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

Background and Issues to Discuss at the Workshop

The genomes of tens of thousands of people have been sequenced, and in the near future many times that number will be added. The sequence, phenotype, and (where available) environmental data from these research participants have been accessible through each individual study, and most analyses have been done within each study. However, if the data could be combined in a broadly accessible central database, so that analyses could be done across many studies, the community would have more information to address many questions, including discovering common variants with small effect sizes and rare variants. An initial impetus for this effort was to look across large numbers of individuals for loss-of-function variants, especially beneficial ones, as target validation for drug development. However, many analyses would be facilitated by combined data sets.

This resource could support studies of disease biology:

- The phenotypes of people with loss-of-function variants or other interesting genotypes.
- The genotypes of people with extreme phenotypes.
- Large-scale GWAS and validation of previous results.
- GWAS on disease characteristics such as progression, response to drugs, and co-morbidities.
- Finding rare variants by sequencing many samples and obtaining more precise allele frequency estimates.
- Gene x gene and gene x environment interactions, including environments protective against genetic risk; detecting interactions typically requires larger sample sizes than detecting main effects.
- Shared genes/variants or environments across diseases.

How could the resource be used for drug target validation?

- Examining the phenotypes of people with variants in regions already associated with disease.
- Other ways?
- What data types would be needed?
- In what circumstances would it be important to be able to re-contact participants?
- How would study designs for drug target validation differ from ones for disease studies?

Generally, data sets are hard to access, they are in various databases, the variant and phenotype/environment data are not comparable across studies, and identifying and accessing all relevant data sets is even more difficult. How should consent processes, policy, ELSI, and database design be handled to allow the data from many studies to be combined for analysis with consent for broad access and use to address many questions?

The possible solutions may differ for data sets that already exist, where the participants have already provided consent and the sequence and phenotype/environment data have already been collected, and for new data sets where consent, sequencing, and collection of phenotype/environment data have not yet happened and standard processes could be developed. The goal is to maximize the value of the data sets that we already have, and to develop processes that will maximize the value of data sets to be collected in the future.

This workshop aims to discuss the scientific questions that these data could address, the problems with obtaining and analyzing multiple datasets, options for dealing with these problems, and the costs and tradeoffs of these options. This workshop should discuss the possible options; later focused workshops can address the details of implementing the recommendations. The major challenges to analyzing multiple data sets fall in four areas. Feasible solutions may involve several approaches.

1. **Data access.** Current NIH policy requires that researchers access any data set that contains both individual-level sequence and phenotype data through a data access committee (DAC). This process provides some protection to the participants, but is cumbersome, reduces the use of the data sets, and must be done one data set at a time. When the consent form has specified use restrictions, such as for particular diseases, more general studies cannot use those data. Several alternative approaches have been proposed to address the access issue; the appropriate ones for a data set may depend on conditions on use of the data:
   a. Provide open data access: participants consent to open release of sequence and phenotype/environment data and no restrictions on data use, and all users have open access to the data.
   b. Streamline the current process: modify the current process to make it less cumbersome to obtain access to data from one or particularly from multiple studies.
   c. Create a "research commons": approved researchers could study all the data sets in a commons. What criteria should be used to decide which data sets should be included in such a resource? Should data sets with disease restrictions be included, with researchers agreeing to abide by the restrictions, or should the resource include data only from participants who provided consent for broad use?
   d. Develop central analysis servers: groups could develop central servers that provide summaries and analysis results but not underlying data. This option could be used in addition to the other options.

   For each option, what policy changes would be needed, and what elements of consent would be needed or should be addressed (broad use, whether participants could be recontacted to obtain more data, whether individual results should be returned to participants), and under what conditions could already-existing data sets be included? A related policy issue is what summary data, such as allele frequencies, could be released publicly or without data use restrictions. Should any other data access options be considered?

2. **Processing sequence data.** Sequence data processing involves mapping the sequence reads and making variant calls. Calling SNPs is relatively robust for most of the genome, but calling indels and structural variants is still difficult. How much centralization of data processing would be needed to allow variants to be analyzed across studies? Options include:
   a. Simply keeping the variants called by each project.
   b. Each project re-calling or filtering variants in standard ways.
   c. A central group re-mapping each project's sequence read data and re-calling all variants with a standard pipeline.

3. **Phenotype and environment data.** How much harmonization of phenotype and environmental data would be needed (retrospectively and prospectively) so they could be analyzed across studies? Options include:
   a. Simply using the data provided by each study
   b. Performing some easy harmonization of variables across studies (retrospectively).
   c. Developing more comprehensive approaches for harmonizing data (retrospectively and prospectively).
   d. For new genomic projects, encouraging use of standard phenotype and environmental data when possible. When would it be useful to measure a standard set of phenotypes on many people?
   e. Obtaining additional phenotype data from participants. When would it be useful to re-phenotype people of interest based on their genetic data (such as for drug target validation)? What are the ELSI issues related to recontact?

4. **Database, tools, and analysis.** Depending on the design of this effort, the data may be compatible with NCBI/dbGaP and EBI. The database(s) would need to be able to accept and distribute data from many projects consistent with any use constraints, describe the types of data available for each project to allow researchers to find data types of interest easily, and allow queries across the data for multiple projects. What tools will be needed for analyzing the data? Some standard analyses (variant calling and imputation, loss-of-function variants, disease associations, gene x gene and gene x environment interactions) could be provided centrally, and some deeper data analyses could be done separately. Should the central database support analysis tools, and should it provide the computing resources needed to analyze these large data sets? What could other groups provide?

The discussion of these topics should include these issues:

1. What are the pros and cons of the approaches, and the tradeoffs among approaches?
2. What are the costs and cost tradeoffs of the approaches?
3. What differences are there among new data collections (new consents, new sequence and phenotype/environment data), partially new data collections (new consents for existing data, or recontacting people for deeper phenotyping), or

already existing data collections (current consents and sequence and phenotype/environment data)?

4. What policy changes would be needed?
5. What are the priorities of various options, and the short-term actions, longer-term actions, and actions that will require a lot of work but should be done?
6. What staging of the effort could be done? Could phenotyping be separated from sequencing? If the sequence data, already-available phenotype/environment data, and pre-computed analyses were accessible in one place, then deeper analyses and further phenotyping (broadly or based on particular genotypes) could be done separately.
7. What next steps are needed? (Having follow-up meetings, setting up working groups, supporting particular areas of research, modifying policy, other (?)).