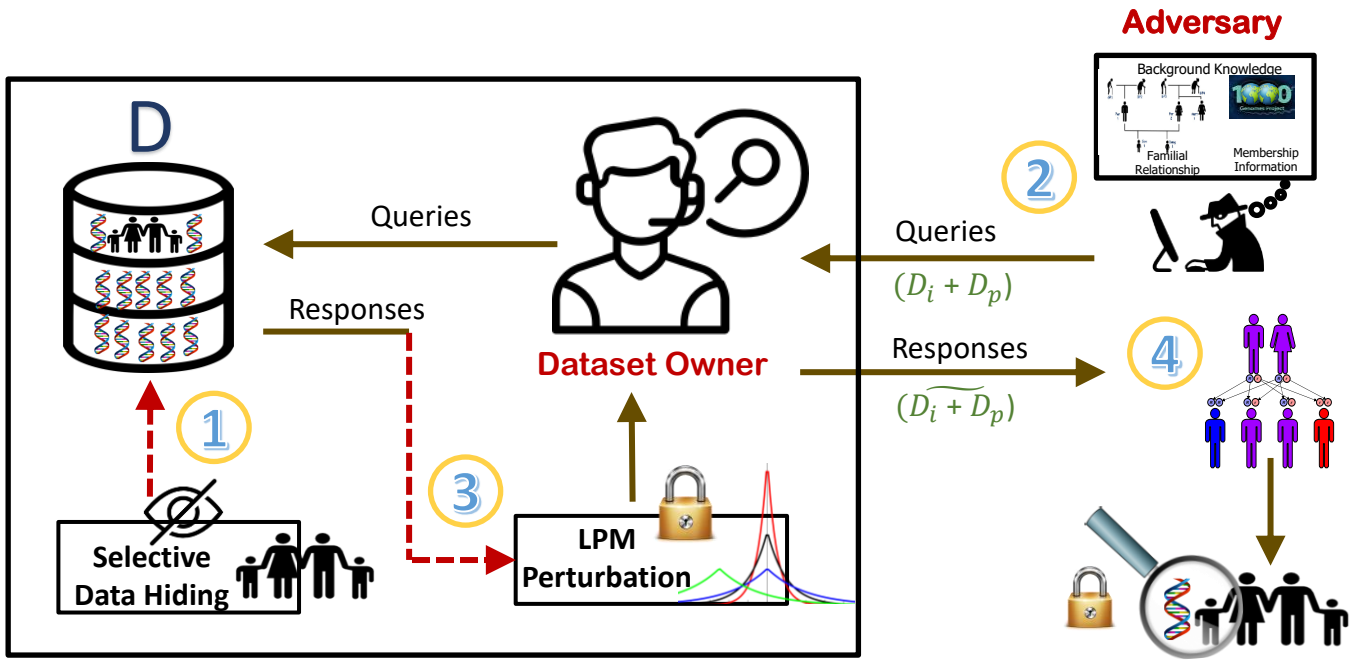
ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Inference attacks against differentially private query results from genomic datasets including dependent tuples

