Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 14: Genomic Data Sharing and Privacy Risks

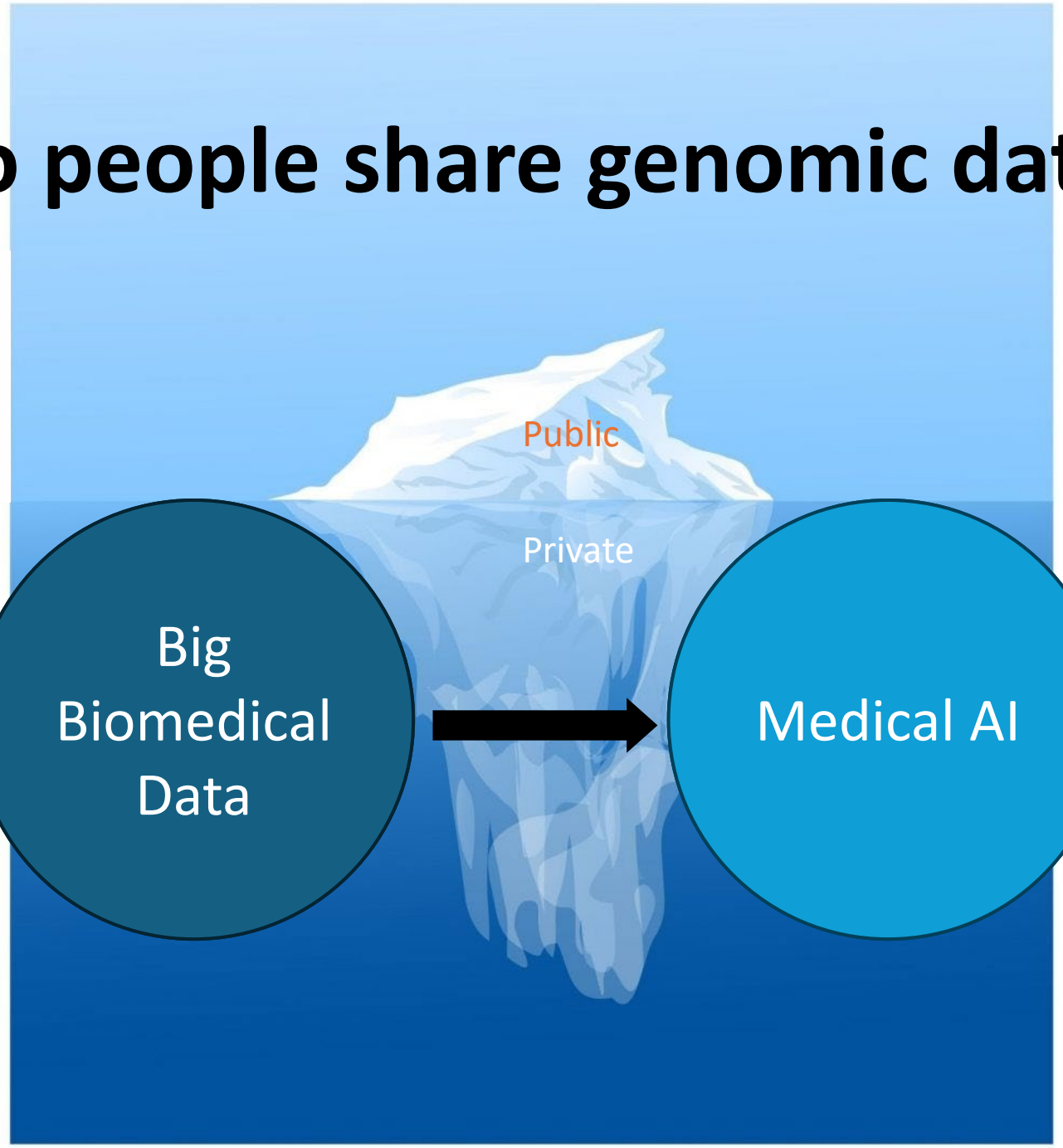
Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)
Assistant Professor of Biomedical Engineering
ShanghaiTech University

November 26, 2025

Learning Objectives of This Lecture

- Know the benefits of sharing genomic data
  - Advancing research and scientific knowledge
  - Help curing diseases
  - Genome-wide association studies
  - Basic research and discovery
  - Reproducibility
  - Genealogical search
- Understand why sharing genomic data is risky and know two types of privacy attacks on genomic data
  - Re-identification attacks
  - Membership/Attribute inference attacks

Why do people share genomic data?



CHINA KADOORIE BIOBANK
中国慢性病前瞻性研究
512,000

BIOBANK JAPAN
270,000

华大基因 BGI
3,000,000
Non-invasive prenatal testing

Vanderbilt BioVU
278,000

uk biobank
500,000

All of Us℠ The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU
1,000,000

23andMe®
12,000,000
(5,000,000 share)

Sharing genomic data is beneficial

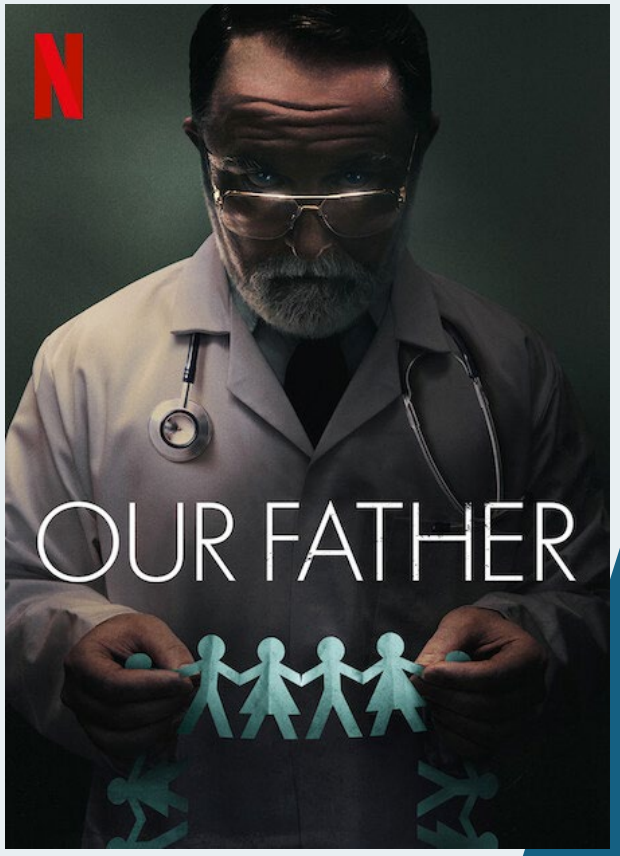
5/14/2013

- Diagnosis of diseases
  - "It's estimated that 55 - 65% of women with the BRCA1 mutation will develop breast cancer before age 70"
  - Accelerate the discovery of associations between genes and diseases

> 1,800 diseases

- Discovering relatives and ancestral origins

> 1,000,000 users



5/11/2022

https://www.nationalbreastcancer.org/what-is-brca
https://www.ebi.ac.uk/gwas
https://www.netflix.com/tudum/our-father-title-page

VANDERBILT UNIVERSITY

4

Advancing research and scientific knowledge

- The foundation of biomedical and healthcare studies relies on the data collected.
- **Genome sequencing technologies** can help identify genetic variations in humans that cause or influence diseases ranging from Huntington disease to cancer.
- It is extremely difficult and expensive to collect all types of genomic data at one site or institution.
- Therefore, data sharing across institutions and labs is essential for data integration.

Responsible Genomic Data Sharing Challenges and Approaches Ch.1

Help curing diseases

- **Rare disease**
  - In the US, a disease is characterized as rare if it affects fewer than 200,000 Americans at any given time. Globally, common definitions include affecting no more than 50 per 100,000 people (EU) or 1 in 2,000 people (Canada).
  - There are 30 million people in the United States and 350 million people in the world currently suffering from a rare disease.
  - There are currently 6000+ rare diseases identified and cataloged in the world.
  - 80% of rare diseases have been determined to have a genetic origin.
  - There is great value in collecting and sharing genetic data on a worldwide scale in the context of rare diseases.
  - By means of pharmacogenomics, one can look at how genetic variation affects the response of a patient to a drug.

Help curing diseases (Cont.)

- **Cancer**
  - The inheritance and accumulation of mutations in the genome is the main cause of many cancer types.
  - The current situation is that such data are confined within the particular hospital database and not shared widely with the research community.
  - The National Cancer Institute Genomic Data Commons (GDC), launched in 2016
  - There are various genetic indicators that predict the probability of effectiveness of immunotherapy treatment
  - The Pediatric Cancer Data Commons (PCDC)

Basic research and discovery

- **International HapMap project**, started in 2002
  - develop a haplotype map of the human genome to provide a resource for researchers to find disease associating genes and their response to drugs.
- **1000 Genomes Project**, first released in 2008
  - 2504 samples
  - finding most genetic variants with frequencies of at least 1% in the populations studied using some of the samples from the HapMap project
- **The Cancer Genome Atlas (TCGA)**, initiated in December 2005
  - to catalog the genomic changes underlying multiple cancer types
  - focused on three different cancer types: brain, lung, and ovarian
  - over $300 million in total funding

Basic research and discovery (cont.)

- **UK Biobank**, established in 2006
  - 500,000 volunteers between the ages of 40 and 69
- **ENCODE, the encyclopedia of DNA elements**, is a public research consortium supported by The National Human Genome Research Institute (NHGRI), started in 2003.
  - to identify the functional elements in the human and mouse genome beyond coding sequences.
- **Genotype-Tissue Expression (GTEx) project**, launched in 2010 by the NIH
  - to investigate how variation in the human genome affects tissue expression.

Reproducibility

- In a poll launched by Nature in 2016, 70% of 1500 scientists claimed that they failed to reproduce at least one other publication's experiment.
- Although access to data is not the only reason causing replication problems, in computational sciences, encouraging access to data will increase reproducibility.

