Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 11: Risks of Genomic Data Sharing

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

April 11, 2025

Learning Objectives of This Lecture

- Understand why sharing genomic data is risky
  - Re-identification attacks
  - Membership/Attribute Inference attacks

Outline

- 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] 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] 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

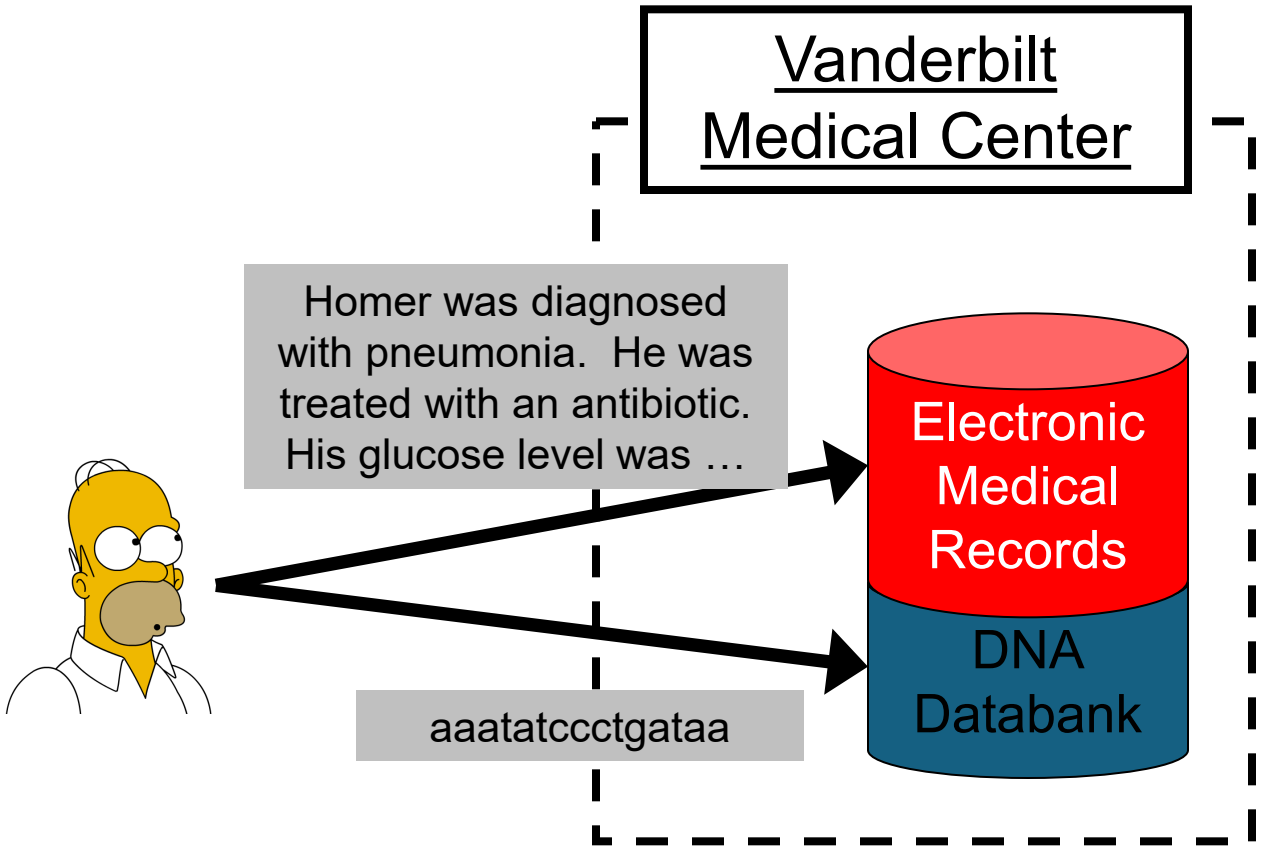
Jill Pulley [1], Ellen Clayton [2], Gordon R Bernard [3], Dan M Roden [4], Daniel R Masys [5]

