P&S Mobile Genomics

Lecture 10: Genomic Data Sharing Under Differential Privacy

Dr. Nour Almadhoun Alserr

ETH Zurich

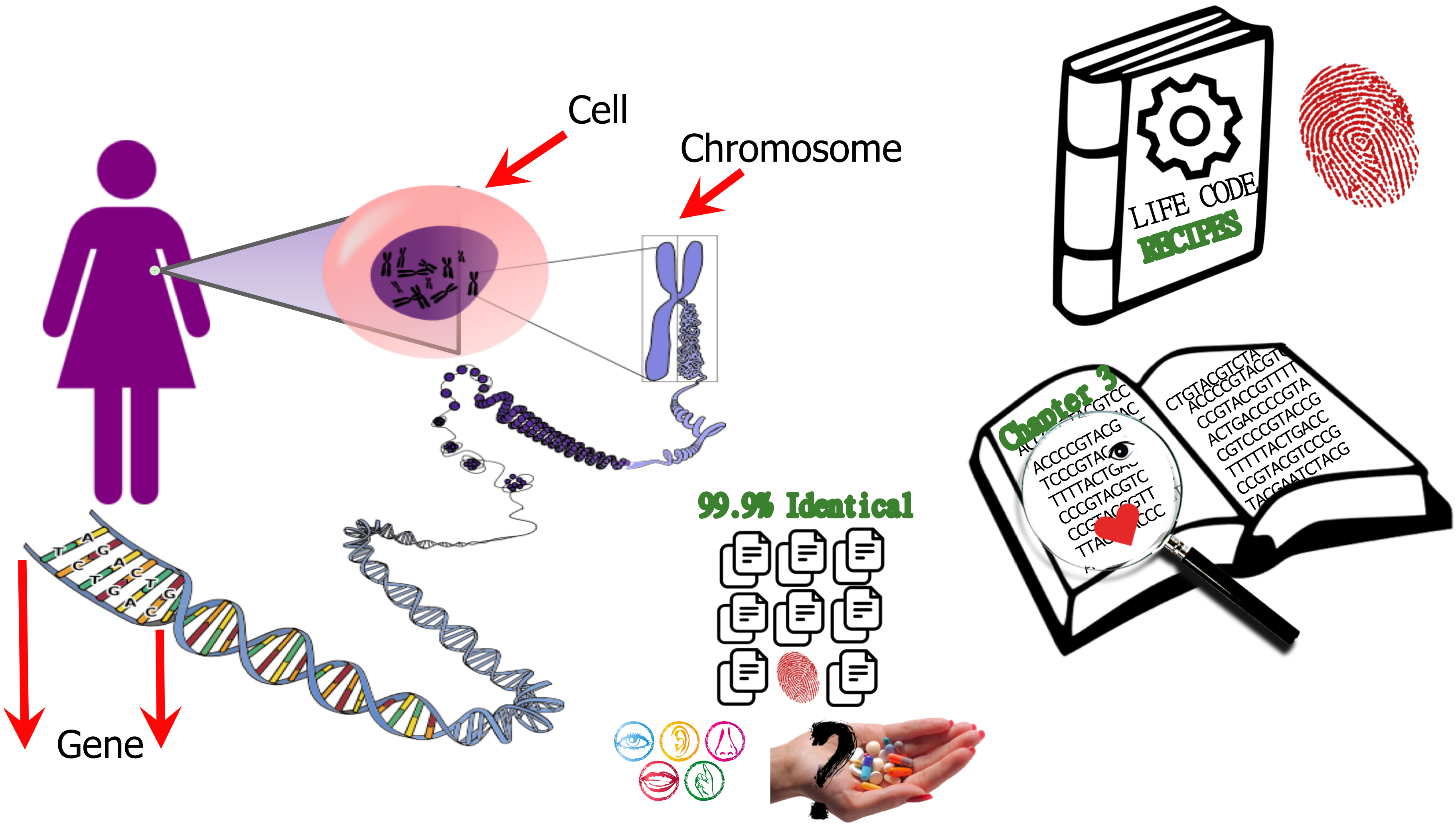
Spring 2022

17 May 2022

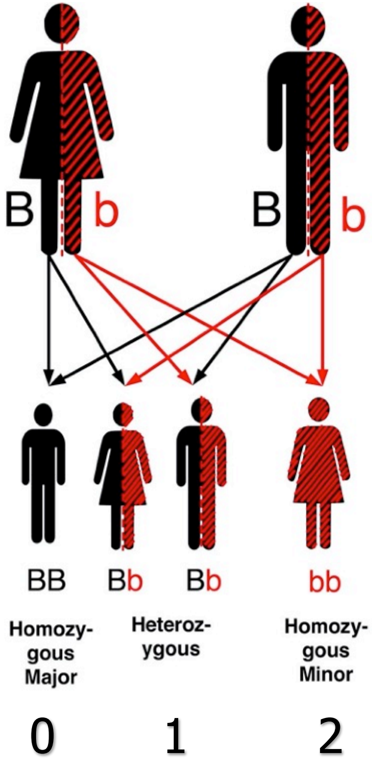
SAFARI

ETH zürich

Genome


Cell
Chromosome
Gene
LIFE CODE
RECIPES
Chapter 3
CTGTACGTCTA
ACCCCGTACGTC
CCGTACCGTTTT
ACTGACCCCGTA
CGTCCCGTACCG
TTTTTACTGACC
CCGTACGTCCCG
TACGAATCTACG
99.9% Identical

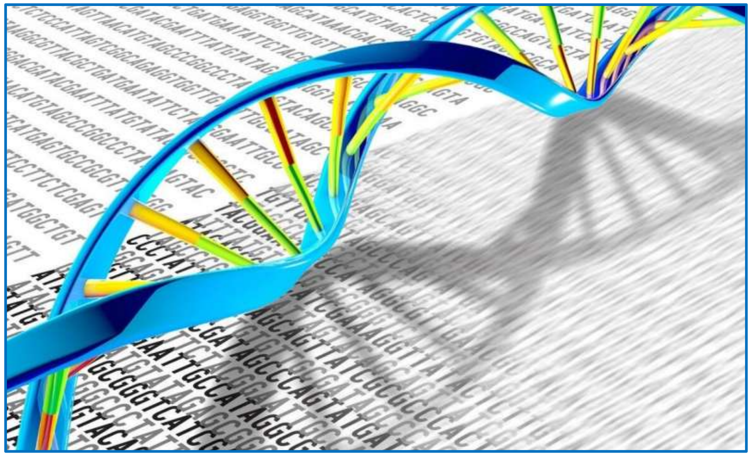
Mendel's Law



	Father		
Mother	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)

	Child		
Mother	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



© Medical Press

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

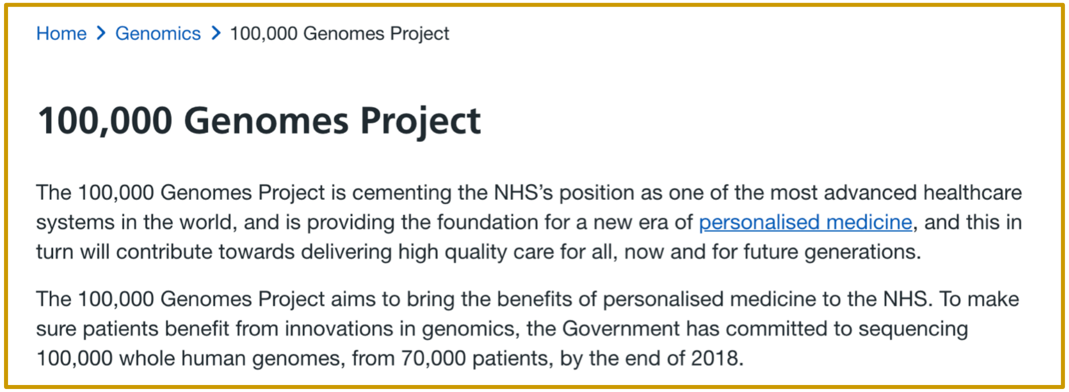
105 Million Sequenced Human genome

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

The Genomic Era

aviesan
alliance nationale
pour les sciences de la vie et de la santé
aviesan
alliance nationale
pour les sciences de la vie et de la santé
FRANCE MÉDECINE GÉNOMIQUE 2025

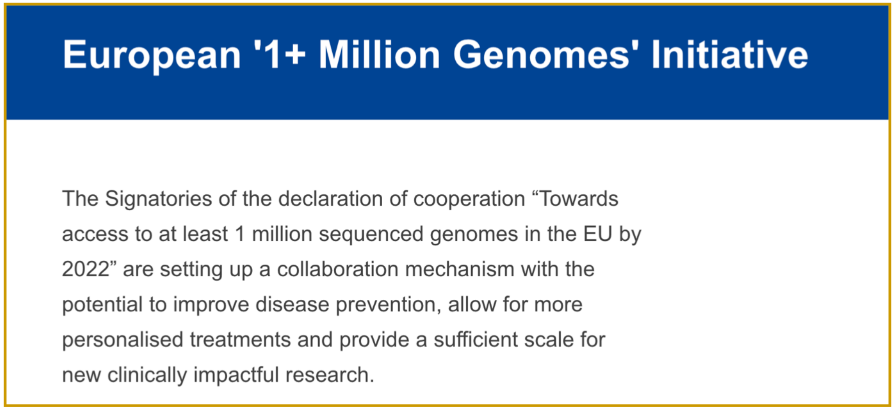
Plan France médecine génomique 2025



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.

