Protecting Patient Privacy in Genomic Analysis

David Wu

based on joint works with:
Gill Bejerano, Bonnie Berger, Johannes A. Birgmeier,
Dan Boneh, Hyunghoon Cho, and Karthik A. Jagadeesh

The Era of “Big Data”

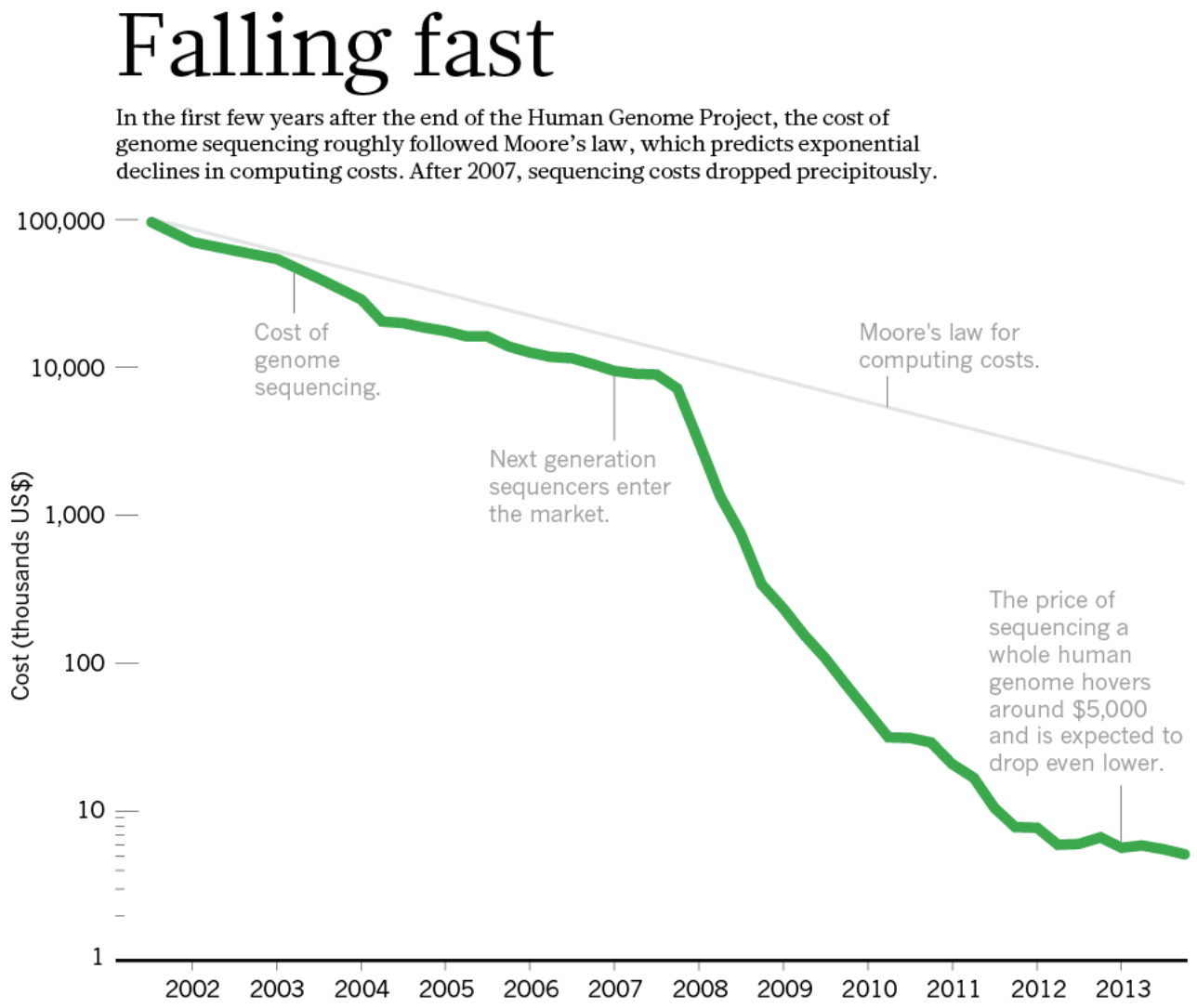
Data Collection and Data Breaches

Entries	Database	Category	Dump Date
358,676,097	Myspace.com	Social Media	2013-06
153,004,874	Adobe.com	Software	2013-10
117,046,470	LinkedIn.com	Social Media	2012
77,039,888	Edmodo.com	Education	2017-05
68,743,269	Neopets.com	Gaming	2013-10
36,397,296	AshleyMadison.com	Dating	2015-07
16,500,334	Zomato.com	Food & Drink	2017-05
6,054,459	Xat.com	Chatroom	2015-11
5,960,654	Adobe.com Common Passwords	Software	2013-10

Database breaches have become the norm rather than the exception…

[Data taken from Vigilante.pw]

