

are in fact quite different from one another and at various stages of maturity. “The principle that we use when applying these new technologies is that there is a lot of expensive sequencing that we do with Applied Biosystem’s 3730xl system and anything that we can move over to the new technologies, as long as it is effective, is bound to be cheaper.” The systems available now from Roche, Illumina and Applied Biosystems do seem to be effective, as the Broad Institute and other organizations are using them for various sequencing-based applications.

### Assembling the future

By the end of last year, 454 Life Sciences, which was founded in 2000 and was recently acquired by Roche, had more than 60 of its sequencing systems placed around the world. “Our technology is in all major US genome centres and some of the international centres,” says Michael Egholm, vice-president of research and development at 454 Life Sciences.

The technology developed by 454 Life Sciences is based on two fundamental principles: emulsion PCR and pyrosequencing. Emulsion PCR side-steps the conventional process of bacterial cloning by attaching fragments of DNA 300 to 500 base pairs long to beads *in vitro*, then amplifying them with PCR to make millions of identical copies. Pyrosequencing allows for a massive parallel reaction format done in 1.6 million wells on a PicoTiterPlate. “Right now, day in and day out, we can perform 400,000 reads of

250 bases each with an accuracy of 99.5% or better,” says Egholm. Although the 454 Life Sciences system is not as accurate as conventional Sanger sequencing, Egholm notes that it is an order of magnitude more productive (see ‘Truth and accuracy’).

This upgraded Genome Sequencer FLX System allows more sequencing cycles and therefore longer reads than the previous Genome Sequencer 20 System. Longer reads help in whole-genome sequencing and assembly applications. “We believe that shortly there will be many more *de novo* assembled genomes due to our technology,” says Egholm. He notes that the genomes of several microorganisms have been assembled from scratch by use of 454 sequencing, and the technology has also been used to supplement Sanger sequencing on a few projects involving larger genomes.

At the Broad Institute, where researchers use two FLX systems and one Genome Sequencer 20, Nusbaum appreciates the ease of the 454 sequencing process. “It is nicer than Sanger sequencing because it is a faster and simpler process.” He points out that at the Broad Institute, sequencing a bacterium can take a month with Sanger methodology, whereas with 454 technology it can be done in a week and without the high degree of clone tracking associated with Sanger sequencing.

Still, for *de novo* sequencing and to assemble larger genomes, such as those of mammals, longer paired reads — that is, two reads that are a known distance apart — will



454 LIFE SCIENCES

The Genome Sequencer FLX, developed jointly by 454 Life Sciences and Roche Applied Sciences, is based on 454 sequencing technology.

be necessary — an issue that Roche and 454 are trying to address. “Whether [454 sequencing] will work with a mammalian genome is a good question, and it is a little way off,” says Nusbaum. But he optimistically notes that 454 Life Sciences has exceeded his expectations in surmounting several other technical hurdles. Egholm, however, is much

## TRUTH AND ACCURACY

Mitch Sogin, director of the Josephine Bay Paul Center for Comparative Molecular Biology and Evolution at Woods Hole Marine Biological Laboratory in Massachusetts, performs environmental sampling of nucleic-acid sequences. “Every sequence has the potential to tell us an important story,” says Sogin, so highly accurate analysis techniques are needed.

But when Sogin’s lab switched over from traditional Sanger-based sequencing to the next-generation sequencing system of 454 Life Sciences in Branford, Connecticut, to study these environmental samples, something strange happened. “The diversity was between ten- and a hundred-fold more divergent than we expected,” recalls Sogin.

So Sogin and his colleagues needed to determine whether the unexpected findings represented



Mitch Sogin tested the accuracy of 454 sequencing.

true biological diversity, or just errors caused by the new sequencing technology. “We had to explore just how good the sequencing technology actually was,” he says.

Sogin and his colleagues did a straightforward experiment in which they resequenced more than 50 templates and cloned sequences on a 454 Life Sciences Genome Sequencer 20 that they had sequenced previously with the Sanger methodology. The work showed that the 454 system was 98% accurate if no culling was used to remove bad bases or reads<sup>6</sup>.

However, by using a very simple set of rules, which caused somewhere between 10% and 20% of the data to be discarded, the accuracy could be pushed up to 99.75%. And discarding up to 20% of the data for this level of accuracy is a trade-off that is fine with Sogin because the latest 454 system — the Genome Sequencer FLX — can produce up to 400,000 reads per run.

Others agree that for some applications the large amount of data generated by the next-

generation sequencing systems could trump the accuracy produced through Sanger sequencing. “For applications such as CHIP-sequencing you can use the 454 or Solexa 1G data even though they have lower base accuracy because you do not need it. What you need is the volume for the experiment,” says Chad Nusbaum of the Broad Institute in Cambridge, Massachusetts.

Sogin now thinks that traditional sequencing methods had been underestimating the biological diversity of the environmental nucleic-acid samples. “Turns out that the diversity is coming from low-abundance nucleic-acid populations that you are not likely to encounter if you sequence only a few hundred molecules. You see these low-abundance molecules only if you sequence many tens of thousands of molecules,” he says.

N.B.