SANGER WHO?

SEQUENCING THE NEXT GENERATION

In November 2008 Elaine Mardis of Washington University in St. Louis and colleagues published the complete genome sequence of an individual with acute myeloid leukemia. Coming just a few years after the decade-long, multibillion dollar Human Genome Project, the paper was remarkable on several levels. For one thing, the team sequenced two human genomes, both cancerous and normal, some 140 billion bases in all. More impressive, though, was what the study omitted: the 50 human genomes Mardis sequenced that year (albeit not as deeply) for the 1,000 Genomes Project. "It's like a whole new world," she says. Welcome to the sequencing frontier. **By Jeffrey M. Perkel**

Elaine Mardis's acute myeloid leukemia work comprised about nine months of collecting 32-base snippets at the rate of about a billion bases per instrument every five days, with five instruments running in parallel, she says. "That seemed like a huge amount [of sequence] at the time," she recalls.

The instruments in question, **Illumina** Genome Analyzers, are one of a cadre of so-called next-generation DNA sequencers. Over the past five years they have wrested control of the high-end sequencing market from the once-dominant Sanger dideoxy sequencing chemistry and its workhorse, the 3730xl from **Applied Biosystems** (now part of **Life Technologies,** formerly **Invitrogen**).

Yet today, says Mardis, those heady gigabase-a-week days seem "sort of like 'ho hum, that took a really long time.'"

Adam Lowe, Illumina's director of life science product marketing, estimates that his company's user base "generates about a thousand gigabases per week," he says, or about 20 times the size of Genbank in 2005.

Harvard University geneticist and next-gen pioneer George Church says the rate of technical improvement in the sequencing arena is unprecedented, about 10-fold per year, and far outpaces Moore's Law. Illumina reads are now at 75 bases standard, and have been pushed as far as 250 with overlapping paired-end reads (about 40 gigabases per run). Life Technologies has doubled the throughput on its next-gen SOLiD instrument every quarter. At those levels, says Mardis, the study that took some "90-ish" Illumina runs to accomplish in 2008 would require just six or eight today.

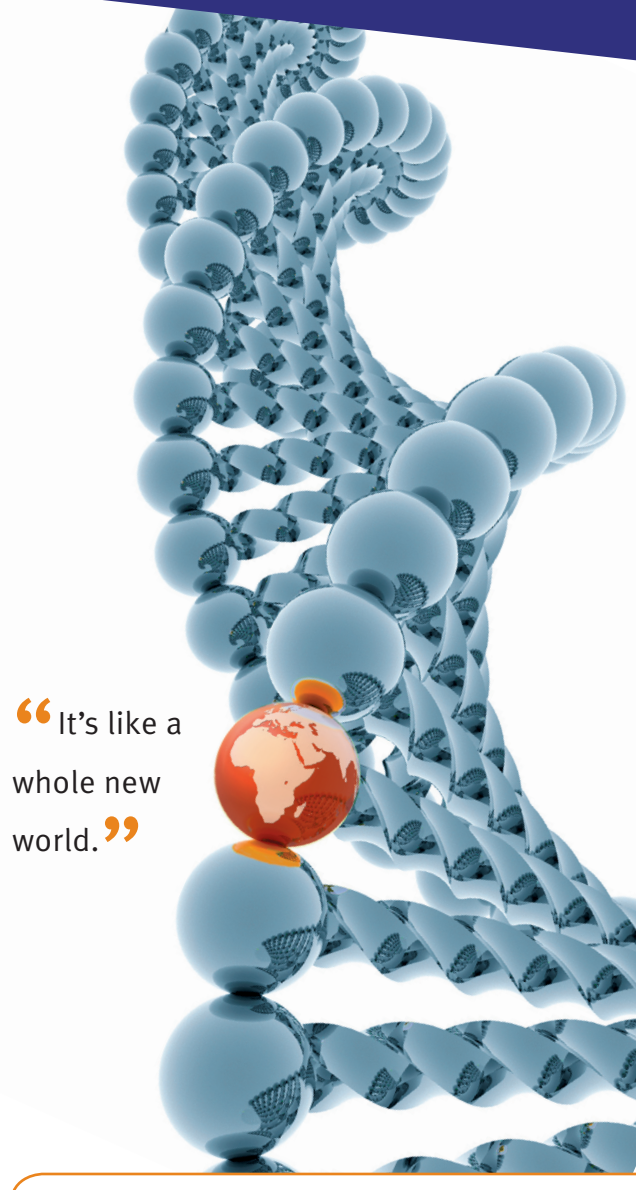
Such is life on genomics' bleeding edge. Rising from relative obscurity in 2005 to *Science's* Breakthrough of the Year in 2007, the next-gen sequencing industry now sports five commercial offerings, with several others nearing release. Employing different strategies and addressing different applications, each promises previously unimaginable data output.

