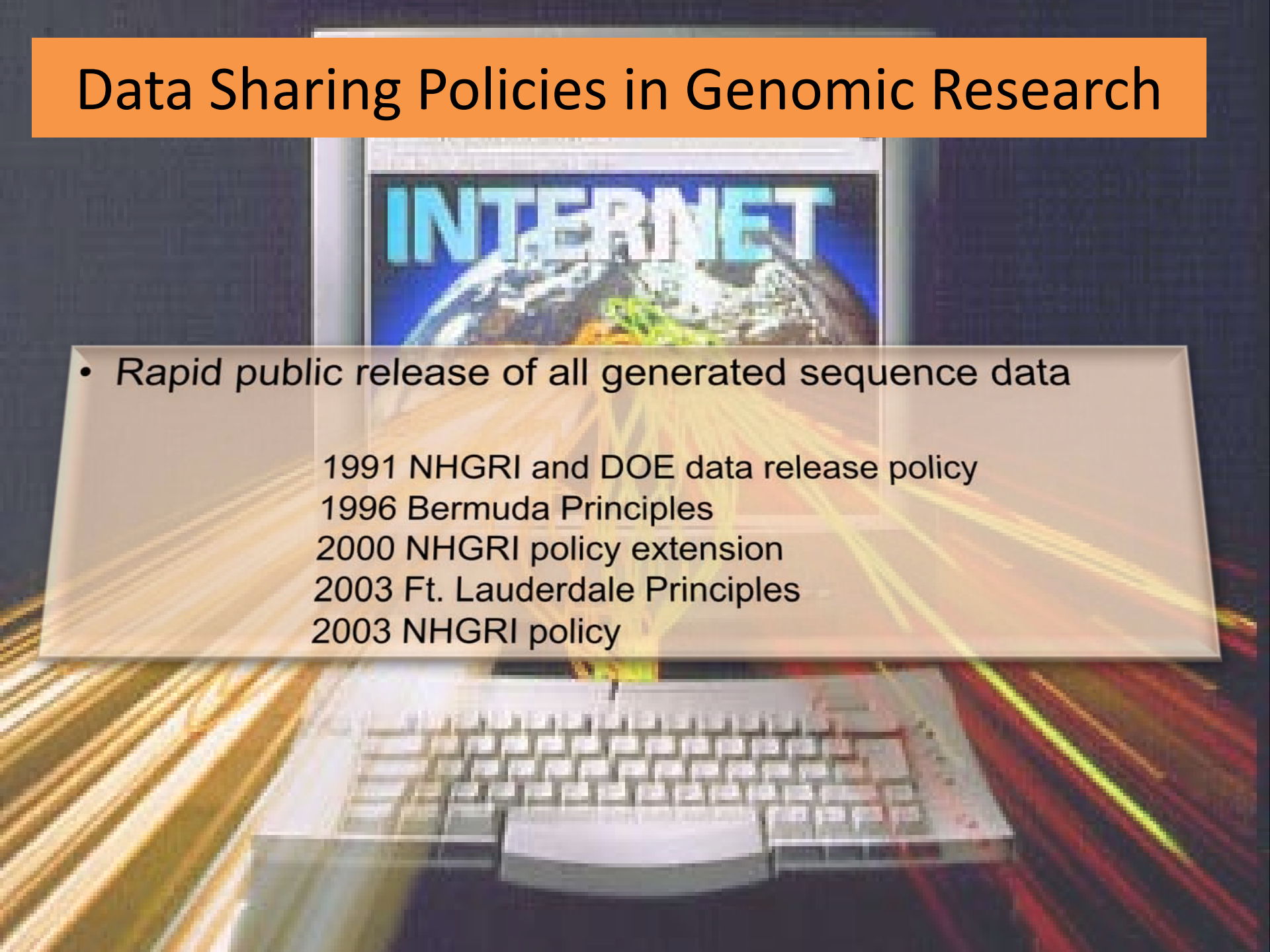
IDENTIFIABILITY OF THE HUMAN MICROBIOME: INVESTIGATOR PERSPECTIVES

Amy McGuire

Center for Medical Ethics and Health Policy

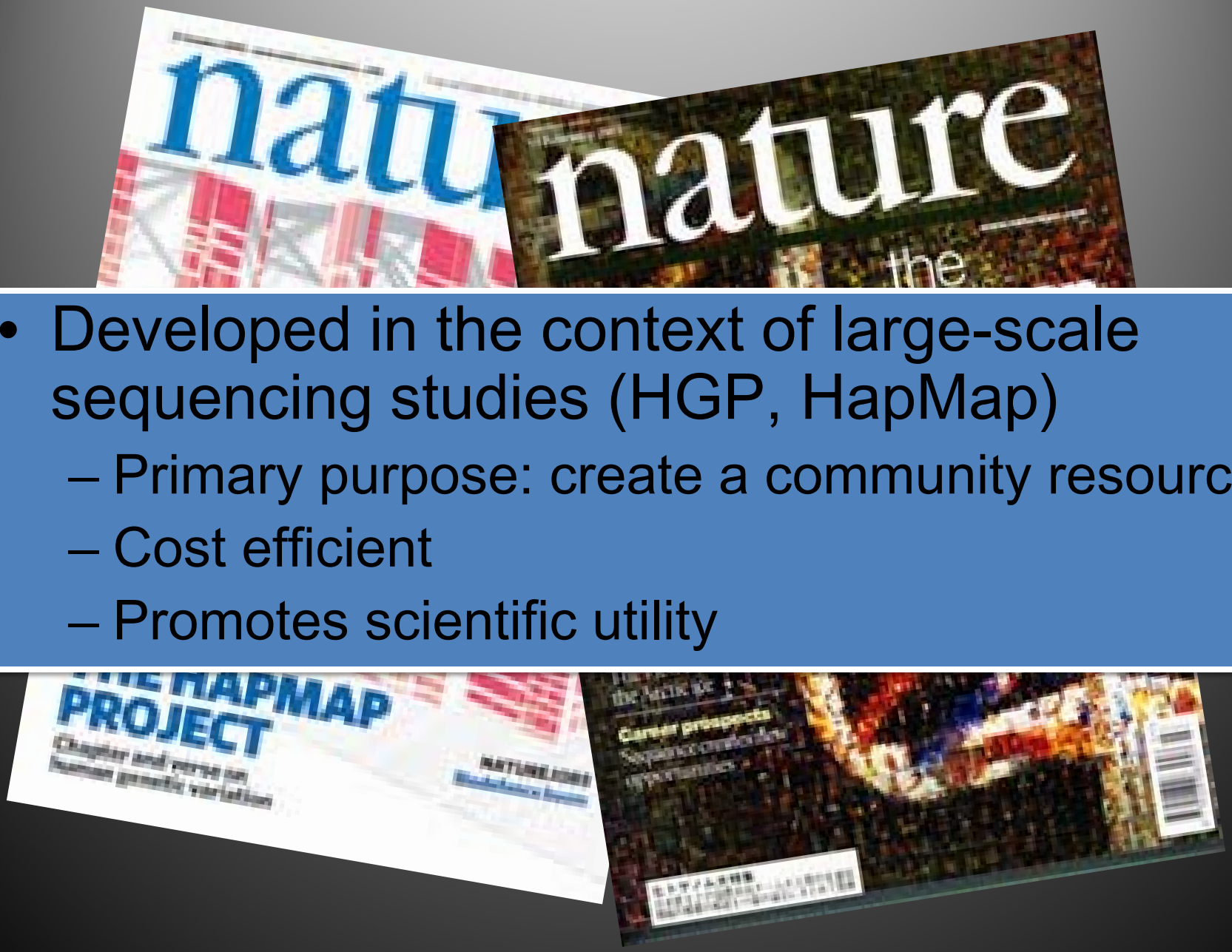
Baylor College of Medicine

Data Sharing Policies in Genomic Research