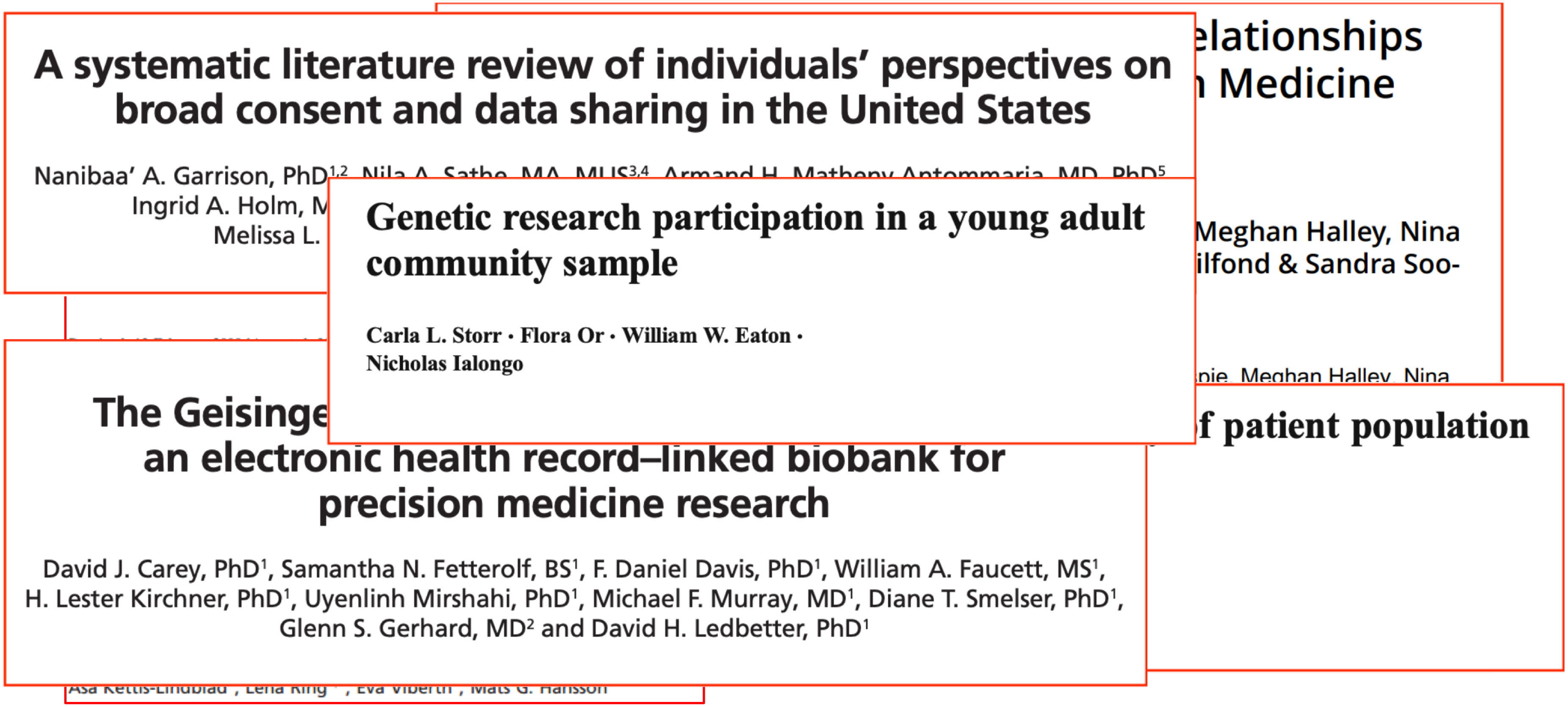
The 100,000 Genomes Project aims to bring the benefits of personalised medicine to the NHS. To make sure patients benefit from innovations in genomics, the Government has committed to sequencing 100,000 whole human genomes, from 70,000 patients, by the end of 2018.



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.



elationships
n Medicine

Meghan Halley, Nina
ilfond & Sandra Soo-

pie, Meghan Halley, Nina

f patient population

A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States

Nanibaa' A. Garrison, PhD[1,2], Nila A. Sathe, MA, MLIS[3,4], Armand H. Matheny Antommaria, MD, PhD[5]
Ingrid A. Holm, M
Melissa L.

Genetic research participation in a young adult community sample

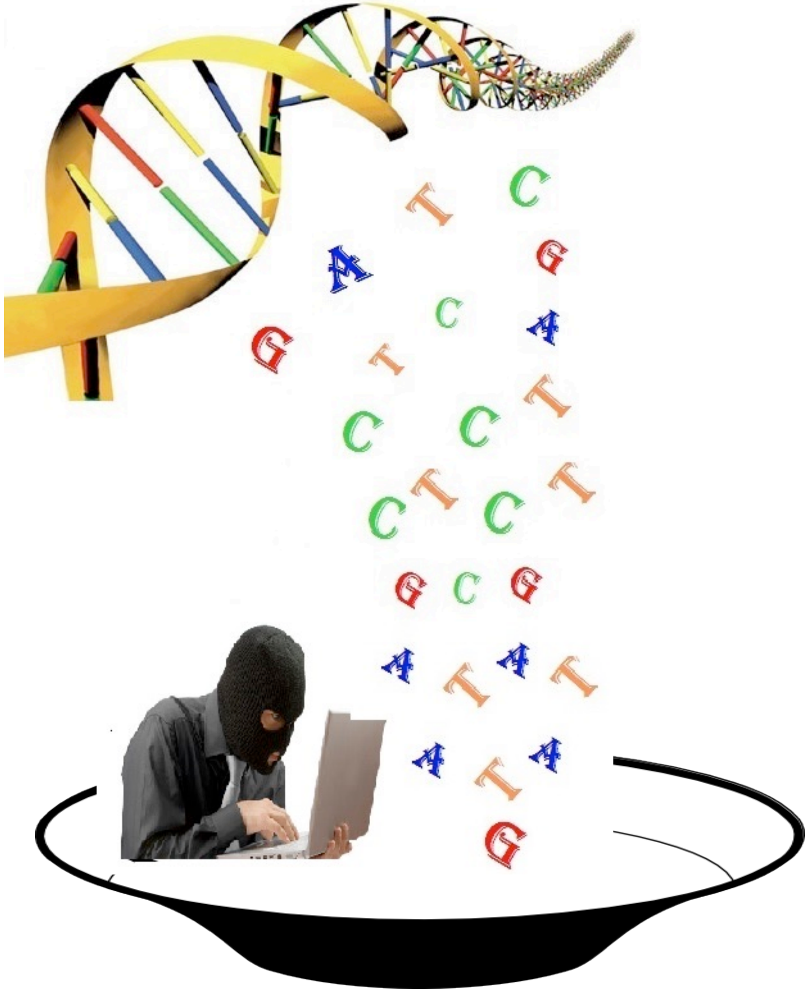
Carla L. Storr · Flora Or · William W. Eaton · Nicholas Ialongo

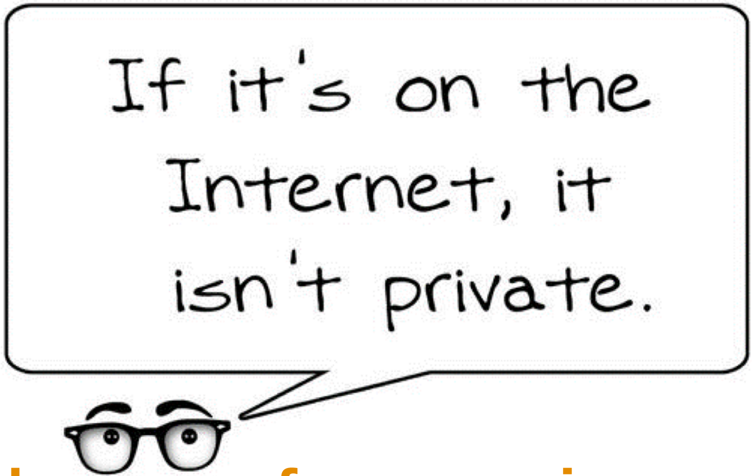
The Geisinge
an electronic health record–linked biobank for precision medicine research

David J. Carey, PhD[1], Samantha N. Fetterolf, BS[1], F. Daniel Davis, PhD[1], William A. Faucett, MS[1], H. Lester Kirchner, PhD[1], Uyenlinh Mirshahi, PhD[1], Michael F. Murray, MD[1], Diane T. Smelser, PhD[1], Glenn S. Gerhard, MD[2] and David H. Ledbetter, PhD[1]

Åsa Kettis-Lindblad , Lena Ring , Eva Viberth , Mats G. Hansson

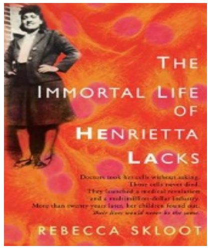
Privacy Risks





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

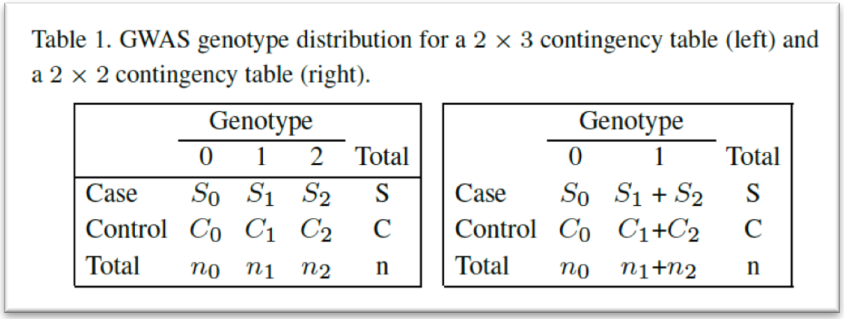
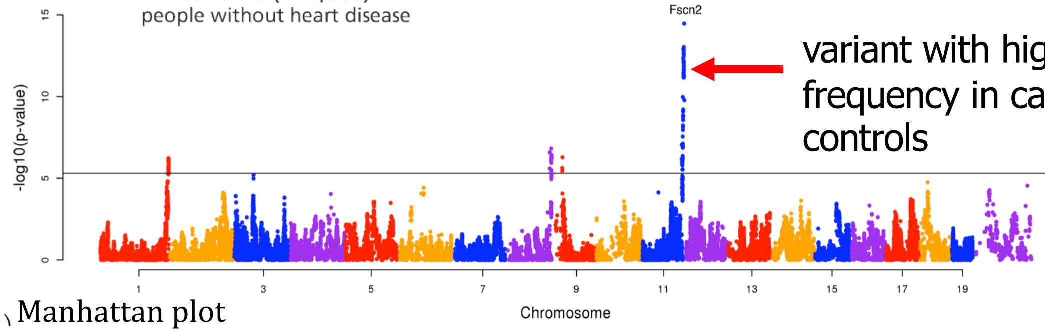


Table 1. GWAS genotype distribution for a 2 × 3 contingency table (left) and a 2 × 2 contingency table (right).

