



## Next-Generation Sequencing

New sequencing technologies speed up large sequencing projects—for scientists prepared for a flood of data.

GAIL MANDEL, AN HHMI INVESTIGATOR AT THE OREGON Health & Science University (OHSU), wanted to understand the role of a particular protein in nerve cell function. She wanted to prove that it was a master regulator for neuronal genes. Using techniques she developed with colleagues at Brookhaven National Laboratory and OHSU, her lab prepared to sequence all the genes in the mouse genome that bound this protein. Before they could determine the extensive sequence of the thousands of DNA nucleotides, they had to prepare numerous pools of DNA, each representing a particular section of the region to be sequenced. The entire process took more than a year.

Meanwhile, another group of researchers at Caltech studying the same protein came up with virtually the same result. “Our data were very similar. But it took them much less time to get more sequence information,” says Mandel.

While her lab’s approach relied on traditional sequencing methods that yield a maximum of 96 short stretches of DNA sequence at a time, the other group used the “next-generation” Solexa sequencing platform from Illumina, Inc., which produces tens of millions of DNA sequences in a single run.

Welcome to the new world of warp-speed DNA sequencing.

Researchers agree that Solexa and two competing systems (the Roche (454) GS FLX sequencer and the SOLiD sequencer from Applied Biosystems) represent a breakthrough in sequencing that is speeding the pace of discovery, making it feasible for researchers

to conduct experiments once considered too expensive or simply impossible.

In 1975, biochemist Frederick Sanger developed one of the first manual sequencing systems that enabled scientists to determine the order of the nucleotides—known by the letters A, T, C, and G—that make up DNA. The process became automated in the 1980s. In today’s version of Sanger sequencing, each type of nucleotide is labeled with a different colored fluorescent tag. DNA fragments differing by only a single nucleotide are separated on the basis of size. Special optics detect the fluorescent nucleotides, creating images that can be “read” as a DNA sequence.

Next-generation systems also analyze fragments of DNA but step up the process by multiplying the number of sequencing reactions that occur on each piece of template DNA (see figure), yielding vastly more data. Solexa produces shorter DNA sequence reads than traditional methods of sequencing, but many more of them—up to 50 million sequences in a single two-day run. 454 and SOLiD use different techniques for sequencing but likewise produce large amounts of data.

HHMI has purchased 15 of the \$500,000 Solexa sequencers for its investigators, and its investment in next-generation sequencing equipment is already paying dividends.

Joe DeRisi, an HHMI investigator researching malaria and emerging viral diseases at the University of California, San Francisco, gives the example of a fragment of a novel virus discovered in his lab a couple of years ago.

**A**

Next-generation DNA sequencers are designed to read sequences from a very large number of individual DNA molecules within a mixed sample. However, not even the new techniques are sensitive enough to truly read a sequence from a single molecule of DNA. Instead, individual **DNA strands must be immobilized and amplified** in a fixed location to form a group of many identical strands that can provide enough signal for a sequence to be read. As a first step, DNA mixtures are prepared by fragmenting the DNA to be sequenced into small sizes (~100–500 bases); then, defined snippets of DNA are added to the two ends of each resulting fragment.

**B**

In the Solexa sequencing system (Illumina, Inc.) two kinds of amplification primers—short pieces of DNA—are fixed to a glass surface, much like a microscope slide. Then, the mixture of template DNA fragments is added to the slide, where individual molecules bind at random positions to the surface-bound primers. Multiple cycles of amplification yield double-stranded DNAs, which are created by bridging between the two types of surface-bound primers. Because each molecule of DNA can reach only so far, this approach tends to **create a “forest” of fragments at a given spot on the slide**. By controlling the number of input DNA molecules, the density of the amplified spots can be set as desired.

**C**

The Solexa sequencing strategy uses modified nucleotides: each base bears a different fluorescent group that also blocks further nucleotide addition. **All four nucleotide bases are flowed simultaneously over the surface**, and each template can incorporate only a single nucleotide. A laser activates each fluorescent group in turn and optics collect images of the surface. The pattern of colored light emission can be used to read the sequence directly from each spot (see magnification at left). Once the fluorescent group and blocking group are cleaved from the growing DNA strand, the next base can be read by precisely the same strategy.

**D**

Though it may seem complex, the process works remarkably well. The Solexa platform permits short sequence reads, with reliable reports of 50 nucleotides being obtained from each template. Moreover, the system **can read tens of millions of DNA sequences simultaneously**.

“This virus was very different than anything that had ever been published before,” he says. “A very talented postdoc threw everything but the kitchen sink at this project, trying to sequence a complete copy of the virus’ genome and was unable to do so.”

The DeRisi team was stumped. Last year, they began working with the Solexa sequencer and decided to revisit the problem. They attempted to sequence the whole pool of DNA from which the virus fragment was isolated.

“After a single run, we had recovered the entire genome in one go. In two days we were able to accomplish in totality what we couldn’t in over a year of hard trying,” says DeRisi, whose team is preparing to publish their discovery.

DeRisi points out that the new technologies are not a replacement for traditional sequencing methods, which are ideal for sequencing a very specific piece of DNA. But because of the ability of the new technologies to tackle large projects or complex mixtures of DNA, he and other HHMI researchers have great expectations.

Cancer researcher Bert Vogelstein, an HHMI investigator at Johns Hopkins University, can now go through millions of genes to find the rare, tumor-specific ones that provide clues about the origins of disease—and about targets for therapy or cancer tests.

“Before, it was just too expensive to study a lot of patients, but now we can,” Vogelstein says.

The technology isn’t without its challenges, however. For example, the new systems spew out data with the force of a fire hose.

HHMI investigator Greg Hannon, whose team uses the Solexa instrument at Cold Spring Harbor Laboratory to study gene regulation, says it produces a terabyte, or one trillion bytes, of raw data. That’s more data than can be transferred easily via the Internet. So his team started filling hard drives with the data and “transferring” the information by foot to cars for delivery to a data center. He calls this a “sneaker protocol.” That’s sneaker as in shoes.

“Many investigators are not prepared to deal with this large amount of data,” says Thomas Tuschl, an HHMI investigator at Rockefeller University who studies the role of RNA in gene silencing. “They do two runs and then they’ll spend a year trying to build up the software to interpret it.”

Over the years, Tuschl has built relationships with experts in bioinformatics, who have built software to handle his deluge of data. He expects the new platforms will double the pace of discovery in his lab, while lowering costs fivefold.

Hannon, also, isn’t put off. “This is the natural evolution of every technology. It goes from the effort required to learn to drive a car to thinking about where you’re going to go.”

As for Mandel, her experience convinced her to purchase a Solexa system. She says Solexa will not only give her more data more quickly, it will also enable her to piece together more easily the big picture of the proteins involved in neurological function. ■

—HOWARD WOLINSKY