Sequencing Multimegabase-Template DNA with BigDye Terminator Chemistry

Cheryl R. Heiner,[1] Kathryn L. Hunkapiller,[1] Shiaw-Min Chen,[2] John I. Glass,[3] and Ellson Y. Chen[1,4]

[1]Advanced Center for Genetic Technology and [2]Genetic Analysis Department, PE-Applied Biosystems, Foster City, California 94404 USA; [3]Department of Microbiology, University of Alabama at Birmingham, Birmingham, Alabama 35294 USA

Using the recently introduced BigDye™ terminators, large-template DNA can be directly sequenced with custom primers on automated instruments. Cycle sequencing conditions are presented to sequence DNA samples isolated from a number of microbial genomes including 750-kb *Ureaplasma urealyticum,* 1.2-Mb *Mycoplasma fermentans,* 2.3-Mb *Streptococcus pneumoniae,* and 4.6-Mb *Escherichia coli.* Average read lengths of >700 bp from unique primer annealing sites are often sufficient to fill final gaps in microbial genome sequencing projects without additional manipulations of template DNA. The technique can also be applied to sequence-targeted regions, thereby bypassing tedious subcloning steps.

In microbial genome or large-insert clone sequencing projects that use the predominant random subclone sequencing strategy, progress tends to decrease dramatically at late stages as one confronts gaps. At these points, DNA is under-represented or unstable in subclones (E.Y. Chen et al. 1996; Chissoe et al. 1997). Further sequencing with additional random subclones is then inefficient at best, and one must frequently employ alternative cloning systems or additional methods like long-range PCR to recover missing DNA (C.N. Chen et al. 1996). The variability of performance of these methods and the necessity for custom-tailored work tend to hamper the late stages of sequencing efforts. In contrast, if one can sequence directly from genomic DNA (or large-insert clones such as BACs or PACs) with walking primers, cumbersome work to fill gaps could be completed in a much shorter time.

As an example, in a recent project to sequence the 750-kb genome of *Ureaplasma urealyticum* (J. Glass, in prep.) assemblage of ~13,000 sequence reads and combinatorial PCR reactions to join contigs left two gaps. No λ pUC, or M13 subclones were recovered that spanned the gaps, nor were PCR products derived with any of several sets of flanking primers. The difficulty of cloning these segments is probably attributable to repeated sequences in and near the two gaps, but the high sensitivity of the recently introduced BigDye terminator (Rosenblum et al. 1997) permitted direct sequencing of the gap regions on genomic *U. urealyticum* DNA templates. Using the conditions described in this report, two gaps of 259 and 121 bp were sequenced from both strands with walking primers to complete the project of 751,723 bp.

Direct sequencing was further tested for larger templates, and good results were reproducibly obtained with 1.2-Mb *Mycoplasma fermentans,* 2.3-Mb *Streptococcus pneumoniae,* and 4.6-Mb *Escherichia coli* genomic DNA (see example in Fig. 1). In addition, several difficult gaps in sequencing projects with BAC clones, ranging in size from 140 to 250 kb, have also been filled in this manner. Essentially the method is applicable whenever 2–3 μg of high-quality large-template DNA is available.

RESULTS AND DISCUSSION

Figure 1 shows an example of the results from these experiments. Although the signal intensity tends to be low—only ~10%–20% compared to the data from regular M13 or pUC templates—base-calling quality remains high, because the baseline noise is sharply reduced by the increased brightness and improved spectral resolution of the BigDye terminators (Rosenblum et al. 1997). Lower signal strength is expected considering the molarity of microbial template DNA, which is several hundred to a thousand times less than that of the regular plasmid templates. Higher level of primers (2×–5×) and greater number of cycles (from 45 to 60, more cycles for

[4]Corresponding author.
E-MAIL cheney@perkin-elmer.com; FAX (650) 638-6177.

