# Génétique, famille et vie privée

# Kin Genomic Privacy

Prof. Kévin Huguenin
University of Lausanne

March 16th, 2021

UNIL | Université de Lausanne

école normale supérieure rennes

# Our research

Privacy

Security

Interdependence

Trade-off

Location

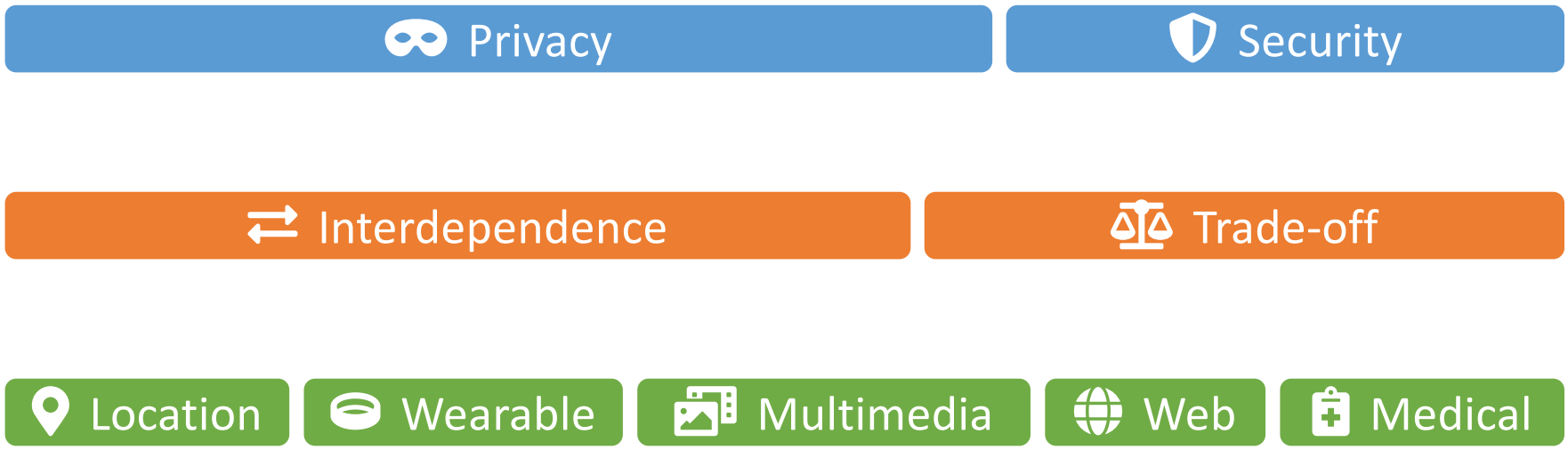
Wearable

Multimedia

Web

Medical

# Our research

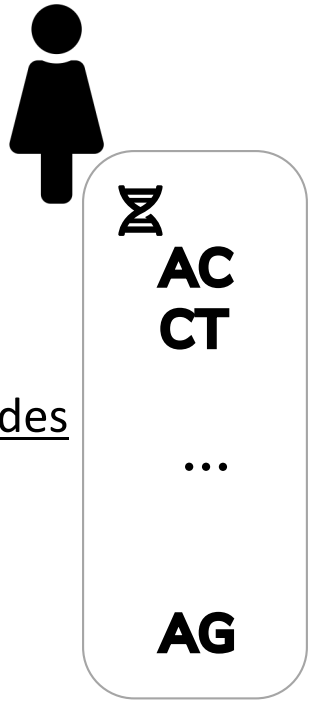



What is security and privacy research?

# Quantification of Kin Genomic Privacy for the Masses

M. Humbert, D. Dupertuis, M. Cherubini, and K. Huguenin. Quantification of Kin Genomic Privacy for the Masses. Major revision for USENIX Security Symp. 2021.

# Genomics



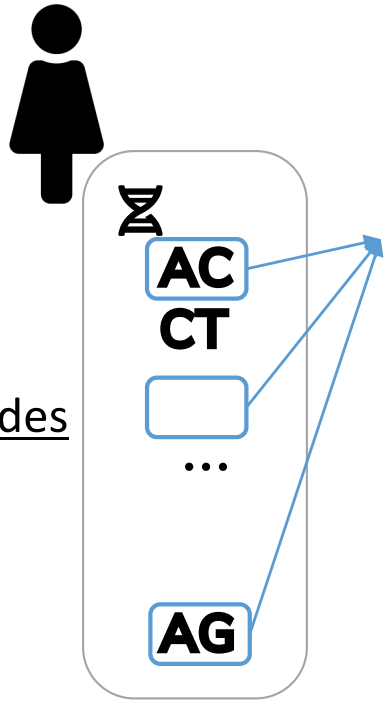

**Genome / DNA:**

≈3B pairs of nucleotides
(A, C, G, T)

# Genomics

**Genome / DNA:**

≈3B pairs of nucleotides
(A, C, G, T)



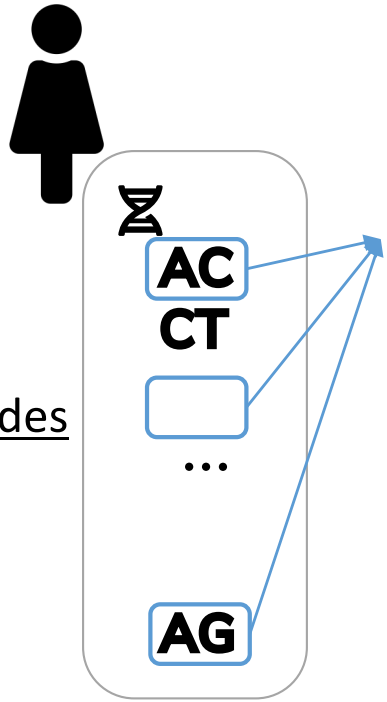

**SNP**:

- Position where individuals' genomes can differ (≈150M SNPs)
- Combination from two values: "m" (minor allele) and "M" (major allele).
  Three possibilities "MM", "Mm", "mm" (ordering does not matter)
- Population statistics on allele frequencies (MAF)

# Genomics

**Genome / DNA:**

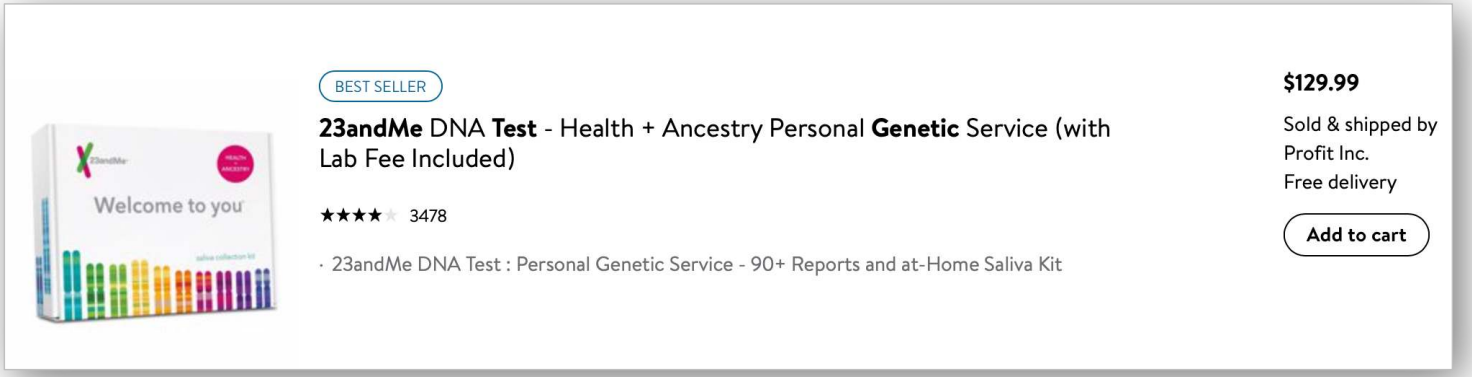
≈3B pairs of nucleotides
(A, C, G, T)




**SNP**:

- Position where individuals' genomes can differ (≈150M SNPs)
- Combination from two values: "m" (minor allele) and "M" (major allele).
  Three possibilities "MM", "Mm", "mm" (ordering does not matter)
- Population statistics on allele frequencies (MAF)

Source: https://www.ancestry.com/dna/




Source: https://www.23andme.com

# Genomic privacy

Your genome influences your physical appearance, but also your non-visible attributes, such as your predisposition to certain diseases or behaviors. It is is immutable.

# Genomic privacy

Your genome influences your physical appearance, but also your non-visible attributes, such as your predisposition to certain diseases or behaviors. It is is immutable.

This information can also be used to discriminate against you, for example, to deny you a health insurance, a job or a loan. It also creates (national) security threats.

Médecine Modifié le 15 octobre 2019 à 07:19

## Des tests ADN pour connaître les risques de développer une maladie

Faire analyser son ADN peut être risqué 19h30 / 4 min. / le 13 octobre 2019

**Découvrir dans l'ADN les risques de développer une maladie, la possibilité existe et est accessible en ligne pour une centaine de francs. Des millions de personnes ont fait le test, dont de nombreux Suisses. De quoi attiser l'intérêt des assurances.**

Les assureurs mènent un combat politique pour rendre obligatoire la transmission des analyses génétiques. Une première tentative a été menée sans succès l'an dernier. La commission du National avait pourtant émis une recommandation allant dans le sens de la demande des assurances, mais le Parlement n'avait pas suivi.

Parmi les arguments des opposants, la crainte que des personnes soient refusées par les assurances, ou renoncent à des analyses par peur de l'être. Mais pour les assureurs, le dossier est loin d'être refermé. "Nous referons exactement les mêmes demandes dans 3 ou 5 ans, affirme Valérie Bourdin. On verra, c'est finalement au législateur de prendre la décision".

Céline Brichet

Publié le 14 octobre 2019 à 06:46 - modifié le 15 octobre 2019 à 07:19

Source [FR]: https://www.rts.ch/info/sciences-tech/medecine/10779718-des-tests-adn-pour-connaitre-les-risques-de-developper-une-maladie.html

# Genomic privacy

Your genome influences your physical appearance, but also your non-visible attributes, such as your predisposition to certain diseases or behaviors. It is is immutable.

This information can also be used to discriminate against you, for example, to deny you a health insurance, a job or a loan. It also creates (national) security threats.

Médecine Modifié le 15 octobre 2019 à 07:19

**Des tests ADN pour connaître les risques de développer une maladie**

Faire analyser son ADN peut être risqué 19h30 / 4 min. / le 13 octobre 2019

**Découvrir dans l'ADN les risques de développer une maladie, la possibilité existe et est accessible en ligne pour une centaine de francs. Des millions de personnes ont fait le test, dont de nombreux Suisses. De quoi attiser l'intérêt des assurances.**

Les assureurs mènent un combat politique pour rendre obligatoire la transmission des analyses génétiques. Une première tentative a été menée sans succès l'an dernier. La commission du National avait pourtant émis une recommandation allant dans le sens de la demande des assurances, mais le Parlement n'avait pas suivi.

