Biomedical privacy and security

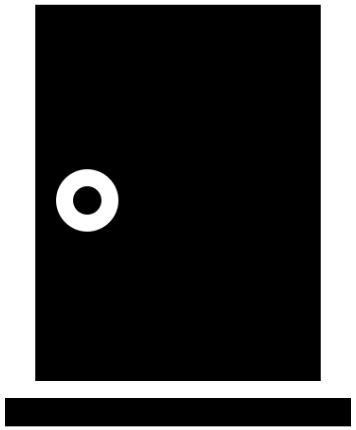
Yun William Yu

02752 – Spring 2026

April 20, 2026

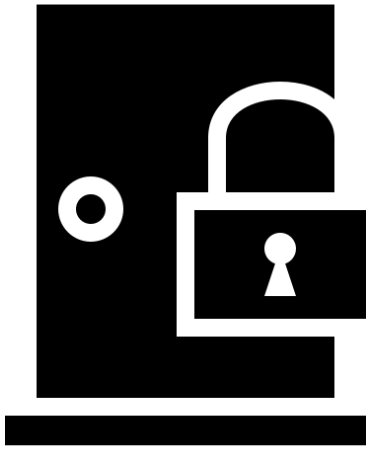
Privacy vs. Security

"Privacy discourse involves difficult normative decisions about competing claims to legitimate access to, use of, and alteration of information."



Clipart: delapouite.com (CC)

"Security implements those choices"



Clipart: delapouite.com (CC)

Derek E. Bambauer, *Privacy Versus Security (*2013). Journal of Criminal Law and Criminology, Vol 103, Issue 3

Banking records

- Privacy question:
  - Who should have access to your credit card bill?
  - Bank? Employer? Insurance company? Advertisers? Your parents?
- Security question:
  - How do we ensure that only people who should access your credit card bill can?
  - Envelopes? Encryption? Legal fines? Bad publicity?

MR JOHN SMITH
1 NERD HOUSE
WALLET ROAD
BOROUGH
NR1 2DS

Your Bank Credit Card

Credit card number	1234 5678 9123 4567
Cardholder	Mr John Smith
Your credit limit	£XXXX
Available to spend	£XXXX
Next month's estimated interest	£X.XX

Your credit card statement
13 January 20XX

Remember, when you purchase items with a credit card, you are protected under Section 75 of the Consumer Credit Act. This means that any purchase between £100 and £30,000 made with a credit card may be refunded by your credit card company if you were to find fault with the item or the seller.

Summary of your account

Previous balance	£XXX
Payments received	£XXX
New transactions, fees and charges	£XXX
Your new balance	**£XXX**
Minimum payment due to reach your account by	**£X.XX 20/01/20XX**

Date of transaction	Merchant name	Description	£Amount
25/12/20XX	XXXX XXXX	XXXX XXXX XXXX	£XX.XX
29/12/20XX	XXXX XXXX	XXXX XXXX XXXX	£X.XX
07/01/20XX	XXXX XXXX	XXXX XXXX XXXX	£XXX
NEW BALANCE			**£XXX**

Breakdown of balance

Balance Type	Simple Annual Rate (%)	Effective Annual Rate (%)	Outstanding Balance (£)	Interest (£)
XXXX XXXX	XX.XX	XX.XX	XXX	XX.XX

* This template is for illustrative purposes only, and will not reflect how every credit card statement will look.

https://www.nerdwallet.com/uk/credit-cards/credit-card-statement/

Privacy tensions: data is power

Who here thinks that records of things you own should not be made public?

TECHNOLOGY

THE WALL STREET JOURNAL.

Tech Leaders Emerge Behind Plan to Build New City Near California Air Base

Group has spent nearly $1 billion to buy thousands of acres northeast of San Francisco

THE ROANOKE TIMES

roanoke.com

Tuesday, March 13, 2007

Current Weather in Roanoke

27°F

Home | News | Sports | Entertainment | Webcast | Classified | Opinion | Blogs | Photos | Multimedia | Message Board

The Roanoke Times Removes Database of Handgun Permit Holders

March 12, 2007 — The Roanoke Times has decided to remove the online database of registered concealed handgun permit holders from its website.

The newspaper is requesting the Virginia State Police, which provided the information, verify the data.

"When we posted the information, we had every reason to believe that the data the State Police had supplied would comply with the statutes. But people have notified us that the list includes names that should not have been released," said Debbie Meade, president and publisher of The Roanoke Times. "Out of a sense of caution and concern for the public we have decided to take the database off of our website."

The database was posted on roanoke.com on Sunday as part of a New River Valley editorial page column about open records. This column, as well as others that will be published this week, is part of a special focus on Sunshine Week, a national initiative to raise awareness about open government and freedom of information.

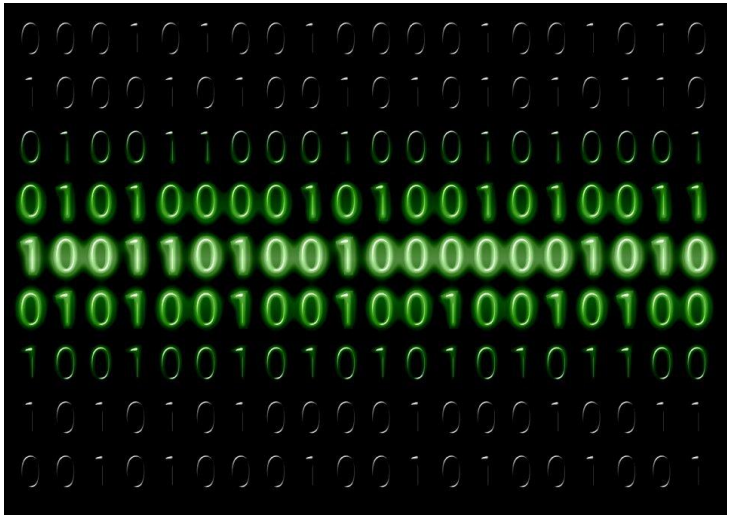
Kelsey M. Swanson. *The Right to Know: An Approach to Gun Licenses and Public Access to Government Records*. https://www.uclalawreview.org/pdf/56-5-14.pdf

Rights to data and privacy

- Many parties may have an interest in accessing data, or in preventing data from being accessed.
- Each society decides as a question of shared values who should have access to data.
- NOT a technical question.

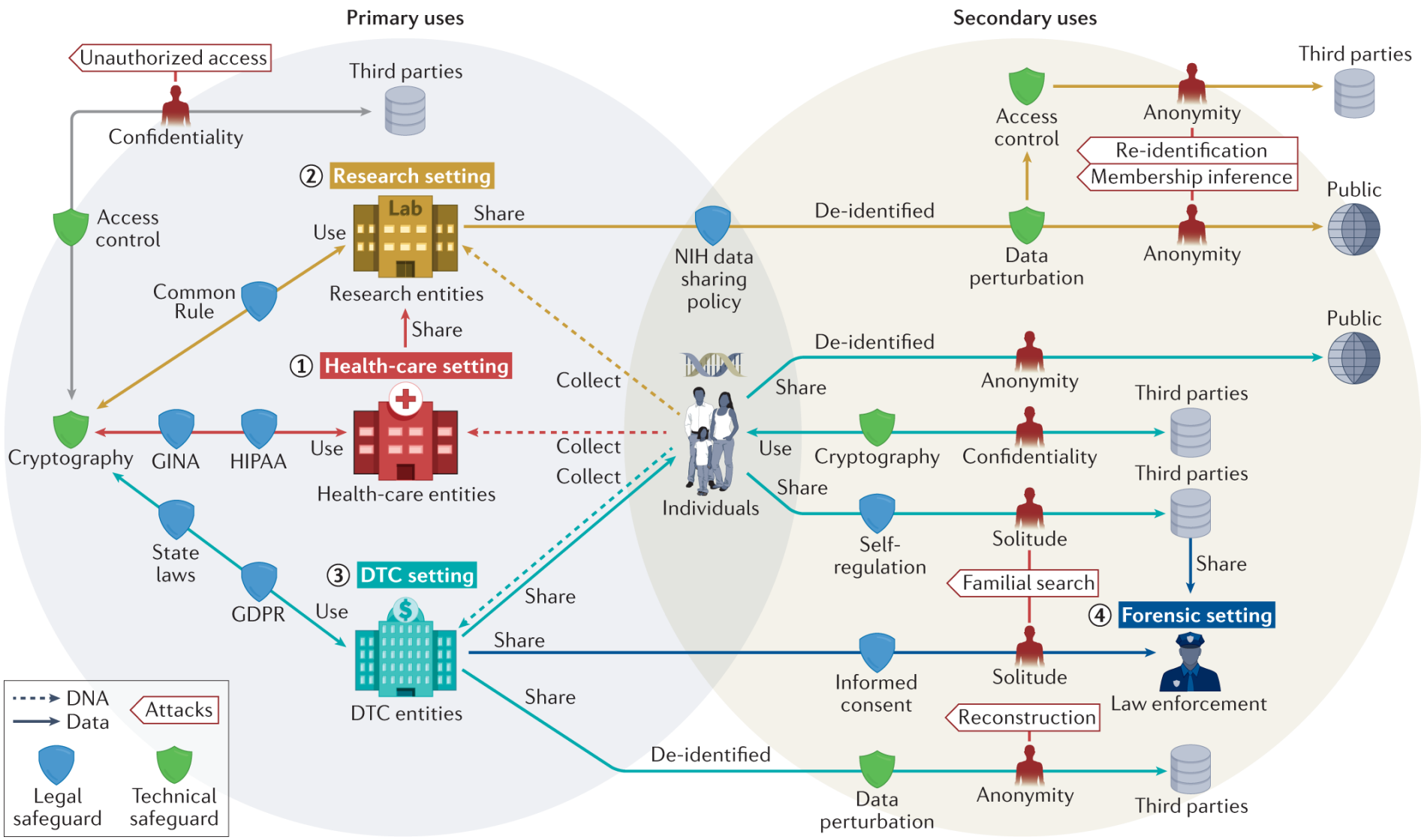
Security

- Legal controls
  - Data use agreements
  - Laws (HIPAA, GDPR, GINA)
  - Informed consent
  - Institutional Review Boards
- Technical
  - Access control lists
  - Cryptography
  - Deidentification
  - Synthetic data



Types of important data for us

- Personally identifiable information (PII)
  - SSN, name, address, phone number, etc.
  - Anything that can identify a specific individual.
- Medical records (or other consumer/behavioral records)
  - A clinical diagnosis, like hypertension, or COVID.
- Individual-level genomes or genotype information



Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nature Reviews Genetics. 2022 Jul;23(7):429-45.

Data = power. For good?

- Genome-wide association studies (GWAS) have allowed us to find genetic causes of disease.
- Tracing ancestry is only possible using genomes if other people share their genomes.
- Tracking criminals by DNA testing may create a safer society. Alternately, identifying the remains of victims of massacres.
- Knowing that a friend has COVID may allow you to prevent the spread of disease.

Danger: re-identification

Identifying Participants in the Personal Genome Project by Name

Latanya Sweeney, Akua Abu, Julia Winn

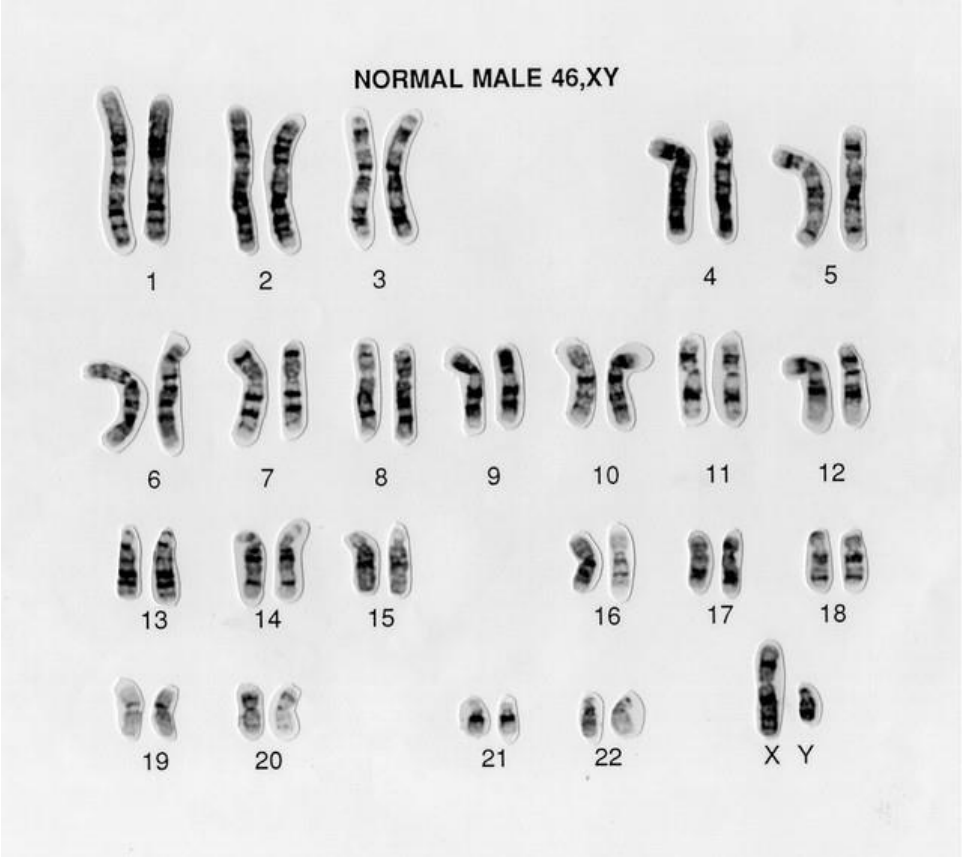
Sweeney L, Abu A, Winn J. Identifying participants in the personal genome project by name (a re-identification experiment). arXiv preprint arXiv:1304.7605. 2013 Apr 29.

"By linking demographics to public records such as voter lists, and mining for names hidden in attached documents, [...] identified 84 to 97 percent of the profiles for which we provided names."

Separate pieces of information can be linked together to reidentify people:
"87% of the U.S. population is uniquely identified by data of birth, gender, and postal code".

Genome background

Human Genome / Chromosomes



https://wellcomecollection.org/works/zd7rdetc/images?id=ysp6z895

Reference Genome

- Original Human Genome Project reference a chimera of volunteers from Buffalo in 1997.
- Also competed with private company Celera, helmed by Craig Ventner.
- Published in 2000.
- At each location reflects sequence of a single donor.
  - RP-11 matches about 70% of that reference.
- Not perfect: 150,000 gaps in reference.

antithesis of the violent militia movement. Such non-violence can serve as an antidote to government oppression, he added.

"If a law is unjust or you're given an order without moral or legal authority,

by Samuel Radford III, a critic of public education who was arrested and pleaded guilty to reduced charges following a 1993 disturbance at the City Campus of Erie Community College.

nt

ed by

PS

ART

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

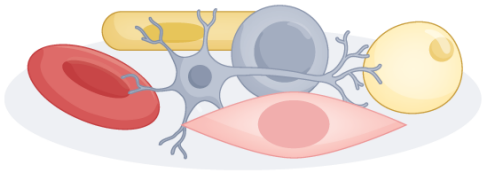
Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

ist be at least 18 years of age.
ergone chemotherapy are not eligible.

For more information please contact the
Clinical Genetics Service
845-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Genomic analysis pipelines

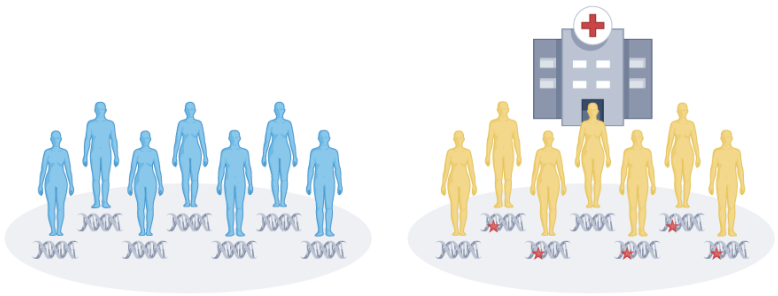
a Biological data: sources of genomic data



b Primary data generation: sequencing of DNA or RNA

c Secondary processing: computational methodology to reconstruct genomic facts about the biological sample

d Tertiary processing: statistics to answer biological questions

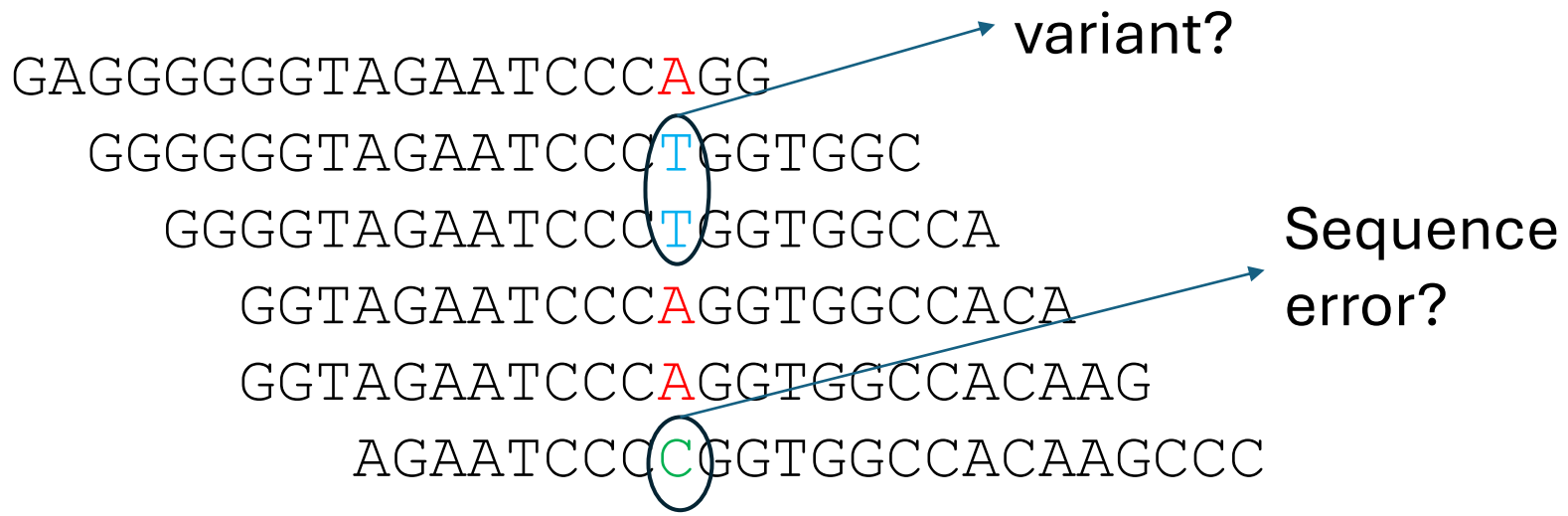


GWAS: identify genetic variants associated with traits such as diseases

Berger B, Yu YW. Navigating bottlenecks and trade-offs in genomic data analysis. Nat Rev Genet. 2023 Apr;24(4):235-250. doi: 10.1038/s41576-022-00551-z. Epub 2022 Dec 7. PMID: 36476810; PMCID: PMC10204111.

