Next-generation sequencing transforms today's biology

Stephan C Schuster

A new generation of non-Sanger-based sequencing technologies has delivered on its promise of sequencing DNA at unprecedented speed, thereby enabling impressive scientific achievements and novel biological applications. However, before stepping into the limelight, next-generation sequencing had to overcome the inertia of a field that relied on Sanger-sequencing for 30 years.

In 1977 Fred Sanger and Alan R. Coulson published two methodological papers on the rapid determination of DNA sequence[1,2], which would go on to transform biology as a whole by providing a tool for deciphering complete genes and later entire genomes. The method dramatically improved earlier DNA sequencing techniques developed by Maxam and Gilbert[3] published in the same year, and Sanger and Coulson's own 'plus and minus' method published 2 years earlier[4]. The obvious advantages of reduced handling of toxic chemicals and radioisotopes rapidly made 'Sanger sequencing' the only DNA sequencing method used for the next 30 years.

With the ultimate goal of deciphering the human genome, the throughput requirement of DNA sequencing grew by an unpredicted extent, driving developments such as automated capillary electrophoresis. Laboratory automation and process parallelization resulted in the establishment of factory-like enterprises called sequencing centers that house hundreds of DNA sequencing instruments operated by cohorts of personnel. However, even successful completion of the two competing human genome projects did not satisfy biologists' hunger for even greater sequencing throughput and, most importantly, a more economical sequencing technology.

Stephan C. Schuster is at Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, 310 Wartik Building, University Park, Pennsylvania 16802, USA.
e-mail: scs@bx.psu.edu
PUBLISHED ONLINE 19 DECEMBER 2007;
DOI:10.1038/NMETH1156

Sequencing centers producing the Sanger sequence data for mammalian genome projects are factory-like outfits with a large number of personnel.

The first signs of what might revolutionize the sequencing market appeared in 2005 with the landmark publication of the sequencing-by-synthesis technology developed by 454 Life Sciences[5] and the multiplex polony sequencing protocol of George Church's lab[6]. Both groups used a strategy that greatly reduces the necessary reaction volume while dramatically extending the number of sequencing reactions. The strategy entailed arraying several hundred thousand sequencing templates in either picotiter plates or agarose thin layers, so that these sequences could be analyzed in parallel—a huge increase compared to the maximum of 96 sequencing templates on a contemporary Sanger capillary sequencer.

Although even the first version of 454's instrument could easily generate a throughput equivalent to that of more than 50 Applied Biosystem's 3730XL capillary sequencers at about one-sixth of the cost, the reaction of the scientific community was surprisingly reserved. Instead of embracing the new technology and rapidly adapting to use its enormous potential, many scientists accustomed to using Sanger sequencing raised issues such as sequencing fidelity, read length, infrastructure cost or simply objected to the need to handle the large volume of data generated using the new technology. This skepticism, initially echoed by funding agencies, may have been driven by the fear that substantial investments in Sanger-type sequencing hardware would become obsolete.

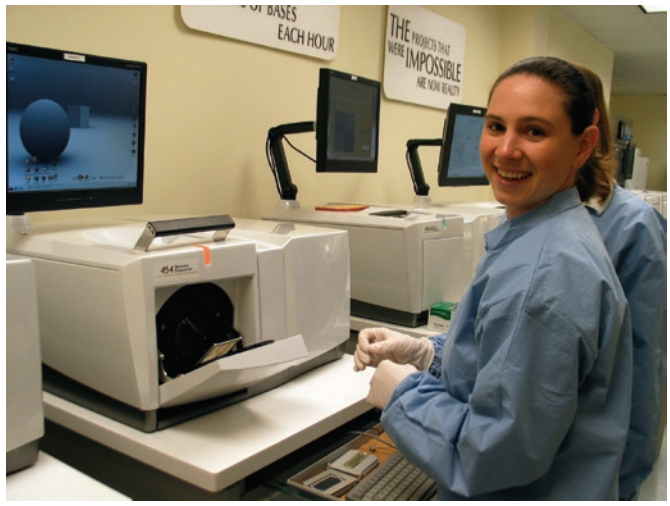
Most critics, however, overlooked the fact that many obstacles they attributed to next-generation sequencing were experienced in much of the same way by Sanger sequencing in its early stages, when read length rarely exceeded 25 bp and attained 80 bp only with the arrival of Fred Sanger's dideoxy terminator chemistry. The sequencing-by-synthesis technology, which uses pyrosequencing for readout, initially started with a read length of 100 bp, which after 16 months on the market had increased to 250 bp. Recent developments have raised the mark again to more than 400 bp, approaching