	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

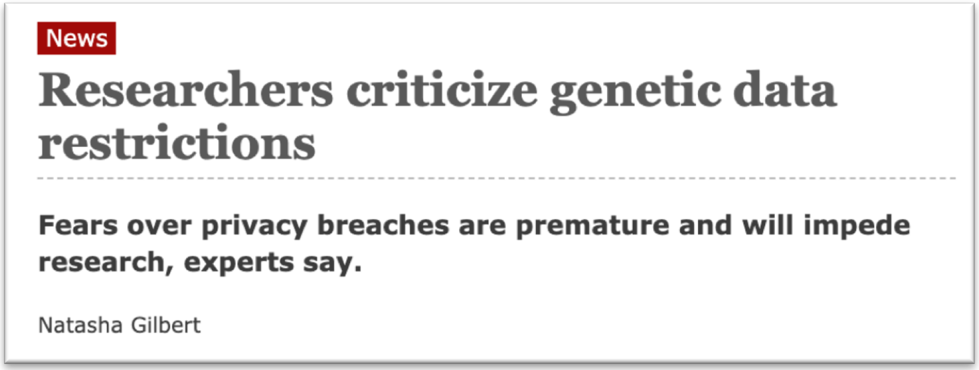
	Genotype		
	0	1	Total
Case	S_0	$S_1 + S_2$	S
Control	C_0	C_1+C_2	C
Total	n_0	n_1+n_2	n



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.

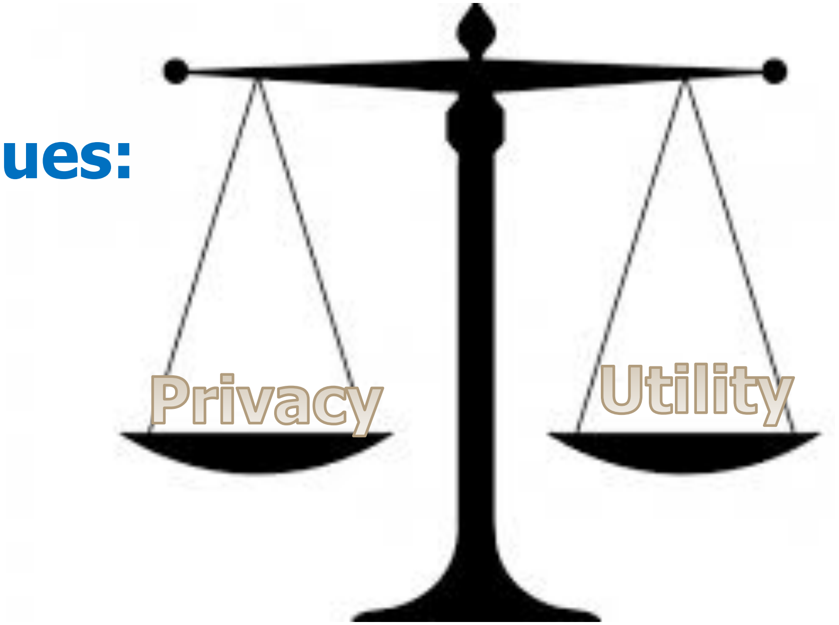
- NIH restricts the access to key results and data of GWAS to only trusted individuals.

Privacy-Utility Tradeoff

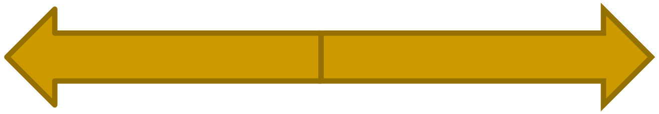
- Hiding some important data needs to tradeoff between **privacy** and **utility**.

➢ **Privacy preserving techniques:**

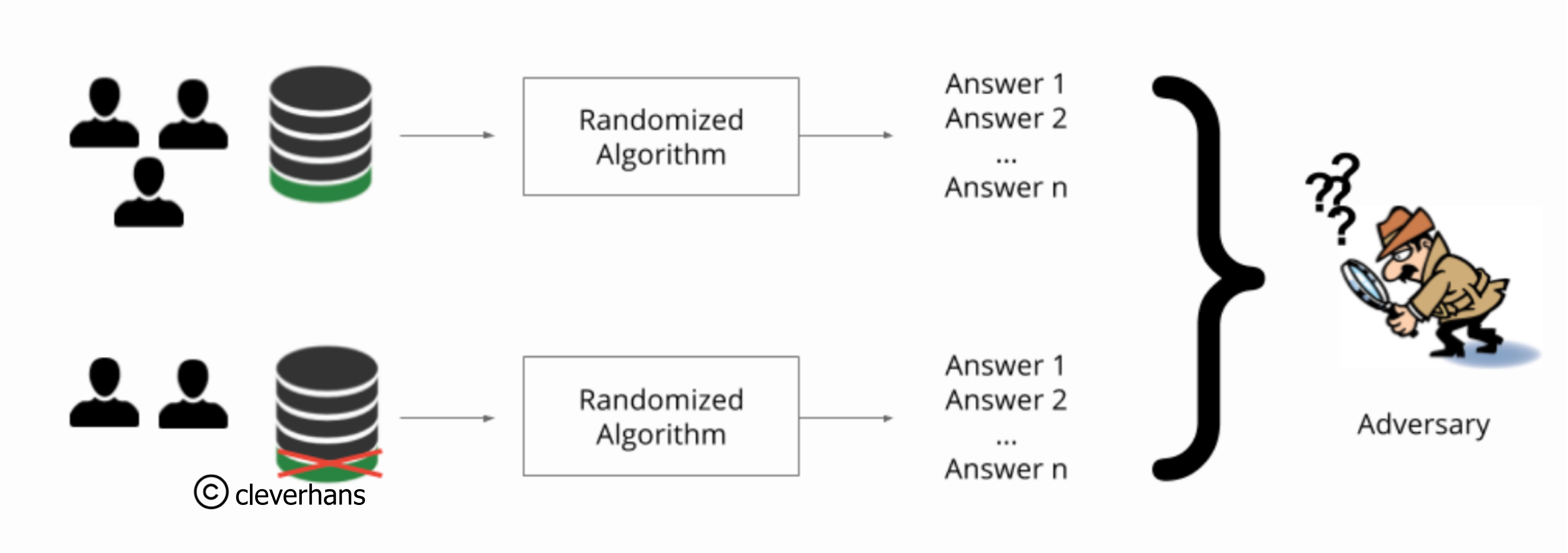
- K-anonymity.
- l-diversity.
- t-closeness.
- **Differential privacy.**
- Crypto-based techniques.



