On the Future of Genomic Data
Scott D. Kahn
Science **331**, 728 (2011);
DOI: 10.1126/science.1197891

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

PERSPECTIVE

On the Future of Genomic Data

Scott D. Kahn

Many of the challenges in genomics derive from the informatics needed to store and analyze the raw sequencing data that is available from highly multiplexed sequencing technologies. Because single week-long sequencing runs today can produce as much data as did entire genome centers a few years ago, the need to process terabytes of information has become de rigueur for many labs engaged in genomic research. The availability of deep (and large) genomic data sets raises concerns over information access, data security, and subject/patient privacy that must be addressed for the field to continue its rapid advances.

The study of genomics increasingly is becoming a field that is dominated by the growth in the size of data and the responses by the broader scientific community to effectively use and manage the resulting derived information. Genomes can range anywhere from 4000 bases to 670 Gb (*1*); organisms that reproduce sexually have two or more copies of the genome (ploidy). Humans have two copies of their inherited genome of 3.2 Gb each. Full sequence data has been archived for many thousands of species (*2*), and more than 3000 humans have been sequenced to some substantial extent and reported in the scientific literature; new sequencing is expanding at an exponential pace.

Output from next-generation sequencing (NGS) has grown from 10 Mb per day to 40 Gb per day on a single sequencer, and there are now 10 to 20 major sequencing labs worldwide that have each deployed more than 10 sequencers (*3*). Such a growth in raw output has outstripped the Moore's Law advances in information technology and storage capacity, in which a standard analysis requires 1 to 2 days on a compute cluster and several weeks on a typical workstation. It is driving a discussion about the value and definition of "raw data" in genomics, the mechanisms for sharing data, the provenance of the tools that effectively define the derived information, and the nature of community data repositories in the years ahead (Fig. 1). A second challenge is analyzing all these data effectively. The pace of innovation in genomic data creation is much higher than the rate of innovation within genomic informatics; this widening gap must be addressed before the overall field of genomics can take the leap forward that the community has foreseen and is needed for many applications, spanning from evolution to medicine.

Central to the challenge is the definition of raw data. Many current sequencing technologies capture image data for each base being sequenced; these must be parsed into a set of intensities for each of the bases that are subsequently interpreted as a specific base call and an assigned quality value (the likelihood that the base call is correct). The quality values currently represent more storage space than the base. The size of these images for many labs is currently greater than 5 TB of information per day if they are stored; the impracticality of using and archiving image data has motivated the development of real-time processing of the images directly to output only the base calls and the quality values (*4*). The ability to process the images in near real time has allowed the speed of sequencing to advance independently from the speed of disk storage devices, which would have otherwise been rate limiting. Although there are computational challenges with such near real-time analysis, this processing affords a two-orders-of-magnitude reduction in data needing to be stored, archived, and processed further. Thus, raw data has been redefined to be bases and qualities, although the data formats here are still a source of ongoing development. Newer and under-development "third-generation" sequencing methods also output bases and base qualities (*5*). Improvements in determining the base from the raw reads, and thus the quality, are ongoing, even though the downstream analysis tools often do not lever the increased precision of the estimated values. Looking ahead, the ways in which base quality scores are captured, compressed, and archived will optimize storage and improve analysis.