Variant calling

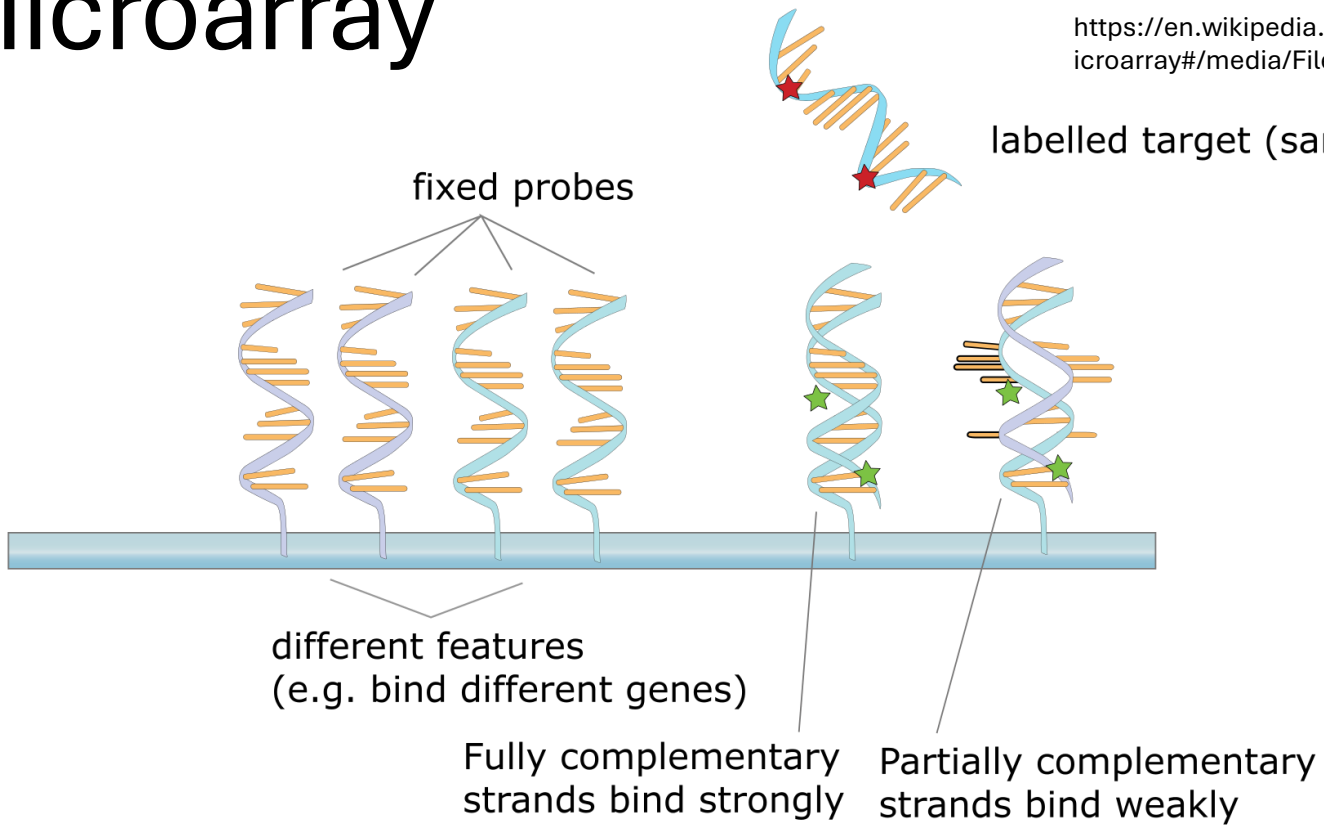
- Given aligned reads to a reference genome, is a read position a variant?



…TAGAGGGGGGTAGAATCCCAGGTGGCCACAAGCCCTCACACGG…

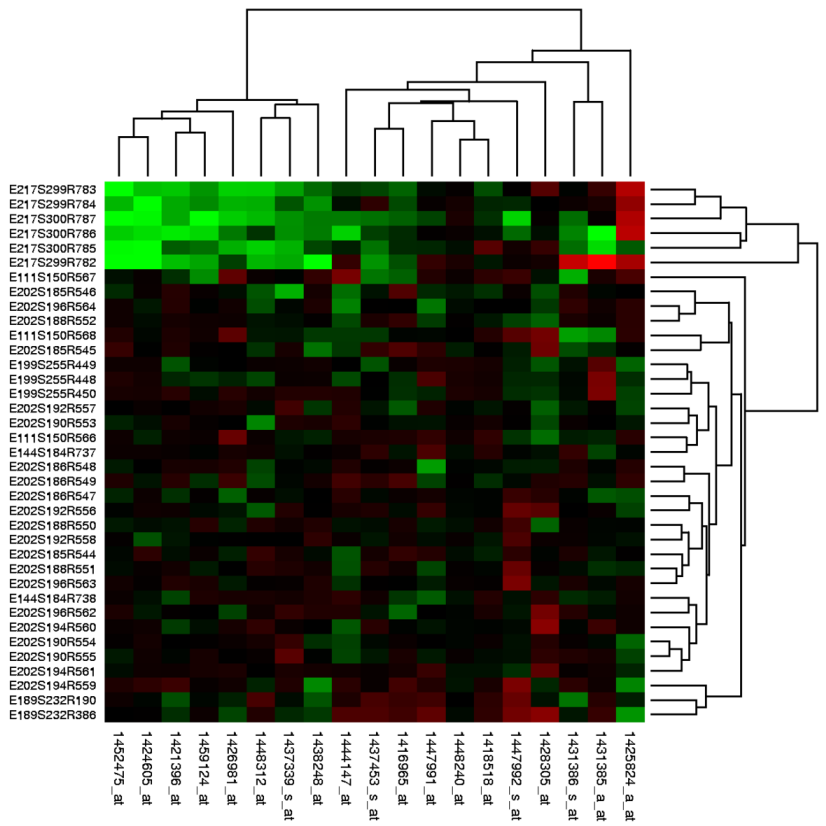
Microarray

https://en.wikipedia.org/wiki/DNA_microarray#/media/File:NA_hybrid.svg



- DNA binds to its reverse complement using hydrogen bonds (A-T, C-G).
- Can construct probes on a chip and then fluorescently label sample to hybridize.

Microarray

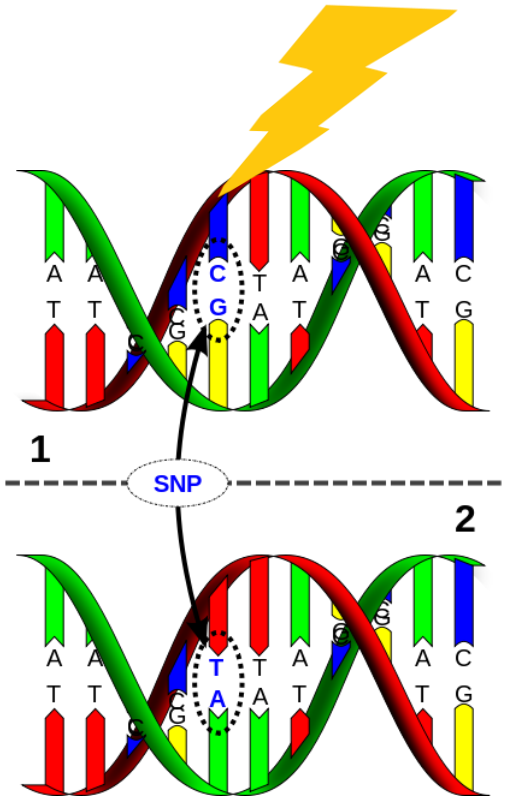


https://en.wikipedia.org/wiki/DNA_microarray#/media/File:Heatmap.png

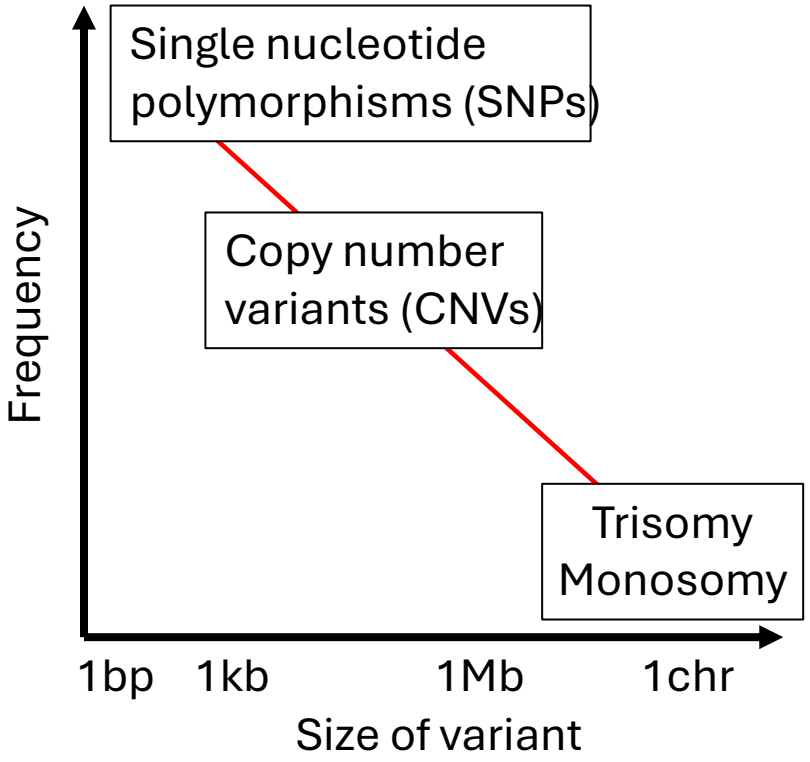
https://en.wikipedia.org/wiki/DNA_microarray#/media/File:Affymetrix-microarray.jpg

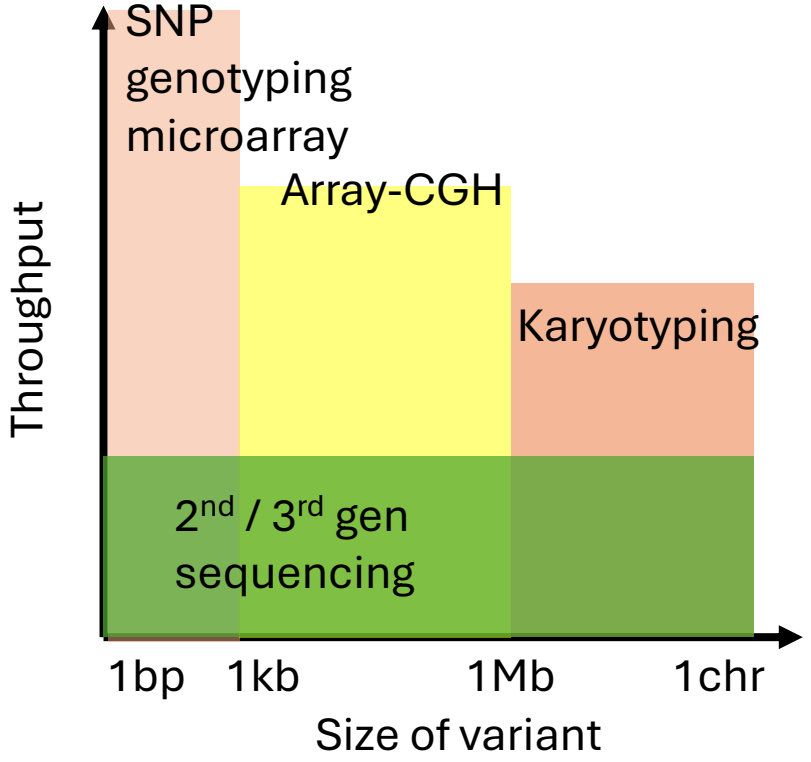
Mutations

- “Typos” occur roughly every 100,000 nucleotides during the cell copying process
- “germline” mutations are inherited by offspring
- This is the source of brand new sequence.
- Can include many types of genetic variation (SNPs, indels, copy number variations, etc.)



Human genetic variation

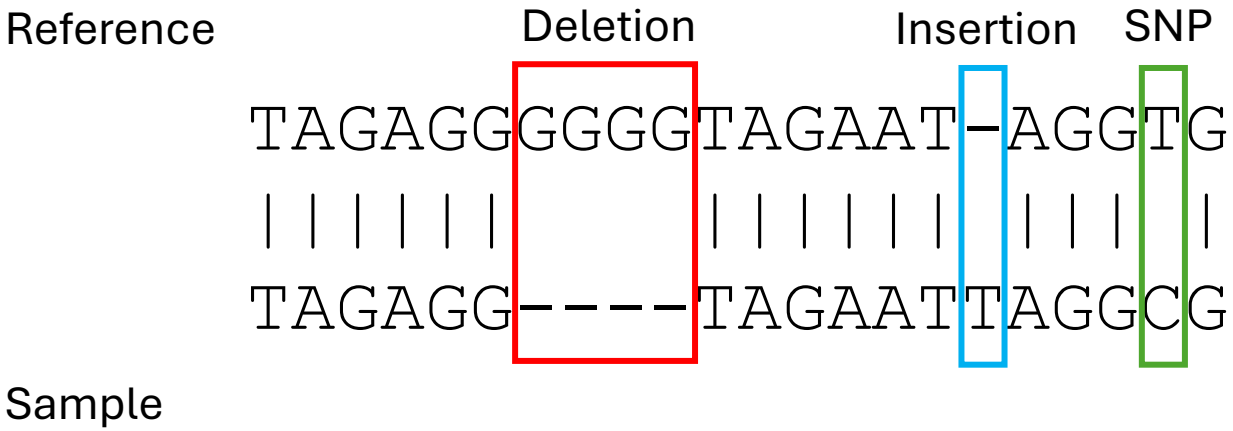




SNPs & short indels

SNP: single nucleotide polymorphism (substitution)

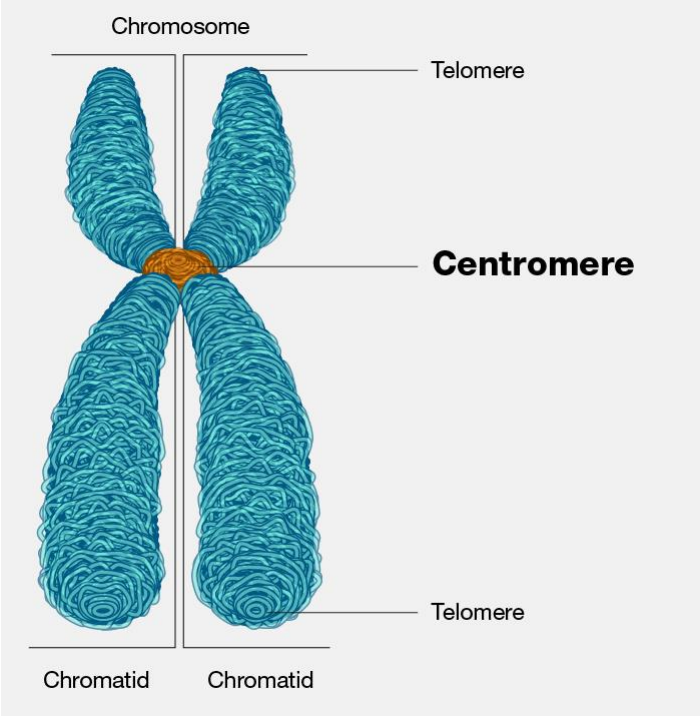
Short indel: insertions or deletions of sequence of length 1 to 50 basepairs



Short tandem repeats

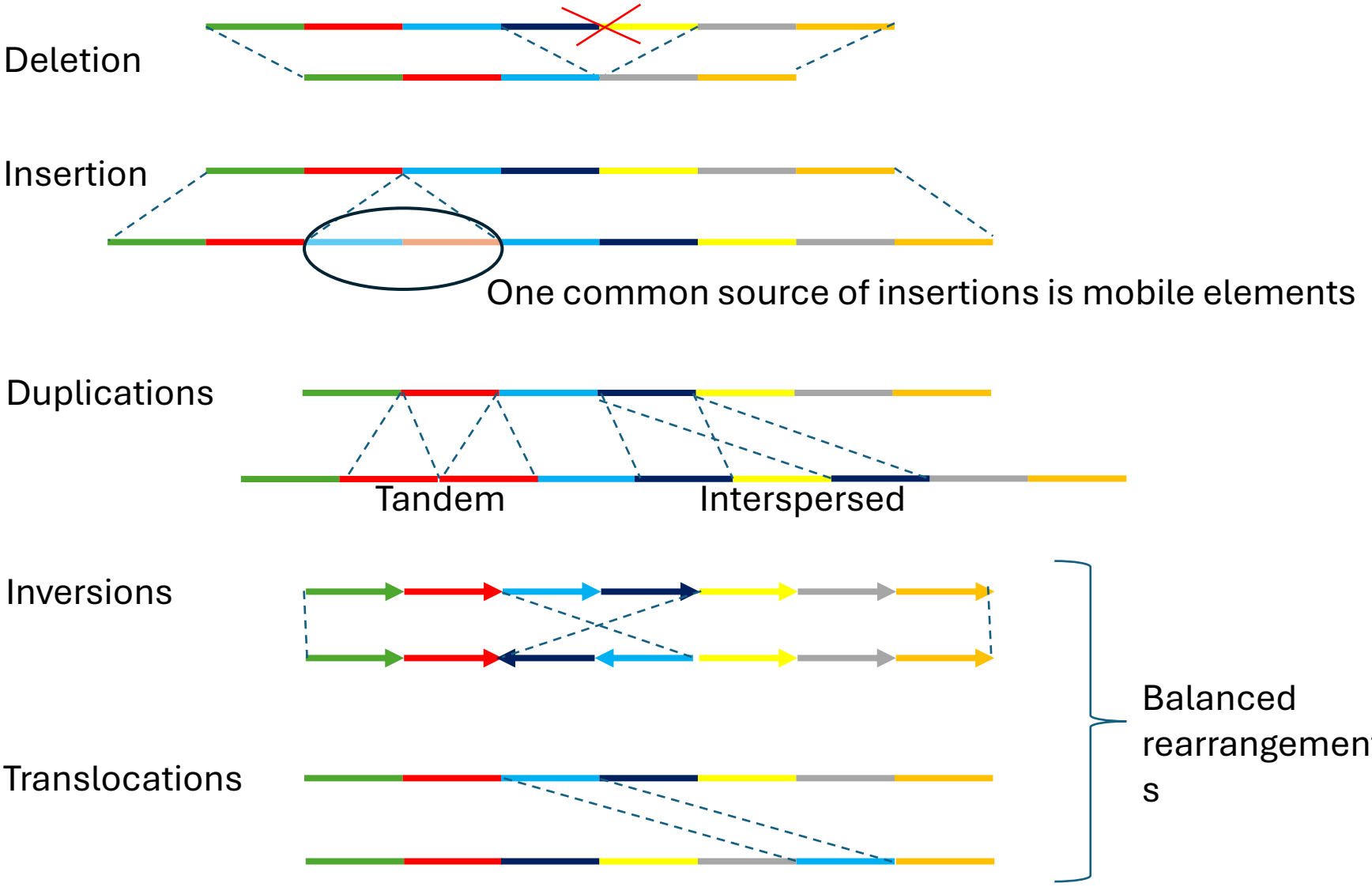
```
Reference:    CAGCAGCAGCAGCAG
Sample:       CAGCAGCAGCAGCAGCAG
```