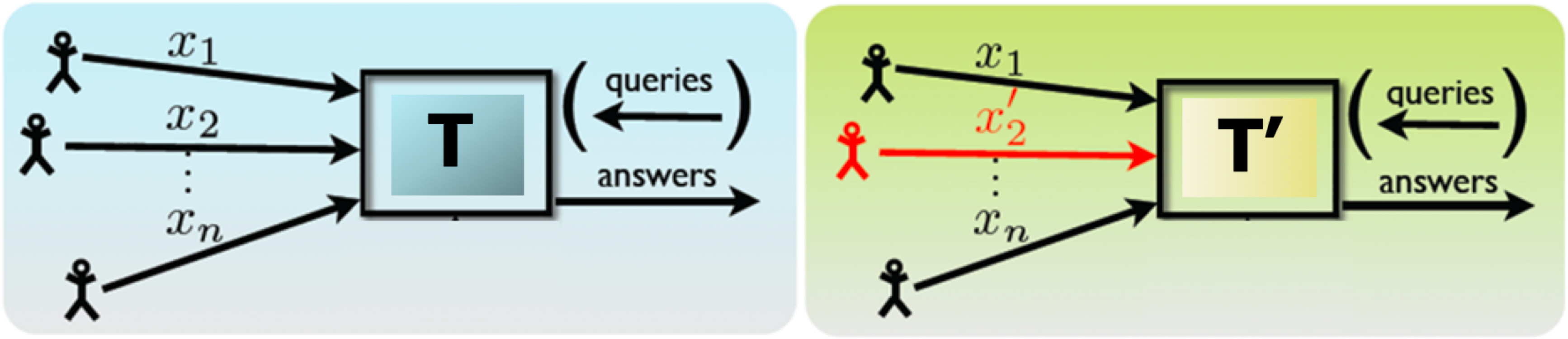
1000
0101
1101
0101
1011
© NIH


Privacy-Preserving
Computing
Sharing

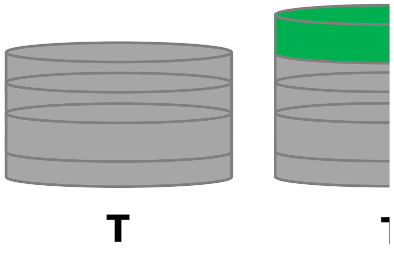
Differential Privacy

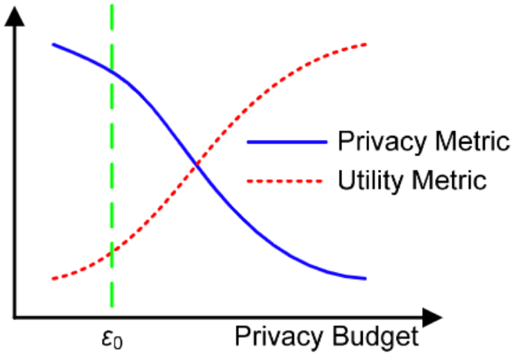


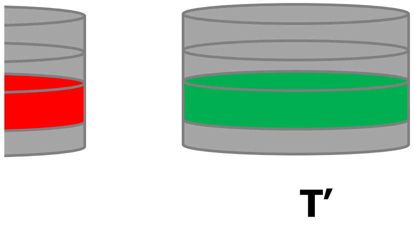
Differential Privacy



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







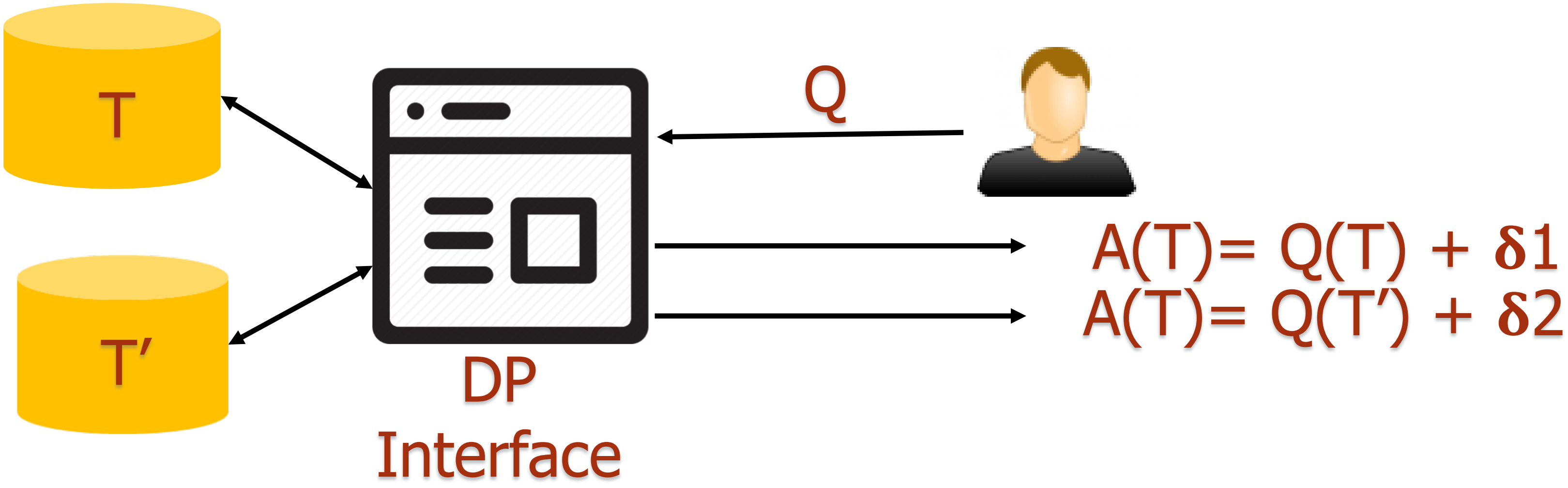
Unbounded Differential

d Differential Privacy (BDP)

© 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



A(T)= Q(T) + **δ**1
A(T)= Q(T′) + **δ**2

Differential Privacy

Google

Differential privacy library

Google Developer (2019)

Privacy

Differential Privacy Team, Apple (2017)

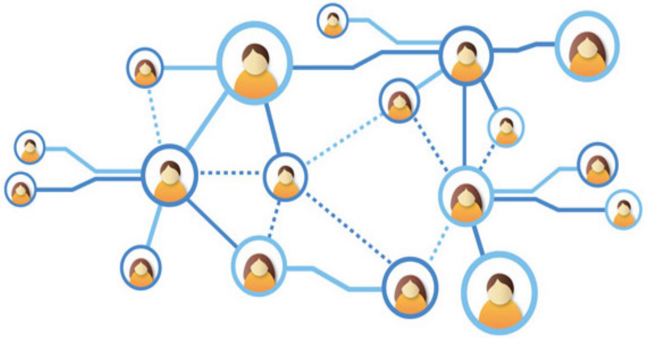
Microsoft Research

Collecting Telemetry Data Privately (2017)

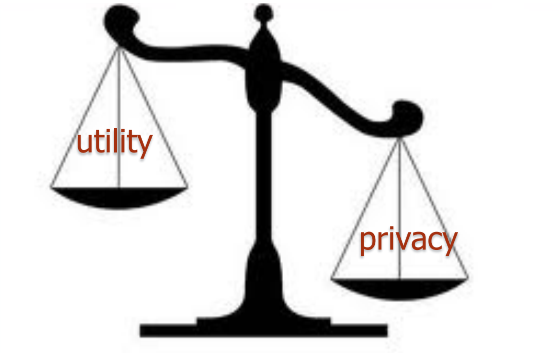
Differentially Private Publication System (2018)

Research Problem

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

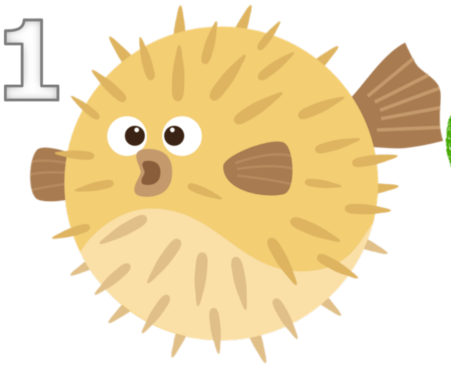


- Current DP-based mechanisms which consider the tuples correlation, provide poor accuracy.



Related Works

1



✓ A generalization of DP

✗ No perturbation algorithm

Pufferfish Framework
(Kifer and Machanavajjhala, 2012)

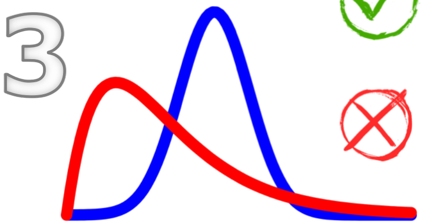
2

✓ Perturbation mechanisms

✗ Deterministic constraints for the adversary

Blowfish Framework
(He et al., 2014)

3



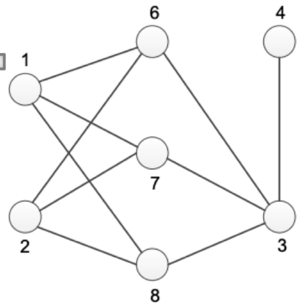
✓ Perturbation mechanisms

✗ 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)

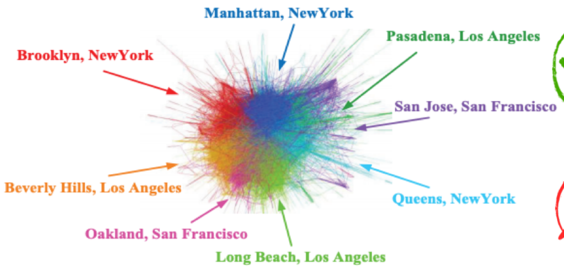
5

✓ 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]

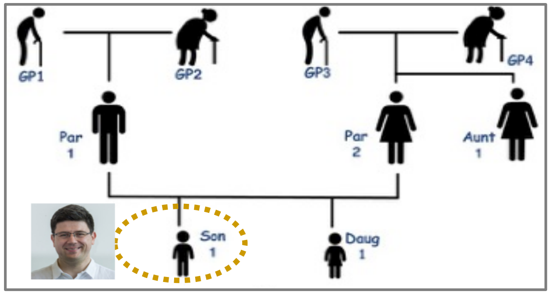
Our Contributions

Countermeasures

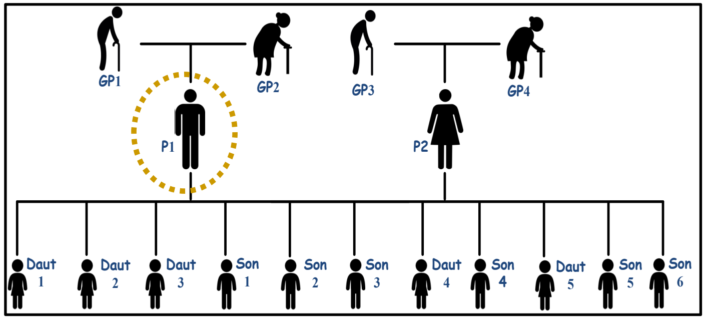
4 Ɛ-differential privacy for sharing genomic datasets with dependent tuples . **[Bioinformatics’19]**