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



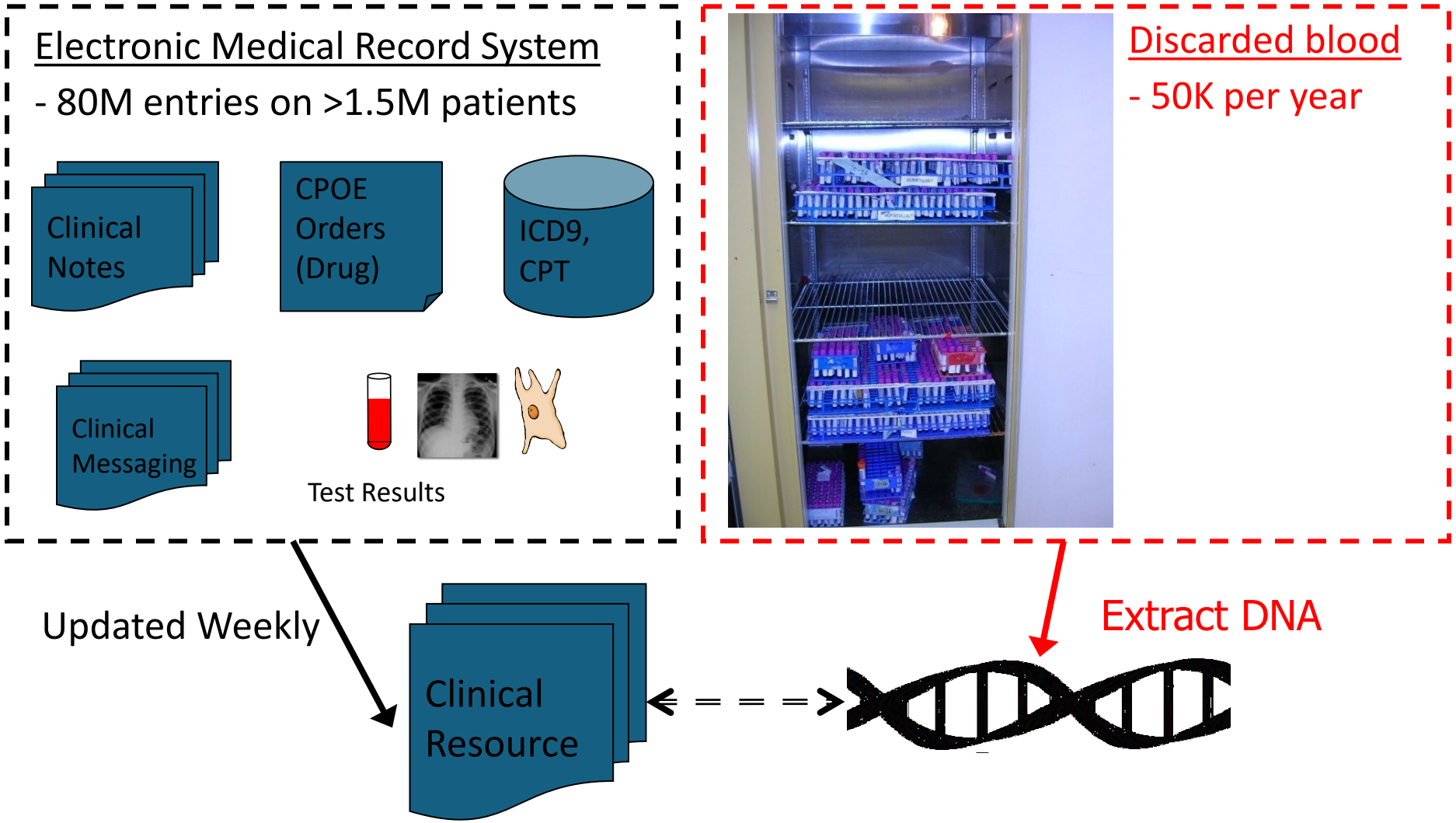
Clinical Goal:

Tailor Homer's treatment to his DNA

Scientific Challenge:

Need to learn associations between DNA clinical features

Information Integration

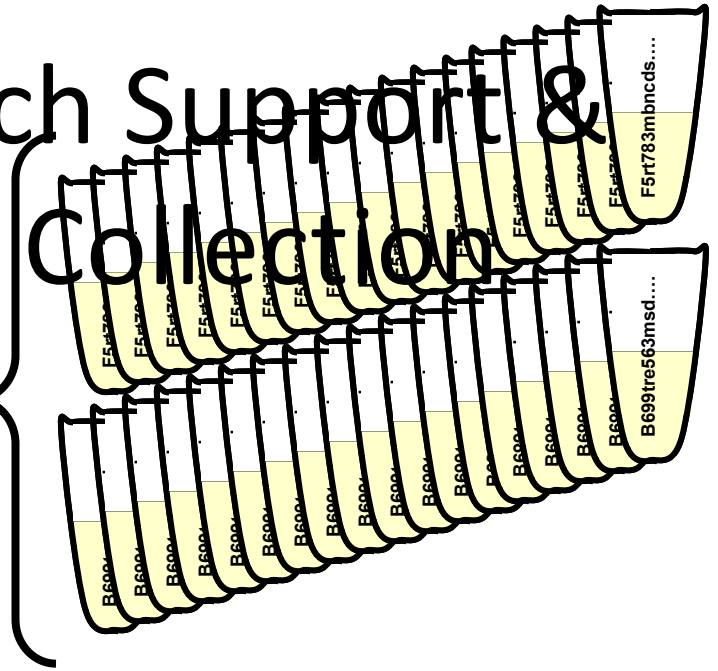


Associations

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



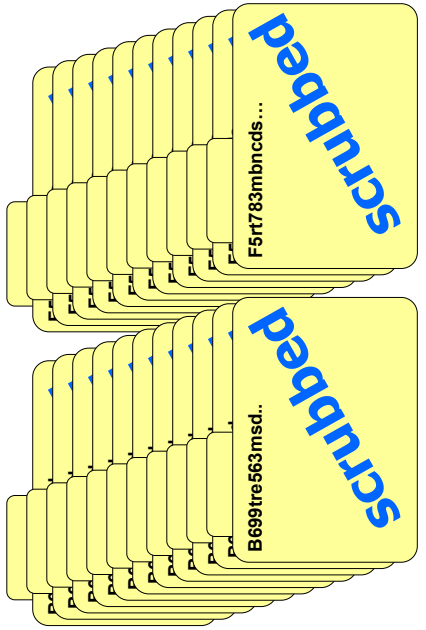
cases

controls

cases

controls

Investigator query



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**

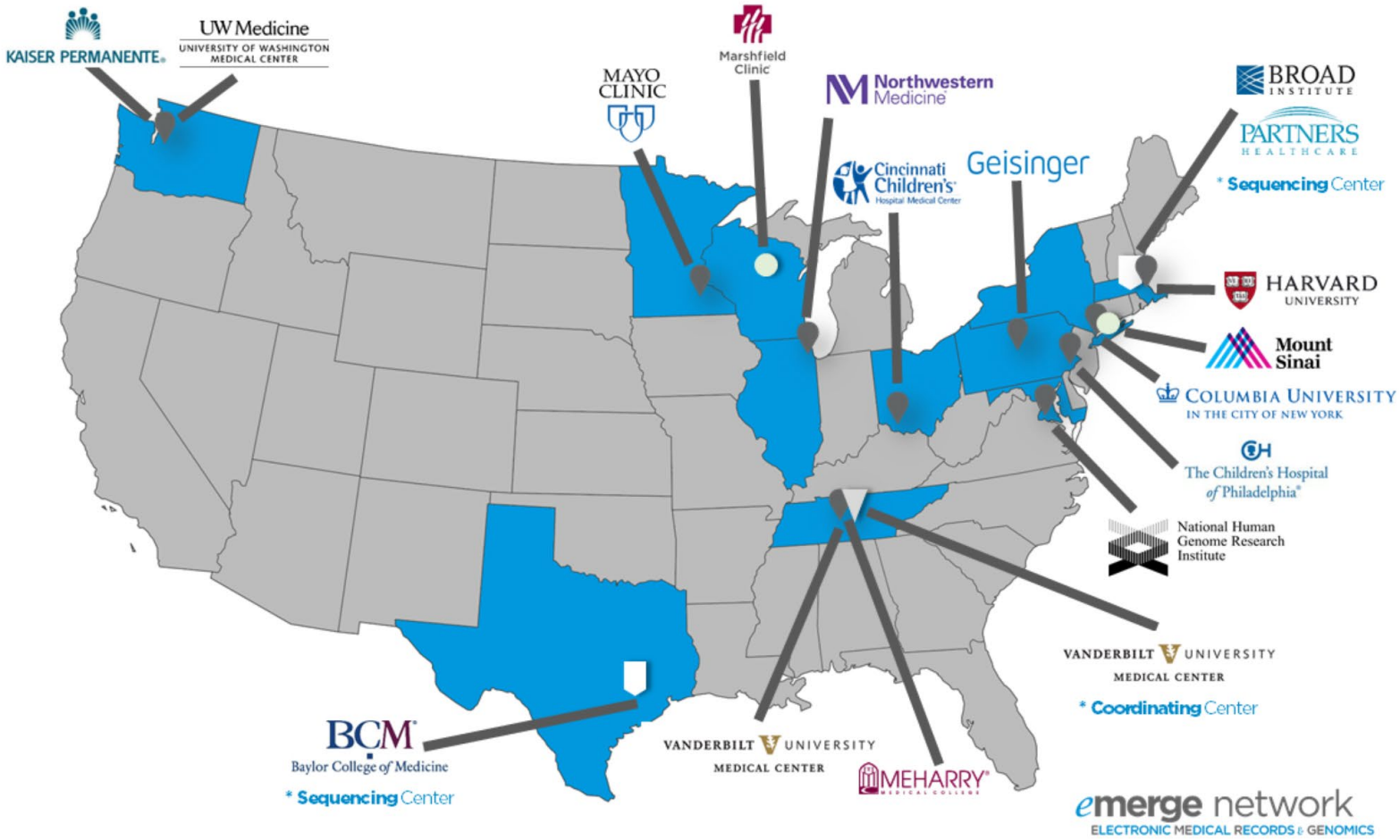
Vanderbilt DNA Databank Model

- 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

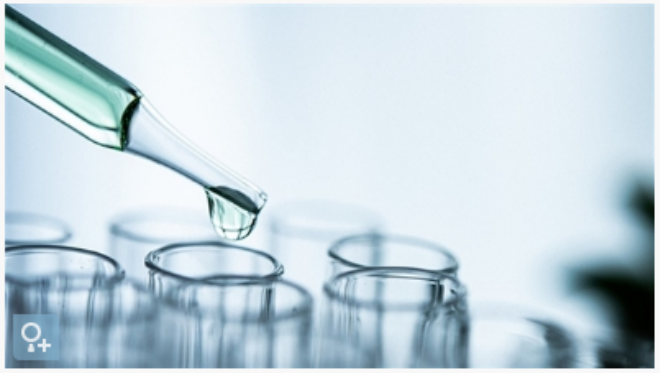
Data Sharing Policies

- 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

Outline

- 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:

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

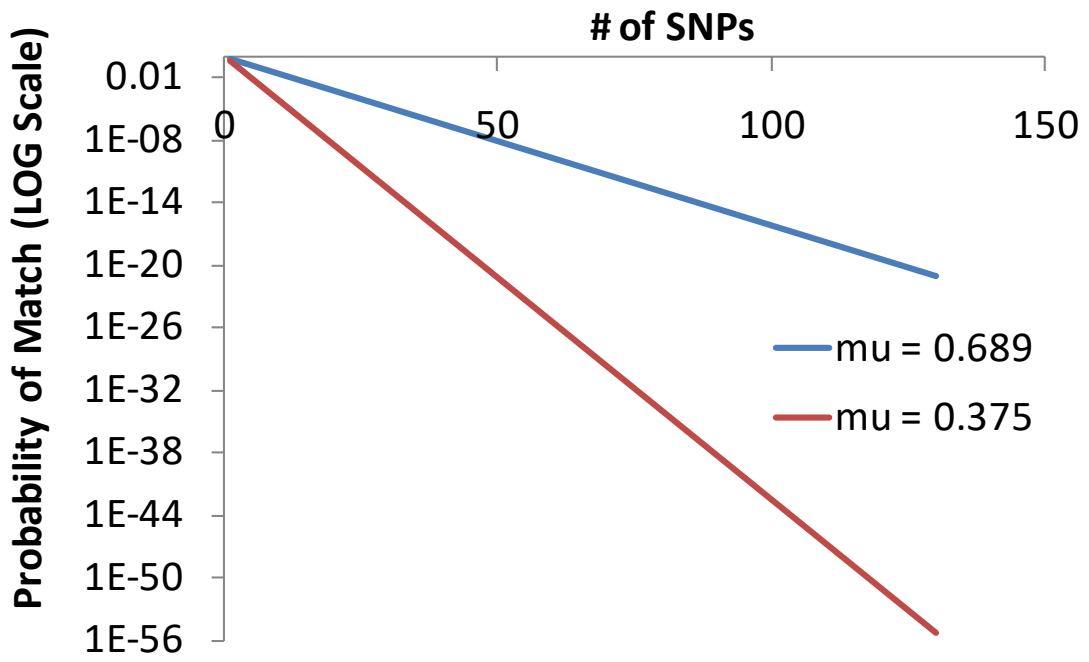
- If dominant allele probability ≤ 0.99:

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

$p_i = 0.5$ $p_i = 0.99$

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!

Outline

- DNA? Who Cares?

