

Preliminary results from the DNA Sequencing Research Group (DSRG) 2000 Study: An evaluation of methods used to sequence and isolate Bacterial Artificial Chromosomes (BACs)

James VanEe¹, George Grills², Susan Hardin³, Dina Leviten⁴, Margaret Robinson⁵, and Theodore Thannhauser¹.
¹Cornell University, Ithaca, NY; ²Albert Einstein College of Medicine, Bronx, NY; ³University of Houston, Houston, TX;
⁴ICOS Corp., Bothell, WA; ⁵University of Utah, UT.

Preliminary results from the DNA Sequencing Research Group (DSRG) 2000 Study: An evaluation of methods used to sequence and isolate Bacterial Artificial Chromosomes (BACs)

James VanEe¹, George Grills², Susan Hardin³, Dina Leviten⁴, Margaret Robinson⁵, and Theodore Thannhauser¹.

¹Cornell University, Ithaca, NY; ²Albert Einstein College of Medicine, Bronx, NY; ³University of Houston, Houston, TX; ⁴ICOS Corp., Bothell, WA; ⁵University of Utah, UT.

Introduction:

Traditional approaches to genome wide sequencing employ strategies that require a detailed physical map for each region to be sequenced. Recently, new strategies have been proposed that eliminate the need for any prior physical mapping (1-3). A prerequisite to these new approaches is a collection of "end sequences" (sequences adjacent to the insertion sites) from all clones in a deep coverage library of bacterial artificial chromosomes (BAC). These end sequences are known as "sequence-tagged connectors" (STCs) and are used to select the path of minimally overlapping BAC clones to sequence.

This change in the sequencing paradigm promises to reduce the cost and accelerate the pace at which genomes can be sequenced. However, its emphasis on BACs as a primary sequencing template has created significant challenges for those who operate DNA sequencing facilities. Due to their large size (80-350 kb) BACs behave differently in sequencing reactions than do plasmid clones. To achieve optimal performance, standard sequencing protocols must be modified. Due to their low copy number it is difficult to isolate sufficient quantities of BAC template using standard minipreps. This is particularly true in high throughput formats. Furthermore, standard templates that are commonly used as controls in sequencing protocols (e.g. pGEM) are not adequate when sequencing BAC templates. There is a need for a standard that can be used to develop and refine BAC sequencing and isolation protocols and that can serve as an adequate control when sequencing BAC clones.

Here we present the preliminary results of a two-part study focusing on the methodologies in current use to sequence and isolate BAC DNA. In this study the methods used are correlated with sequencing outcome with a view towards developing optimized standard procedures.

Materials and Methods:

The standard BAC used for this study was selected at random from a library of *Brassica oleracea*. A stock of this clone was used to inoculate and grow cultures from which the BAC DNA was purified using the Large-Construct Kit from Qiagen. This material was evaluated by A_{260}/A_{280} gel electrophoresis and by sequencing. Only batches that were judged to be acceptable by these criteria were utilized for Part 1 of this study. For part 2, agar stabs were generated from an overnight culture that had been inoculated with stock.

Participation in this study was solicited through a general mailing to the ABRF membership and by posting the study announcement to appropriate electronic bulletin boards and discussion groups. Those interested in participating in part 1 were supplied with 2 μ g of the purified Standard BAC DNA and 200 pmoles of T7 primer. Those interested in participating in part 2 of the study were supplied with a bacterial stab and the information necessary to grow the cells.

The participants in part 1 were asked to sequence the Standard BAC template with the primer provided. Submissions were restricted to this template/primer combination, but no restrictions were imposed regarding the sequencing conditions, protocols and instrumentation used. The sequence data was submitted electronically and the details of the sequencing conditions were collected on web-based survey forms. Great care was taken to ensure the anonymity of all participants.

The performance level of each submission to part 1 was judged by sequence quality. To determine sequence quality the unedited sequences were analyzed by phred (4,5). The total number of base calls with a quality value >20 ("phred score") for each sequence was determined with the program "quat" (written by J. V.), and was used to rank the submissions for each machine type and configuration represented in the study.

The participants in part 2 were asked to grow cultures from the bacterial stab provided, and isolate the BAC DNA by the methodology currently utilized in their laboratories. Information concerning the growth conditions and purification protocols was collected on web-based survey forms and samples of the purified BAC DNA were sent to the DNA Sequencing Research Group for sequencing. These samples were treated identically, using standard protocols for thermocycling, terminator removal, and sequencing. Each prep was judged by the quality of the sequence obtained from it (as determined by the "phred score"). Furthermore, the various preps were compared on the basis of yield, "hands on" time and cost.

About the Template:

Figure 1 is a plot of the fraction of G+C for the standard BAC template as a function of sequence position calculated over sliding 50 base windows. It is presented here to emphasize that this study was intended to explore the effects of template size on sequencing success. This template does not represent a significant challenge from the point of view of GC content.