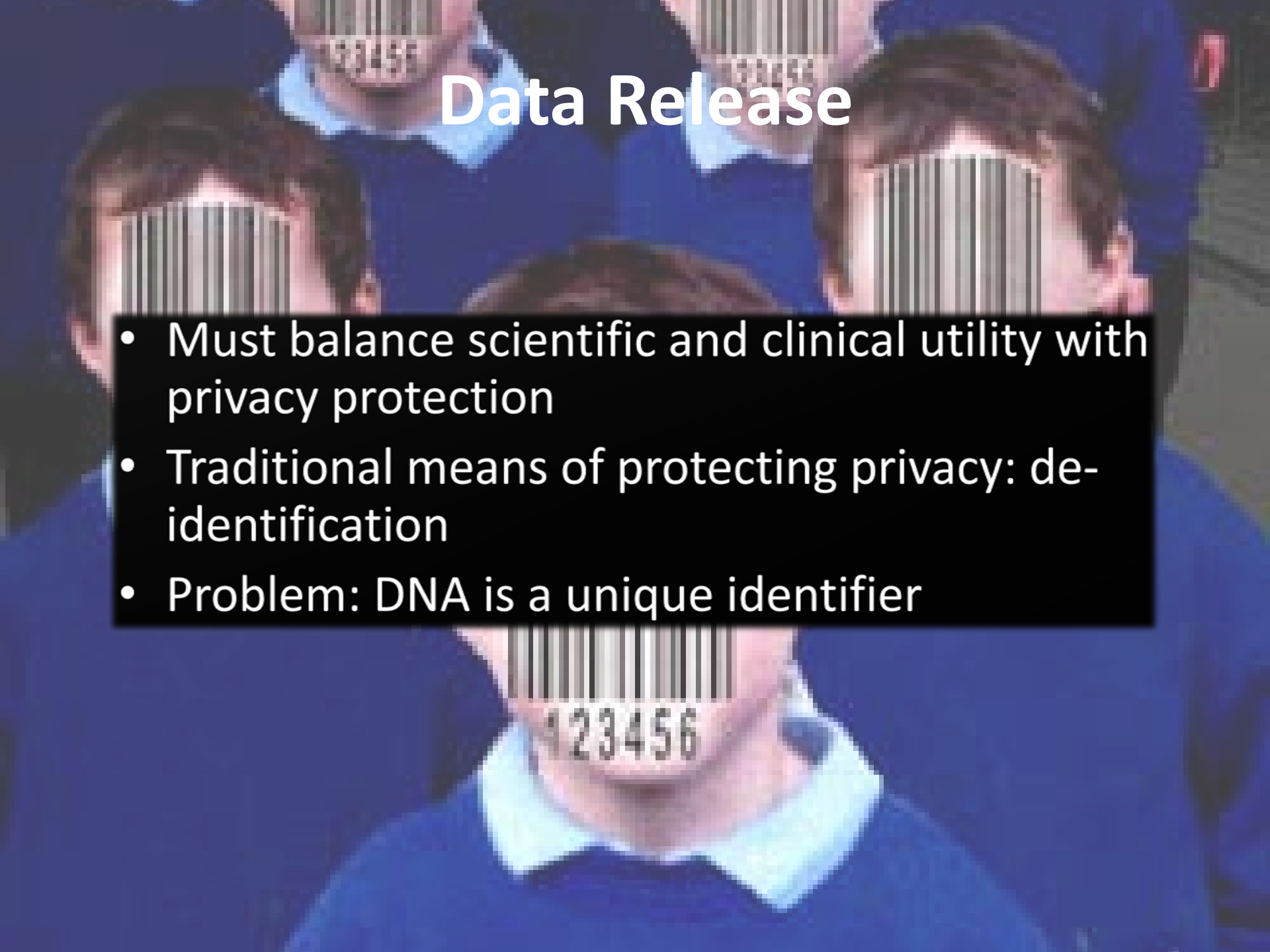
- *Rapid public release of all generated sequence data*

1991 NHGRI and DOE data release policy
1996 Bermuda Principles
2000 NHGRI policy extension
2003 Ft. Lauderdale Principles
2003 NHGRI policy