5 Selective hiding mechanism and differential privacy. **[arXiv’21]**

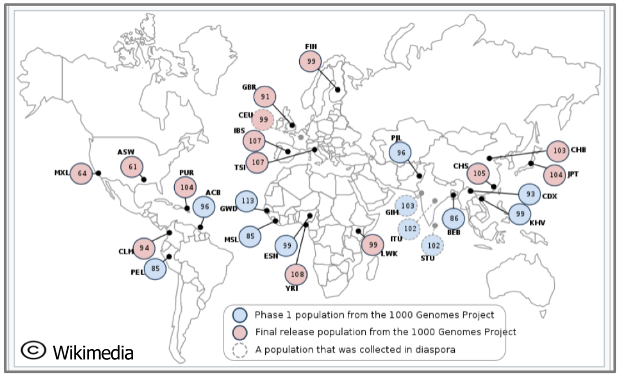
Dataset Description



Manuel Corpas Family

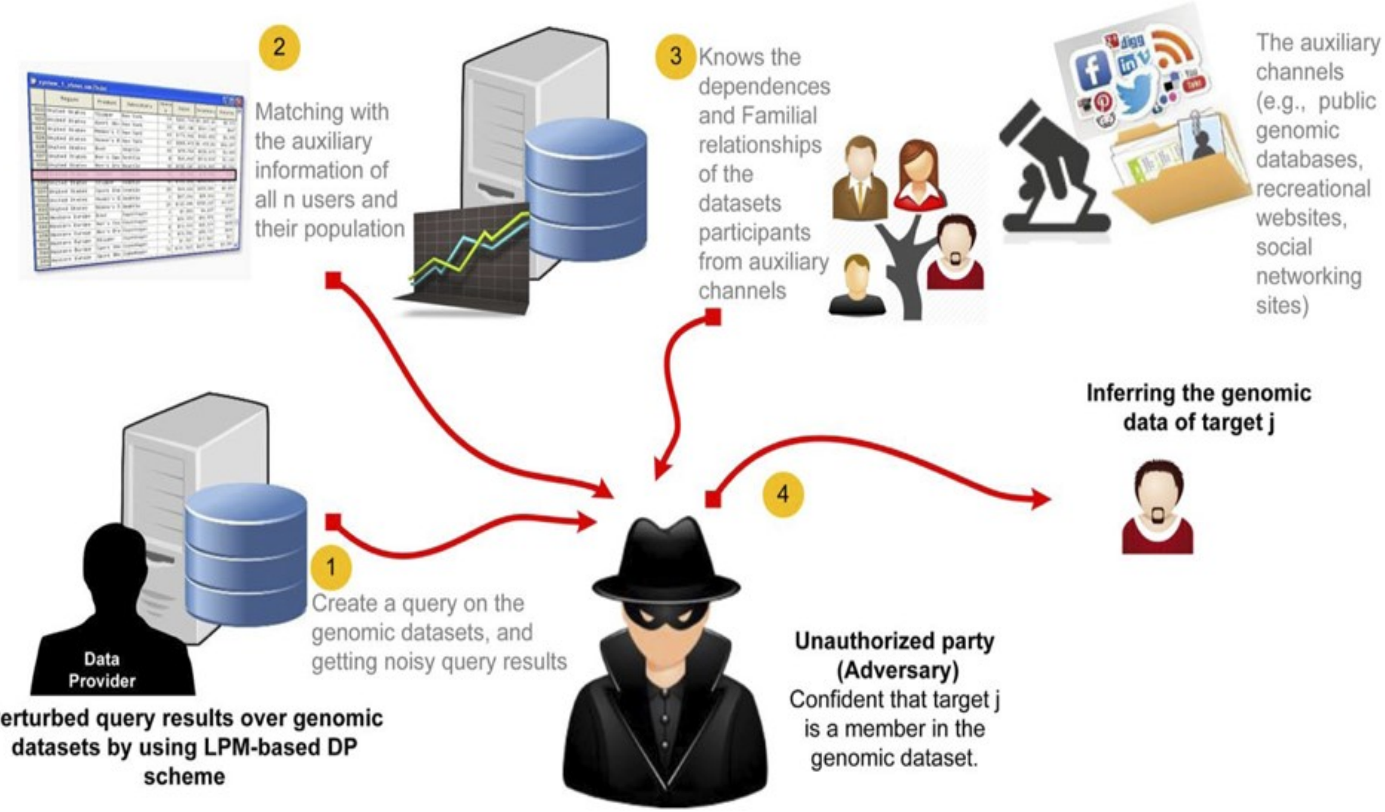


CEPH/Utah Family



n= 2514 or n= 2508, m =100

DP Inference Attacks



Data Provider

Perturbed query results over genomic datasets by using LPM-based DP scheme

1 Create a query on the genomic datasets, and getting noisy query results

2 Matching with the auxiliary information of all n users and their population

3 Knows the dependences and Familial relationships of the datasets participants from auxiliary channels

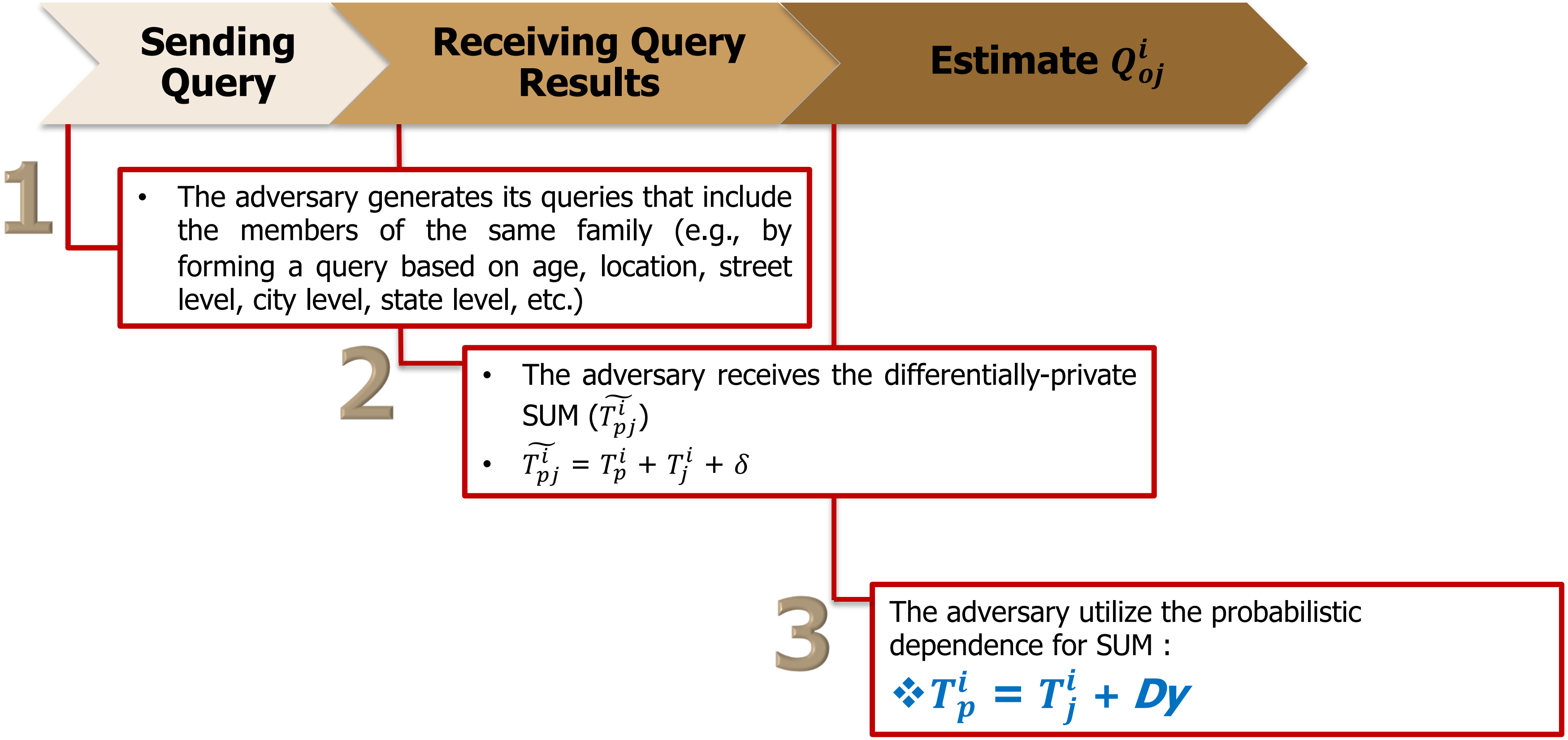
The auxiliary channels (e.g., public genomic databases, recreational websites, social networking sites)

Unauthorized party (Adversary)
Confident that target j is a member in the genomic dataset.

4

Inferring the genomic data of target j

Attribute Inference Attack



Sending Query

Receiving Query Results

Estimate Q_{oj}^{i}

1

- The adversary generates its queries that include the members of the same family (e.g., by forming a query based on age, location, street level, city level, state level, etc.)

2

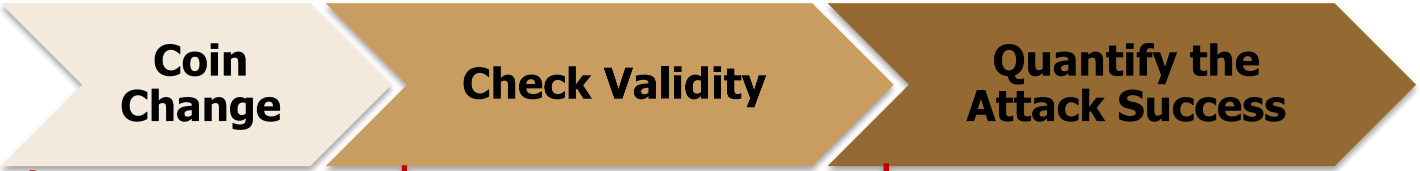
- The adversary receives the differentially-private SUM ($\widetilde{T_{pj}^{i}}$)
- $\widetilde{T_{pj}^{i}} = T_p^i + T_j^i + \delta$

3

The adversary utilize the probabilistic dependence for SUM :

❖ $\boldsymbol{T_p^i = T_j^i + Dy}$

Attribute Inference Attack



4

- The adversary obtains all the possible partitions of T^i_{pj} (each partition will include (p+1) individuals).

	T^i_{pj} (Sum)	p+1 participa			
✓	6	4			
✓	6	4			
✗	6	4	1	3	1

- The adversary uses Mendel's law to find the valid permutations for each partition. Then, he computes the probability by considering potential values of SNP i (0, 1, 2) for target j.

	Father	Mother	Son	Son
✓	1	2	2	1
✗	1	2	1	0

6

- Estimation error metric:

$$E = \sum_{i=1}^{m} P\left(x_j^i \mid X_j\right) \left|Dist\left(x_j^i, {x'}_j^i\right)\right|$$