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

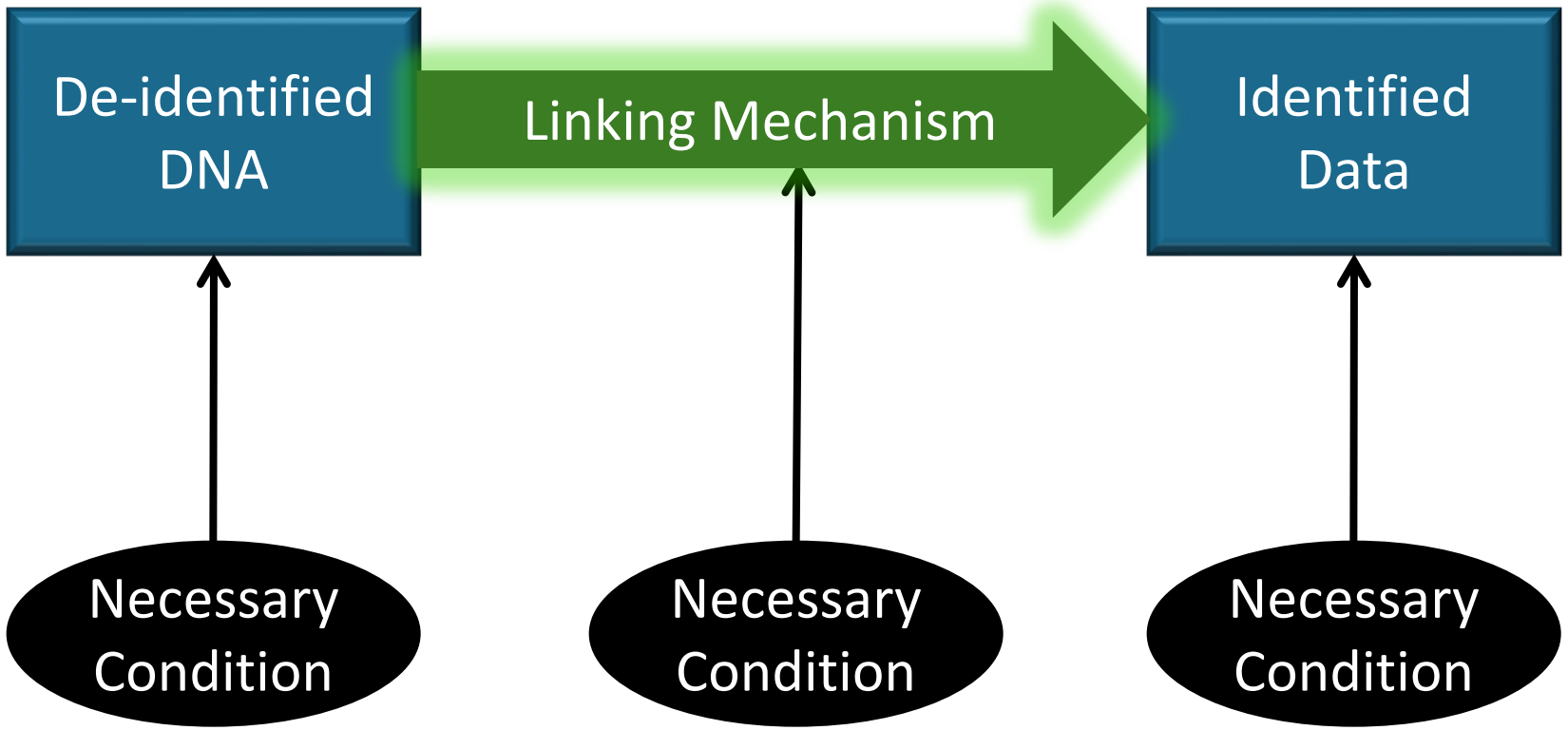
*omics Data is High-Dimensional!

- 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

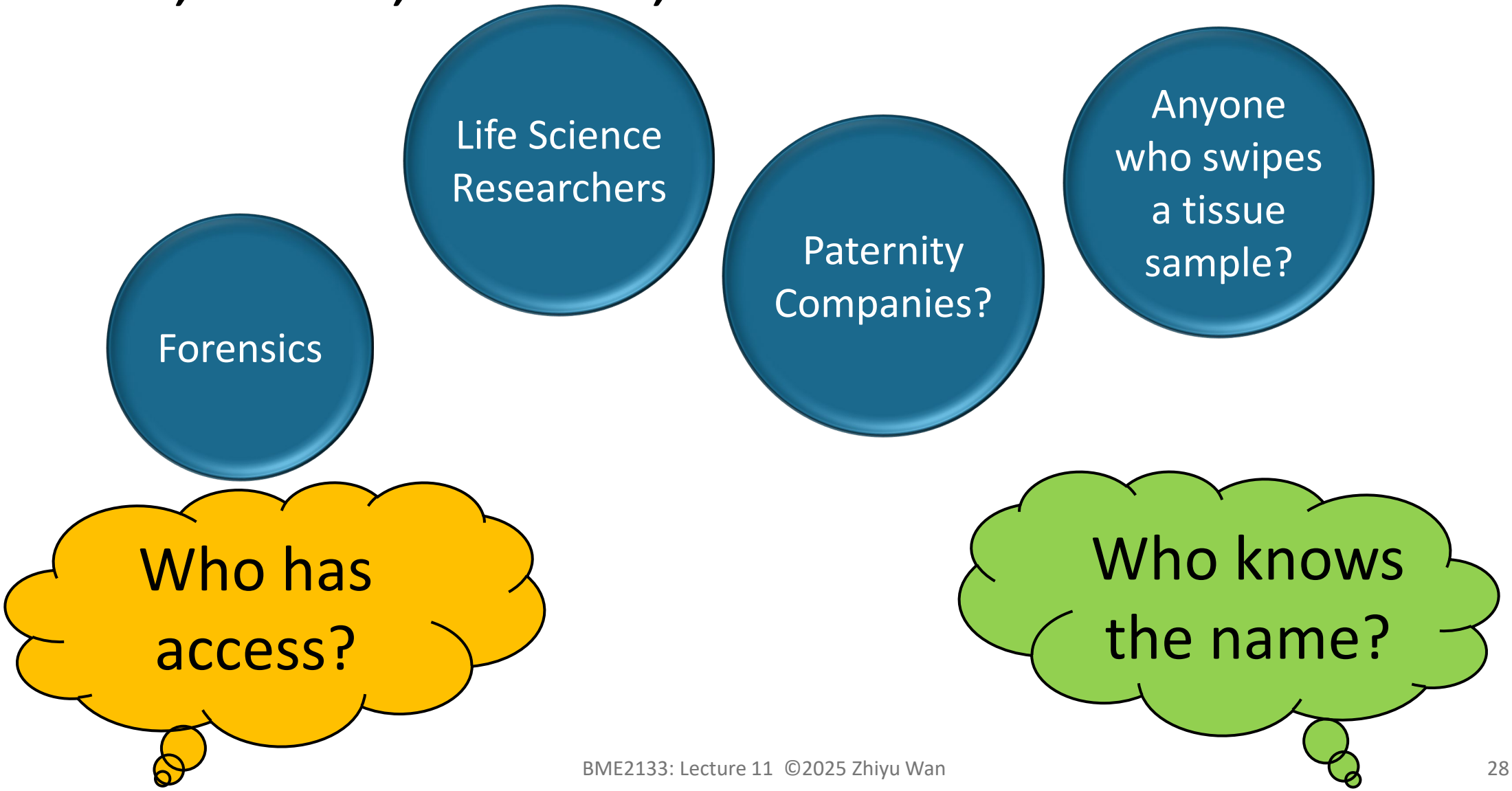
But Who is This?

