Cell

# Next-Generation Sequencing: The Race Is On

**The $1000 genome may still be years away, but with the arrival of next-generation sequencing (NGS) technologies that are much faster and cheaper than the traditional Sanger method, large-scale sequencing of hundreds or even thousands of human genomes is fast becoming reality.**

Sequencing of the 3 billion base pairs of the human genome took 3 to 4 years using conventional Sanger sequencing machines and cost about $300 million. But the commercial availability of next-generation sequencing (NGS) technologies that are up to 200 times faster and cheaper than conventional Sanger machines is spawning a flurry of ambitious new sequencing projects. "A drop in price of 200-fold doesn't happen very often," says geneticist George Church of Harvard Medical School. Church is cofounder of Knome, which announced in November last year that it would offer sequencing of a whole human genome for $350,000. Church has also started the "personal genome project," which aims to sequence relevant regions of the genomes of 100,000 people over the next two years.

In January, an international consortium that includes the National Institutes of Health announced the 1,000 Genomes Project, which aims to sequence the genomes of 1,000 people worldwide including individuals from Africa, Japan, China, the US, and Italy. "Such a project would have been unthinkable only two years ago," says Richard Durbin of the Wellcome Trust Sanger Institute in the UK, which is part of the consortium. The Beijing Genomics Institute in Shenzhen, China, another participant in the 1,000 Genomes Project, has started its own initiative to sequence the genomes of 100 Chinese people. And Craig Venter, founder and president of the J. Craig Venter Institute in Rockville, Maryland, says his institute plans to use the new technologies to sequence between 10 and 50 human genomes over the next year and up to 10,000 over the next 10 years. Finally, a few weeks ago, a company called Pacific Biosciences announced that it is working on a next-generation sequencing instrument that ultimately will be capable of sequencing a diploid human genome at 1-fold (1×) coverage in about 4 min. The company plans to start selling a first-generation version of the system by 2010. The race is now on for cheaper and faster DNA sequencing technologies capable of handling an ever greater number of human genomes.

## Shopping for Hardware

Several companies have commercially available NGS platforms. 454 Life Sciences, now owned by Roche, started selling its NGS sequencing machines in 2005. Next came Illumina's Genome Analyzer, which became commercially available in January 2007. And in October last year, Applied Biosystems (ABI) formally launched their NGS sequencer. A fourth company, Helicos BioSciences, announced last month that it received its first order.

Other than buying a commercial platform, users also have the option to build their own NGS system using components provided by George Church, in an open-source approach Church likens to the Linux operating system. According to Church, building such a system costs about $120,000, or about a quarter the cost of most of the commercially available NGS platforms.

Most NGS technologies eliminate the bacterial cloning step used in traditional Sanger sequencing and instead amplify single isolated DNA molecules and analyze them in a massively parallel way. Hundreds of thousands or even tens of millions of single-stranded DNA molecules are immobilized on a solid surface like a glass slide or on beads. In the case of 454, single DNA strands are attached to beads and then amplified by PCR in separate water droplets in oil for each DNA-loaded bead, such that there is a separate PCR and resulting clone for each DNA strand. The beads are then mixed with DNA polymerase and deposited in plates containing more than 1 million wells with one bead per well. Nucleotides then flow sequentially over the wells, and as each nucleotide is added to form complementary DNA strands, pyrophosphate is released and detected in a chemiluminescent flash.

The NGS platforms differ in several ways, such as read length and the number of DNA molecules they sequence in parallel. Traditional Sanger sequencing machines analyze terminally labeled DNA strands and can read about 800 bases of 100 DNA molecules simultaneously. In contrast, the NGS platforms read many more DNAs in parallel but have shorter read lengths. For example, 454's GS FLX machine reads 400,000 DNAs that are each about 250 bases in length; Illumina's Genome Analyzer and ABI's SOliD platform can read tens of millions of DNAs up to about 35–50 bases in length. These three NGS platforms also differ in cost per base: the 454 machine sequences are about ten times cheaper than traditional Sanger technology, and Illumina and ABI are 100 times cheaper, says Steven Jones, head of bioinformatics at the Genome Sciences Centre in Vancouver, Canada.

## Forging New Research Paths

NGS technologies have made a variety of ambitious sequencing projects feasible. Sequencing the Neanderthal genome has been "made possible totally by the high throughput sequencing technologies," says Svante Pääbo of the Max Planck Institute for evolutionary anthropology in Leipzig, Germany, who heads the project. A big challenge is that only 1%–6% of the DNA in ancient Neanderthal bones is Neanderthal, the remaining DNA is from bacterial contamination. "I think the key with 454 is that it's actually PCR of a single isolated DNA molecule on a bead," Pääbo says. This, he points out, gives even rare DNAs a chance to be detected. "The polymerase has to

work with the one molecule that is there," Pääbo says. With traditional sequencing, the DNAs would have to be amplified first in a mixture, competing with millions of nonhuman DNAs.

