Workshop on Establishing a Central Resource of Data from Genome Sequencing Projects

Phenotype and Exposure Data

Leslie Lange, Lisa Brooks, Erin Ramos, Winifred Rossi

The utility of data from sequencing projects beyond their original purpose is limited by issues in comparing phenotype and exposure data across studies. Great effort has been made to build dbGaP as a repository for genetic, phenotypic, and exposure data that are widely accessible to the scientific community. One of the most important features of dbGaP is that it facilitates analysis of many samples, which has become increasingly critical for identifying factors that explain heritability beyond genetic variants already identified. Several steps could be taken to increase the utility of data from large-scale sequencing projects.

Harmonize data for existing studies: Harmonizing phenotype and exposure data that have already been collected would facilitate analysis across existing data sets. Retrospective harmonization of existing phenotype and exposure data is usually a multi-step, iterative process. Ideally, there is input from researchers with expertise in the individual phenotypes, as well as from representatives from each study contributing data. In consortium settings, a "point person" often first collects information from the sources above to define variable categories, such as "smoking behavior". There is then a very general search to identify all variables from each study related to each of the categories. The lists of variables that map to each category are often very large and some have limited descriptive information available. In order to narrow down the list, the point person then may go back and forth with phenotype experts and study representatives. Examination of data distributions and samples sizes is often required in this process. Documentation of efforts that have already occurred is not currently widely available to users of dbGaP. Development of an information resource would go a long way towards addressing many researchers' needs and could be facilitated by creating a position or a working group with the role of collecting and integrating various harmonization efforts for use in dbGaP. Harmonization efforts used in past and ongoing consortia such as ESP, the Gene Environment Association Studies (GENEVA)[1], and PhenX[2] could be leveraged to aid this effort.

Harmonize data for future studies: For future and ongoing studies, developing standard ways of collecting phenotype data would facilitate analyses across data sets. The first step to developing a panel of standardized variables for future studies would be to obtain input from researchers with expertise across a range of phenotypes and representatives from some of the larger existing studies, in order to balance the need for utilizing the most modern measurement tools with the ability to integrate data from new studies with data previously collected (sometimes many years previously). Once such a standardized panel of phenotypes and exposures is designed, NIH-sponsored funding opportunities could mandate that applicants agree to collect some minimal subset of phenotypes using this protocol.

Use a panel of standardized phenotype measures: Adding a set of standardized phenotypic measures across at least some of the existing large epidemiological studies and including them in future prospective studies would build a core set of comparable data on large numbers of people, making them even more valuable for a wider range of research questions by taking advantage of the existing genetic data and enriching them with additional phenotypic information. Some such standardized measures exist, (e.g., the PhenX Toolkit https://www.phenxtoolkit.org/), and others are being developed (e.g., an NIA effort to develop a standard core set of phenotypes to be collected across existing and new studies). These measures could be collected on existing study participants during an additional one-hour visit.

Obtain all study phenotype and exposure data for existing participants: Many studies in dbGaP have only a limited set of phenotypic and environmental exposure data; often, only the phenotypic data that are the main focus of a study are submitted. Increasing the amount of these data for the existing participants would increase the utility of the data sets. Additional data could be obtained through ancillary studies and study visits that occur after initial data submission with relatively minimal cost, as well as by collecting new data from participants. However, acquiring these data would require funding and coordination with the parent studies.

Collect new phenotype data for existing participants: Many important traits do not yet have sufficient samples sizes across studies in dbGaP. Rather than funding new studies to genotype and phenotype study participants, a more cost-effective solution would be to phenotype participants with existing genotype data. Many ancillary studies involve phenotyping new biomarkers for samples with existing genotype data. Acquiring the data this way takes time. While the potential value is clear, there are formidable hurdles to consider when designing new data collection efforts, including re-contacting participants and obtaining appropriate informed consent, the additional burden to participants, and cost.

Provide further information on all variables in each data set: Identifying the variables for phenotypes in databases is often cumbersome. There is considerable variability in the quality of phenotypic data documentation. Too often basic information, such as measure units and assay descriptions, are not provided. Sometimes different measuring units are used for the same variable in different participants. We strongly recommend that studies include data dictionaries that provide, minimally, a brief description of every variable, the measure units, the assay used, and the formula for each calculated variable. NIH should consider requiring standard units (e.g., metric system) for commonly measured phenotypes. Existing studies would need to update all documentation, a time-consuming and therefore costly process. In addition, NIH should continue to invest in user-friendly tools that help users identify studies that contain combinations of variables of interest (e.g., studies that have exome chip genotype data on African-Americans measured for diabetes). Such tools would be invaluable for users to achieve the full potential of a resource containing many studies.

Develop a limited standardized phenotype and exposure data set for wide use: Some studies include thousands of phenotype variables, many of which are study-derived and measure similar features (e.g., many different measures related to smoking behavior). Even with a data dictionary, it is difficult for users to identify which variables they should use. Developing a data set that includes a few widely used phenotypes across all cohorts

would require relatively limited resources and address many researchers' needs quickly.

References:

1. Bennett SN, Caporaso N, Fitzpatrick AL *et al*. Phenotype Harmonization and Cross-Study Collaboration in GWAS Consortia: The GENEVA Experience. *Genetic Epidemiology* 35:159–173, 2011.
2. Pan H, Tryka KA, Vreeman DJ *et al*. Using **PhenX** measures to identify opportunities for cross-study analysis. *Hum Mutat*. 33(5):849-57, 2012.