- Developed in the context of large-scale sequencing studies (HGP, HapMap)
  - Primary purpose: create a community resource
  - Cost efficient
  - Promotes scientific utility

Data Release



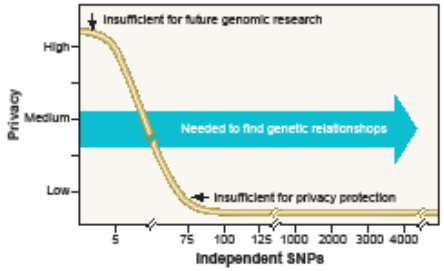
- Must balance scientific and clinical utility with privacy protection
- Traditional means of protecting privacy: de-identification
- Problem: DNA is a unique identifier

GENETICS

Genomic Research and Human Subject Privacy

Zhen Lin,[1] Art B. Owen,[2] Russ B. Altman[1]*

Interest in understanding how genetic variations influence heritable diseases and the response to medical treatments is intense. The academic community relies on the availability of public databases for the distribution of the DNA sequences and their variations. However, like other types of medical information, human genomic data are private, intimate, and sensitive. Genomic data have raised special concerns about discrimination, stigmatization, or loss of insurance or employment for individuals and their relatives (*1*, *2*). Public dissemination of these data poses nonintuitive privacy challenges.



Trade-offs between SNPs and privacy.

Unrelated persons differ in about 0.1% of the 3.2 billion bases in their genomes (*3*). Now, the most widely used forms of forensic identification rely on only 13 to 15 locations on the genome with variable repeats (*4*, *5*). Single nucleotide polymorphisms (SNPs) contain information that can be used to identify individuals (*5*, *6*). If someone has access to individual genetic data and performs matches to public SNP data, a small set of SNPs could lead to successful matching and identification of the individual. In such a case, the rest of the genotypic, phenotypic, and other information linked to that individual in public records would also become available.

The world population is roughly 10^{10}. Specifying DNA sequence at only 30 to 80 statistically independent SNP positions will uniquely define a single person (*7*). Furthermore, if some of those positions have SNPs that are relatively rare, the number that need to be tested is much smaller. If information about kinship exists, a few positions will confirm it. Thus, the transition from *private* to *identifiable* is very rapid (see the figure).

Tension between the desire to protect privacy and the need to ensure access to scientific data has led to a search for new technologies. However, the hurdles may be greater than had been suspected. For example, one approach to protecting privacy is to limit the amount of high-quality data released and randomly to change a small percentage of SNPs for each subject in the database (*8*). Suppose that 10% of SNPs are randomly changed in a sequence of DNA, a fairly major obfuscation that would not please many genetics researchers. Our estimates (*7*) show that measuring as few as 75 statistically independent SNPs would define a small group that contained the real owner of the DNA. Disclosure control methods such as data suppression, data swapping, and adding noise would be unacceptable by similar arguments.

A second approach is to group SNPs into bins. Disregarding exact genomic locations of SNPs increases the number of records that share the same values, thus increasing confidentiality. Our calculations (*7*) show that such strategies do not protect privacy, because the pattern of binned values is unlikely to match anyone other than the owner of the DNA. Data analysis would be greatly complicated by binning, and the information content would be severely reduced or even eliminated.

Until technological innovations appear, solutions in policy and regulations must be found. We are building the Pharmacogenetics and Pharmacogenomics Knowledge Base (*8*, *9*), which contains individual genotype data and associated phenotype information. No genetic data will be provided unless a user can demonstrate that he or she is associated with a bona fide academic, industrial, or governmental research unit and agrees to our usage policies (including audit of data access) (*10*). Although this does not prevent data abuse, it provides a way to monitor usage.

Social concerns about privacy are intricately connected to beliefs about benefits of research and trustworthiness of researchers and governmental agencies. In the United States, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the associated Privacy Rules of 2003 (*11*) generally forbid sharing identifiable data without patient consent. However, they do not specifically address use or disclosure policies for human genetic data. Recent debates in Iceland, Estonia, Britain, and elsewhere (*12*–*15*), reveal a range of views on the threats posed by genetic information. The United States may be at one end of this spectrum, as its citizens seem to strongly desire health privacy. Whatever the setting, we recommend explicit clarifications to rules and legislation (such as HIPAA), so that they explicitly protect genetic privacy and set strong penalties for violations. These clarifications should define entities authorized to use and exchange human genetic data and for what purposes.

[1]Department of Genetics, Stanford University School of Medicine, CA 94305-5120, USA. [2]Department of Statistics, Stanford University, CA 94035-4065, USA.