today's Sanger sequencing read length of ~750 bp.

Besides read length, the number of sequencing reads (or sequence tags) that can be produced in a single instrument run for a given cost is another important factor. These issues have been targeted by 454's competitors, whose systems generate up to tenfold more reads, albeit at the cost of a much shorter read length of 35 or fewer base pairs. Today three commercial next-generation DNA sequencing systems are available: namely Roche's (454) GS FLX Genome Analyzer marketed by Roche Applied Sciences, Illumina's Solexa 1G sequencer, and most recently Applied Biosystem's SOLiD system. Additional contenders, who are believed to be poised to enter the market within 1 to 2 years, are the '3rd generation' (also called 'next-next-generation') sequencing systems based on single-molecule analysis and developed by the companies VisiGen and Helicos.

Although the proof-of-principle study by Margulies *et al.*[5] demonstrated that small- and medium-sized bacterial genomes can be sequenced with one or two instrument runs, it was not immediately accepted that pyro-sequencing would enable sequence-based biological projects that were not feasible with Sanger sequencing. Initial projects involving Roche's 454 GS20 instrument either resequenced existing bacterial genomes or used the data to complement ongoing large 'Sanger projects'. Other initial studies immediately focused on metagenomics as this line of research, besides requiring datasets that can be larger than the human genome, has traditionally suffered from library-construction and cloning-host biases. From this point of view, the 454 technology rapidly appeared to present a key advantage in the combination of emulsion PCR and pyrosequencing. Emulsion PCR allows for bias-free amplification of single DNA molecules by entrapping them in lipid microreactors, thereby avoiding competition by multiple templates for a limited number of DNA polymerases. Pyrosequencing, in turn, can easily be performed in parallel and generate a visual signal for processing by a computer system. Early studies published in 2006 demonstrated the versatility of next-generation sequencing for unraveling the microbial diversity of a deep mine[7], rare biospheres in the deep sea[8] or marine viromes in multiple oceans[9].

A study in late 2005 combined metagenomics analysis with the field of ancient DNA research[10]. With a single run of a Roche (454) GS20 instrument, the analysis of 13 Mb of sequence from the nuclear genome of a 28,000-year-old mammoth became possible[10], thereby paving the way for the even more challenging project of deciphering the Neanderthal genome[11,12]. The ancient humanoid genome project faces even more difficult hurdles than the ancient elephantid project, as the amount of Neanderthal DNA that can be obtained from available samples is less than 5% of the total DNA extracted, when compared to a sample from a modern human. Therefore, 20 times more sequencing is required for the ancient project than for a modern human genome.



The latest next-generation sequencing instruments can generate as much data in 24 h as several hundred Sanger-type DNA capillary sequencers, but are operated by a single person.

Further, the combination of DNA damage occurring in samples with an ambient-temperature storage history and the next-generation sequencing error often exceeds the sequence divergence determined for modern humans and Neanderthals. The assertion that a given sequence is from the ancient specimen, and not a modern contaminant, is therefore easier for mammoth, as modern elephants, unlike humans, generally do not frequent the laboratory environment. The obstacles of obtaining genuine ancient mammalian sequences on a genome-wide scale will therefore require multiple-fold coverage of a given region or resequencing with a combination of methods to ascertain the origin. Both can only be achieved through additional dramatic cost-cutting for projects of this scale. This, together with the breakthroughs made for sequencing complex DNA mixtures from most diverse sources will allow for the study of any ecosystem of this planet at the sequence level. It will also open a window to the flora and fauna of at least the last 100,000 years, in ways far beyond what would have been deemed possible only a short while ago.