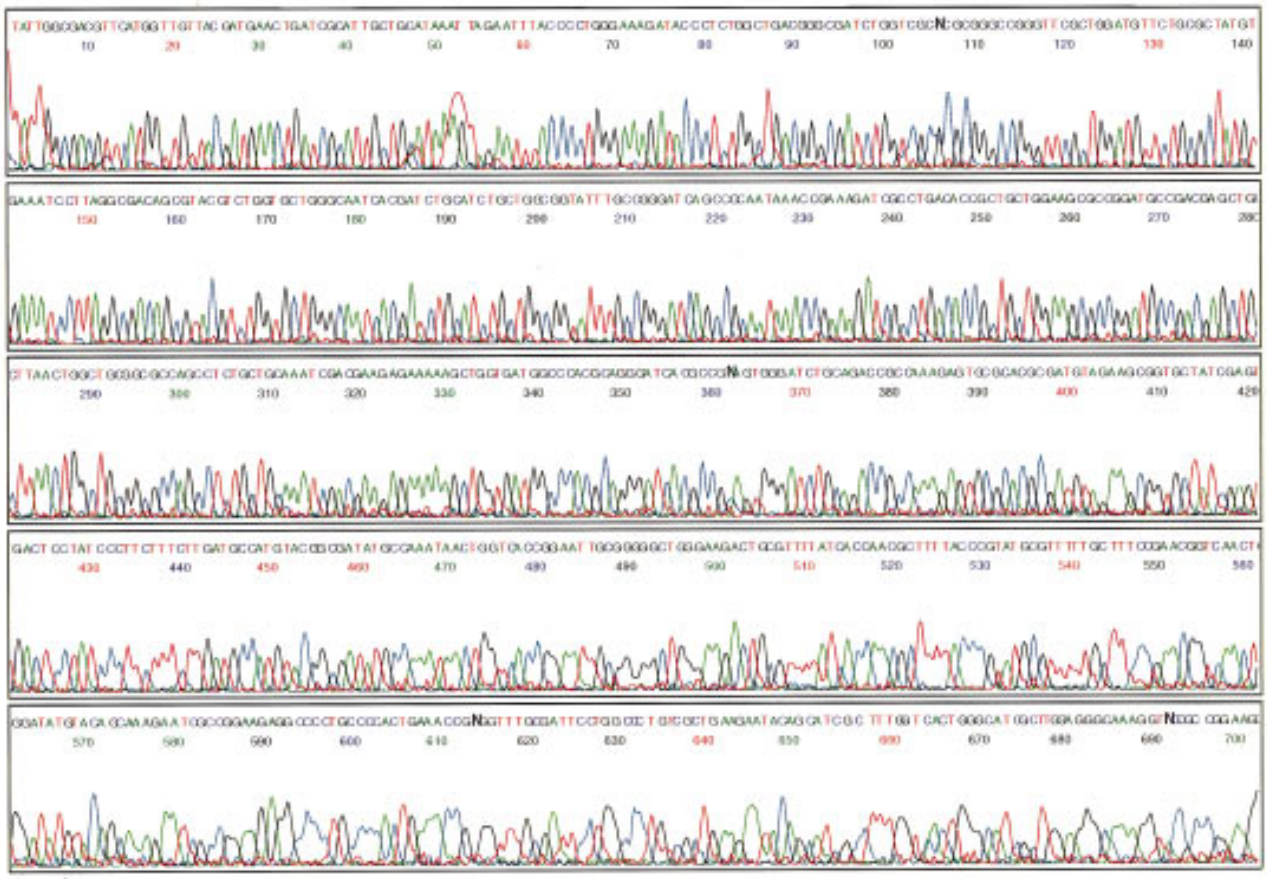

Figure 1 Sequencing of *E. coli* K12 strain genomic DNA with BigDye terminators. Approximately 3 µg of *E. coli* DNA was sequenced with an *apaG* gene primer (5′-GTTCCCACACTCATTCATTA) using the conditions described in the text.

larger templates) as described in Methods helped to boost the signal intensities. The addition of cycles (up to 99) has been found to increase the signal strength and decrease the readable range (see Table 1). Accurate quantitation of template DNA to within 2–4 µg is essential. Too much template (>5 µg) produced much lower quality results (see Fig. 2 for an example), whereas too little DNA also gave rise to weak signal and low-quality results (data not shown).