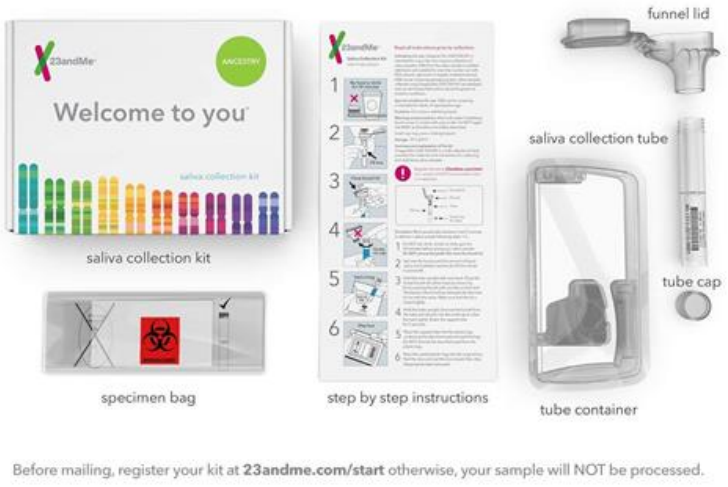
Genomics in the Era of Big Data



Genome sequencing around $1000!

Source: *Nature*, 2014

Genomics in the Era of Big Data



GENETICS

Genealogy Databases Enable Naming Of Anonymous DNA Donors

CAMBRIDGE, MASSACHUSETTS—One afternoon in March last year, Yaniv Erlich sat down at his computer to do an experiment. Before

Privacy concerns have been raised about publicly accessible genome data before. A study 5 years ago showed that individuals

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

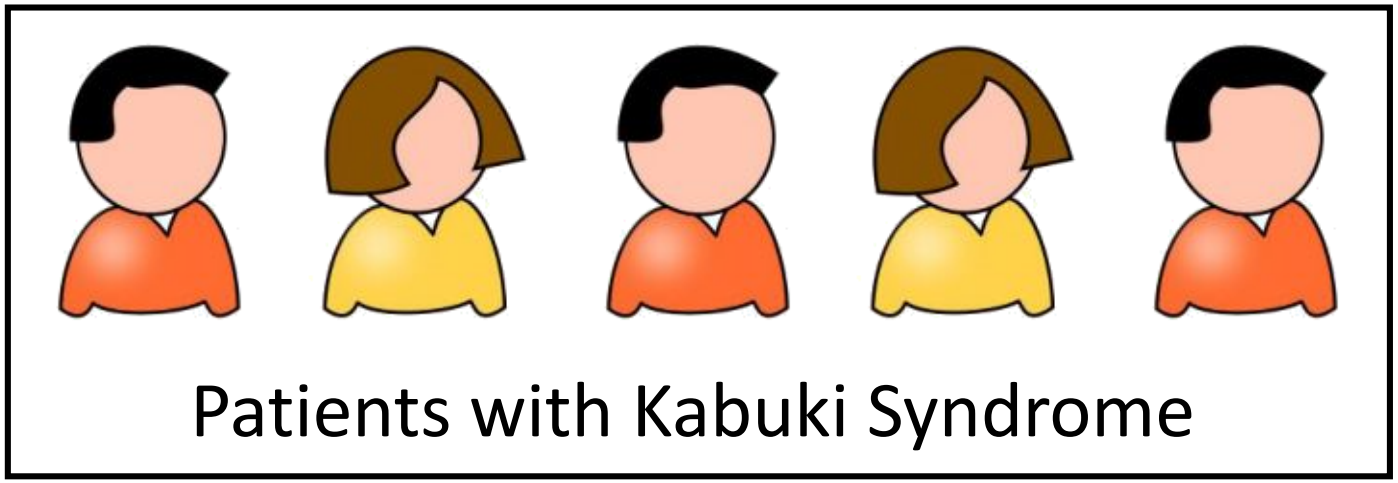
Privacy-Preserving Genomics

Finding a tradeoff between functionality and privacy

Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

What gene causes a specific (rare) disease?