- Leaked information metric

$$L = \sum_{i=1}^{m} 1 - |sgn(Dist(x_j^i, {x'}_j^i))|$$

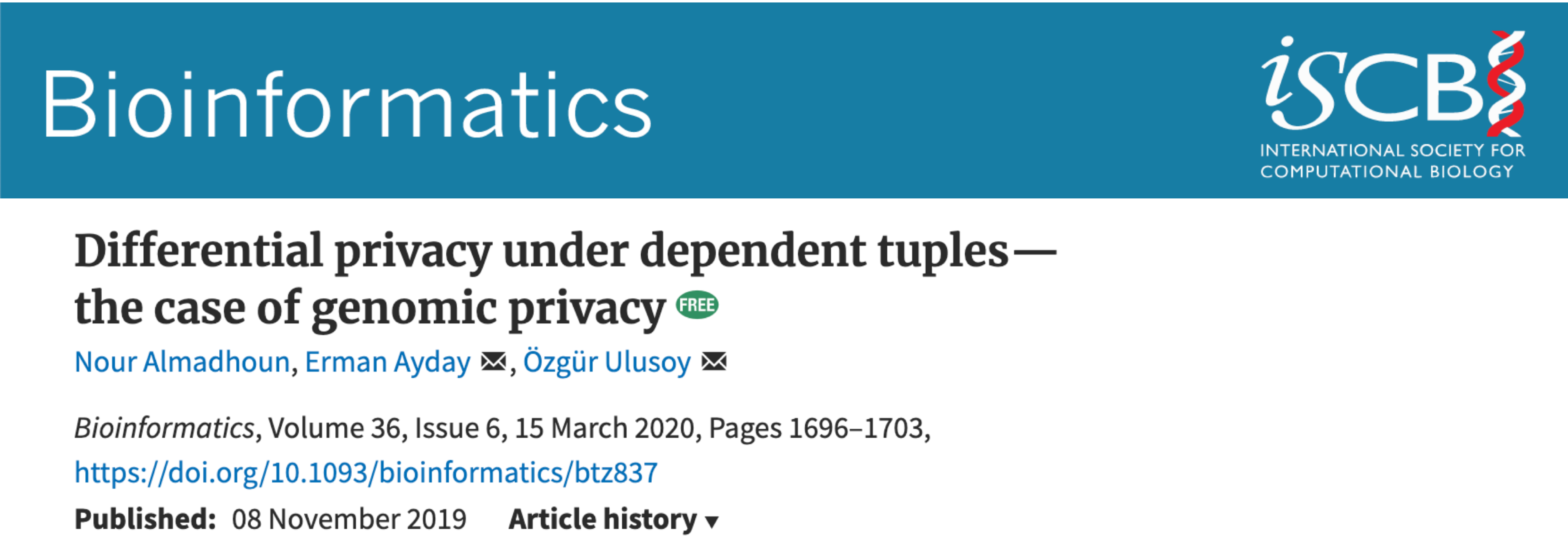
Key Results

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

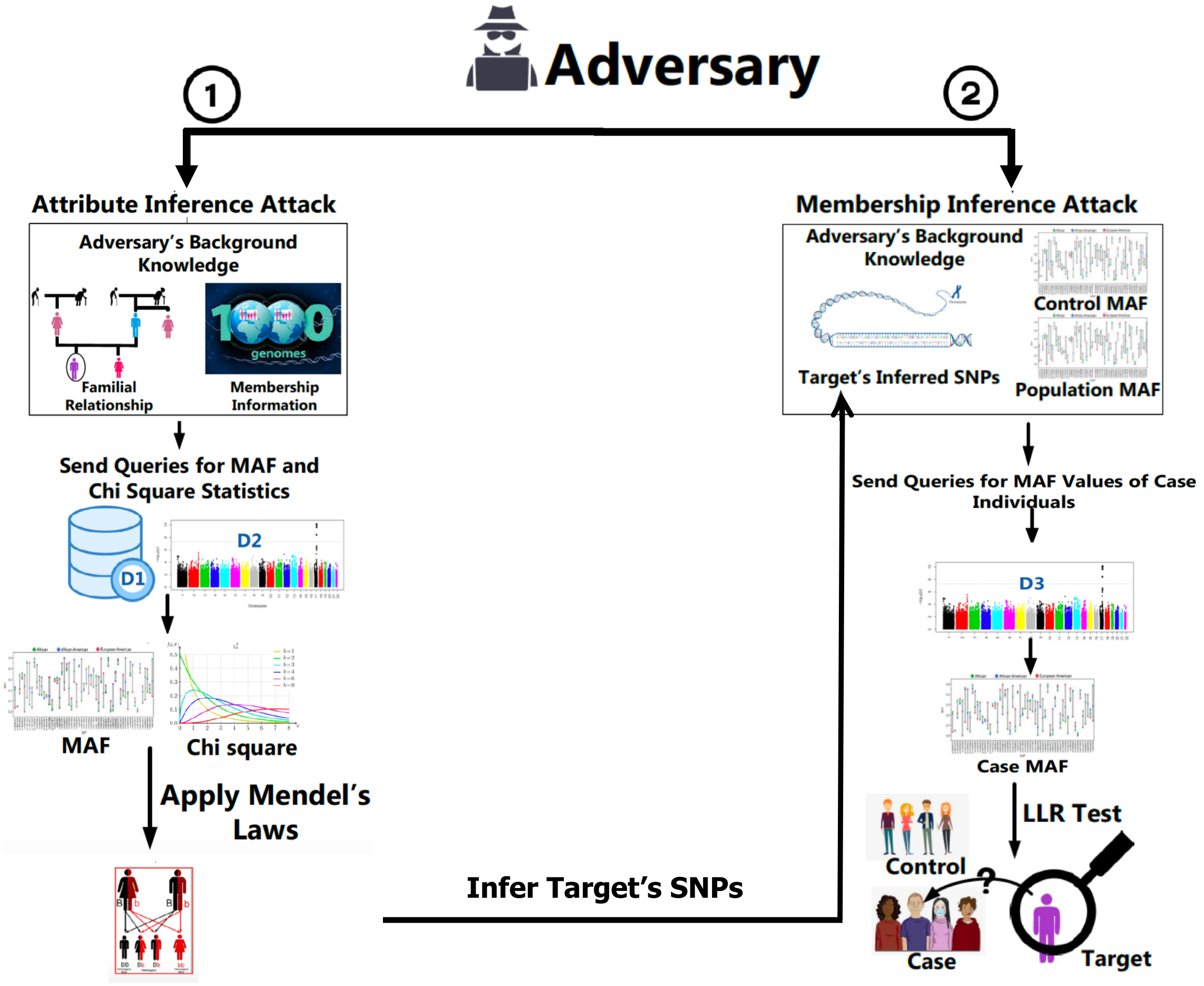
Differential privacy under dependent tuples—the case of genomic privacy FREE

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



Adversary

1

2

Attribute Inference Attack

Adversary's Background Knowledge

Familial Relationship

Membership Information

Send Queries for MAF and Chi Square Statistics

MAF

Chi square

Apply Mendel's Laws

Infer Target's SNPs

Membership Inference Attack

Adversary's Background Knowledge

Target's Inferred SNPs

Control MAF

Population MAF

Send Queries for MAF Values of Case Individuals

Case MAF

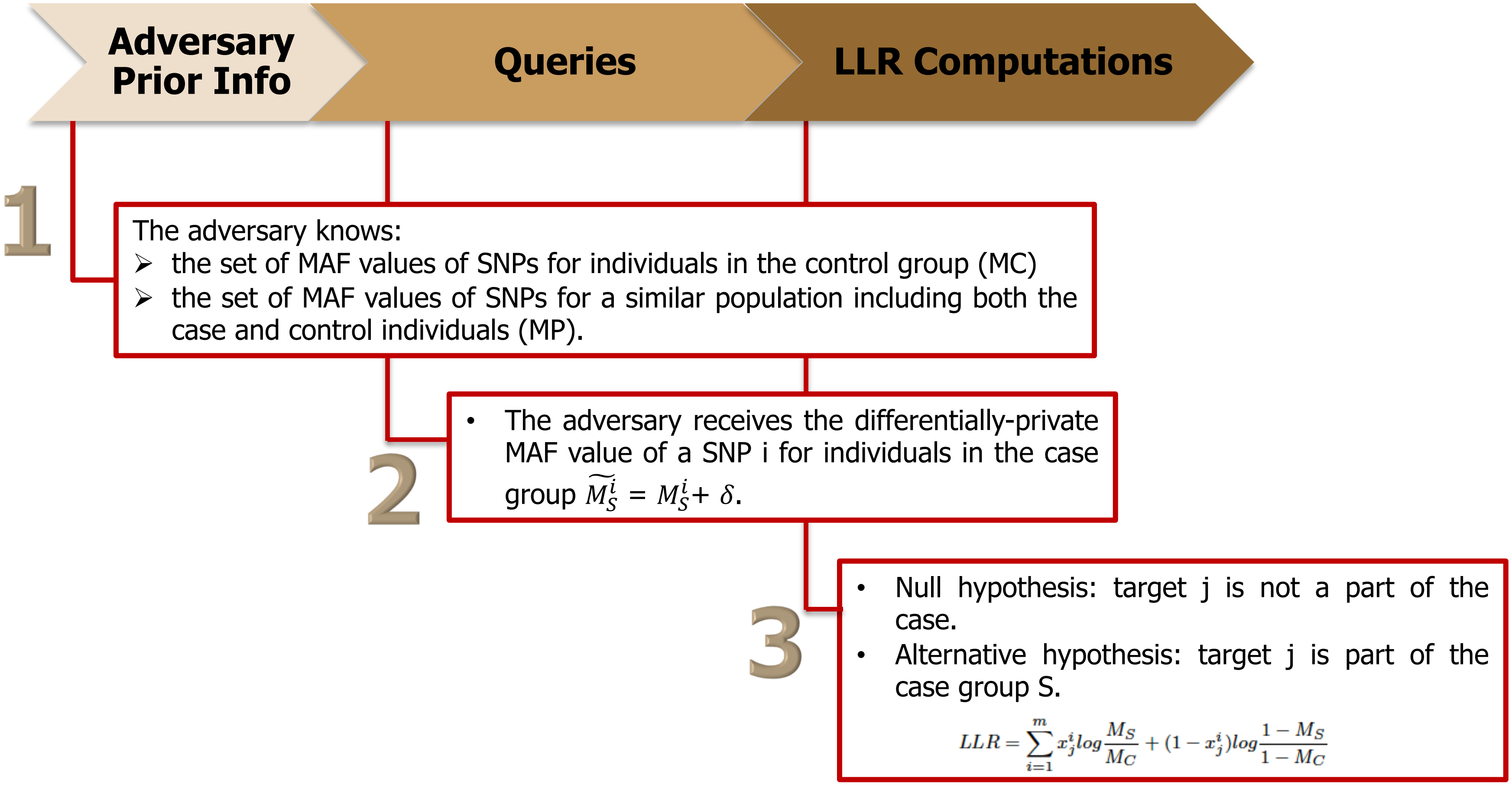
LLR Test

Control

Case

Target

Membership Inference Attack



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).