At the cellular level, next-generation sequencing has been applied to the resequencing of previously published reference strains, but it also allowed for the first time the identification of all mutations in an organism at the genomic level. Initial studies in 2005 identified drug-resistance alleles in *Mycobacterium tuberculosis*[13], while Velicer *et al.*[14] were the first to identify all mutations in a 9-Mb bacterial genome taken from a strain that had evolved for 1,000 generations. Through these early attempts it became clear that the new technology not only has the power to detect new mutations and allow identification of errors in published literature[14], but that it also has to deal with challenges, namely sequencing errors, such as homopolymer errors in pyrosequencing or rapidly deteriorating 3′ sequence quality in next-generation technologies with short read length.

The initial solutions were strategies that mixed Sanger and pyrosequencing data[15]. As the cost and effort of the Sanger component in any project still is prohibitively expensive, many laboratories now rely solely on next-generation sequencing data or combine the advantages of relatively long reads from pyrosequencing with the low operating costs of Illumina's Solexa or Applied Biosystem's SOLiD platforms, thereby independently verifying each system's performance. With the availability of more non-Sanger sequencing methods, it now becomes possible to assess both the next-generation sequencing accuracy and the correctness of the vast majority of Sanger-based reference sequences in the public databases.

The goal of generating large amounts of sequence data from closely related organisms is driving the application known as

resequencing, which handles the sequence data in different ways than *de novo* assemblies of genomes. In resequencing, the assembly is guided by a reference sequence and requires much less coverage (8–12×) than assembling genomes *de novo* (25–70×). One study using this approach sequenced 10 mammalian mitochondrial genomes[16], thus enabling population-genetic studies based on complete mitochondrial genomes rather than just short sequence intervals. Currently, many microbial sequencing projects are underway that will not only help to expand the number of available genomes, but also enable many comparative studies that will link genotype and phenotype at the genomic level.

Even the study of organisms that are not scheduled now for genomic sequencing will benefit from next-generation sequencing approaches that allow direct access to deciphering the cell's transcripts on the sequence level. Characterizing transcripts through sequences rather than through hybridization to a chip is advantageous in many ways. Most importantly, the sequencing approach does not require the knowledge of the genome sequence as a prerequisite, as the transcript sequences can be compared to the closest annotated reference sequence in the public database using standard computational tools. Knowing the sequence of transcripts will therefore truly revolutionize the research of organisms that are not now in line for genomic sequencing, and in some instances never will be. Initial examples for this line of research have shown that it is possible to align cDNA sequences to reference genomes as distant as the legume *Meticago truncatula* and the plant reference *Arabidopsis thaliana*[17] and revealed a large number of previously undescribed expressed sequence tags in *Zea mays* (maize)[18].

A similar transcriptomics approach could circumvent the problems posed by extremely large genomes. Despite having successfully enabled viral, microbial and large-scale mammalian sequencing projects, Sanger sequencing left the task of unraveling genomes of polyploidic plants to its successors. These gigantic genomes, often associated with crop plants, such as wheat with its 16-Gb hexaploid genome, have made previous sequencing attempts futile. However, the promise of much lower sequencing cost with the now proven concept of next-generation expressed-sequence-tag sequencing will allow assessment of plant genomes at least at the functional level[18].