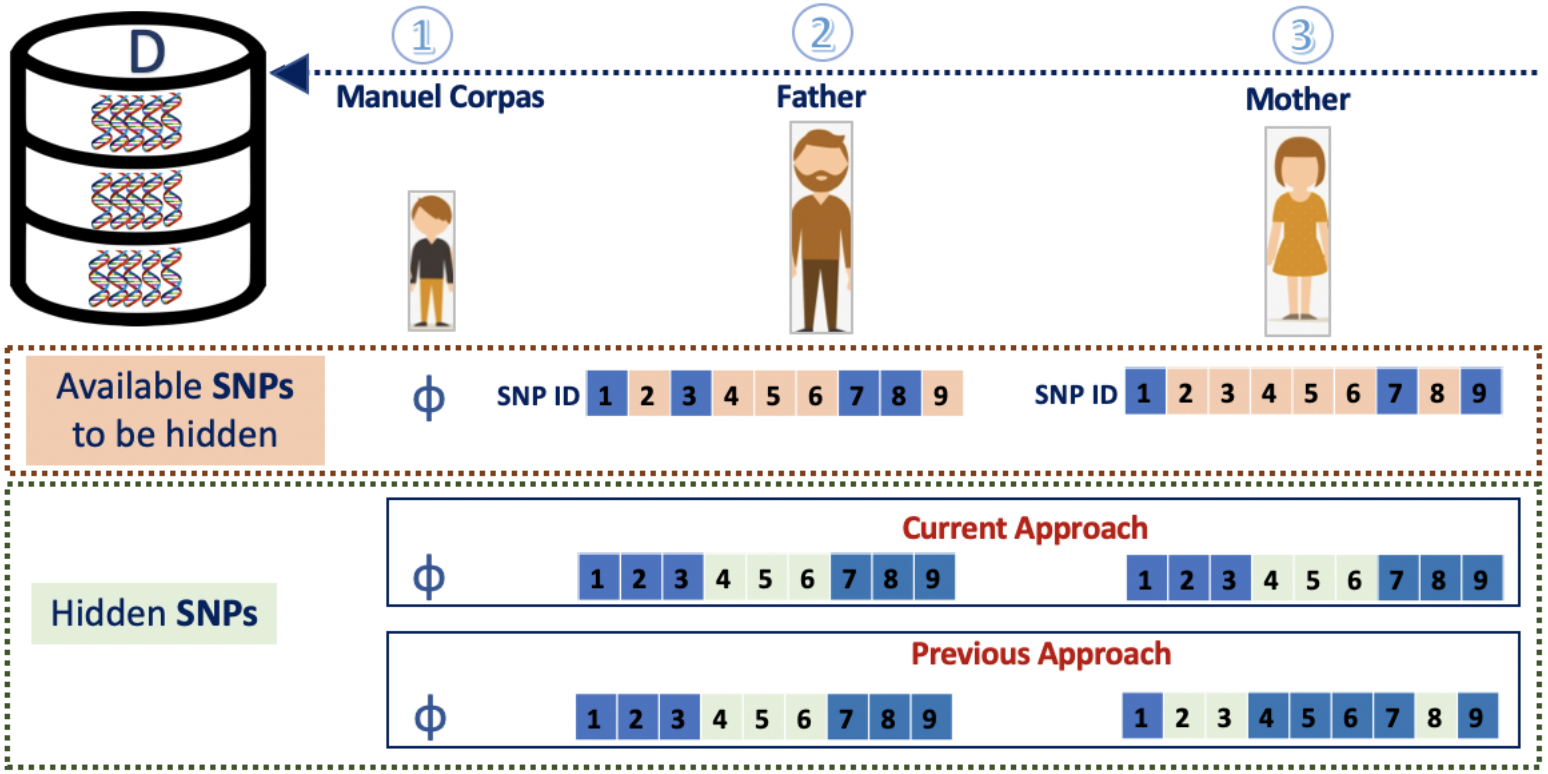
Nour Almadhoun, Erman Ayday ✉, Özgür Ulusoy ✉

Bioinformatics, Volume 36, Issue Supplement_1, July 2020, Pages i136–i145,
https://doi.org/10.1093/bioinformatics/btaa475
Published: 13 July 2020

Selective Hiding Model



Selective Hiding Model



Key Results

We provide **similar privacy guarantees** of Ɛ-differential privacy, with **higher utility** than the state-of-the-art schemes.

