RESEARCH ETHICS

The Complexities of Genomic Identifiability

Recent work reveals the need to re-examine the current paradigms for managing the potential identifiability of genomic and other "omic"-type data.

Laura L. Rodriguez,[1] Lisa D. Brooks,[1] Judith H. Greenberg,[2] Eric D. Green[1*]

Sharing research data has long been fundamental to the advancement of science. In today's scientific culture, making research data available broadly and efficiently via the internet has become the standard for many data types, including genomic and some other "omic"-type data produced by high-throughput methods. The acceleration of research progress and the resulting public benefit achieved through such broad data-sharing have been transformative for the scientific enterprise (*1*–*3*). However, sharing data generated from human research participants must be done in a manner that appropriately protects participant interests.

Several recent studies have suggested that some analyses of high-dimensional molecular data can raise more risks to privacy than had been appreciated. For instance, it is possible to determine whether data from a person with a known genotype are in a gene-expression database (*4*) or in aggregated data sets of allele frequencies (*5*) or of phenotype regressions in genome-wide association studies (GWAS) (*6*). It has also been suggested that people may have a stable microbiome-variation profile (*7*), which, theoretically, could be matched to a sample from a known individual.

In this issue of *Science*, Gymrek *et al*. (*8*) describe how various public data sets developed for both research and nonresearch purposes can be analyzed to deduce the individual identity of some research participants by leveraging information pertaining to distant patrilineal relations. Specifically, Gymrek *et al*. attempted surname identification for 10 males in the Center for Study of Human Polymorphisms (CEPH) family collection whose genomes were sequenced as part of the 1000 Genomes Project (*9*). The researchers used sequence data for Y-chromosome short tandem repeats (STRs) and databases linking STRs to surname information, with the resulting information used to query public genealogy data and other available information [e.g., from the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository at the Coriell Institute, which distributes biological materials from the CEPH family collection, and obituary archives]. Using these steps, the authors were able to establish the identity of close to 50 of the CEPH participants (women as well as men). It is important to note that the authors do not reveal the names of the participants (or violate any known policies), but merely demonstrate their ability to identify them through the analysis of publicly available information.

Limitations and Broader Implications of the Study

The approach described by Gymrek *et al*. highlights vulnerabilities in efforts to protect the privacy of participants in genomics (and other 'omics) research. At this time, this methodology is particularly relevant to participants in the CEPH family collection because of the richness of publicly available research data and genealogic information derived from these individuals and their relatives (*10*, *11*). The CEPH participants whose samples were included in the HapMap Project (and then in the 1000 Genomes Project) underwent a process of re-consent to inform them about the plans for providing very broad and open access to the genomic data derived from their samples and for the in-depth genomic analyses that would be performed on those data. The inability to guarantee privacy and the possibility—then seen as remote—that individual identification might eventually become feasible were described explicitly. Despite this hypothetical and assumed low risk of identification, Gymrek *et al*. have now shown that it is possible to identify some participants of a genomics research study even in the absence of a second (matching) DNA sample.

Although additional studies are needed to assess more fully the generalizability of these findings to the broader population, this report—along with previous studies exposing other potential vulnerabilities in the current approach for protecting participant identity (*4*–*7*)—raises broader issues about how to protect participant privacy as more information becomes readily accessible to the public. These issues will become even more challenging in the future, as genomic technologies and information are used increasingly outside of research and health-care settings. It is thus prudent for the research community to reflect on the implications of these various studies (*4*, *6*) in considering how best to move forward.

The willingness of individuals and communities to assume some risk to participate in biomedical research depends on the scientific community's ability to maintain the public's trust. Indeed, it was this spirit that prompted the authors of the Gymrek *et al*. paper to contact staff at the National Institutes of Health (NIH) about their findings before publication. NIH staff, in turn, consulted with *Science* and the local institution for the CEPH study in

[1]National Human Genome Research Institute, National Institutes of Health (NIH), Bethesda, MD 20892, USA. [2]National Institute of General Medical Sciences, NIH, Bethesda, MD 20892, USA.

*Author for correspondence. E-mail: egreen@nhgri.nih.gov

Published by AAAS

Utah. In consultation with the authors, NIH staff acted swiftly to mitigate future risks by working with the NIGMS repository to shift age information, which had been available for some of the participants on the repository's public Web site, into controlled-access portions of the resource.

Shifting Concepts of Identifiability and Privacy

The recent set of papers exploring the potential for identifying individuals using genomic and other types of data, culminating with this latest report, calls into question whether the goal of complete deidentification of many types of human data is realistic in today's information-rich society. The ability to establish an acceptable threshold for "identifiability" has been debated vigorously since whole-genome analyses became feasible on a large scale with the introduction of GWAS (*12*–*14*). The approaches developed by Gymrek *et al*. and others call for reconsidering whether a simplistic distinction between identifiability and nonidentifiability remains adequate as a metric for describing expectations about participant protections. Some have suggested framing the risk of identifiability along a continuum (*13*, *15*) rather than as an absolute.

The general expectations of the public about privacy and confidentiality may be subtly shifting as well. In addition to social media outlets (e.g., Facebook) that have led to more pervasive sharing of personal details, patient-centric organizations (e.g., PatientsLikeMe) now provide the means to share in-depth information about health status and to identify research opportunities for motivated individuals (*16*). There are many perspectives about how to incorporate these potentially shifting norms into the systems for protecting research-participant interests in a manner that promotes maximum public benefit. Examples include an increasing number of "citizen science" initiatives [such as the Sage Bionetworks Commons (*17*) and Genomera (*18*)], which use informatics tools and social-media strategies to build research models for integrating participant preferences about privacy protection and future research use in an iterative and dynamic way. These initiatives can promote participants' long-term investment in and commitment to research, thereby gaining public trust through transparency and accountability (*19*). Although important questions remain regarding the scalability and feasibility of such approaches across populations and for various types of research projects, it is important for the research community to consider the options and potential advantages for the scientific enterprise as a whole.