- Microsatellites (STR=short tandem repeats) 1-10bp
  - Used in population genetics, paternity tests, forensics
- Minisatellites: 10-60bp
- Other satellites
  - Alpha satellites: centromeric/pericentromeric, 170bp in humans
  - Beta satellites: centromeric (some), 68bp in humans
  - Satellite 1 (25-48bp), 2 (5bp), 3 (5bp)
    - E.g. CATTC, imperfectly repeated in 2 and 3.
- Disease relevance:
  - Fragile X Syndrome (CGG)
  - Huntington's disease (CAG)

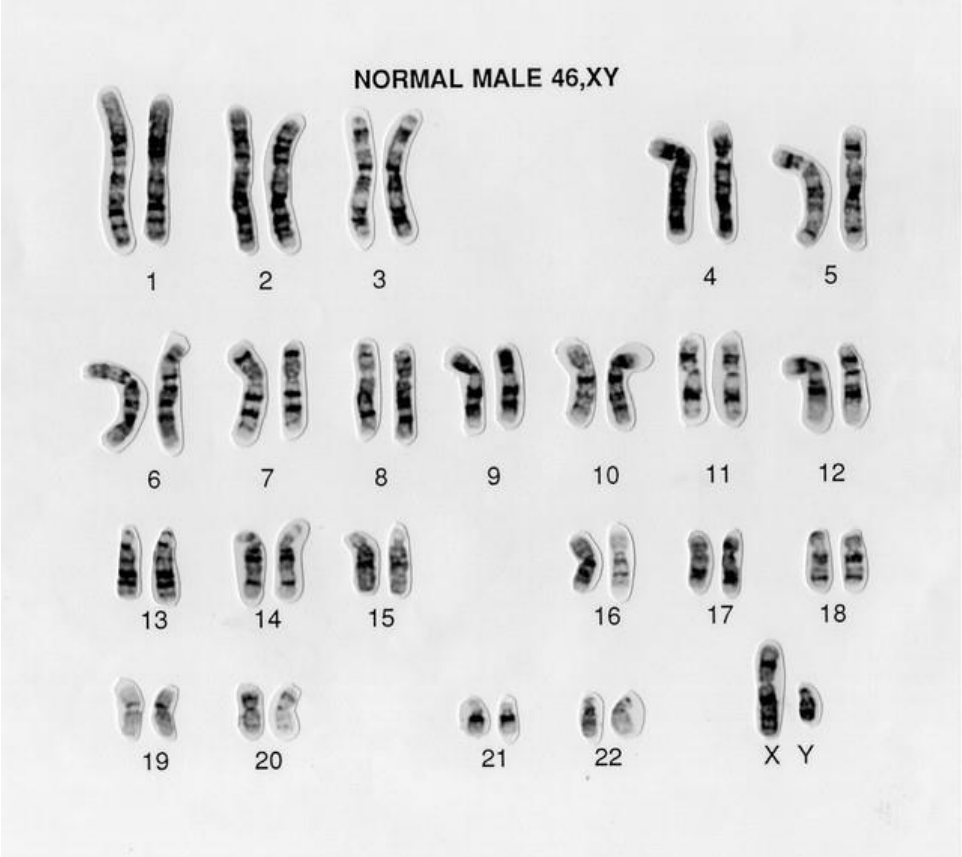


https://www.genome.gov/genetics-glossary/Centromere

Structural variation types



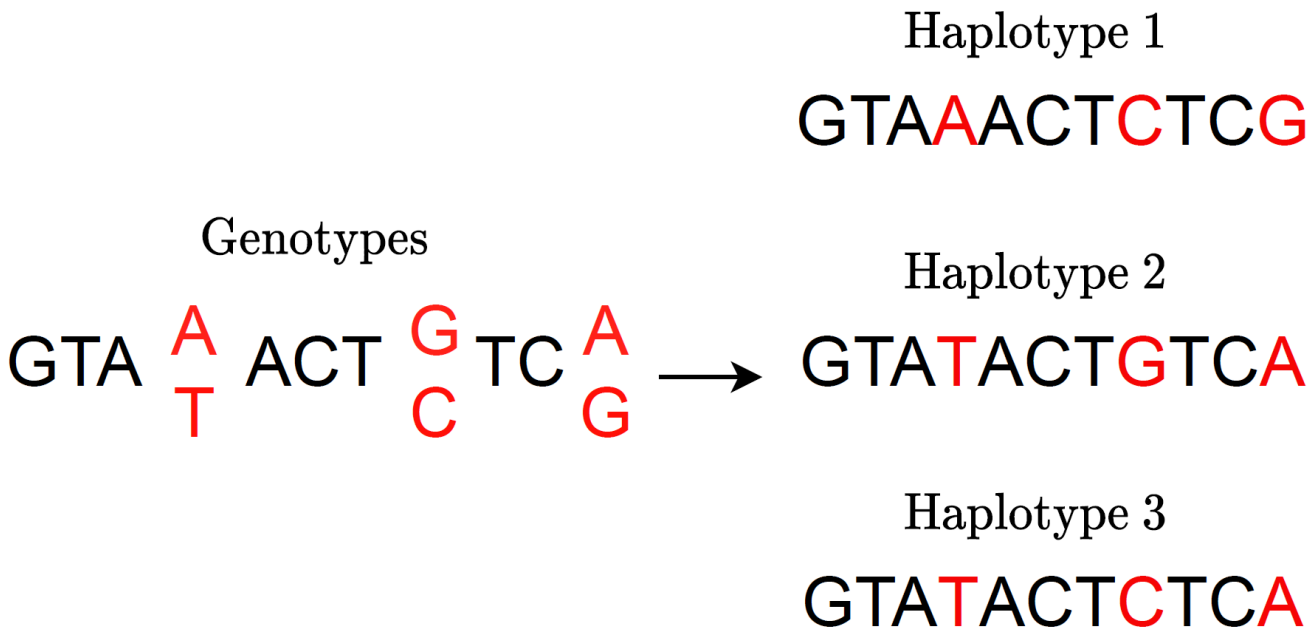
Human Genome / Chromosomes



https://wellcomecollection.org/works/zd7rdetc/images?id=ysp6z895

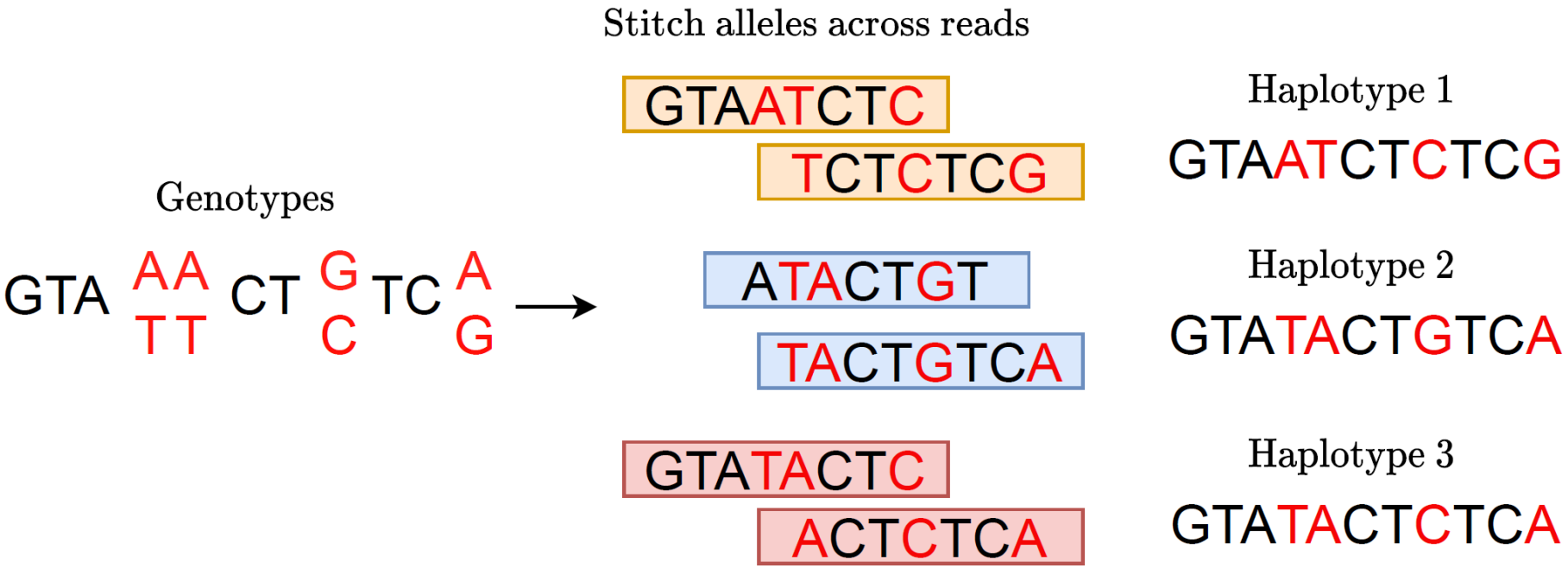
Computational haplotype phasing

- Computationally determining sequence of alleles on each chromosome.



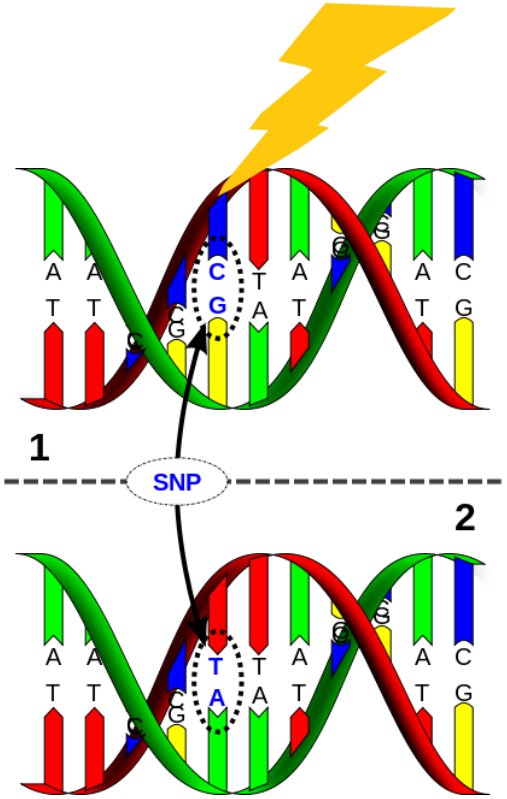
Single individual phasing

- Phasing using aligned reads from single individual (no population information)

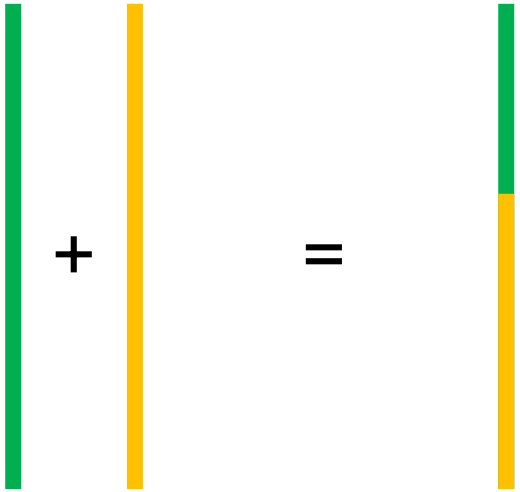


Population level approach

- Micro-scale: mutations slowly change a genome, and are passed down to progeny

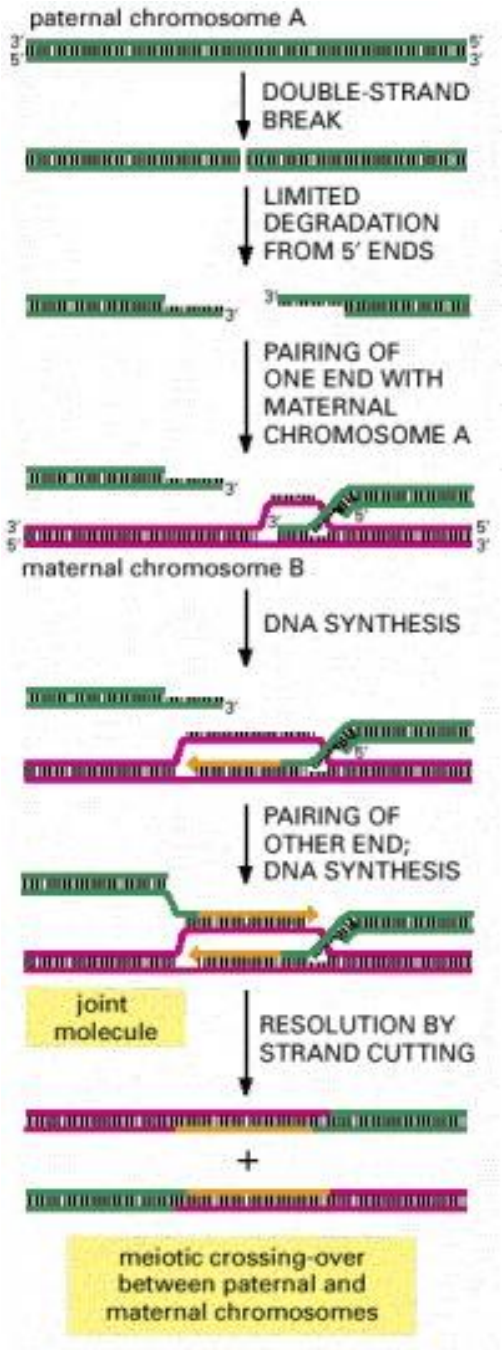


- Macro-scale: DNA from one individual can be mixed together with DNA from another.
  - Mobile Genetic Elements
  - Recombination



Recombination

- DNA recombination is when genetic material is exchanged between chromosomes (or regions of the same chromosome), typically in homologous regions.
- When this happens during meiosis, there can be crossing-over between paternal and maternal chromosomes.



Molecular Biology of the Cell. 4th edition.