Adapted from Brad's slides

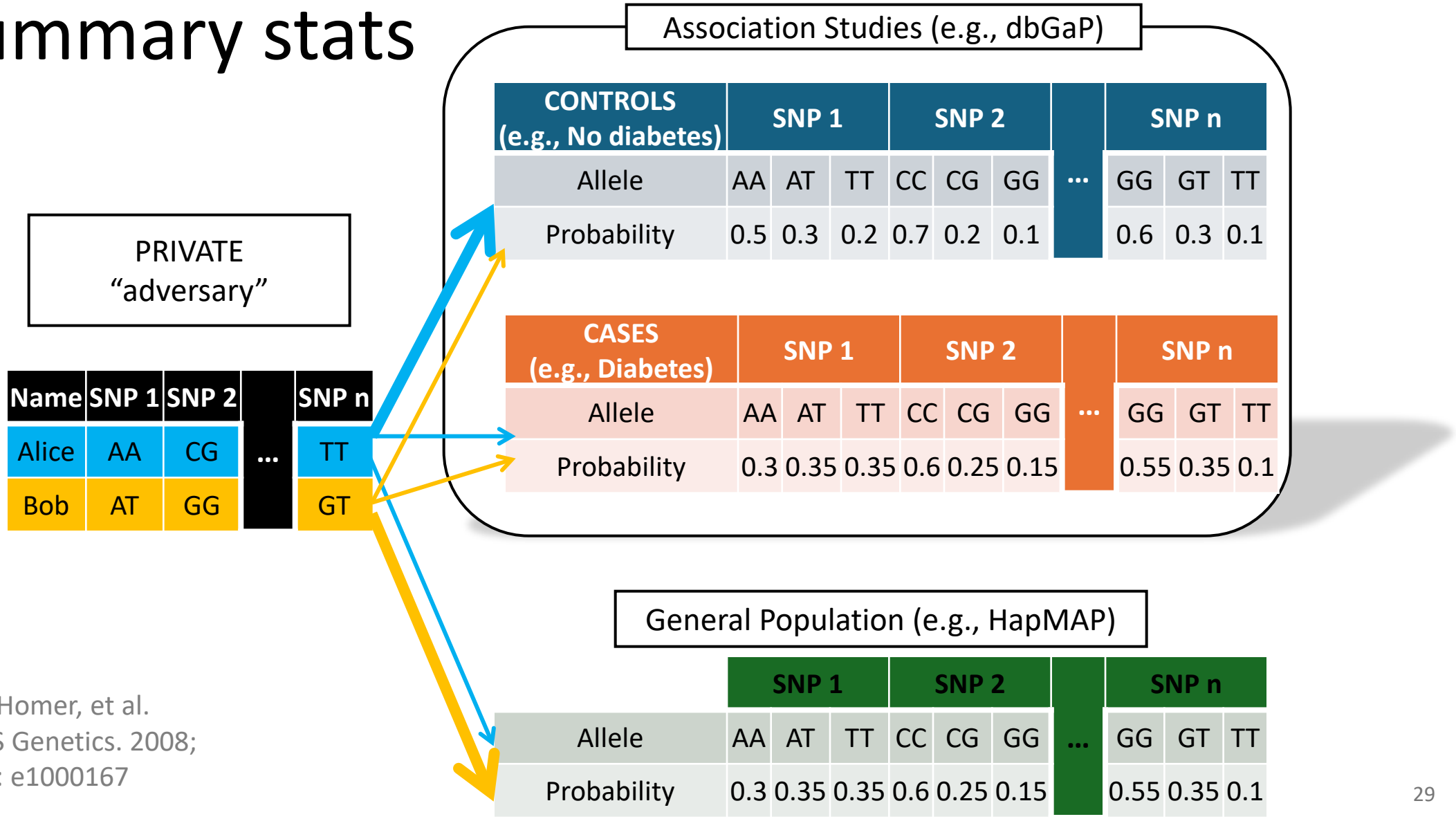
Uniqueness is **NOT** Sufficient



Who, What, Where, …

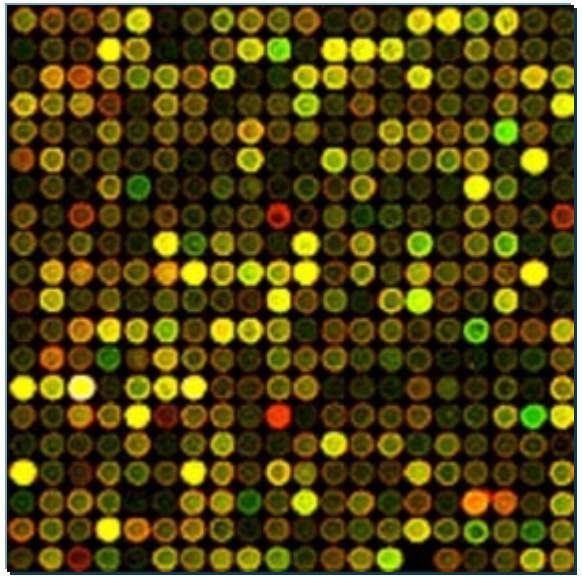


It’s not just unique sequences… summary stats



*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

Number of SNPs Necessary

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

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.

ZEYI YANG LOUISE MATSAKIS CAROLINE HASKINS SECURITY APR 2, 2025 1:31 PM