From an oversight perspective, several bills proposing to convey various forms of property rights to genetic or genomic information have been introduced in U.S. state legislatures over the past few years (*20*). Such proposals should be considered in the face of an already inconsistent array of privacy protections at the state level that address access to or use of genetic/genomic information (*21*). The current patchwork of extant and potential legal standards for acquiring and managing such information adds more uncertainty to the considerations.

The Value of Broad Data Sharing

Gymrek *et al*. argue against placing barriers to accessing genomic and other high-density 'omic data derived from human participants. Substantial differences are seen in the use of data sets available through open-access versus controlled-access mechanisms. For example, the open-access HapMap and 1000 Genomes data sets of human genomic variation are used by many more researchers each year than related data sets in the controlled-access database of Genotypes and Phenotypes (dbGaP). In addition, recent NIH meetings on the scientific needs and opportunities for "big data" in stimulating basic and translational research recommended expanding open-access mechanisms for human research data with appropriate governance (*22*). Although the research community must be realistic and mindful of identifiability concerns, there are also ethical responsibilities to ensure that data contributed by participants for research are maximally utilized and that public research funding stimulates the greatest public good.

It is thus time for the research community to engage in a rigorous and open discussion about data identifiability and how to balance most effectively the benefits of broad data sharing and the imperative to respect and protect research participants. This dialogue should involve the full range of stakeholders, including participants, researchers, clinicians, database managers, advocacy groups, journal editors, and public representatives. Developing sustainable models that promote both continued willingness to participate in research and ongoing public trust will require a panoply of approaches.

Conclusion

We are at a crucial juncture brought about by the confluence of new technologies for data generation, bioinformatics, and information access on the one hand, which seem to create new risks to privacy, and the public's desire to benefit from these advances for a variety of personal and health reasons on the other hand. In light of this changing landscape, it is time to re-examine how to balance the protection of research participants (individuals, families, and groups) with the societal benefits likely to be gained through the enhanced research that broad data sharing facilitates. In doing so, we should consider whether there are alternative approaches that provide appropriate participant privacy and allow implementation across various research settings, including genomics and other emerging fields. The ultimate goal must be to develop a robust system that ensures the full societal benefits of biomedical research while respecting both individual needs and the communal good.

References and Notes

1. Batelle Technology Partnership Practice, *Economic Impact of the Human Genome Project* (2011); http://battelle.org/docs/default-document-library/economic_impact_of_the_human_genome_project.pdf.
2. E. Birney *et al*., *Nature* **461**, 168 (2009).
3. The Wellcome Trust, Sharing data from large-scale biological research projects: A system of tripartite responsibility, meeting, Fort Lauderdale, FL, 14 to 15 January 2003; www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf.
4. E. E. Schadt, S. Woo, K. Hao, *Nat. Genet.* **44**, 603 (2012).
5. N. Homer *et al*., *PLoS Genet.* **4**, e1000167 (2008).
6. H. K. Im, E. R. Gamazon, D. L. Nicolae, N. J. Cox, *Am. J. Hum. Genet.* **90**, 591 (2012).
7. S. Schloissnig *et al*., *Nature* **493**, 45 (2013).
8. M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, Y. Erlich, *Science* **339**, 321 (2013).
9. 1000 Genomes Project Consortium, *Nature* **491**, 56 (2012).
10. Coriell Institute for Medical Research, Centre de'Etude du Polymorphism Humain (CEPH) Resources; www.ccr.coriell.org/Sections/Collections/NIGMS/CEPHResources.aspx?PgId=525&coll=GM.
11. Sorenson Molecular Genealogy Foundation, Overview of the Sorenson Database; www.smgf.org/pages/sorenson-database.jspx.
12. L. Curren *et al*., *Eur. J. Health Law* **17**, 329 (2010).
13. W. W. Lowrance, F. S. Collins, *Science* **317**, 600 (2007).
14. A. L. McGuire, R. A. Gibbs, *Science* **312**, 370 (2006).
15. C. Heeney, N. Hawkins, J. de Vries, P. Boddington, J. Kaye, *Public Health Genomics* **14**, 17 (2011).
16. PatientsLikeMe, www.patientslikeme.com.
17. Sage Bionetworks Commons, www.sagebase.org/commons/.
18. Genomera, http://genomera.com.
19. J. Kaye *et al*., *Nat. Rev. Genet.* **13**, 371 (2012).
20. J. K. Wagner, D. Vorhaus, On genetic rights and states: A look at South Dakota and around the U.S. *Genomics Law Rep.* (2012); www.genomicslawreport.com/index.php/2012/03/20/on-genetic-rights-and-states-a-look-at-south-dakota-and-around-the-u-s/.
21. Presidential Commission for the Study of Bioethical Issues, *Privacy and Progress in Whole Genome Sequencing* (Commission, Washington, DC, 2012); http://bioethics. gov/cms/sites/default/files/PrivacyProgress508.pdf.
22. Establishing a central resource of data from genome sequencing projects, NIH workshop, Rockville, MD, 5 and 6 June 2012; www.genome.gov/27549169.

Acknowledgments: The authors thank J. McEwen and M. Guyer for consultation and feedback, and A. Bailey for editorial assistance.

10.1126/science.1234593