GENETICS

No Longer De-Identified

Sequencing human DNA to discover genetic variation should be governed by existing regulations for human subjects.

Amy L. McGuire[1]* and Richard A. Gibbs[2]

As DNA sequencing becomes more affordable and less time-consuming, scientists are adding DNA banking and analysis to research protocols, resulting in new disease-specific DNA databases. A major ethical and policy question will be whether and how much information about a particular individual's DNA sequence ought to be publicly accessible.

Without privacy protection, public trust will be compromised, and the scientific and medical potential of the technology will not be realized. However, scientific utility grows with increased access to sequenced DNA. At present, ethical concerns about the privacy of subjects whose sequenced DNA is publicly released have largely been addressed by ensuring that the data are "de-identified" and that confidentiality is maintained (*1*–*2*). There is a large literature on the various data-management models and computer algorithms that can be used to provide access to genetic data while purportedly protecting privacy (*3*–*6*). We believe that minimizing risks to subjects through new developments in data and database structures is crucial and should continue to be explored, but that additional safeguards are required.

Scientists have been aware for years of the possibility that coded or "anonymized" sequenced DNA may be more readily linked to an individual as genetic databases proliferate (*1*, *3*, *7*, *8*). In 2004, Lin and colleagues demonstrated that an individual can be uniquely identified with access to just 75 single-nucleotide polymorphisms (SNPs) from that person (*9*). Genome-wide association studies routinely use more than 100,000 SNPs to genotype individuals. Although individual identification from the public release of these data would currently require a reference sample, the privacy risk associated with public data release is fueled by the extraordinary pace of technological developments and the rapid proliferation of electronic databases. If protective measures are not adopted now, public trust will be compromised, and genomic research will suffer.

Genetic sequencing typically involves three phases of investigation: (i) subject recruitment and sample collection (primary clinical investigation), (ii) DNA sequencing and data broadcast (genomic sequencing study), and (iii) data retrieval and analysis (secondary-use research)

From subject to data analysis. A typical medical genomic sequencing study.

(see figure, above). Institutional Review Board (IRB) oversight and informed consent are unambiguously required for the first phase of sample collection, because it clearly involves human subjects research. There are also detailed consent requirements for some large-scale sequencing studies, such as the HapMap project, that cover the second and third phases. However, it is our experience that, in general, the consent process for most disease-specific genetic research is not protective for these phases and that the privacy risks associated with public data-sharing are not stated. Consent for these studies is highly variable, and in most cases, subjects are simply told that genetic analysis will be performed, without any explanation of what that means or with whom the resulting data will be shared. Further, participants are typically not offered the opportunity to participate in the research if they do not want their data publicly broadcast (*10*).

In the United States, there are now two federal regulations that could potentially apply to such studies—the Common Rule, which regulates all federally funded research and sets forth the federal policy for the protection of human research subjects (*11*) and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, which restricts certain unauthorized uses and disclosures of patients' identifiable protected health information by covered entities (*12*). Neither one specifically mandates IRB oversight or subject consent for the public release of sequenced data. The Common Rule would not apply if genomic sequencing studies were not considered to constitute human subjects research. Human subjects research is defined under the Common Rule as research involving "an individual about whom the investigator … obtains data through intervention or interaction with the individual, or identifiable private information" (*11*). According to a guidance document published in 2004 by the Office for Human Research Protections (OHRP), because the data are collected and coded by the primary clinical investigator, and the sequencing investigator is prohibited from deciphering the code, the data are not considered identifi-

[1]Center for Medical Ethics and Health Policy, Baylor College of Medicine, [2]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Suite 310D, Houston, TX 77030, USA. *Author for correspondence. E-mail: amcguire@bcm.edu

Published by AAAS

able, and the sequencing study is not subject to federal regulation (*13*). Brief IRB review may be necessary to confirm that the research does not involve human subjects, but once that determination is made, no IRB oversight or informed consent is mandated.

Similarly, HIPAA does not provide unambiguous protection because it has not been clear that genomic data constitute readily identifiable protected health information (*14*–*16*), and even if they do, institutions vary in whether the scientists who conduct sequencing studies are considered "covered entities" who must comply with HIPAA. Institutions may choose to impose more stringent requirements on investigators, but there is no federal mandate to do so now.

To resolve the tension between privacy protection and access to DNA data needed for progress in medical research, Lin and colleagues recommend a tiered data-access approach, with sensitive data layers masked from full public view (*9*). This approach has the advantage of minimizing privacy risks without unduly sacrificing progress, but suffers from a lack of flexibility to respond to individual preferences and judgments. It also threatens to slow the pace of research, because current policy calls for sequenced data to be released publicly within 24 hours of their generation (*17*), whereas obtaining approval to access restricted databases could take weeks, if not months. We believe that restricted access should be offered as an option to subjects, but that it not be adopted as a general approach for all genomic sequencing studies.

Kohane and Altman propose that researchers specifically seek out volunteers who are most willing to have their health data publicly shared and that these subjects have explicit control over who has access to their data (*18*). Relying only on such "information altruists" to participate in genetic studies would potentially create subject bias, influencing the ability of investigators to identify disease alleles relevant to the population at large.

We propose that general safeguards be put in place to encourage understanding of and trust in genomic sequencing studies. As an essential first step, genomic sequencing studies should be recognized as human subjects research and brought unambiguously under the protection of existing federal regulation. This would have the effect of mandating informed consent for public release of potentially identifiable sequenced data and requiring IRB oversight to ensure that risks to subjects are minimized and that informed consent, or a waiver of the requirement to obtain informed consent, is obtained (*11*).

