

Special News Feature

Sequencing's new race

Illustration credit: Tavares Jones

Next-generation sequencing has pushed forward the boundaries of genetic research and enabled the completion of a rapidly growing number of whole-genome sequencing projects. But the impending arrival of third-generation DNA sequencing technology could change the landscape yet again.

In the late 1990s, scientists J. Craig Venter and Francis Collins became household names during the epic race to sequence the human genome. Eventually that race ended in a tie in 2000, but similar levels of fame and publicity—not to mention millions in potential revenue—await the company or individual that wins sequencing's latest race: to develop and implement a so-called third-generation sequencing system.

The sequencing community turned its sights toward the promise of third-generation sequencing instruments less than five years ago, when a number of companies started promising single-molecule sequencing instruments that could provide faster, cheaper genome sequencing—instruments that would finally enable the burgeoning field of personalized medicine and take new fields of study such as transcriptomics to the next level.

So strong is the interest in new sequencing technologies that the National Institutes of Health (NIH) is supporting the race toward third-generation technologies through their \$1000 Genome Initiative, which provides funding to companies and individuals developing innovative solutions aimed at rapid, efficient DNA sequencing. Given the frenzied pace of development and the millions of dollars in financial support, many suspect that it is only a matter of time before someone crosses the finish line, commercializing a third-generation sequencing system that will finally give

researchers a full human genome sequence for less than \$1000.

Slow, long reads: the starting line

Frederick Sanger was awarded part of the 1980 Nobel Prize in Chemistry for his development of a method to sequence DNA. His approach uses gel or capillary electrophoresis to separate DNA fragments of differing lengths that have labeled dideoxynucleotides incorporated at the ends. By identifying each labeled nucleotide in the resulting DNA ladder, the ordered sequence of any DNA fragment can be determined. After nearly 30 years, Sanger sequencing, as it has come to be known, is still being used by many researchers for its simplicity and effectiveness along with its ability to accurately determine the sequence of long stretches of DNA.

A few years after Sanger won his Nobel Prize, developers at Applied Biosystems launched an automated DNA sequencer based on his method, which used fluorescently labeled dideoxynucleotides. Sanger's sequencing-by-synthesis approach to DNA sequencing, along with the newly developed automated instruments, would serve as the enabling technology for the Human Genome Project. Although the method has proved durable, reliable, and accurate, it suffers from being slow, expensive, and relatively low-throughput—key reasons why

Collins' government-funded researchers required \$3 billion over 13 years to generate their draft map of the human genome.

Recent advances in the Sanger methodology and associated computational assembly tools, along with the availability of the draft human genome sequence, enabled the assembly of another human genome sequence in 2007, using the dideoxy approach. This time, though, the genome was of a single individual: J. Craig Venter (1). Despite the obvious success of the Human Genome Project and the other recent whole-genome sequencing efforts, Sanger sequencing is slowly becoming outdated. Costs remain high and throughput is



Applied Biosystems' SOLiD sequencing system was commercialized in 2008, vastly improving upon their Sanger method-based automated sequencing system. Courtesy of Applied Biosystems.

not high enough to support the growing interest in personalized medicine and the desire to uncover the genetic basis of human disease. Since these efforts could require tens of thousands of individual genome sequences, geneticists have been searching for alternatives.

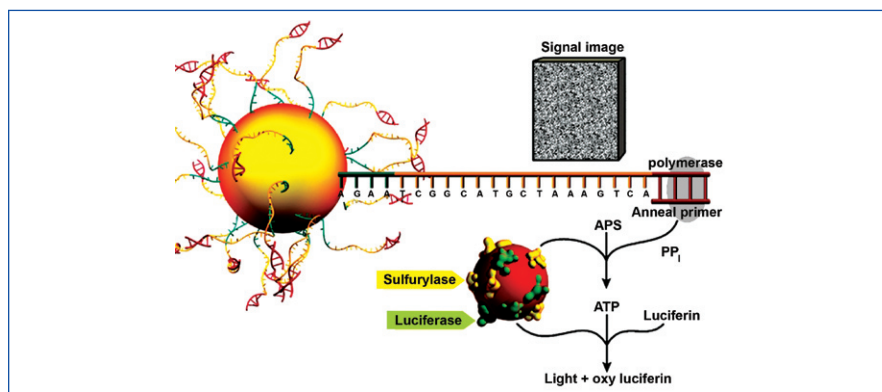
Claire Wade, a professor of veterinary science at the University of Sydney in Australia, recently completed the sequencing of the horse genome using Sanger methodology (2). “The horse was one of the last sequencing projects to be fully conducted using Sanger sequencing,” says Wade. “It represents the peak of our understanding of mammalian assembly on that platform.”

