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

Central Analysis Servers

Mark DePristo, David Altshuler, Lisa Brooks, Carlos Bustamante, Adam Felsenfeld, Pearl O'Rourke, Laura Rodriguez, Stephen Sherry

Executive summary

A central analysis server would make it possible for researchers to answer scientific questions about the relationships between inherited DNA variation and human phenotypes by (a) aggregating in a single location available data on human DNA sequence and phenotype, (b) providing a state-of-the-art computational environment and analysis tools to manage, process, and analyze the data for phenotypic association, and (c) managing security, data use, and user access to ensure that each dataset is used only in manners allowed by the original informed consent and data use agreements.

The central analysis server would aggregate and manage controlled-access data sets and provide a set of core "Apps" (*a la* the iPhone) for data processing and analysis. The server would not provide access to underlying individual-level data, but rather would provide tools that analyze the data and report results. By jointly processing all of the sequence data in standard ways, the server could provide uniformly high quality and directly comparable data on genetic variants, improving power across studies by improving variant calling, and (as allowed) providing access to comparator samples.

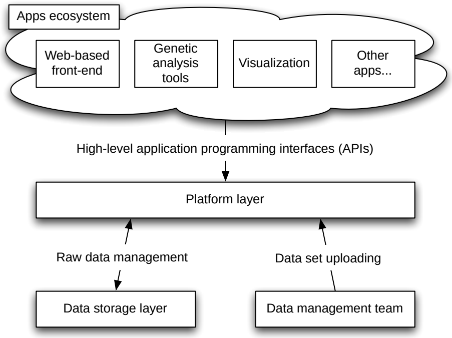
The server would store phenotype data for genotype/phenotype studies but would not include HIPAA identifying information. The server would not redistribute individual-level data. Thus, the Analysis Server would augment but not replace nor duplicate databases such as dbGaP that provide access to individual-level data. However, by having merged data with tools for analysis, and by providing results in a "redacted" form not requiring IRB or DAC approval to view, many more users could obtain analysis results more easily than they can with the current dbGaP DAC access approval system.

To achieve this vision, the server would provide controls to ensure that data are used only in a manner consistent with data use conditions. We envision that users would register (to indicate, for example, whether they work for a company), and the system would track the allowed uses of each dataset. For example, if a dataset could be used to study T2D but not psychiatric disease, then the analysis Apps would allow a user to perform analysis of T2D, but not allow the dataset to be merged as controls in an analysis of schizophrenia. In this way the server would enforce data use conditions for each sample, and would track the changes in these conditions as raw data percolate into high-level summaries that may permit more open sharing, such as allele frequencies and annotation.

This approach benefits from a strong network effect – access to cutting-edge Apps for data processing and analysis, as well as easy access to available comparator datasets, would drive adoption, and thus draw additional App developers and submission of data. This approach would enhance the value of large collections of shareable data, such as for the common NIMH controls and the 1000 Genomes Project, by streamlining access to data on these resources.

To achieve this vision **requires developing a <u>platform infrastructure that manages security and user access</u>, seeding it with** public data available through dbGaP and similar repositories, deploying a number of "Apps" for performing data processing, integration, and analysis, and performing and publishing paradigmatic analyses that demonstrate the value of the approach.

The Platform



We believe that the most critical step is to create a software ***Platform*** (ideally, more than one) designed with enterprise level security to manage data, control access, and interact with Apps. The system would annotate each data element for classes of users who can access them and allowed uses.

Specifically, each data set would be annotated (based on the use conditions) with regard to which phenotypes could be studied, which analyses could be performed, and which answers could be reported. Some datasets could be used only in a limited way (say, for the study of diabetes or of cancer). Most datasets would allow reporting of derived results (such as the frequencies of variants) as long as they are in a form that is not identifying. It should be possible to codify these permissions and to provide software that ensures (to the greatest extent possible) that Apps are unable to access data for uses that are not allowed or to report information that might violate confidentiality or use conditions. This process of matching the analysis to the conditions must be managed in an automated, rather than in a human-intensive, manner.

