How to get genomes at one ten-thousandth the cost

Jeffery A Schloss

The NHGRI's Advanced DNA Sequencing Technology program is spearheading the development of platforms that will bring routine whole-genome sequencing closer to reality.

In 2001, when it was clear that initial sequencing of the human genome would soon be completed, the National Human Genome Research Institute (NHGRI) of the US National Institutes of Health (Bethesda, MD, USA) undertook a formal planning process for the next phase of genomics research[1,2]. Two high-level planning meetings flanked a series of ten topical workshops that, through the involvement of more than 600 scientists and other participants, identified several significant emerging themes and opportunities for genomics research for the next one to two decades. The outcome of this process was a bold plan[3] that included a list of 'technological leaps' that seemed so far off at the time as to be almost fictional. Among these was the reduction of the cost of DNA sequencing by four to five orders of magnitude to enable sequencing of an individual human genome for $1,000 or less. On the fourth anniversary of the first grant awards focused on that goal, it is timely to take stock of the considerable progress that has been achieved, the challenges that remain and the opportunities that have been (and would be) enabled by vastly cheaper sequencing technology.

National Human Genome Research Institute

NHGRI Advanced DNA Sequencing Technology Development Grantee Meeting held in San Diego in March 2007. NHGRI hosts annual meetings for investigators to encourage information exchange, partnerships and brain storming to solve critical challenges for next-generation sequencing platforms.

Aims of the program

Beginning in 2004, the NHGRI's DNA sequencing technology goals were addressed in a series of grant solicitations[4] that comprised a two-phased approach. At that time, the state of the art was Sanger sequencing with capillary array electrophoresis (CAE). As implemented at large-scale sequencing centers supported by the NHGRI[5] or elsewhere, Sanger CAE was capable of producing a high-quality draft of a mammalian genome sequence for about $10 million. The solicitations described the NHGRI's challenge to the scientific community to develop technologies that would reduce the cost by approximately four orders of magnitude—to about $1,000—in ten years, with an intermediate five-year goal of reducing the cost by two orders of magnitude, or about $100,000. Such a trajectory represented a twofold acceleration in the rate at which sequencing costs had been reduced since the beginning of the Human Genome Project, during which it took about ten years to reduce costs 100-fold, from $10.00 to $0.10 per finished base pair[6]. The solicitations also addressed a key issue in determining the true meaning of sequencing by cost by linking it to a measurable standard of sequence quality. The 'gold standard' for genomic sequencing was 99.99% accuracy with essentially no gaps[7], but experience showed that this was achievable only with a considerable amount of manual intervention. So, although the ultimate goal remained the production of extremely accurate sequence, the program goal was framed as the automatic generation of sequence data at least as good as the high-quality draft sequence of the mouse genome at that time, specifically with respect to coverage, per-base quality and contig length[8].

Jeffery A. Schloss is Program Director, Technology Development Coordination, Division of Extramural Research, National Human Genome Research Institute, US National Institutes of Health, Bethesda, Maryland 20892, USA.
e-mail: jeff_schloss@nih.gov

The research community has responded enthusiastically to the initial and subsequent calls for proposals; more than 160 grant applications have been submitted to date. In a total of five rounds of research grant awards, the NHGRI has committed more than $100 million to 50 research teams. The funded projects have ranged from testing the feasibility of new ideas to developing systems, and they have combined fundamental research with engineering. The proportion of awards was initially greater for those that addressed the nearer-term goal.

With the enormous progress that has been made toward the $100,000-genome goal, the balance in NHGRI funding has, more recently, shifted toward technologies for the $1,000 genome. Although some technologies were funded at the outset for a 10-year development period to achieve a $1,000 genome, it is possible that, with innovation, improvement and optimization, some $100,000-genome technologies may ultimately achieve costs substantially lower than their initial goal. The research has produced a rich collection of published results and intellectual property, as is evident from the program's bibliography[9].

Approaches

Several approaches to achieving the technology goals are being pursued. Public abstracts of the $100,000- and $1,000-genome technology grants are available from the NHGRI (http://www.genome.gov/10000368#6). Sequencing by synthesis and sequencing by ligation predominate among the $100,000-genome grants, and nanopore approaches predominate among the $1,000-genome projects. These are the subject of articles in this issue (p. 1135; p. 1146; p. 1117) and have been reviewed elsewhere[10–12].

Other funded approaches include miniaturization and integration of Sanger sequencing sample preparation, capillary array separation and detection (for example, ref. 13); real-time monitoring of the DNA polymerase–directed nucleotide incorporation reaction by observing fluorescence at the polymerase site or in the reagent flow downstream of the enzyme (for example, refs. 14,15 and http://visigenbio.com/technology.html); monitoring of sequential interactions of the protein synthetic machinery as it traverses an mRNA molecule, to read out nucleic acid sequence; sequencing by hybridization, in which libraries of very short oligonucleotides are sequentially annealed to DNA templates[16], their individual locations recorded using fluorescence or other detection means, and the positional information used in combination with known oligo sequences to support the determination of the template sequence; extraction of sequence information by measuring changes in forces on DNA molecules as they increase in length by sequential addition of nucleotides, as oligonucleotide probes bind or are released as a function of degree of match or mismatch, or as DNA is pulled through the limiting circumference of a chemical ring[17]; and finally, use of changes in stretching force to sequentially release DNA molecules from a surface depending on their length.

