National Institute of Health

Modifications to GWAS Data Access

August 28, 2008

The National Institutes of Health (NIH) has modified part of our current policy for data posting and access to genomic data contained within NIH GWAS databases, including those hosted by the National Cancer Institute (NCI), such as the Cancer Genetic Markers of Susceptibility (CGEMS), and the database for Genotype and Phenotype (dbGaP). This fact sheet summarizes the modifications and the reasons for the changes.

The NIH developed a two-tiered access policy for GWAS data. The first level is the public posting (open access) of summary-level information and aggregate genotype data, including allele frequencies by case-control status, association tests odds ratios, and p values for each SNP in the scan. The second level is controlled access to individual-level data (genotypes and phenotypes). The controlled access data are available to investigators from scientific institutions who submit Data Access Request (DAR) packages that are reviewed and approved by the NIH Data Access Committees (DACs).

New statistical techniques for analyzing dense genomic information make it possible to infer the group assignment (i.e., case or control) of an individual DNA sample if one has access to high-density genomic data for that specific individual from another source and the allele frequencies for the case and control groups from publicly available aggregate datasets. The odds ratios can be used in a similar manner. We are also investigating whether these methods and data types could be used to statistically infer whether a blood relative of an individual is a member of the case or control groups of a study. Although it is currently unlikely that dense, individual-level genotype data will be obtained outside the research context, as these statistical techniques evolve and access to and use of genomic technologies expands, we need to minimize the possibility that aggregate data could be used to deduce whether individuals are present in any given dataset as it may infer personal information. Therefore, the NIH made modifications, described below, to the current policy for level-one (open) access to aggregate data.

To address any concerns that may arise related to the possibility of inferring group association from aggregate, publicly available GWAS data, NIH has taken the following preemptive actions:

- We have removed aggregate genotype data for GWAS studies from public access, but may make them available through the controlled access DAR/DAC process.
- We do not currently anticipate reposting the original format of allele frequencies by case/control status or the odds ratios for all the SNPs, but we will consider alternatives.
- We may soon be able to provide a categorical association test p value for each SNP, which could be useful for indicating the strength of association for scientists who do not need individual-level information.
- We are testing various means of data redaction to minimize the possibility that an individual's status as a case or control could be determined using publically available datasets. For example, we may ultimately be able to post more information than just the p value on a much reduced number of SNPs.

- We are adopting this modified data access policy for GWAS studies already posted in CGMES and dbGaP, those studies approved but not yet posted, and for all future GWAS studies.
- We are coordinating policy development across NIH.

We note that these new statistical approaches have implications beyond NIH policies, as aggregate GWAS data have been provided in many other ways in publicly available form, including other research databases and websites, journal articles and other publications, and scientific presentations. NIH will be working with the wide range of stakeholders related to genomic data sharing over the coming months to further explore and address the policy implications of this finding.