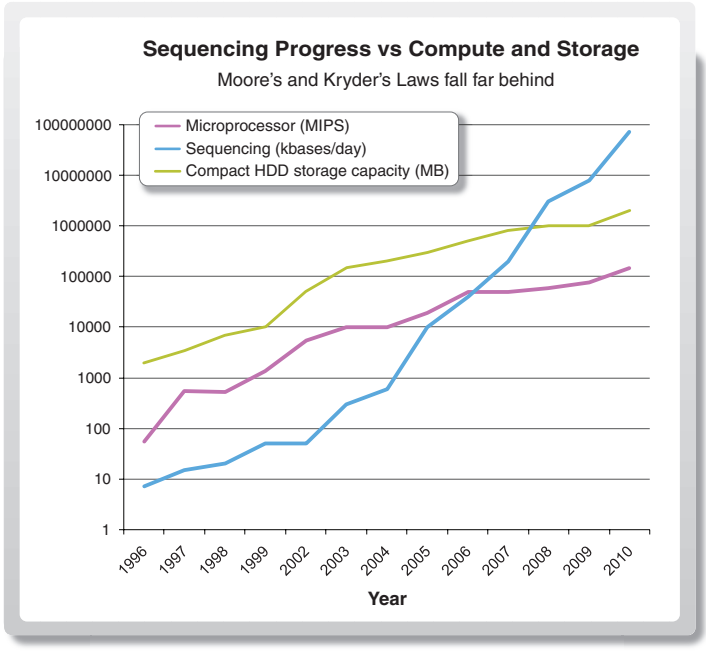

Fig. 1. A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

The size of the collective data emerges as a major concern as one moves downstream of data creation on the sequencer to the analyses and comparisons that constitute the transition into biologically and/or medically relevant information. For example, the current size of the 1000 Genomes Project (www.1000genomes.org) pilot data, representing a comparative analysis of the full genomes of 629 people, is roughly 7.3 TB of sequence data. The ability to access this data remotely is limited by network and storage capabilities; the download time for a well-connected site within North America will range between 7 and >20 days. Having this data reside within a storage cloud does not entirely mitigate the longer-term challenges, in which aggregation of multiple data stores (data stored within different clouds) will be required to perform the comparative analyses and searching of information that are envisaged. This fundamental inability to move sequence data in its current form argues for considerable changes in format and approach. Without a solution, these downstream informatics challenges will gate advancements of the entire field; a substantial leap in informatics capability and concomitant changes in the definition of data must take place to support movement of the field forward. Centralization of data on the Cloud is a positive start.

One approach that is being explored is to move computation to the data rather than moving the data to the computation. This model is made possible through so-called service-oriented archi-

Illumina, 9885 Towne Centre Drive, San Diego, CA 92121, USA.

tectures (SOAs), which encapsulate computation into transportable compute objects that can be run on computers that store targeted data. SOA compute objects function like applications that are temporarily installed on a remote computer, perform an operation, and then are uninstalled. This solution poses a challenge around how compute costs are shared when the computers performing the work are maintained by the data owners rather than the researchers performing the analysis. Collaborative compute grids may offer a solution here (*6*).

There are additional concerns for human data that include data security, protection of subject/patient privacy, and the use of the information that is consistent with informed consent. Although HIPAA provides guidelines to de-identify the data, researchers have shown that the genomic data are inherently identifiable (*7*) and that additional safeguards are required. This concern resulted in the NIH temporarily removing all access to several genomic databases until the risks to privacy could be evaluated and processes put in place to minimize these risks (*8*). The passage of the Genetic Information Non-discrimination Act (GINA) acknowledges some of these fundamental challenges with genomic information and attempts to provide additional regulation so as to discourage inappropriate use of such data.

One proposed solution to minimize data storage is to use reference genomes so that ultimately all that needs to be stored in a new analysis are the differences from the reference. Rather than storing every base being sequenced, only base mutations that are distinct from the reference need to be saved; typically, these differences represent just 0.1% of the data. This large reduction in data size offers a solution to the dilemma around publication of results, even though it departs from the standard of submission of discrete sequence reads. However, with analysis methods still under active development it may be premature for the transition to referential formats. Referential formats can also pose problems with capture of data quality throughout a genome. Knowledge of data quality is most needed when evaluating derived information (such as genomic regions of putative function) in order to provide a contextual basis for the certainty of the assignment (or assignments). Once the physical challenges in storage and access of genomic data are solved, the issues involving the quality and provenance of the derived information will persist. This is particularly an issue for published works and aggregated databases of derived information, if the semantic of the information in the source data changes over time. There may be no automatable mechanism to revise conclusions or redact records.