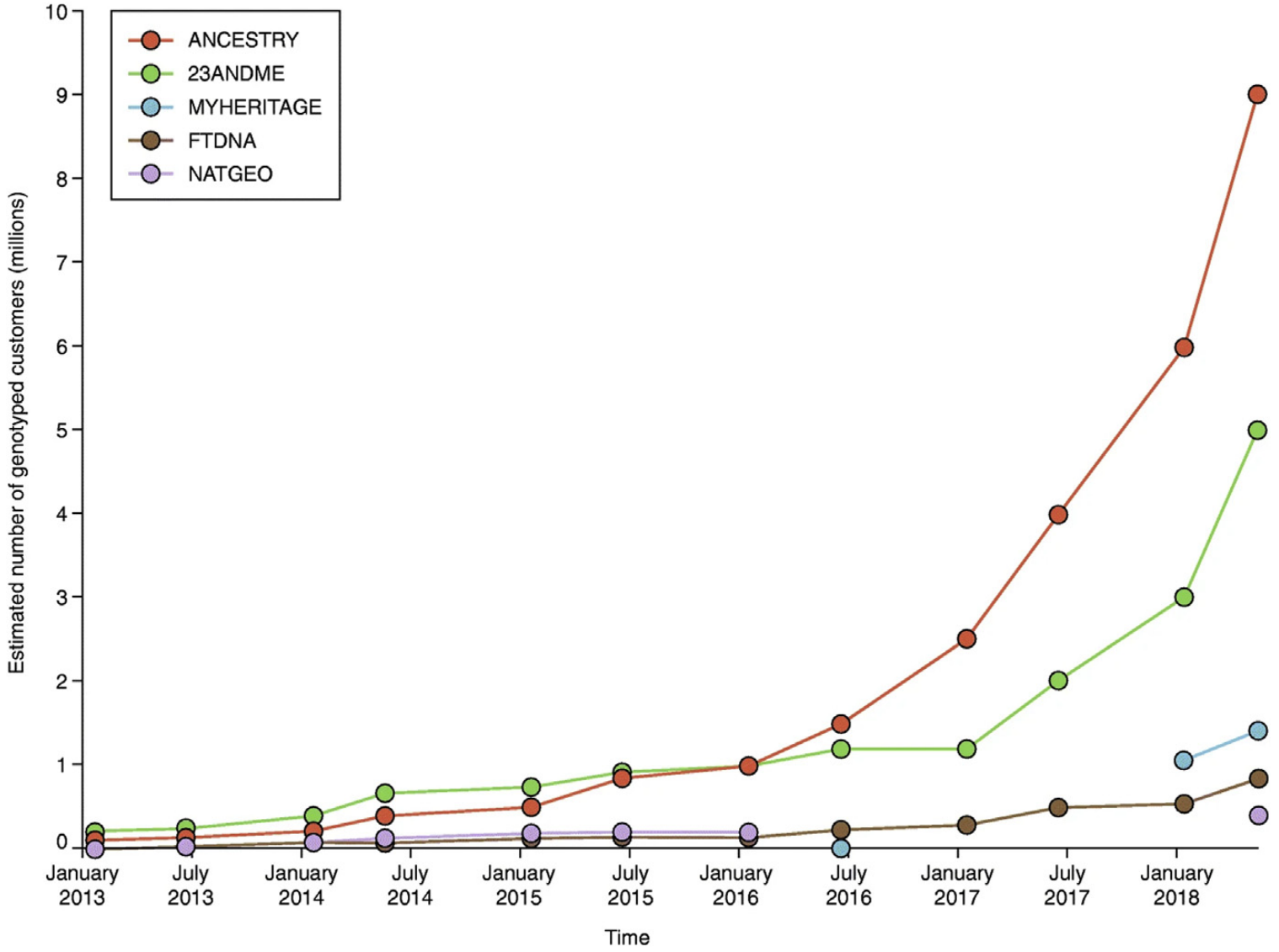
Public's views on genomic data sharing

- **Patients or individuals who are extremely sick** do not seem to think about who owns and controls their genomic data. They are interested instead on sharing their data quickly and with multiple institutions in case one can find novel ways of treating their medical condition.
- On the other hand, **healthy individuals** tend to think more about their data ownership, and the potential for leaking their identifying information.

Direct-to-consumer genetic companies

- **Family Tree DNA**, started in 2000
- **23andMe**, started in 2000
  - sell its kit at $99 starting December of 2012
  - Reaches 1-million customers in 2015.
  - In 2015, it gained FDA approval for a genetic test to predict Bloom syndrome
  - In 2017, it received FDA approval to sell a genetic risk test. The test provided information on 10 health conditions, most notably, Parkinson's disease and late-onset Alzheimer's disease.
  - As of early 2018, they had tested over 3 million customer samples
- **AncestryDNA**, started in 2012
  - Provides ancestry data
  - low price, $59, and access to genealogy tools
  - Reaches 1-million customers served by the end of 2015
  - Reaches an astounding 7 million customers by the beginning of 2018

Responsible Genomic Data Sharing Challenges and Approaches Ch.4



Autosomal DNA database growth

File format

- Variant call format (VCF) is a tab-delimited text file format storing gene sequence variations.
  - It was originally developed to support the 1000 Genomes Project.

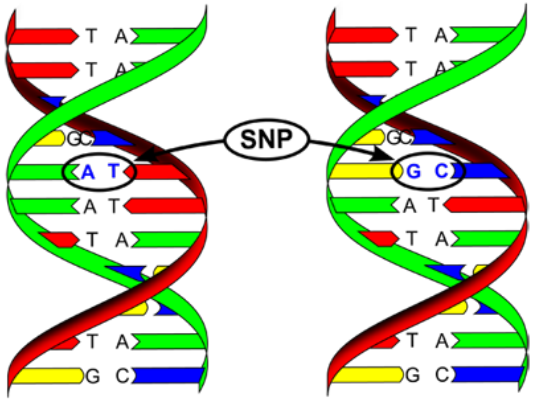
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=20150218
##reference=ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
##source=1000GenomesPhase3Pipeline
##contig=<ID=1,assembly=b37,length=249250621>
##...
##ALT=<ID=DEL,Description="Deletion">
##...
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##...
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth; only low coverage data were counted towards the DP, exome data were not used">
##INFO=<ID=VT,Number=.,Type=String,Description="indicates what type of variant the line represents">
##...

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00097	HG00099
22	16050075	rs587697622	A	G	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=8012;VT=SNP	GT	0|0	0|0	0|0
22	16050115	rs587755077	G	A	100	PASS	AC=32;AF=0.00638978;AN=5008;NS=2504;DP=11468;VT=SNP	GT	0|0	0|0	0|0
22	16050213	rs587654921	C	T	100	PASS	AC=38;AF=0.00758786;AN=5008;NS=2504;DP=15092;VT=SNP	GT	0|0	0|0	0|0
22	16050319	rs587712275	C	T	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=22609;VT=SNP	GT	0|0	0|0	0|0
22	16050527	rs587769434	C	A	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=23591;VT=SNP	GT	0|0	0|0	0|0
22	16050568	rs587638893	C	A	100	PASS	AC=2;AF=0.000399361;AN=5008;NS=2504;DP=21258;VT=SNP	GT	0|0	0|0	0|0
22	16050607	rs587720402	G	A	100	PASS	AC=5;AF=0.000998403;AN=5008;NS=2504;DP=20274;VT=SNP	GT	0|0	0|0	0|0
22	16050627	rs587593704	G	T	100	PASS	AC=2;AF=0.000399361;AN=5008;NS=2504;DP=21022;VT=SNP	GT	0|0	0|0	0|0

A sample VCF file.

DTC format

- The human genetic data formats used by DTC companies are very similar. They are all tab-delimited text files.



Single Nucleotide Polymorphism (SNP)

#Genetic data is provided below as five TAB delimited columns. Each line
#corresponds to a SNP. Column one provides the SNP identifier (rsID where
#possible). Columns two and three contain the chromosome and basepair position
#of the SNP using human reference build 37.1 coordinates. Columns four and five
#contain the two alleles observed at this SNP (genotype). The genotype is reported
#on the forward (+) strand with respect to the human reference.

rsid	chromosome	position	allele1	allele2			
rs587697622	22	16050075	A	A			
rs587755077	22	16050115	G	G			
rs587654921	22	16050213	C	C			
rs587712275	22	16050319	C	C			
rs587769434	22	16050527	C	C			
rs587638893	22	16050568	C	C			
rs587720402	22	16050607	G	G			
rs587593704	22	16050627	G	G			

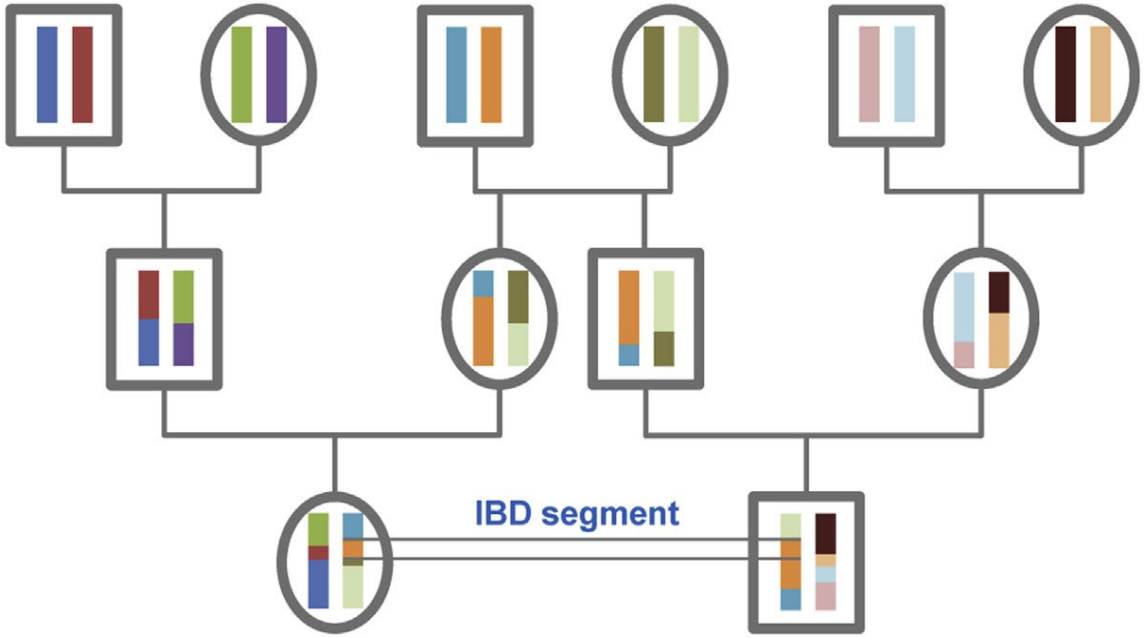
A sample DTC file.

Compressed format

- PLINK format (BED)
  - PLINK is a genetic data toolset developed by Shaun Purcell et al. The goal of PLINK format is to handle large datasets and perform analysis on them in a computationally efficient manner.
- UK Biobank format (BGEN)
  - BGEN is a data format storing either typed or imputed genotype data with the imputed genotype probability (i.e., dosage value). It was developed by Gavin Band and Jonathan Marchini.
- GDS format
  - Xiuwen Zheng et al. propose an array-oriented data format to store genome-wide data, named Genomic Data Structure (GDS).

What is IBD

- Identical by descent (or identity by descent) (IBD) is a biological terminology proposed by Gustave Malecot.
- Originally, it serves as an indication of two homologous alleles descending from a common ancestor.
- Contemporarily, from the genome-scale perspective, IBD is redefined as two homologous chromosome segments being inherited from a common ancestor.