Recombination

- Inheritance of genetic material without recombination

- Inheritance of genetic material with recombination

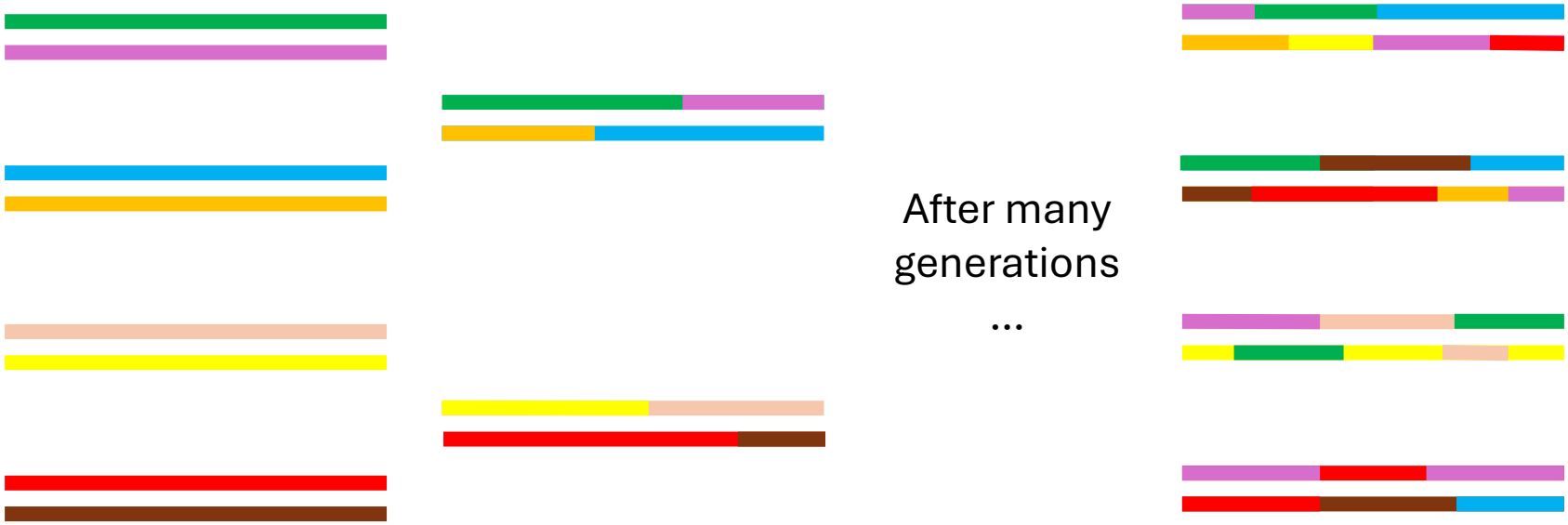
Recombination shapes genome structure

Ancestral population

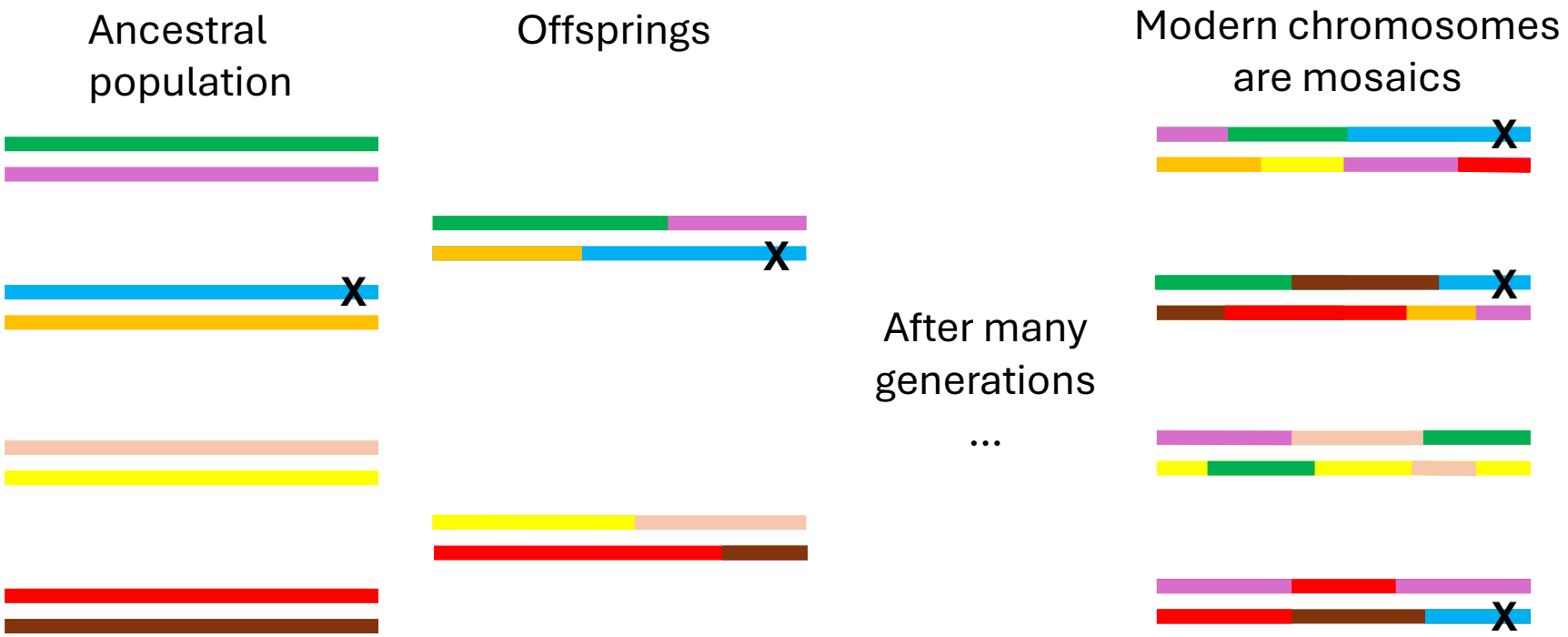
Offsprings

After many generations ...

Modern chromosomes are mosaics

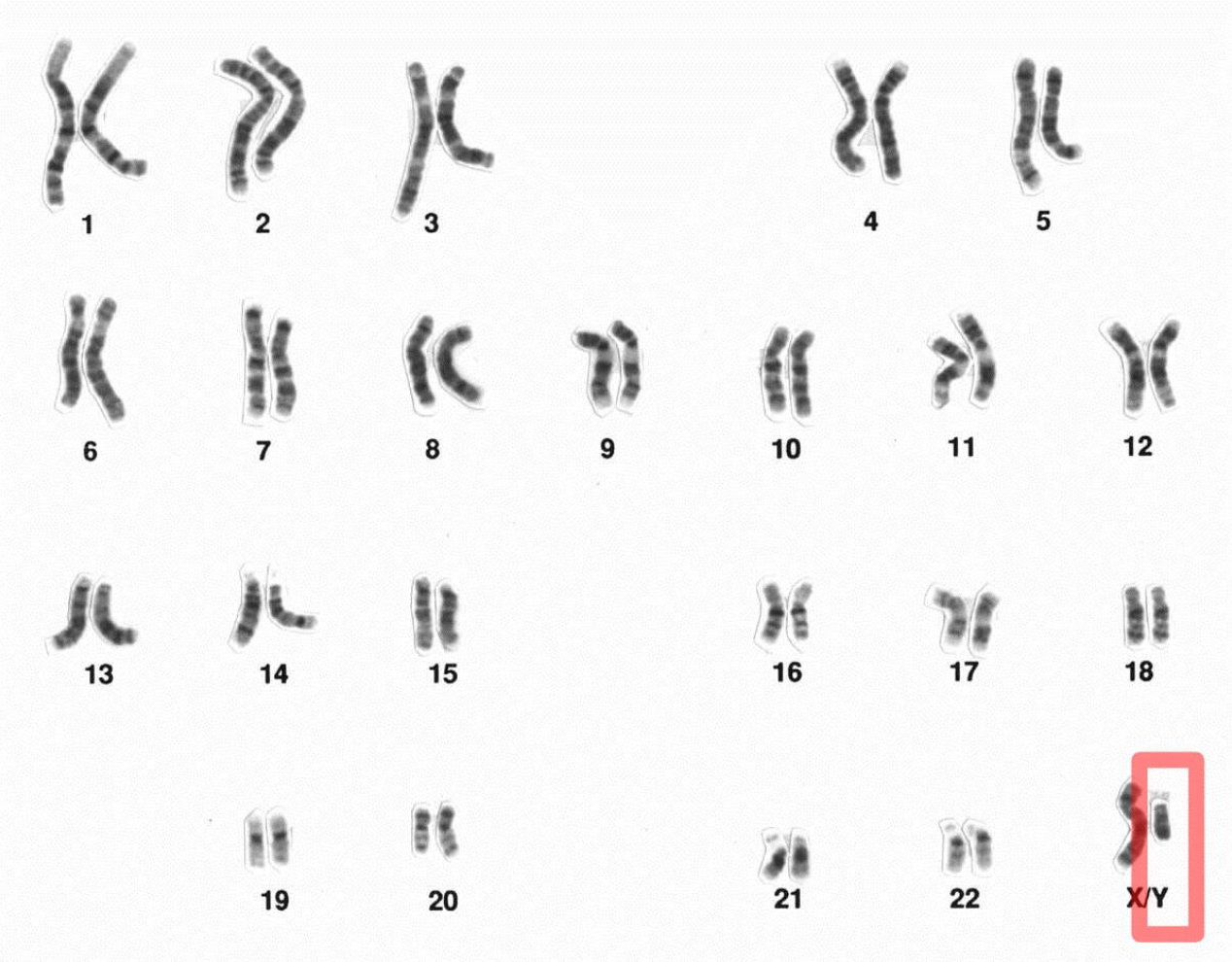


Mutations get mixed throughout population



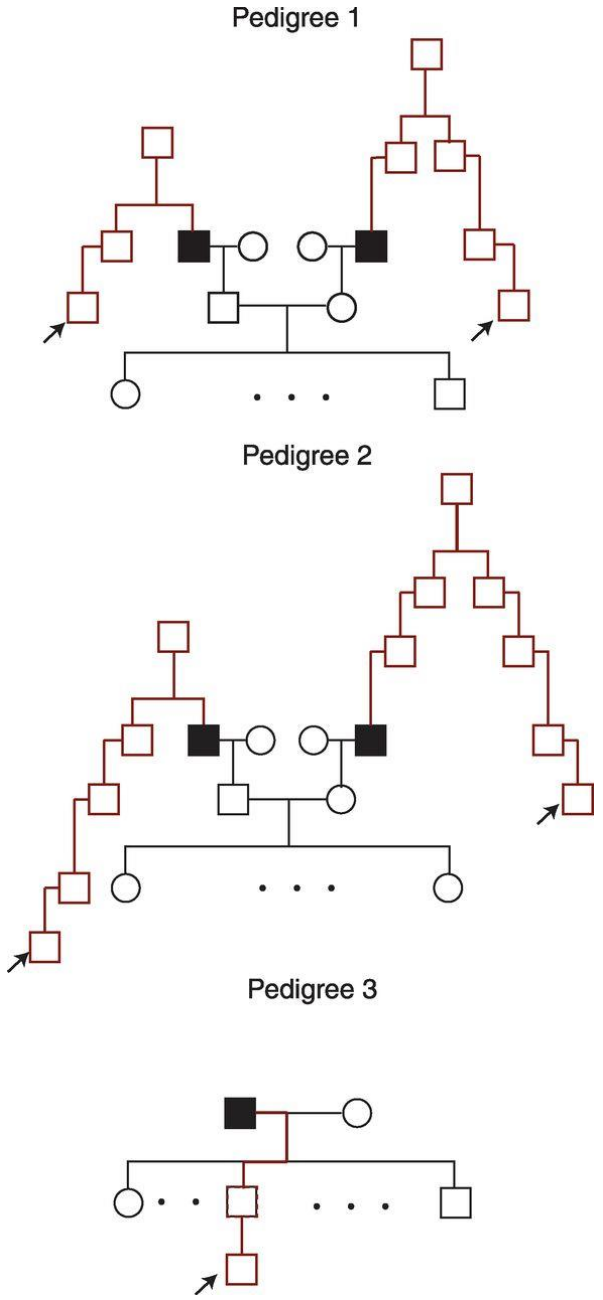
Majority of human sequence variation is due to mutations that have appeared **once** in the history of mankind, and then been passed down through inheritance.

Y-chromosome is special



Last names via Y-STR

- Y-chromosome short tandem repeats can often be used to infer surnames of distant relatives.
- Starting from STRs of 911 individuals, study projected success rate of 12% within U.S. males of Caucasian ancestry.



Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).

Tension: haplotypes are shared

- Maybe bank records belong to you.
- Maybe even medical records like whether or not you've ever broken a leg.
- But what about your genome? Your genome is simultaneously unique to you, while also shared with all of your relatives in part.

Genotype imputation

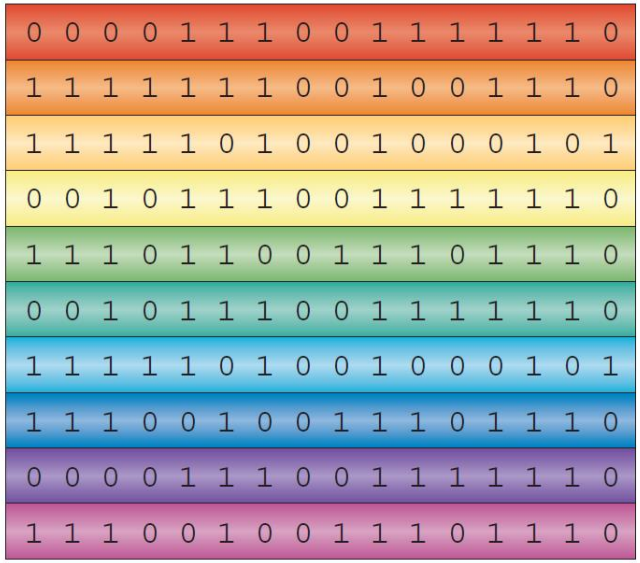
- Consider SNP array genotype data
- Can we infer the untyped genotypes?

Genotype data with missing data at untyped SNPs (question marks)

1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

Marchini & Howie, 2010, NRG

Reference haplotypes (e.g. HapMap)



Genotype imputation

Marchini & Howie, 2010, NRG

Each sample is phased and haplotypes modelled as mosaic of reference panel

Reference haplotypes used to impute alleles into the sample

Small data → big data

- Revealing a few SNPs for genealogical studies may reveal your disease proclivities.

- The Genetic Information Nondiscrimination Act of 2008 (GINA) only prohibits genetic discrimination for health insurance and employment for asymptomatic individuals.

- No protection for life and long-term disability. (though there may be other applicable laws, such as ADA)

RANDY PENCH Sacramento Bee/TNS

Joseph James DeAngelo, the suspect in the Golden State Killer/East Area Rapist case, is arraigned in a Sacramento courtroom and charged with murdering Katie and Brian Maggiore in Rancho Cordova, Calif., in 1978 on Friday.

Police using DNA in ingenious, maybe disturbing ways

lead police to his door.

Police began exploring DNA fingerprinting in 1985, a year before the Golden State Killer raped and bludgeoned his last known victim. It was used to help convict Florida rapist Tommie Lee Andrews two years later, Time magazine reported – then a string of other crim-

The News Tribune, Tacoma, Washington, Sun, Apr 29, 2018

Golden State Killer

- FBI used genomic data from cold case to identify serial killer active between 1974 and 1986, responsible for 13 murders, 51 rapes, and 120 burglaries.
- Did not use traditional FBI DNA database.
- Instead used public GEDmatch ancestry website to find a distant relative to narrow down results.
- GEDmatch allows users to upload their own DNA results from DTC (direct-to-consumer) services like ancestry.com or 23andMe.

Membership Inference

- Sometimes, participating in a study may itself be sensitive information.
  - E.g. knowing you participated in a GWAS targeting individuals who are autistic.

- GWAS only typically publishes summary statistics, and not individual level data.

- But, one can sometimes infer whether or not you participated in a GWAS based just on the published statistics!
  - E.g. The summary statistics suggest that Ashkenazi Jewish lineage is correlated with a disease, and the study was done in a place with a very small Jewish population.

Danger: study volunteers' privacy

- Membership inference leaks a basic amount of information, just a single bit on participation.
- Much more genomic information can be leaked!

Mosca and Cho *Genome Biology* (2023) 24:271
https://doi.org/10.1186/s13059-023-03105-6

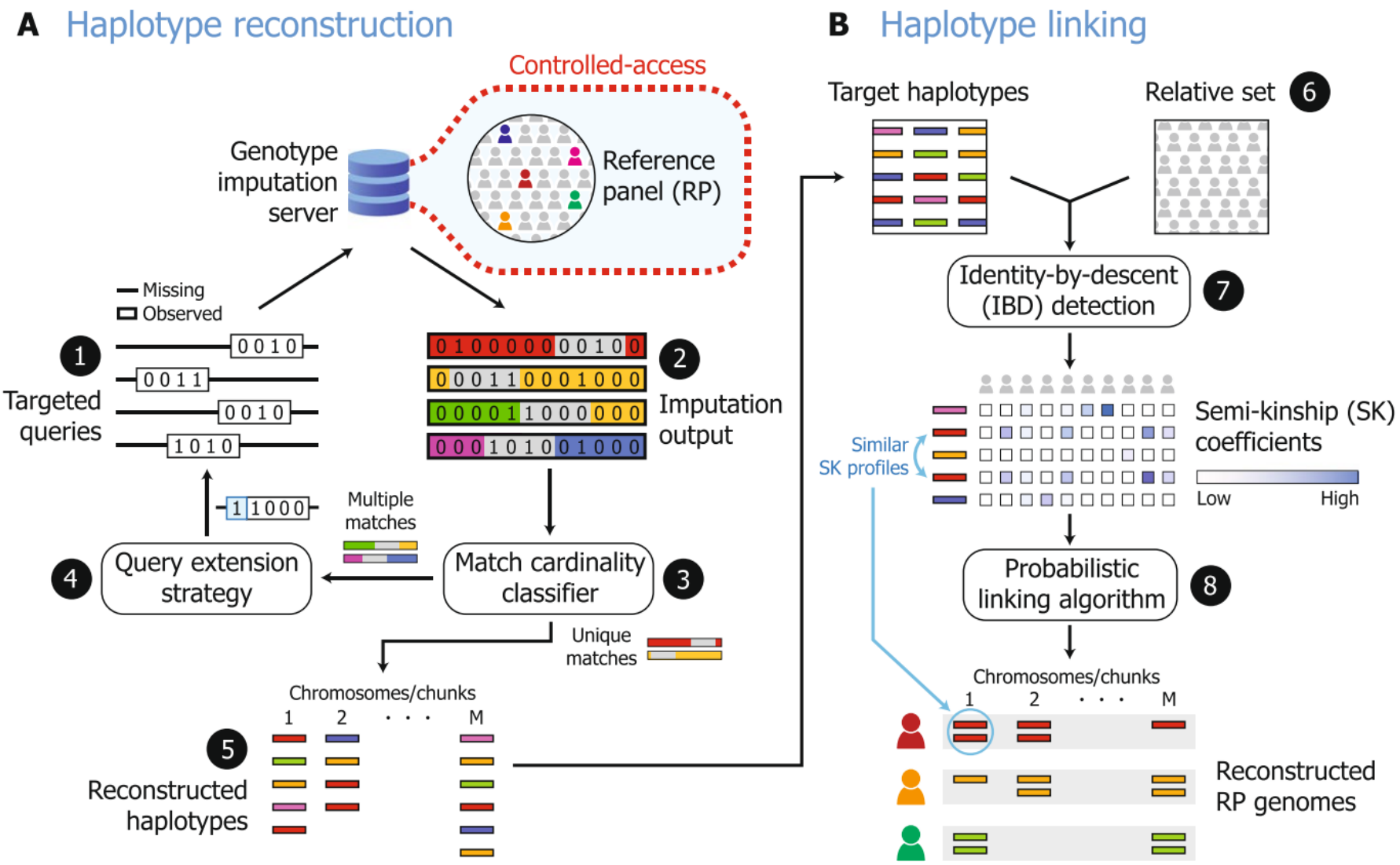
Genome Biology

RESEARCH Open Access

Reconstruction of private genomes through reference-based genotype imputation

Matthew J. Mosca[1] and Hyunghoon Cho[1,2*]

Access controlled databases for imputation leak the original database!

