Sciencexpress Letter

Protecting Aggregate Genomic Data

Elias A. Zerhouni[1] and Elizabeth G. Nabel[2]

[1]Director, National Institutes of Health, Bethesda, MD 20892, USA. [2]Co-Chair, Senior Oversight Committee, NIH Policy for Sharing GWAS Data, and Director, National Heart, Lung, and Blood Institute, Bethesda, MD 20892, USA.

A paper published last week in *PLoS Genetics* (*1*) describes a statistical method for resolving individual genotypes within a mix of DNA samples or data sets containing aggregate single-nucleotide polymorphism data. This scientific advance may have important implications for forensics and for genome-wide association studies (GWAS). It has also changed our understanding of the risks of making aggregate genomic data publicly available. While we assess the broader scientific, ethical, and policy implications of this development, the NIH has moved swiftly to remove aggregate genomic data from our publicly available Web sites. Further information about changes in NIH open-access policies for GWAS is available on the NIH's GWAS Web site (*2*).

The paper by Homer *et al*., showed that a new statistical technique applied to aggregate data can determine whether a specific individual's genomic data are part of a given data set, including whether they are in the control group or the case (affected) group. It may also be possible to statistically infer whether a relative of the individual is a member of the case or control groups. The method requires having an individual's high density genotype data in hand from another source. Though the specific identity of the individual who was the source of the data could only be determined if that source were known through other means or reference data, this discovery nonetheless has implications for how these summary data should be protected. As a result, the NIH has removed from open-access databases the aggregate results (including *P* values and genotype counts) for all the GWAS that had been available on NIH sites (such as dbGaP and CGEMS). NIH intends to move the aggregate genotype data to the controlled-access database, where there is a firewall as well as protections and policies in place for appropriate data access, including review and approval of data access requests. The new finding does not have the same implications for data available through controlled access, and NIH access policies for individual-level genotype and phenotype data have not changed.

Sharing genomic data and, particularly, allele frequencies has become common practice, if not an imperative, in science. Yet, the protection of participant privacy and the confidentiality of their data are of paramount importance. These new statistical approaches have implications far beyond NIH data sharing policies, as aggregate GWAS data have been provided in publicly available form in many other ways, including other research databases and Web sites, journal articles and other publications, and scientific presentations. NIH urges the scientific community to consider carefully how these data are shared and take appropriate precautions to secure aggregate GWAS data in order to protect participant privacy and data confidentiality.

In short order and over the coming months, the NIH will work with our advisory groups and the wide range of stakeholders related to GWAS to further explore and address the policy implications of this finding. We call on our colleagues in the scientific community to join us in these important deliberations.

References

1. N. Homer *et al*., *PLoS Genet.* **4**, e1000167 (2008).
2. NIH Genome-wide Association Studies Web site: http://grants.nih.gov/grants/gwas/.

2 September 2008; accepted 3 September 2008
Published online 4 September 2008;
10.1126/science.1165490
Include this information when citing this paper.