*To whom correspondence should be addressed. E-mail: russ.altman@stanford.edu

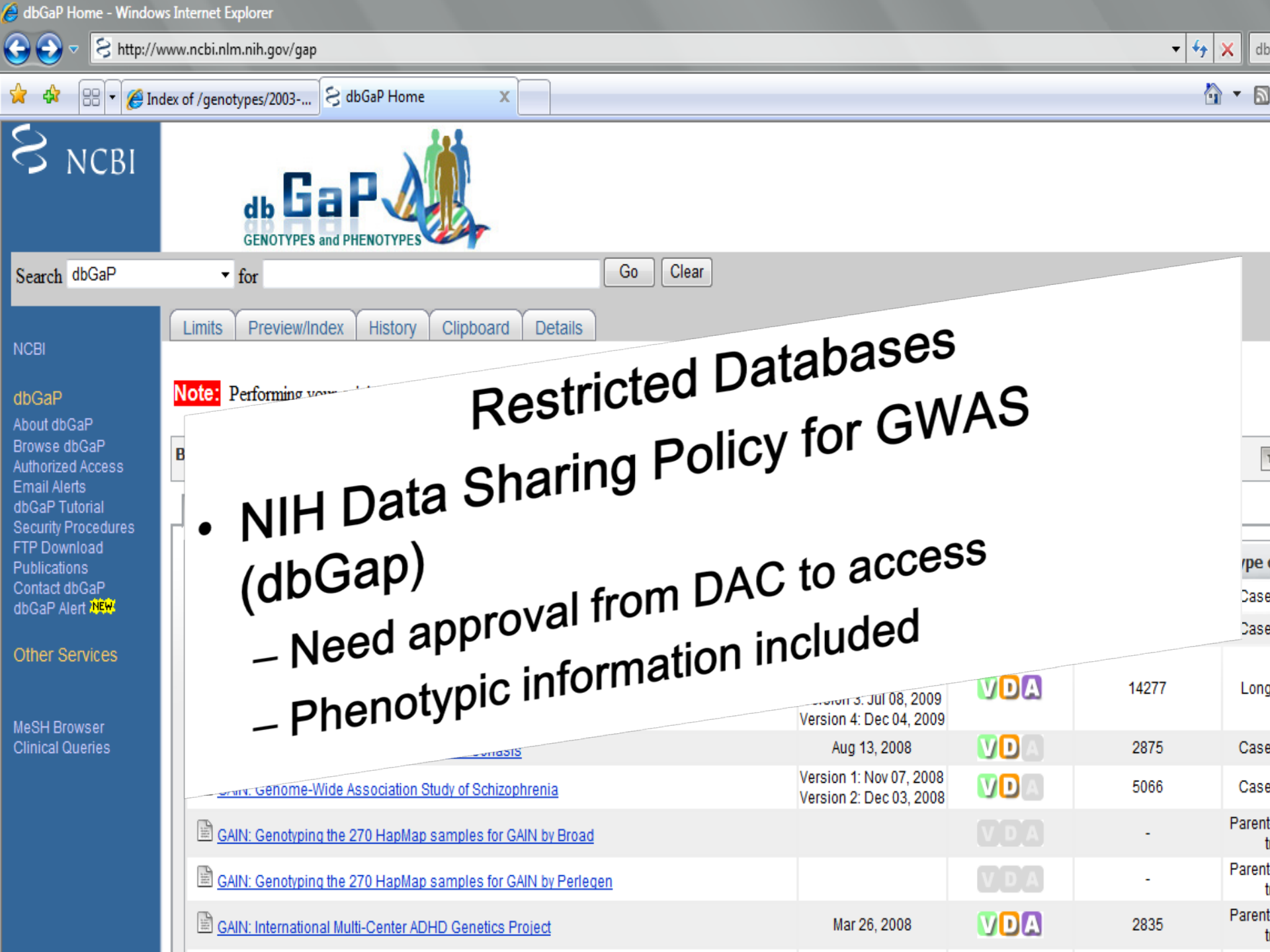
References and Notes

1. M. R. Anderlik, M. A. Rothstein, *Annu. Rev. Genomics Hum. Genet.* 2, 401 (2001).
2. P. Sankar, *Annu. Rev. Med.* 54, 393 (2003).
3. W. H. Li, L. A. Sadler, *Genetics* 129, 513 (1991).
4. L. Carey, L. Mitnik, *Electrophoresis* 23, 1386 (2002).
5. H. D. Cash *et al.*, *Pac. Symp. Biocomput.* 2003, 638 (2003).
6. National Commission on the Future of DNA Evidence, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group* (National Institute of Justice, U.S. Department of Justice, Washington, DC, 2000).
7. See supporting online material for further discussion.
8. L. C. R. J. Willenborg, T. D. Waal, *Elements of Statistical Disclosure Control* (Springer, New York, 2001).
9. T. E. Klein *et al.*, *Pharmacogenomics J.* 1, 167 (2001).
10. www.pharmgkb.org/home/policies/index.jsp
11. *Fed. Regist.* 67, 53181 (2002).
12. R. Chadwick, *BMJ* 319, 441 (1999).
13. L. Frank, *Science* 290, 31 (2000).
14. M. A. Austin *et al.*, *Genet. Med.* 5, 451 (2003).
15. V. Barbour, *Lancet* 361, 1734 (2003).

