Report on the

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

June 5 – 6, 2012

Background and Goals of the Workshop

The use of genome sequencing in human research is growing rapidly, creating unprecedented opportunities for advancing scientific knowledge. The genomes and exomes of tens of thousands of people have been sequenced, and hundreds of thousands more are expected in the next few years, many with phenotype and exposure data available. These genetic, environmental, and clinical data represent the product of major investments by the NIH and other agencies, with the expectation that the data will yield new insights to guide the understanding, prevention, diagnosis, and treatment of human disease.

Realizing the potential benefit of this research effort requires overcoming technical, organizational, and ethical barriers, many of them related to the need to aggregate data from multiple studies. Data sets are often difficult to find, large, cumbersome to access, and may be stored in different databases. Computing on large data sets can be expensive and demanding; portals and additional tools are needed for analysis. Software and analytical methods are rapidly evolving and non-standardized, and needed computational infrastructure is expensive to create and maintain; progress is slowed because tools and data are not brought together efficiently or widely. The underlying sequence, phenotype, and exposure data often have not been harmonized, making it difficult to carry out meaningful comparisons and aggregation across data sets. Current ethics protocols were not designed to facilitate large-scale data-sharing repositories and cross-study analyses.

Large sample sizes are needed for reproducible genetic findings; combining data from different studies increases statistical power. Some diseases and genotypes are so rare that only by combining data across individual studies can findings be obtained. Genes and pathways influence multiple diseases (e.g., protein misfolding for Alzheimer's and Parkinson's diseases, inflammation for cancer and heart disease, and autoimmune attack for type 1 diabetes and rheumatoid arthritis); looking across seemingly unrelated diseases can illuminate these shared pathways. The common history of humans means that many variants and haplotypes are shared within and between populations, and yet subtle population differences can confound analysis; by analyzing disease studies together with common reference data, power can be maximized and confounders avoided. Specifically, combining data increases sensitivity and accuracy to detect rare variants, to impute variants missing from sequence or GWAS data, and to infer haplotypes, thus providing substantially greater power to detect associations between genetic variants and diseases.

Overcoming these barriers is necessary to honor the social contract among research participants, geneticists, epidemiologists, translational investigators (including pharmaceutical companies), and funders. Specifically, the public participates in genetic research out of the desire to contribute to new knowledge and to help others facing health challenges; participants expect that their data will be used to advance biomedical research while respecting and honoring their autonomy and privacy. Geneticists and epidemiologists want to make discoveries to expand knowledge and improve health care; access to data and analytical methods is their lifeblood. Pathophysiology of human disease is the foundation of the diagnostics and pharmaceutical industries; genetics and epidemiology are building

blocks to understand disease processes. Funders want to ensure that they get the greatest return on investment in the form of scientific understanding of disease.

Enhanced methods for data stewardship and governance of data repositories are needed. As data sharing increases in scope, research participants are no longer asked to consent to a single study, but rather to make their data available to a large number of researchers, likely from many different countries. Public trust in the procedures used to store and access data is essential. Transparency is needed about methods to ensure data security, policies used to approve researcher access to data, auditing methods to ensure compliance with policies, and consequences that result from any missteps or compliance failures. Effective methods to ensure public input on each of these critical areas of governance are needed. As a component of transparency, communication is needed to inform the public about the scientific value of broad data sharing, without generating false expectations about the speed with which translation to health benefits will occur.

This workshop aimed to (a) review the scientific questions that sequence, phenotype, and exposure data can address; (b) outline challenges to obtaining and analyzing data sets from multiple independent studies; and (c) consider options for dealing with these problems, including the costs and tradeoffs of multiple options. Individuals from a wide range of groups participated, including genetics and disease researchers, ethicists, government funders, biopharmaceutical leaders, and journal editors.

The workshop was characterized by an unusual level of agreement among participants that enabling data access as broadly as possible will expedite understanding of human health and disease, and should be pursued actively. Participants viewed positively solutions that would work for most but not all existing data, acknowledging that while there may be sound reasons to exempt some data from broad sharing, this should not stop progress from being made on other fronts. The group thought that multiple solutions should be pursued, rather than forcing "one size fits all" solutions. In particular, different solutions may be needed for existing data as compared to data not yet collected; the accelerating nature of sequence data generation means that solutions that can apply only to new studies but not to existing studies will still be extremely valuable.