Parmi les arguments des opposants, la crainte que des personnes soient refusées par les assurances, ou renoncent à des analyses par peur de l'être. Mais pour les assureurs, le dossier est loin d'être refermé. "Nous referons exactement les mêmes demandes dans 3 ou 5 ans, affirme Valérie Bourdin. On verra, c'est finalement au législateur de prendre la décision".

Céline Brichet

Publié le 14 octobre 2019 à 06:46 - modifié le 15 octobre 2019 à 07:19

Source [FR]: https://www.rts.ch/info/sciences-tech/medecine/10779718-des-tests-adn-pour-connaitre-les-risques-de-developper-une-maladie.html

The New York Times

***Pentagon Warns Military Personnel Against At-Home DNA Tests***

The tests, from companies such as 23andMe and Ancestry, have become popular holiday gifts, but the military is warning service members of risks to their careers.

Source: https://www.nytimes.com/2019/12/24/us/military-dna-tests.html

# Kin genomic privacy

**Addressing the Concerns of the Lacks' Family: Quantification of Kin Genomic Privacy**

Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux
Laboratory for Communications and Applications
EPFL, Lausanne, Switzerland
firstname.lastname@epfl.ch

Amalio Telenti
Institute of Microbiology
University Hospital of Lausanne
Lausanne, Switzerland
amalio.telenti@chuv.ch

**ABSTRACT**

The rapid progress in human-genome sequencing is leading to a high availability of genomic data. This data is notoriously very sensitive and stable in time. It is also highly correlated among relatives. A growing number of genomes are becoming accessible online (e.g., because of leakage, or after their posting on genome-sharing websites). What are then the implications for kin genomic privacy? We formalize the problem and detail an efficient reconstruction attack based on graphical models and belief propagation. With this approach, an attacker can infer the genomes of the relatives of an individual whose genome is observed, relying notably on Mendel's Laws and statistical relationships between the nucleotides (on the DNA sequence). Then, to quantify the level of genomic privacy as a result of the proposed inference attack, we discuss possible definitions of *genomic privacy* metrics. Genomic data reveals Mendelian diseases and the likelihood of developing degenerative diseases such as Alzheimer's. We also introduce the quantification of *health privacy*, specifically the measure of how well the predisposition to a disease is concealed from an attacker. We evaluate our approach on actual genomic data from a pedigree and show the threat extent by combining data gathered from a genome-sharing website and from an online social network.

**Categories and Subject Descriptors**

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection*; J.3 [**Life and Medical Sciences**]: *Biology and genetics*; K.4.1 [**Computer and Society**]: Public Policy Issues—*Privacy*

**Keywords**

Genomic Privacy; Inference Algorithms; Metrics; Kinship

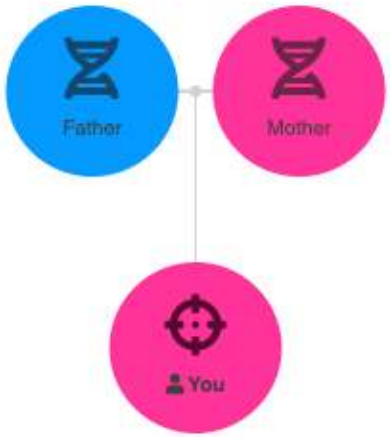
"The family of Henrietta Lacks (August 1, 1920 - October 4, 1951), whose DNA was sequenced and published online without the consent of her family.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CCS'13, November 4–8, 2013, Berlin, Germany.
Copyright 2013 ACM 978-1-4503-2477-9/13/11 ...$15.00.
http://dx.doi.org/10.1145/2508859.2516707



Source: Humbert et *al.*, CCS, Nov 2013
doi 10.1145/2508859.2516707

In each pair of nucleotides in your genome, one nucleotide/allele is inherited from your mother and the other from your father.



# Kin genomic privacy

Addressing the Concerns of the Lacks' Family: Quantification of Kin Genomic Privacy

Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux
Laboratory for Communications and Applications
EPFL, Lausanne, Switzerland
firstname.lastname@epfl.ch

Amalio Telenti
Institute of Microbiology
University Hospital of Lausanne
Lausanne, Switzerland
amalio.telenti@chuv.ch

ABSTRACT

The rapid progress in human-genome sequencing is leading to a high availability of genomic data. This data is notoriously very sensitive and stable in time. It is also highly correlated among relatives. A growing number of genomes are becoming accessible online (e.g., because of leakage, or after their posting on genome-sharing websites). What are then the implications for kin genomic privacy? We formalize the problem and detail an efficient reconstruction attack based on graphical models and belief propagation. With this approach, an attacker can infer the genomes of the relatives of an individual whose genome is observed, relying notably on Mendel's Laws and statistical relationships between the nucleotides (on the DNA sequence). Then, to quantify the level of genomic privacy as a result of the proposed inference attack, we discuss possible definitions of *genomic privacy* metrics. Genomic data reveals Mendelian diseases and the likelihood of developing degenerative diseases such as Alzheimer's. We also introduce the quantification of *health privacy*, specifically the measure of how well the predisposition to a disease is concealed from an attacker. We evaluate our approach on actual genomic data from a pedigree and show the threat extent by combining data gathered from a genome-sharing website and from an online social network.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection*; J.3 [**Life and Medical Sciences**]: *Biology and genetics*; K.4.1 [**Computer and Society**]: Public Policy Issues—*Privacy*

Keywords

Genomic Privacy; Inference Algorithms; Metrics; Kinship

"The family of Henrietta Lacks (August 1, 1920 - October 4, 1951), whose DNA was sequenced and published online without the consent of her family.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CCS'13, November 4–8, 2013, Berlin, Germany.
Copyright 2013 ACM 978-1-4503-2477-9/13/11...$15.00.
http://dx.doi.org/10.1145/2508859.2516707

1. INTRODUCTION

With the help of rapidly developing technology, DNA sequencing is becoming less expensive. As a consequence, the research in genomics has gained speed in paving the way to personalized (genomic) medicine, and geneticists need large collections of human genomes to further increase this speed. Furthermore, individuals are using their genomes to learn about their (genetic) predispositions to diseases, their ancestries, and even their (genetic) compatibilities with potential partners. This trend has also caused the launch of health-related websites and online social networks (OSNs), in which individuals share their genomic data (e.g., OpenSNP [1] or 23andMe [2]). Today, already today, thousands of genomes are available online.

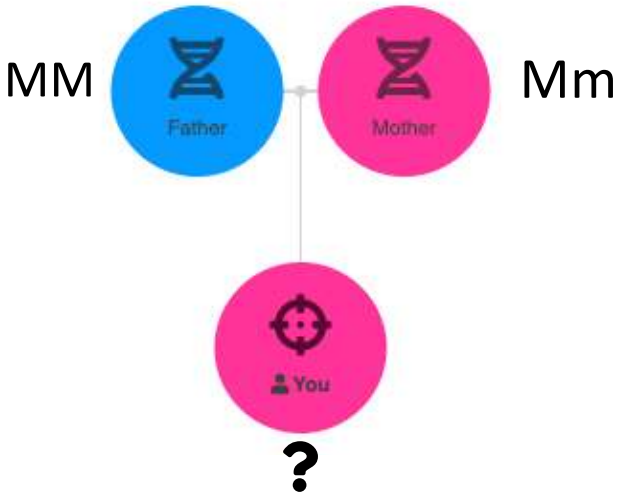
Even though most of the genomes on the Internet are anonymized, it is possible to find genomes with the identifiers of their owners (e.g., OpenSNP [1]). Furthermore, it has been shown that anonymization is not sufficient for protecting the real identities of the genome donors [29, 47]. Once the owner of a genome is identified, he is faced with the risk of discrimination (e.g., by employers or insurance companies) [8]. Some believe that they have nothing to hide about their genetic structure, hence they might decide to give full consent for the publication of their genomes on the Internet to help genomic research. However, our DNA sequences are highly correlated to our relatives' sequences. The DNA sequences between two random human beings are 99.9% similar, and this value is even higher for closely related people. Consequently, somebody revealing his genome does not only damage his own genomic privacy, but also puts his relatives' privacy at risk [46]. Moreover, currently, a person does not need consent from his relatives to share his genome online. This is precisely where the interesting part of the story begins: *kin genomic privacy*.

A recent New York Times' article [3] reports the controversy about sequencing and publishing, without the permission of her family, the genome of Henrietta Lacks (who died in 1951). On the one hand, the family members think that her genome is private family information and it should not be published without the consent of the family. On the other hand, some scientists argued that the genomes of current family members have changed so much over time (due to gene mixing during reproduction), that nothing accurate could be told about the genomes of current family members by using Henrietta Lacks' genome. As we will also show in this work, they are wrong. Minutes after Henrietta Lacks' genome was uploaded to a public website called SNPedia, researchers produced a report full of personal information

1141

Source: Humbert et *al.*, CCS, Nov 2013
doi 10.1145/2508859.2516707

In each pair of nucleotides in your genome, one nucleotide/allele is inherited from your mother and the other from your father.



# Kin genomic privacy

Addressing the Concerns of the Lacks' Family: Quantification of Kin Genomic Privacy

Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux
Laboratory for Communications and Applications
EPFL, Lausanne, Switzerland
firstname.lastname@epfl.ch

Amalio Telenti
Institute of Microbiology
University Hospital of Lausanne
Lausanne, Switzerland
amalio.telenti@chuv.ch

ABSTRACT

The rapid progress in human-genome sequencing is leading to a high availability of genomic data. This data is notoriously very sensitive and stable in time. It is also highly correlated among relatives. A growing number of genomes are becoming accessible online (e.g., because of leakage, or after their posting on genome-sharing websites). What are then the implications for kin genomic privacy? We formalize the problem and detail an efficient reconstruction attack based on graphical models and belief propagation. With this approach, an attacker can infer the genomes of the relatives of an individual whose genome is observed, relying notably on Mendel's Laws and statistical relationships between the nucleotides (on the DNA sequence). Then, to quantify the level of genomic privacy as a result of the proposed inference attack, we discuss possible definitions of *genomic privacy* metrics. Genomic data reveals Mendelian diseases and the likelihood of developing degenerative diseases such as Alzheimer's. We also introduce the quantification of *health privacy*, specifically the measure of how well the predisposition to a disease is concealed from an attacker. We evaluate our approach on actual genomic data from a pedigree and show the threat extent by combining data gathered from a genome-sharing website and from an online social network.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection*; J.3 [**Life and Medical Sciences**]: *Biology and genetics*; K.4.1 [**Computer and Society**]: Public Policy Issues—*Privacy*

Keywords

Genomic Privacy; Inference Algorithms; Metrics; Kinship

"The family of Henrietta Lacks (August 1, 1920 - October 4, 1951), whose DNA was sequenced and published online without the consent of her family.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CCS'13, November 4–8, 2013, Berlin, Germany.
Copyright 2013 ACM 978-1-4503-2477-9/13/11 ...$15.00.
http://dx.doi.org/10.1145/2508859.2516707

1. INTRODUCTION

With the help of rapidly developing technology, DNA sequencing is becoming less expensive. As a consequence, the research in genomics has gained speed in paving the way to personalized (genomic) medicine, and geneticists need large collections of human genomes to further increase this speed. Furthermore, individuals are using their genomes to learn about their (genetic) predispositions to diseases, their ancestries, and even their (genetic) compatibilities with potential partners. This trend has also caused the launch of health-related websites and online social networks (OSNs), in which individuals share their genomic data (e.g., OpenSNP [1] or 23andMe [2]). To date, already today, thousands of genomes are available online.