There are several other factors that are also important for optimal results. First, template DNA should be free of salts, detergents, proteins, and cell debris that might interfere with the primer annealing. Second, as may be expected, having a good primer is critical. Successful primers have been typical 21–25-mers with appropriate GC contents from a unique site. Only 1 of the 10 primers tested failed to generate a useful sequence because that primer annealed at two different locations on the *S. pneumoniae* genome. Third, special care should be taken in the elimination of excess dye terminators before loading samples on gels (note that carryover of dyes can be seen in Fig. 1 in the region of bases 45–55; apparently the system is very sensitive to residual dye when signals are so low). Fourth, to get high-quality, low-signal data, it is important to have a well-tuned sequencing instrument equipped with a good multicomponent matrix and base-calling software capable of analyzing data with weak signals. Here we used version 3.0.1b3 software, which requires no minimal signals.

There appears to be no correlation between the quality of the sequence data and the size of the template, as shown in Table 2, in which nine different sequences obtained from bacterial genomic templates are compared with corresponding GenBank sequences that were presumably derived from subcloned templates. Using the unedited sequence files of up to 600 bases beyond the first legible base, genome-derived sequences have an average fidelity of 98.2% (1.1% *N* and 0.7% discordant base-calls). Af-

Table 1. Sequence Quality and Signal Strength as a Function of the Number of Cycles

No. of cycles	Percent accuracy and no. of ambiguous bases in 100-base intervals			Useful data range (bases)	Relative signal strength			
	1–100	101–700	701–800		G	A	T	C
35	98 (1)[a]	100 (0)	99 (1)	900	361	269	260	357
45	99 (0)[a]	100 (0)	98 (1)	850	477	311	229	403
75	100 (0)	100 (0)	99 (2)	850	814	568	348	670
99	100 (0)	100 (0)	96 (4)	780	1123	778	444	918

An amount of 50 ng of pGEM was sequenced with BigDye terminators with 35, 45, 75, or 99 cycles. The accuracy of each sequence is listed as percentage agreement with the known sequence. The number following the percentage value is the number of ambiguous bases (*N*) for each interval. The usable data range is the lengths of sequence that are accurate after minor human editing of the computer generated base calls.

[a]These *N*s and errors are due to incomplete removal of residual dye terminators, obscuring sequencing data where signals are somewhat weak.

ter minimal manual editing of the sequences, the average useable read length was 712 ± 18 bases. It is interesting that the signal strength remains relatively constant among these nine genomic templates ranging from 0.75 to 4.6 Mb. This suggests that the present protocol, after some modifications, may be applicable to even larger templates. The method, nevertheless, does require 2–3 µg of high-