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

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

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

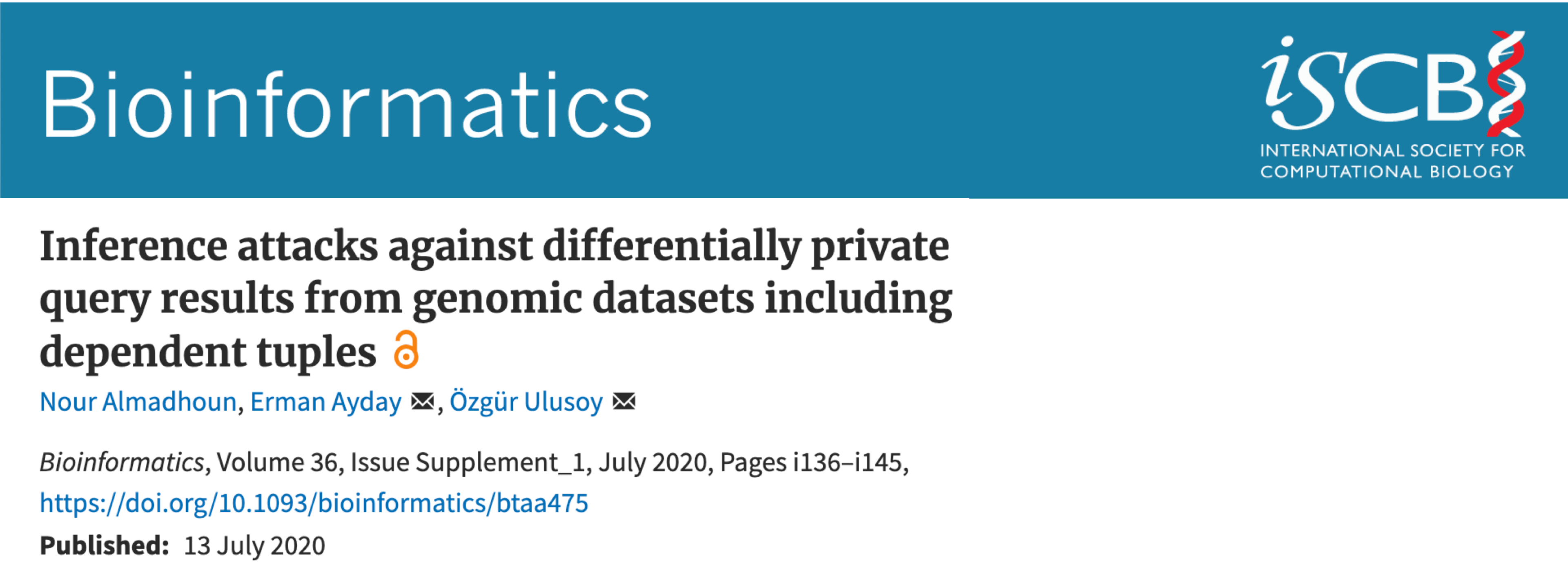
Key Results

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]

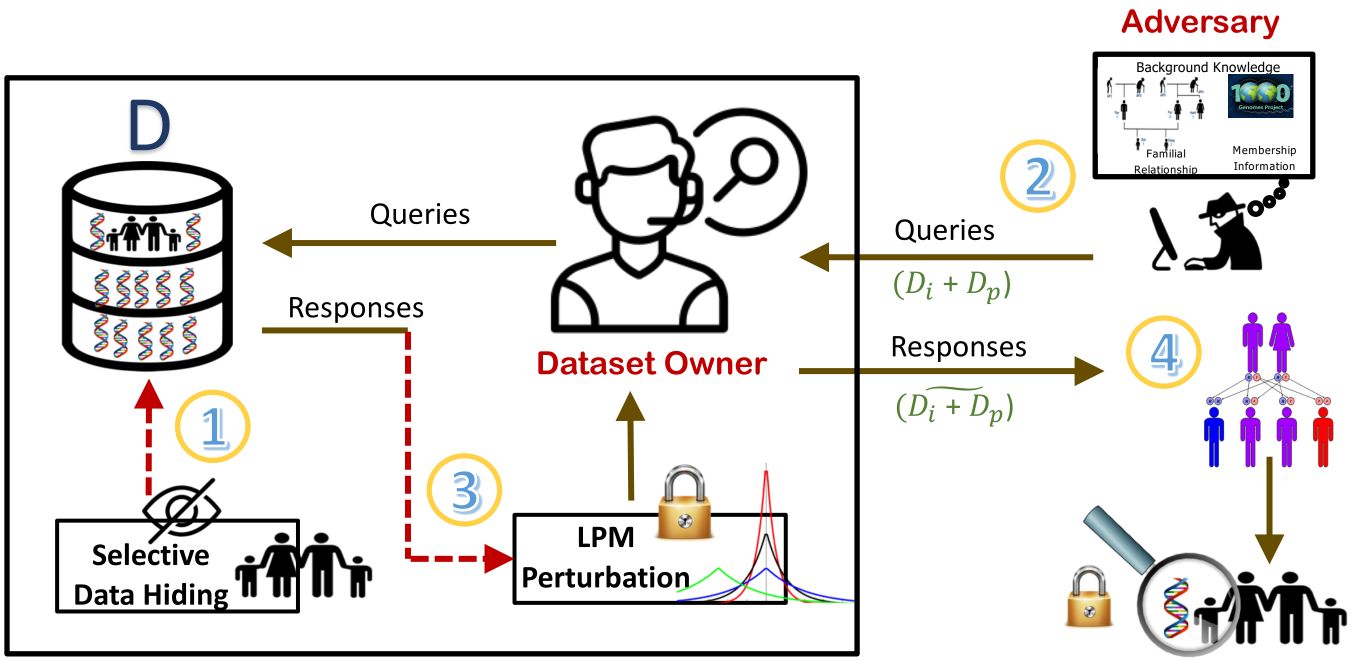


Inference attacks against differentially private query results from genomic datasets including dependent tuples

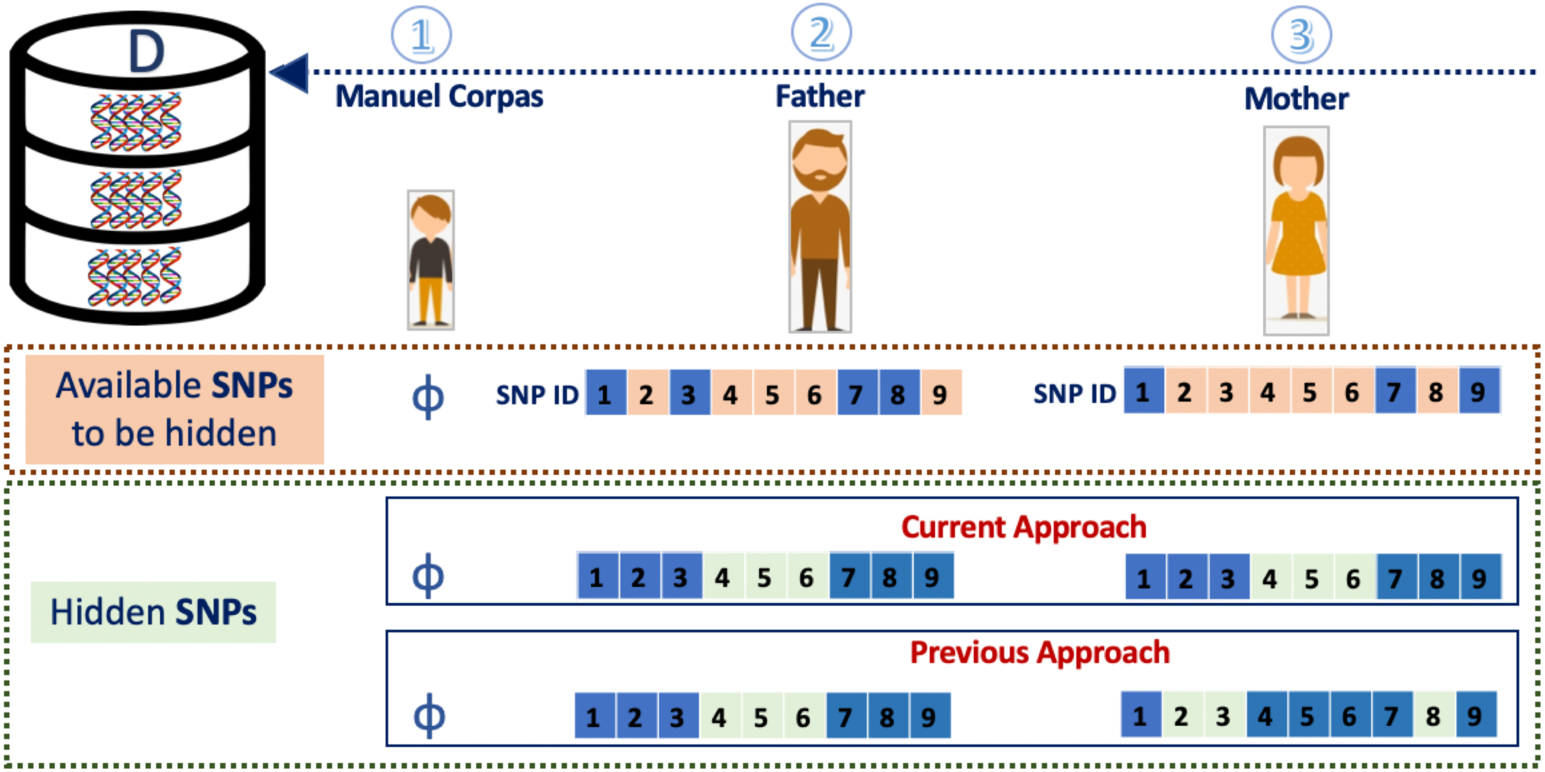
Nour Almadhoun, Erman Ayday ✉, Özgür Ulusoy ✉

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]

arXiv.org > cs > arXiv:2106.05211

Computer Science > Cryptography and Security

[Submitted on 9 Jun 2021]

Near-Optimal Privacy-Utility Tradeoff in Genomic Studies Using Selective SNP Hiding