Even though most of the genomes on the Internet are anonymized, it is possible to find genomes with the identifiers of their owners (e.g., OpenSNP [1]). Furthermore, it has been shown that anonymization is not sufficient for protecting the real identities of the genome donors [29, 47]. Once the owner of a genome is identified, he is faced with the risk of discrimination (e.g., by employers or insurance companies) [8]. Some believe that they have nothing to hide about their genetic structure, hence they might decide to give full consent for the publication of their genomes on the Internet to help genomic research. However, our DNA sequences are highly correlated to our relatives' sequences. The DNA sequences between two random human beings are 99.9% similar, and this value is even higher for closely related people. Consequently, somebody revealing his genome does not only damage his own genomic privacy, but also puts his relatives' privacy at risk [46]. Moreover, currently, a person does not need consent from his relatives to share his genome online. This is precisely where the interesting part of the story begins: *kin genomic privacy*.

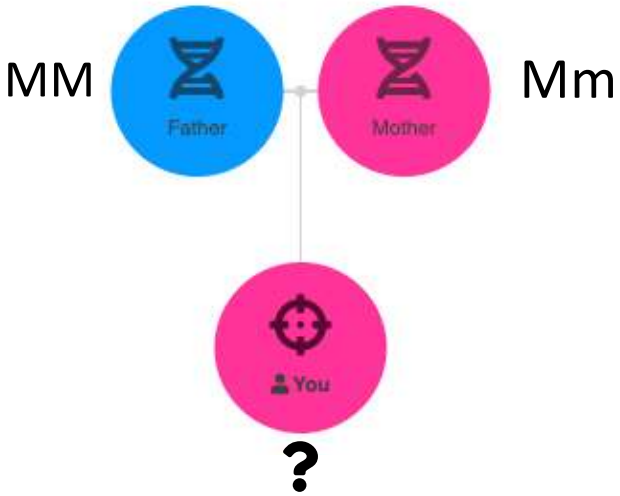
A recent New York Times' article [3] reports the controversy about sequencing and publishing, without the permission of her family, the genome of Henrietta Lacks (who died in 1951). On the one hand, the family members think that her genome is private family information and it should not be published without the consent of the family. On the other hand, some scientists argued that the genomes of current family members have changed so much over time (due to gene mixing during reproduction), that nothing accurate could be told about the genomes of current family members by using Henrietta Lacks' genome. As we will also show in this work, they are wrong. Minutes after Henrietta Lacks' genome was uploaded to a public website called SNPedia, researchers produced a report full of personal information

1141

Source: Humbert et *al.*, CCS, Nov 2013
10.1145/2508859.2516707

In each pair of nucleotides in your genome, one nucleotide/allele is inherited from your mother and the other from your father.

Hence, your genome is directly linked to those of your parents and to those of your children, but also indirectly to those of all your family members.



# Kin genomic privacy

MIT Technology Review

Biotechnology / DNA Testing

## Do Your Family Members Have a Right to Your Genetic Code?

When a woman gets her genome sequenced, questions about privacy arise for her identical twin sister.

by **Emily Mullin** Nov 22, 2016

Source: https://www.technologyreview.com/s/602946/do-your-family-members-have-a-right-to-your-genetic-code/

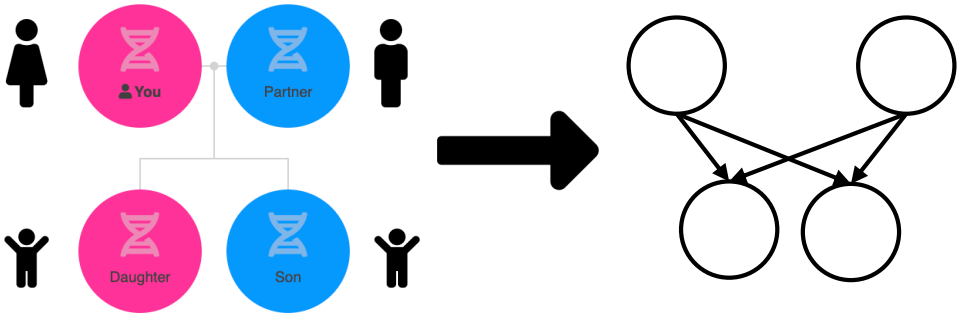
# Kin genomic privacy

- **Concept**:
    1. Model family tree as a knowledge graph (Bayesian network)

# Kin genomic privacy

- **Concept**:
  1. Model family tree as a knowledge graph (Bayesian network)

| | mm | Mm | MM |
|---|---|---|---|
| | $p^2$ | $2p(1-p)$ | $(1-p)^2$ |

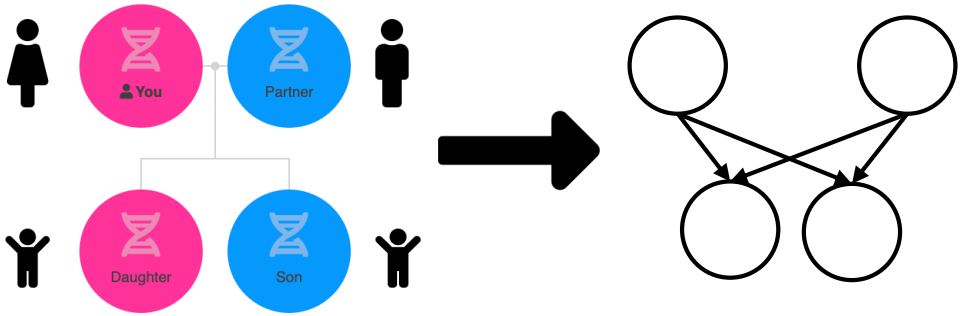



| | | mm | Mm | MM | mm | Mm | ... |
|---|---|---|---|---|---|---|---|
| | | mm | mm | mm | Mm | Mm | |
| mm | | 1 | 0.5 | 0 | 0.5 | 0.25 | |
| Mm | | 0 | 0.5 | 1 | 0.5 | 0.5 | ... |
| MM | | 0 | 0 | 0 | 0 | 0.25 | |

# Kin genomic privacy

- **Concept**:
  1. Model family tree as a knowledge graph (Bayesian network)

| | mm | Mm | MM |
|---|---|---|---|
| | $p^2$ | $2p(1-p)$ | $(1-p)^2$ |



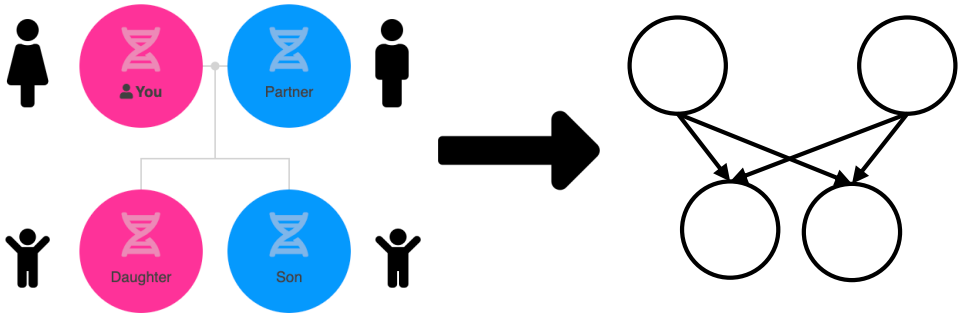

| | | mm | Mm | MM | mm | Mm | ... |
|---|---|---|---|---|---|---|---|
| | | mm | mm | mm | Mm | Mm | |
| mm | | 1 | 0.5 | 0 | 0.5 | 0.25 | |
| Mm | | 0 | 0.5 | 1 | 0.5 | 0.5 | ... |
| MM | | 0 | 0 | 0 | 0 | 0.25 | |

  2. Input the data of the sequenced relatives

# Kin genomic privacy

- **Concept**:
  1. Model family tree as a knowledge graph (Bayesian network)

| | mm | Mm | MM |
|---|---|---|---|
| | $p^2$ | $2p(1-p)$ | $(1-p)^2$ |



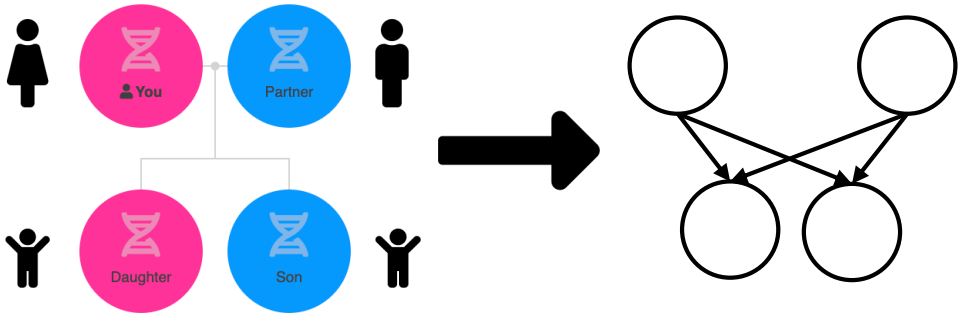

| | | mm | Mm | MM | mm | Mm | ... |
|---|---|---|---|---|---|---|---|
| | | mm | mm | mm | Mm | Mm | |
| mm | | 1 | 0.5 | 0 | 0.5 | 0.25 | |
| Mm | | 0 | 0.5 | 1 | 0.5 | 0.5 | ... |
| MM | | 0 | 0 | 0 | 0 | 0.25 | |

  2. Input the data of the sequenced relatives
  3. Perform inference using the graph

# Kin genomic privacy

- **Concept**:
  1. Model family tree as a knowledge graph (Bayesian network)

| | mm | Mm | MM |
|---|---|---|---|
| | $p^2$ | $2p(1-p)$ | $(1-p)^2$ |




| | | mm | Mm | MM | mm | Mm | ... |
|---|---|---|---|---|---|---|---|
| | | mm | mm | mm | Mm | Mm | |
| mm | | 1 | 0.5 | 0 | 0.5 | 0.25 | |
| Mm | | 0 | 0.5 | 1 | 0.5 | 0.5 | ... |
| MM | | 0 | 0 | 0 | 0 | 0.25 | |

2. Input the data of the sequenced relatives
3. Perform inference using the graph
4. Quantify privacy according to inference error