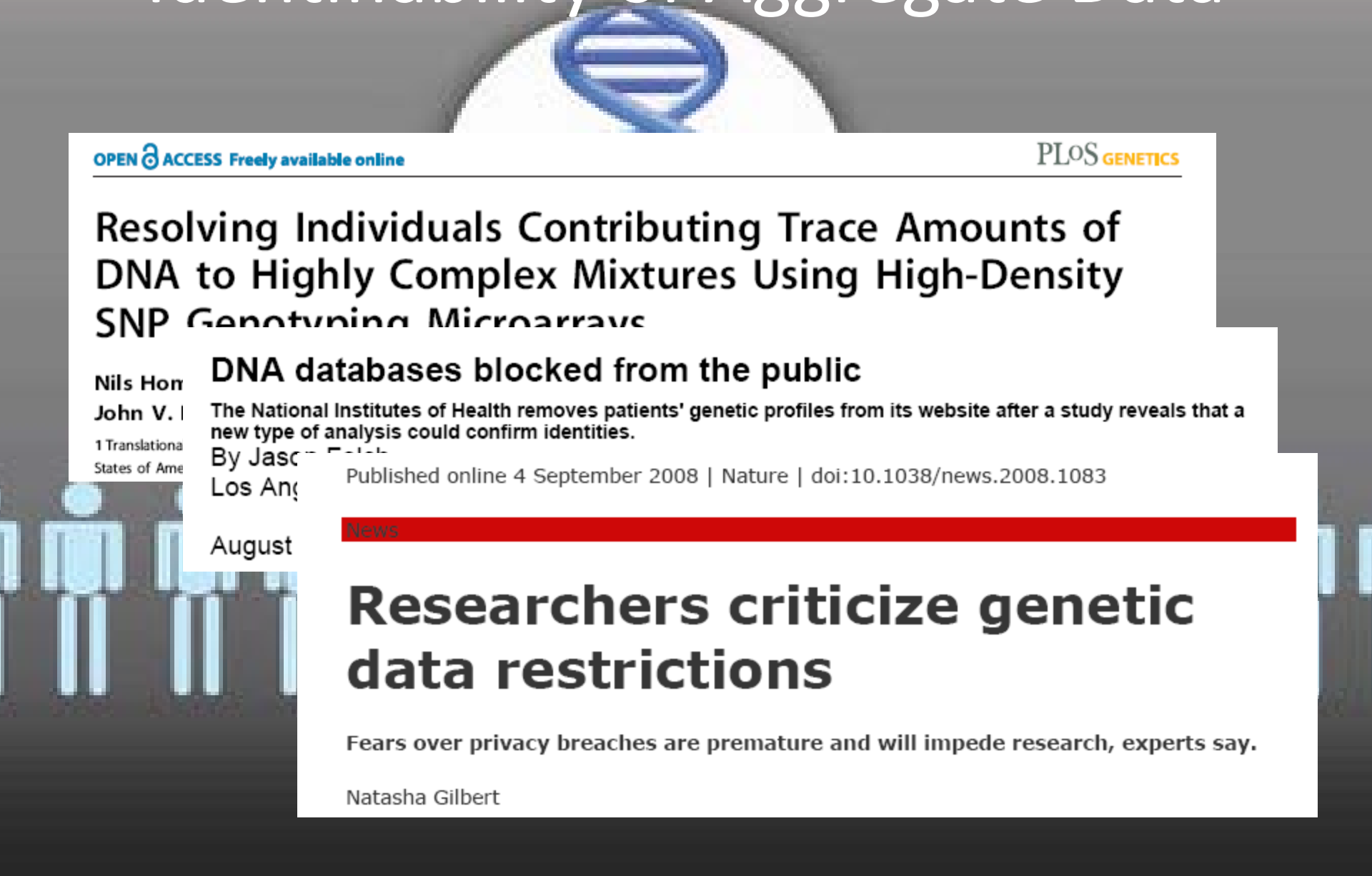
16. Supported in part by NIH/NLM Biomedical Informatics Training Grant LM007033 (Z.L.), NSF Grant DMS-0306612 (A.B.O.), and the NIH/NIGMS Pharmacogenetics Research Network and Database U01-GM61374 (R.B.A.). We thank J. T. Chang, B. T. Naughton, T. E. Klein, and reviewers.

Supporting Online Material
www.sciencemag.org/cgi/content/full/305/5681/183/DC1

"Specifying DNA sequence at only 30 to 80 statistically independent SNP positions will uniquely identify a single person."


dbGaP Home - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/gap
Index of /genotypes/2003-...
dbGaP Home
NCBI
dbGaP
GENOTYPES and PHENOTYPES
Search dbGaP for
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
NCBI
dbGaP
About dbGaP
Browse dbGaP
Authorized Access
Email Alerts
dbGaP Tutorial
Security Procedures
FTP Download
Publications
Contact dbGaP
dbGaP Alert NEW
Other Services
MeSH Browser
Clinical Queries
Note:
Restricted Databases
• NIH Data Sharing Policy for GWAS (dbGap)
– Need approval from DAC to access
– Phenotypic information included
Version 3: Jul 08, 2009
Version 4: Dec 04, 2009
14277
Aug 13, 2008
2875
Version 1: Nov 07, 2008
Version 2: Dec 03, 2008
5066
GAIN: Genotyping the 270 HapMap samples for GAIN by Broad
GAIN: Genotyping the 270 HapMap samples for GAIN by Perlegen
GAIN: International Multi-Center ADHD Genetics Project
Mar 26, 2008
2835

Identifiability of Aggregate Data



OPEN ACCESS Freely available online

PLoS GENETICS

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Hon
John V.
1 Translationa
States of Ame

DNA databases blocked from the public

The National Institutes of Health removes patients' genetic profiles from its website after a study reveals that a new type of analysis could confirm identities.

By Jasc
Los Ang

August

Published online 4 September 2008 | Nature | doi:10.1038/news.2008.1083

News

Researchers criticize genetic data restrictions

Fears over privacy breaches are premature and will impede research, experts say.

Natasha Gilbert

HMP Data Release and Resource Sharing Guidelines

- Guiding principle: pre-publication metagenomic and associated data released to scientific community as rapidly as possible via deposition into public databases.
- Potentially identifying data submitted to dbGaP.
- http://commonfund.nih.gov/hmp/datareleaseguidelines.aspx

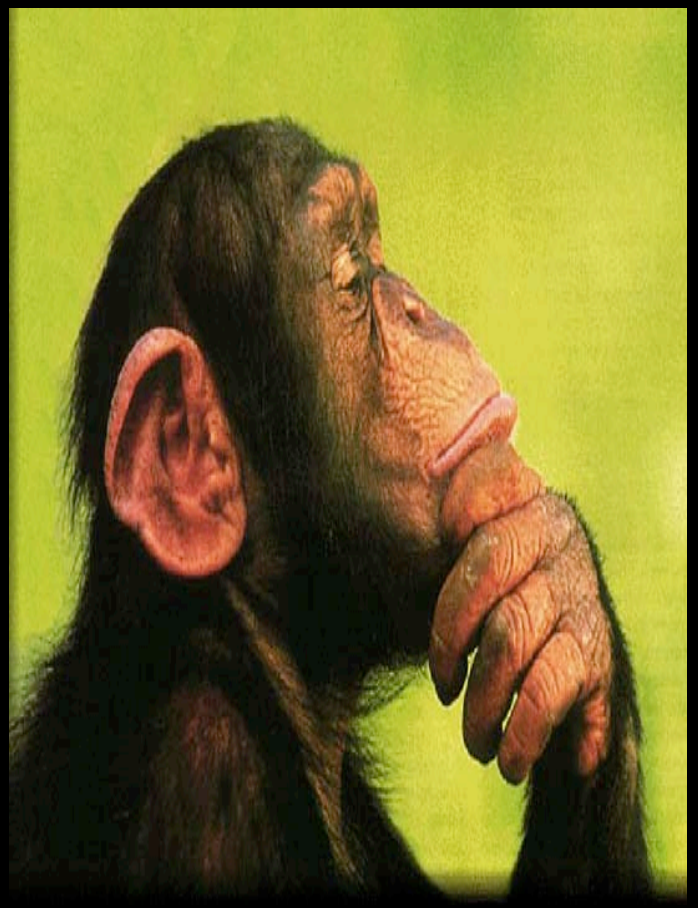
HMP Consent Form

- Clinical data: coded and submitted to dbGaP
- Microbiome data: coded and placed in open access (public) database
  - "Every effort [will be made] to remove any human DNA data from the microbe genetic data"
- Human DNA from blood: aggregate data publicly released; individual level data in dbGaP

"There is a small risk that someone outside the project could learn some information about you."

