nature medicine

Privacy and protection in the genomic era

The establishment of an NIH working group managing access to HeLa cell genomic data highlights the limitations of assuring the privacy of participants in genomics studies. If, as this case illustrates, anonymity cannot be guaranteed, informed consent rules may need to be revised.

Two recent studies describing the genomic sequence of several HeLa cell lines prompted the US National Institutes of Health (NIH) to form the HeLa Genome Data Access working group in August. Sequence data from both publications have been placed under controlled access in the NIH-sponsored database of genotypes and phenotypes (dbGaP), and NIH-funded researchers are required to apply for access through the working group, which is composed of researchers and two members of the family of Henrietta Lacks, from whose tumor cells the original HeLa cell line was developed without consent. Once approved, researchers are granted access for one year and have to abide by a set of terms, including (but not limited to) uploading genomic data to dbGaP and not sharing the sequence data with unapproved users.

The NIH should be commended for taking prompt action to protect the privacy of the Lacks family, even though these measures may have been implemented too late. In an opinion piece published in *The New York Times* (23 March 2013), writer Rebecca Skloot quotes a member of the Lacks family as stating that the genomic data were private and consent should have been obtained prior to publication.

Francis S. Collins, director of the NIH, and Kathy L. Hudson, deputy director for science, outreach and policy at the NIH, have directly acknowledged a number of caveats with this arrangement (*Nature* **500**, 141–142, 2013). Given that approximately 1,300 gigabytes of partial genomic data on HeLa cells already exist in publicly accessible databases, the whole genome could theoretically be independently assembled. Moreover, data from one study (*G3* **3**, 1213–1224, 2013) were made publicly available in March and were downloaded by at least 15 people (according to Skloot) before being taken offline. In this context, restricting access to these whole-genome sequences may not be sufficient to protect the privacy of the family, as no efforts have been made to remove the previously published data from the online repositories.

Restricting access to the HeLa genome sequence may also have undesirable effects on biomedical research. The number of labs that can retrieve the genomic information might be limited; data in controlled access sites receive fewer visitors than those from the freely accessible HapMap and 1000 Genomes projects (*Science* **339**, 275–276, 2013). Publishers often request that genomic data be deposited in publicly accessible repositories to facilitate their independent confirmation, fostering reproducibility and confidence in the data. Restricting access may weaken some of these efforts.

Although NIH researchers are required to abide by these rules, the NIH cannot enforce them with non–NIH funded researchers. Labs can affordably sequence the genome of other HeLa-derived cell lines and may decide to openly share their data rather than post it to dbGAP, which could ultimately render this repository obsolete.

In terms of sharing data, how will researchers manage large-scale studies with numerous collaborators while abiding by the terms of access? Clarification is also needed regarding compliance, in terms of who will monitor whether new genomic data are uploaded to other sites, if and how breaches (by non–NIH funded researchers) will be penalized, and whether publishers or funding agencies will be required to suggest that researchers upload their data to dbGAP.

The notion of privacy and restricting access to certain data was recently tested following publication of a study this year in *Science* (**339**, 321–324, 2013). Multiple individuals in public sequencing projects were identified using freely available information from these projects, genetic genealogy databases and demographic data (including year of birth and/or residency information) that are not protected by the United States Health Insurance Portability and Accountability Act (HIPAA). In response to this study, the National Human Genome Research Institute at the NIH moved age information for certain participants (which was previously openly accessible) into controlled access.

Despite these efforts, the ease with which de-identified information can be unmasked calls into question broad informed consent clauses stating that participants will remain anonymous. Current technology is rendering this notion somewhat outdated, and approaches to obtaining informed consent should be modified to explicitly state that anonymity cannot be guaranteed. In the Personal Genome Project (*Proc. Natl. Acad. Sci. USA* **109**, 11920–11927, 2012), study participants sign an open-consent form, which states that data are deposited in an open-access database and may be re-identified and that participants can choose to withdraw from the study at any time. An additional (but not mutually exclusive) model, termed participant-centered initiatives, makes patients active participants in the study and adopts a more 'adaptive' informed consent model (*Nat. Rev. Genet.* **13**, 371–376, 2012). By leveraging information technology, researchers can notify patients of consent and study protocol changes, results related to their samples, and how to contribute additional samples for longitudinal studies, while providing them with the choice to opt out and have their data removed from the database at any time.

It is imperative, in light of these potential revisions and updates in informed consent and of the proliferation of publicly available data, that unauthorized disclosure of information be prevented and the use of these data be regulated. Although in the United States some legislation (including the Genetic Information Nondiscrimination Act (GINA)) is in place to prohibit insurance companies and employers from using genetic information on individuals, additional laws are needed to protect study participants from discriminatory practices. Such changes would help researchers gain the trust of study participants, encouraging enrollment and retention in studies and increase the pool of genomic data openly accessible to researchers, while ensuring that participants are aware of how their information is being used.