The next generation: ready to go

As it turns out, the desire to improve upon the Sanger approach and related instrumentation over the last decade has led to the development of a new generation of sequencing platforms that are currently used in labs around the world. But even these newer and faster ‘next-generation’ platforms might merely be holding scientists over until the arrival of third-generation technologies.

In 2005, Roche 454 Life Sciences became the first company to commercialize a next-generation sequencing platform (3). Their core technology maps genomes by attaching beads to DNA fragments which are then amplified, enriched, and loaded onto a multi-well plate for sequencing. Labeled nucleotides are then flowed across the plate wells, where each well contains a single bead. When a complementary nucleotide crosses the template strand and is incorporated, a chemiluminescence reaction produces a signal that is imaged and recorded. By performing the nucleotide flow step repeatedly with each nucleotide, the current GS-FLX system can generate 400 million raw bases of data per 10-hour instrument run, from sequence read lengths in excess of 400 bases.

Illumina’s Genome Analyzer (GA) IIx system was introduced in 2006 and takes a slightly different route. The GA generates millions of clone clusters on a solid support, which are then sequenced using reversible dideoxy terminator-based sequencing chemistry, which is akin to the Sanger methodology. Read lengths using Illumina’s system are shorter than those from Roche’s GS-FLX (averaging 100 bp), but the number of sequence reads per run is dramatically greater (the GA IIx generates 300 million reads per flow cell, in comparison to the GS-FLX’s 1 million per run).



Roche 454 Life Sciences’ approach to sequencing relies on attaching DNA fragments to beads which are amplified and placed into wells on a microwell plate. Nucleotides are flowed over the well and addition of a specific nucleotide to the DNA fragment on the bead results in a chemiluminescent reaction that can be imaged, allowing for the determination of the DNA sequence of the fragment attached to the bead. Courtesy of Roche 454 Life Sciences.

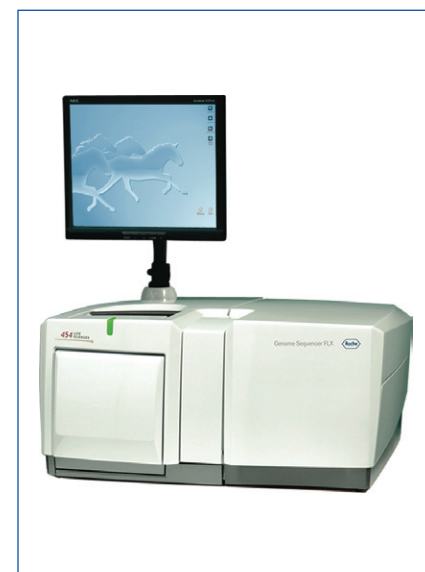
Applied Biosystems developed the SOLiD system in 2008, which combines elements of various approaches to sequence clonally amplified DNA fragments linked to beads. While the amplification and attachment to beads are similar to the Roche and Illumina platforms, Applied Biosystems’ platform relies on a unique sequencing by ligation approach using dye-labeled oligonucleotides, rather than sequencing-by-synthesis. The method actually provides two-base redundancy in sequencing reads, which the company says results in a higher accuracy. Similar to Illumina’s platform, the latest SOLiD 3 system generates large numbers of shorter reads: the current SOLiD system can generate more than 60 gigabases of raw data from more than one billion sequence tags per instrument run.

While shorter read lengths can cause problems with de novo genome assembly, the increased number of reads per instrument run—along with new computational tools designed specifically for next-generation sequencing analysis—have made it possible to decode whole genomes rapidly on these platforms, at a significantly reduced cost. In August 2009, Illumina announced the mapping of a single human genome at a cost of \$48,000.