The average and the range of shared IBD segments per relationship

Relationship group (cluster)	Average shared IBD (percentage)	Range of shared IBD (percentage) [112]	Average shared IBD (centiMorgan)	Range of shared IBD (centiMorgan) [126]
Identical twin	100%	Not available	6800.00 [127]	Not available
Parent, child	50%	Variable	3400.00	Variable
Sibling	50%	Variable	2550.00 [125]	2209.00–3384.00
Grand parent, aunt (or uncle), niece (or nephew), grand child	25%	Variable	1700.00	1294.00–2230.00
Great grand parent, great aunt (or uncle), first cousin, great niece (or nephew), great grand child	12.5%	7.31%–13.80%	850.00	486.00–1761.00
second great grand parent, great grand aunt (or uncle), first cousin once removed, great grand niece (or nephew), second great grand child	6.25%	3.30%–8.51%	425.00	131.00–851.00
Third great grand parent, second great grand aunt (or uncle), first cousin twice removed, second cousin, second great grand niece (or nephew), third great grand child	3.125%	2.85%–5.04%	212.50	47.00–517.00

Genealogical search

- Genealogy search involves constructing a family tree, locating relatives, and sometimes providing historical records to the customer

Cousin relationship	Probability of having detectable shared DNA segment			
		In practice [129]		
	In Theory [128]	By 23andMe	By AncestryDNA	By FamilyTreeDNA
First cousin	100.00%	100.00%	100.00%	100.00%
Second cousin	100.00%	100.00%	100.00%	99.00%
Third cousin	97.70%	89.70%	98.00%	90.00%
Fourth cousin	69.30%	45.90%	71.00%	50.00%
Fifth cousin	30.20%	14.90%	32.00%	10.00%
Sixth cousin	10.10%	4.10%	11.00%	2.00%

Genomic Data Sharing

- DNA? Who Cares?

- DNA and the Quasi-Identifier Dilemma

*omics Data is High-Dimensional!

- AG AT CA AA CA TG GC AA TA AT CA GT AG GG AT AT CC AC AG TT AA TA CT CC AG
  CA GC AG AT AA CA GG TA GG GC GA AA GA GA GC CA TA CT CC CC CC CA TT AG TT
  GT AA GG GG AT AC CA GC GG AC AC AC AG TT TC AA CC AC AT CG TG AC AT GG TC
  AT CC GC GA TC AC CT CC CG TG TT TT GC TC TC AC AC AT AT GA CA TG CT TA AC
  CT AG TC CC AC TT CA AA CG TG TA AA TT GA TA TC CT AC AT CA CT CA CC TC AG
  TC TA CA CC CG TG TA TT AG GT CC TT TA TG TA CC GA GG CC GG GG AG TT GG TT
  CA CG TT CC CC AA CG CC TC AT TG AT AT AT CC CC GC CA CA TC AA GT CT GG CT
  CC AC CT TA GT TA AC [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] TT CC CA AG GG AG CC AC
  AT AG GG GT GG TA GG GA AG TT TT TA CG AC AC GT TG TT CG AT CG GC AC CC AT
  CT AC TG AT TT TT AG CT GA CG CA TT GC TG CG AG CG GG TG TC AA AC TA TG TC
  CG GT GT CG AC AA [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] TA CA TA AC AC AC AG
  TC TG AC TG TA GT AC GT TG TG AA GG TG CG TG CT AG TG CG GG CC AA CC GG AG
  GA GG GT CG TC TG CC GT AC TT CC AG CT AT AT TG TT AG GG TA AA AG AA GT CA
  AC GC CC TT TC GA GC GT GG AT GT CG TC AC AT GA AC AT CT GA TC CA GA CA GA
  CA AT AT TG AC GG CC CA AT GT CT TT GT GT CT AG CT CG CA GT AC CC CT GA CG
  CC GT GC TA CC CC AG TG GG GG TA GA TT CG AC CG CG GG TC AG TA GC TC CT GA
  AG GG AC TA AT TA CT TC TG TT GT GT GC GT CC CT TC GC TA GG AC GA AT CT TG
  GA TG GG AA AA TA AT CT CA AA AC CC TC AT CG AG GG TG CC GC GT AG AC CT TC
  CA AG GC GA TA TG AC TA AA CG AA GA CA GA TA TG AT AT CA CC AC AC GG TT TA
  GG GC CG CC TG TA TG TT GT CC CT AC TT CC CT TA CA AC TG CG GG TC AT GA CC

There's only one person who matches on 500 SNPs

Adapted from Bradley Malin's slides

Associations

Age	Race	Sex	Clinical Phenotype	Drug Administered	Adverse Reaction?
42	White	M	Deep Vein Thrombosis Diabetes Type II	Warfarin 7.1mg	No
12	White	M	Pulmonary Embolism Pneumonia	Warfarin 8.2mg	Yes
65	White	F	Deep Vein Thrombosis Stroke	Warfarin 4.8mg	No
58	Black	F	Pulmonary Embolism Broken Arm	Warfarin 5.2mg	No
32	Asian	M	Pulmonary Embolism Dementia	Warfarin 7.8mg	No
56	White	F	Deep Vein Thrombosis Diabetes Type II	Warfarin 4.5mg	No
23	Black	M	Deep Vein Thrombosis HIV-Positive	Warfarin 7.2mg	Yes
37	Asian	F	Blood Clot Hypercholesterolemia	Warfarin 5.7mg	No
19	White	F	Blood Clot Hyperlipidemia	Warfarin 6.3mg	No
24	Black	F	Blood Clot Shortness of Breath	Warfarin 7.4mg	Yes

TUESDAY, JUNE 20, 2006

THE TENNESSEAN

SECTIONS 3

VU to put patient DNA in vast research pool

Blood samples included unless people opt out

By CLAUDIA PINTO
Staff Writer

DNA from as many as 400,000 people will be fed over five years into a database at Vanderbilt University Medical Center under a $5 million research program expected to launch in the fall.

Patients at the hospital and its clinics will have the option to call a hot line and opt out.

The data will be extracted from blood that would otherwise be thrown out, from lab tests or other uses. Researchers primarily will use the data to conduct research on how to eliminate adverse drug reactions, which kill 100,000 people nationally each year.

"We find that a one-size-fits-all approach to therapy or diagnosis is often inadequate," said Dr. Jeff Balser, Vanderbilt's associate vice chancellor for research. "A good example might be can-

AJMG AMERICAN JOURNAL OF medical genetics PART A

Research Article

Two large-scale surveys on community attitudes toward an opt-out biobank†

Kyle B. Brothers, Daniel R. Morrison, Ellen W. Clayton

First published: 07 November 2011 | https://doi.org/10.1002/ajmg.a.34304 | Citations: 52

▸ Clin Transl Sci. 2010 Feb 24;3(1):42–48. doi: 10.1111/j.1752-8062.2010.00175.x

Principles of Human Subjects Protections Applied in an Opt-Out, De-identified Biobank

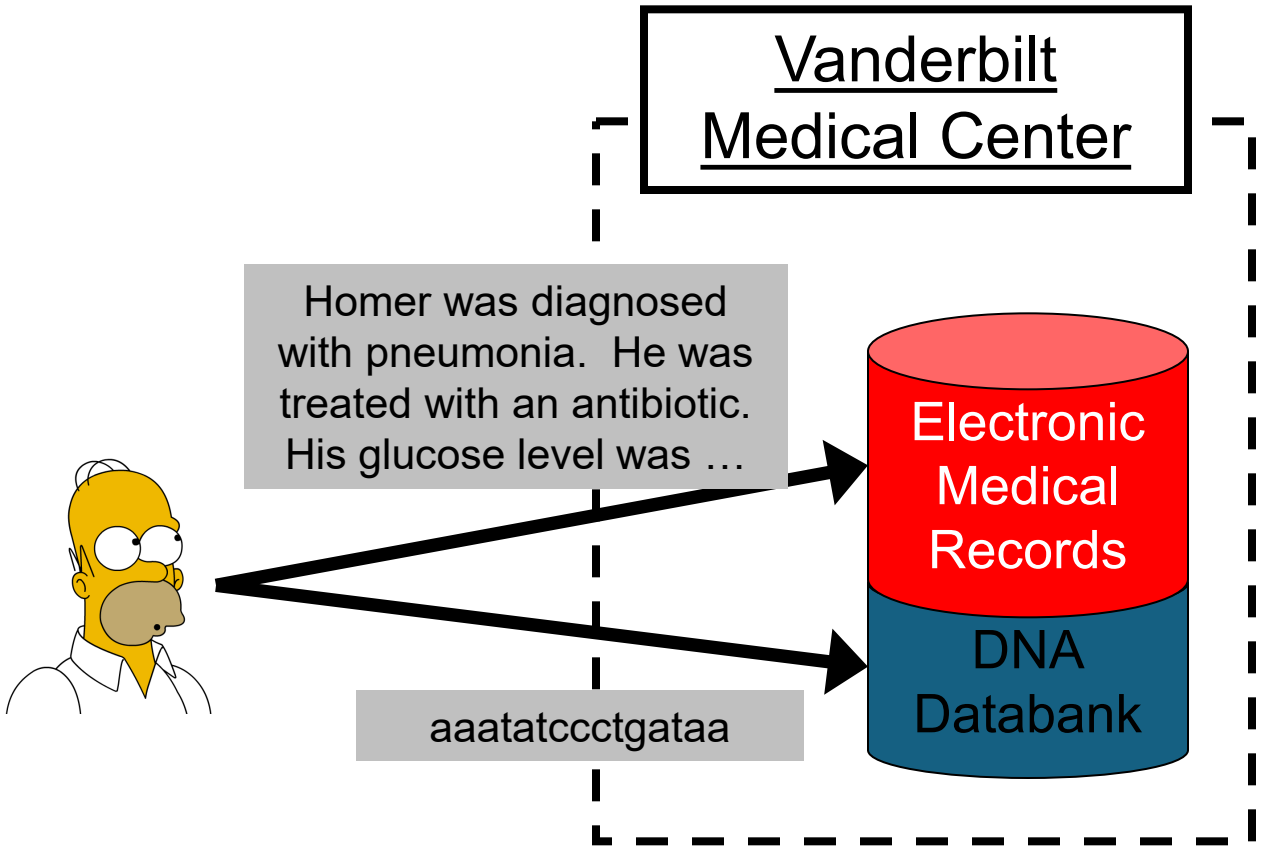
Jill Pulley [1], Ellen Clayton [2], Gordon R Bernard [3], Dan M Roden [4], Daniel R Masys [5]

January 29, 2015

Consent process for BioVU participation updated

As of last week, Vanderbilt has updated the process used to facilitate patient participation in BioVU, the Medical Center's DNA repository.