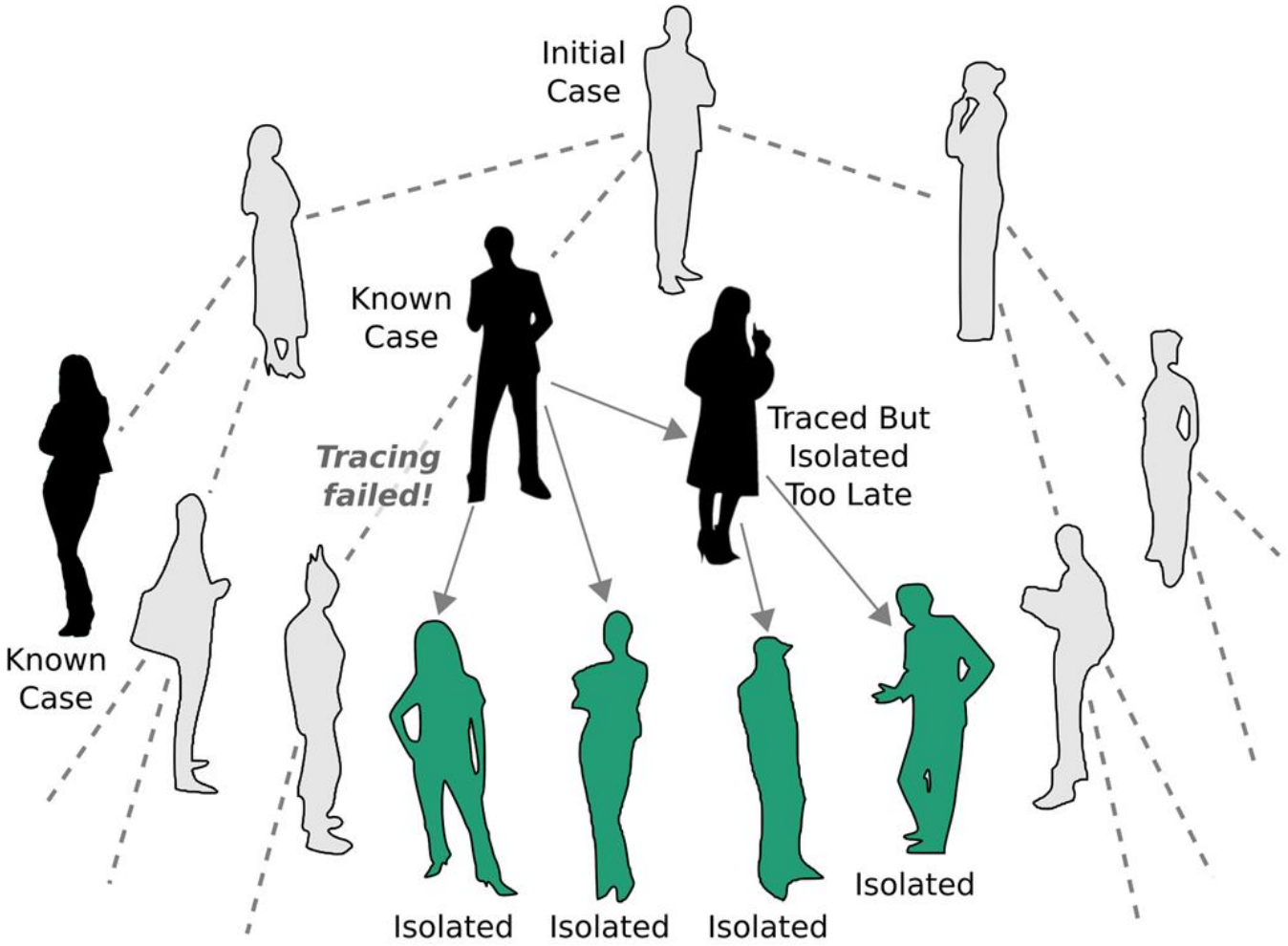


Tension: privacy vs utility

- Medical and genomic data can be used to protect other people.

- No matter what protections we put on the data, simply knowing the answers to the queries we care about reveals private information.

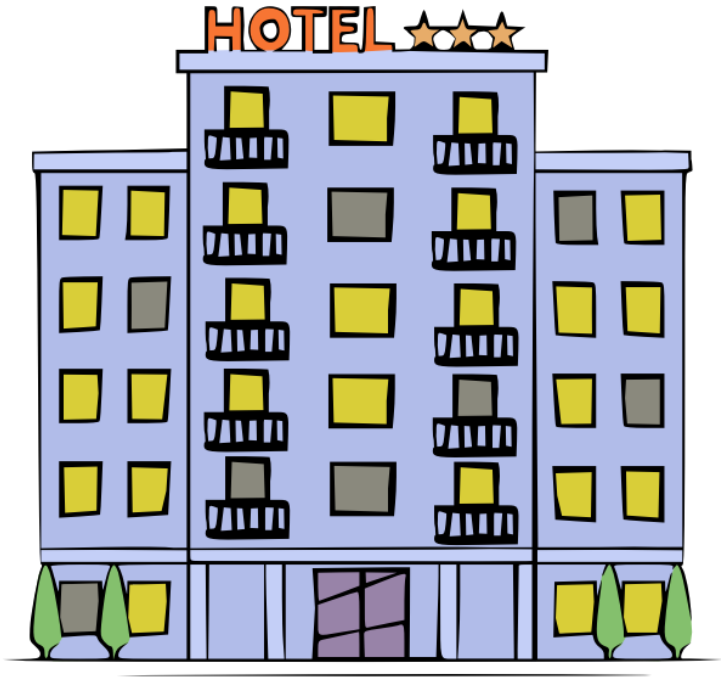
Digital contact tracing



Bradshaw WJ, Alley EC, Huggins JH, Lloyd AL, Esvelt KM. Bidirectional contact tracing could dramatically improve COVID-19 control. Nature communications. 2021 Jan 11;12(1):232.

One privacy threat

- Threat models
  - Government, business, strangers, acquaintances, foreign state actors …
  - Misuse by authorities vs. information leaks vs. disinformation campaigns
  - Passive vs. active attacks
- Example: hotel learning guest diagnosis status

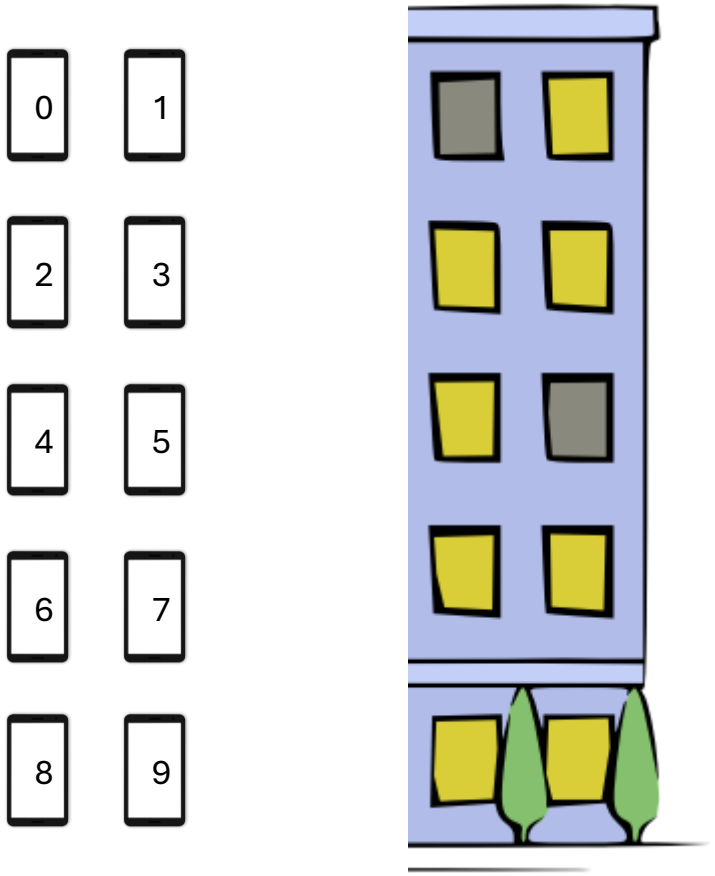


Bengio Y, Ippolito D, Janda R, Jarvie M, Prud'homme B, Rousseau JF, Sharma A, Yu YW. Inherent privacy limitations of decentralized contact tracing apps. Journal of the American Medical Informatics Association. 2021 Jan;28(1):193-5.

One privacy threat

- Threat models
  - Government, business, strangers, acquaintances, foreign state actors …
  - Misuse by authorities vs. information leaks vs. disinformation campaigns
  - Passive vs. active attacks
- Example: hotel learning guest diagnosis status

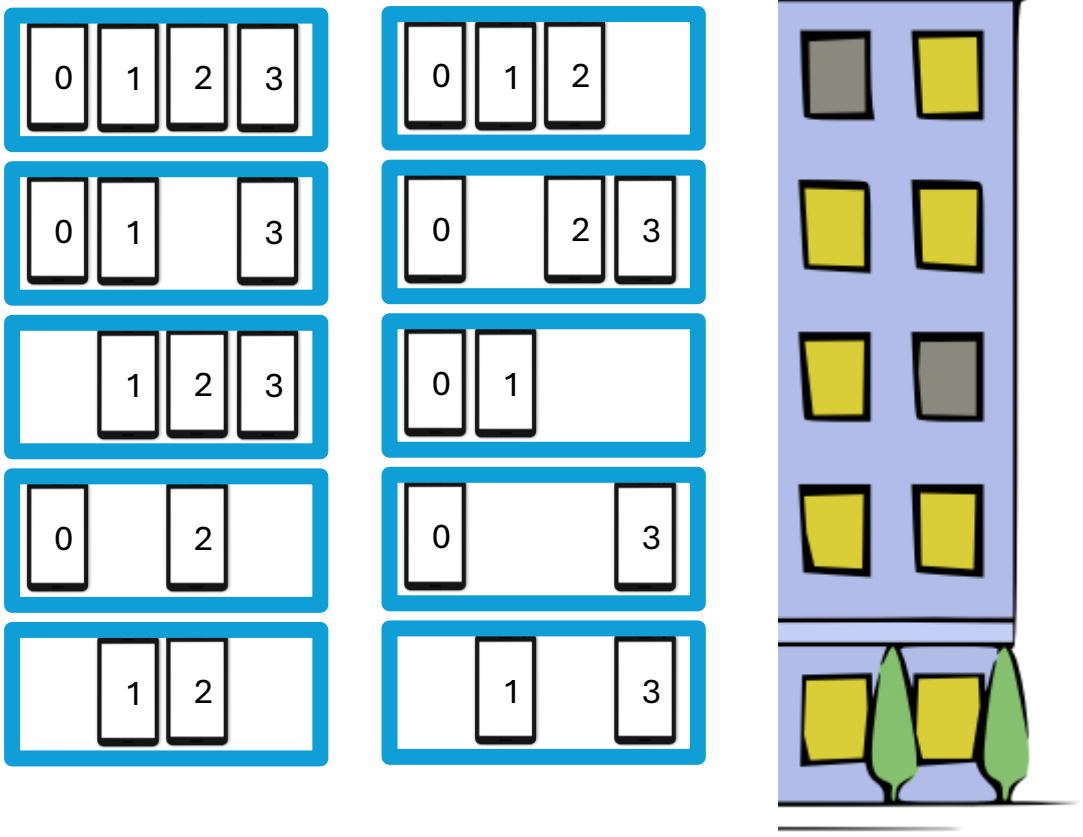
One smartphone per room/day



Bengio Y, Ippolito D, Janda R, Jarvie M, Prud'homme B, Rousseau JF, Sharma A, Yu YW. Inherent privacy limitations of decentralized contact tracing apps. Journal of the American Medical Informatics Association. 2021 Jan;28(1):193-5.

One privacy threat

Binary search through rooms



Bengio Y, Ippolito D, Janda R, Jarvie M, Prud'homme B, Rousseau JF, Sharma A, Yu YW. Inherent privacy limitations of decentralized contact tracing apps. Journal of the American Medical Informatics Association. 2021 Jan;28(1):193-5.

Inherent utility / privacy trade-offs

Feature	Potential privacy loss
Manual contact tracing - Health authority calls exposed users	• Social graph of infected user to health authority • Exposure status of contacts to health authority
Health passport - App informs 3rd parties that a user is "safe"	• Any 3rd party with control over premises may learn current medical status of guests
Geographical infection heatmaps - Health authority generates map of infections	• Infected users who live in remote areas may have their medical status exposed to anyone with the heatmap
Digital exposure notification - App informs user if they had a recent exposure	• Hotels learn positive diagnosis status of guests • Contacts learn when exactly they were exposed
Early warning signals through risk propagation - Apps send around background intermediate risk values to compute user risk levels	• All users reveal partial medical information to hotels, instead of only diagnosed users. • Social networks may be partially revealed through risk propagation.

Data perturbation strategies for privacy protection

- De-identification
  - Obviously, names and other identifiers should be removed.
- Removing sensitive data
  - E.g. rare SNPs
  - K-anonymity
- Adding noise to either original data or to the released statistics.
  - Randomized response
  - Differential privacy

K-anonymity [Sweeney, 2002]

- If every record in a released data set is equivalent to k-1 other records with the same quasi-identifiers, then individuals have plausible deniability.
- Don’t want to release information about a rare disease if only 4 patients in the world have it.
- Difficulty in that it reduces utility of dataset if done well.
- If done poorly, other weak identifiers may be combined to relink individuals.

Survey on sensitive topics

- Premise: sensitive question that someone might not want to admit to.

- Have you ever cheated on an exam?

- A cheater might not want to answer this question honestly, but we might want to know the total proportion of cheaters.

Randomized response [1965, Warner]

- Intuition: add some randomness to give everyone plausible deniability.
- Have you ever cheated on an exam?
  - Flip a coin.
  - If coin is Heads, flip it again, ignore the result, and give the real answer.
  - If coin is Tails, flip it again, and answer Yes if Heads, No if Tails.
- Probability of responses:
  - If 0% cheated, then 25% of respondents answer Yes.
  - If 100% cheated, then 75% of respondents answer Yes.
  - If 50% cheated, then 50% of respondents answer Yes.

Differential Privacy [Dwork, et al, 2006]

- Consider database D with sensitive information managed by a trusted data custodian Trent.
- An analyst Alice asks queries to Trent, who replies.
- We want Trent to reply with helpful answers, but without revealing sensitive information.

Neighboring databases

- Two databases D_1 and D_2 are neighbors if they agree except for one entry.
- Idea: if Trent replies nearly identically for D_1 and D_2, then an attacker can't tell which one was used (and hence can't learn much about the individual.
- Definition:
  - A mechanism M is ϵ-differentially private if for any two neighboring databases D_1 and D_2 and any set R of possible responses

$$\Pr(M(D_1) \in R) \leq \exp(\epsilon) \Pr(M(D_2) \in R)$$

- Thus, for any possible response y

$$\exp(-\epsilon) \leq \frac{\Pr(M(D_1) = y)}{\Pr(M(D_2) = y)} \leq \exp(\epsilon)$$

Differential privacy properties

- EXERCISE: Composability: If you query an ϵ-DP mechanism t times, the result is ϵt-DP.
- Robustness: For any deterministic or random function F you use to postprocess the result from a mechanism M, if M is ϵ-DP, so is $F(M)$.
- How to achieve: Apply random noise to results in a particular way (Laplace noise).

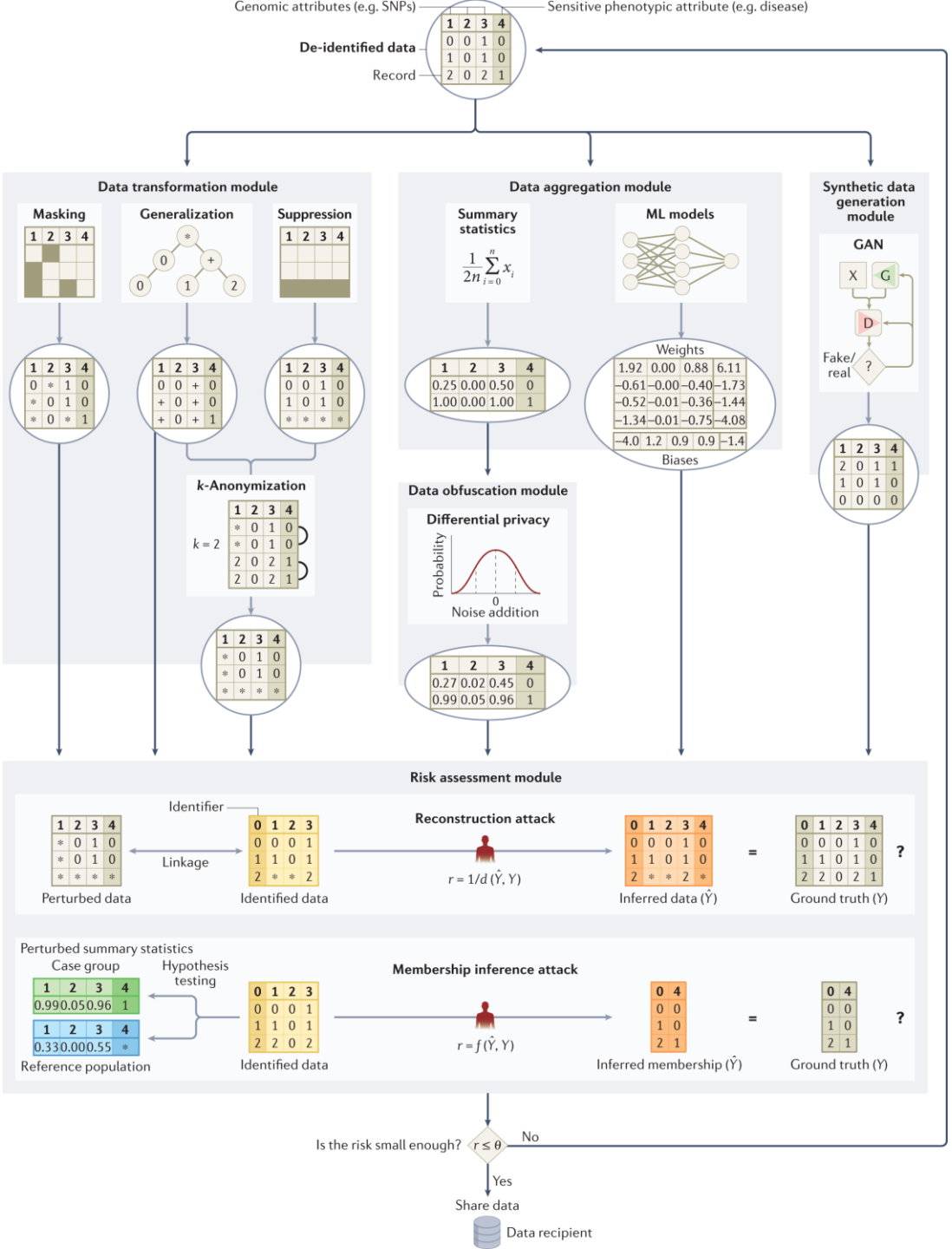
- Problems: noise can mess up utility of data, and can also be harder to explain to researchers.
  - Trade-off of privacy and utility.

Synthetic data

- Alternate approach is to generate new data probabilistically similar to old data.

- If synthetic data is similar enough, can still train models / perform analysis, but without revealing original patient information.

- Difficulty in making sure how similar is similar.



Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nature Reviews Genetics. 2022 Jul;23(7):429-45.

Privacy vs utility

- The amount of noise added simultaneously improves accuracy and decreases utility.

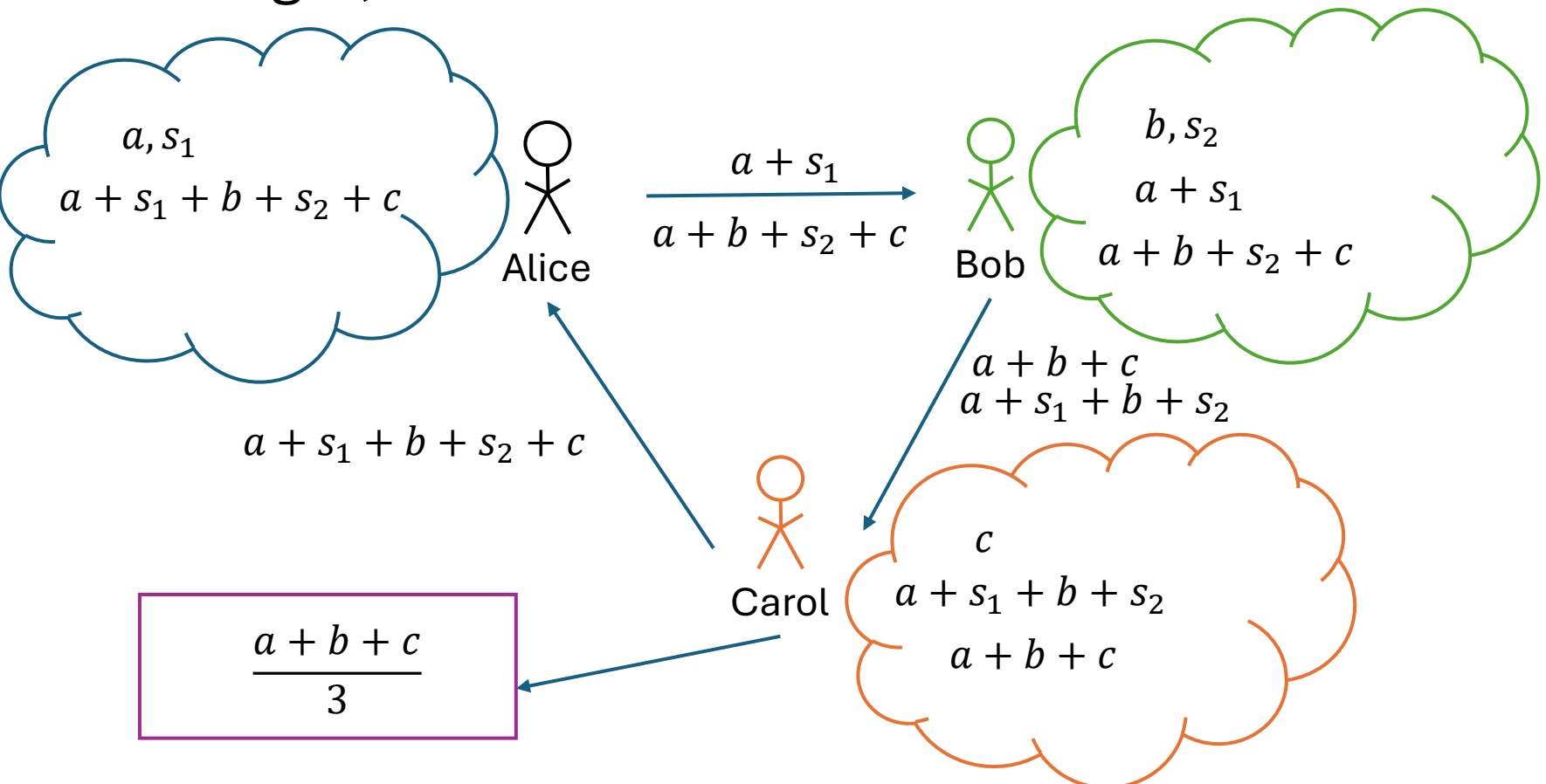
- This trade-off is true for any perturbation-based method for release of data.

Multi-party computation problem

- When many untrusted parties are collaborating to do an analysis, computing information may inadvertently reveal private information.
- One solution is legal data use agreements and controls.
- Another solution is adding noise (e.g. differential privacy or randomized response), but this comes at cost to accuracy.
- Another is homomorphic encryption and secure multiparty computation (MPC).

MPC example: secret shares

- Alice, Bob, and Carol want to know their average weight, but not reveal it to each other.

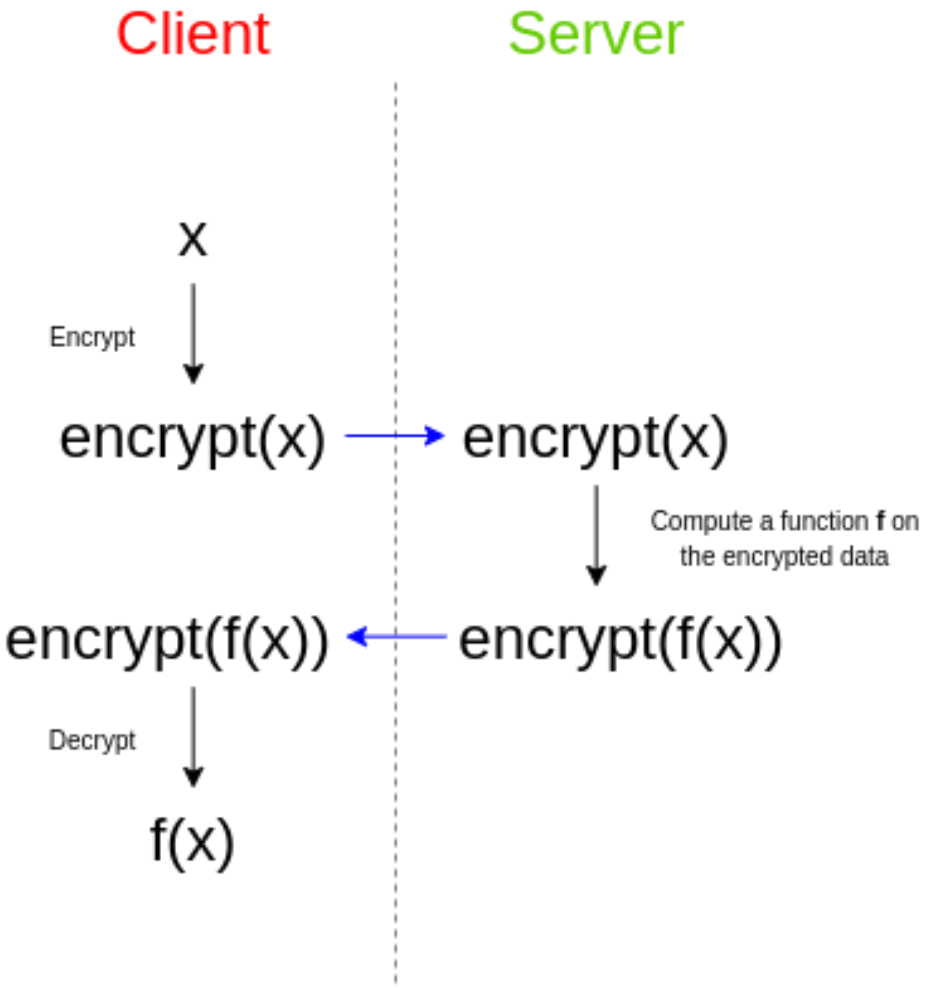


Exercise

- This protocol is fully decentralized. How secure is this protocol if you centralize all communication through a central server?
- Is this protocol secure against an eavesdropper?
- If not in either case, can you adapt this protocol to secure it?

Homomorphic Encryption

- Client has data x. Server has an algorithm f. Compute $f(x)$ without revealing the data or algorithm to each other (except $f(x)$).
- Trick is to have a cryptosystem that allows operating directly on encrypted data.



https://vitalik.eth.limo/general/2020/07/20/homomorphic.html

Simple RSA example for multiplication

- Find triple e, d, n such that $(m^e)^d \equiv m \pmod{n}$.
  - Public key (e, n). Encrypt message m by $m^e \pmod{n}$.
  - Private key d. Decrypt ciphertext c by $c^d \pmod{n}$.
- If the data are x and y, we can find the product $f(x, y) = xy$ homomorphically.
- Client sends messages x^e, $y^e \pmod{n}$ to Server.
- Server multiples them together without decrypting to get $c = x^e y^e = (xy)^e$, and sends it back.
- Client decrypts with by taking $c^d \equiv \left((xy)^e\right)^d \equiv xy$

Exercise

- Break simple RSA.
- More precisely, provide a method to decrypt at least one encrypted value without knowing the private key. *Hint: perhaps a value trivially encrypts*.
- *Giving up the homomorphic property*, come up with a workaround to the attack above.
  - Note: your solution doesn't need to preserve the homomorphic property.

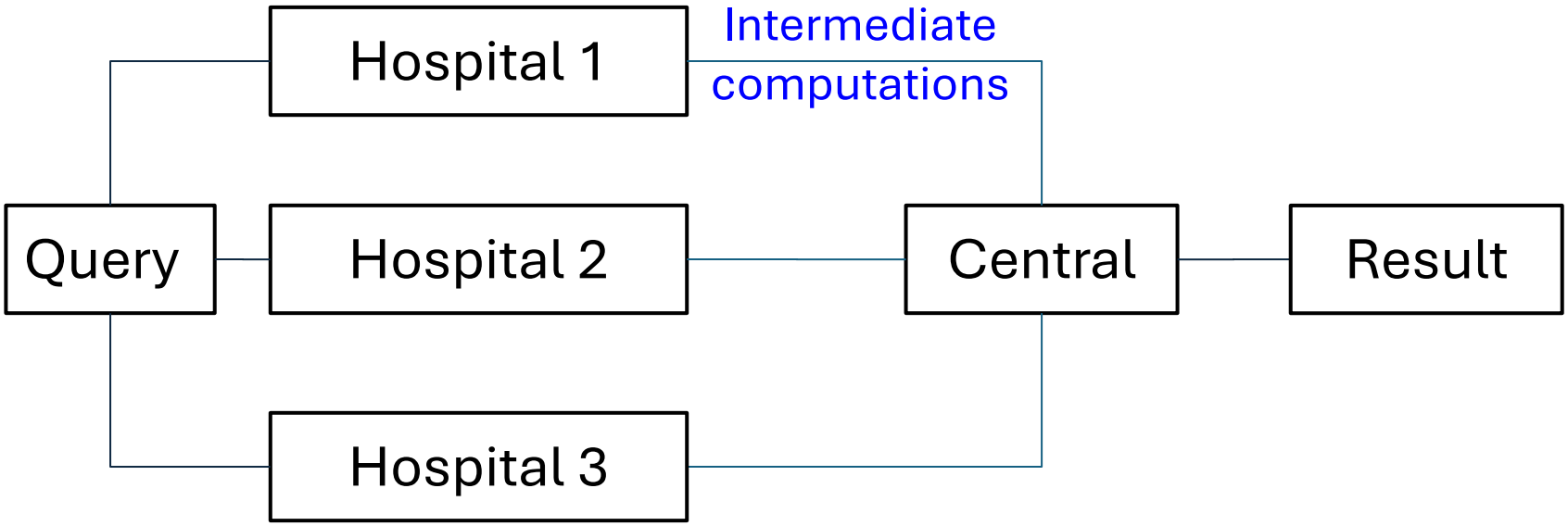
Secure Federated Boolean Count Queries Using Fully-Homomorphic Cryptography

Alexander T. Leighton, Brandeis University

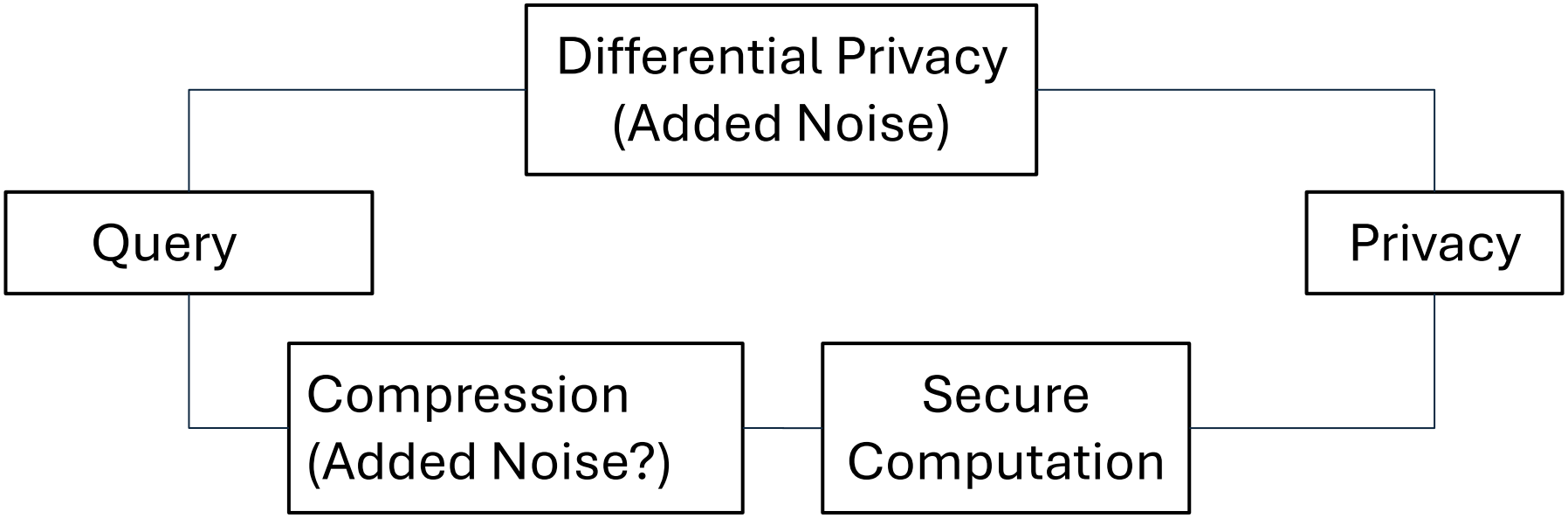
Yun William Yu, PhD, Carnegie Mellon University

Research in Computational Molecular Biology: 28th Annual International Conference, RECOMB 2024, Cambridge, MA, USA, April 29–May 2, 2024, Proceedings Pages 54 - 67 https://doi.org/10.1007/978-1-0716-3989-4_4