Major Recommendations (timeframe for implementation; groups to be consulted)

1. Given the added value of data that can be deposited in central databases, the default expectation should be that **sequence/phenotype/exposure data sets are deposited in one or several central databases**. The greatest utility will be obtained if both germline (inherited) and somatic (cancer) data are aggregated in the same or interoperable databases. Exceptions to central deposition may be appropriate for data sets where additional participant protection is needed. (Short- to long-term; discussions with researchers and funders.)

2. Given the added value of broad use, funders should support and encourage the development and use in future sequencing studies of **consent procedures that seek permission for broad data use.** Exceptions may be appropriate for data sets where additional participant protection is needed; these data sets with use conditions are also valuable and should be included in central databases, with the use conditions flagged. (Short- to long-term; discussions with researchers and funders.)

3. New governance procedures for the central databases should be created, including methods and policies that support responsible access to individual data sets for different users and different uses. Governance should include public disclosure of policies and procedures; public

participation in policy-making, e.g., public representation on an oversight committee and opportunities for public review and input on policies under consideration; and clear systems for accountability, to include specific penalties for data misuse or failure to comply with policies of the central resources. New policies should include procedures for approval to access data; justifications for exempting data from submission to a central resource or broad data sharing; and auditing methods to monitor researcher compliance with data-use policies. (Short- to long-term; discussions with researchers and funders.)

4. Given that the results of analyses are highly valuable, and their risk to participants at this point theoretical and limited, **summary statistics (variant names, genotype counts, allele frequencies, effect size estimates and standard errors, p-values) from GWAS and sequence-based association studies should be released publicly**, except for studies of particular sensitivity. Release of summary statistics should be the default position. (Short- to medium-term; NIH, Wellcome, other; policy discussions and researcher input.)

5. **The process for accessing data from dbGaP should be streamlined as soon as feasible, with enhanced efforts to inform the public about dbGaP and seek public input regarding dbGaP procedures.** (Short-term; steps are already being taken, and NIH policy discussions are occurring; public consultation.)

6. To streamline data access while ensuring accountability, **a system should be developed so that users of non-public data are registered as a prerequisite for access, and access can be provided in one step in a granular manner to multiple available data sets**, such as those with broad consent and those with consents consistent with a given user's research goals. Users who violate the terms of data use should be denied future access to public data. (Medium-term; NIH policy discussions and consultation with the public and research communities.)

7. **Multiple approaches to aggregating analyses of the sequence data sets should be supported, including those referred to below as "research commons", "data analysis servers", and "disease-specific portals"**. (Short- to medium-term; funding agencies could support.)

8. Because of the considerable added value to uniform data processing of large sequence data sets, **central processing of sequence data from sequence reads to variant calls should be encouraged**. (Short- to medium-term; funding agencies could support.)

9. **Harmonization of phenotype and exposure data retrospectively should be encouraged and all such harmonization efforts should be captured in the central databases**. **Future projects should include phenotype and exposure harmonization;** these should be described in the funding application and resource sharing plans. (Short- to medium-term; funding agencies could support.)

Summary of the Discussions

In this section, we summarize the topics discussed during the workshop, providing the bases for the recommendations listed above. We consider central databases, data set use conditions, data access models and computing on aggregated data sets, DNA sequence data processing, phenotype harmonization, recontact of research participants, public summaries of data sets, governance, and public engagement.

Central databases

Given the value of aggregating data from many studies, almost all data sets with underlying individual-level sequence/genotype data and phenotype/exposure data from NIH studies should be deposited in a central database. Similarly, data from the UK and other countries carrying out large-scale genetic studies should be in central databases. There should be international coordination of policy and requirements and a global inventory of studies.

Segregating data by disease or institute should be avoided, although disease-specific portals could be useful to disseminate information to particular research communities. Exceptions may be appropriate for data sets where additional participant protection is needed; policies defining appropriate justifications for exception are needed.