With sequencing projects that once took more than a decade and billions of dollars now being done in just over 2 weeks for a little less than \$50,000, genomics researchers are applying next-generation instrumentation to projects that 7 years ago seemed impossible. Indeed, the NIH-funded 1000 Genomes Project, launched in 2008, involves labs around the globe and seeks to create an improved map of the variation found in human DNA by sequencing genomes from 1200 individuals. Even more ambitious is the Genome 10K project, an international effort to sequence

10,000 vertebrate species, or approximately one genome for every vertebrate genus. The high throughput and digital nature of next-generation sequencing technologies have enabled their application within emerging fields like transcriptomics, where it is important to quantify numbers of low-abundance messenger RNA molecules and understand how they regulate protein expression.

But in the end, according to Wade, it might be the developing ‘third-generation’ sequencing methods that have the potential to reveal even more about the composition and dynamics of complex mammalian genomes. “There needs to be sufficient read length to span longer repetitive sequences, and this represents a challenge with short read length massively parallel methods,” she explains.



In 2005, Roche 454 Life Sciences’ sequencing-by-synthesis instrument was the first next-generation platform to be commercialized. Courtesy of Roche 454 Life Sciences.

The third generation: almost there

Cambridge, Massachusetts-based Helicos Biosciences is sitting at the nexus of current next-generation sequencing systems and the emergence of third-generation sequencing. The company is the first to develop a platform based on single-molecule sequencing (4), an approach the company calls True Single Molecule Sequencing (tSMS). Capable of sequencing several billion bases in one instrument run in real time, the true advantage of single-molecule approaches like Helicos' might just lie in the ability to sequence without amplifying template DNA, permitting accurate quantification of specific RNA or DNA molecules rapidly from very small starting template amounts.

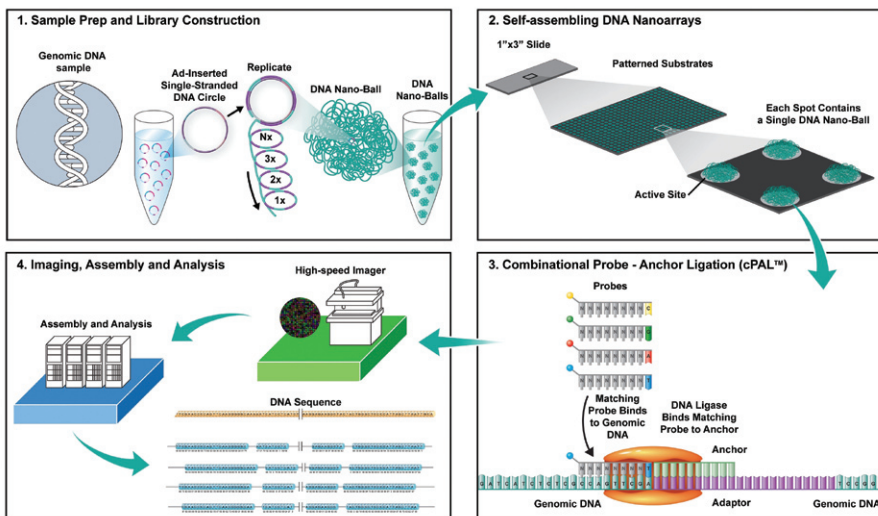
While Helicos' tSMS may have been the first approach out of the single-molecule gate, other methods are beginning to emerge.

"We anticipate that by 2013, the SMRT sequencer will be available and able to sequence a human genome for a few hundred dollars—and in a matter of minutes," says Sejal Sheth, director of product marketing at Pacific Biosciences. Bold statements such as this are common in the world of third-generation development.

Pacific Biosciences made a splash at the 2008 Advances in Genome Biology and Technology (AGBT) conference when the company's chief technology officer, Stephen Turner, showed early data on the effectiveness of SMRT, their approach to single-molecule sequencing. Turner's talk at AGBT was quickly followed by two articles describing the method in greater detail (5,6). Pacific Biosciences' approach works by sequencing strands of DNA on



Clifford Reid cofounded the company Complete Genomics in 2006 in an effort to develop a faster method to sequence DNA. Courtesy of Complete Genomics.



The Complete Genomics sequencing method includes the creation of novel clusters of DNA fragments and pieces of known sequence called DNA nanoballs. Courtesy of Complete Genomics.

