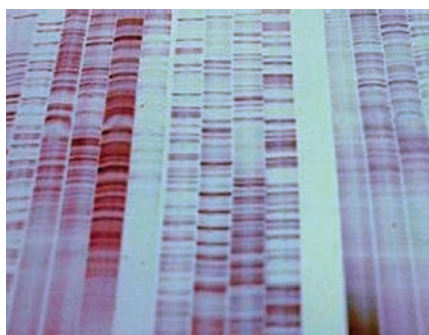


The year of sequencing

In 2007, the next-generation sequencing technologies have come into their own with an impressive array of successful applications. Kelly Rae Chi reports.

In the toxicology building of North Carolina State University in Raleigh, Nigel Deighton, head of a small genome research facility, and a few others unpack the facility's first next-generation sequencing machine, a 454 GS FLX, on loan from Roche Diagnostics for three months. They train for a few days, nebulize a colleague's bacterial DNA and PCR-amplify "the living daylights out of it," Deighton recalls. They load the bead-bound PCR products onto a plate with holes that are not visible to the naked eye, pop the plate into the machine and close the drawbridge-like door. Then they go out for a beer.

The next day, Deighton scrolls through graphs from the newly sequenced bacterial genomes. On its first test run, the GS FLX gave sixfold coverage of each genome, with read lengths of around 250 bases. "That's not bad, for the first go," Deighton says.



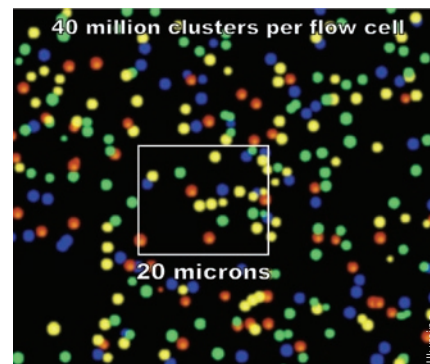
Sanger sequencing becomes the 'old' generation.

In the last few years, the next-generation sequencing platforms have been developed and sprung up in large sequencing facilities. But in 2007, even for smaller facilities such as Deighton's, sequencing started seeming more affordable, says George Church, who has been developing a related technique at Harvard Medical School.

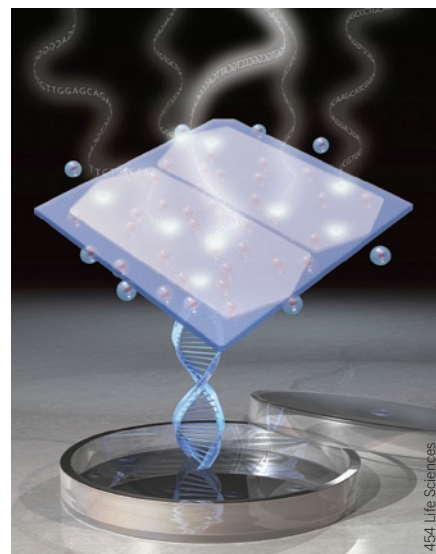
The 454 Life Sciences instrument, released in 2005, was the first platform on the market (see 'Rothberg's recipe for success'). The company was purchased by Roche Diagnostics in 2007. Solexa, Inc., later acquired by Illumina, released its platform, which combines sequencing-by-synthesis chemistry and cluster technology, in June 2006 (see 'Creating the Genome Analyzer'). In October 2007, Applied Biosystems announced availability of the ABI SOLiD system, the only commercial system that uses a ligation-based chemistry (see 'Applied Biosystems' approach to sequencing'). These platforms are being applied in diverse ways, as evidenced at last year's sequencing conferences, Church says. Chad Nusbaum, co-director of genome sequencing and analysis at the Broad Institute, says researchers are increasingly using the sequencing as a general-purpose tool to tackle what he calls low-lying fruit.

In 2007, researchers performed whole-genome human sequencing using old and new platforms. Researchers at Baylor College of Medicine and 454 Life Sciences sequenced James Watson's genome in two months, for about \$1 million. Two other personal genomes were sequenced: Craig Venter's, at the Institute he founded, and that of a Chinese individual, at the Beijing Genomics Institute. The J. Craig Venter Institute used Sanger technology for sequencing Venter's DNA, which cost an estimated \$70 million and took several years, but is now focusing its efforts on sequencing more human genomes, a task that will rely on the next-generation platforms and reference sequences captured using Sanger technology, says Yu-Hui Rogers, scientific director of the J. Craig Venter Institute's Joint Technology Center.

Researchers at Yale University, in collaboration with 454 Life Sciences, combined the 454 sequencing technology with paired-end mapping to detect structural variation in the human genome, for an October 2007 study published in *Science*. "The technology allowed us to get some things we could never get before," says



Color dots replace bands as a sequence is read with reversible terminators on a Solexa (Illumina) sequencer.