Personalizing Medicine



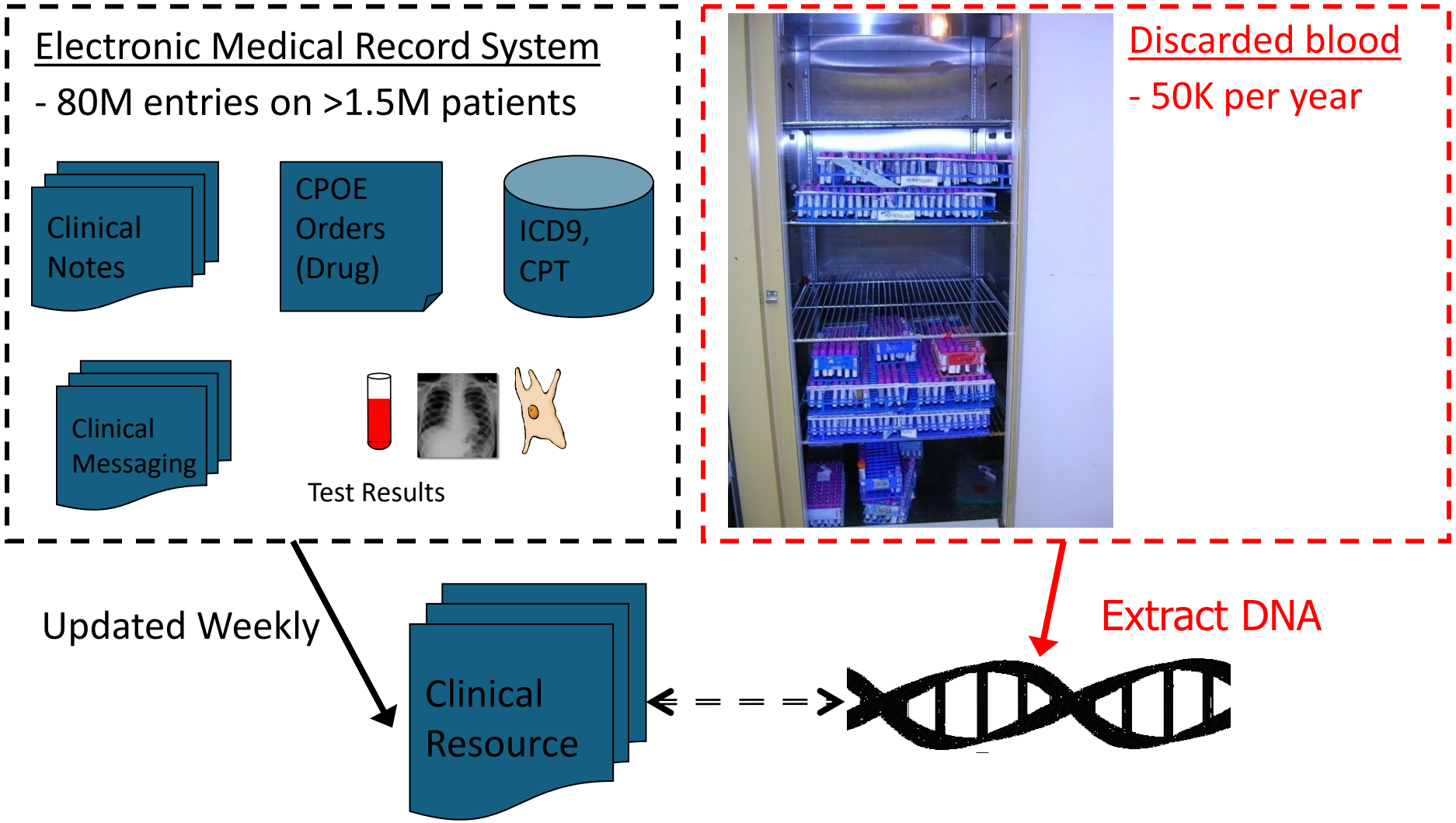
Clinical Goal:

Tailor Homer's treatment to his DNA

Scientific Challenge:

Need to learn associations between DNA clinical features

Information Integration

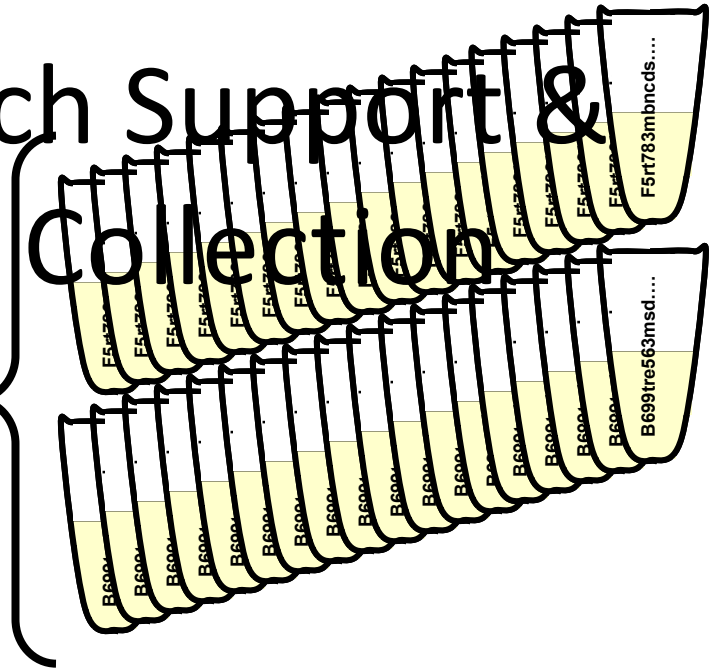


Associations

Age	Race	Sex	Clinical Phenotype	Drug Administered	DNA	Adverse Reaction?
42	White	M	Deep Vein Thrombosis Diabetes Type II	Warfarin 7.1mg	aaca	No
12	White	M	Pulmonary Embolism Pneumonia	Warfarin 8.2mg	cggt	Yes
65	White	F	Deep Vein Thrombosis Stroke	Warfarin 4.8mg	aagt	No
58	Black	F	Pulmonary Embolism Broken Arm	Warfarin 5.2mg	cgca	No
32	Asian	M	Pulmonary Embolism Dementia	Warfarin 7.8mg	agga	No
56	White	F	Deep Vein Thrombosis Diabetes Type II	Warfarin 4.5mg	agct	No
23	Black	M	Deep Vein Thrombosis HIV-Positive	Warfarin 7.2mg	aact	Yes
37	Asian	F	Blood Clot Hypercholesterolemia	Warfarin 5.7mg	cact	No
19	White	F	Blood Clot Hyperlipidemia	Warfarin 6.3mg	cggt	No
24	Black	F	Blood Clot Shortness of Breath	Warfarin 7.4mg	aggt	Yes

Research Support & Data Collection

Genotyping, genotype-phenotype relations



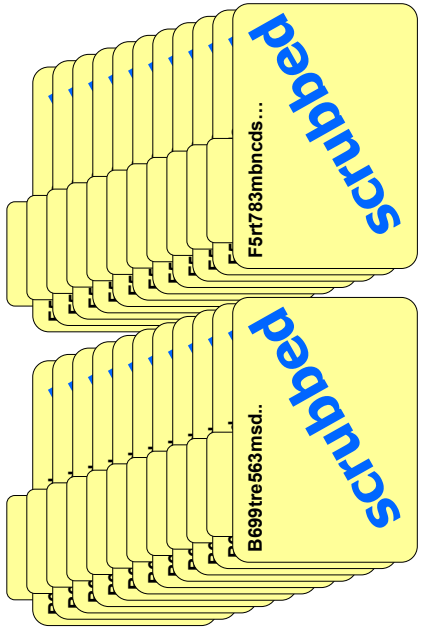
cases

controls

Investigator query

cases

controls



Data analysis

Genetic Association Approaches

Traditional Model

- Disease specific
- Defined population
- Investigator driven
- Specific hypothesis
- Smaller populations
- Research derived samples and information
- Candidate genes specific to disease
- Subjects can be recontacted
- **Hypothesis testing**

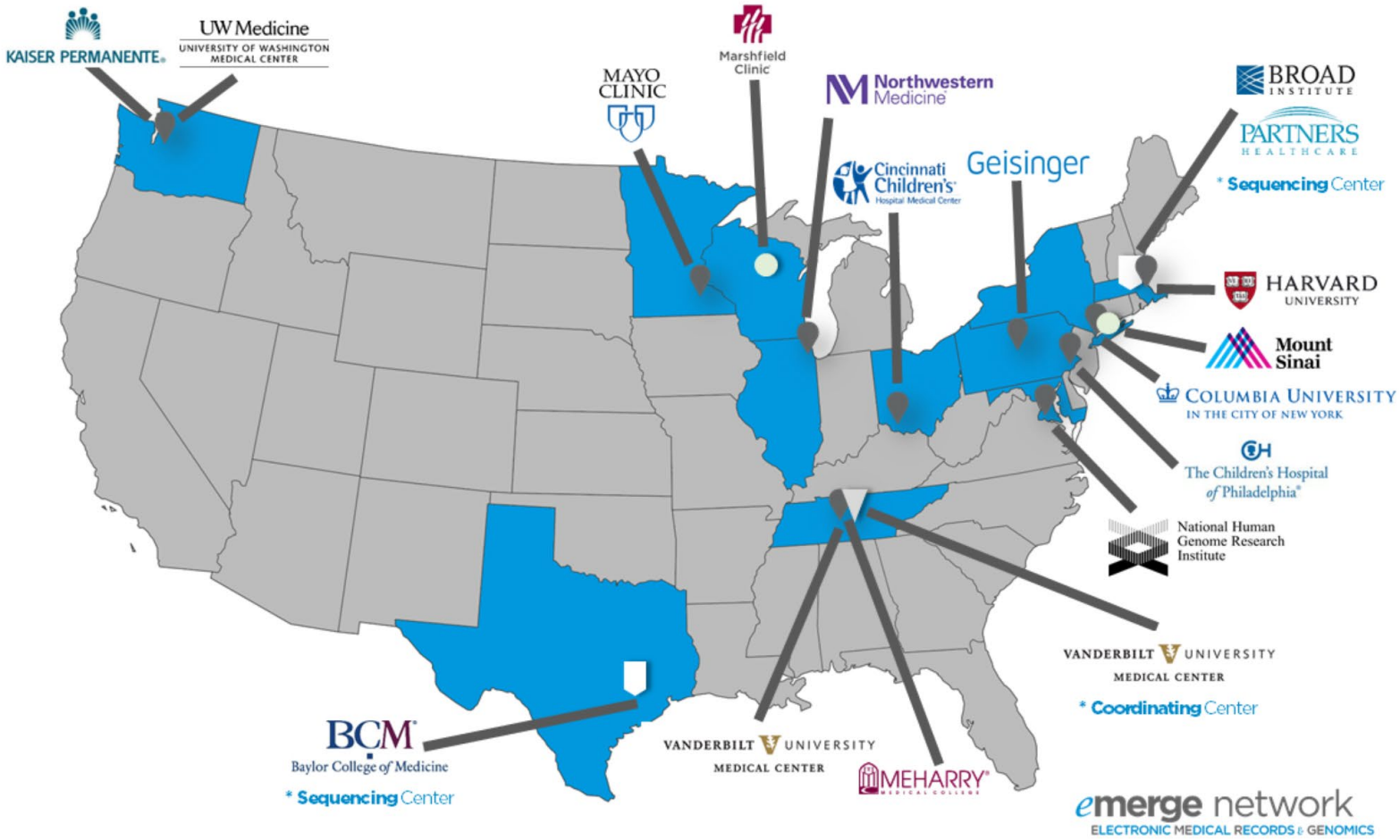
Vanderbilt DNA Databank Model

- Any disease
- All comers
- Institutionally managed
- Multiple/dynamic hypotheses
- Large scale
- Clinically derived samples and information
- Genome scan, shared genotyping database
- De-identified
- Hypothesis generation