Specifically, we recommend a stratified consent process in which all subjects who participate in future genomic sequencing studies are fully informed about how their DNA data may be broadcast and have the authority to decide with whom they want their data shared (*19*). A number of options could be presented; we propose three levels of confidentiality (see table above). A more rigorous assessment of the risks and benefits associated with each of these options will be necessary.

SUGGESTED DATA RELEASE OPTIONS

	Data released	Identifiers	Privacy risk	Benefit
Option 1	Multiple gene loci	>75 SNPs	Higher	Ability to study interactions among different genes
Option 2	Single gene loci	Typically <20 SNPs	Intermediate	Ability to study individual genes
Option 3	No data released	None	Low	Ability to include subjects with low risk tolerance, which increases generalizability and applicability to specific subsets of the population

Some of the practical challenges include providing adequate disclosure and education about a complex risk calculus, ensuring subject comprehension, coordinating a system of restricted access, and managing a complex database that accounts for subjects' informed disclosure preferences. Although it may represent a substantial departure from the traditional model of informed consent in research, stratified consent procedures are commonly used in clinical medicine, where patients frequently make informed choices about treatment options on the basis of individual values and judgments. Stratified consent procedures are also being considered in other areas of research where subjects have to make complicated decisions, such as what type of future research they are willing to participate in and to what extent they want research-related incidental findings reported back to them.

There may be concern about the added burdens on IRBs. McWilliams and colleagues have shown that there is currently considerable variability among local IRBs, particularly in how they deal with DNA banking, risk-benefit analysis, and consent for genetic research (*20*). They recommend centralized IRB oversight for multicenter research.

Although some might fear a negative impact on subject participation in genomic research, stratified consent merely restricts the ability to release sequenced data publicly. If anything, it may boost enrollment by providing an opportunity for even the most risk-averse members of society to participate in research, while ensuring optimal privacy protection.

Empirical study of these and other challenges associated with the implementation of a stratified consent model in research is essential for future policy development. In addition, federal legislation prohibiting genetic discrimination would significantly alter the risk-benefit calculus associated with public data release and should be enacted without delay (*21*). Although it would not obviate the ethical obligation to obtain subject consent, it may foster public trust and positively affect the willingness of subjects to participate in genomic research and to share their genetic data publicly.

References and Notes

1. National Human Genome Research Institute (NHGRI), U.S. National Institutes of Health (NIH), *NHGRI-DOE* [U. S. Department of Energy] *Guidance on Human Subjects Issues in Large-Scale DNA Sequencing* (17 August 1996, updated 27 April 1998) (www.genome.gov/10000921).
2. The International HapMap Consortium, *Nat. Genet.* **5**, 467 (2004).
3. L. Sweeney, *J. Law Med. Ethics* **25**, 98 (1997).
4. B. A. Malin, *J. Am. Med. Inform. Assoc.* **12**, 28 (2005).
5. J. E. Wylie, G. P. Mineau, *Trends Biotechnol.* **21**, 113 (2003).
6. L. Sweeney, Ed., "Navigating computer science research through waves of privacy concerns: Discussions among computer scientists at Carnegie Mellon University" *ACM Comput. Soc.* **34** (April 2004); (http://privacy.cs.cmu.edu/dataprivacy/projects/csresearch1.pdf).
7. B. Malin, L. Sweeney, *J. Biomed. Inform.* **37**, 179 (2004).
8. National Cancer Institute (NCI), NIH, "Confidentiality, data security, and cancer research: Perspectives from the National Cancer Institute" (NCI, Bethesda, MD, 23 March 1999); (www3.cancer.gov/confidentiality.html).
9. Z. Lin, A. B. Owen, R. B. Altman, *Science* **305**, 183 (2004).
10. The International HapMap Consortium, "Template consent form" (www.hapmap.org/consent.html.en).
11. "Protection of human subjects," 45 Code of Federal Regulations (C.F.R.) § 46 (2005).
12. "Security and privacy," 45 C.F.R § 164 (2002).
13. Office for Human Research Protections, U.S. Department of Health and Human Services (HHS), "Guidance on research involving coded private information or biological specimens" (HHS, Washington, DC, 10 August 2004); (www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf).
14. HHS, "Research repositories, databases, and the HIPAA Privacy Rule" (HHS, Washington, DC, January 2004); (http://privacyruleandresearch.nih.gov/pdf/research_repositories_final.pdf).
15. P. A. Roche, *Science* **307**, 1200 (2005).
16. M. A. Rothstein, *J. Law Med. Ethics* **33**, 89 (2005).
17. NHGRI, "Reaffirmation and extension of NHGRI rapid data release policies: Large-scale sequencing and other community resource projects" (NIH, Bethesda, MD, February 2003); (www.genome.gov/10506537).
18. I. S. Kohane, R. B. Altman, *N. Engl. J. Med.* **353**, 2074 (2005).
19. Human Genome Organization (HUGO) Ethics Committee, "Statement on human genomic databases" (HUGO, London, December 2002); (www.gene.ucl.ac.uk/hugo/HEC_Dec02.html).
20. R. McWilliams *et al.*, *JAMA* **290**, 360 (2003).
21. Genetic Information Nondiscrimination Act of 2005, S. 306 (109th Congress).

22. We thank S. E. McGuire, B. A. Brody, L. B. McCullough, M. A. Majumder, and R. R. Sharp. R.A.G. is supported by grants from the NHGRI.

10.1126/science.1125339

No Longer De-Identified

Amy L. McGuire and Richard A. Gibbs

Science **312** (5772), 370-371.
DOI: 10.1126/science.1125339