The concept of next-generation sequencing according to 454: fiberoptic flow cell and emulsion droplets.

ROTHBERG'S RECIPE FOR SUCCESS

Since 1986, Jonathan Rothberg says he has wanted to improve ease and cost of gene sequencing. In 1999, when his son was born, Rothberg spent two weeks on paternity leave and made arguably his biggest leaps toward this goal. From the sequencing ideas he generated in this short time, and with the help of colleagues whose work he admired, he invented a technology that, in 2007, 454 Life Sciences and others used to sequence Watson's DNA and solve a mystery of what was killing honeybees.

Inspired by the idea that cramming more features onto an integrated circuit would yield more power at the same cost—an idea first introduced by Gordon Moore in 1965—Rothberg wanted to make sequencing massively parallel by putting it onto a microreactor chip. For more than a decade, researchers had tried without success to move Sanger sequencing to a chip, but this technique was not scalable, he says.

At the time, he was familiar with pyrosequencing, a technique he believed could be scalable. The technique, invented by Mathias Uhlén, Mostafa Ronaghi and Pål Nyrén in the mid-1990s, was being used to sequence single-nucleotide polymorphisms, only one to seven bases at a time. "I was excited about [pyrosequencing] for its simplicity," he says. It required no moving parts and no separation like Sanger sequencing did. And as a chemical engineer, he thought he could move the sequencing to a small reactor and do millions of reactions in parallel.

So during his two weeks off, Rothberg designed a chip and a flow cell. For the chip, he sliced up fiberoptic bundles and turned each cross-section into a slide, with millions of wells, each of which could contain a pyrosequencing reaction that was modified to work on a solid support. He needed a way to flow nucleotides over the open reactors, so he surfed the net and found a design he could mimic that was used in neuroscience laboratories to irrigate living cells with liquid nutrients.

Making hardware was only half the battle for Rothberg. There was still the lengthy process of sample preparation. He thought he could eliminate "bacterial cloning and picking" by shattering DNA into random pieces and finding a way to clone each piece. Rothberg wanted to put the DNA into water-oil emulsions. Andrew Griffiths and Dan Tawfik published an emulsion experiment the year before, in *Nature Biotechnology*, that helped Rothberg develop the emulsion. (Rothberg says he was so fond of this paper that he carried it in hand everywhere he went for a while.) The idea would eliminate the need for test tubes because DNA pieces would attach to beads, each of which would end up in a separate droplet of water.

He wasn't sure whether these ideas would work, so he ran them by Joel Bader, a close friend and colleague, now at Johns Hopkins University. During the two weeks, Rothberg and Bader talked about limiting dilutions, diffusion and reaction rates. Bader did the math to verify that the chemistry would work. He helped Rothberg write up the sequencing patents. Rothberg founded 454 Life Sciences.

Rothberg credits John Leamon, Jan Berka and the molecular biology team at 454 Life Sciences with the idea to thermostabilize the emulsions using surfactants from the explosives industry. These conditions would allow them to do PCRs.

Rothberg says sequencing Watson's DNA and sequencing a virus that killed honeybees last year have given him great closure. In November 2007, 454 Life Sciences announced its 100th peer-reviewed study. A few years ago, Rothberg says, 454's next-generation sequencer was an invention, but in 2007, it really felt like a method.



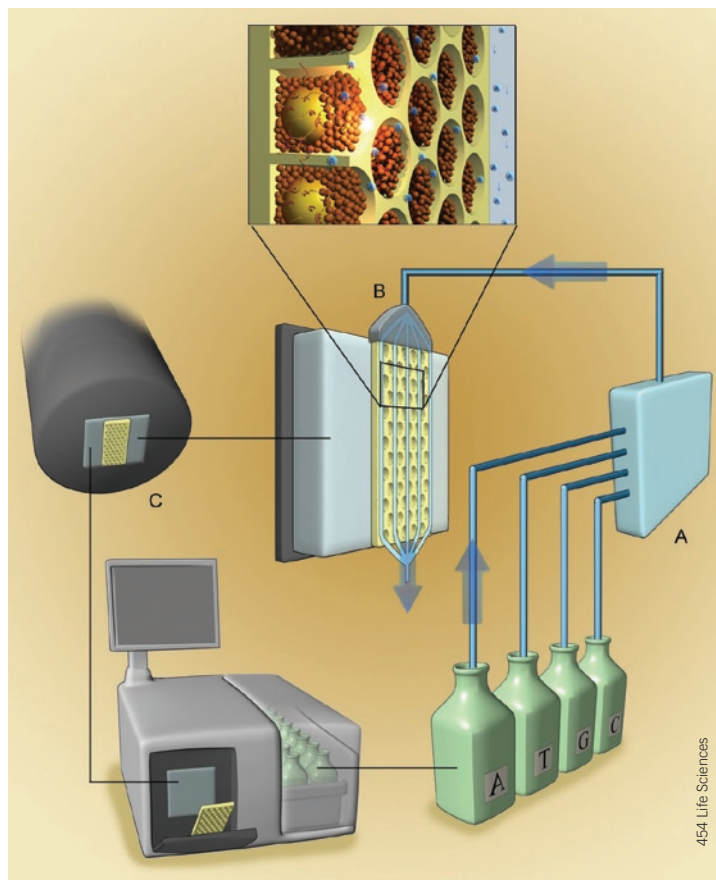
Jonathan Rothberg invented the 454 technology.