Unsurprisingly, the technology is attracting attention. "People can't seem to get enough of it," says Church. "When you get a factor of 10,000 [improvement] in four years, people eventually notice."

Out with the Old...

Prior to 2005, almost all DNA sequencing used a variant of the chemistry first described in 1977 by Fred Sanger.

Sanger's methodology coopts the normal process of DNA synthesis by blocking the growth of new DNA chains using a sort of molecular brake called a continued >



"It's like a whole new world."

Look for these **Upcoming** Articles

Molecular Diagnostics — May 8

Nucleic Acid Purification and Manipulation — May 15

Technologies for Gene Transfer — June 19

Inclusion of companies in this article does not indicate endorsement by either AAAS or Science, *nor is it meant to imply that their products or services are superior to those of other companies.*

"The applications tend to break apart on read length dependency versus tag density."

dideoxynucleotide terminator. The resulting pool of molecules, which on average will terminate at every position, can then be sequenced chromatographically, originally on large polyacrylamide gels, and later in hair-thin capillaries.

The ABI 3730 ran 96 capillaries in parallel, each capable of producing between 500 and 1000 bases of high-quality data per run. At that rate, says Mardis, the system could produce about 1.2 megabases per day. Compare that to the gigabase throughputs being generated on new equipment, and it's clear why sequencing centers are mothballing their old equipment.

But throughput isn't the only factor; Sanger sequencing is labor-intensive and expensive. DNA to be sequenced first must be cloned, and the resulting libraries maintained. That requires instrumentation and labor, not to mention lab real estate.

New sequencing technologies employ completely different paradigms. All avoid size-separation in favor of strategies in which fragmented DNA is immobilized in a fixed position and repeatedly interrogated, like an iterative microarray assay. Most, but not all, use polymerase chain reaction to amplify that DNA; **Helicos Biosciences, Pacific Biosciences,** and **ZS Genetics** actually read single molecules instead. **454 Life Sciences** (part of **Roche Applied Science**), Illumina, Helicos, and Pacific Biosciences use DNA polymerase to drive their sequencing reactions, but Polonator (**Dover Systems**), SOLiD (Life Technologies), and **Complete Genomics** sequence with DNA ligase, and ZS Genetics uses electron microscopy. And whereas most reactions are synchronous—that is, sequential, with a single base interrogated at a time at each position—454 and Pacific Biosciences generate asynchronous reads, such that individual reactions run at their own rates and are not synchronized to one another.

The Longest Read

In 454's process the DNA to be sequenced is fragmented into 500-to-1,000-base-pair pieces and capped on each end with adaptors. The fragments are then amplified on bead surfaces via emulsion PCR (emPCR), a massively parallel strategy that separately amplifies each fragment inside an aqueous microdroplet in oil emulsion to create a sort of run-time sequencing library. Michael Egholm, chief technology officer and vice president of research and development at 454 Life Sciences calls emPCR one of "several key innovations" at the heart of the company's success.

Following amplification, the emulsion is broken and the beads placed into the wells of a PicoTiterPlate (PTP), a 6-by-6-cm support of some 3.5 million packed optical fibers etched with wells on one side. "We deposit the beads in each of these [fibers], and cleverly, the holes on the end of the optical fibers are such that there's only room for one bead," says Egholm. The result: an immobilized array of 3.5 million beads, each containing millions of identical DNA fragments. Each bead is what Church calls a polony, or polymerase colony.