Data set use conditions

Although the workshop attendees recommended that research participants provide consent for broad data sharing in future studies, existing data sets may be based on consents that have conditions on who may use the data (e.g., exclusion for commercial use) and for what the data may be used (e.g., general research use, or specific disease research only). Data sets should be clearly marked with any such use conditions. More discussion may be needed to develop a consensus on what already-existing consent language supports deposition of data, such as consent that is "consistent with" or "not inconsistent with" data deposition. dbGaP is working to harmonize consent group categories (such as for general research use), and these categories should be specified for existing data sets and future data sets; this approach should be used for other databases as well.

Future informed consent documents should address the issue of commercial use of data. Given the increasing frequency of pre-competitive public-private collaboration, many for-profit entities may request access to data with no intent to commercialize the information. Thus informed consent documents should include information on use of data by commercial entities for public release of knowledge and for development of useful drugs, tests, and therapies.

Some extensions to current use conditions should be considered. For example, it seems reasonable for investigators focusing on methods development (e.g., for association analysis) or resource creation (e.g., for haplotype estimation) that will be useful for studies of any disease to be allowed access to disease-specific data, even though they do not focus on a particular disease. Before creating such a policy, views of interested stakeholders should be sought, including researchers involved in depositing and using data; research funders; IRBs responsible for setting data use conditions; research participants; and the general public.

Data access models and computing on aggregated data sets

Innovation and a diversity of approaches to data access and for computing on aggregated data sets should be supported. These would be based on databases with the individual-level sequence, phenotype, and exposure data, and would generally include data from multiple sources. Any entity aggregating data sets should have appropriate governance structures to ensure stakeholder input, transparency, and accountability. Workshop participants endorsed five models for data access and computing on aggregated data, with the preferred model depending on both the scientific questions to be addressed and requirements for data security and confidentiality. Each approach should include governance mechanisms that incorporate transparency, public input, and accountability. When

appropriate, there should be standard procedures for data use and user criteria, to promote consistency and efficiency in dealing with requests.

Open data access: Participants provide appropriate consent so that data can be freely available to all on the internet, e.g., the Personal Genome Project and the 1000 Genomes Project. This option allows the broadest use of data, but will be subject to possible biases due to potential participants who choose not to participate without greater control over privacy and data use.

Streamlined data access: Intended to streamline the procedures for obtaining access to data in dbGaP and similar data repositories. Data would be made available to qualified users after review by an appropriate committee.

Registered user system: Users of non-public data could be registered, to provide accountability and the possibility of recourse for violating any data use conditions. Such users would, after registering, be provided access to all available data sets for which they are eligible. In the US, an NIH "super-DAC" could provide this registration. Ideally, there would be reciprocity of registration between countries. User status would be a matter of public record and an audit system to ensure appropriate data use would be in place. Data could then be used by download and analysis, through a data analysis server or other approved means. A researcher's standing as a registered user should be conditional on required data submissions occurring in a defined timeline for studies they lead.

Research commons: Registered users would be able to examine and carry out analyses on all appropriately-consented data sets in a central database, including **having access to the individual-level data.** More work is needed to determine how to do analyses on large data sets centrally.

Data analysis servers: Data analysis servers would have access to all relevant data sets, provide standard methods for data processing and analysis, and provide computer interfaces for users to make queries and obtain results. **Users would not obtain access to the underlying individual-level data.** A special-purpose server could provide specific analyses, such as certain types of sequence annotation, imputed haplotypes, or admixture analysis based on all available samples.

As a specific example, workshop participants strongly supported NCBI's current efforts to aggregate general research use data sets to allow unified access and queries of data. This effort is useful in itself and can be treated as a pilot of broader efforts in which multiple groups would innovate on ways to manage, process, and analyze the data and to serve the results.

DNA sequence data processing

Processing of DNA sequence data from many studies jointly rather than one study at a time leads to more complete and accurate variant identification and more confident genotype calls. Data processing pipelines remain highly complex and rapidly evolving, such that subtle differences in methods result in substantial incompatibilities in output data. For these reasons, there is value in having sequence datasets uniformly processed with one or more methods. This could be done by having sequencing centers provide BAM files to one or more central data processing services that would uniformly process BAM files to variant calls and track the data freezes and the analysis methods and parameters used to produce them.

Phenotype harmonization

For data already collected, key phenotype measures can be harmonized across studies and this should

be encouraged where possible, recognizing that this requires substantial time and effort. Standards for this harmonization process should be established. Each study should be responsible for harmonizing its data, and all such harmonized information should be captured in the central database.