# ◎ Objective

🛠 Enable everyone to evaluate their kin genomic privacy in a simple and interactive way for real and hypothetical scenarios

# Concept

- **For each SNP (and associated MAF)** :
  1. Consider all possible configurations of SNP values (mm, Mm, MM) for each sequenced individual
  2. Compute the posterior distribution for the target
  3. Quantify privacy according to the remaining uncertainty

# Concept

- **For each SNP (and associated MAF)** :
  1. Consider <u>all possible configurations</u> of SNP values (mm, Mm, MM) for <u>each</u> sequenced individual
  2. Compute the <u>posterior distribution</u> for the target
  3. <u>Quantify privacy</u> according to <u>the remaining uncertainty</u>

$$\text{privacy} = 1 - \frac{H(\text{target}) - H(\text{target} \mid \text{father}, \text{mother}, \cdots)}{H(\text{target})}$$

uncertainty on the target's genome when no information is available

# Concept

- **For each SNP (and associated MAF)** :
  1. Consider <u>all possible configurations</u> of SNP values (mm, Mm, MM) for <u>each</u> sequenced individual
  2. Compute the <u>posterior distribution</u> for the target
  3. <u>Quantify privacy</u> according to <u>the remaining uncertainty</u>

uncertainty on the target's genome when the genomes of others are available

$$\text{privacy} = 1 - \frac{H(\text{target}) - H(\text{target} \mid \text{father}, \text{mother}, \cdots)}{H(\text{target})}$$

uncertainty on the target's genome when no information is available

# Concept

- **For each SNP (and associated MAF)** :
  1. Consider <u>all possible configurations</u> of SNP values (mm, Mm, MM) for <u>each</u> sequenced individual
  2. Compute the <u>posterior distribution</u> for the target
  3. <u>Quantify privacy</u> according to <u>the remaining uncertainty</u>

uncertainty on the target's genome when the genomes of others are available

$$\text{privacy} = 1 - \frac{H(\text{target}) - H(\text{target} \mid \text{father}, \text{mother}, \cdots)}{H(\text{target})}$$

uncertainty on the target's genome when no information is available

averaged over all configurations and all SNPs

# Concept

- **For each SNP (and associated MAF)** :
  1. Consider <u>all possible configurations</u> of SNP values (mm, Mm, MM) for <u>each</u> sequenced individual
  2. Compute the <u>posterior distribution</u> for the target
  3. <u>Quantify privacy according to the remaining uncertainty</u>

**Contribution #1**: Quantify without data by considering all configurations

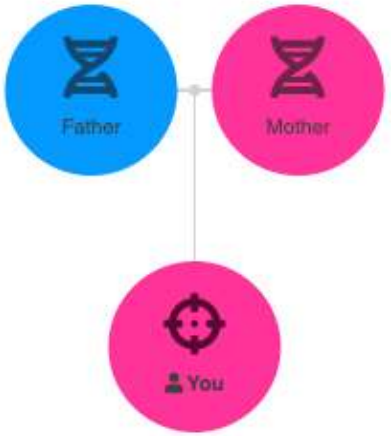
$$\text{privacy} = 1 - \frac{H(\text{target}) - H(\text{target} \mid \text{father}\ \text{mother}\ \cdots)}{H(\text{target})}$$

uncertainty on the target's genome when no information is available

averaged over all configurations and all SNPs

# Concept: Example

- **For SNP rs753426**
  - MAF(rs753426) = 0.1
  - $H(\Xi_{You}) = 0.7580$ [prior]




| $\Xi_{Father}$ | $\Xi_{Mother}$ | $P(\Xi_{Father}, \Xi_{Mother})$ | $P(\Xi_{Target} \mid \Xi_{Father}, \Xi_{Mother})$ | | | $H(\Xi_{Target} \mid \Xi_{Father}, \Xi_{Mother})$ |
|---|---|---|---|---|---|---|
| | | | MM | Mm | mm | |
| MM | MM | **0.6561** | 1.0 | 0.0 | 0.0 | **0.0** |
| MM | Mm | **0.1458** | 0.5 | 0.5 | 0.0 | **1.0** |
| … | | | | | | |

# Concept: Complexity & Execution time

$3^N$ configurations, ≈500k SNPs

$N$: number of sequenced relatives in the family tree

# Concept: Complexity & Execution time

**Contribution #2**: Make quantification "scalable" with optimizations

$3^N$ configurations, ≈500k SNPs

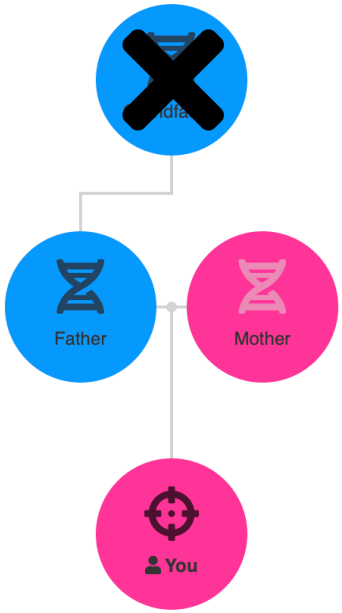
$N$: number of sequenced relatives in the family tree

# Optimization: Simplification

- Remove individuals whose genomes do not matter when inferring the target's genome (knowing those of the others)

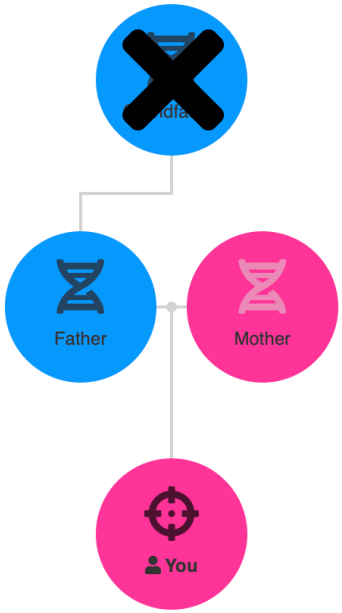
# Optimization: Simplification

- Remove individuals whose genomes do not matter when inferring the target's genome (knowing those of the others)



# Optimization: Simplification

Remove individuals whose genomes do not matter when inferring the target's genome (knowing those of the others)




$3^N \rightarrow 3^n$ configurations

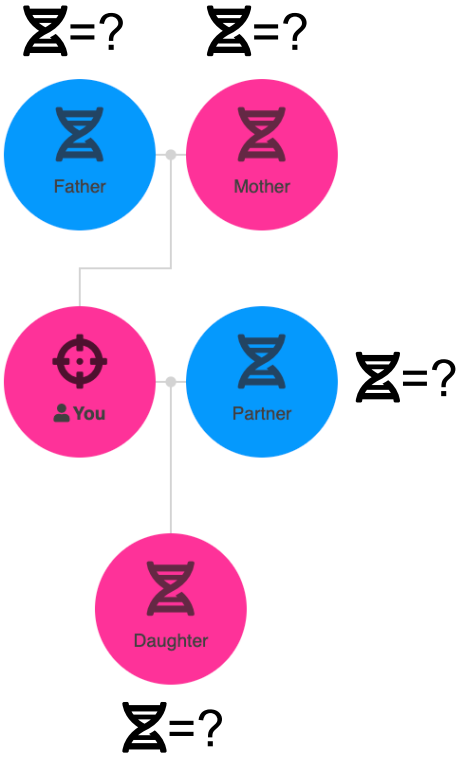
$n$: number of sequenced relatives in the *simplified* family tree