The sequence itself is read via pyrosequencing, which monitors base-incorporation via the resulting release of pyrophosphate. Pyrosequencing converts that pyrophosphate into ATP, which in turn drives luciferase. As a result, a burst of light is produced whenever a new base is incorporated.

In 454's Genome Sequencer FLX, this process occurs in a flow cell. Basically, each base is flowed sequentially over the PTP – first A, then C, then G, then T. If the next base in the template is a G, the polymerase must wait until dCTP flows in. At that point, it will incorporate the base and release pyrophosphate, resulting in a flash of light whose intensity is directly proportional to the number of bases added (CCC will yield light three times as intense as C alone). That light is picked up by the optical fibers and transmitted to a camera, which reads the reaction.

Illumina and Helicos don't use pyrosequencing, yet their processes are largely similar, except the amplification (in the case of Illumina's technology) occurs directly on the flow cell rather than on beads, and the synthesis uses fluorescently labeled, reversible terminators; the reactions thus pause after each incorporation event (as if using a sort of Sanger sequencing 2.0). Helicos eliminates the amplification step, using what it terms true single molecule sequencing.

On the other hand 454 uses only standard DNA building blocks. As a result, says Egholm, it is both fast and free of background. "It's almost biblical, there's light and then there's no light," he says. And, producing by far the longest reads of any next-gen instrument, between 400 and 500 bases per bead, Egholm says the FLX can sequence 50 million bases per hour.

The Power of the Short Read

With read lengths approaching those of the 3730, 454 far outpaces the 125 bases Illumina is rolling out, not to mention the SOLiD's 75, Complete Genomics' 70, or the Polonator's 26. As such, it has become the de facto choice for metagenomics, immunogenomics, viral profiling, whole-transcript sequencing, and especially de novo genome sequencing. The technology has been used to sequence and assemble the *Arabidopsis* and *Drosophila* genomes from scratch—but not the human; when James Watson's DNA was decoded in April 2008, that was resequencing, aligning reads to the preexisting reference framework made possible by the Human Genome Project.

On the other hand, the FLX's 1.25 million reads is a mere fraction of what other instruments can produce. The Polonator yields 200 million to 400 million mappable reads and the SOLiD about 750 million. At the February 2009 Advances in Genome Biology and Technology (AGBT) meeting, Illumina presented data suggesting it could generate 520 million mappable reads per paired-end run. At those levels, a whole different set of applications opens up, including digital RNA profiling, targeted resequencing, and polymorphism discovery.

"The applications tend to break apart on read length dependency versus tag density," says Kevin McKernan, senior director of scientific operations for SOLiD at Life Technologies. At the **Broad Institute of Harvard and MIT,** whose 40 3730s, 20 Genome Analyzers, 10 FLXs, 8 SOLiDs, and one Polonator churned out 3 *petabases* of sequence in 2008, the SOLiD tackles "applications that require tons of data," such as polymorphism discovery and tumor profiling, says continued >

Chad Nusbaum, co-director of the institute's genome sequencing and analysis program.

The SOLiD, along with the Polonator and Complete Genomics' process, is based on sequencing by ligation, a strategy Church first successfully demonstrated in 2005 on *E. coli*.

The process, Church explains, "is directly mappable to sequencing by polymerase. In both cases you've got a template and a primer. In one case polymerase adds a mononucleotide, and in the other case ligase adds an oligonucleotide 6-to-9 bases long, where one of the bases is keyed to the color."

In general, given a primer-template pair, you add a pool of short oligonucleotides whose sequence is completely random, except that one base corresponds to the fluorescent dye attached to the molecule; you then let ligase make the base call.