For future projects, including those based on existing cohort studies, plans for phenotype and exposure harmonization should be included in grant applications and resource sharing plans, and then carried out. Standard methods for collecting phenotype and exposure data that allow harmonization, such as PhenX, should be developed and used.

Recontact of research participants

An important possible use of aggregated data sets is to identify individuals with particularly informative genotypes or phenotypes, for example, individuals with null alleles or extreme phenotypes. Recontacting these individuals for additional sequencing, genotyping, or phenotyping has the potential to be a particularly efficient way to learn about genotype-phenotype relationships, and could be useful for drug target validation and for understanding disease biology. Longitudinal studies with regular recontact are already set up for this. Recontact could be done for data sets already collected if the existing consent forms and policies allow it. When feasible, future studies should ensure that participants are asked for permission for recontact, to facilitate such studies, although several ELSI issues relating to recontact would need to be addressed.

Public summaries of data sets

Where possible, comprehensive data summaries should be released publicly via the internet. For example, for genotype-array or sequence-based association studies, these should include lists of variants, genotype counts, allele frequencies, effect size estimates (e.g., odds ratios or regression coefficients) and their standard errors, p-values, and the calculation methods used. These summary data, especially when aggregated across relevant studies, would provide valuable information for studies of the genetic contribution to disease or of population structure.

For example, the HapMap and 1000 Genomes data sets currently are accessed by hundreds of thousands of users, despite limited sample sizes, to obtain allele frequency and haplotype information for multiple populations. Releasing information on tens of thousands of individuals in GWAS and sequencing studies would provide much more accurate data on allele frequencies and haplotype structure, especially for less common and rare variants.

For some studies these summary data are not appropriate for public release; examples might include studies involving small populations where individuals are at higher risk of being identifiable, populations distrustful of genomic research, or samples with stigmatizing health conditions; further policy-making is needed to define appropriate justifications for exemption. However, for most studies the risk of individual identification and harm from such release is small while the value to research on human health and disease is high.

Portals with aggregated data sets could provide answers to common questions:

- For a disease, what genes are associated with it (above some threshold)?
- For a gene, what diseases/traits are associated with it (above some threshold)?
- For a variant, what is the association to all available traits (quantitative and directional)?
- For a genome sequence, what variants are seen, with frequencies and associations?

Governance

Several stewardship issues need to be addressed as part of a system that uses broad consent: transparency (public disclosure of methods and policies governing data protection and access and accessible reports on data uses), accountability (appropriate consequences for data misuse or other compliance failures), and public input (appropriate public representation in policy-making and oversight). IRBs would need to consider wide sharing of genomic, clinical, and exposure data. For cohort studies that require project-specific approvals (e.g., Framingham), it may be useful to discuss the value of general research use with study participants; this would provide an opportunity for broader community engagement about the consent process and research more generally.

Public engagement

Researchers have a social contract with the public, who pay for public research, and with study participants, who give their samples and data to address questions with the potential to advance science and improve the public's health. Attention needs to be given to informing the public about the research efficiencies gained by data aggregation and to seeking public input into the policies and procedures to be used. Effective dialogue may occur in many venues, for example, with participants and potential participants at health care and research institutions, via outreach activities to community organizations, and through engagement with advocacy groups and other organizations representing various community interests. Given the low participation in genomic research among minority populations, particular efforts to establish dialogue with underrepresented communities are needed.

There is a need for further exploration and articulation of the benefits of genomic research, including a realistic assessment of its potential impact on health care and the likely timeline. There is also a need for attention to what counts as a benefit, and from whose perspective. The dialogue should acknowledge that collection of genomic data provides important research opportunities but that privacy risks cannot ever be removed fully; people are likely to be more willing to accept small privacy risks if they perceive their participation as leading to important societal benefits. They are also more likely to trust the measures in place to protect privacy if governance and oversight procedures are transparent and include input from people or organizations they view as trustworthy.

Conclusion

The workshop brought together a diverse group of stakeholders to discuss strategies to maximize the use of sequence and phenotype/exposure data as a public good. The participants agreed on several specific recommendations on various data access and analysis models. They emphasized that a single model will not cover all aspects of responsible sharing of human sequence and phenotype data, but that multiple models will support obtaining the most value from these data sets for promoting human health.