# ⚙ Optimization: Pruning

✋ Detect impossible configurations early and stop exploring

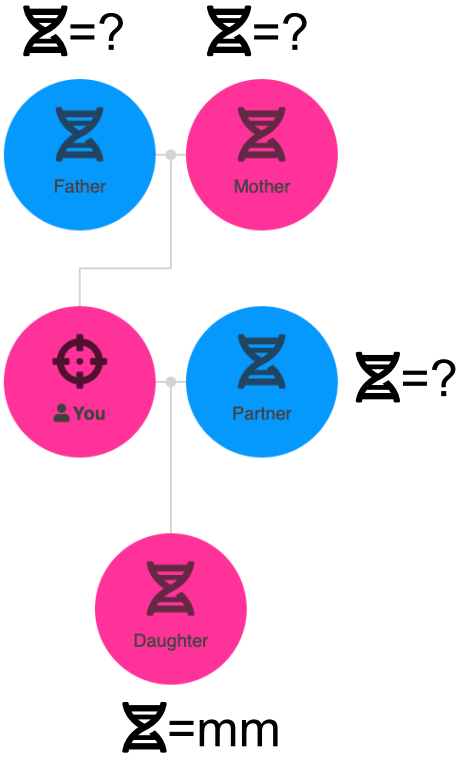
# ⚙ Optimization: Pruning

✋ Detect impossible configurations early and stop exploring



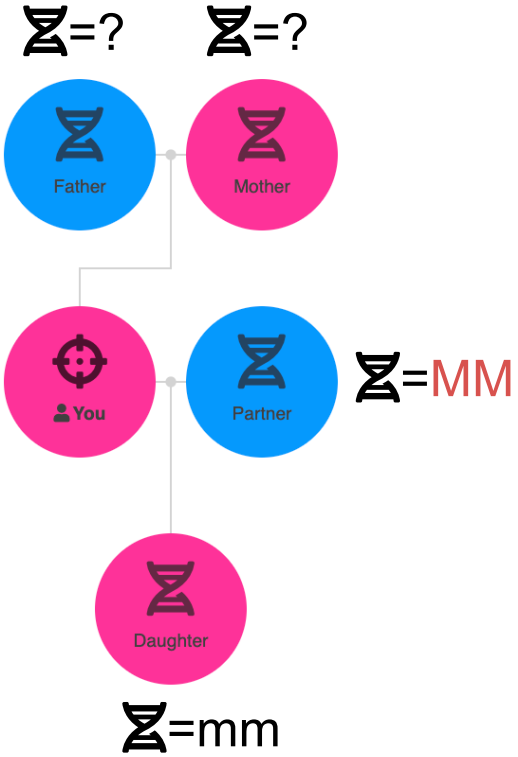
# ⚙ Optimization: Pruning

✋ Detect impossible configurations early and stop exploring



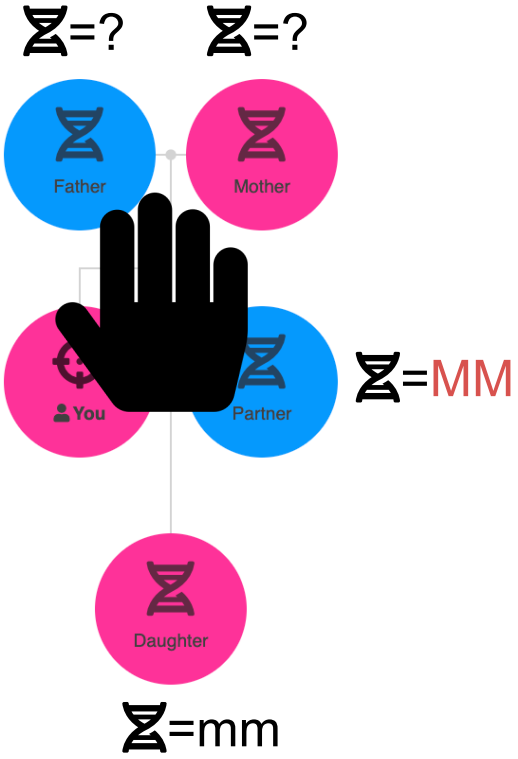
# ⚙ Optimization: Pruning

✋ Detect impossible configurations early and stop exploring



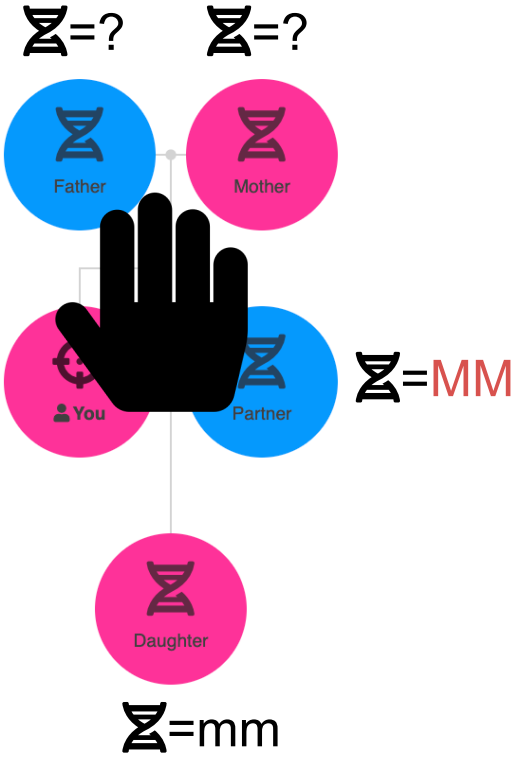
# ⚙ Optimization: Pruning

✋ Detect impossible configurations early and stop exploring



# ⚙ Optimization: Pruning

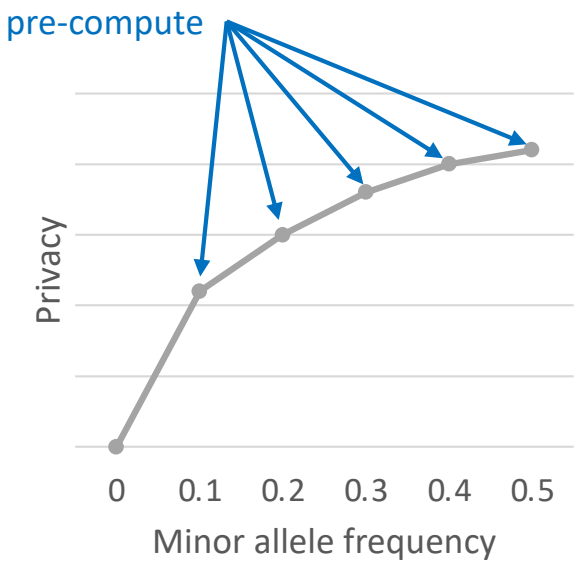
✋ Detect impossible configurations early and stop exploring




$\leq 3^n$ configurations

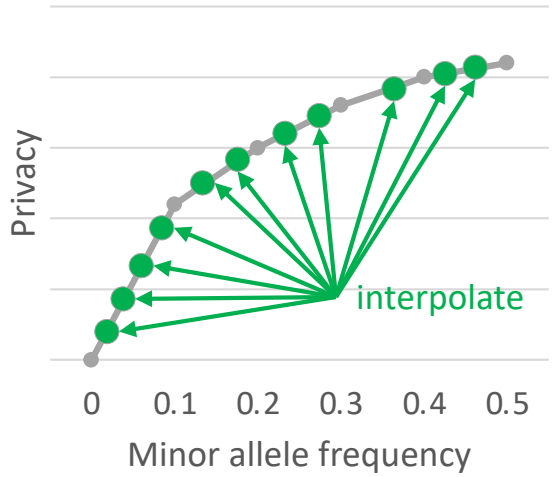
# Optimization: Interpolation

Compute privacy for a few minor allele frequency values and interpolate for the other values (for each SNP)



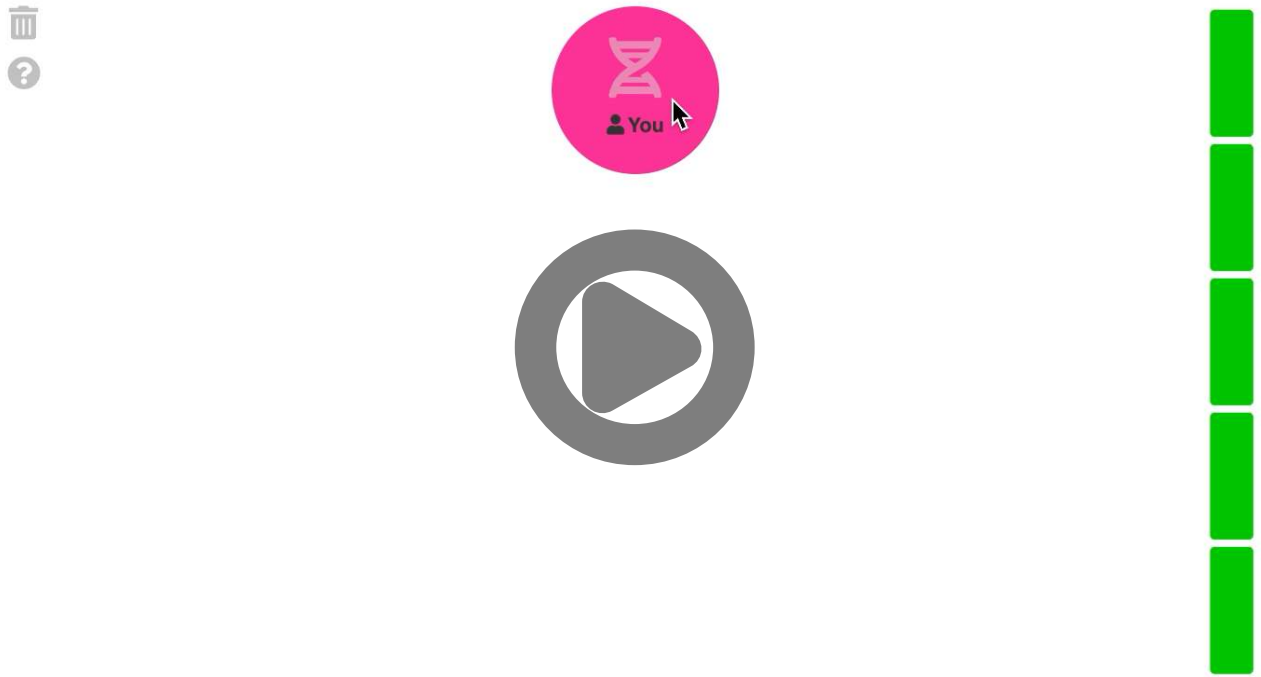
# Optimization: Interpolation

Compute privacy for a few minor allele frequency values and interpolate for the other values (for each SNP)