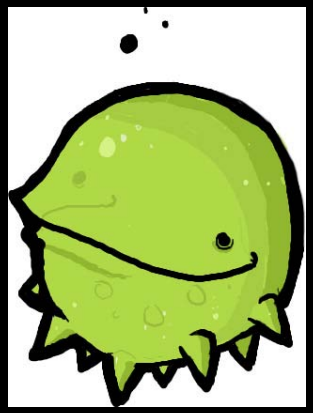
Ethical, Legal, and Social Dimensions of Human Microbiome Research (NIH 1 R01HG004853)

- **Overall Goal: Identify and analyze ethical challenges associated with human microbiome research from the perspective of stakeholders**
- **Method: In-depth interviews**
  - **Investigators/Project Leaders (n=63)**
  - **BCM Jumpstart Recruits (n=50)**



Preliminary Findings: Investigator Perspectives

- 3 issues related to data sharing and identifiability
  - Human contamination
  - Identifiability of the microbiome
  - Linking human DNA, metadata, and microbiome



Human Contamination

- Problem: microbiome DNA contains some human DNA (novel variants)
- Filtering options:
  - Screen against human genome sequence and only release what doesn't hit human genome
  - Screen against known bacterial sequence and only release what hits known bacterial sequence
  - Investigator explanation: "by doing the latter, you're potentially losing a lot of novel microbial information… by doing the former approach you're potentially releasing some human genome sequence to the public…"

Investigator Perspectives on Human Contamination

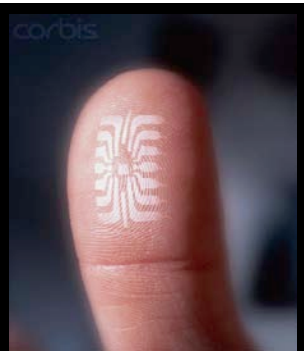
- Many feel filtering efforts are appropriate
- But others are more critical
  - They consent to the risks: "I go back and forth on how much I worry about it—because I think, in general … they give up their DNA samples and they sign a consent, and they have a certain amount of understanding of what's going to happen with their information."
  - The risks are small: "I know that in theory if someone wanted to invest a lot of effort that maybe they could trace it back to the individual. But it's like, come on… what are the real consequences of that?"
  - It's a waste of money: "we're not achieving the de-identification, and we're spending a lot of money pretending we are."

Identifiability of Microbiome: Is My Microbiome Part of My Identity?

- Some think of the microbiome as part of who we are
  - "I think your microbiome is you. It has a huge impact on you, so it's really you."
- Others think of it as separate from us
  - "[My subjects] don't view *yet* that their microbial DNA is part of their genetic landscape. They still have the sense of 'other'to it, so that they don't sense that we are learning something about them." (emphasis added)

Identifiability of Microbiome: Is My Microbiome Unique To Me?

- Some think it is unique and thus identifiable
  - "I'm quite convinced that the day will come when microbiome analysis will be fairly unique to an individual just like their own DNA is."
- Others don't or are uncertain
  - "I have a hard time believing that you could identify an individual from the microbiome."
  - "I guess one of the issues is stability of the microbiome."





Forensic identification using skin bacterial communities

Noah Fierer[a,b,1], Christian L. Lauber[b], Nick Zhou[b], Daniel McDonald[c], Elizabeth K. Costello[c], and Rob Knight[c,d]

[a]Department of Ecology and Evolutionary Biology, [b]Cooperative Institute for Research in Environmental Sciences, and [c]Department of Ch
Biochemistry, University of Colorado, Boulder, CO 80309; and [d]Howard Hughes Medical Institute

Edited by Jeffrey I. Gordon, Washington University School of Medicine, St. Louis, MO, and approved February 13, 2010 (received for review J

Recent work has demonstrated that the diversity of skin-associated bacterial communities is far higher than previously recognized, with a high degree of interindividual variability in the composition of bacterial communities. Given that skin bacterial communities are personalized, we hypothesized that we could use the residual skin bacteria left on objects for forensic identification, matching the bacteria on the object to the skin-associated bacteria of the individual who touched the object. Here we describe a series of studies demonstrating the validity of this approach. We show that skin-associated bacteria can be readily recovered from surfaces (including single computer keys and computer mice) and that the structure of these communities can be used to differentiate objects handled by different individuals, even if those objects have been left untouched for up to 2 weeks at room temperature. Furthermore, we demonstrate that we can use a high-throughput pyrosequencing-based approach to quantitatively compare the bacterial communities on objects and skin to match the object to the individual with a high degree of certainty. Although additional work is needed to further establish the utility of this approach, this series of studies introduces a forensics approach that could eventually be used to independently evaluate results obtained using more traditional forensic practices.

bacterial forensics | human microbiome | pyrosequencing | skin microbiology | microbial ecology

The human skin surface harbors large numbers of bacteria that can be readily dislodged and transferred to surfaces upon touching, hence the importance of proper hand hygiene by health care practitioners (1, 2). These skin bacteria may persist on touched surfaces for prolonged periods because many are highly resistant to environmental stresses, including moisture, temperature, and UV radiation (3, 4). Therefore, we likely leave a persistent "trail" of skin-associated bacteria on the surfaces and objects that we touch during our daily activities.