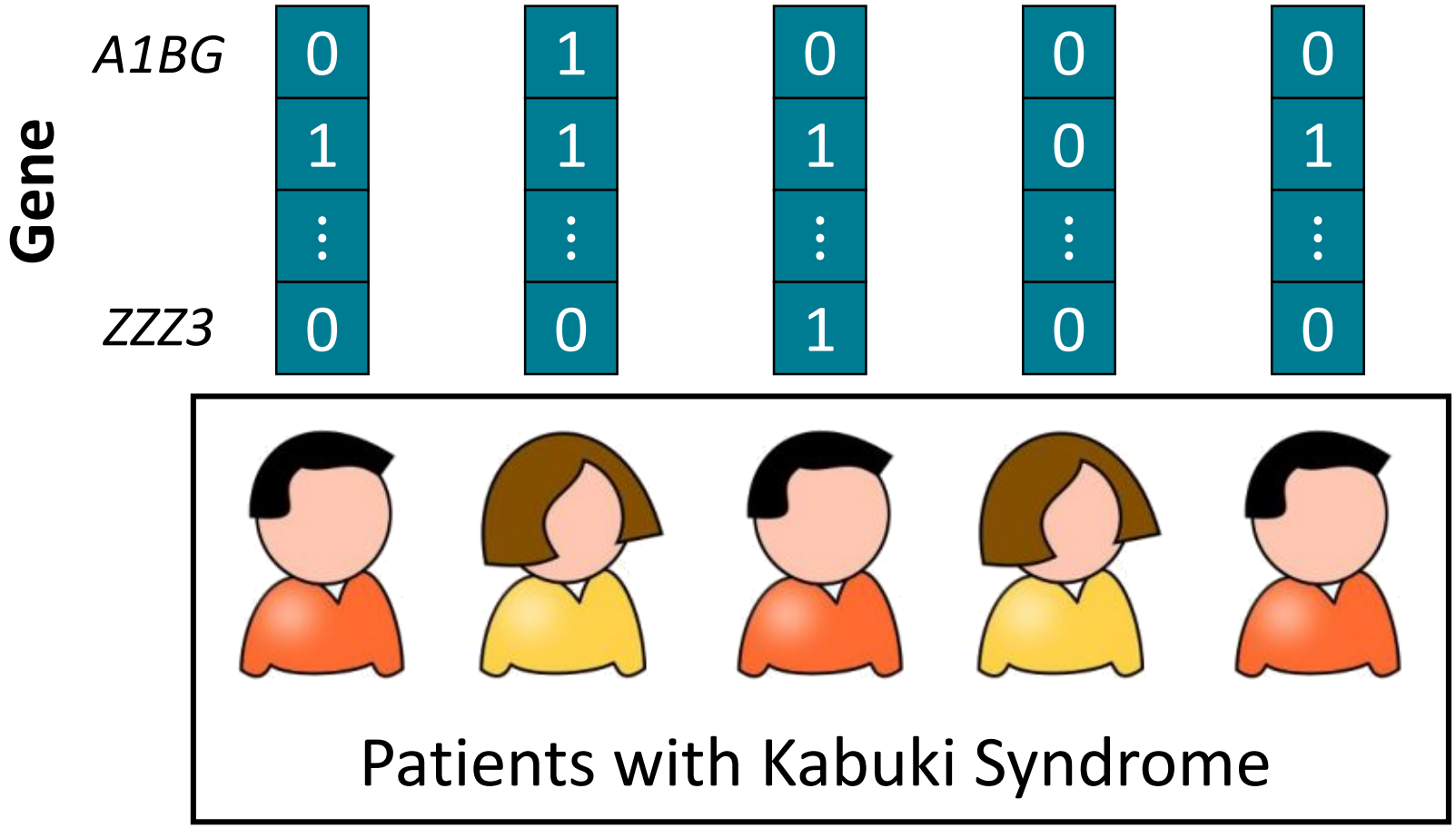


Patients with Kabuki Syndrome

Each patient has a list of 200-400
rare variants over ≈20,000 genes

Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]



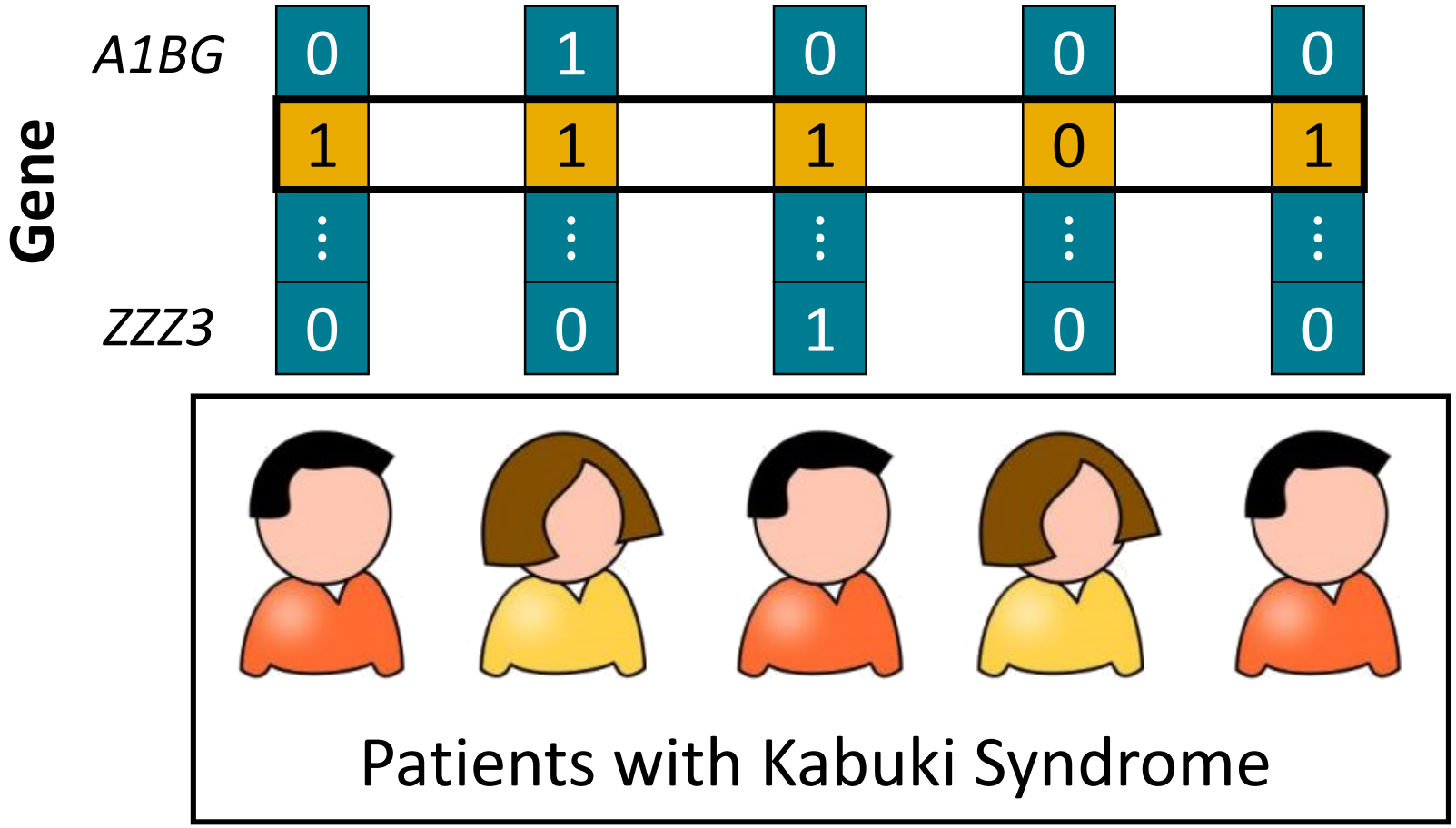
Each patient has a list of 200-400 rare variants over ≈20,000 genes