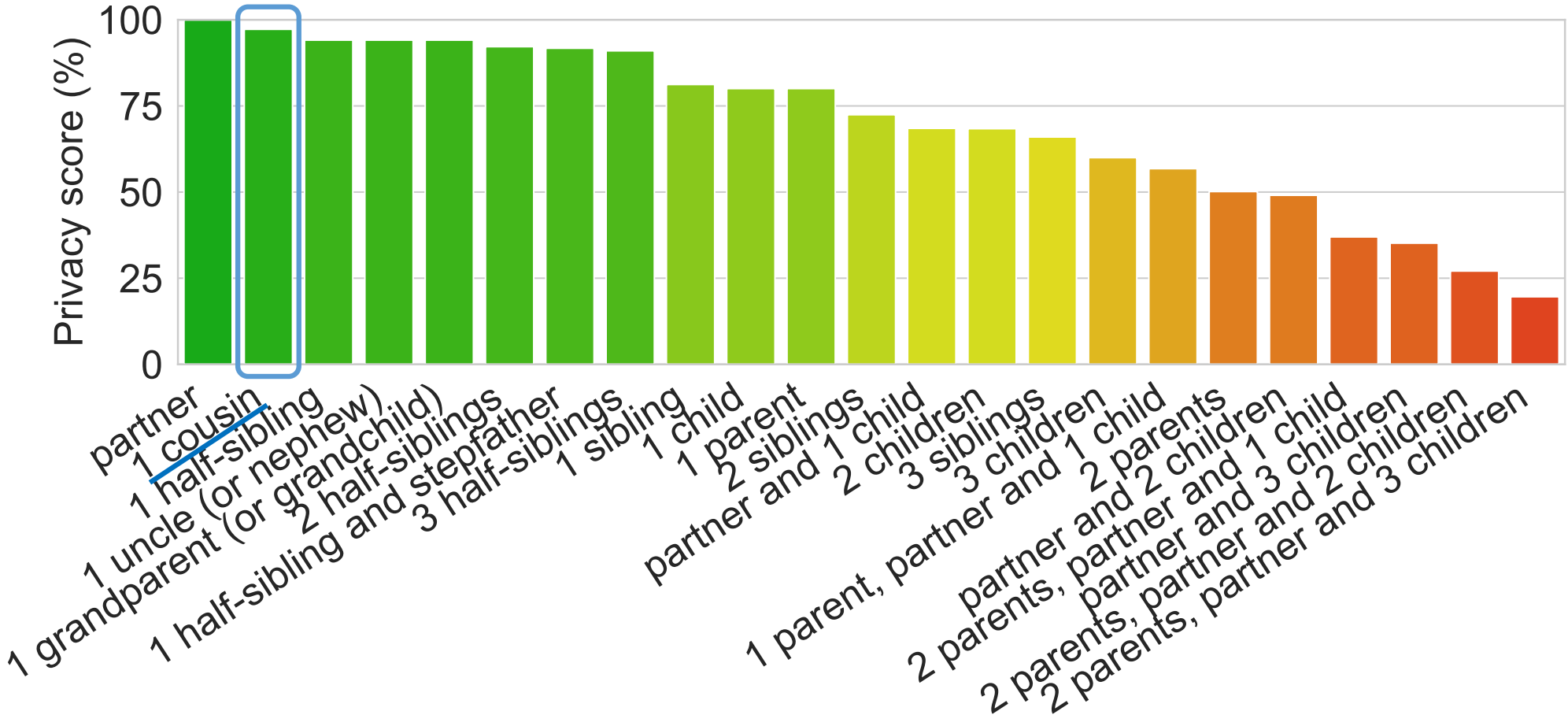


# The tool

https://santeperso.unil.ch/privacy

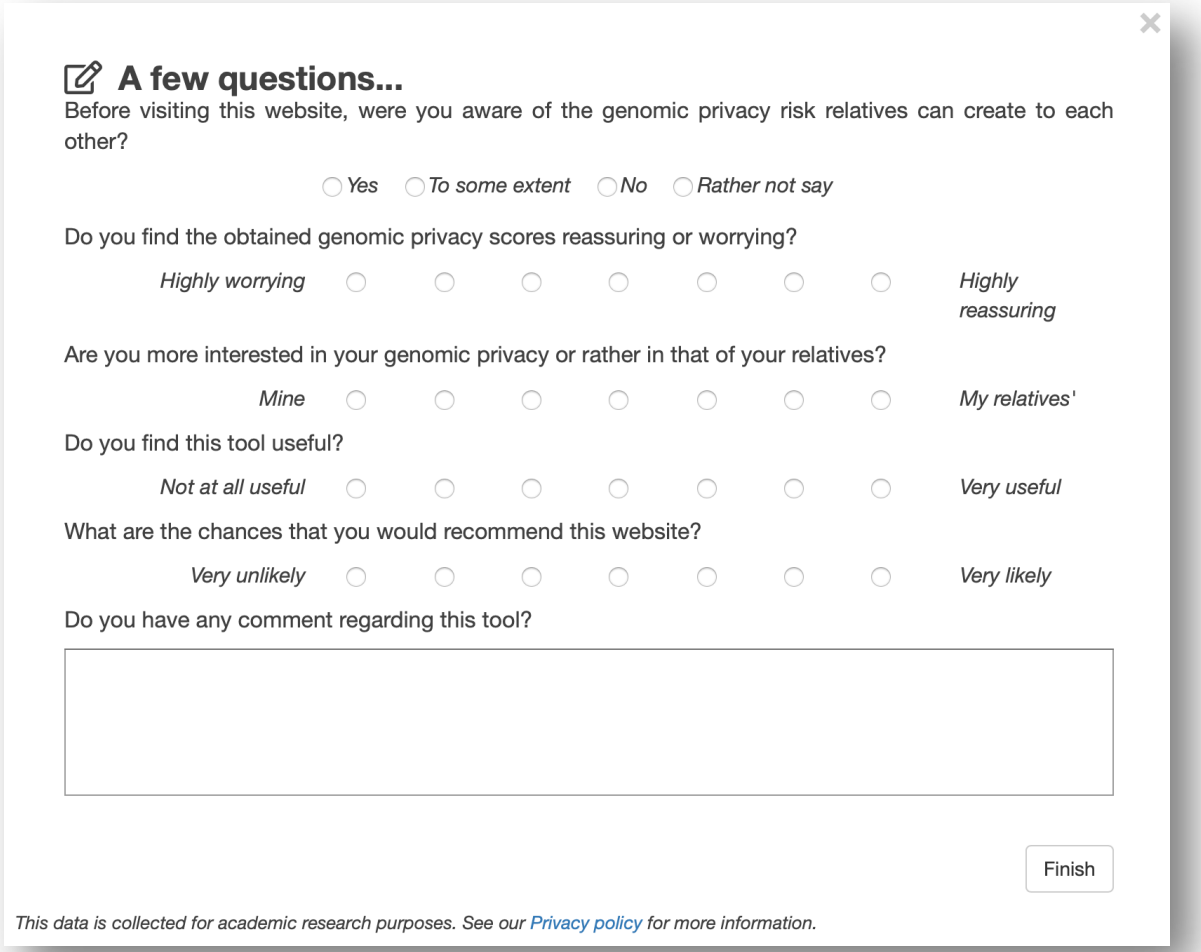


# The tool


Privacy score (%)
100
75
50
25
0
partner
1 cousin
1 half-sibling
1 uncle (or nephew)
1 grandparent (or grandchild)
2 half-siblings
1 half-sibling and stepfather
3 half-siblings
1 sibling
1 child
1 parent
2 siblings
partner and 1 child
2 children
3 siblings
3 children
1 parent, partner and 1 child
2 parents
2 parents, partner and 2 children
partner and 1 child
2 parents, partner and 3 children
partner and 2 children
2 parents, partner and 2 children
2 parents, partner and 3 children

# User study

+ | 413 users, 68 respondents | 2019



**A few questions...**

Before visiting this website, were you aware of the genomic privacy risk relatives can create to each other?

*Yes* *To some extent* *No* *Rather not say*

Do you find the obtained genomic privacy scores reassuring or worrying?

*Highly worrying* — *Highly reassuring*

Are you more interested in your genomic privacy or rather in that of your relatives?

*Mine* — *My relatives'*

Do you find this tool useful?

*Not at all useful* — *Very useful*

What are the chances that you would recommend this website?

*Very unlikely* — *Very likely*

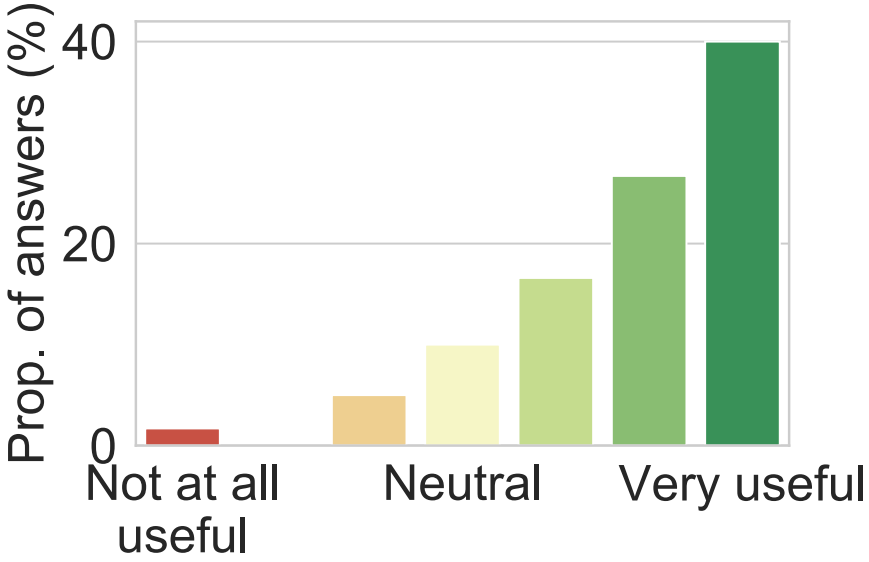
Do you have any comment regarding this tool?

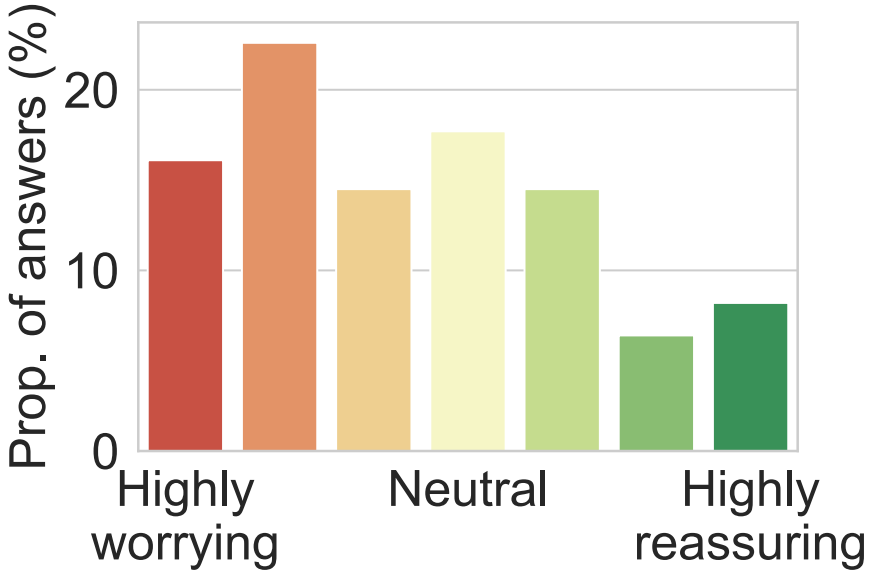
Finish

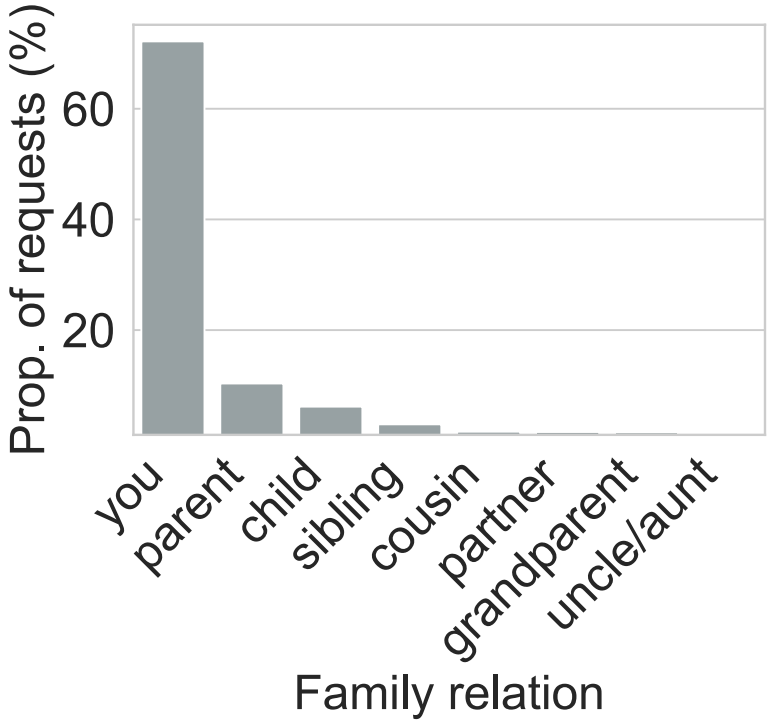
*This data is collected for academic research purposes. See our Privacy policy for more information.*