Results, Part 1:

We received requests from 45 individuals for the material necessary to participate in part 1. In response we received 77 data files from 37 of the 45 respondents before the deadline for inclusion in this preliminary report. The response rate of 82% is the highest achieved for any DSRG study. A variety of machine types and configurations were represented in the data set as is demonstrated by Figure 2. Many different thermocycling protocols were also included in the data. The range of the conditions used is represented graphically in Figure 3.

Figure 1. Fraction G+C (Window=50)

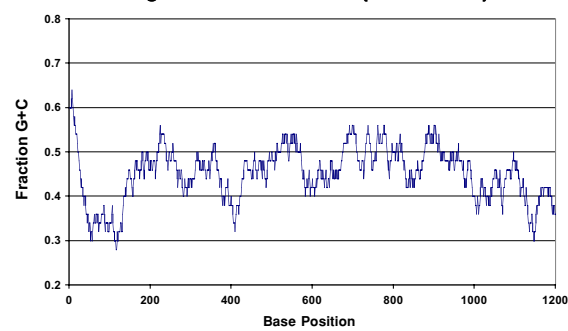


Figure 2. Number of Submissions by Machine Type

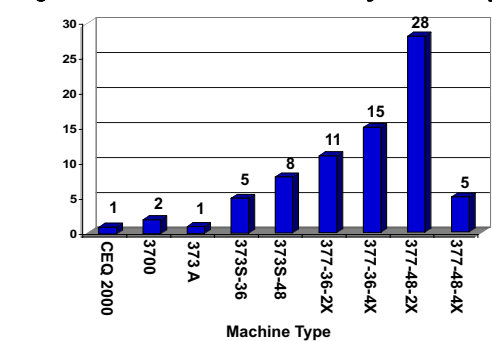
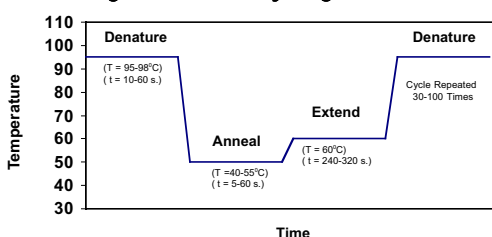


Figure 3. Thermocycling Profiles Used



The submissions were organized on the basis of machine type and configuration and then ranked by their phred score. The ranking of all submissions is given in Table 1. Although the ranking was done on the basis of the phred score, information concerning numbers of errors over defined sequence intervals is also given. The results given in table 1 suggest that the BAC template utilized here will be useful as a control when sequencing BACs and possibly other large construct templates.

Much effort was expended in an attempt to correlate sequencing and thermocycling conditions with sequencing outcome. Although these were largely unsuccessful, the following observations were made:

- Strong correlation between WTR and run time with higher phred scores
- Tenuous correlation between amount of template used with phred scores
- Weak correlation between number of cycles used with phred scores

Conclusions, Part 1:

This study has demonstrated a wide variety of protocols and results obtained from sequencing a large construct BAC template. Interest has been high with a response rate of over 80% in just 27 days! It is generally perceived that there is a need in the core facility for standard templates beyond the regular pGEM controls already run in many laboratories. The BAC template used in this study would be very suitable as a standard and could be made available to the ABRF community for use in controlling the conditions for large construct sequencing. The DSRG intends to investigate the best means for making this BAC standard available to interested users. Although there may not yet be clear indications from the data to recommend a standard sequencing protocol, the data collected in this study will form a foundation for future research. The DSRG will use the data as a basis for an internal research group study that will try to examine some of the more critical conditions suggested in this study, but performed under stringent controlled conditions until a robust standard protocol can be recommended. In the meantime, the study remains open for further submissions and more BAC DNA may be requested for further experimentation.