Technology + Policy

- Databank access restricted to Vanderbilt employees
  - it is NOT a public resource
- Databank users sign Data Use Agreement that prohibits use of data for re-identification
- Access approved on project-specific basis by Operations Advisory Board (OAB) and Institutional Review Board
- Project-specific user ID and password; all data access logged and audited by OAB

The eMERGE Network brings together researchers with a wide range of expertise in genomics, statistics, ethics, informatics, and clinical medicine from leading medical research institutions across the country. Each center participating in the consortium is uniquely situated to provide critical resources to this highly collaborative and productive network. Each site combines a biobank or study cohort with extensive genomic data and access to clinical data derived from electronic medical records. Sites are geographically dispersed and have diverse patient populations, including two sites focusing specifically on pediatrics. Member sites include:



Participant Sites: Project Overview

Data Sharing Policies

- Feb '03: National Institutes of Health Data Sharing Policy
  - ***"data should be made as widely & freely available as possible"***
  - ***researchers who receive >= $500,000*** *must develop a data sharing plan* ***or describe why data sharing is not possible***
  - Derived data must be shared in a manner that is devoid of "identifiable information"
- Aug '07: NIH Supported Genome-Wide Association Studies Policy
  - Researchers who received > $0 for GWAS
- Aug '14: NIH Genomic Data Sharing Policy
  - For any genomic sequencing data
- Funding condition: contribute de-identified genomic and EMR-derived phenotype data to database of genotypes and phenotypes (dbGAP) at NCBI, NIH

Pharmas Plummet as US NIH Bans CN from Accessing Genetic & Disease Databases

Close

2025/04/07 09:55 CST | 39 60 46 A⁻ A⁺ STOCK INFO SHORT SELL

Enhanced security measures regarding data access management have been announced on the website of the US National Institutes of Health (NIH) Office of the Director, according to reports from several Chinese media outlets.

Starting last Friday (4th), institutions from China, Russia, Iran, and other countries of concern are banned from accessing NIH's controlled access data repositories and related data, which, as indicated by the reports, include key data platforms such as dbGaP and AnVIL.

促进会 EN

数据"断供"：中国医药创新遭遇"卡脖子"新战场

产业资讯 药渡 2025-04-09 76

4月2日，美国国立卫生研究院(NIH)发布文件——《实施更新：增强NIH受控访问数据的安全措施》：**自2025年4月4日起，NIH禁止中国、俄罗斯、伊朗等“受关注国家”的机构访问其受控数据存储库，包括人类基因型-表型数据库(dbGaP)、基因数据分析云平台AnVIL等。**该政策基于美国司法部2024年2月28日发布的第14117号行政命令，旨在限制敏感数据交易，最终规则于2025年4月8日正式生效。

Implementation Update: Enhancing Security Measures for NIH Controlled-Access Data

Notice Number:
NOT-OD-25-083

Key Dates

Release Date: April 2, 2025

Genomic Data Privacy Risks

- DNA? Who Cares?

- DNA and the Quasi-Identifier Dilemma
  - Genome Basics
  - Identity Disclosure (Re-identification)
  - Membership/Attribute Inference

Some Basic Genetics

- SNPs usually have two alleles (major E & minor *e*)
  - Usually don't care which parent the SNP is from

- Barring rare events: $SNP_i \in \{EE, Ee, ee\}$

- Probabilities of these events

$$\{\pi_{i1}, \pi_{i2}, \pi_{i3}\} \rightarrow \{\pi_{i(EE)}, \pi_{i(Ee)}, \pi_{i(ee)}\}$$

Some Probabilities

- p_i = Probability of observing dominant allele for i^{th} SNP

		Mother	
		E	e
Father	E	p_i^2	$p_i(1-p_i)$
	e	$p_i(1-p_i)$	$(1-p_i)^2$

Hardy-Weinberg Assumption

EE	p_i^2
Ee	$2p_i(1-p_i)$
ee	$(1-p_i)^2$

It's beyond today's class, but the model scales to any number of alleles

Outline

- DNA? Who Cares?

- DNA and the Quasi-Identifier Dilemma
  - Genome Basics
  - Identity Disclosure (Re-identification)
  - Membership/Attribute Inference

Defining Uniqueness

- Chance that 2 ***unrelated*** people match at a single SNP i:

$$\mu_i = \sum_{l=1}^{3} \pi_{il}^2$$

- If dominant allele probability ≤ 0.9:

$$0.375 \leq \mu_i \leq 0.689$$

$p_i = 0.5$ $p_i = 0.9$

Defining Uniqueness

- Chance that 2 ***unrelated*** people match at a single SNP i:

$$\mu_i = \sum_{l=1}^{3} \pi_{il}^2$$

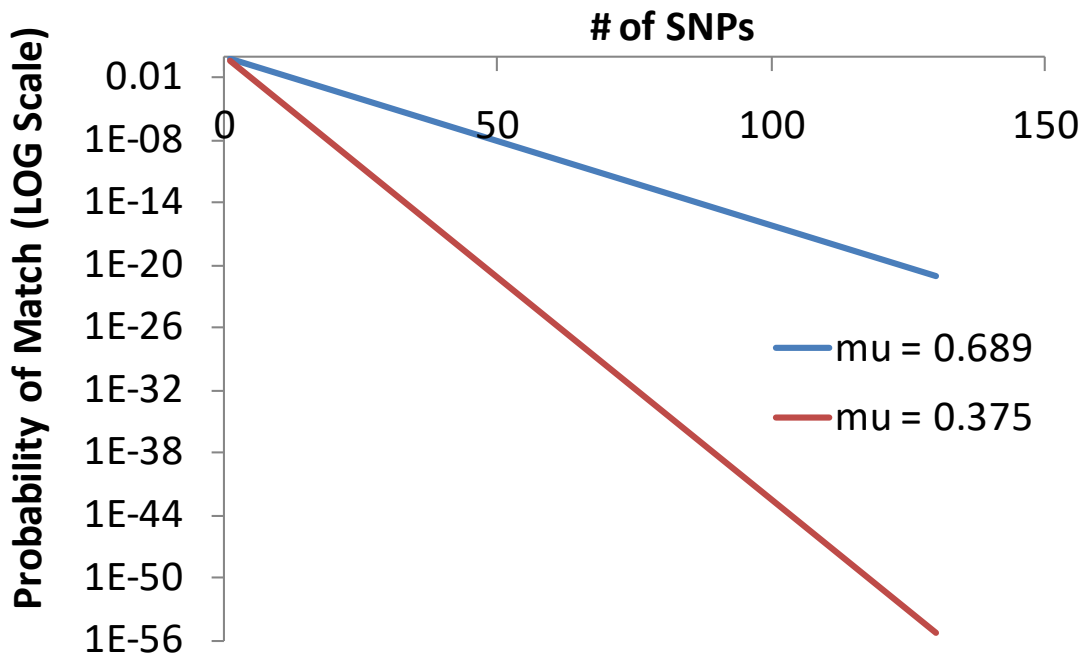
- If dominant allele probability ≤ 0.99:

$$0.375 \leq \mu_i \leq 0.961$$

$p_i = 0.5$ $p_i = 0.99$

Defining Uniqueness

- Assume independence of SNPs (not always the case)
- Prob. 2 people match on set of SNPs S: $\prod_{i=1}^{|S|} \mu_i$



Leveraging Prior Knowledge

- You suspect person is selected from a population of N people

- Probability it's the same individual given your sample AND record in dataset is a "match" over set of SNPs is:

$$P(same \mid match) = \frac{P(match \mid same)P(same)}{P(match \mid same)P(same) + P(match \mid \neg same)P(\neg same)}$$

$$P(same|match) = \frac{1(1/N)}{1(1/\mathrm{N}) + \prod_{i=1}^{|S|} \mu_i \, (1 - 1/N)}$$

Independent SNPs?

- Chromosome 21
  - ~ 24,047 SNPs
  - Summarize into ~ 4,563 SNPs

- But we only need around 80 to uniquely represent you!

Privacy risks in sharing individual-level data

- Sharing de-identified individual-level data with sensitive attribute (e.g., disease) is risky
- Attackers can collect and sequence DNA samples from identified targets
- Linkage attack (Re-identification)

Targeted DNA samples

Name	SNP					
	1	**2**	**3**	**4**	**…**	**m**
Angelina	0	1	1	0	…	2
Bradley	2	0	0	0	…	1

Shared Genomic Records

SNP						Disease
1	**2**	**3**	**4**	**…**	**m**	
1	0	2	0	…	1	HIV
1	0	1	0	…	0	HIV
1	1	2	1	…	1	Cancer
0	0	0	0	…	0	Diabetes
0	1	1	0	…	2	Cancer
0	0	0	0	…	1	Miscarriage

SNP (single nucleotide polymorphism) is the most common type of genetic variation.

42

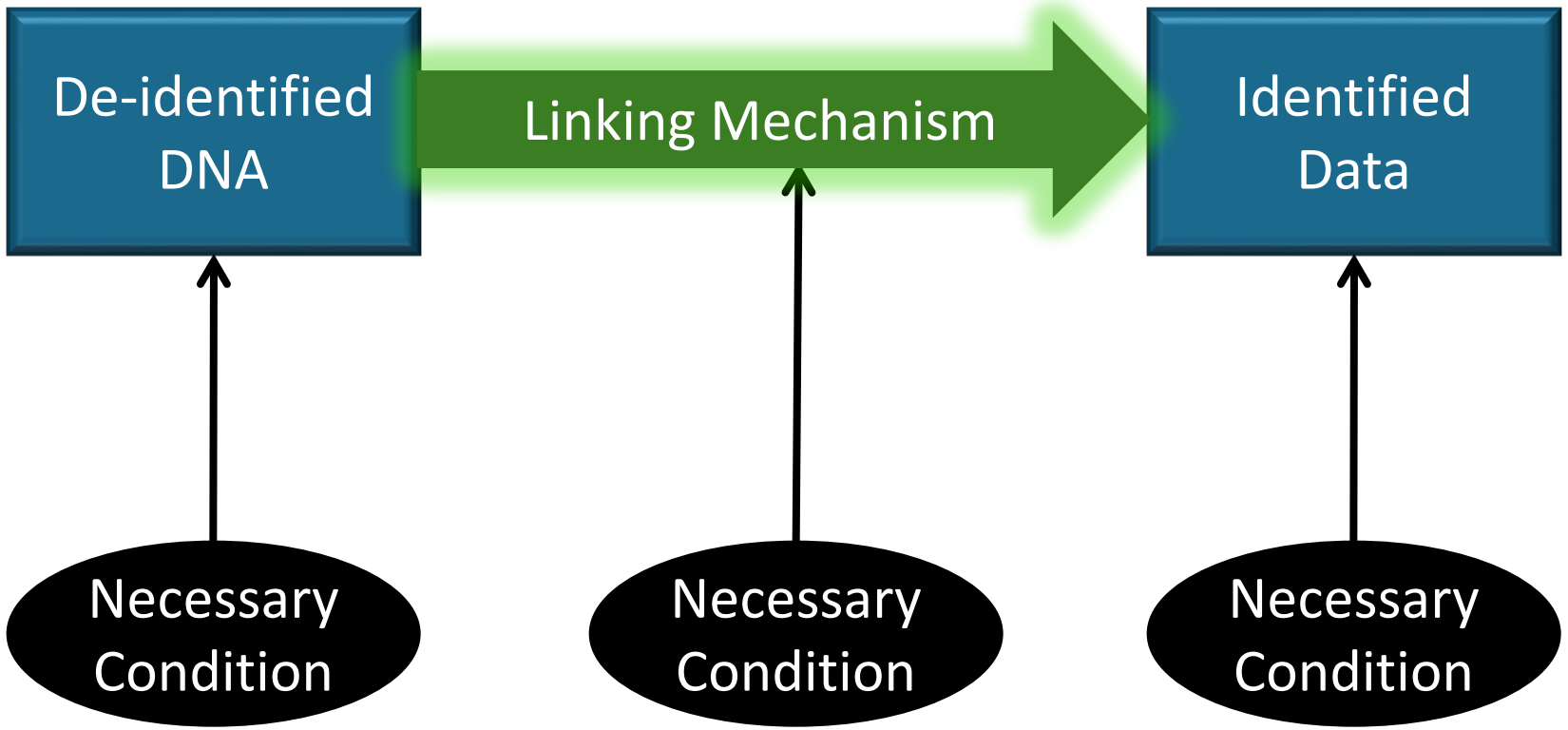
*omics Data is High-Dimensional!