the study's lead author Michael Snyder, referring to inversions and breakpoints that would be missed using comparative genome hybridization and fosmid paired-end sequencing. John West, general manager of Illumina's DNA sequencing business unit and Kevin McKernan, an inventor of the Applied Biosystem's SOLiD System technology, both say their companies are playing with paired-end sequencing to detect structural variations of different sizes.

According to Nusbaum, the platforms are also clearly a boon for sequencing shorter pieces of material, such as DNA fragments that bind to histones and transcription factors. Taking advantage of Solexa's shorter read lengths, a 2007 *Cell* report of a study led by Keji Zhao at the US National Heart, Lung and Blood Institute analyzed histone modifications by chromatin immunoprecipitation (ChIP) in one of the first published studies using Solexa technology. Zhao says that using the traditional ChIP coupled with serial analysis of gene expression (SAGE) method (which uses Sanger sequencing of concatenated DNA tags) would have cost close to \$100,000 per histone modification, a much higher price tag than the \$2,000 it cost them with the Solexa machine. With the savings, they have been able to sequence more. "Without this machine, we can't do anything at the scale we're working on," Zhao says. Elaine Mardis, co-director of the Genome Sequencing Center at the University of Washington in St. Louis, says the new platforms are ideal for ChIP studies like Zhao's because researchers can sequence and characterize the bound fragments instead of inferring the sequences using hybridization arrays.

Metagenomics will also benefit from the new platforms, Mardis says, with the US National Institutes of Health's Human Microbiome Project as one example. In September 2007, researchers at Columbia University, led by W. Ian Lipkin, and 454 Life Sciences published a study in *Science* that used metagenomics to identify a virus implicated in the death of billions of honeybees. "Metagenomics was one of those things that was heinously expensive to do in the past and troublesome as well because it wasn't easy," Mardis says. Now, with the next-generation platforms, it is very straightforward to do these experiments and extract meaning from them, she adds.

KRC

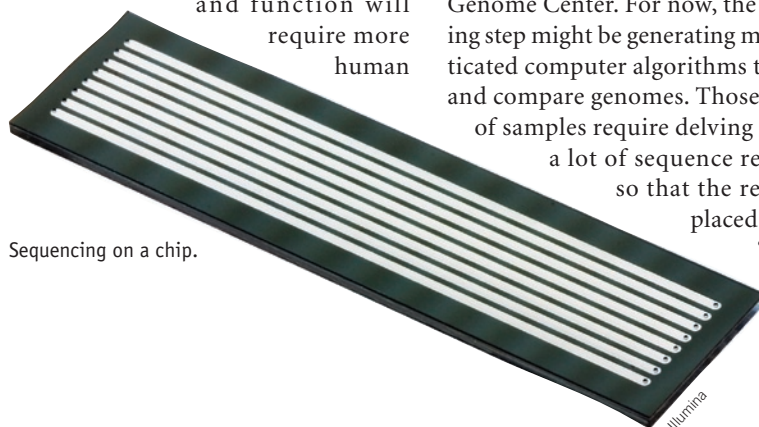


The inner workings of the 454 Life Sciences sequencer.

Church predicts that, in 2008, people will use the platforms to see a clearer connection between sequence and function. Already, association studies linking genetic sequence and function are getting significant results at a much higher rate—from about one a year in previous years to 60 in 2007. “That’s a pretty significant increase, but very few of those have actually been hunted down to the causative allele,” Church says. Seeing the connections between sequence and function will require more human

sequences and phenotypic data. Len Pennacchio, a geneticist from the US Department of Energy’s Joint Genome Institute, says: “At this point, Watson’s DNA is a parts list, but it hasn’t told us what makes Jim Jim.”

Though the platforms are not producing \$1,000 human genomes yet—a research initiative the National Institutes of Health tried to spark in a 2004 request for proposals—they are on their way, says Richard Myers, director of the Stanford Human Genome Center. For now, the next limiting step might be generating more sophisticated computer algorithms to assemble and compare genomes. Those thousands of samples require delving deep to get a lot of sequence redundancy, so that the reads can be placed accurately. “It is a hard



Sequencing on a chip.