Table 1. Ranking of all submissions by machine type and configuration

Instrument	Filename	Phred Score	Errors				RunTime(hr)	Results?
			41-240	41-440	41-640	41-840		
CEQ 2000	A7444A.SCF	432	3	3	12	85	3	normal
3700	C3807A.ABI	586	5	5	5	20	4.5	normal
	C3807B.ABI	461	20	21	27	95	4.5	normal
373A	A5140A.ABI	466	0	1	11	NC	14	excellent
373S-36	A1234A.ABI	499	0	0	4	NC	14	normal
	B5710B.ABI	484	0	0	9	NC	12.05	norma3
	B5710A.ABI	479	6	6	10	NC	12.05	normal
	A3303B.ABI	192	1	39	104	NC	14	poor
A3303A.ABI	66	NA	NA	NA	NA	14	poor	
373S-48	A9923A.ABI	712	0	0	0	0	16.5	normal
	A4666B.ABI	599	0	1	6	37	18	normal
	A6295B.ABI	592	1	1	1	16	18	normal
	A1919A.ABI	566	12	12	15	29	15.5	normal
	1925A.ABI	551	1	1	5	18	18	normal
	A4666A.ABI	423	6	8	21	69	18	poor
A6295A.ABI	362	1	4	33	95	18	normal	
A1919B.ABI	291	29	45	83	112	15.5	normal	
377-36-2X	C7553B.ABI	649	0	0	2	NC	9	excellent
	C7553D.ABI	586	2	3	4	NC	9	excellent
	T4132A.ABI	566	0	0	0	NC	9	normal
	B5854A.ABI	539	0	0	11	NC	8	normal
	C0216A.ABI	530	0	0	1	NC	7.5	excellent
	C8980B.ABI	526	0	0	0	19	9	excellent
	7720A.ABI	512	2	2	4	NC	8	normal
	A8488A.ABI	451	6	6	28	NC	8	poor
	C0216B.ABI	327	124	126	132	NC	7.5	poor
	C8980A.ABI	279	4	9	37	NC	9	normal
	A5140B.ABI	79	NA	NA	NA	NA	9	poor
	377-36-4X	A9636C.ABI	455	3	3	38	NC	3
C4717B.ABI		455	6	6	37	NC	3	normal
A9636A.ABI		448	2	2	43	NC	3	normal
B2941D.ABI		448	0	0	36	NC	3	normal
C4717E.ABI		447	3	3	25	NC	3	normal
B2941F.ABI		436	0	0	37	NC	3	normal
B2941B.ABI		434	0	0	37	NC	3	normal
B2941C.ABI		431	3	5	40	NC	3	normal
B2941E.ABI		427	0	0	39	NC	3	normal
A9636D.ABI		425	2	2	45	NC	3	normal
C4717D.ABI		422	5	5	34	NC	3	normal
C4717A.ABI		420	6	6	41	NC	3	normal
A9636B.ABI	417	0	0	47	NC	3	normal	
C4717C.ABI	397	5	5	36	NC	3	normal	
B2941A.ABI	358	2	4	75	NC	3	normal	
377-48-2X	A2001B.ABI	720	0	0	0	1	16	normal
	C6459.ABI	696	0	0	0	2	12	normal
	A2001A.ABI	678	0	0	0	1	16	normal
	A0715.ABI	638	0	0	1	26	10	normal
	C7952.ABI	627	0	0	0	15	11	normal
	C2724G.ABI	606	0	0	0	50	10	normal
	B1255B.ABI	575	1	1	1	16	10	normal
	C9322K.ABI	554	4	4	4	33	10	normal
	C9322H.ABI	544	6	6	7	35	10	normal
	C9322I.ABI	541	7	7	8	27	10	normal
	B5622J.ABI	539	2	2	4	26	10	normal
	B5622I.ABI	538	3	3	9	50	10	normal
C7610A.ABI	529	0	0	6	100	10	normal	
A2724C.ABI	523	0	0	6	67	10	normal	
B5622G.ABI	521	12	12	14	37	10	normal	
C9322J.ABI	509	4	4	5	24	10	normal	
A0674B.ABI	508	0	0	8	112	10	normal	
B1255A.ABI	470	0	0	3	35	10	normal	
C8677F.ABI	466	6	6	8	36	10	normal	
A2724B.ABI	457	3	5	12	46	10	poor	
C8677G.ABI	454	6	6	14	53	10	normal	
X5677A.ABI	450	16	18	22	41	16	poor	
C2724F.ABI	424	0	3	15	79	10	normal	
C5558A.ABI	350	2	16	40	115	10	poor	
A0674A.ABI	315	2	7	30	127	10	normal	
B2724D.ABI	246	9	18	41	100	10	poor	
B2724E.ABI	190	NA	NA	NA	NA	10	poor	
A2724A.ABI	1	NA	NA	NA	NA	10	poor	
377-48-4X	T4132B.ABI	552	0	0	0	16	10	normal
	C7228C.ABI	523	0	0	2	43	10	normal
	C7228B.ABI	378	17	17	28	112	10	normal
	C7228D.ABI	368	12	14	22	93	10	normal
	C7228A.ABI	272	3	11	39	140	10	normal

Results, Part 2:

Twenty individuals requested material for part 2. The lower number of requests reflects the fact that only 25% of laboratories that provide DNA sequencing also offer a template purification. Seven responded by the deadline for inclusion in this preliminary report (participation rate 35%). The purification protocols represented in this study break down into two basic types: commercial, single tube format approaches that emphasize product quality over other factors, and a 96 well plate format that maximizes throughput, minimizes cost but sacrifices product quality to some extent. The samples submitted to this part of the study were sequenced by a standard protocol which involved the thermocycling profile represented in Figure 4. Terminator removal was performed by ethanol precipitation and the samples were sequenced on an ABI model 377 sequencer utilizing a 48cm WTR and 2X run conditions. The submissions are summarized in Table 2.