- AG AT CA AA CA TG GC AA TA AT CA GT AG GG AT AT CC AC AG TT AA TA CT CC AG
  CA GC AG AT AA CA GG TA GG GC GA AA GA GA GC CA TA CT CC CC CC CA TT AG TT
  GT AA GG GG AT AC CA GC GG AC AC AC AG TT TC AA CC AC AT CG TG AC AT GG TC
  AT CC GC GA TC AC CT CC CG TG TT TT GC TC TC AC AC AT AT GA CA TG CT TA AC
  CT AG TC CC AC TT CA AA CG TG TA AA TT GA TA TC CT AC AT CA CT CA CC TC AG
  TC TA CA CC CG TG TA TT AG GT CC TT TA TG TA CC GA GG CC GG GG AG TT GG TT
  CA CG TT CC CC AA CG CC TC AT TG AT AT AT CC GC GC CA CA TC AA GT CT GG CT
  CC AC CT TA GT TA AC GC GT TG AA CG TG GC CA TT CC CA AG GG AG CC AC
  AT AG GG GT GG TA GG GA AG TT TT TA CG AC AC GT TG TT CG AT CG GC AC CC AT
  CT AC TG AT TT TT AG CT GA CG CA TT GC TG CG AG GC GG TG TC AA AC TA TG TC
  CG GT GT CG AC AA AT GG GT CC GT TT CG CA GA AT TT GC TA CA TA AC AC AC AG
  TC TG AC TG TA GT AC GT TG TG AA GG TG CG TG CT AG TG CG GG CC AA CC GG AG
  GA GG GT CG TC TG CC GT AC TT CC AG CT AT AT TG TT AG GG TA AA AG AA GT CA
  AC GC CC TT TC GA GC GT GG AT GT CG TC AC AT GA AC AT CT GA TC CA GA CA GA
  CA AT AT TG AC GG CC CA AT GT CT TT GT GT CT AG CT CG CA GT AC CC CT GA CG
  CC GT GC TA CC CC AG TG GG GG TA GA TT CG AC CG CG GG TC AG TA GC TC CT GA
  AG GG AC TA AT TA CT TC TG TT GT GT GC GT CC CT TC GC TA GG AC GA AT CT TG
  GA TG GG AA AA TA AT CT CA AA AC CC TC AT CG AG GG TG CC GC GT AG AC CT TC
  CA AG GC GA TA TG AC TA AA CG AA GA CA GA TA TG AT AT CA CC AC AC GG TT TA
  GG GC CG CC TG TA TG TT GT CC CT AC TT CC CT TA CA AC TG CG GG TC AT GA CC

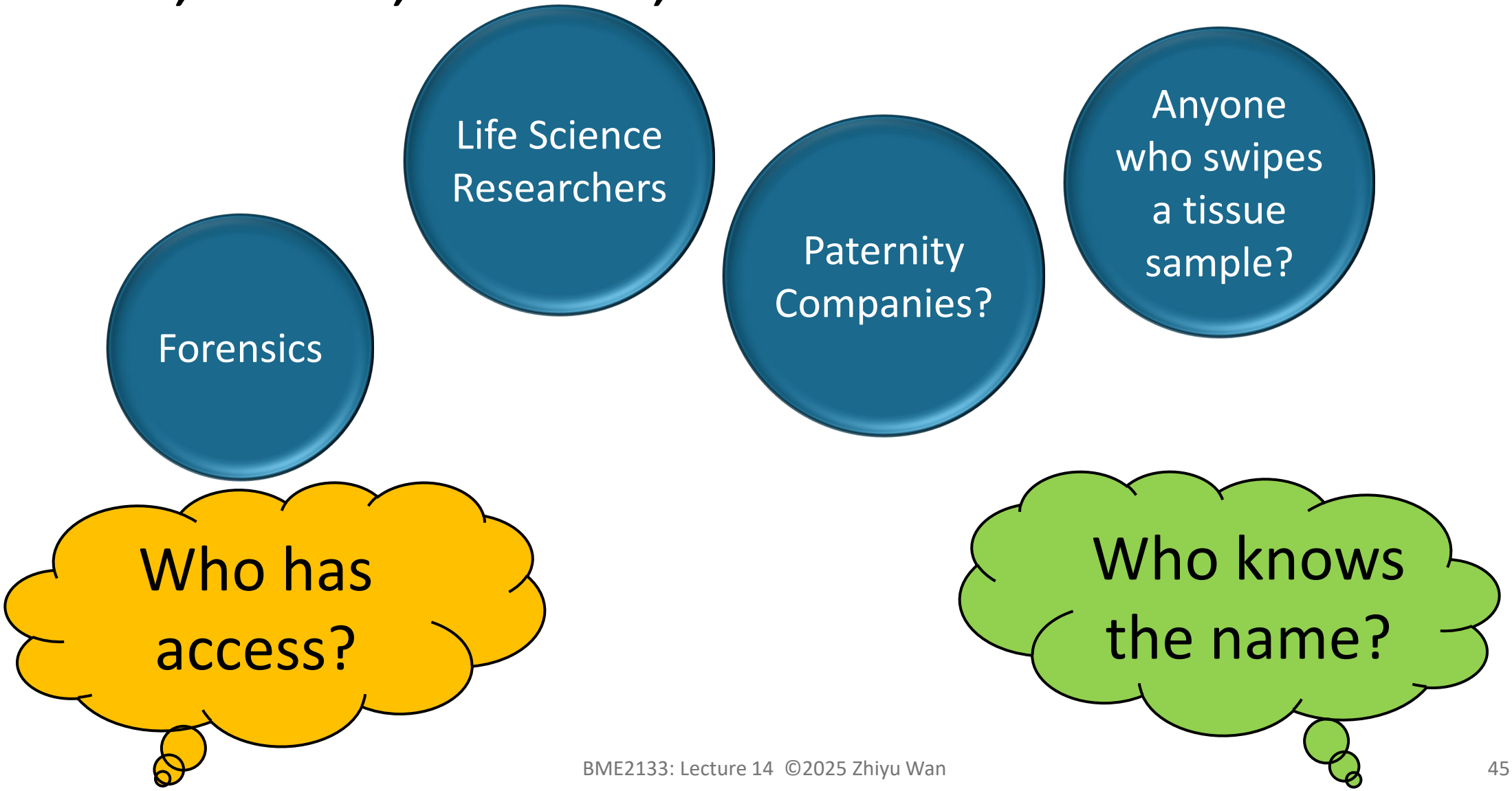
But Who is This?

Adapted from Brad's slides

Uniqueness is **NOT** Sufficient



Who, What, Where, …



Outline

- DNA? Who Cares?

- DNA and the Quasi-Identifier Dilemma
  - Genome Basics
  - Identity Disclosure (Re-identification)
  - Membership/Attribute Inference

Privacy risks in sharing summary statistics

- Sharing individual-level genomic data is useful, but risky
- Sharing allele (variant of genomic region) frequencies about a pool of genomes is still useful, but also (less) risky
- In 2008, Homer et al. introduced an attack…

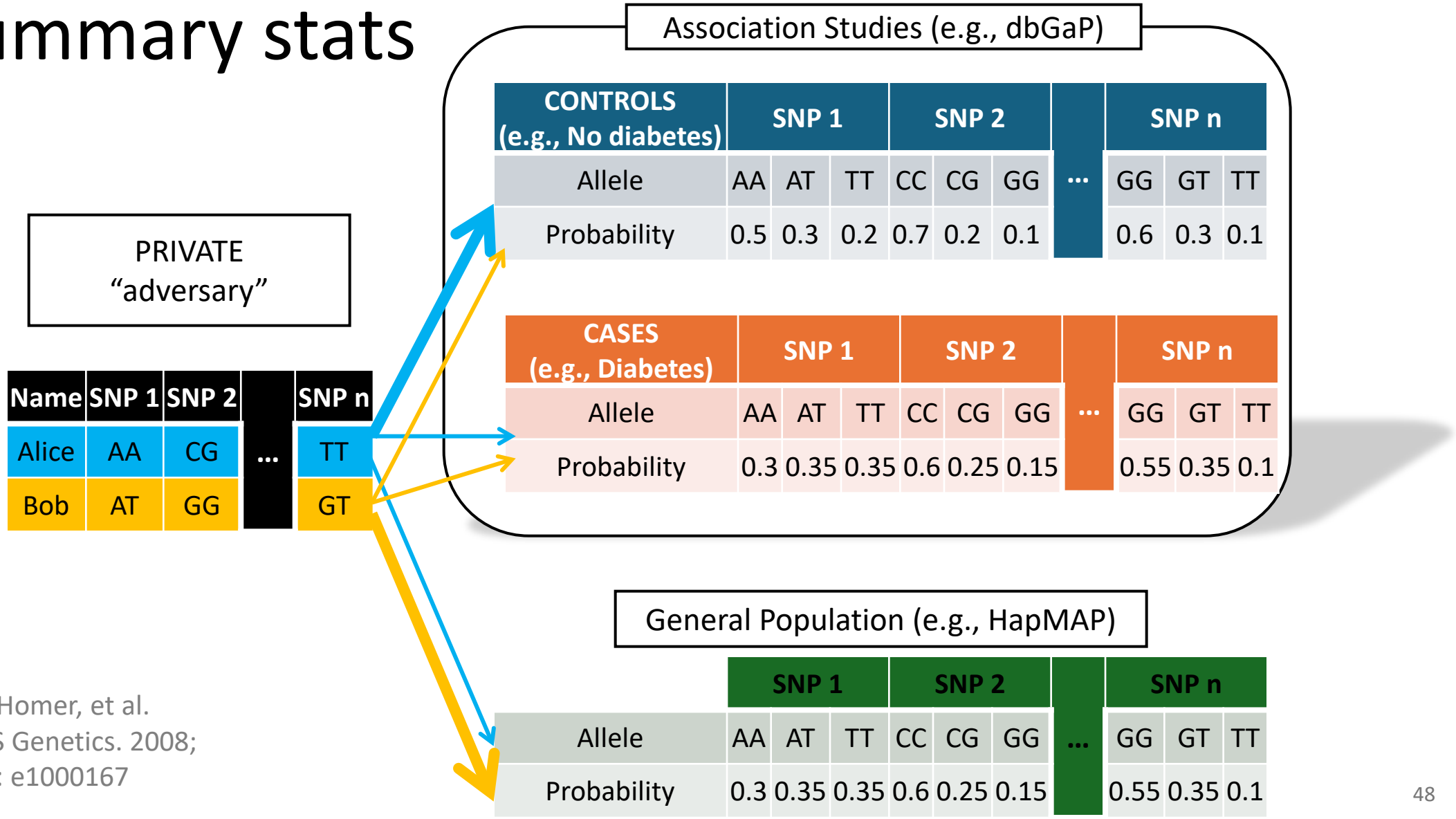
Shared Allele Frequencies

SNP						Disease
1	**2**	**3**	**4**	**…**	**m**	
0.1	0.2	0.3	0.1	…	0.5	Cancer/HIV

Homer N, et al. PLoS Genetics. 2008; 4(8): e1000167.