Cybersecurity Professor Faced China-Fundi Disappearing, Sources Say

A lawyer for Xiaofeng Wang and his wife says they are "safe" after FBI searche from Indiana University, where he taught for over 20 years.

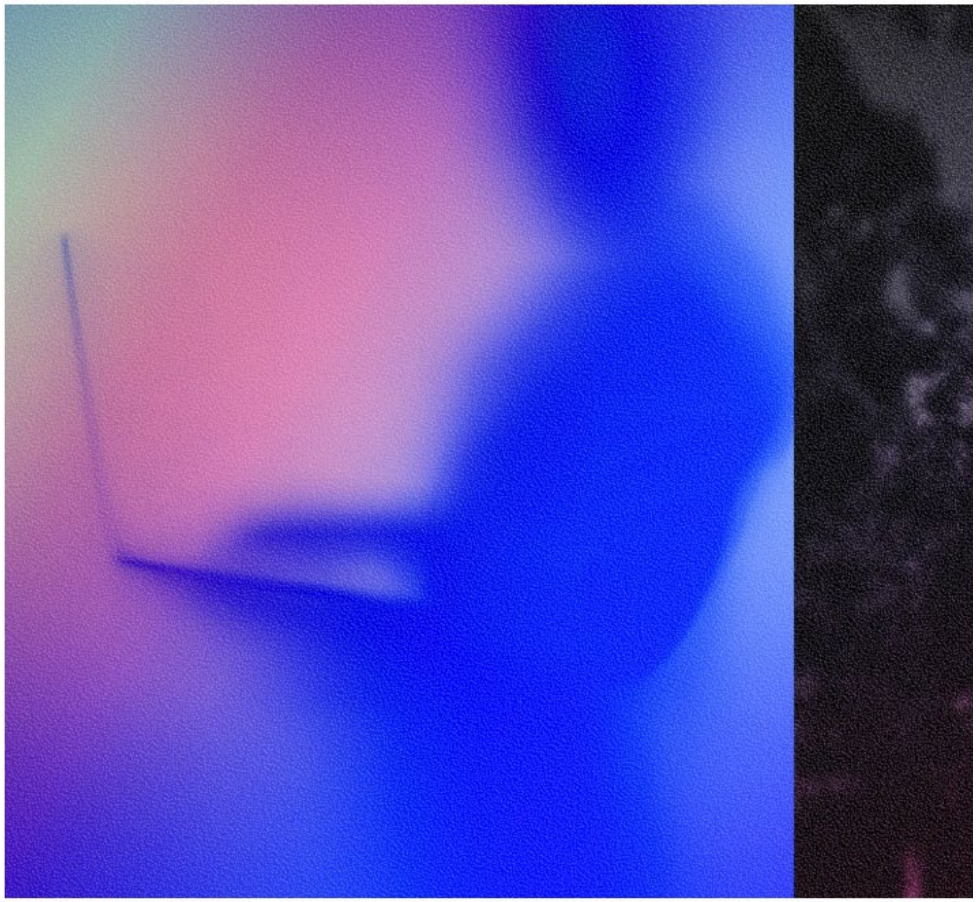


PHOTO-ILLUSTRATION: WIRED; GETTY IMAGES

Science *China / Science*

China Initiative 2.0? Raids on scientist Wang Xiaofeng revive spectre from first Trump era

Two Indiana homes of cybersecurity researcher and professor raided by FBI and Homeland Security but no reason cited, local media report