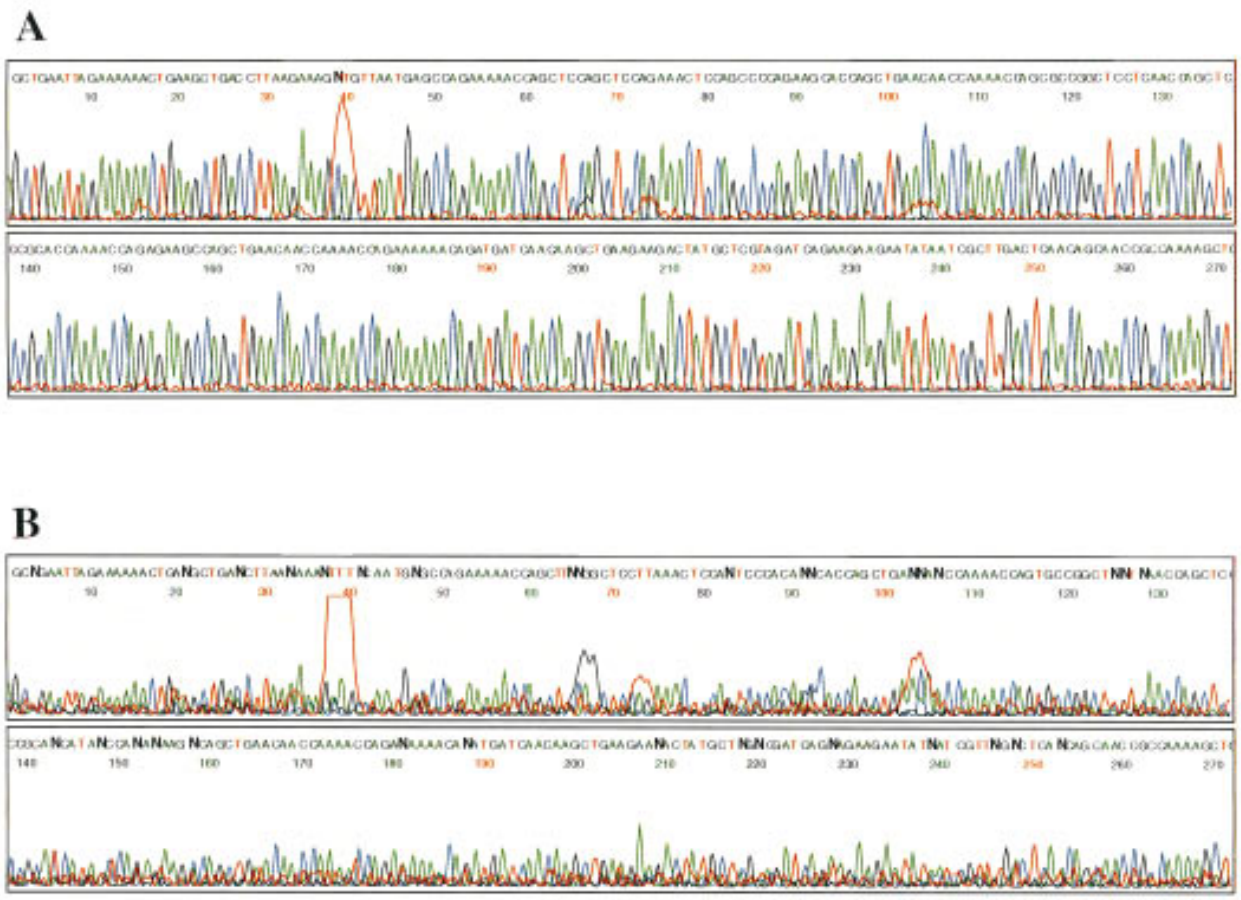

Figure 2 Sequencing of *S. pneumoniae* genomic DNA with BigDye terminators using either 2.5 µg (*A*) or 5.0 µg (*B*) of template and a SKH1 primer (23-mer 5′-AACAATAATGTAGAAGACTACTT). All other conditions are as described in the text.

Table 2. Sequence Quality as a Function of Read Lengths

Organisms sequenced	Percent fidelity of sequence and no. of ambiguous bases within 100-base intervals from first readable base							Useful data range (bases)
	1–100	101–200	201–300	301–400	401–500	501–600	601–700	
E. coli apaG gene	99 (0)	99 (1)	100 (0)	99 (1)	100 (0)	99 (0)	95 (2)	700
E. coli HtpG gene	99 (1)	99 (1)	99 (0)	98 (2)	95 (5)	87 (5)	64 (14)	560
E. coli ldhA gene	97 (2)	100 (0)	100 (0)	98 (2)	98 (2)	97 (3)	91 (5)	740
S. pneumoniae pspA gene	98 (0)	100 (0)	100 (0)	100 (0)	100 (0)	99 (1)	91 (5)	790
U. urealyticum #1	96 (4)	100 (0)	100 (0)	100 (0)	99 (1)	96 (3)	90 (3)	655
U. urealyticum #2	93 (3)	98 (1)	100 (0)	100 (0)	99 (1)	99 (1)	89 (6)	689
U. urealyticum #3	93 (3)	98 (1)	100 (0)	100 (0)	100 (0)	99 (0)	77 (8)	674
U. urealyticum #4	96 (3)	99 (1)	99 (1)	97 (2)	92 (8)	96 (3)	68 (8)	650
M. fermentans[a]	100 (0)	0[a]	0[a]	0[a]	0[a]	1	4	725
Average	96.8 (1.6)	99.1 (0.6)	99.8 (0.1)	99.1 (0.6)	97.8 (1.8)	96.5 (1.9)	83 (6.0)	712 ± 18

Each unedited sequence determined from genomic DNA template is aligned and compared with its corresponding GenBank database sequence. The fidelity of each sequence is listed as percentage agreement in a 100-bp interval. The number following the percentage value is the number of ambiguous bases (*N*) for each interval. The usable data range is the length of sequence that would be employed after human editing of the computer-generated base-calls.