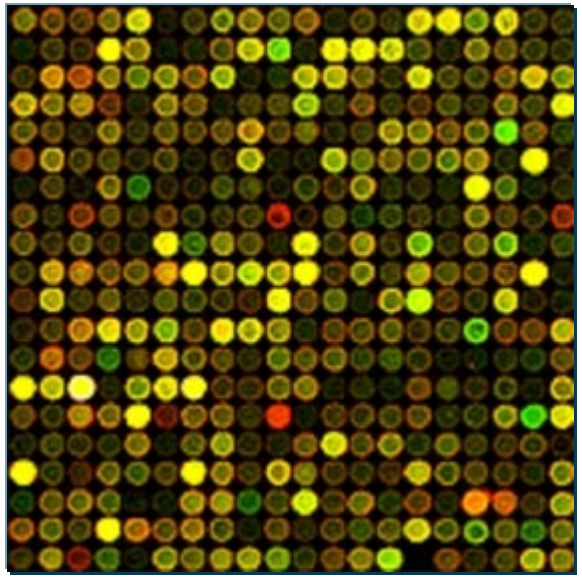
47

It’s not just unique sequences… summary stats



*N. Homer, et al. PLoS Genetics. 2008; 4(8): e1000167

More Specifically



- Use microarray technology to measure intensity of allele.
- For a single individual i:
  - Each SNP_j allele has intensity of 0, 0.5, or 1
  - call this $Y_{i,j}$
- For a “mixed” study population:
  - Each SNP_j allele has intensity proportional to study population’s contribution
  - Call this M_j
- For a “reference” population:
  - Same concept as M
  - Call this R_j

So, where's the Target?

- $|Y_{ij} - M_j|$ ← difference between individual & mixed study
- $|Y_{ij} - R_j|$ ← difference between individual & reference pop.

$$D(Y_{ij}) = |Y_{ij} - R_j| - |Y_{ij} - M_j|$$

- Null Hypothesis: Individual is not in mixed study.
  - $D(Y_{ij})$ should be approaching 0 [due to "ancestral similarity" in M and R]
- Alternative Hypothesis
  - $D(Y_{ij}) > 0$ because M_j is shifted away from reference by Y_j's contribution to the mixture
  - $D(Y_{ij}) < 0$ because Y_j is more similar to reference population than the mixture

Testing

$$T(Y_i) = \frac{E(D(Y_i)) - \mu_0}{SD(D(Y_i))/\sqrt{s}}$$

- μ_0 : Mean of $D(Y_i)$ of all individuals **not** in the mixture
- $SD(Y_i)$: St. Dev. of $D(Y_{i,j})$ for all SNPs j and individual Y_i
- s: number of SNPs
- Can assume $\mu_0 = 0$ [random individual equidistant to M & R]
- Null hypotheses T = 0. Alternative is that T > 0

Homer's attack in a nutshell

The attacker knows:

- The genome of the target (her set of genomic variants) - Y_{ij}
- The allele frequencies of the Mixture (pool of genomes) he's attacking - M_j
- Population allele frequencies - Pop_j

Snp	Allele Frequency ($Y_{i,j}$)	Distance Measure	Interpretation at the given SNP
	0.0 0.25 0.50 0.75 1.0	$D(Y_{i,j}) = \lvert Y_{i,j} - Pop_j \rvert - \lvert Y_{i,j} - M_j \rvert$	
j		= |1.0 - 0.25| - |1.0 - 0.75| = 0.75 - 0.25 = 0.50	most likely to be in the Mixture

Homer N, et al. PLoS Genetics. 2008; 4(8): e1000167.

52

Number of SNPs Necessary

- Approximately 10,000 – 25,000 SNPs necessary to determine if person in a particular study.

Privacy risks in sharing summary statistics

- Sharing individual-level genomic data is useful, but risky
- Sharing allele (variant of genomic region) frequencies about a pool of genomes is useful, but also (less) risky
- In 2008, Homer et al. introduced an attack…

… that led the NIH to removing summary statistics from NIH Database of Genotypes and Phenotypes (dbGaP)

- And more powerful attacks have emerged (e.g., Wang, Sankararaman)

Homer N, et al. PLoS Genetics. 2008; 4(8): e1000167.
Wang R, et al. ACM CCS. 2009: 534-544.
Sankararaman S, et al. Nature Genetics. 2009: 965-967.

54

Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study

Rui Wang, Yong Li, XiaoFeng Wang, Haixu Tang, Xiaoyong Zhou
Indiana University Bloomington
Bloomington, IN
{wang63,yonli,xw7,hatang,zhou}@indiana.edu

Abstract

Genome-wide association studies (GWAS) aim at discovering the association between genetic variations, particularly single-nucleotide polymorphism (SNP), and common diseases, which have been well recognized to be one of the most important and active areas in biomedical research. Also renowned is the privacy implication of such studies, which has been brought into the limelight by the recent attack proposed by Homer et al. Homer's attack demonstrates that it is possible to identify a participant of a GWAS from analyzing the allele frequencies of a large number of SNPs. Such a threat,

1. INTRODUCTION

The rapid advancement in genome technology has revolutionized the field of human genetics by enabling the large-scale applications of genome-wide association study (GWAS) [7], a study that aims at discovering the association between human genes and common diseases. To this end, GWAS investigators determined the genotypes of two groups of participants, people with a disease (cases) and similar people without (controls) in an attempt to use statistical testing to identify genetic markers, typically single-nucleotide polymorphisms (SNP), that are associated to the disease suscepti-

Proc. 2009 ACM Conference on Computers & Communications Security

Homer needed ~10,000 SNPs… Wang needs around 200!

Leverages linkage disequilibrium and some nifty integer programming.

Linkage Disequilibrium

- Non-random association of alleles at different loci (i.e., different regions of genome)

- Occurs when loci are not independent

And then some…

- Homer (and others) fail to provide an upper bound on the power of detection
- Given
  - n: number of people in mixed sample
  - β: maximal allowable power
  - α: false positive level

  the Likelihood Ratio (LR) test provides the bound

$$z_{\alpha} + z_{1-\beta} = \sqrt{|S|/n}$$

*S Sankararaman, G. Obozinski, M. Jordan, E. Halperin. Nature Genetics. 2009; 41(9): 965-967.

Valid When…

- n > 500
- Minor allele frequency > 0.05

*S Sankararaman, G. Obozinski, M. Jordan, E. Halperin. Nature Genetics. 2009; 41(9): 965-967.

And then some…

- Provides an upper bound on the number of SNPs that can be “safely” disclosed for chosen
  - False positive rate
  - Power of detection

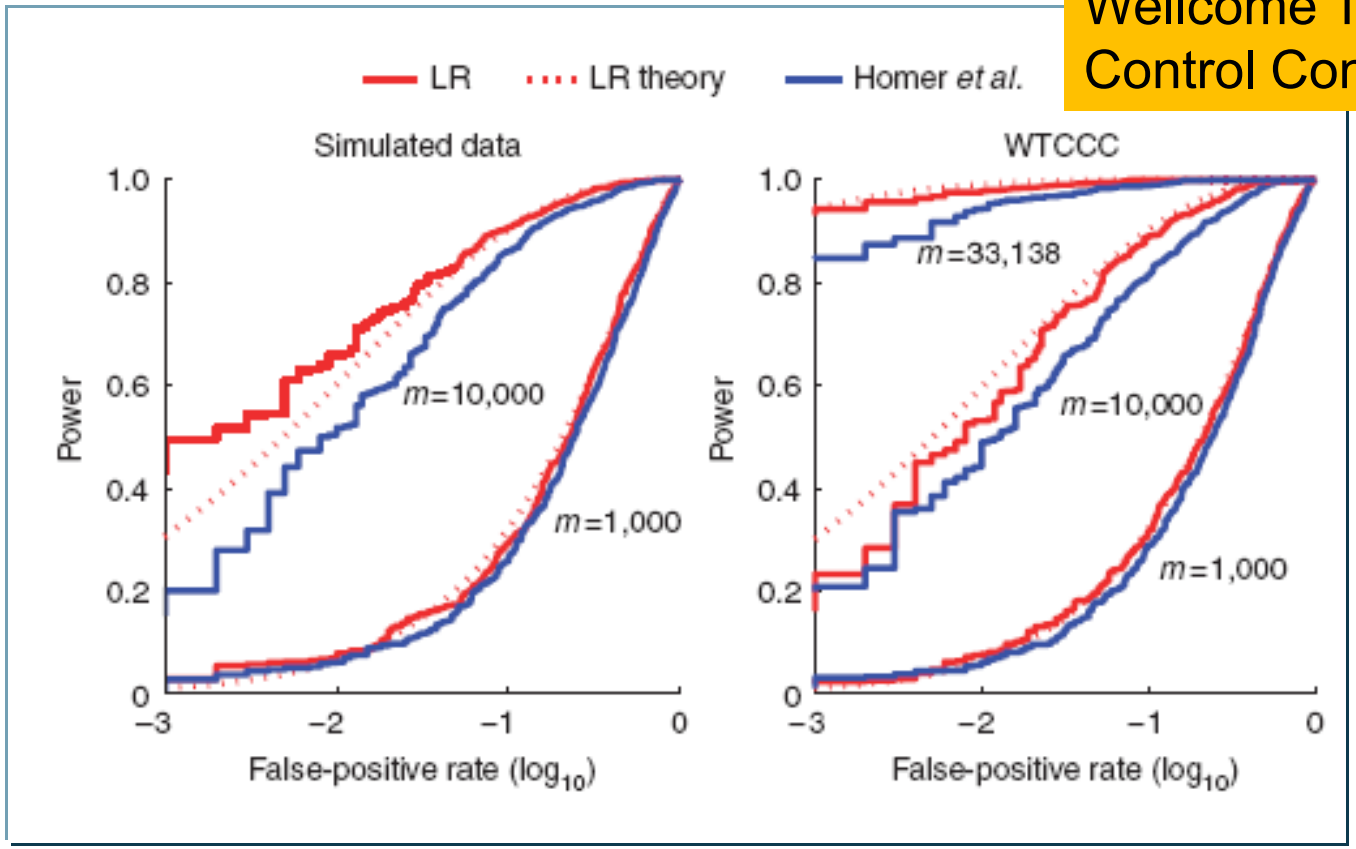
- Implies
  - |S| is linear in *n* for a fixed false positive and negative rate
  - Power of the test does NOT depend on allele frequencies (if the recessive allele is large enough)!