The 454 sequencer, with its relatively long reads of about 250 bases, "is the ideal read length for ancient DNA," Pääbo says. That is because it takes at least 30 base pair long high-quality reads to ensure that bacterial sequences do not show up as apparent matches with human sequences. In 2006, Pääbo's group reported the sequence of 1 million bases of Neanderthal DNA after sifting through a total number of 20 million bases. Pääbo has calculated that it will take 2,000 runs on a 454 machine to get 1× coverage of the full Neanderthal genome. Pääbo says he is ready to publish 1%–2% of the Neanderthal genome sequence soon using the original Neanderthal specimen from Croatia, as well as two additional specimens discovered in Germany and Spain.

The relatively longer reads of the 454 platform are also the choice for sequencing regions of bacterial 16S rRNA genes to study bacterial diversity. Mitchell Sogin of the Woods Hole Marine Biological Laboratory says he uses 454 because he needs a stretch of at least 100 bases to reliably determine how a sequence is related to known bacterial rRNA sequences. "We need to have on the order of 100 base pairs of continuous information to make this work," Sogin says. He notes that for his analyses, sequencing with 454 is about 100 times cheaper than traditional sequencing. In 2006, for example, he sequenced 130,000 16S rRNA genes from sea water and found that one liter contains at least 25,000 different kinds of microbes. Rob Knight at the University of Colorado also uses 454 machines to sequence fragments of bacterial 16S rRNA genes. Among other sequencing projects, Knight is collaborating with Jeffrey Gordon of Washington University in St. Louis to compare bacterial diversity in the guts of obese and nonobese twins. The idea is to use twins to untangle diet and environmental factors from each other and from genetic factors, Knight says. "The goal is to get twins that are concordant and discordant for obesity and [look at] systematic changes."

Whereas 454's longer reads are the best choice for some research projects, the shorter read lengths generated by the Illumina machines work better for other projects, such as combining sequencing with chromatin immunoprecipitation (ChIP) assays in which DNA molecules bound by proteins are isolated with antibodies. Until recently, these DNA pieces were analyzed by hybridization to DNA probes on microarray chips (called ChIP-chip). But several recent studies instead use Illumina's NGS platform to sequence and count all of the isolated DNAs in a new approach dubbed ChIP-Seq. Keji Zhao of the National Heart, Lung, and Blood Institute in Bethesda, Maryland, and his group used ChIP-Seq for the first human genome-wide mapping of 20 different types of histone modifications. Illumina's machines are ideal for ChIP-Seq because they can read about 30 million DNAs per run, says Zhao, whereas 454 only sequences several hundred thousand DNAs per run. "The most important feature we need is the sequence tag number," Zhao says. "We just need 25 base pairs to identify [any unique] sequence in the genome."

ChIP-Seq is an example of an NGS application that only requires resequencing (instead of de novo sequencing) of genes that are already mapped to a sequenced and assembled genome. Shorter read lengths are sufficient for such applications because the reads only need to be long enough to find a unique match in the assembled genome. But NGS read lengths are not long enough for de novo sequencing projects because it is too difficult to assemble the shorter reads, and so traditional Sanger sequencing still has its place. "There is no question that, today, one cannot assemble human genomes as effectively by using the new methods, as can be achieved by using the Sanger technology," says Jeffery Schloss of the National Human Genome Research Institute. Eric Lander, director of the Broad Institute of MIT and Harvard University, agrees. The new sequencing technologies do not have the ability right now to assemble a mammalian genome, he says.

But at least with 454, it is possible to do de novo assembly of bacterial sequences, points out Chad Nusbaum, codirector of the genome sequencing and analysis program at the Broad Institute in Cambridge, Massachusetts. "When you sequence a bacterium by traditional methods, you do 6–8× coverage and you get a really good assembly," Nusbaum says, adding that with the 454 sequencer, about 15× coverage is needed to get good assembly. Because of the required additional coverage, 454 is probably less than ten times cheaper than traditional Sanger sequencing, Nusbaum points out. And with the even shorter reads of the Illumina machine, de novo assembly is not yet possible but should be soon, he adds. Nusbaum is combining traditional sequencing with the 454 and Illumina platforms to sequence de novo the genome of the bacterial pathogen *Listeria monocytogenes*.

## Next-Generation Sequencing Gets Personal

In October 2006, the X Prize Foundation announced a $10 million prize for the first team to sequence 98% of 100 human genomes at 99.999% accuracy over a 10 day period for $10,000 per genome. The current technologies are nowhere near that, Venter says. "I don't think any of the existing commercial technologies have a chance of winning the X Prize," emphasizes Venter, co-chair of the scientific advisory board for the genomics X Prize.

"The current cost to sequence a human genome is hundreds of thousands of dollars," says David Altshuler of the Broad Institute. And Lander points out that "the $1,000 genome is going to take another generation or beyond." Others are more optimistic. With next-next-generation sequencing technologies such as nanopore sequencing on the horizon, some say that the $1,000 genome is less than ten years away. "If you just graph things statistically, it's going to happen within five years," Venter says. That price will put the sequencing of an entire genome within everybody's reach, emphasizes Church. "Suddenly, millions of people can afford to get their own genome analyzed and correlations with medical and non-medical traits explored," Church says, adding that there will be a lot to sequence. "There are 6 billion humans and each of them has 6 billion base pairs," he says. Church believes

that most of these sequencing efforts will be for resequencing, using the blueprint of the assembled human genome for comparison.