[a]The *M. fermentans* initiation factor database sequence only overlapped the new sequence for 150 bases, limiting the comparison to that stretch.

quality template DNA for each sequencing reaction—an amount that could be difficult to obtain from certain microorganisms.

In summary, we have demonstrated that template DNAs up to 4.6 Mb in size can be directly sequenced with automated sequencers using custom primers and BigDye terminators. This protocol could expedite large-scale sequencing projects by facilitating gap closure, especially for difficult areas that are refractory to cloning and PCR methods. The technique can also be applied to bypass tedious subcloning steps in bacterial sequencing projects that focus on individual genes.

METHODS

Preparation of DNA Samples

Multiple methods, including Easy-DNA kit (Invitrogen, USA), SDS–proteinase K lysis procedure, and isopycnic banding in CsCl gradients (Wilson 1994), all worked well to prepare microbial genomic DNA samples for sequencing. BAC DNA was purified as described by Rosenblum et al (1997). Sequencing primers were designed using Oligo 5.0 (National Biosciences, USA). Sizes varied from 21 to 25 bases and T_m from 60°C to 74°C (GC% method).

Sequencing Conditions

Each cycle sequencing reaction contained 16 µl of BigDye Terminator mix (PE-Applied Biosystems, USA), 13 pmoles of primer, and 2–3 µg of microbial DNA (or 6 pmoles of primer and 0.4 µg of BAC DNA) in a total of 40-µl volume. In some experiments 1 µl of ThermoFidelase I (Fidelity Systems, USA) was included to improve data quality, but the results were inconsistent, so it is not a standard component of current reaction mixtures. The cycle conditions were initial denaturation at 95°C for 5 min followed by 30–60 cycles (30 for BACs, 45 for all microbial DNAs, except *E. coli* samples, which get 60 cycles; see Results/Discussion) at 95°C for 30 sec, 55°C for 20 sec, and 60°C for 4 min. Excess dye terminators were removed with a spin column (PE-Applied Biosystems User Manual) and reaction mixtures were dried in a SpeedVac system. Each sample was resuspended in 2 µl of formamide solution and denatured by heat, and the entire volume loaded on an ABI 377 automated DNA sequencing instrument with a 36-well

comb and a 48-cm WTR (well-to-read) gel. Electrophoresis was at 2400 V for 10–11 hours. Sequence data were analyzed by ABI version 3.0.1b3 software modified for weak signal base-calling (available through PE-ABI ftp site).

ACKNOWLEDGMENTS

We thank Dan Allison for modifying the base-calling software, Jennifer Glass for help in preparing the microbial DNAs, and Chun-Nan Chen for valuable discussions. This work is supported, in part, by the National Institutes of Health (grants HG00201 and RO1 AI28279).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Chen, C.N., Y. Su, P. Baybayan, A. Siruno, R. Nagaraja, R. Mazzarella, D. Schlessinger, and E. Chen. 1996. Ordered shotgun sequencing of a 135 kb Xq25 YAC containing ANT2 and 4 possible genes, including three confirmed by EST matches. *Nucleic Acids Res.* 24: 4034–4041.

Chen, E.Y., M. Zollo, R. Mazzarella, A. Ciccodicola, C. Chen, L. Zuo, C. Heiner, F. Burough, M. Ripetto, D. Schlessinger, and M. D'Urso. 1996. Long-range sequence analysis in Xq28: Thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. *Hum. Mol. Genet.* 5: 659–668.

Chissoe, S.L., M.A Marra, L. Hillier, R. Brinkman, R.K. Wilson, and R.H. Waterston. 1997. Representation of cloned genomic sequences in two sequence vectors: Correlation of DNA sequence and subclone distribution. *Nucleic Acids Res.* 25: 2960–2966.

Rosenblum, B.B., L.G. Lee, S.L. Spurgeon, S.H. Khan, S.M. Menchen, C.R. Heiner, and S.-M. Chen. 1997. New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res.* 25: 4500–4504.

Wilson, K. 1994. Preparation of genomic DNA from bacteria. In *Current protocols in molecular biology* (ed. F.M. Ausubel, R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith, and K. Struhl), Vol. 1, pp. 2.4.1–2.4.5. John Wiley & Sons, New York, NY.

Received November 26, 1997; accepted in revised form March 18, 1998.