*S Sankararaman, G. Obozinski, M. Jordan, E. Halperin. Nature Genetics. 2009; 41(9): 965-967.

Analysis

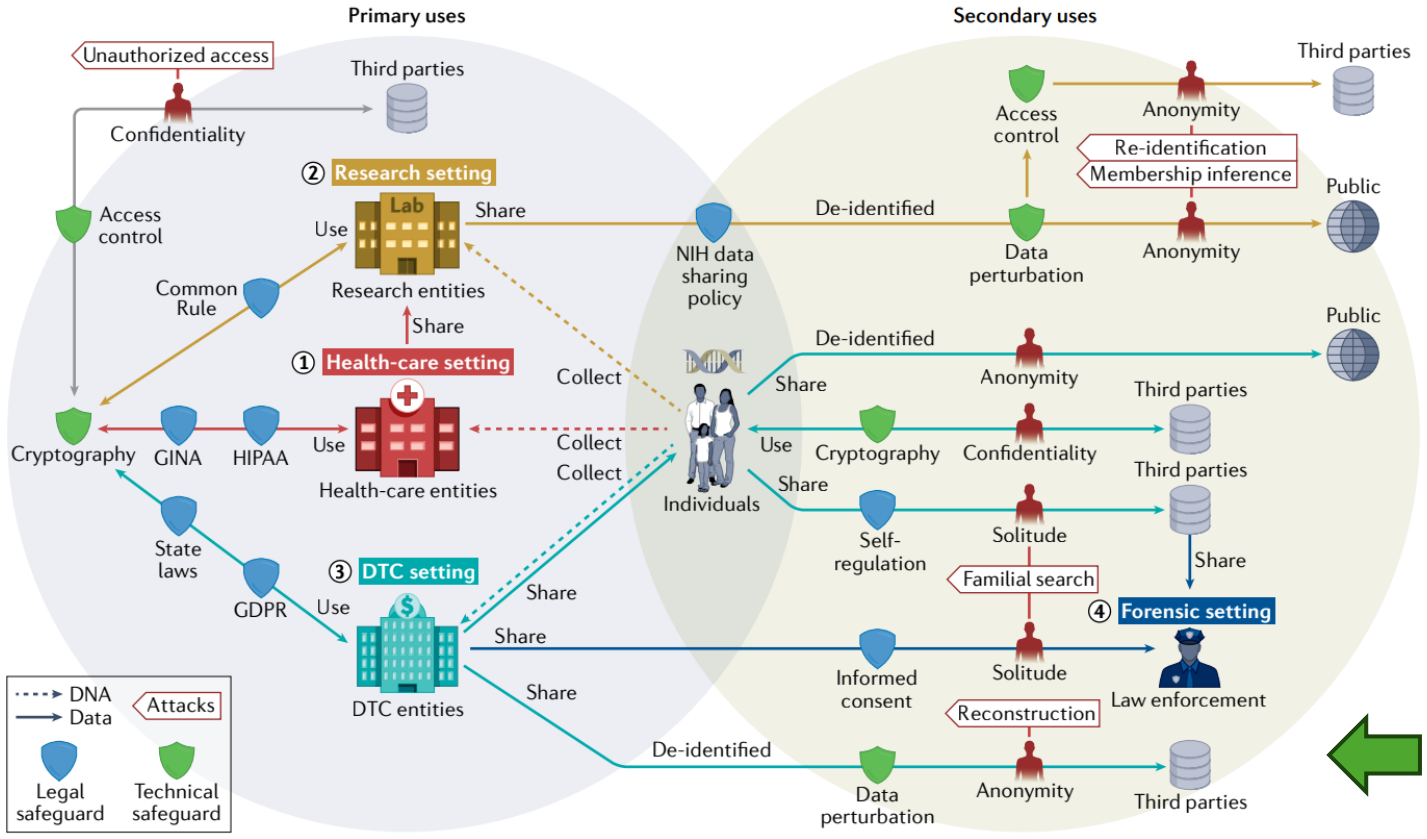
- Note: “m” is “|S|”

Wellcome Trust Case Control Consortium



*S Sankararaman, G. Obozinski, M. Jordan, E. Halperin. Nature Genetics. 2009; 41(9): 965-967.

Genomic Data Protection Methods



nature reviews genetics

Explore content · About the journal · Publish with us

nature › nature reviews genetics › review articles › article

Review Article | Published: 04 March 2022

Sociotechnical safeguards for genomic data privacy

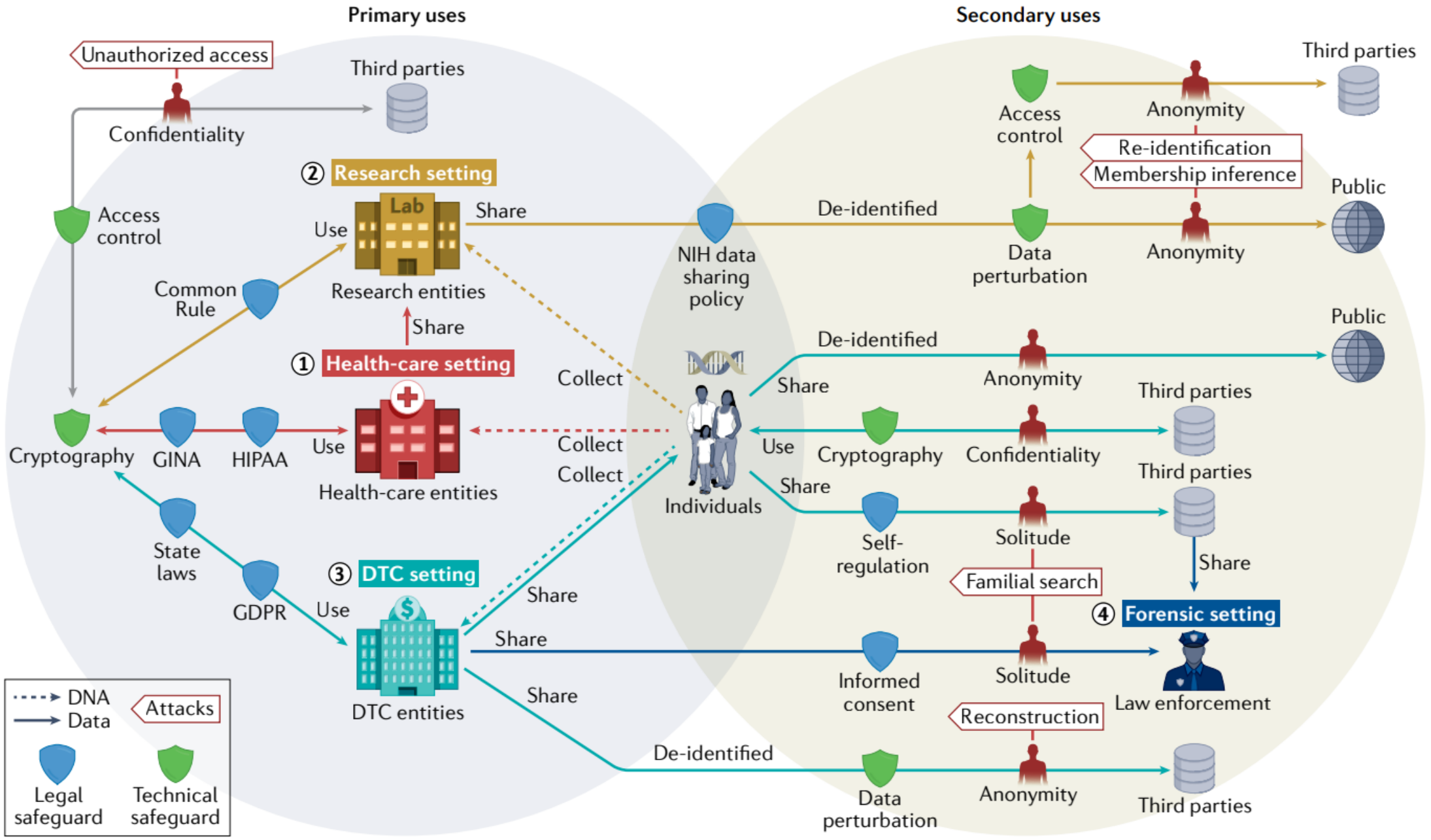
Zhiyu Wan, James W. Hazel, Ellen Wright Clayton, Yevgeniy Vorobeychik, Murat Kantarcioglu & Bradley A. Malin

Nature Reviews Genetics **23**, 429–445 (2022) | Cite this article

Abstract

Recent developments in a variety of sectors, including health care, research and the direct-to-consumer industry, have led to a dramatic increase in the amount of genomic data that are collected, used and shared. This state of affairs raises new and challenging concerns for personal privacy, both legally and technically. This Review appraises existing and emerging threats to genomic data privacy and discusses how well current legal frameworks and technical safeguards mitigate these concerns. It concludes with a discussion of remaining and emerging challenges and illustrates possible solutions that can balance protecting privacy and realizing the benefits that result from the sharing of genetic 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.



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.

Groups working on genomic privacy

Readings due on December 3rd

- 1. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Malin B. **Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach**. *The American Journal of Human Genetics*. 2017 Feb 2;100(2):316-22. https://www.cell.com/action/showPdf?pii=S0002-9297%2816%2930526-2
- Optional
  - ❑ 2. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Ganta R, Heatherly R, Malin BA. **A game theoretic framework for analyzing re-identification risk**. *PloS one*. 2015 Mar 25;10(3):e0120592. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0120592&type=printable
  - ❑ 3. Wan Z, Vorobeychik Y, Xia W, Liu Y, Wooders M, Guo J, Yin Z, Clayton EW, Kantarcioglu M, Malin BA. **Using game theory to thwart multistage privacy intrusions when sharing data**. *Science Advances*. 2021 Dec 10;7(50):eabe9986. https://www.science.org/doi/pdf/10.1126/sciadv.abe9986

Feedback Survey

- One thing you learned or felt was valuable from today's class & reading

- Muddiest point: what, if anything, feels unclear, confusing or "muddy"

- https://www.wjx.cn/vm/hX0mIro.aspx