Recent work has demonstrated that our skin-associated bacterial communities are surprisingly diverse, with a high degree of interindividual variability in the composition of bacterial communities at a particular skin location (5–9). For example, only 13% of the bacterial phylotypes on the palm surface are shared between any two individuals (8), and a similar level of interpersonal differentiation is observed at other skin locations (5, 9). In addition, skin bacterial communities are relatively stable over time: palm surface bacterial communities recover within hours after hand washing (8); and, on average, interpersonal variation in community composition exceeds temporal variation within people, even when individuals are sampled many months apart (5, 9). Given that individuals appear to harbor personally unique, temporally stable, and transferable skin-associated bacterial communities, we hypothesized that we could use these bacteria as "fingerprints" for forensic identification.

To demonstrate that we can use skin bacteria to link touched surfaces to specific individuals, the following criteria must be met: *(i)* bacterial DNA recovered from touched surfaces allows for adequate characterization and comparison of bacterial communities; *(ii)* skin bacterial communities persist on surfaces for days to weeks; and *(iii)* surfaces that are touched can be effectively linked to individuals by assessing the degree of similarity between the bacterial communities on the object and the skin of the individual who touched the object. To establish these criteria and to demonstrate the potential utility of the approach for forensic identification, we carried out three interrelated studies that combine recent developments in phylog
nity analyses (10) with high-throughput pyrosequer
(11). First, we compared bacterial communities on in
three computer keyboards to the communities found o
the keyboard owners. Second, we examined the sim
skin-associated bacterial communities on objects st
(a standard method for storing samples before DN
versus those objects stored under typical indoor envi
ditions for up to 14 days. Finally, we linked objects t
viduals by comparing the bacteria on their computer
database containing bacterial community information
250 hand surfaces, including the hand of the owner.

Results and Discussion

To establish criteria *i* and *iii*, we swabbed individual k
personal computer keyboards and compared the c
those keys to the bacterial communities on the finger
board owners. We also sampled individual keys from
and public computer keyboards so that we could quan
of correspondence between the bacterial communities
fingers and keyboard versus other keyboards never t
person. Bacterial DNA was extracted from the swab
community composition was determined using the t
sequencing procedure described previously (8), obtain
of over 1,400 bacterial 16S rRNA gene sequences p
found that bacterial communities on the fingertips or keyboard of a given individual are far more similar to each other than to fingertips or keyboards from other individuals (Fig. 1 and Fig. 2). Likewise, the bacterial communities on the fingers of the owner of each keyboard resembled the communities on the owner's keyboard (Fig. 1 and Fig. 2), which suggests that differences in keyboard-associated communities are likely caused by direct transfer of fingertip bacteria. The discrimination between individuals is stronger with the unweighted UniFrac metric than with the weighted metric, suggesting that differences in community membership (rather than community structure) discriminate best among individuals. The patterns evident in Fig. 1 are confirmed by ANOSIM analyses, which demonstrate that each keyboard harbors a distinct bacterial community, the finger-associated bacterial communities are unique to each of the three individuals, and that the interindividual differences in fingertip and keyboard communities exceed the differences between bacterial communities on the fingers and keyboards belonging to a given individual (Table S1). Together these results demonstrate that bacterial DNA can be recovered from relatively small surfaces, that the composition of the keyboard-associated communities are distinct across the three keyboards, and that individuals leave unique bacterial 'fingerprints' on their keyboards.

Author contributions: N.F., C.L.L., N.Z., and R.K. designed research; N.F., C.L.L., N.Z., and E.K.C. performed research; D.M. contributed new reagents/analytic tools; N.F., C.L.L., D.M., E.K.C., and R.K. analyzed data; and N.F. and R.K. wrote the paper.

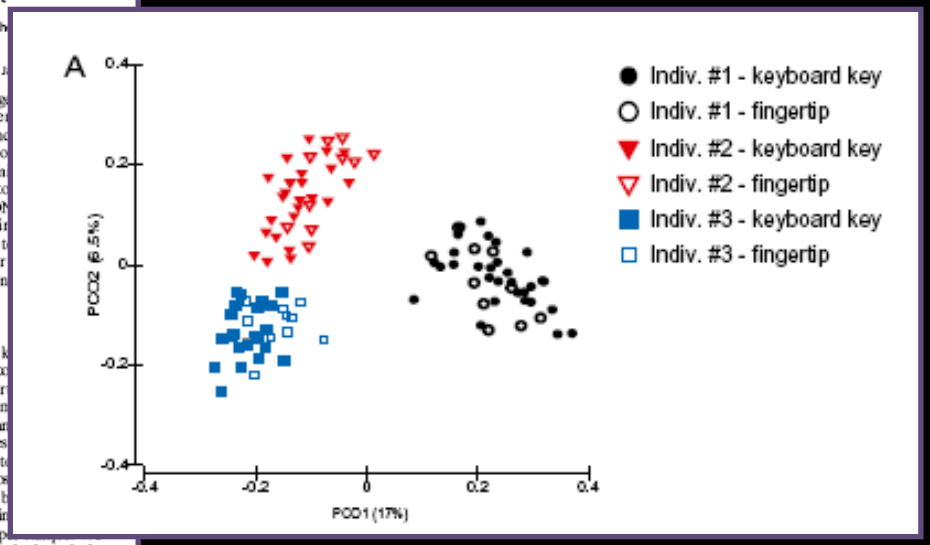
The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data have been deposited in the GenBank Short Read Archive (SRA010034.1).

[1]To whom correspondence should be addressed. E-mail: noah.fierer@colorado.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/1000162107/DCSupplemental.