As a consequence of support from NHGRI and additional support from other sources, several of these projects show great promise. Recent grant solicitations[18] have explicitly encouraged new approaches, such as the development of additional physical, chemical, biochemical, spectroscopic or microscopic measurements that can sequentially read out nucleotide identity along a DNA strand.

Community insights

An integral feature of this research grant program has been the annual investigators' meetings. Designed to encourage information sharing, and engendering a judicious mix of collaboration and competition, these meetings have focused on discussions of progress and shared understanding of the state of the art and of real or apparent roadblocks and possible solutions. Investigators have discussed the availability of resources and expertise. Several partnerships have been established, for example, between investigators who have particular technical expertise, have developed a particular chemistry or have worked on a crucial component with others for whom that might solve a technical hurdle.

> Wider dissemination of discussion of the challenges resulting from such discussion will be essential for galvanizing progress.

At this year's meeting, attendees identified several key barriers to accelerated development of vastly less expensive technologies. For example, for sequencing by synthesis, crucial control of surface chemistry is required to ensure efficient and stable binding of templates while minimizing nonspecific binding, particularly of fluorescent species. Fluorescent labeling of nucleotides must be accomplished with attention to fluorescence yield and photostability of the dyes, high synthetic yield and methods for purification of the labeled nucleotides, and polymerase compatibility, among many other concerns. Optical detection requires high resolution, speed, sensitivity with low noise, color discrimination and width of field. Ensemble methods require maintenance of synchrony and signal strength among large numbers of templates.

Although each of today's next-generation sequencing systems has been optimized with varying degrees of success around the current state of the art for each of these challenges, fundamental innovations to meet these or other underlying technology challenges could result in significant performance improvements, including longer read length, lower miscall rate and faster data collection. For nanopore sequencing, challenges include reliable and reproducible pore fabrication, sensor and electronics integration, control of the motion and orientation of DNA as it traverses the sensor, and the key challenge of concepts and methods to distinguish between the bases.

Among the grantees are individuals and groups who understand subsets of these challenges that are most relevant to their own approach. A clear articulation of those challenges in discussions at these meetings should help the grantees, individually and as a group, to refine their strategies and make more rapid progress. Equally important is the potential to increase the opportunity for other scientists and engineers, who do not currently study sequencing technology but have relevant knowledge or ideas, to contribute their expertise to overcoming those challenges. For this reason, the wider dissemination of discussion of the challenges resulting from such discussion will be essential for galvanizing progress (for example, see p. 1146).

Looking forward

Since the program began in 2004, several new sequencing platforms have been commercialized and put into use in laboratories worldwide. NHGRI provided grant support toward the development of the systems from Roche/454 (Basel; http://www.454.com/enabling-technology/index.asp)[19] and Applied Biosystems (Foster City, CA, USA; http://marketing.appliedbiosystems.com/mk/get/SOLID_KNOWLEDGE_LANDING). Those systems and the Illumina/Solexa system (San Diego; http://www.illumina.com/pages.ilmn?ID = 250)[20] have been implemented in NHGRI-supported and many other laboratories and sequencing centers, and feedback from these expert users has stimulated numerous system improvements. NHGRI also provided grant support for the more recently introduced Helicos system (Cambridge, MA; http://www.helicosbio.com), for which there is less experience among research laboratories.

Although the throughput and quality of data produced by current 'next-generation' technologies are substantial, they come with tremendous bioinformatics challenges. One is simply the large amount of data that must be collected, stored and analyzed. Another more complex issue is sequence assembly. By the time the original Human Genome Project sequencing goals were completed, the Sanger CAE technology was able to produce individual reads of 700–900 base pairs in length, with accompanying error metrics on each base. Bioinformatics tools for assembling and analyzing those reads are now rather well established[21]. For the new technologies, however, the nature and format of the read data differ substantially from those of Sanger CAE data, and the read length is significantly shorter[22]. Widely accepted quality metrics, although being developed, are not yet available, and the significantly shorter reads require that many more reads and much higher depth of coverage are used to assemble a sequence of acceptable quality. Some of the sequencing laboratories have found that combining data from more than one of the new technologies yields the highest sequence quality per cost; even so, additional bioinformatics work is needed to develop and optimize these approaches using multiple data types. The difficulties in handling next-generation data will be reduced as recently introduced paired-end read protocols become routinely established, more effective algorithms and methods (for example, refs. 23,24) are developed and implemented, and the characteristics of the underlying data improve through ongoing protocol and technology improvements by the vendors, users and research community that lead to increased read length, read mapping and per-base read quality. These improvements will result from a complex interplay of refinements in nucleotide chemistry, polymerase activity, fluorescence and its detection, surface chemistries and image analysis, and base-calling software (see p. 1125).