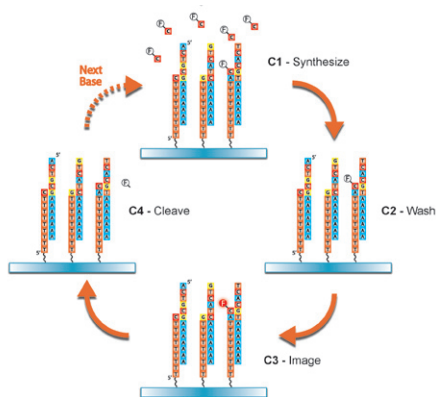
chips containing thousands of zero-mode waveguides (ZMWs). Each ZMW functions as a nanophotonic visualization chamber, providing an incredibly small detection volume of 20 zeptoliters (10^{-21} liters). At this volume, the activity of a single molecule can be detected against thousands of labeled nucleotides, which provides a window for watching DNA polymerase as it performs sequencing-by-synthesis. In each chamber, a single DNA polymerase molecule is attached to the bottom surface so that it permanently resides in the detection volume. Phospholinked nucleotides, each labeled with a different colored fluorophore, are introduced into the reaction solution at concentrations that promote enzyme speed, accuracy, and processivity. As the single molecule of DNA polymerase incorporates complementary nucleotides, each base is held within the detection volume for tens of milliseconds. During this time, the engaged fluorophore emits fluorescent light whose color corresponds to the identity of a particular base. The polymerase then cleaves the bond holding the fluorophore in place and the dye diffuses out of the detection volume. Following incorporation and cleavage, the signal immediately returns to baseline and the process repeats. The result of this process could be very fast sequencing—on the order of 50 nucleotides per second—with longer read lengths than is currently available with any next-generation or Sanger-based system.

Expectations at Pacific Biosciences are high. Sheth believes that with their current development timeline, the company will be the first to reach the \$1000 genome. "We firmly believe that we have the leading third-generation single-molecule DNA sequencing technology," she says.

Others disagree. "The technology that will deliver a true \$1000 genome will not rely on light and optics," says Zoe McDougall, director of communications at Cambridge, UK-based Oxford Nanopore Technologies. Oxford Nanopore is advancing the use of protein nanopores for DNA sequencing in a platform that uses an electronic rather than an optical signal to identify DNA bases (7). According to McDougall, this places Oxford Nanopore's approach in the unique position to drive down sequencing costs.

Using nanopores—small openings in the lipid bilayer of a cell—to sequence DNA was proposed years ago, but the technology to make the idea feasible has proved elusive. McDougall says Oxford Nanopore's approach makes nanopore sequencing a viable option. The company uses silicon chips containing a series of microwells where DNA samples are introduced. The lipid bilayer that lies across the top of the well gives a high-resistance electrical seal across which voltage is sent to drive a current toward the bottom of the well. The nanopores are the only point across which the current can flow. As each DNA base passes through a nanopore, it must first bind to an adaptor cyclodextrin. During this binding event, each base blocks the flow of current across the nanopore to a different degree; the variations allow researchers to identify which bases are passing through the pore. Similar to Pacific Biosciences, Oxford Nanopore demonstrated with initial data that their nanopore sequencing approach will be able to map as many as 25 bases per second (7) with the potential to sequence very long fragments of DNA without interruption.

Another company is taking their third-generation offering one step further, not



Helicos Biosciences' approach to single molecule sequencing involves binding of targets to a flow cell followed by sequential rounds of nucleotide addition, imaging, and cleavage. Courtesy of Helicos Biosciences.

only by creating a novel technology platform, but offering it as a sequencing service. In 2006, Clifford Reid, Radoje Drmanac, and John Curson came together to develop a new approach to how DNA is sequenced and how researchers obtain their sequences of interest. For the next 3 years, their company, Complete Genomics, worked on developing their proprietary sequencing technology. In 2009, Complete Genomics announced their intention to offer whole-genome sequencing services to researchers, and in February of that same year, the company released their first complete human genome sequence (8). The Complete Genomics model differs from other next-generation sequencing companies in that they do not sell their instruments; instead, the company sequences DNA samples that customers provide, and then reports back the results.