Although there is a widespread focus on human DNA sequencing and its application to improving clinical understanding and outcome, genomic data can be even more complex. A further problem is that much of the sequencing data being collected is dynamic and is and will be collected at many times, across many tissues, and/or at several collection locations, where standards in quality and data vary or evolve (over the lifetime of each datum). Much sequence data, both affecting humans and not, is not of human origin (for example, of viruses, bacteria, and more). The challenges with analysis and comparison across organisms are exacerbated by these issues. Fields such as metagenomics are actively engaged in scoping the data, and metadata requirements of the problems are being studied, but standards have not yet been agreed upon. The informatics demands of epigenetics data will be more burdensome because of the dynamic nature of gene regulation. Whereas there are ideas being formulated to compress (human) DNA data through the use of the human reference genome as noted above, no such reference exists within the metagenomic and epigenomic fields.

The centrality of reference data and standards to the advancement of genomics belies the limited research investments currently being made in this area. Large intersite consortia have begun to develop standard references and protocols, although a broader call to action is required for the field to achieve its goals (for example, the development of standardized and approved clinical grade mutation look-up tables). This is an activity that would benefit from input from the broader informatics community; several such interdisciplinary workshops and conferences have been organized, and these are having modest success in capturing a shared focus to address the challenges presented. One exemplar is the current state of electronic medical records (EMRs) and their inability to capture genomic data in a meaningful manner despite the widespread efforts to apply sequencing information in order to guide clinical diagnoses and treatment (*9*–*14*). These efforts require large cross-functional teams that lack the informatics tools to capture the analysis and diagnostic process (or processes) and thus have limited means to build a shared knowledge base. Discussions around personalized medicine rarely focus on the data and information challenges, even though these challenges are substantial technically, institutionally, and culturally. Although it is early still for the impact that NGS will have on the practice of medicine, taking action to define and implement a comprehensive, interoperable, and practical informatics strategy seems particularly well timed.

The future of genomic data is rich with promise and challenge. Taking control of the size of data is an ongoing but tractable undertaking. The issues surrounding data publication will persist as long as sequence read data are needed to reproduce and improve basic analyses. Future advances with use of referential compression (*16*) will improve data issues, although most of the analysis methods in use will need to be substantially refactored to support the new format. More difficult will be the challenges that emerge with practical curation of the wealth of information derived from genomic data in the years ahead. The nature of derived information used for clinical applications also raises issues around positive and negative controls and what must be stored as part of the medical record. Similarly, the evolution of informatics frameworks (such as EMRs) and scalable informatics implementations (such as SOA) to handle genomic data will probably be a hard requirement for advancing the biological and medical sciences made possible by the advances in sequencing technologies.

References and Notes

1. http://en.wikipedia.org/wiki/Genome
2. www.ncbi.nlm.nih.gov/genbank/
3. The output of the first 454 (*15*) and the current HiSEq. 2000 output, assuming 300 Gbase over a 7- to 8-day run.
4. J. Karow, *GenomeWeb*, 28 July 2009.
5. J. Clarke *et al.*, *Nat. Nanotechnol.* **4**, 265 (2009).
6. https://portal.teragrid.org/
7. N. Homer *et al.*, *PLoS Genet.* **4**, e1000167 (2008).
8. P. Aldhous, *PLoS Genet.* (2008).
9. E. A. Worthey, *et al.*, *Genet. Med.*, PMID: 21173700 (2010).
10. A. N. Mayer *et al.*, *Genet. Med.*, PMID: 21169843 (2010).
11. J. E. Morgan *et al.*, *Hum. Mutat.* **31**, 484 (2010).
12. T. Tucker, M. Marra, J. M. Friedman, *Am. J. Hum. Genet.* **85**, 142 (2009).
13. V. Vasta, S. B. Ng, E. H. Turner, J. Shendure, S. H. Hahn, *Genome Med.* **1**, 100 (2009).
14. M. Choi *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19096 (2009).
15. M. Margulies *et al.*, *Nature* **437**, 326 (2005).
16. M. Hsi-Yang Fritz, R. Leinonen, G. Cochrane, E. Birney, *Genome Res.*, 10.1101/gr.114819.110 (2011).

10.1126/science.1197891