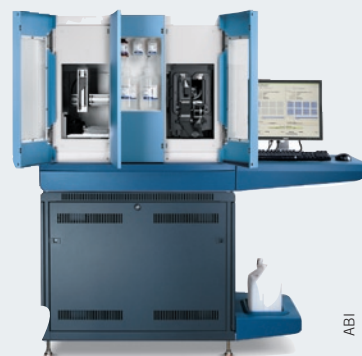
APPLIED BIOSYSTEMS’ APPROACH TO SEQUENCING

Applied Biosystems’ SOLiD system, which launched in October 2007, has had a successful start, says Kevin McKernan, an inventor of the technology. The company has been able to double DNA sequence throughput every quarter for the last year.

Applied Biosystems adopted a ligation approach that was developed by Agencourt Personal Genomics, a sequencing company that Applied Biosystems acquired in 2006. While working on a polymerase-based approach in 2003, McKernan and others had trouble incorporating the fluorescent tags efficiently, prompting them to move toward the ligation approach.

The groundwork for a ligase approach was already there, McKernan says, with work dating back to the 1980s that used ligases for genotyping in a sequence-specific manner. What was not there was how to go about it in a serial way, acquiring information about DNA sequences one base after another. “That took a lot of white-boarding and sketch-work,” he recalls.

The resulting ligation system used an error correction system that they could not envision with polymerases, he says. In the process of recording these bases, every base is read with a different probe. “What we can do then is see how many times the probes corroborate the same data. That gives us a tremendous amount of error correction that you can’t get with polymerase,” he says. *KRC*



Applied Biosystems’ SOLiD instrument.

CREATING THE GENOME ANALYZER

When John West started as CEO of Solexa Ltd. in August of 2004, the longest stretch of DNA that the company could sequence was only six bases long. "That was a little bit intimidating," recalls West, now the vice president and general manager of Illumina's DNA sequencing business unit following the acquisition of Solexa Ltd. by Illumina. "The problem was we never had a commercial platform to be able to sequence it on."

The next-generation sequencing platform they developed and commercialized by June 2006, West says, has been a huge success. Since the commercial release of the platform they have sold 100 instruments and increased the scale of what they have tackled using the technology—from a 5,300-base-pair viral genome to a 150-million-base-pair human X chromosome. But the machine was a challenge to develop. The developers had to bring together key elements of chemistry, people and technology to make it work.

By the time Solexa Ltd. announced its plans to merge with Lynx Therapeutics in August 2004, a lot of the core sequencing-by-synthesis chemistry and molecular biology had already been done, West says. The early chemistry work, done by Solexa Ltd. in the United Kingdom, led to the creation of a reversible terminator nucleotide and a polymerase that would incorporate it. Solexa Ltd. and Lynx Therapeutics had bought cluster technology for solid phase amplification together from the Swiss company Manteia SA, and did some instrumentation design.

Still, the company was in a state of flux in late 2004 and needed to figure out how to combine the chemistry and technology into a complete system. The researchers needed to meet and brainstorm, operating on an eight-hour time difference between the two branches, one in the UK and one in the US. Ligation chemistry developed by Lynx Therapeutics was an option. There were differing opinions about which chemistry, ligation or polymerase would work best, but ultimately they took a gamble on the Solexa polymerase because it could complete reactions faster. "It was risky at the time because it wasn't all proven," West remembers. "We basically put all our eggs in one basket."

And it worked. It became easier, with the cluster technology and polymerase chemistry, to increase the sequence lengths. By February 2005, they increased the read length to 25 bases. They sequenced the 5,300-base-pair virus PHIX174 genome, the same genome

that Sanger first sequenced. In October 2005, they sequenced an 180,000-base-pair bacterial artificial chromosome.

Illumina acquired Solexa, Inc. in January of last year, and commercial sales increased, West says. In 2007, the company completed sequencing the human X chromosome and has since moved on to sequencing the human genome using paired-end sequencing. "I think [2007] has been a great year for us," West says.

KRC



Illumina's (Solexa) Genome Analyzer.

problem, but people are starting to apply [the technology] that way," he says.

In the meantime, Deighton will try to apply the 454 Life Sciences technology to the existing research at North Carolina State University. They have assembled the bacterial genomes and are going back to

get better coverage, about 20-fold, so that they can fill in the gaps. About 15 researchers have signed up to use the machine to sequence everything from trees to butterflies. By the end of the three-month loan period, Deighton will do a cost analysis to show his colleagues that the machine's

benefit outweighs its \$500,000 price tag. "I'm convinced that it'll be worth keeping," he says.

Kelly Rae Chi is a freelance science journalist based in Chapel Hill, North Carolina, USA (kellyraechi@gmail.com).