Each patient has a vector v where $v_i = 1$ if patient has a rare variant in gene i

Goal: Identify gene with most variants across all patients

Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]



Each patient has a list of 200-400 rare variants over ≈20,000 genes

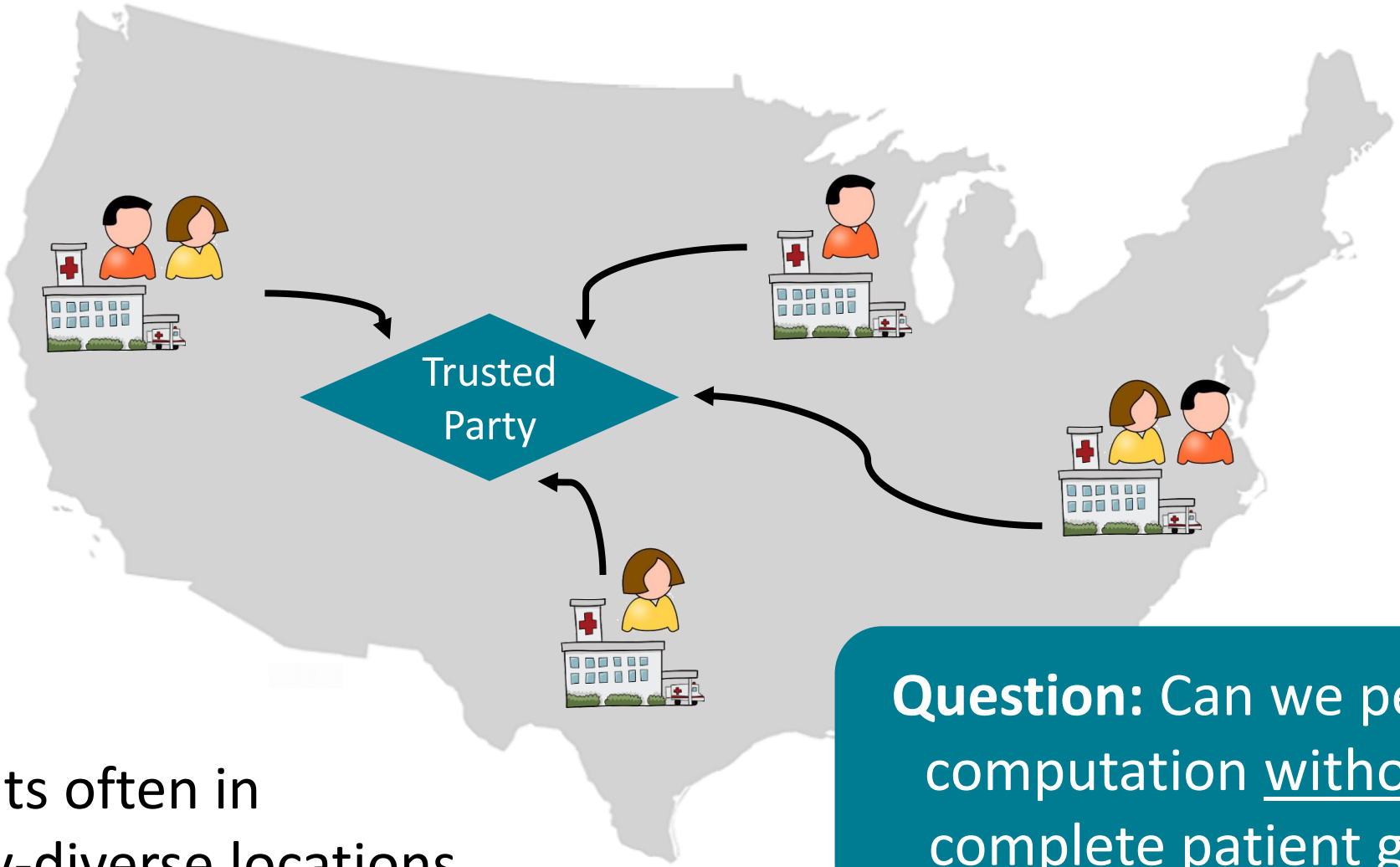
Each patient has a vector v where $v_i = 1$ if patient has a rare variant in gene i