Nour Almadhoun Alserr, Gulce Kale, Onur Mutlu, Oznur Tastan, Erman Ayday

GenShare Model

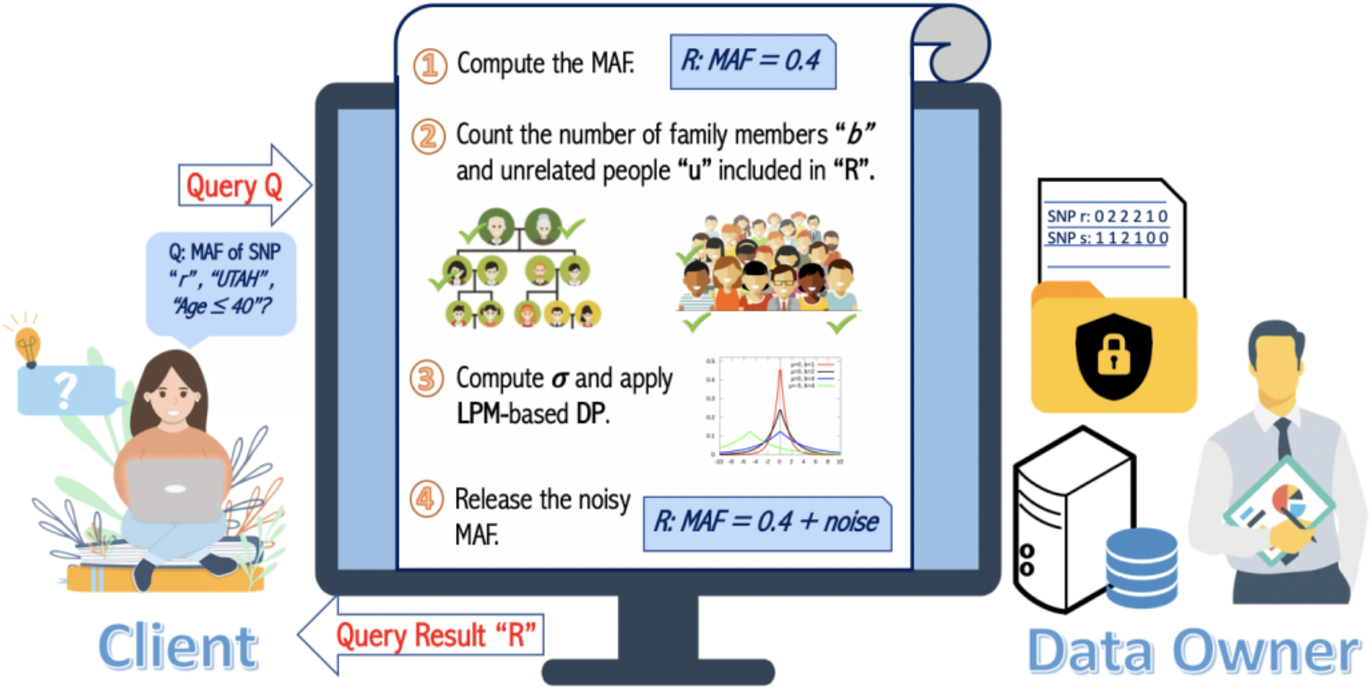
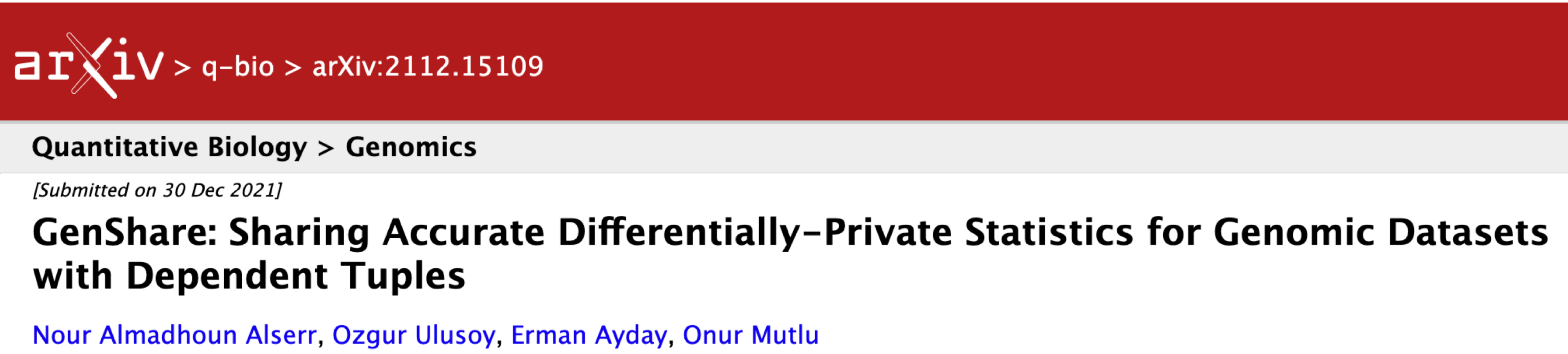

Query Q
Q: MAF of SNP "r", "UTAH", "Age ≤ 40"?
?
Client
Query Result "R"
① Compute the MAF.
R: MAF = 0.4
② Count the number of family members "b" and unrelated people "u" included in "R".
③ Compute σ and apply LPM-based DP.
④ Release the noisy MAF.
R: MAF = 0.4 + noise
SNP r: 0 2 2 2 1 0
SNP s: 1 1 2 1 0 0
Data Owner

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



arXiv > q-bio > arXiv:2112.15109

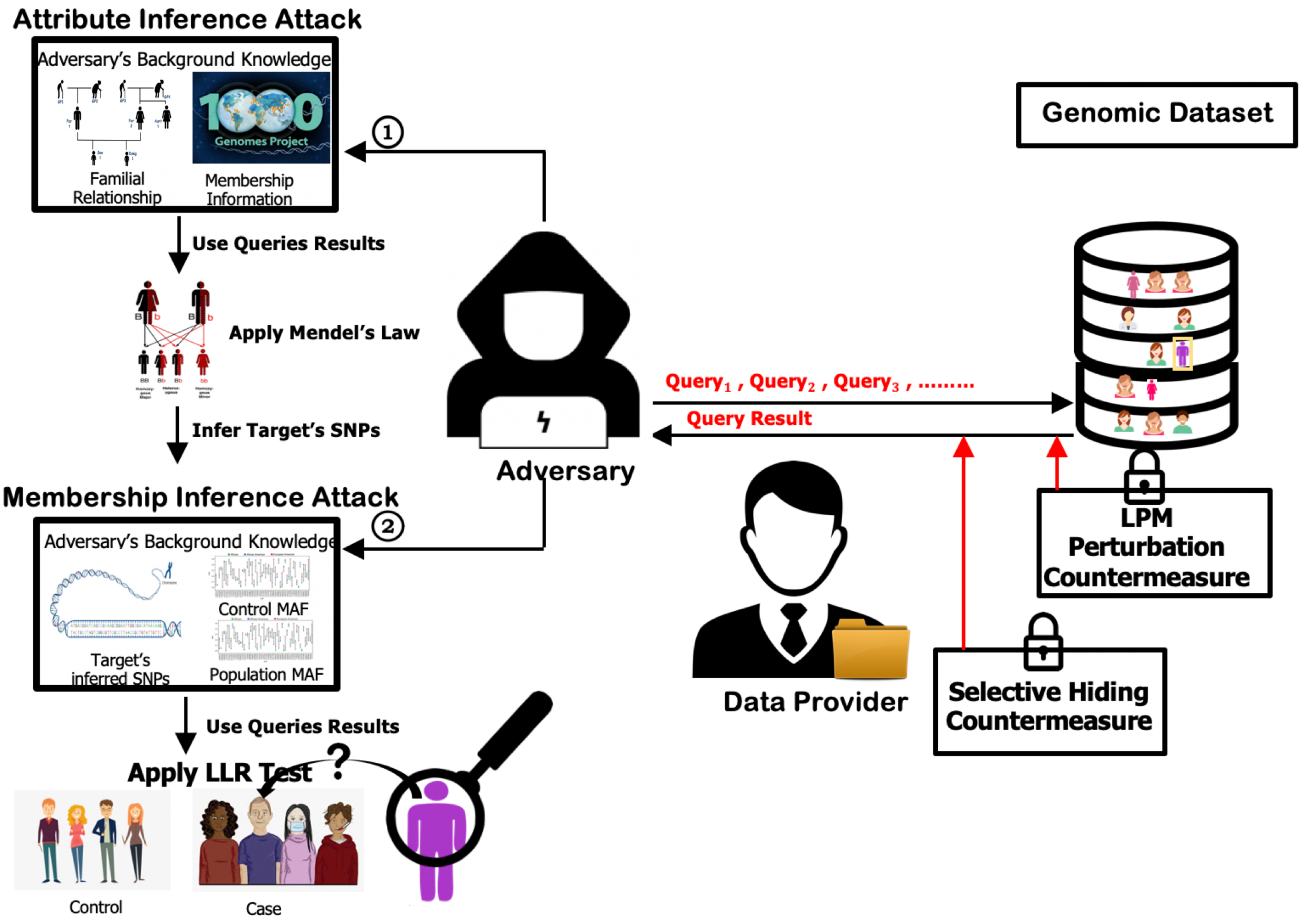
Quantitative Biology > Genomics

[Submitted on 30 Dec 2021]

GenShare: Sharing Accurate Differentially-Private Statistics for Genomic Datasets with Dependent Tuples

Nour Almadhoun Alserr, Ozgur Ulusoy, Erman Ayday, Onur Mutlu

Full Model



Attribute Inference Attack

Adversary's Background Knowledge

Familial Relationship

Membership Information

Use Queries Results

Apply Mendel's Law

Infer Target's SNPs

Membership Inference Attack

Adversary's Background Knowledge

Target's inferred SNPs

Control MAF

Population MAF

Use Queries Results

Apply LLR Test

Control

Case

Adversary

Query$_1$, Query$_2$, Query$_3$, ...

Query Result

Genomic Dataset

Data Provider

LPM Perturbation Countermeasure

Selective Hiding Countermeasure

P&S Mobile Genomics

Lecture 10: Genomic Data Sharing Under Differential Privacy

Dr. Nour Almadhoun Alserr

ETH Zurich

Spring 2022

17 May 2022

SAFARI

ETH zürich