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

Rationale for Aggregating Data Sets

David Altshuler

One of the central goals in biomedicine is to define the relationships between biological pathways and the etiology of human disease, as this provides the foundation of knowledge for the rational design of diagnostics, prevention, and therapy. The study of variation in DNA sequence — whether inherited in the germline or somatically mutated in cancer — can provide an unbiased approach to forge links between genes, diseases, and other clinical traits and to characterize the phenotypic consequence in humans of perturbation in gene function. Such information can provide critical information to guide therapeutic development, where the fate of major investments in clinical trials hinges on predictions about the consequences (both positive and negative) of modulating target function in patients.

Specifically, the human genetics community should provide clear answers to the following questions:

- Given a **disease** of interest, what is the subset of human genes that harbor DNA variation that is robustly and reproducibly associated with risk of disease?
- Given a **gene** of interest, what phenotypes (if any) are associated with DNA variation?
- Given a **genome sequence** of interest, what are the population frequencies and known disease associations of variants (common or rare) observed in an individual?

The challenge

Technically and scientifically, the answers to such questions are in reach; unfortunately, there are many organizational and structural barriers to overcome. Specifically, data about each disease is collected and analyzed in silos, one phenotype and one sample set at a time. Informed consent and data use restrictions limit the free sharing of data on the internet, and yet in many cases would support much broader use than is currently achieved. The current approach makes it impossible to leverage the tremendous sample size generated by the national and international investment in genotyping and sequencing, and the decades of clinical research and phenotyping on which these genetic studies are built. As attention turns to rare variants, sample size and sharing of information become even more critical, as any given study will have limited ability to distinguish a rare mutation from a low frequency polymorphism, and any given study might have only one or a few observations of a rare variant and phenotype. The current model makes it difficult to answer the simple question of whether a gene variant identified as associated with a given disease is also associated with other disease phenotypes. The current model necessitates that when a new statistical method is developed, it can be applied only to data that are locally available. These organizational factors limit statistical power to make new discoveries and the scope of the answers that can be obtained.

A distinct but equally important challenge is the gulf between the disciplinary background and means of communication of the genomics community and those of the much larger community of basic scientists, physicians, and scientists in industry. Specifically, genomic scientists work with very large datasets, and often develop answers that are ***statistical*** in nature. The analysis of genome sequence data now requires large investments in computational hardware and software, and sophisticated expertise in the analysis of "big data." In contrast, biology has historically been an ***experimental*** science, and few biologists have the computational capability to analyze large-scale genome sequence data even if they obtained access to the raw data. Moreover, the languages of statistical and experimental scientists have evolved separately. For these reasons, the international investment in genotype/phenotype studies, which has largely been reported in statistical terms, is opaque to many life scientists and physicians. This, too, must change if we are to bridge from genomic discovery to biological insight and medical benefit.

Until recently, there existed **fundamental** barriers to such integrated analysis of genotype and phenotype: the incomplete nature and inadequate scale of genotype data, the storage of data at many sites without clear paths to access, unmeasured technical confounding due to different genomic platforms, the lack of comparability of phenotype data in different studies, and the lack of computational methods instantiated in software required to perform (let alone automate) analysis. The language and visualization tools required to report such results were in their infancy.

Four recent developments make it possible to consider bringing together large-scale data on DNA sequence and phenotype in an environment with robust analytical tools: (1) next-generation sequence data intrinsically encode genome location (based on sequence alignment) and data quality, enabling integration of distinct datasets and empirical modeling of errors; (2) NIH and other funders have supported the generation of data at the breadth and scale required, with over 50,000 genome sequences funded by NIH through 2012, and many more to follow; (3) NIH has mandated for a number of years that all human (including cancer) sequence data be deposited in a public repository, dbGaP, aggregating and providing a route to access data; and (4) computational methods have been developed that can integrate data collected at different sites and with different platforms, perform data processing and association analysis in an increasingly automated manner, and report results in a manner that is clearer to biologists and doctors.

At the same time, the potentially identifying nature of DNA sequence data, and the sensitive nature of the genetic and phenotypic data, place great responsibilities on investigators and institutions to ensure compliance with informed consent and security of human subjects data. It is critical that any environment created to support such data analysis be secured with enterprise-level software systems, and intrinsically support and enforce the commitment that data is used only for the purposes for which informed consent was given.

The vision

These recent developments provide a potentially ***transformative opportunity*** to develop a much richer set of answers about the relationship of genetic variation to disease based on aggregating existing and rapidly emerging data, by developing a software platform that provides enterprise-level management of security and user access, that instantiates centralized oversight over data use for consistency with informed consent, and melds these to a powerful environment for deploying analytical methods without exposing potentially identifying information about individual-level genotypes.

In the long term, the desired outcome is a vibrant and sustainable ecosystem in which many investigators provide data (based on submission to public archives, for their use and for use by others), many developers contribute methods, many scientists make discoveries, and an even broader community is able to access the answers obtained in easily understood form.

The genomes of tens of thousands of research subjects have been sequenced, and in the coming years that number will likely grow to hundreds of thousands. Each of these individuals will have been richly characterized based on DNA sequence, many phenotypes, and (in some cases) environmental exposures. If this wealth of data could be combined and integrated, such that analyses could be done across many studies, the scientific community could answer many more questions, and systematically characterize the phenotypic consequences across many traits of common variants and of rare mutations.

An initial impetus for this effort was to enable the identification across many individuals of rare loss-of-function variants, especially ones that are protective against disease, as an approach to target validation for drug development. However, a vast array of analyses would be facilitated by combining data sets and increasing power and comprehensiveness of analyses:

- Phenotypes associated with any loss-of-function variants observed in every human genome.
- Genotypes of people with extreme (and often rare) phenotypes.
- Characterization of the full phenotypic consequence of variants found in GWAS.
- Accurate estimates of allele frequencies of rare variants across multiple studies.
- Power for gene x gene and gene x environment interactions.