The system would support chaining of multiple tools to create analytical pipelines. Initial "Core Apps" would provide a baseline of needed functionality such as alignment of sequence to the genome, calibration of error models, variant calling, functional annotation, association analysis, and cancer genome analysis. The API could include a "developer toolkit" so others could develop tools compatible with the Platform's architecture, security, and user management, and could support an "App store" where applications could be registered, run on the aggregated data, and downloaded.

The technologies to generate, process, and analyze NGS data are rapidly advancing. The Platform would be regularly upgraded with new data, Apps for data processing, and analyses. Methods would need to be developed to enable existing data to be integrated with new data, such that all samples could be compared in analysis without unrecognized technical bias, and to perform ongoing quality assurance and quality control on the data and resulting outputs.

Critically, the Platform would make it possible for users to run analysis tools without accessing the underlying individual-level data. Study results would be served (in redacted or limited form, if needed to ensure privacy) on a website available to the entire scientific community. Funding agency policies would determine which types of information could be reported, and the data uses allowed for each dataset. The Platform could also support an interface such as BioMart from Ensembl for data processing and queries. A query for this system involves choosing datasets to query, attributes, and filters to restrict the query. One can imagine pull down "radio buttons" that implement common queries such as estimating allele frequencies, calculating odds ratio and population heterogeneity statistics, or running new analysis apps. The website could be queried by phenotype ("show me all genes associated with T2D at P<0.00001"), by gene ("show me all phenotypes associated with PPARG at P<0.0001"), and by variant ("show me all the phenotypic associations, regardless of P value, for rs1234567"). A user could ask the system to "process raw sequence data from studies 12 and 14 using 1000 Genomes Project pipeline #3, filter out any variants seen in the 30,000 samples that allow sharing of variant frequency information, and then perform a burden test for association of private loss of function mutations using Analysis Pipeline 72, outputting a list of genes in which private loss of function mutations are more common in controls that in cases, suggesting a protective effect." When evaluating this request, the Platform would check each dataset for annotation allowing its inclusion in the proposed analysis. Standard questions (such as the associations between genetic variation and specific diseases, or somatic mutations in cancer types) could be pre-computed on a regular basis, such that answers could be made instantly available without custom compute jobs.

We propose that multiple such Servers be developed, as competition and diversity are virtues. Some Servers might be comprehensive; others might specialize. Different Apps built on these servers might provide particular capabilities; for example, a server may provide haplotypes and methods to impute variants in submitted GWAS datasets, or may provide ancestry deconvolution for admixed samples in data sets submitted by researchers. Such a system could lower the barriers to new analysis tools being rapidly applied to large datasets, and could provide researchers with the benefits of combining data from many samples without having to obtain access to each of the data sets individually.

ELSI and policy considerations

The central analysis server would obtain datasets from dbGaP and similar databases, which do not contain HIPAA identifying information. The server would support various levels of data sharing. At one end of the spectrum are 1000 Genomes data, which are publicly available, and the server would allow anyone to operate on them in any way given the server's apps. At the other end of the spectrum are data that permit only a single use (Schizophrenia by non-commercial users). Here investigators studying Schizophrenia could still use the server for data storage and analysis infrastructure, and benefit from shared controls, but no studies of other phenotypes could benefit from these data. Of course, most data live in the middle, with many analyses that could be performed and shared.

Thus, it is key that the platform can support and enforce data use conditions, which will vary at an atomic level among the samples in the system.

The central server would likely be considered a research protocol and would need to be reviewed and approved by an IRB. The IRB protocol should include details on the "business rules" such as:

a. Requirements for submission of data to the server
    i. All data will have been generated from tissue obtained under an IRB-approved informed consent form (ICF) process. The consent forms should say that coded data will be placed in a central repository for future research by other researchers.
    ii. Server data managers would confirm standards before accepting a dataset, such as that any HIPAA identifiers were removed, and would obtain the data use conditions from the DAC that approves use of the dataset or from the originating institution.

b. Requirements on use of data
    i. High-interest analyses will be pre-computed, with results made available to all users
    ii. Requests for custom analyses will be submitted in a uniform manner that allows automated systems to confirm that requests abide by data use conditions.