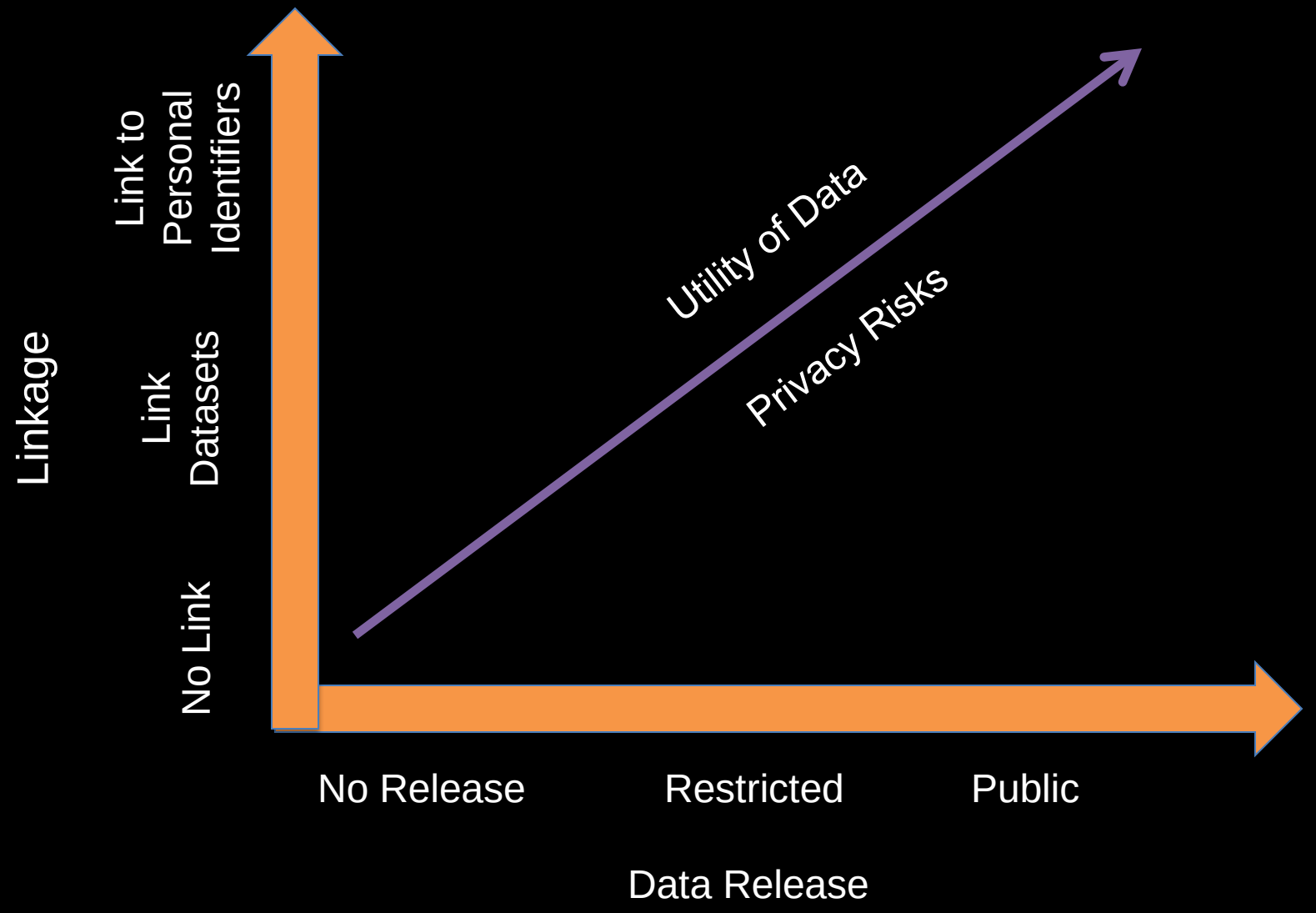
Together these results demonstrate that bacterial DNA can be recovered from relatively small surfaces, that the composition of the keyboard-associated communities are distinct across the three keyboards, and that individuals leave unique bacterial 'fingerprints' on their keyboards.

Linking Human DNA, Metadata, and Microbiome

The Missing Link

- Linking microbiome to human DNA
  - "when you understand something about the microbiome, you need to know something about the host eventually…at some point linking those pieces of information clearly makes sense"
- Linking microbiome to metadata
  - "I think the main issue related to data sharing is the lack of metadata… associated with the sample…it's hindering data sharing…. there are efforts on-going trying to establish standards of data sharing among groups of researchers, but I think that effort has been…initiated a little too late."

Policy Options



Investigator Perspectives

- Risks of harm are low but uncertain so protect
  - "I don't think it's going to be a problem, but because we don't know for sure, that's the kind of thing that makes us anxious."
  - "I wouldn't necessarily assume that anyone would have both the motivation and the time… as well as just the lack of ethics, to use it for some undesirable purpose. But… it is always a possibility even though it is remote. So one does not want to be in a situation where I wasn't careful enough about patient data."
  - "You really wouldn't want to see that on the front page that if we messed up and someone got identified."

Investigator Perspectives

- Benefits of public release
  - "there's about a ten to a hundredfold factor of the number of people who use open access and controlled access, so we clearly want this data in open access where it can be most used by the research community."
- Benefits outweigh the risks
  - "I think we've complicated things a whole lot by raising all kinds of concerns about privacy…if we inadvertently sequence human DNA…"so what?" that's the only risk that's entailed in this study, and to me that's a very remote risk, greatly, greatly outweighed by actual potential benefit to society."

- They know what they are getting into
  - "[The subjects] signed consent forms, knowing that their sequence could be released into the public domain..."
- You can't plan for everything
  - "You know, there are things we don't know. Okay, so you don't know what you don't know."
- Need greater accountability for misuse
  - "I think the data can and should be disseminated, but those individuals need legal protection... In the event that their identity is revealed, the onus should be on the people who misuse the information, rather than the people who are providing the information."

Lessons From Genomics and HMP Participants

- Participants worry about their privacy
- They generally trust researchers
- They want their samples/data used
- They are comfortable with broad data sharing and linkage to clinical data, but not to personal identifiers
- They want to be asked
- It's really about *RESPECT*

- **Funding**
  - **NHGRI-ELSI**
- **Collaborators**
  - **Sheryl McCurdy, Simon Whitney, Laura Achenbaum, Melody Slaskinski**
  - **Wendy Keitel, Jim Versalovic, Richard Gibbs**
  - **Invesgtigators, project leaders, and HMP recruits who participated in this study. THANK YOU!**