Reading Time: **3 minutes**

Why you can trust SCMP

Listen

Holly Chik and Dannie Peng in Beijing

Published: 10:03pm, 31 Mar 2025 | Updated: 9:45am, 1 Apr 2025

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

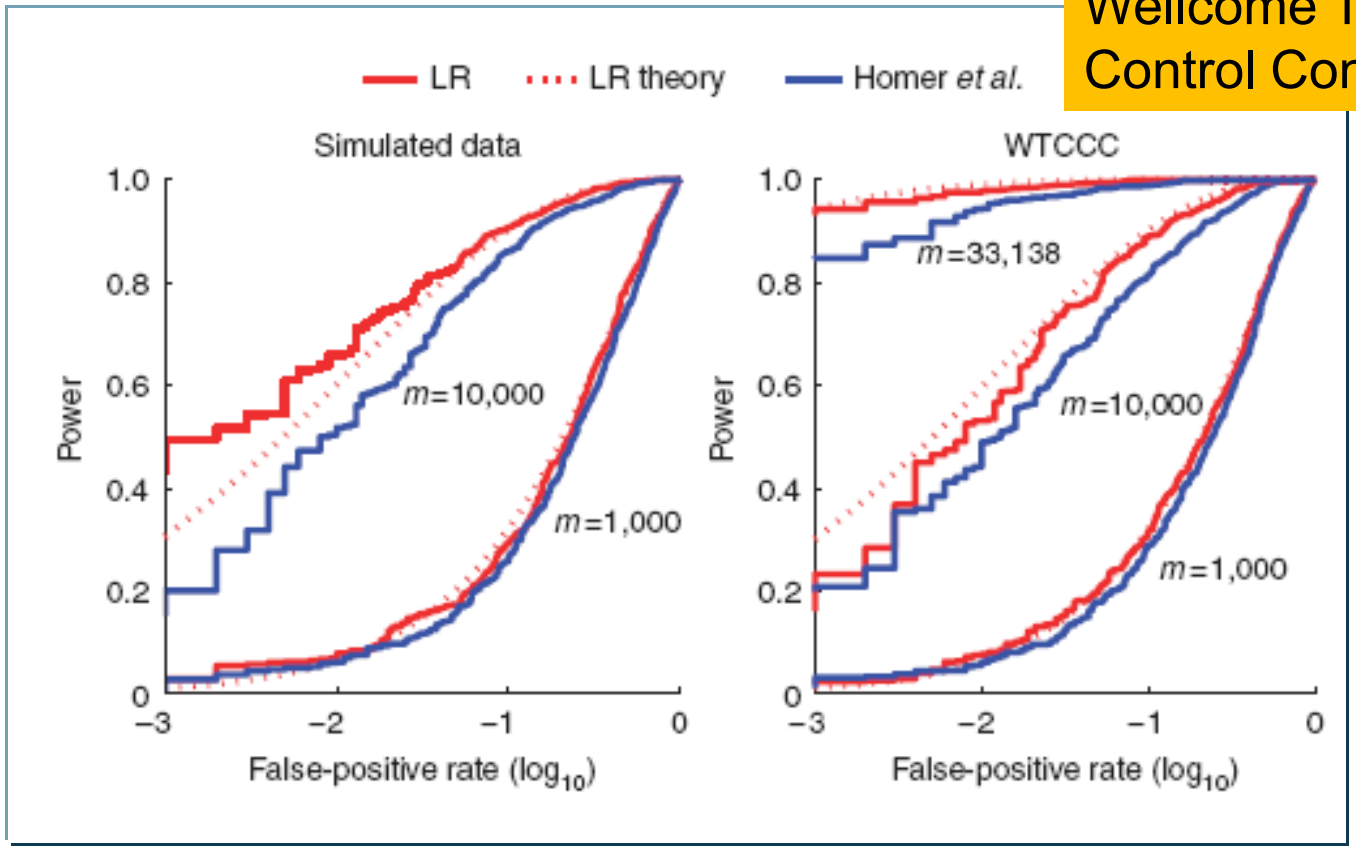
- Implies
  - |S| is linear in *n* for a fixed false positive and negative rate
  - Power of the test does NOT depend on allele frequencies (if the recessive allele is large enough)!

*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.

Readings for the Next Week

- 1. Sweeney L. **k-anonymity: A model for protecting privacy**. *International journal of uncertainty, fuzziness and knowledge-based systems*. 2002 Oct;10(05):557-70.

- Optional
  - ❑ 2. Dankar FK, El Emam K. **A method for evaluating marketer re-identification risk**. *InProceedings of the 2010 EDBT/ICDT Workshops* 2010 Mar 22 (pp. 1-10).
  - ❑ 3. Newton EM, Sweeney L, Malin B. **Preserving privacy by de-identifying face images**. *IEEE transactions on Knowledge and Data Engineering*. 2005 Jan 10;17(2):232-43.

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