Complete Genomics published schematics of their sequencing system in November 2009 (9). The approach is to create 500-bp libraries with genomic DNA fragments of known sequence interspersed at regular intervals. These are then amplified in solution, in a single reaction chamber, which enables higher density and lower reagent usage. These resulting DNA nanoballs (DNBs) consist of more than 200 copies of the head-to-tail concatemers in a ball-like configuration. The DNBs are then transferred onto patterned silicon substrates with arrays of spots that are activated to capture and hold the DNBs in place, essentially allowing for self-assembly of the DNBs into DNA nanoarrays.

The next step involves Complete Genomics' combinatorial probe-anchor ligation (cPAL), which is an unchained, non-iterative, base-reading sequencing assay. According to Reid, this method has the advantage of dramatically reducing the required probe and enzyme concentrations and reducing imaging time, which

substantially cuts reagent and imaging costs, respectively. The collected images are then assembled with Complete Genomics' computing software.

Currently, Complete Genomics offers sequencing at \$20,000 per genome for projects with a minimum of eight genomes, and volume discounts for projects with more than 24 genomes. For those researchers with projects of even larger volume, the cost per genome is around \$5,000.

But McDougall cautions against "comparing apples to pears" when it comes to advancing toward the \$1000 genome. "When we [at Oxford Nanopore] speak about the cost of the genome, we like to consider the full cost. That means reagents, amortization of our instruments, labor, IT, informatics, project management, and sample preparation," she says. Oxford Nanopore's technology does not require reagents, which McDougall believes gives them an advantage when it comes to reducing total costs.

With fame and funds abounding, the third-generation sequencing race has attracted numerous competitors. The Archon X-Prize for Genomics was established in October 2006 and offers a \$10-million prize to the first team able to sequence 100 human genomes in 10 days. Research teams and companies from around the world have stepped up to the challenge. Participants include team cracker from Taiwan, base4innovation from the UK, and the US-based companies Visigen Biotechnologies, Reveo, ZS Genetics, and 454 Life Sciences. Also from the US is the Foundation for Applied Molecular Evolution and George Church's Personal Genome X.

At the finish line

Which company will be the one to break the \$1000 genome barrier first? What will be included in the total cost of that \$1000? How will a winner actually be decided? Though these questions remain unanswered, Pacific Biosciences, Oxford Nanopore, and Complete Genomics all predict there will be a winner soon—and according to each company, it's going to be them.

From the development of the Sanger method to the completion of the Human Genome Project, geneticists have made significant strides in understanding and accessing the information stored in our genes. Whether using optics, nanopores, nanoballs, or some yet-unannounced sequencing methodology, the completion of the third-generation sequencing race will mark another milestone in the history of genetics. Yet like all previously lauded achievements, it will be improved upon.

Speed, accuracy, easily assembled long reads, and reduced cost will only satisfy for so long. Once the \$1000 goal is reached, developers are likely to set their sights on the \$100 genome, and, perhaps, someday even the \$1 genome.

References

1. Levy, S., G. Sutton, P.C. Ng, L. Feuk, and A.L. Halpern. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* 5:e254.
2. Wade, C.M., E. Giulotto, S. Sigurdsson, M. Zoli, S. Gnerre, F. Imsland, T.L. Lear, D.L. Adelson, et al. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326:865-867.
3. Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, and L.A. Bembem. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
4. Harris, T.D., P.R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, et al. 2008. Single molecule DNA sequencing of a viral genome. *Science* 320:106-109.
5. Korlach, J., P. Marks, R. Cicero, J. Gray, D. Murphy, D. Roitman, T. Pham, G. Otto, et al. 2008. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. USA* 105:1176-1181.
6. Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, and S. Turner. 2008. Real-time DNA sequencing from single polymerase molecules. *Science*. 323:133-8.
7. Clarke, J., H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4:265-270.
8. Rosenbaum A.M., J.V. Thakuria, X. Wu, A.W. Zaranek, J. Li, P. Hulick, M. Murray, M.F. Browning, et al. 2009. Clinical analysis of individual genomes. *Nature*. (In press).
9. Drmanac, R. and C.A. Reid. 2009. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. 327:78-81.

Written by Erin Podolak.

Supplementary material for this article is available at www.BioTechniques.com/article/113371.

BioTechniques 48:105-111 (February 2010)

doi 10.2144/000113371