NIH Background Fact Sheet on GWAS Policy Update

August 28, 2008

A research team, led by David W. Craig, Ph.D. at the Translational Genomics Research Institute (TGen) in Phoenix AZ, has developed a new bioinformatics method that allows the detection of a single person's SNP profile in a mixture of 1,000 or more individual DNA samples. In other words, bioinformatics techniques have progressed to the point that with enough genomic data on an individual from another source, it is now possible to determine whether that individual participated in a study by analyzing only the pooled summary data.

SNP stands for single nucleotide polymorphism, which is a change in a genetic letter in a specific location on a DNA molecule when compared to other DNA molecules. SNPs are used to study human genetic variation and are a powerful way to investigate genetic predispositions to health or disease. Large-scale genomic studies of human variation – called genome-wide association studies or GWAS – have recently provided important clues to the genetic roots of numerous common diseases. Because of the power of this technology, many institutes and centers at the National Institutes of Health support or are involved in such studies to understand the genetics of common maladies.

This new bioinformatics method is powerful, but it is still not simple to detect a specific individual's SNP profile in a pooled dataset. To find a specific profile within a set, the inquirer would first need to already have a highly-dense genomic profile (currently at least 10,000 SNPs) from an individual. Then this SNP profile would need to be statistically compared against the study dataset to measure how similar or different it is. Prior expectations were that individual profiles would have to be compared one to one to confirm a match; however, this new statistical analysis can now be used to detect a profile even in pooled data.

Although the technique has been demonstrated to work, the NIH is unaware that it has been used to compromise any information within NIH GWAS datasets. The technology to obtain the required genomic profile is not commonly used outside of the research community. And, even if an individual's SNP profile was found within a pooled dataset, all that would be learned is that this profile was contained in the dataset and, thus, it could then be associated with the characteristics of that dataset (e.g., disease or control population). The NIH GWAS databases do not contain the names or other identifiable information about individual study participants, so there is no risk to an individual participant's financial accounts or other personal information.

This discovery, however, has important policy implications for the way the scientific community shares such pooled sets of genetic data. For example, scientific journals have required researchers to make available aggregate data from GWAS studies when the results are published as a means to ensure the quality of the data. And, because use of these pooled datasets can speed up disease gene discovery, NIH – as well as

other research institutions and individual laboratories – developed public databases that allow researchers to freely download the datasets into their computers for analysis.

Because individual SNP profiles can now be detected within aggregate data, the NIH has moved quickly to assure continued protection of research participant privacy in genomics studies by controlling access to pooled datasets. For example, on Monday, Aug. 25, 2008, the NIH removed aggregate statistics files of individual GWAS studies from the public portion of the databases it manages, such as the Database of Genotypes and Phenotypes (dbGaP), operated by the National Center for Biotechnology Information, and the Cancer Genetic Markers of Susceptibility (CGEMS), operated by the National Cancer Institute. The data remains available for researcher use, but researchers must now apply for access to the data and agree to protect the confidentiality of the data in the same way that has been done all along for individual-level study data.

In addition, NIH is aware that others operating databases with similar types of datasets, including the Wellcome Trust Case Control Consortium in England and the Broad Institute of MIT and Harvard in Boston, have removed aggregate data from public availability.

NIH will continue to focus on this fast-moving field of research and on the development of policies to appropriately manage its databases and to promote policies that protect the confidentiality of all those who participate in NIH-sponsored research studies.