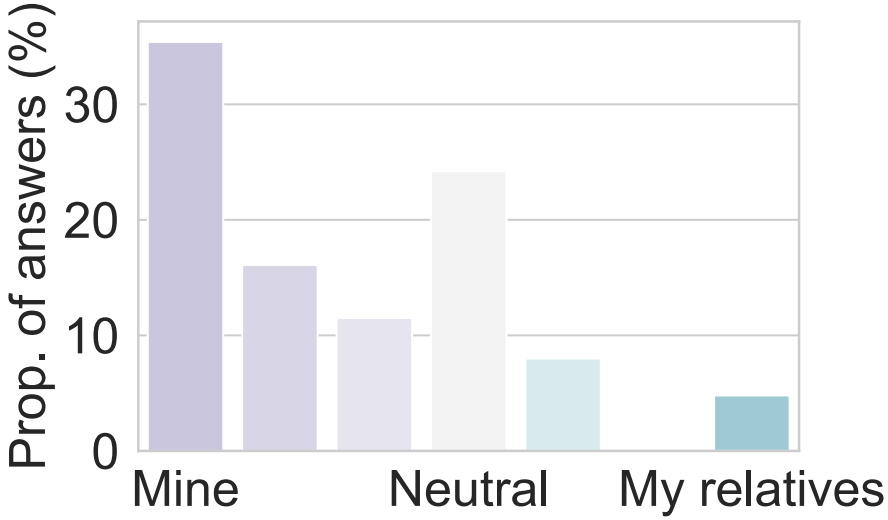
Source: https://santeperso.unil.ch/

# User study

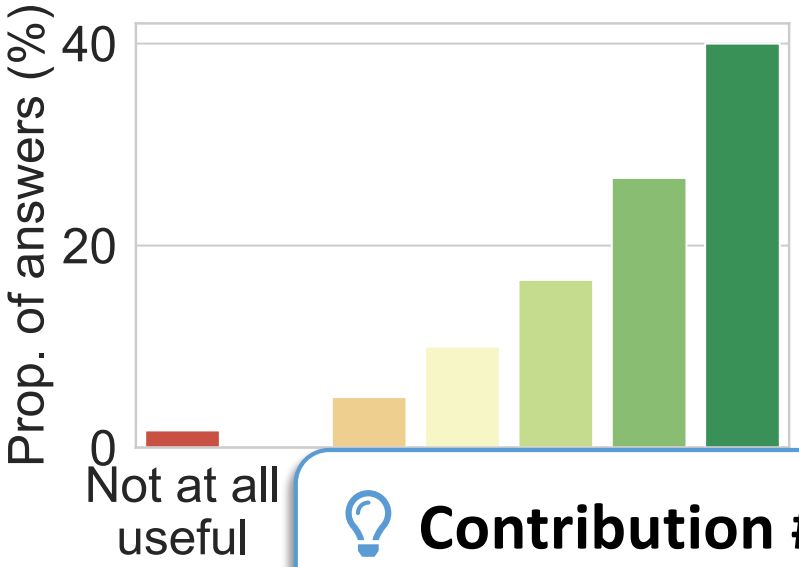

Prop. of answers (%)
40
20
0
Not at all useful
Neutral
Very useful


Prop. of answers (%)
20
10
0
Highly worrying
Neutral
Highly reassuring

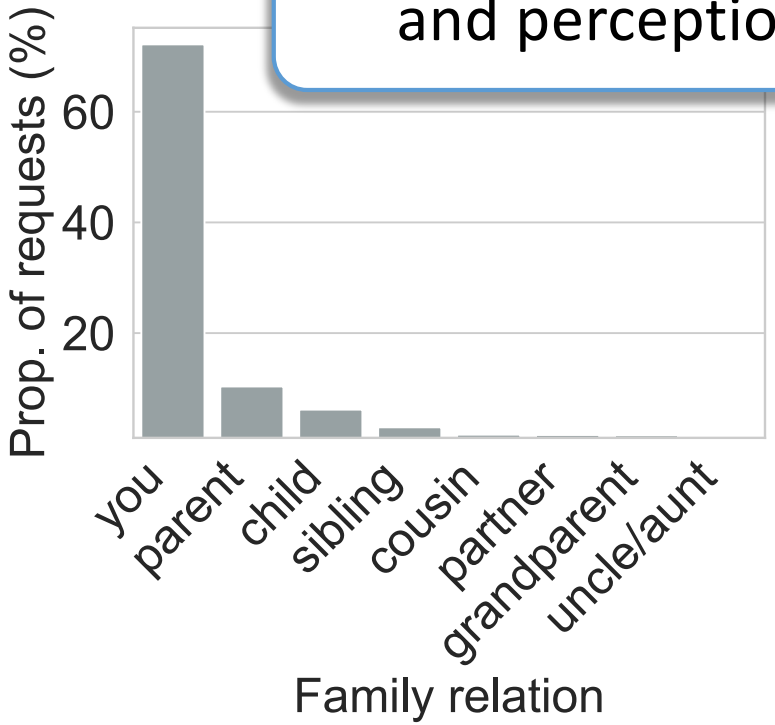

Prop. of requests (%)
60
40
20
you
parent
child
sibling
cousin
partner
grandparent
uncle/aunt
Family relation

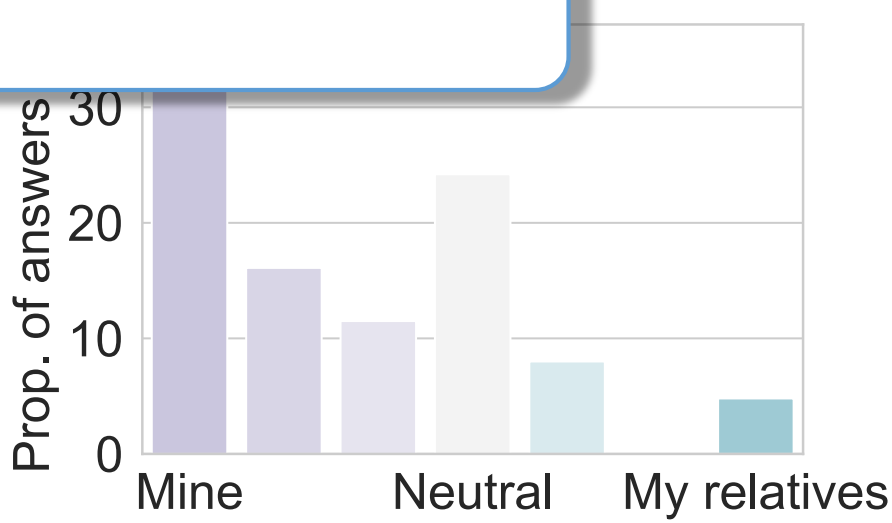

Prop. of answers (%)
30
20
10
0
Mine
Neutral
My relatives

# User study



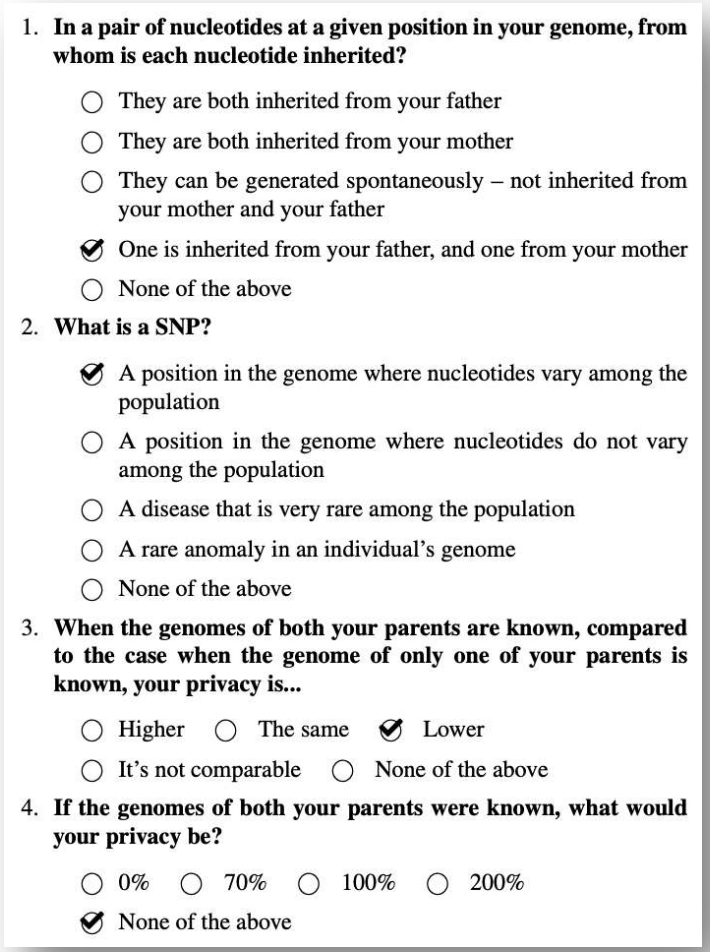
**Contribution #3**: Study users' behaviors and perceptions



# User study

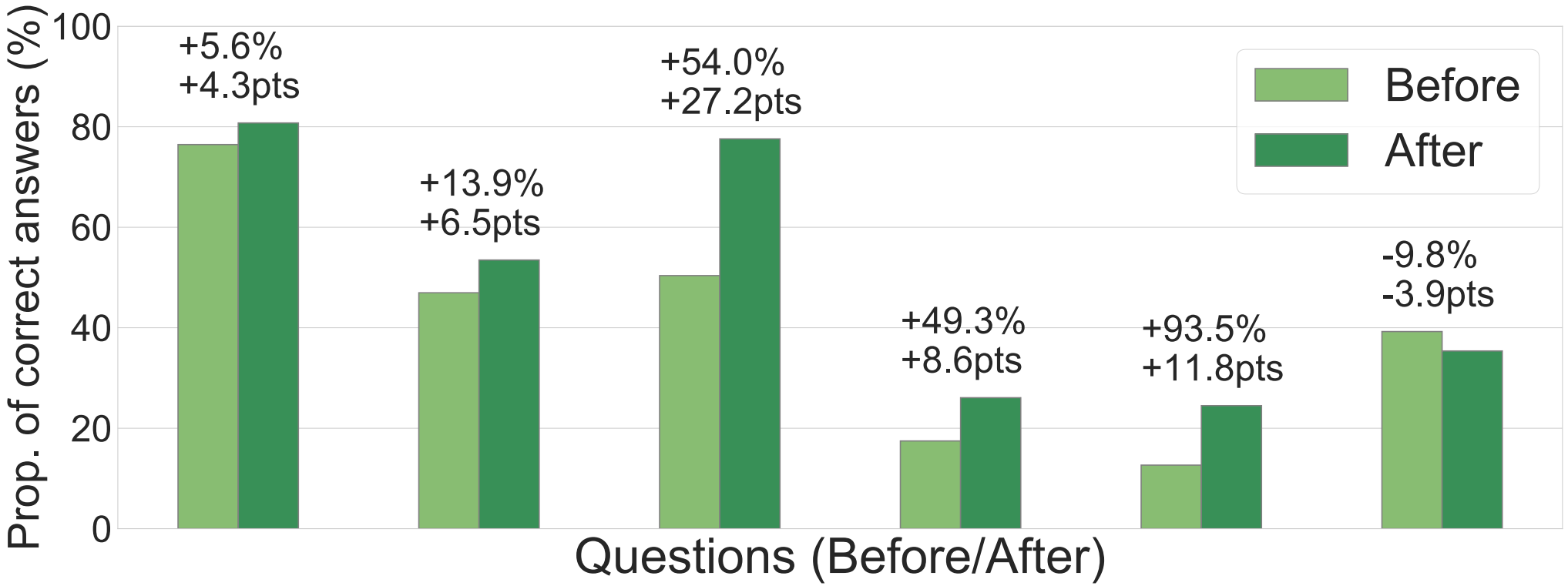
+ | 1822 users/respondents (Prolific) | 2020



