TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

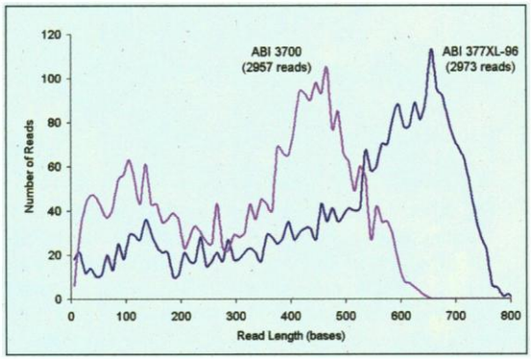

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 μm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

We sequenced 31 96-well plates of DNA with the ABI 3700. The DNA sequence data were extracted from the ABI 3700 Oracle database and subsequently processed by the base-calling program "phred"(*5*). This program calculates a base sequence for the traces and assigns a quality value, Q, to each base, where the probability of assigning an erroneous base is $P_{error} = 10^{(-Q/10)}$. We evaluated Q against finished sequence and found that this measure underestimates the true accuracy for ABI 3700 data; at a 1.0% error rate, the value phred assigns to Q is 16 instead of 20.

With Q properly calibrated, read length is computed by totaling the number of bases with a Q value greater than or equal to 20. Figure 1 shows a histogram of read length for the 31 96-well plates sequenced with the ABI 3700 and 31 different 96-well plates sequenced with the ABI 377XL-96, both using ABI Big Dye chemistry. Note that the mode of the histogram for the ABI 3700 is 465 bases, whereas the ABI 377XL-96 mode is 655 bases. Furthermore, there are a total of 954,191 $Q \geq 20$ bases from the ABI 3700 data set with three failed lanes and 1,448,468 $Q \geq 20$ bases from the ABI 377XL-96 data set with 17 failed lanes.

As a further test, a cloned 100 kbp section of human chromosome 22 was selected for shotgun sequencing, and 15 96-well plates of DNA were sequenced with the ABI 3700 and another 15 96-well plates sequenced with slab gel sequencers. All of the sequence data were assembled and corrected manually to an accuracy of less than one error in 10,000 bp to give a finished sequence of this region of the genome. Sequence assembly and editing were equivalent for both types of data. The percentage of edited bases was 0.60% for the assembled ABI 3700 Big Dye Terminator reads, 0.54% for the ABI 377XL Big Dye Terminator reads, and 2.15% for the ABI 377XL ET Primer reads. Thus, the effort required to edit the reads was nearly the same for the Big Dye Terminator reads run on either machine, and was significantly more for the ET Primer reads.

A comparison of two regions of sequence from these clones produced by the ABI 3700 and by the ABI 377XL (with other types of chemistry) reveals the difficulties that the ABI 3700 has with such regions (Fig. 2). Reads 1, 2, and 3 are all from the ABI 3700, as noted by the read name extension of ".ab1," and show a situation where a base is missed, denoted by the letter N. The second difficult region, a run of 24 thymidine bases, is illustrated in Fig. 2 as reads 4, 5, and 6.

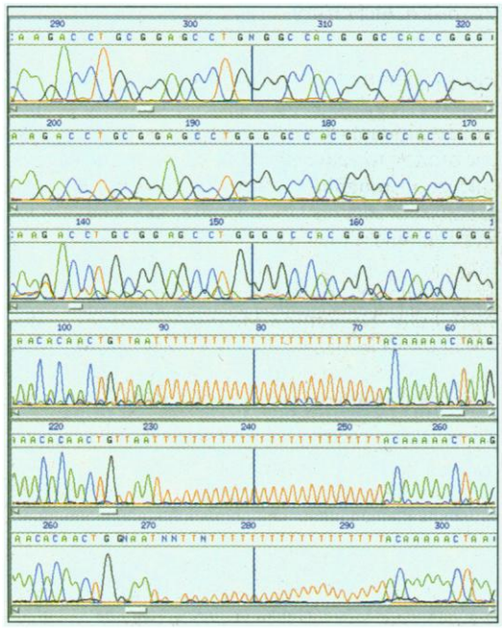

Fig. 2. Examples of sequence data from the ABI 3700 machine. Reads 1 through 6 are shown from top to bottom. The vertical bar indicates the undetermined base in read 1. Reads 1 and 3 are both displayed in the order that they emerged from the capillaries, but read 2 is reverse complemented (A swapped with T and G swapped with C and the order reversed), because its subclone contained the other complementary strand of DNA. This means that the four G's were read as four C's on the sequencer, which gives an independent sampling of the sequence. In read 1, the N should be a G but was missed because it is weak in relation to the surrounding G's, as can be seen in read 3. Base calling succeeds in read 3 because the resolution is better closer to the start of a read; note that the N is located at base position 304 in read 1 and is correctly identified as a G at position 152 in read 3. Reads 4, 5, and 6 report a run of 24 T bases. Read 6 shows that the signal strength from the T-labeled bases drops to a level where the base caller chooses N because it cannot find an observed peak. In read 5, the base caller correctly reads this region, although the weak regions in this read correspond to the missing bases in read 6. Read 4 from a slab gel machine with ABI Big Dye Terminator chemistry supports read 5.

The ABI 3700 is extremely user friendly in its operation. It is still a new machine, however, and we anticipate that improvements will be made, particularly in its software. It has the potential to be an extremely high-throughput machine and as such needs larger reagent reservoirs to take full advantage of its potential capabilities. It is too early to comment on capillary lifetime, although this has a specification of 300 sequencing runs (~40 days at 8 runs/day); similarly, we have so far had no opportunity to evaluate cleaning schedules and maintenance. Another operational milestone for this machine will be when it passes from our Research and Development department into routine production use. From our experience with other hardware and software transitions to production, we anticipate that new issues will arise at this time.

At present, the ABI 3700 can only sequence samples prepared with one commercially available chemistry. Our experience with sequencing the genomes of a number of organisms shows that it is often necessary to use other sequencing chemistries to confirm ambiguous sequence data. This facility is available in all slab gel sequencers and would make the ABI 3700 more flexible in its use. Sequencing reaction volumes required for ABI 3700 sequencer are identical to those for slab gel machines. However, the ABI 3700 sequencer only loads one-tenth of the prepared sample volume into its capillary injection wells, leaving nine-tenths to be discarded. The software also does not allow variation of many of its operating parameters. It would be advantageous to have access to more settings so that we could optimize the operating conditions.

Our results show that read length obtained from the ABI 3700 sequencer is currently shorter than that obtained from the same type of samples sequenced on slab gel machines. This, coupled with the fact that the ABI 377XL-96 (at roughly half the price of the ABI 3700) has a throughput of three plates per day as opposed to the current maximum throughput of six plates per day for the ABI 3700, means that there is no immediate gain in throughput in terms of capital investment. Capillary sequencers already established in the marketplace, however, have attained read lengths comparable to those from slab gel machines; the Molecular Dynamics MegaBACE 1000 can read 600 to 700 bases and Perkin-Elmer's single capillary sequencer, the ABI 310, produces read lengths of around 600 bp. Optimization of the ABI 3700 will likely lead to increased read lengths, and we expect that this machine will become a widely used instrument in large-scale sequencing operations.

References and Notes

1. See 12 January 1999 press release at www.celera.com
2. J. L. Weber and E. W. Meyers, *Genome Res.* **7**, 401 (1997).
3. http://www.sanger.ac.uk/Info/Statistics/
4. H. Kambara and S. Takahashi, *Nature* **361**, 565 (1993).
5. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998); B. Ewing and P. Green, *ibid.*, p. 186.