Goal: Identify gene with most variants across all patients

Works well for Mendelian (monogenic) diseases (estimated to affect ≈10% of individuals)

Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

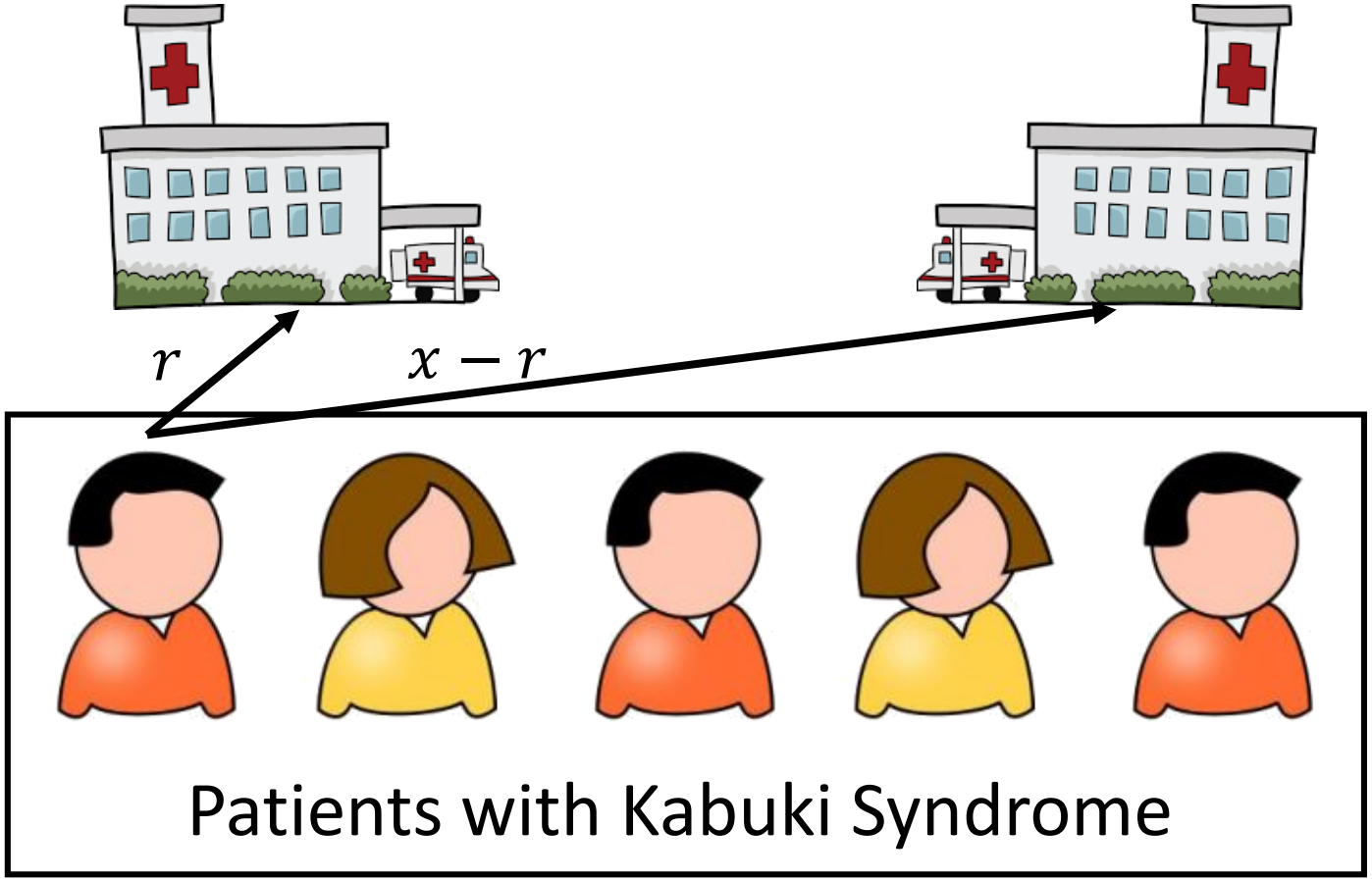


Patients often in geographically-diverse locations

Question: Can we perform this computation without seeing complete patient genomes?

Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

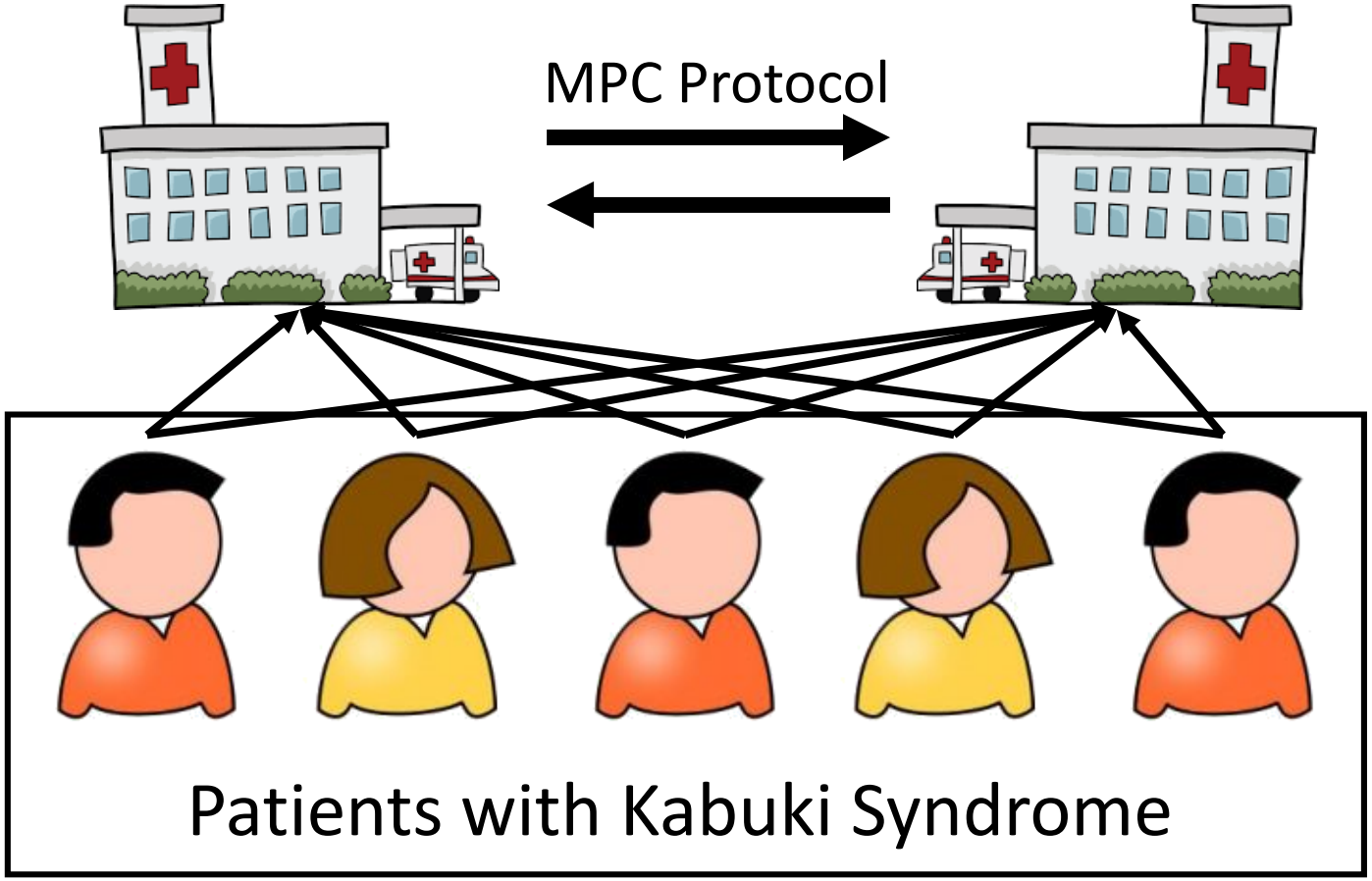


Each patient has a list of 200-400 rare variants over ≈20,000 genes

Patients "secret share" their data with two non-colluding hospitals

Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]