Finally, next-generation sequencing has applications that are immediately relevant to the medical field. In cancer genetics, for example, specific cancer alleles can now be detected in tissues through ultra-deep sequencing of genomic DNA, in instances where previous Sanger-based trails have failed[19]. Short read length, initially deemed a major drawback of next-generation sequencing, becomes a blessing when the Sanger-based 700-bp read length is traded for a much larger number of sequence reads. As cancer genetics does not follow the path of Mendelian inheritance, laser-capture microdissection must be used for enrichment of the alleles of interest and targeted by sequencing of PCR products and/or amplicon sequencing while avoiding the traditional cloning and PCR biases.

Despite having already enabled a plethora of studies using next-generation sequencing, scientists and engineers who are working on this technology—and the companies that commercialize the applications—still have a long to-do list of improvements. High on the list is cost reduction: a reduction of 1–2 orders of magnitude is needed to deliver on the promise of personal genomics, which targets a cost of $1,000 for the resequencing of a human genome. Additionally, a reduced sequencing error rate would be highly welcome, not only for all present next-generation sequencing technologies, but also for Sanger sequencing, which clearly will continue to make valuable contributions in the immediate future. This might come in the form of custom-tailored DNA polymerases that provide a direct readout of DNA sequence in the form of emitted light, but even with these improvements we are unlikely to see a digital translation of DNA sequence into machine-readable code. As cost comes down, the amount of data are likely to skyrocket, creating an analytical bottleneck. Therefore much of the gain provided by future generations of sequencing instruments will be offset by increased costs and efforts on the bioinformatics front.

With the publication of more than 100 research articles in less than two years, next-generation sequencing has demonstrated its enormous potential for anyone working in the life sciences, at a time when many believed the age of post-genomics had arrived. It also has brought the field of genomics back into the laboratories of single investigators or small academic consortia, as is evidenced by the fact that the majority of next-generation sequencing publications originate from sites other than the large genome centers. One therefore will wonder, when looking back from the not too distant future, why the application of next-generation sequencing technologies initially was not more cheerfully welcomed in the scientific community and, more importantly, by the public funding agencies. Hopefully this lesson will have been learned when the 3rd generation of sequencing instruments is introduced, as by then the success of the current initiatives should have broken the ice that 30 years of Sanger sequencing have cast over the sequencing landscape.

ACKNOWLEDGMENTS

I thank W. Miller and N. Wittekindt for suggestions on the manuscript. The author is funded in part by the Gordon and Betty Moore Foundation.

1. Sanger, F. *et al. Nature* **24**, 687–695 (1977).
2. Sanger, F., Nicklen, S. & Coulson, A.R. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
3. Maxam, A.M. & Gilbert, W. *Proc. Natl. Acad. Sci. USA* **74**, 560–564 (1977).
4. Sanger, F. & Coulson, A.R. *J. Mol. Biol.* **94**, 441–448 (1975).
5. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
6. Shendure, J. *et al. Science* **309**, 1728–1732 (2005).
7. Edwards, R.A. *et al. BMC Genomics* **7**, 57 (2006).
8. Sogin, M.L. *et al. Proc. Natl. Acad. Sci. USA* **103**, 12115–12120 (2006).
9. Angly, F.E. *et al. PLoS Biol.* **4**, 2121–2131 (2006).
10. Poinar, H.N. *et al. Science* **311**, 392–394 (2006).
11. Green, R.E. *et al. Nature* **444**, 330–336 (2006).
12. Noonan, J.P. *et al. Science* **314**, 1113–1118 (2006).
13. Andries, K. *et al. Science* **307**, 223–227 (2005).
14. Velicer, G.J. *et al. Proc. Natl. Acad. Sci. USA* **103**, 8107–8112 (2006).
15. Goldberg, S.M. *et al. Proc. Natl. Acad. Sci. USA* **103**, 11240–11245 (2006).
16. Gilbert, M.T.P. *et al. Science* **317**, 1927–1930 (2007).
17. Cheung, F. *et al. BMC Genomics* **7**, 272 (2006).
18. Ohtsu, K. *et al. Plant J.* **52**, 391–404 (2007).
19. Thomas, R.K. *et al. Nat. Genet.* **39**, 347–351 (2007).