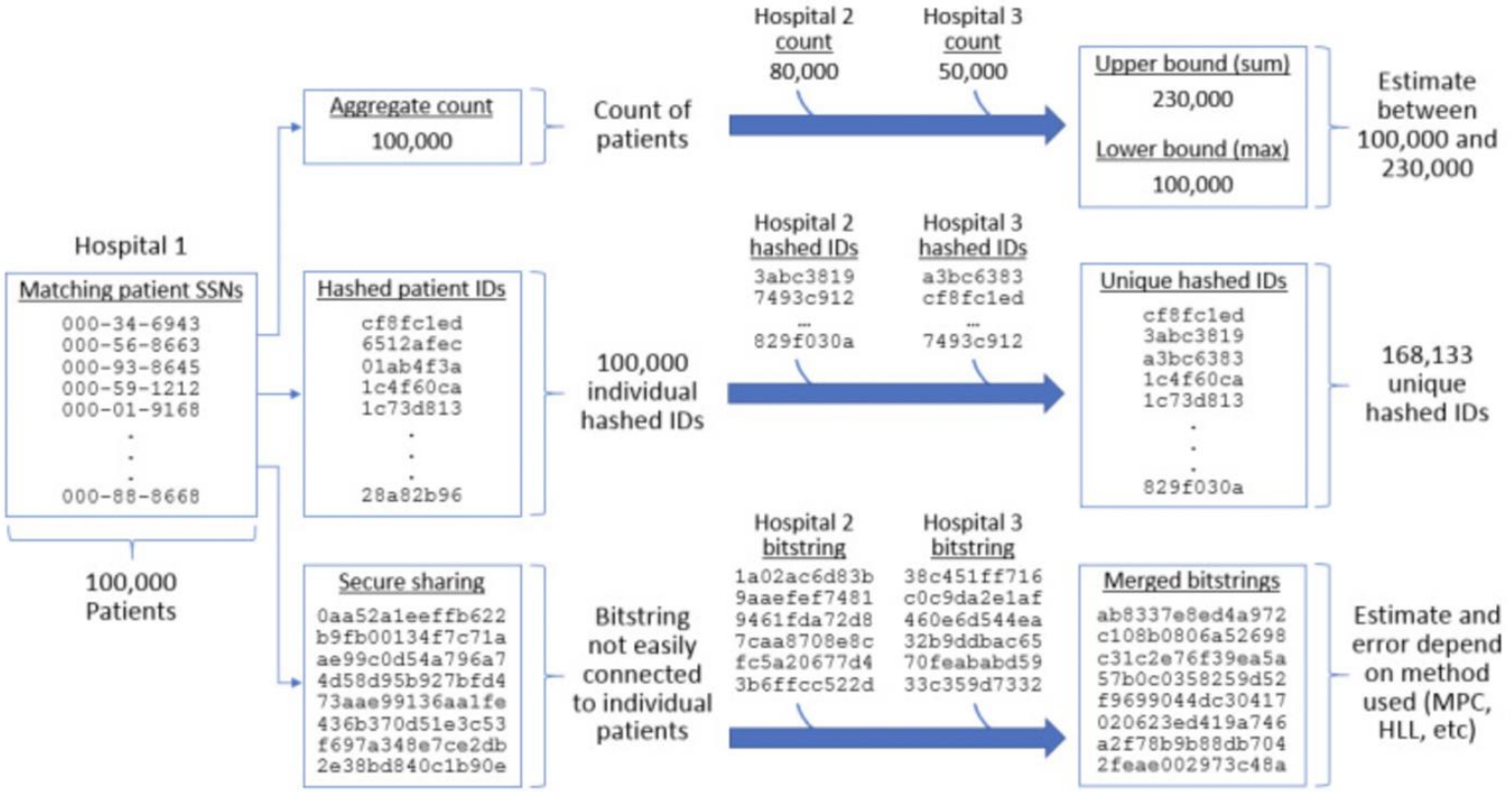
Federated Queries



Approaches to Privacy of Intermediate Computations

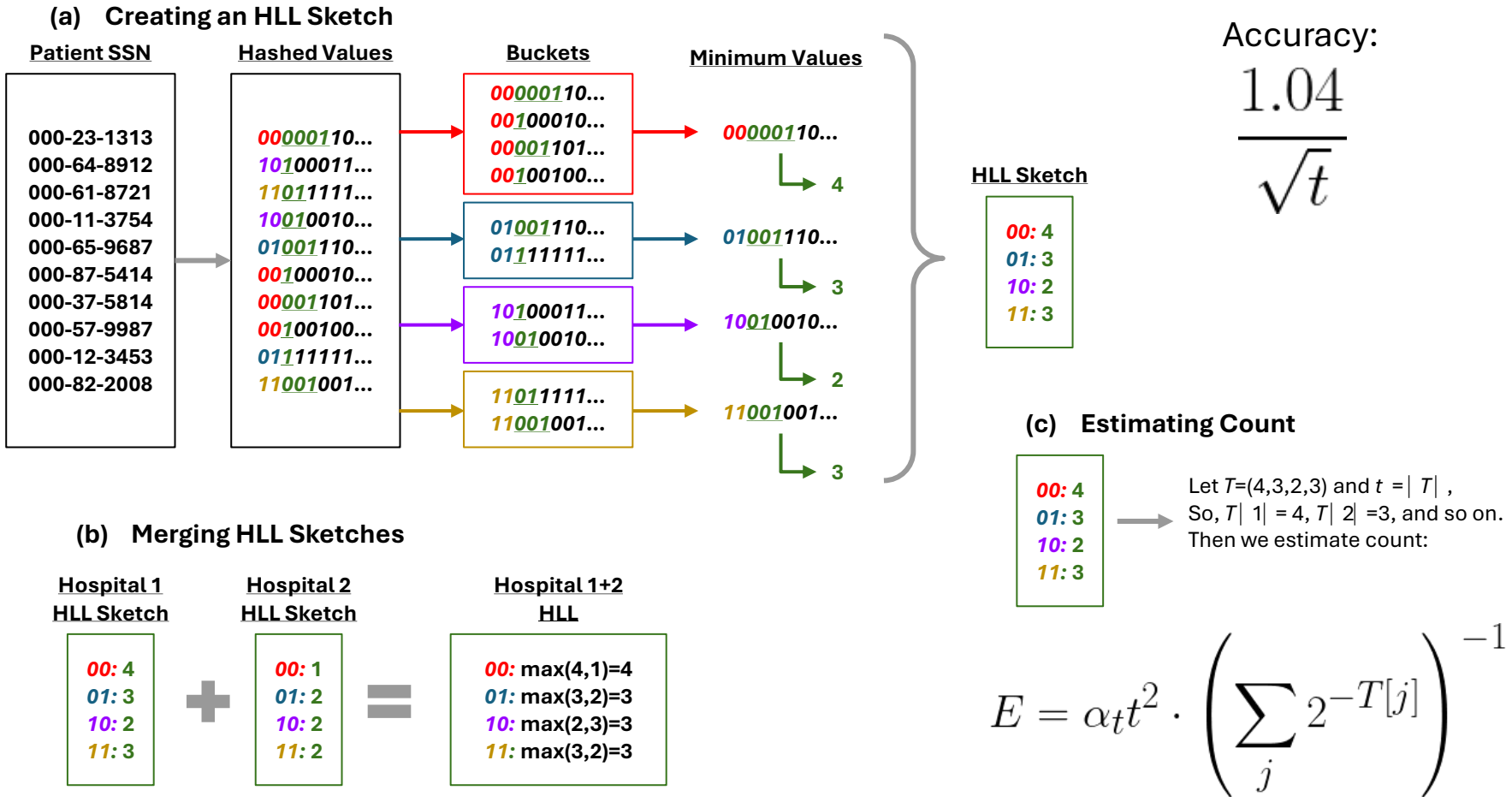


Count Queries



Yu YW, Weber GM. Balancing Accuracy and Privacy in Federated Queries of Clinical Data Repositories: Algorithm Development and Validation. *J Med Internet Res*. 2020;22(11):e18735. Published 2020 Nov 3. doi:10.2196/18735

HyperLogLog Sketching [Flajolet et al. 07]



Accuracy:

$$\frac{1.04}{\sqrt{t}}$$

(b) Merging HLL Sketches

(c) Estimating Count

00: 4
01: 3
10: 2
11: 3

Let T=(4,3,2,3) and $t = |T|$,
So, $T[1] = 4$, $T[2] = 3$, and so on.
Then we estimate count:

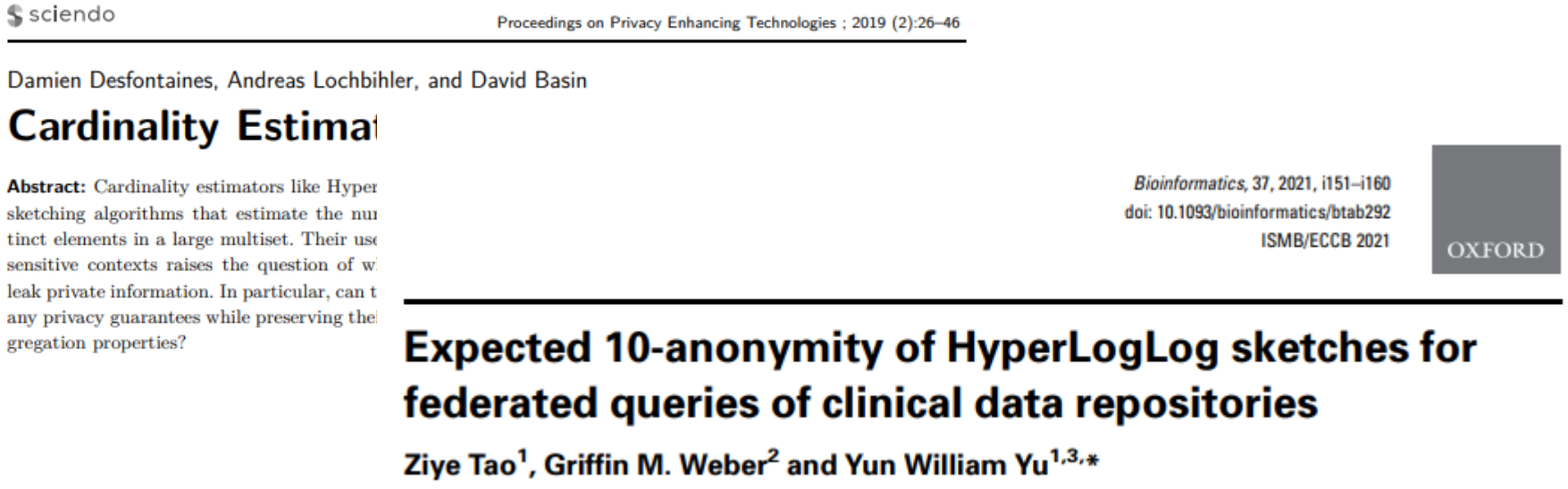
$$E = \alpha_t t^2 \cdot \left(\sum_j 2^{-T[j]} \right)^{-1}$$

(Lack of) Privacy of HLL Sketching

- Bob is Y years old and has rare condition Z and I want to know if he has Diabetes
  - Send query {Y and Z}
  - Then send queries {Y and Z and Diabetes} and {Y and Z and NOT Diabetes}
- Adversary knows what bucket Bob hashes to, and if that bucket changes in either of the latter queries, the adversary immediately learns Bob's disease state.

Deduplication is antithetical to privacy

- Intuition: if a hospital publishes information that allows a patient Bob to NOT be double counted, then an adversary can simply try counting Bob and see if anything changes.



sciendo

Proceedings on Privacy Enhancing Technologies ; 2019 (2):26–46

Damien Desfontaines, Andreas Lochbihler, and David Basin

Cardinality Estimat

Abstract: Cardinality estimators like Hyper
sketching algorithms that estimate the nu
tinct elements in a large multiset. Their us
sensitive contexts raises the question of w
leak private information. In particular, can t
any privacy guarantees while preserving the
gregation properties?

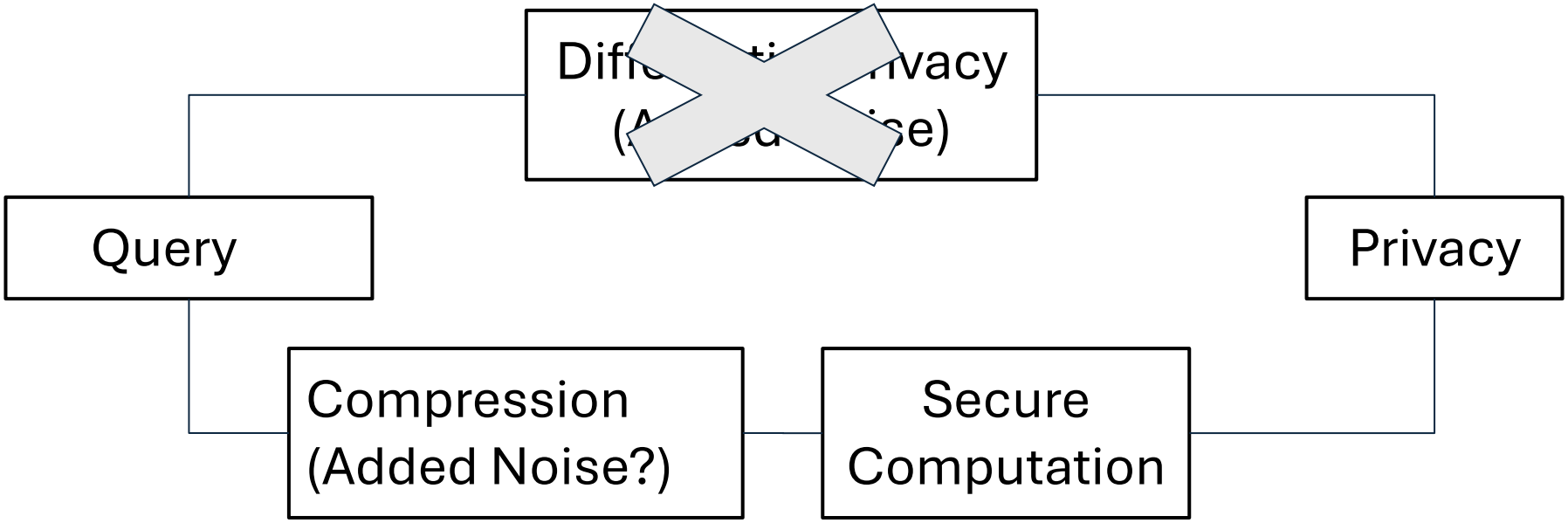
Bioinformatics, 37, 2021, i151–i160
doi: 10.1093/bioinformatics/btab292
ISMB/ECCB 2021

OXFORD

Expected 10-anonymity of HyperLogLog sketches for federated queries of clinical data repositories

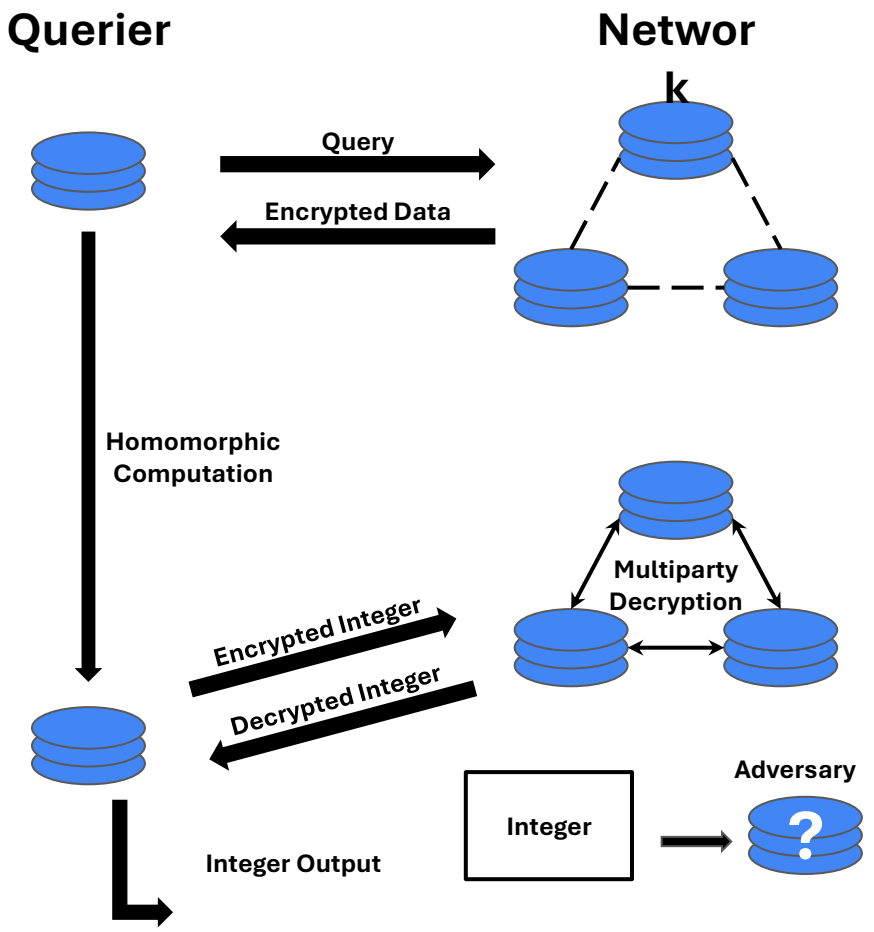
Ziye Tao[1], Griffin M. Weber[2] and Yun William Yu[1,3,*]

Approaches to Privacy of Intermediate Computations



Protocol

1. Compute the estimator with secure multiparty computation
   a. Fully Homomorphic Encryption (FHE)
   b. Don't reveal any hash values!
   c. Be clever about merging ciphertext



Fully homomorphic encryption scheme

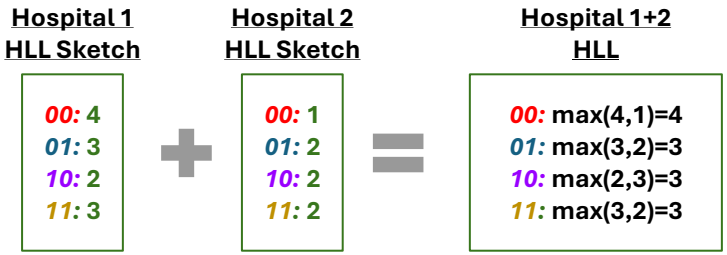
Used Palisade's implementation of BFVrns (Brakerski, Fan, Vercauteren residue number system), which gives fast addition and slow multiplication of integers.

- In theory, can do any computation, because you can rewrite everything as combination of addition and multiplication.
- In practice, slow if there are too many multiplications, and not feasible for complicated functions like exponentials and inverses.

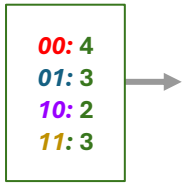
Key is to rewrite count-queries as addition and multiplication.

A Caveat

Merging HLL Sketches:



Estimating Count HyperLogLog

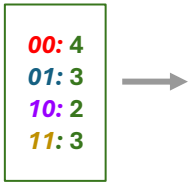


Let T=(4,3,2,3) and $t = |T|$,
So, $T[1] = 4$, $T[2] = 3$, and so on.
Then we estimate count:

$$E = \alpha_t t^2 \cdot \left(\sum_j 2^{-T[j]} \right)^{-1}$$

Accuracy: $\frac{1.04}{\sqrt{t}}$

Estimating Count LogLog

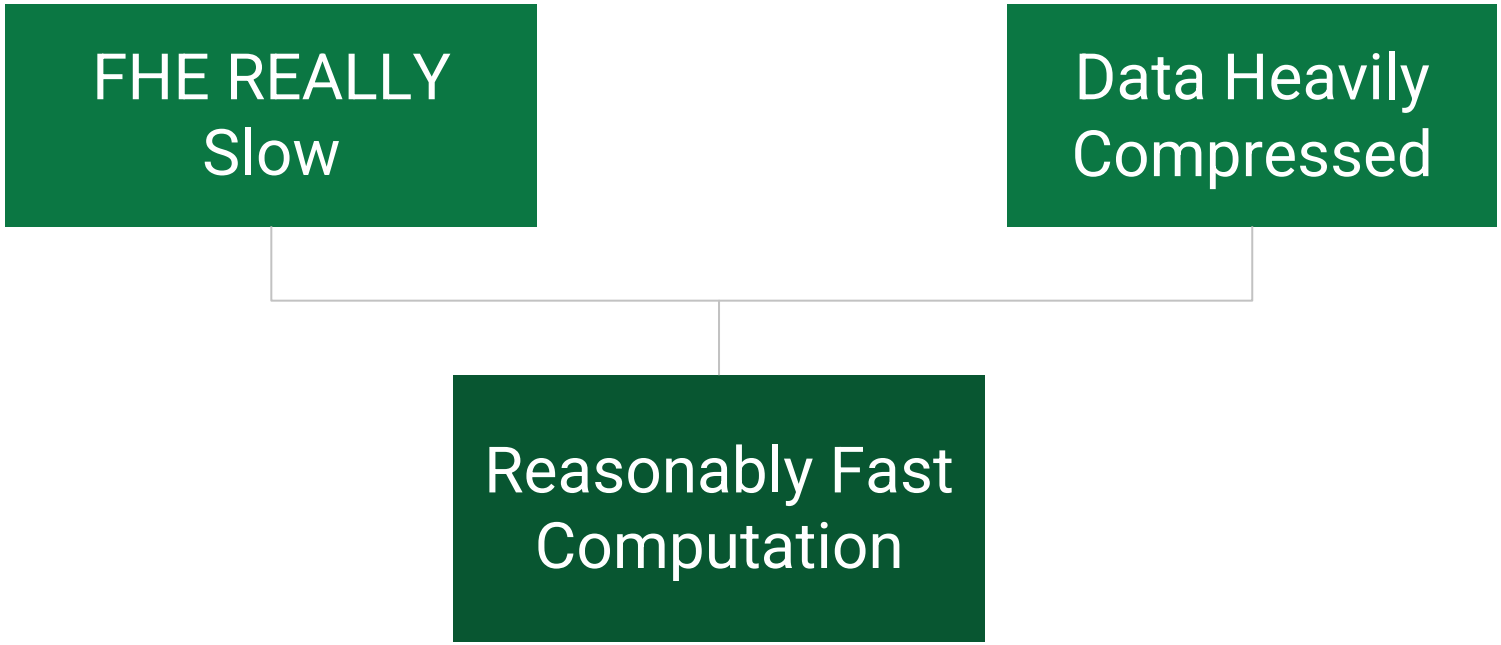


Let T=(4,3,2,3) and $t = |T|$,
So, $T[1] = 4$, $T[2] = 3$, and so on.
Then we estimate count:

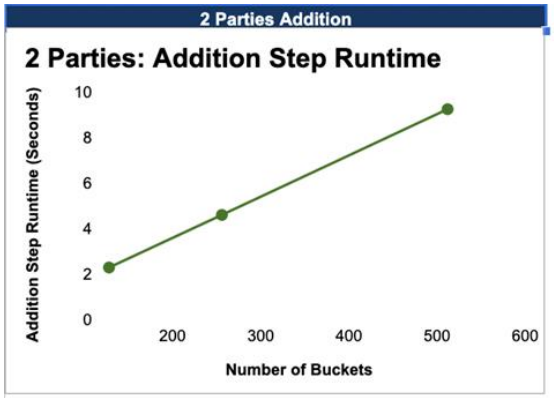
$$E = \alpha_t t \cdot 2^{\frac{1}{t}\sum_j T[j]}$$

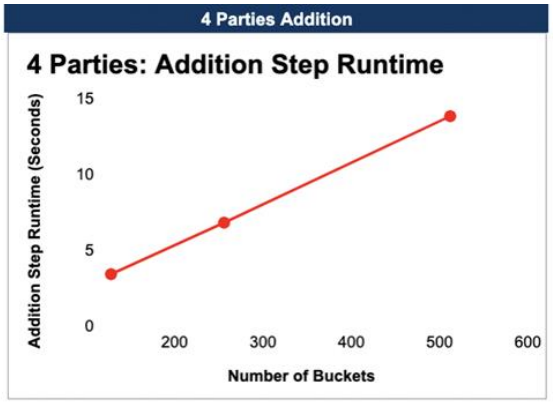
Accuracy: $\frac{1.30}{\sqrt{t}}$

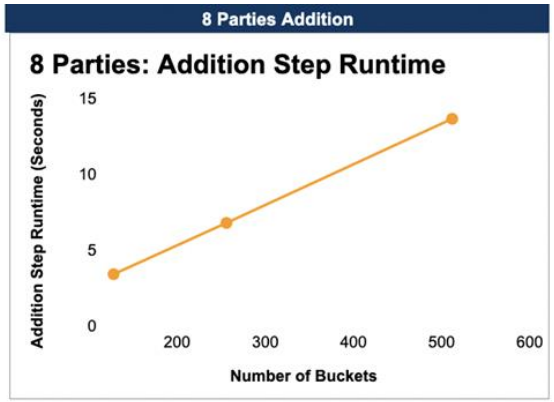
To Summarize

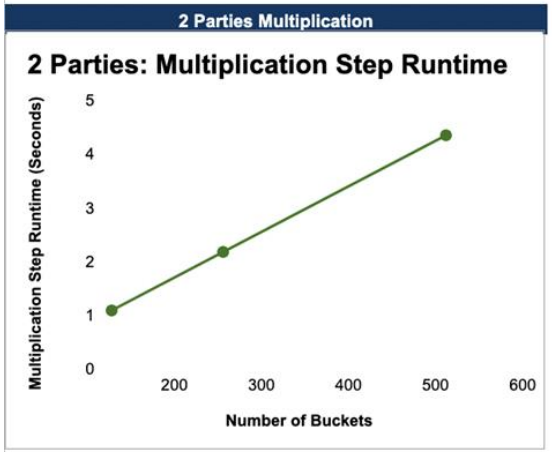


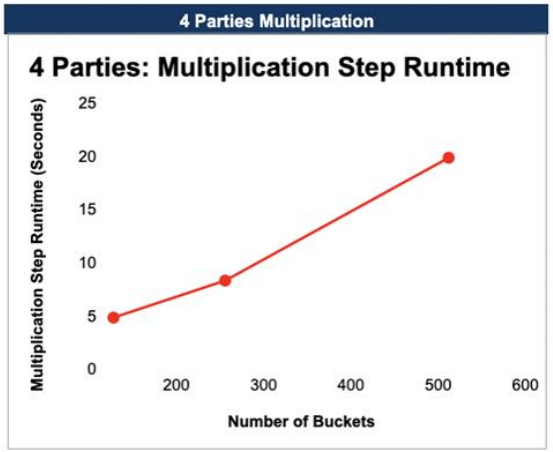
Runtimes

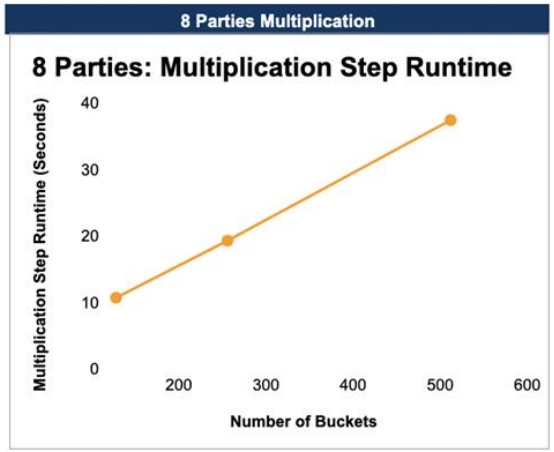

2 Parties Addition
2 Parties: Addition Step Runtime
Addition Step Runtime (Seconds)
0
2
4
6
8
10
200
300
400
500
600
Number of Buckets


4 Parties Addition
4 Parties: Addition Step Runtime
Addition Step Runtime (Seconds)
0
5
10
15
200
300
400
500
600
Number of Buckets


8 Parties Addition
8 Parties: Addition Step Runtime
Addition Step Runtime (Seconds)
0
5
10
15
200
300
400
500
600
Number of Buckets


2 Parties Multiplication
2 Parties: Multiplication Step Runtime
Multiplication Step Runtime (Seconds)
0
1
2
3
4
5
200
300
400
500
600
Number of Buckets

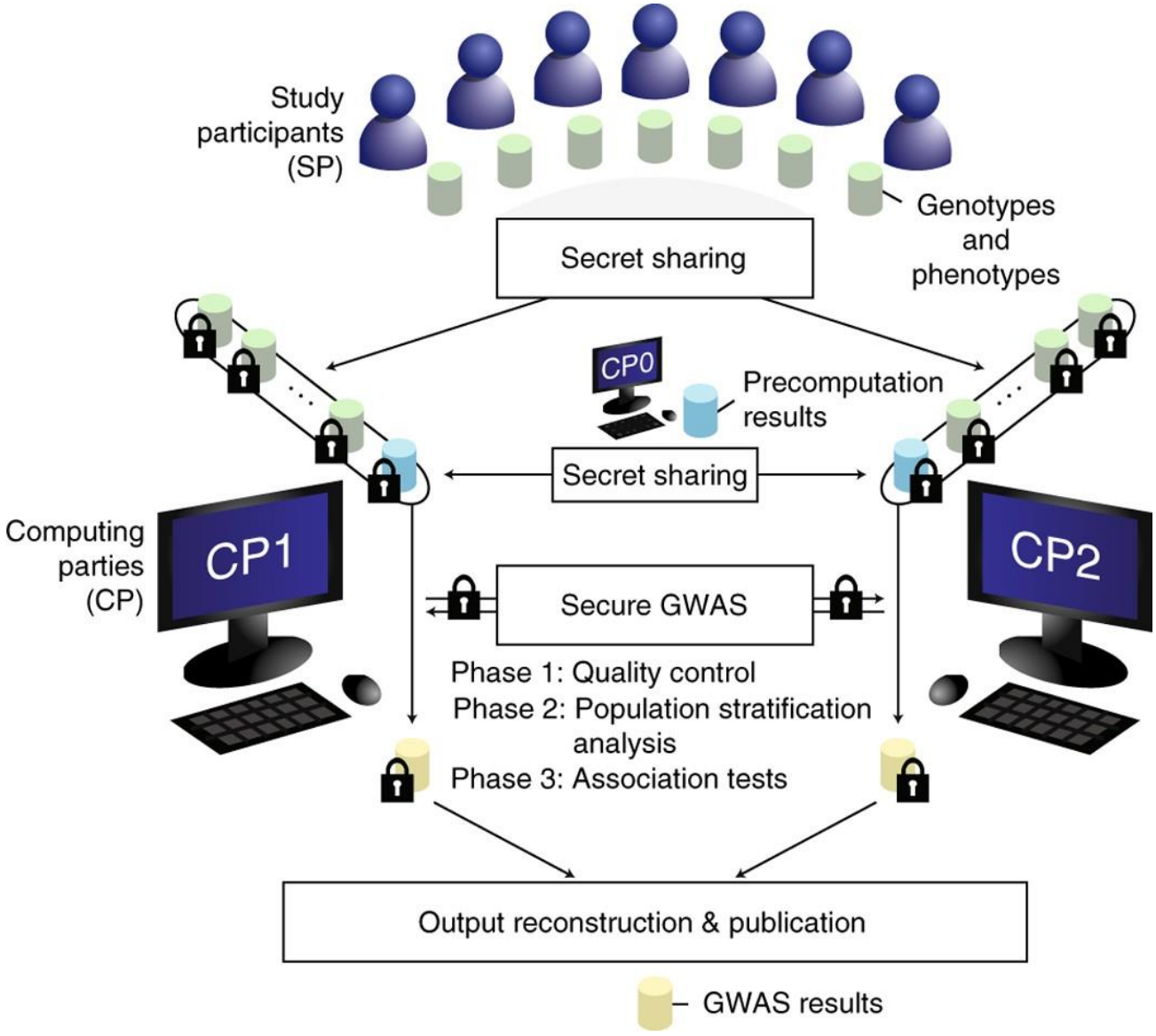

4 Parties Multiplication
4 Parties: Multiplication Step Runtime
Multiplication Step Runtime (Seconds)
0
5
10
15
20
25
200
300
400
500
600
Number of Buckets


8 Parties Multiplication
8 Parties: Multiplication Step Runtime
Multiplication Step Runtime (Seconds)
0
10
20
30
40
200
300
400
500
600
Number of Buckets

Rewriting computations

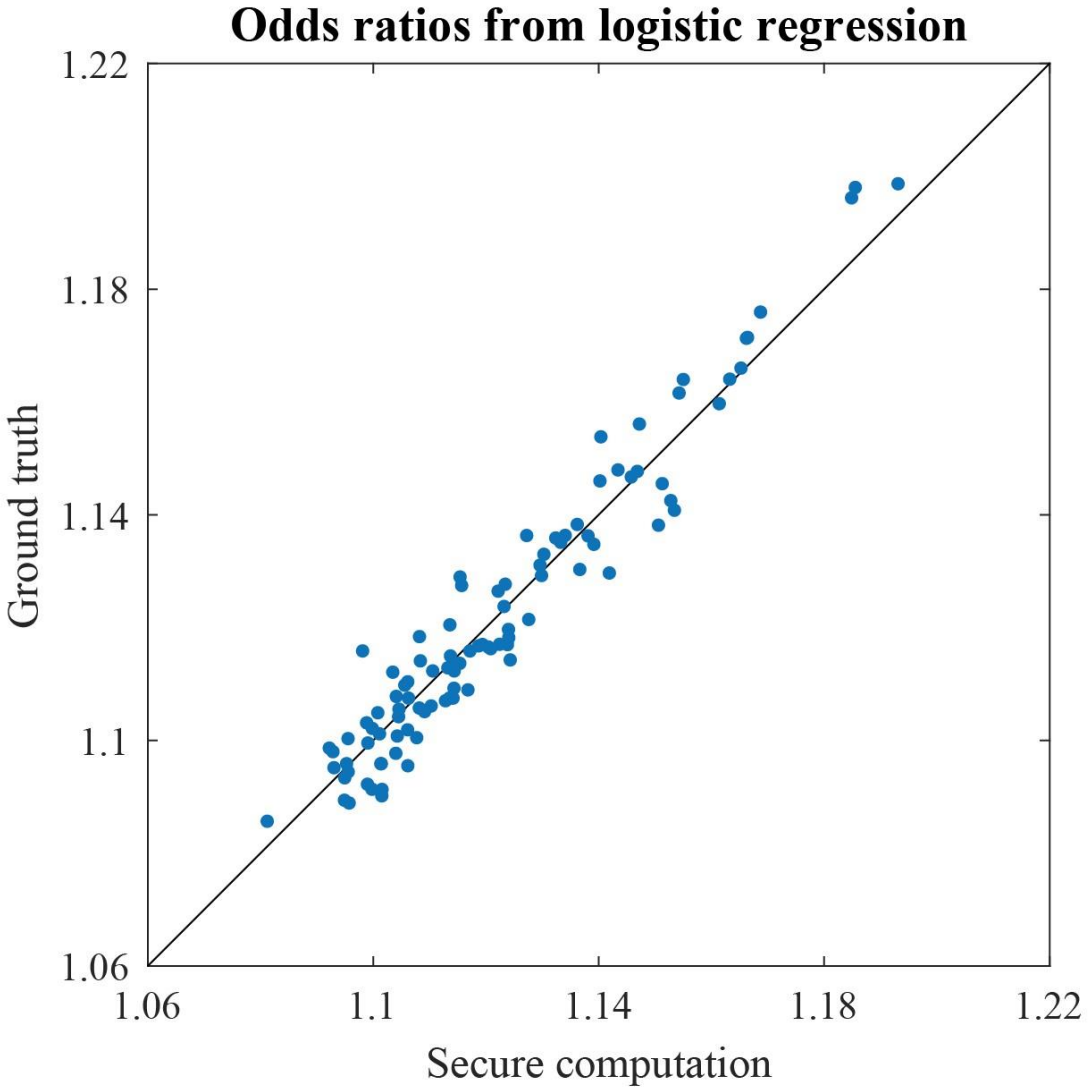
- Sometimes cannot easily rewrite complicated analysis as basically just a summation.
- For example, GWAS and linear regression are substantially more complicated, especially when applying population correction.
- Can make use of MPC and precomputation at important key steps.

GWAS across institutions



Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nature biotechnology. 2018 Jul;36(6):547-51.

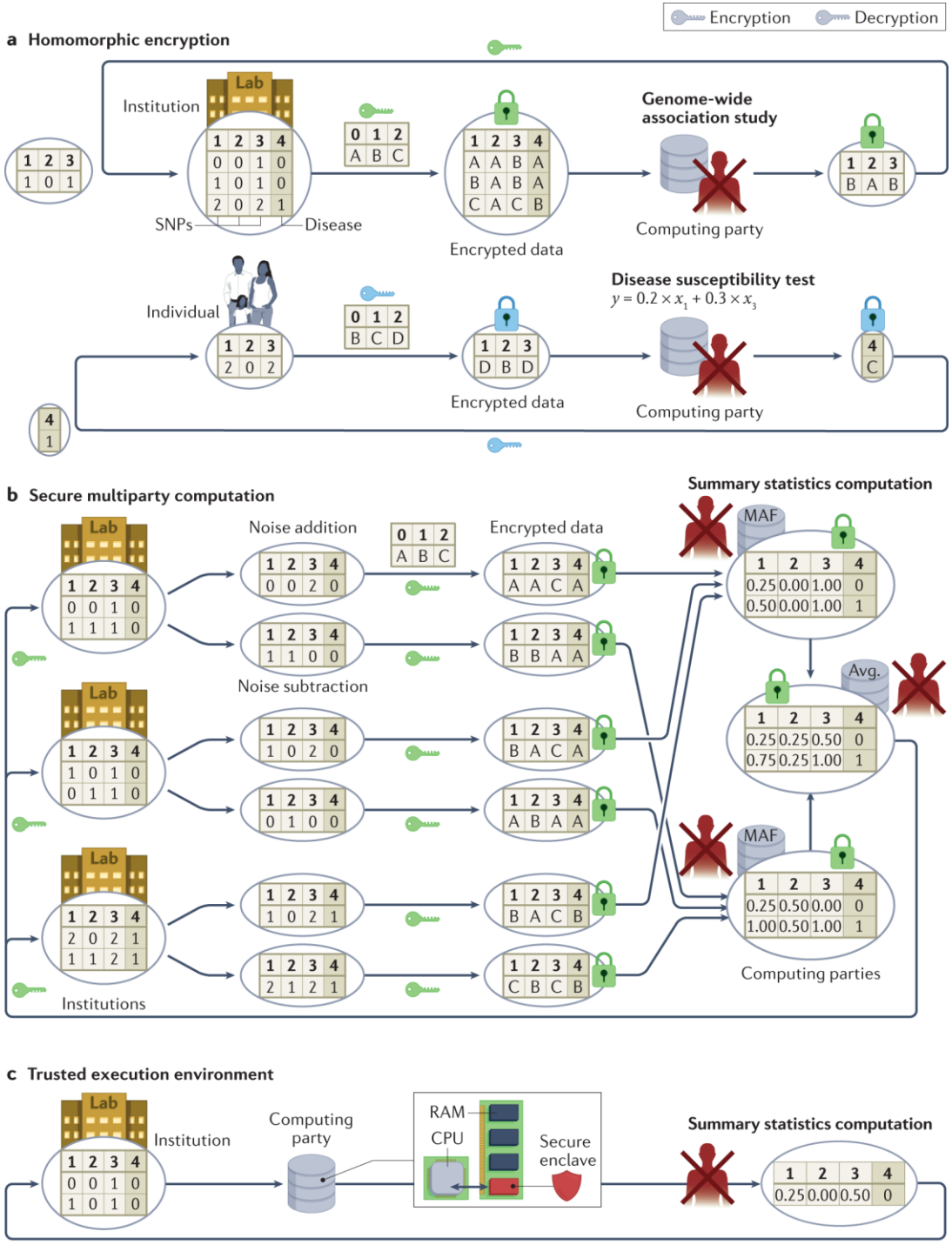
GWAS across institutions (lung cancer)



Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nature biotechnology. 2018 Jul;36(6):547-51.

Trusted execution environment

- An alternate approach to homomorphic encryption and MPC is to use a trusted execution environment (e.g. Intel SGX).
- You need a trusted 3rd party (e.g. Intel), who then locks down a server (e.g. on AWS) cryptographically, so you can be guaranteed that only your code will run and nothing else.



Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nature Reviews Genetics. 2022 Jul;23(7):429-45.

Conclusion

- Privacy questions who should have data.
- Security can be used to implement privacy decisions.
- Problem is complicated because genomic data is shared, and many people have legitimate uses for it.
- Can use combination of de-identification, noise addition, and cryptography to try to implement privacy.