The next generation of sequencing technologies is here today and is being used to produce large amounts of biomedically relevant data[22]. What does the future hold? Technologies even beyond those of the next generation are being actively developed. A vital need for these even newer technologies will be longer read length. Although there will be further improvements to today's next-generation platforms, investigation of technologies—such as real-time read-out from active DNA polymerase molecules[14,15] and nanopore sequencing (p. 1146)—is already beginning to offer evidence of the potential to read thousands of bases and to reread the same template for the purpose of increasing data quality[25–27]. There remains considerable uncertainty as to whether and when these technologies will become available as practical tools, although Pacific Biosciences (Menlo Park, CA, USA) has projected that its zero-mode waveguide confinement system for detection of single DNA molecules[28] will be available in 2010. Thus, both the challenges and the promise are substantial.

The ultimate goal of the NHGRI and its sequencing technology development program is of course not simply to produce better technology, but to improve human health through technology.

The ultimate goal of the NHGRI and its sequencing technology development program is of course not simply to produce better technology, but to improve human health through technology. The program is based on the conviction that acquiring vastly greater amounts of highly accurate DNA sequence information at very low cost will become a crucial step on the path toward improved health. The Human Genome Project has laid the groundwork and offered the first returns on the promise of genomic medicine. Since the initial human genome sequence was published, an explosion of information, enabled in large part by recent improvements in sequencing technologies that were stimulated by the NHGRI technology development program, are revealing new insights into human sequence variation and its relations to disease, the functional elements of the human genome, and the composition of microbial communities in the environment and in our bodies, to list a few examples[22]. These and many other areas of research that hold great promise will build on the increasing availability of sequence information. Beyond research, a major challenge will be personalized medicine, in which genomes of very large numbers of individuals must be routinely sequenced at a data quality substantially higher than that offered by current technology, and sequence information must be interpreted in a way that is relevant to an individual's health and wellness. The prospect of sequencing technologies that would enable individuals to be diagnosed and treated in large part on the basis of detailed knowledge of their own genetic inheritance is momentous. And for the technology and biomedical communities, the enabling $1,000 genome presents the challenge of a lifetime.

ACKNOWLEDGMENTS

I thank NHGRI colleagues, advisors and grantees for insights that have contributed so substantially to the development of this program, and M. Guyer for valuable comments on the manuscript.

1. http://www.genome.gov/10005717
2. http://www.genome.gov/12010624
3. Collins, F.S. *et al. Nature* **422**, 835–847 (2003).
4. http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html
5. http://www.genome.gov/11508922
6. Service, R.F. *Science* **311,** 1544–1546 (2006).
7. http://www.genome.gov/10000923
8. Mouse Genome Sequencing Consortium. *Nature* **420**, 520–562 (2002).
9. http://www.genome.gov/Pages/Research/DER/GTP/GTPPubs&Patents.pdf
10. Shendure, J. *et al. Nat. Rev. Genet.* **5**, 335–344 (2004).
11. Bayley, H. *Curr. Opin. Chem. Biol.* **10**, 628–637 (2006).
12. Zwolak, M. & Di Ventra, M. *Rev. Mod. Phys.* **80**, 141–165 (2008).
13. Blazej, R. *et al. Anal. Chem.* **79**, 4499–4506 (2007).
14. Korlach, J. *et al. Proc. Natl. Acad. Sci. USA* **105**, 1176–1181 (2008).
15. Bashford, G. *et al. Opt. Express* **16**, 3445–3455 (2008).
16. Lizardi, P. *Nat. Biotechnol.* **26**, 649–650 (2008).
17. Ashcroft, B.A. *et al. Small* **4**, 1468–1475 (2008).
18. http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-08-008.html
19. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
20. Bentley, D.R. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006).
21. International Human Genome Sequencing Consortium. *Nature* **431**, 931–945 (2004).
22. Mardis, E. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
23. Brockman, W. *et al. Genome Res.* **18**, 763–770 (2008).
24. Zerbino, D.R. & Birney, E. *Genome Res.* **18**, 821–829 (2008).
25. Korlach, J. *et al. Nucleosides Nucleotides Nucleic Acids* **27**, 1072–1083 (2008).
26. Williams, J.G.K. *et al. Nucleic Acids Res.* published online, 22 August 2008 (doi:10.1093/nar/gkn531)..
27. Gershow, M. & Golovchenko, J.A. *Nat. Nanotechnol.* **2**, 775–779 (2008).
28. Foquet, M. *et al. J. Appl. Phys.* **103,** 034301–1–034301–9 (2008).