Selective SNP Hiding

Nour Almadhoun Alserr, Gulce Kale, Onur Mutlu, Oznur Tastan, Erman Ayday
"Near-Optimal Privacy-Utility Tradeoff in Genomic Studies Using Selective SNP Hiding"
arXiv, 2021
[Source code]

GenShare Model

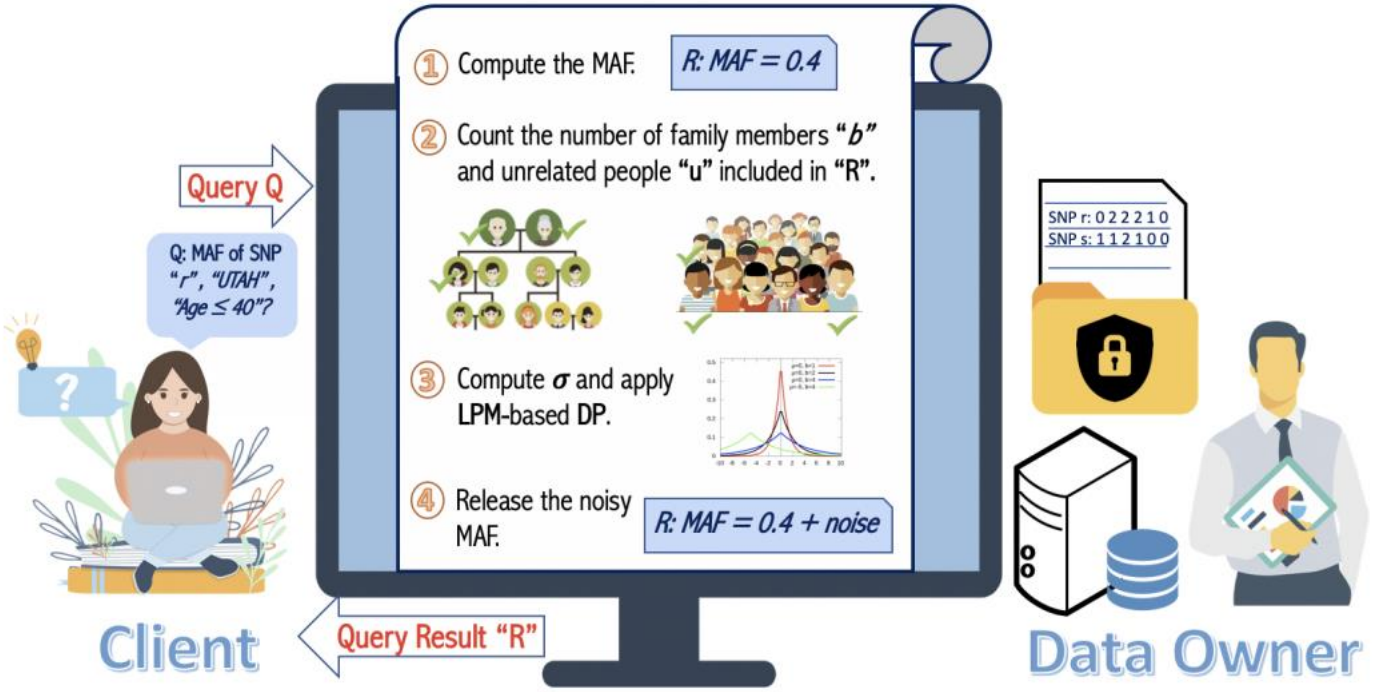
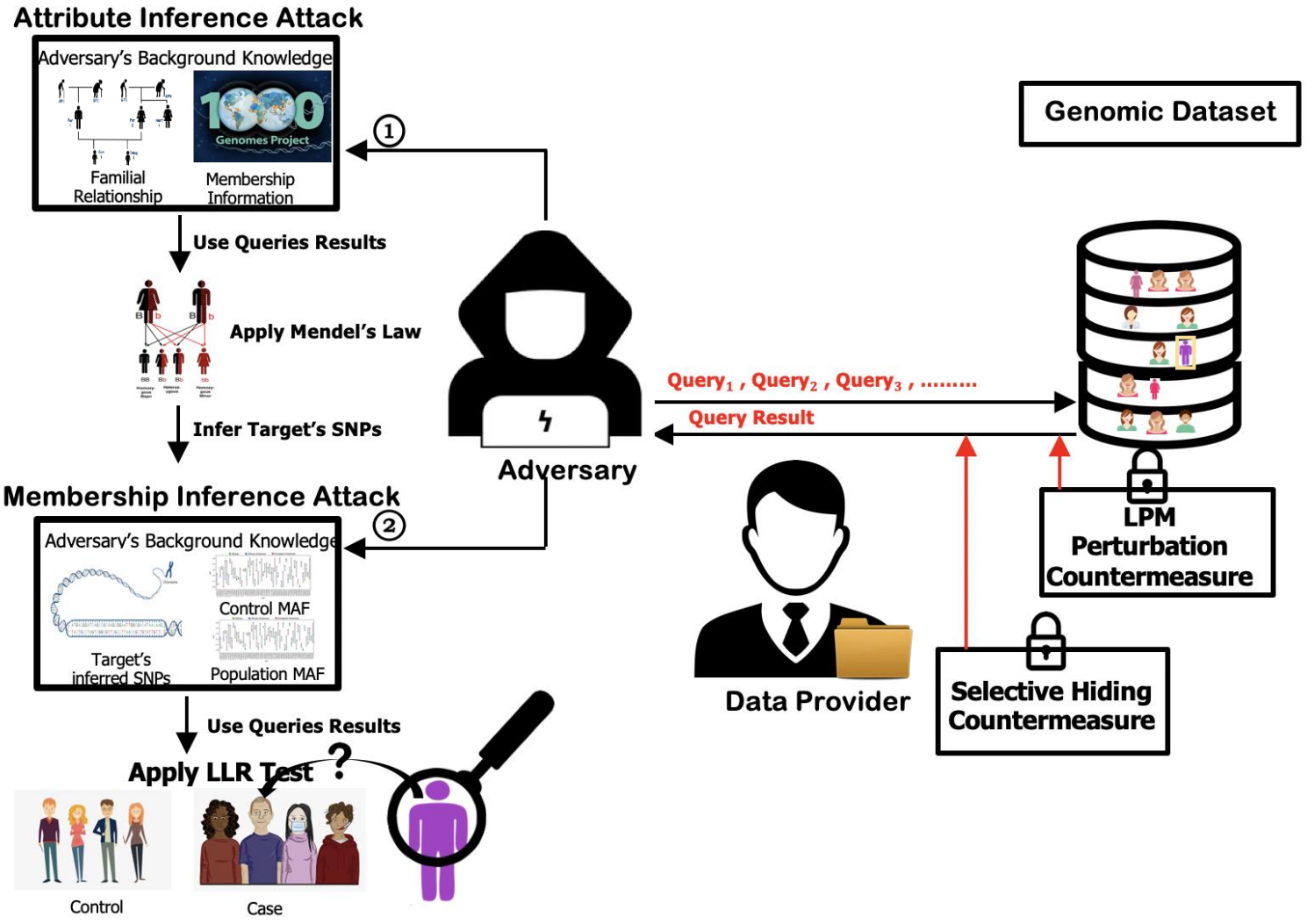


Fig. 1: Our proposed GenShare model

GenShare

Nour Almadhoun Alserr, Ozgur Ulusoy, Erman Ayday, Onur Mutlu
"**GenShare: Sharing Accurate Differentially-Private Statistics for Genomic Datasets with Dependent Tuples**"
arXiv, 2021

Full Model


Attribute Inference Attack
Adversary's Background Knowledge
Familial Relationship
Membership Information
1000 Genomes Project
①
Use Queries Results
Apply Mendel's Law
Infer Target's SNPs
Membership Inference Attack
Adversary's Background Knowledge
②
Target's inferred SNPs
Control MAF
Population MAF
Use Queries Results
Apply LLR Test
Control
Case
Adversary
Genomic Dataset
Query1 , Query2 , Query3 , ...
Query Result
Data Provider
LPM Perturbation Countermeasure
Selective Hiding Countermeasure

P&S Accelerating Genomics

Lecture 11: Genomic Data Sharing Under Differential Privacy

Dr. Nour Almadhoun Alserr

ETH Zurich

Fall 2022

12 January 2023

SAFARI

ETH zürich