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

Streamlining Current Access Policies and Processes

Laura Lyman Rodriguez, Lisa Brooks, Adam Felsenfeld, Nicole Lockhart

The access policies and processes for controlled access databases were established to ensure that researchers have access to genomic data for research use in a manner that promotes respect for research participants and establishes consistent processes to protect participant interests. While controlled access models (e.g., dbGaP and EBI) have been successful, as the data from increasing numbers of studies are deposited there, the processes to deposit and access data are becoming increasingly cumbersome. Data resources utilizing a controlled access model, and researchers depositing data to them, have invested substantial staff time and resources to build infrastructure to manage the data submission and access processes. As use of these resources grows, existing approaches to implement the data sharing policy are not likely to scale to the degree needed to facilitate analyses across the number of data sets necessary to achieve many scientific aims.

The areas identified below represent "pressure points" within current controlled access systems where process or policy modifications might result in increased efficiencies within the existing framework. Enhancements in these areas are not suggested as comprehensive solutions, but rather as short-term mechanisms to mediate challenges or bottlenecks in data submission or data access, or as potential means to achieve a more appropriate balance between necessary participant protections and broad accessibility of human genomic data for research. Implementation of some of these recommendations may require policy modifications.

1. **Consolidate Data Access Committees (DACs) to increase efficiency and consistency in the review of data access requests:** This change would reduce the number of DACs (potentially to one) that need to review an access request, especially when many data sets are requested.
2. **Develop a consistent lexicon for data use limitations to define consent groups:** A standard vocabulary of data use limitations in data repositories could increase the uniformity of language used across consent groups, which could promote consistent interpretation of appropriate research uses by submitting institutions, requesting investigators, and DACs. Improved consistency would diminish the substantial time spent by all parties to understand and communicate about the specific parameters of each consent group.
3. **Establish categories of common types of data users and research uses that are appropriate:** If a common understanding of appropriate research uses for different types of data use limitations (based on consistent consent groups) were developed, ideally across data repositories, DAC review of certain requests could be expedited. This would enable DACs to focus their time on access requests that raise complex participant protection questions.

4. **Promote the use of consents that enable broad data sharing without data use limitations:** Broad language and discussions in consent processes such as "all biomedical research" or "any research purpose" would increase data accessibility. Limitations on research uses, such as disease-specific conditions or restrictions on the sector of the research community able to access the data (e.g., academic, non-profit, corporate, government, etc.) diminish the potential for public benefit to be realized through the data. Additionally, more complex and time-consuming processes for data submission and data access must be implemented. Restrictions on fields of use (e.g., insurance research, or population relatedness) can similarly limit the utility of the data and extend review times. To implement these goals, sufficient resources to recruit members of population groups that are less accepting of broad data sharing or use will need to be provided.
5. **Simplify data submission:** Providing improved tools, automated data submission methods, and funding for the data repositories and the institutions submitting data would lower the barriers to data deposition into central repositories.
6. **Enhance the ability of users to find data in the data repository:** Repositories should be organized so that users can quickly find all studies related to particular diseases or containing particular phenotypes. It should be possible to sort data sets by consent group, so that data appropriate for research uses of interest can be identified quickly, enabling investigators to tailor their access requests.
7. **Modify policies and procedures to enable efficient data exchange in research consortia:** The expectation for investigators to obtain separate data access approvals by standard processes to participate in the analysis of common data sets creates substantial burdens for consortium participants and often causes substantial delays in producing analyses. Systems for group pre-certification, for example, may expedite processing of requests without diminishing participant protection or institutional obligations for investigator conduct.
8. **Allow data uses of general value:** Repository policies should identify certain types of studies likely to advance the study of all data sets through innovative methodologies or tools available for application to many data sets. If clear and consistent standards for designating studies anticipated to benefit genomic analyses broadly were established (e.g., software improvements for calling variants, or meta-analyses applicable to all diseases), then researchers would know what types of studies would be in this category and DACs would have a basis to approve researchers doing such studies for access to data with disease-specific data use limitations. It is reasonable to expect that the output of such studies would advance analyses in disease-specific areas, though the initial secondary use may not directly pertain to any particular condition.

Data that can be released publicly under the current policies:

- List of studies in dbGaP: including the sample sizes, ethnicities, and data variables.
- Variants: can be sent to dbSNP and dbVar (but not their frequencies). This includes variants likely to affect gene expression (e.g., nulls).
- Disease-associated variants: lists of these variants, and their statistical significance.

Data that would be valuable to release publicly, which would require changes in policy:

- Genotype numbers: these are used to calculate allele frequencies. The release should be by population/ethnicity for control groups, case groups, and general population samples, except for case samples in potentially stigmatizing studies (e.g., drug use, HIV).
- Haplotypes: how extensive should haplotype data be allowed to be released, i.e., haplotypes of certain sizes, or for some extent around genomic elements?