Say you are using six-base-long oligos and interrogating base No. 3. Of the 4,096 possible hexamers, 1,024 have an A at position 3 and a corresponding color, 1,024 have a C at that position and a different color, and so on. Only that one oligo whose sequence precisely matches the template will bind strongly enough to be ligated, so that, when the unbound molecules are washed away, the reaction will glow a uniform color. Then, to read the next base, simply denature the primer-template pair, add new primer, and repeat.

One advantage of this approach is that, unlike polymerase-based methods, the bases may be read out of order, thereby eliminating polymerase-induced errors. "In a way, it's better than the polymerase, where you go sequentially, in the sense that there's a certain element of random access to this," Church says. Another advantage: unlike with polymerases, ligase can sequence in both the 5'-to-3' and 3'-to-5' directions.

But ligation strategies also produce extremely short fragments, which must then be aligned to a reference. Complete Genomics generates 70 bases by reading a few bases from each of eight start sites; the Polonator interrogates 26 bases by reading two sets each of six and seven bases, respectively, from either end of a longer DNA fragment (a strategy called paired-end sequencing, also supported by Illumina and Life Technologies, which improves the mappability of short sequences by adding phase information). Life Technologies actually garners the longest contiguous reads of any ligation strategy – up to 75 bases, according to data presented at AGBT—by reading two bases at a time at five-base increments, resetting, and repeating the process with a one-base frameshift.

The Next Next-Generation?

Other companies are pushing alternate strategies. Like Helicos, Pacific Biosciences is pursuing single molecule sequencing. The company arrays DNA polymerases on the surface of a plate, relying on zero mode waveguides to isolate the individual enzymes and watch as they add base after fluorescent base using a highly multiplexed confocal fluorescence microscope built for the purpose, says founder and chief technical officer Stephen Turner.

"The differentiation with Pacific Biosciences [compared to other polymerase-based strategies] is that we don't stop the action of the polymerase," says Turner. "We let it go at its native speed, and we watch in real time and simply record the activity of the polymerase."

The technology potentially could produce reads far longer than 454's. The company announced at AGBT the sequencing of the *E. coli* genome with reads averaging 586 bases and as long as 2,805; a commercial launch is planned for 2010.

ZS Genetics literally reads sequences using transmission electron microscopy (EM). A DNA-copying step is used to substitute the normal bases of DNA with variants containing proton-rich atoms (such as 5-bromo-dCTP), which then are visualized directly in the EM. Though still in development, says William Glover III, company president and vice president of research and development, "We expect to have, when we launch, in the range of five-to-8,000-base pair reads or better."

Whatever the future, next-generation sequencing has entered the scientific zeitgeist. It has its own X-Prize challenge, and garnered two spots in *The Scientist* magazine's 2008 top 10 technologies list. Knome is actively selling consumer genomics at $99,500 apiece, while Complete Genomics talks of sequencing one million human genomes in the next five years. Church's Personal Genome Project has signed up some 10,000 volunteers to have their genomes sequenced and released into the public domain.

Meanwhile, technology development continues on the *next* next-gen, based on such ideas as nanopores and fluorescence resonance energy transfer between nucleotide and polymerase.

That's not to say Sanger chemistry is disappearing; some applications simply don't need next-gen throughput. And with their on-the-fly library generation, next-gen strategies don't support clone reanalysis. Though the need likely exists for multiple technologies (say, long versus short reads) what remains to be seen is whether the market exists for so many competing technologies. One thing is certain: at the current pace of development, 2009 should be a very interesting year.

Jeffrey M. Perkel is a freelance science writer based in Pocatello, Idaho.

DOI: 10.1126/science.opms.p0900033

Featured Participants

Applied Biosystems (Life Technologies)
www.appliedbiosystems.com

Broad Institute of Harvard and MIT
www.broad.mit.edu

Complete Genomics
www.completegenomics.com

Dover Systems
www.polonator.org

Harvard University
www.harvard.edu

Helicos Biosciences
www.helicosbio.com

Illumina
www.illumina.com

Invitrogen (Life Technologies)
www.invitrogen.com

Pacific Biosciences
www.pacificbiosciences.com

Roche Applied Science
www.roche-applied-science.com

ZS Genetics
www.zsgenetics.com