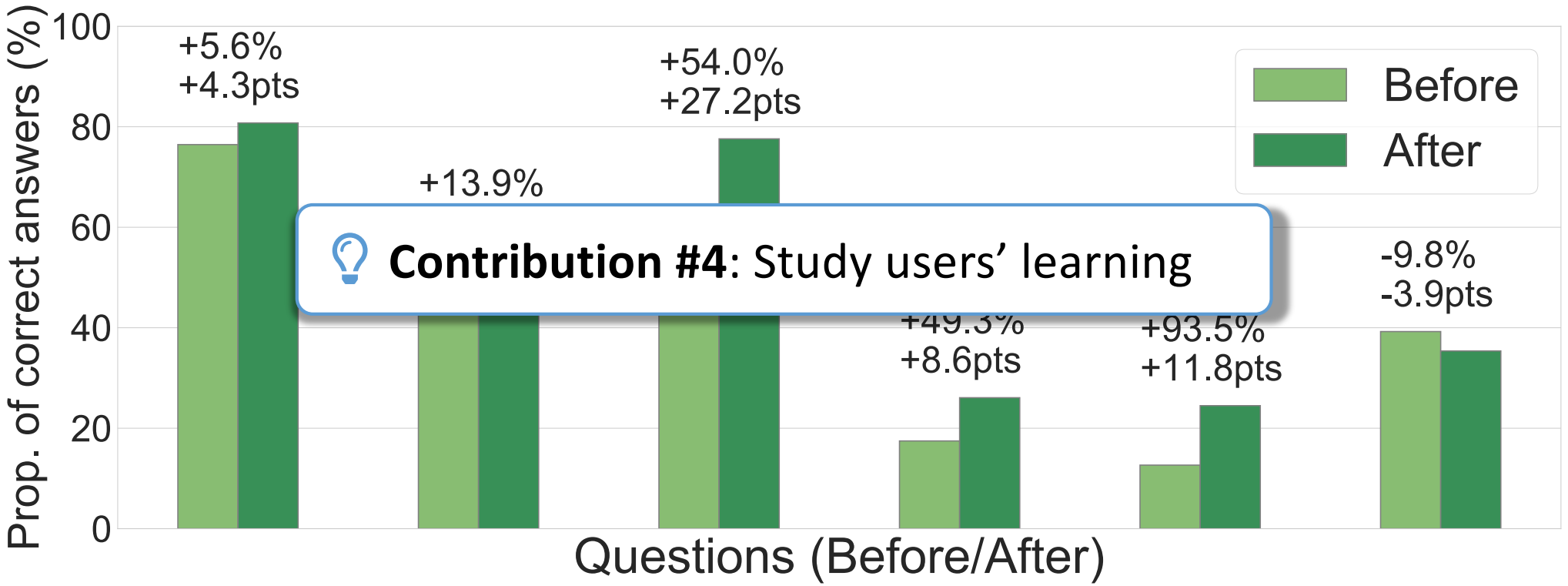
1. **In a pair of nucleotides at a given position in your genome, from whom is each nucleotide inherited?**
   - ○ They are both inherited from your father
   - ○ They are both inherited from your mother
   - ○ They can be generated spontaneously – not inherited from your mother and your father
   - ☑ One is inherited from your father, and one from your mother
   - ○ None of the above
2. **What is a SNP?**
   - ☑ A position in the genome where nucleotides vary among the population
   - ○ A position in the genome where nucleotides do not vary among the population
   - ○ A disease that is very rare among the population
   - ○ A rare anomaly in an individual's genome
   - ○ None of the above
3. **When the genomes of both your parents are known, compared to the case when the genome of only one of your parents is known, your privacy is...**
   - ○ Higher ○ The same ☑ Lower
   - ○ It's not comparable ○ None of the above
4. **If the genomes of both your parents were known, what would your privacy be?**
   - ○ 0% ○ 70% ○ 100% ○ 200%
   - ☑ None of the above

Asked before and after using the tool

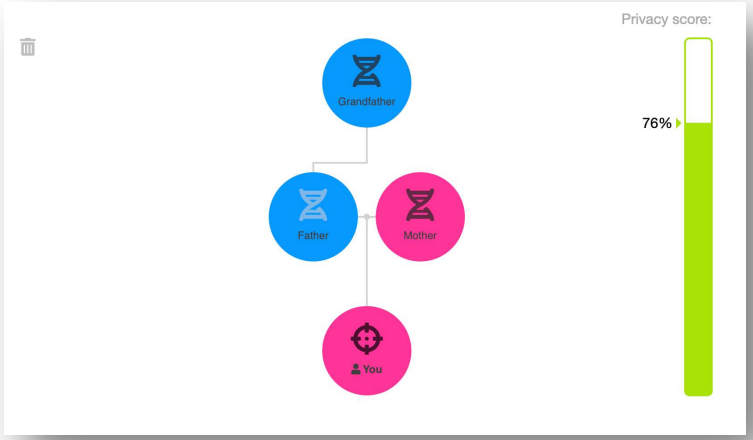
# User study


Prop. of correct answers (%)
100
80
60
40
20
0
+5.6%
+4.3pts
+13.9%
+6.5pts
+54.0%
+27.2pts
+49.3%
+8.6pts
+93.5%
+11.8pts
-9.8%
-3.9pts
Before
After
Questions (Before/After)

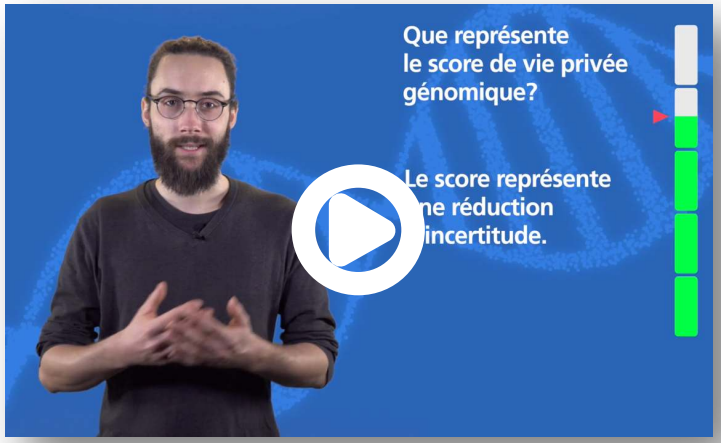
# User study


Prop. of correct answers (%)
100
80
60
40
20
0
+5.6%
+4.3pts
+13.9%
+54.0%
+27.2pts
+49.3%
+8.6pts
+93.5%
+11.8pts
-9.8%
-3.9pts
Before
After
Contribution #4: Study users' learning
Questions (Before/After)

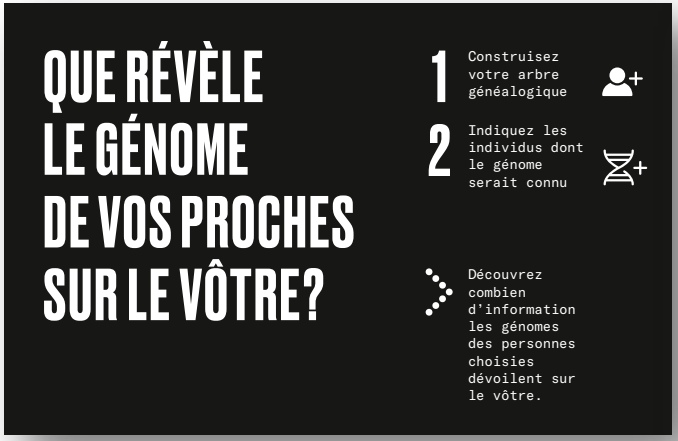
# Dissemination

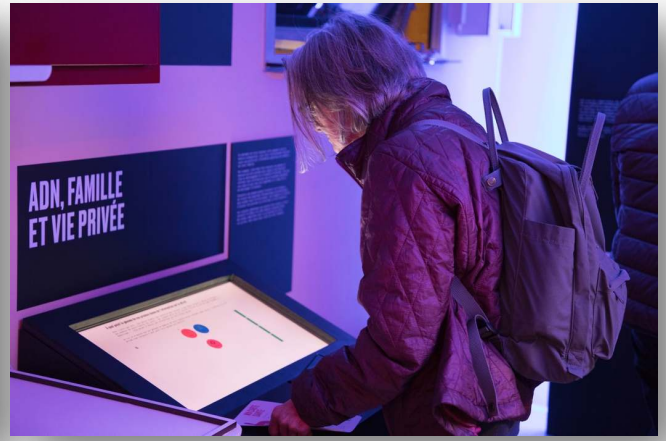



Source: https://santeperso.unil.ch/




Source [FR]: Video produced in collaboration with UNIL's support center for teaching next-generation technologies (RISET)




Source [FR]: Mobile exhibition on personalized medicine https://www.santeperso.ch/Projets/A-notre-sante

Available on GitHub

# Conclusion

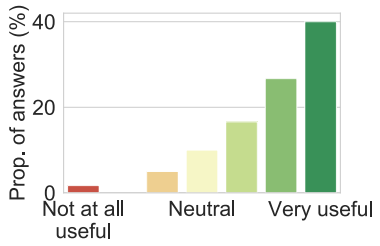
- Tackled an important interdependent privacy issue: kin genomic privacy
- Produced an online tool and software library for quantification (without data)
- Conducted user studies for validation
- Disseminated results and raised awareness

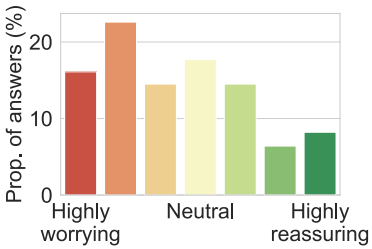
# ⍰ Questions

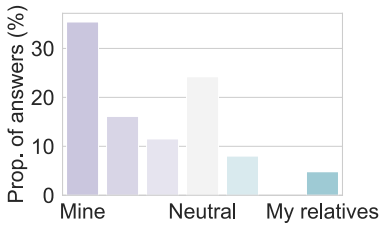
- Tackled an important interdependent privacy issue: kin genomic privacy

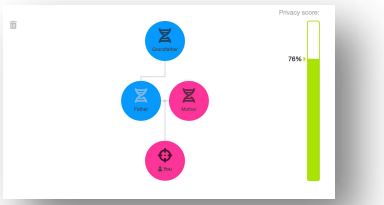
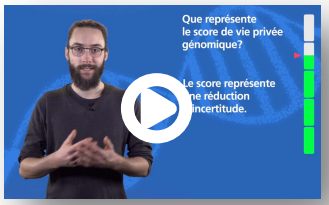
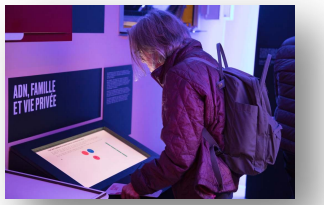
- Produced an online tool and software library for quantification (without data)

- Conducted user studies for validation




- Disseminated results and raised awareness