Still, it may be too soon to focus on resequencing without obtaining more detailed de novo sequences of more than just a few individual humans, says Venter. We still do not know enough about how much variation there is between individual human genomes, he adds. To improve the quality of the available draft sequence, Venter has used the more accurate, longer reads of the traditional Sanger sequencers. He recently published an improved draft of his own genome sequence, with 99% coverage of both sets of chromosomes. Given the variations between genomes, it is also important to compare many people to be able to tell which variations are the relevant ones, says Lander. "The real way that you end up doing this is you look at many people and then you say aha, I am seeing a mutation in the same gene."

## A SNP at the Price?

NGS technologies are also being used to confirm and extend associations between single nucleotide polymorphisms (SNPs), identified in genome-wide association studies, and certain diseases. More than 100 associations have been identified so far for a variety of diseases including rheumatoid arthritis, prostate cancer, and type 1 and 2 diabetes. The Broad Institute's Altshuler leads one of the groups that have identified ten associations between SNPs and diabetes type 2. Until now, these studies used microarrays that tested patients for SNPs largely derived from the HapMap project, which has identified millions of SNPs across the human genome. The big challenge is to find the biologically relevant variations that are the underlying cause of disease. To do this, Altshuler is using the Illumina NGS platform to "deep sequence" the regions around the SNPs associated with type 2 diabetes in hundreds of people with and without the disease. "We are sequencing lots of people around these diabetes loci to figure out not just the common SNPs but rare [variations] as well to get a fuller picture of how [they] are actually altered to affect disease," Altshuler says. "In almost no cases do we know the actual causal genetic change." This means using sequencing to find every possible variation to recognize mutations that cause the disease, he says. In another approach, Jeremy Edwards of the University of New Mexico in Albuquerque is planning to use NGS to sequence the entire chromosome 6 of several thousand melanoma patients. "We will try to uncover new SNPs that are related to poor survival in people that get melanoma," Edwards says.

Altshuler notes that additional sequencing results will also be used to improve currently available SNP arrays. "What's going to happen is that hundreds or thousands of people will be sequenced in the coming years because the technology makes it possible," Altshuler says, "and those SNP chips will then be upgraded to have all the common genetic variations on them." Eventually, sequencing will replace SNP chips, but only when it becomes cheap enough. Currently, it costs hundreds of dollars to test one million SNPs, Altshuler says, much less than the hundreds of thousands of dollars to sequence a human genome. "If you could sequence people with great accuracy and at a cost that was hundreds of dollars then I am sure it will replace the genotyping," Altshuler says.

With his "personal genome project," Church hopes to sequence 1% of the genomes of 100,000 people over the next two years. "That's the sort of number that you need in order to get statistical significance," he says. The project will correlate genetic information with physical characteristics and medical traits, and the sequencing will focus on coding regions. "About 95 percent of trait-affecting mutations impact coding regions," says Church. The first ten volunteers have enrolled and given blood, saliva, and skin tissue samples, and enrollment of the general public will begin soon.

## A Glimpse of the Future

So what will sequencing look like in the future? Both companies and academic groups are working on next-next-generation sequencing technologies. Nanopore sequencing, for example, measures the change in current as a single DNA molecule is pulled through a tiny pore. This approach could achieve long read lengths, says David Deamer of the University of California at Santa Cruz, who is developing nanopore sequencing. "We can pull DNA that's several thousand bases long through a pore," he says. Deamer predicts that the sequencing devices should be quite cheap, in the range of several thousand dollars in larger production runs, and nanopore sequencing could also save costs for sample preparation, as the only step will be purifying the DNA, without the need for amplification. His group is using natural nanopores made of a bacterial protein, and others are using artificial nanopores to pull through the DNA strands. Deamer expects that sequencing with nanopores will be demonstrated in about a year, although surpassing the speed of the current NGS platforms is still several years away. "We are aiming for the $1,000 genome," Deamer says.

Still, it remains unclear who will win the race for the $1,000 genome, or the X Prize, for that matter. But as Lander points out, "Race implies that the first person over a finish line wins, but that's not how science is done," he says. "It's a never ending progression." Lander adds that as sequencing technology improves, it will be possible to use it for increasingly demanding tasks. After ChIP-Seq, "the next simplest problem is mutation detection," Lander says. "The hardest problem is de novo assembly. Everybody is working up that chain right now." Lander, for his part, says that he does not have any favorite technology. "I see potential advantages in all of these," he says. "Science will be served by the diversity of approaches. All of these [will] contribute to progress. Keep it coming."

**Andreas von Bubnoff**
New York, NY, USA
DOI 10.1016/j.cell.2008.02.028