Each patient has a list of 200-400 rare variants over ≈20,000 genes

Hospitals run a multiparty computation (MPC) protocol on pooled inputs

Patients “secret share” their data with two non-colluding hospitals

Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

MPC Protocol

Patients with Kabuki Syndrome

Each patient has a list of 200-400
rare variants over ≈20,000 genes

Top variants (sorted):
KMT2D, COL6A1, FLNB

Known cause of disease

Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

MPC Protocol

Patients with Kabuki Syndrome

Each patient has a list of 200-400 rare variants over ≈20,000 genes

Top variants (sorted):
KMT2D, COL6A1, FLNB

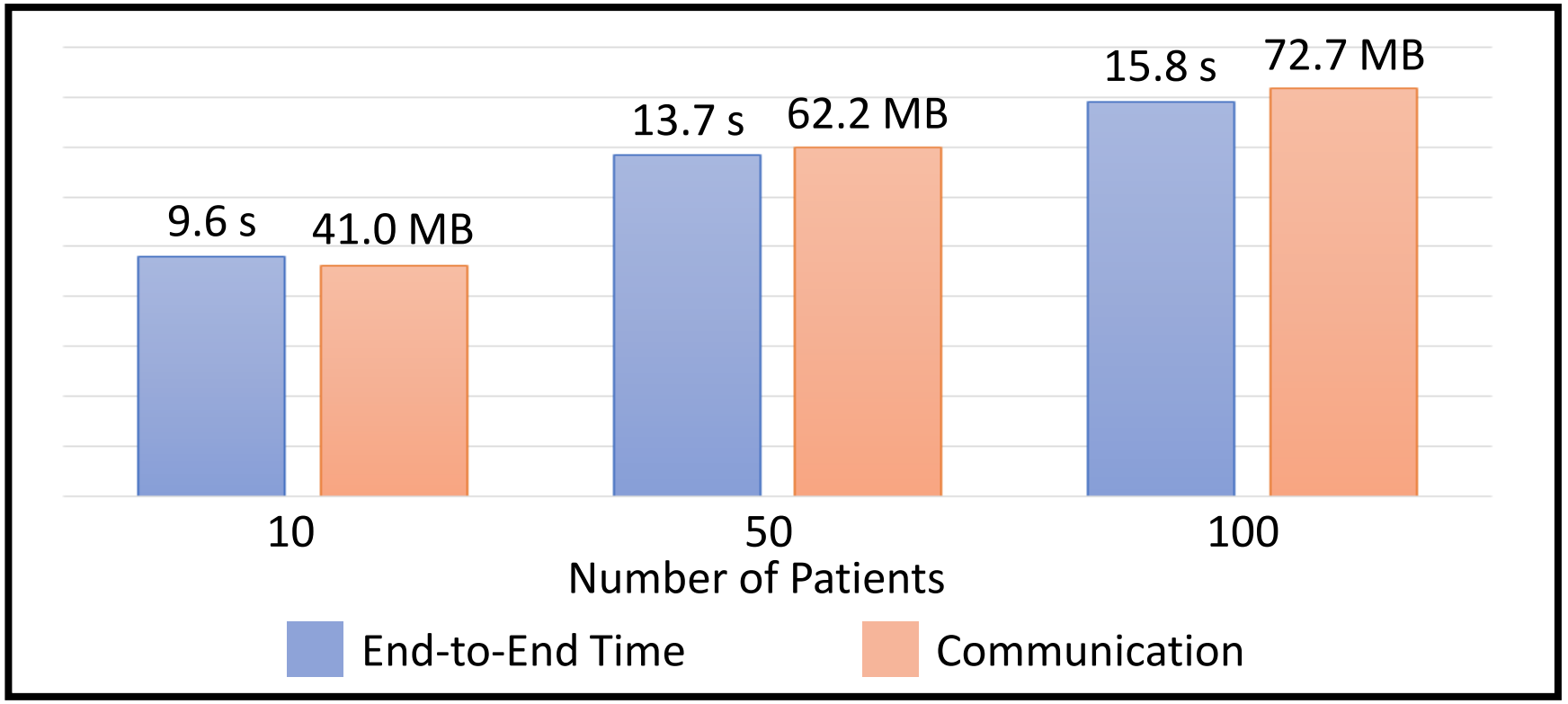
Other variants that the patients possess are kept hidden

Rare Disease Diagnosis

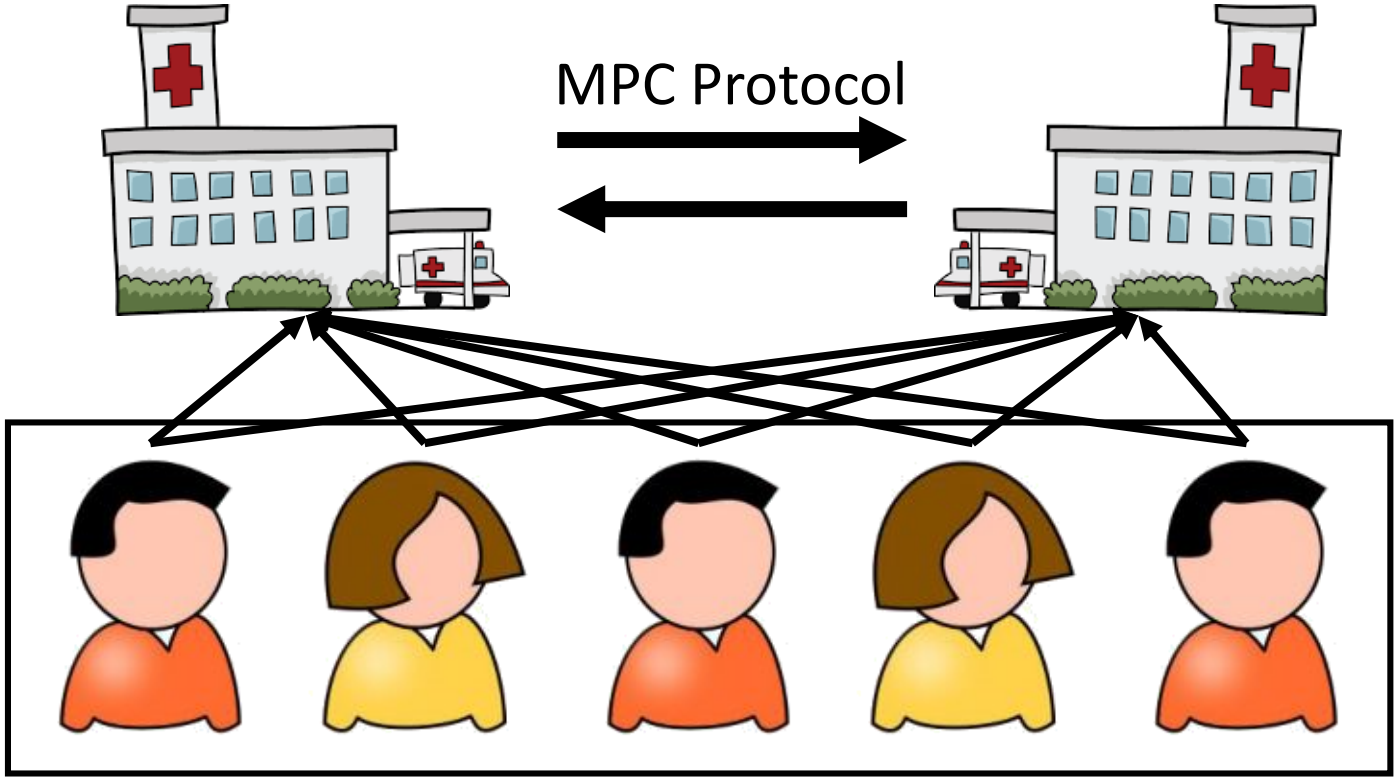
Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

Experimental benchmarks for identifying causal gene in small disease cohort

- Simulated two non-colluding entities with 1 server on East Coast and 1 on West Coast

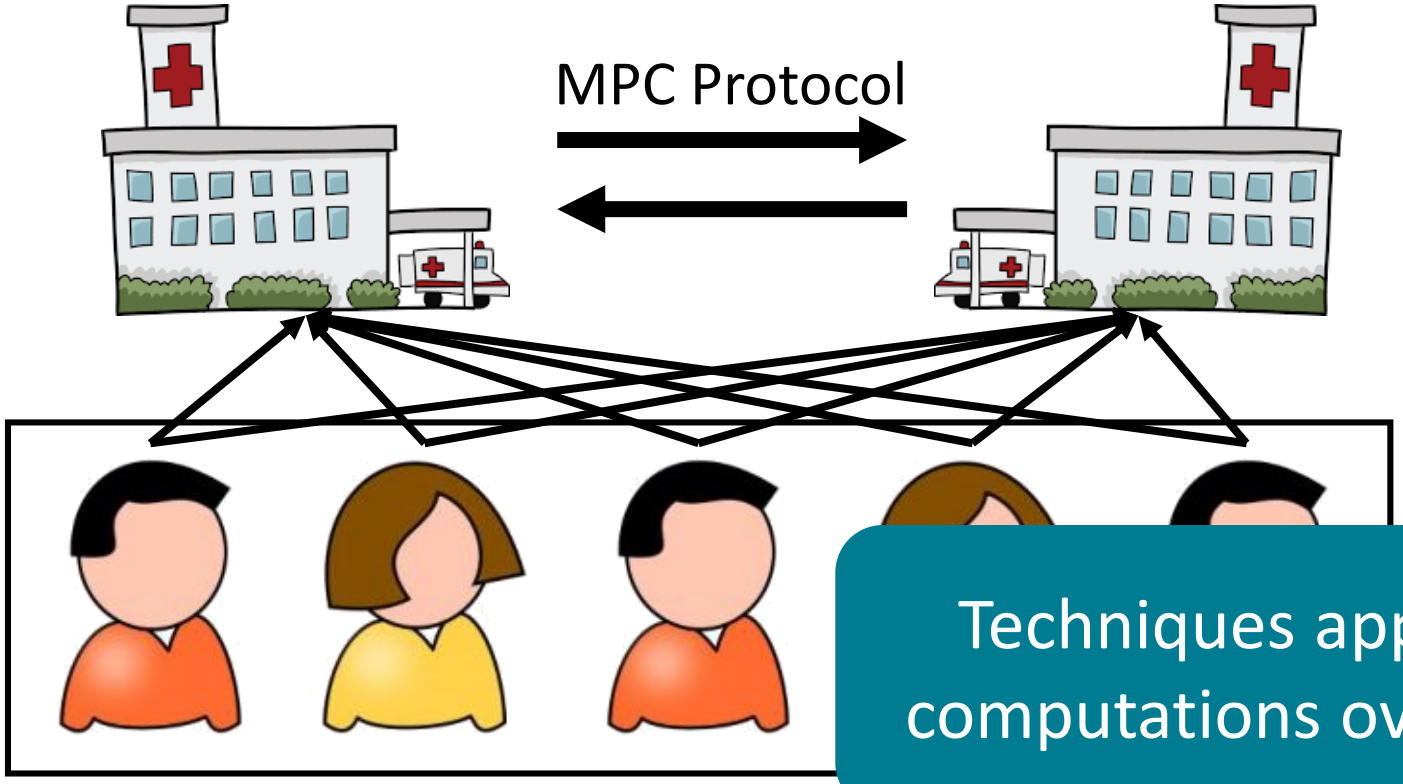


Secure Genome Computation



Modern cryptographic tools enable useful computations while protecting the privacy of individual genomes

Secure Genome Computation



Techniques apply to general computations over private data

Modern cryptographic tools enable useful computations while protecting the privacy of individual genomes

Conclusions

- Privacy and functionality are <u>not</u> inherently incompatible
- Modern cryptographic tools enable computation on private data
- **Question:** What privacy challenges are there in your area, and what kind of cryptographic tools can we use to address them?

Project Website:

`https://crypto.stanford.edu/~dwu4/genomepriv-project.html`

Thank you!