Figure 4. Standard Thermocycler Profile

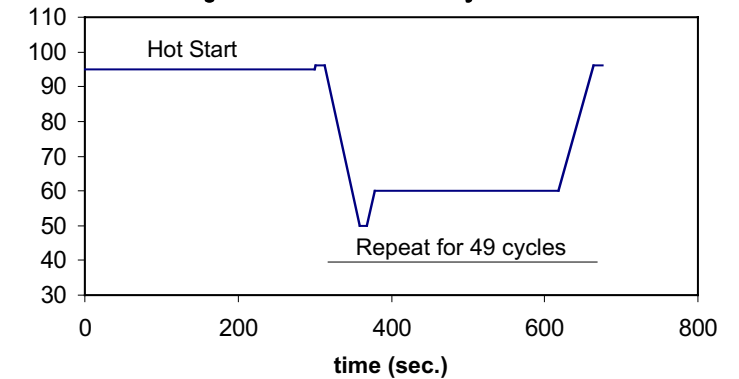


Table 2. Summary of Submissions, Part 2

Prep Type	Sample ID	Average Phred Score ¹ (+/-SD)	Kit Manufacturer (type)	Format	Yield	Quantitation	time/prep	cost/prep ³
Large Scale	1919	607 (+/-25)	Qia.(Midi plasmid)	single tube	40 ug	OD 260	4 hr.	\$187.20
	1974	494 (+/-164)	Qia.(Midi plasmid)	single tube	100-200ug	OD 260	2 hr.	\$97.20
	2724	599 (+/-28)	Gen. Sys.(MKB-100)	single tube	40-200ug	OD 260	2 hr.	\$110.00
	9312d	410 *	Qia.(L.Construct)	single tube	15ug	OD 260	1.25 hr.	\$58.45
	9312nd	292 *	Qia.(L.Construct)	single tube	15ug	OD 260	1.25 hr.	\$58.45
Small Scale	0839d	411 (+/-65)	in house method	96 well	0.2-1 ug	ND	1 hr.	\$0.75
	0839nd	226 (+/-57)	in house method	96 well	0.2-1 ug	ND	1 hr.	\$0.75
	5622d	405 (+/-97)	in house method	96 well	0.2-1 ug	ND	1 hr.	\$0.75
	5622nd	403 (+/-151)	in house method	96 well	0.2-1 ug	ND	1 hr.	\$0.75
	4519A	479 *	P. S. (Psi cine BAC)	single tube	1ug	OD 260	2 hr.	\$91.88
4519B	495 *	P. S. (Psi cine BAC)	single tube	1ug	OD 260	2 hr.	\$91.88	

- 1) The average (n=5-7) of the total number of base calls per sequence with an associated quality value of > than 20 as determined by phred.
- 2) This number represents a single value.
- 3) The cost was figured assuming a labor rate of \$45/hr.

Conclusions, Part 2:

- Commercial, single tube format DNA purification kits can be used effectively to isolate large quantities (40-200 μ g) of high quality BAC template that are suitable for sequencing. Submissions that were purified by this type of purification protocol gave the highest quality sequence as judged by average phred scores.
- The 96 well, high throughput approach to BAC isolation has definite advantages relative to the single tube formats involving both cost and speed. However, BAC DNA isolated by this procedure gave significantly lower phred scores when sequenced by our standard procedure. It is unclear at this point whether this difference is due to a lower quality product or variable yields from the small volume cultures.

Acknowledgements:

The authors would like to thank the following people for their many contributions to this study: Professor June Nasrallah (Cornell University) for providing the BAC clone used in this study; Tom Stelick, Bill Enslow, Tatyana Pynitkova and Farhad Dhabhar (Cornell University) for their excellent technical assistance.

References:

- 1) Weber, J. L. and Myers, E. W. (1997) Human whole-genome shotgun sequencing, *Genome Research* 7: (5) 401-409.
- 2) Batzoglou, S., Berger, B., Mesirov, J. and Lander, E. S. (1999) Sequencing a genome by walking with clone-end sequences: A mathematical analysis *Genome Research* 9: (12) 1163-1174.
- 3) Siegel, A. F., Trask, B., Roach, J. C., Mahairas, G. G., Hood, L. and Engh, G. V. D. (1999) Analysis of sequence-tagged-connector strategies for DNA sequencing, *Genome Research* 9: (3) 297-307.
- 4) Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998) Base-calling of automated sequencer traces using phred: I. Accuracy assessment. *Genome Research* 8:175-185.
- 5) Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred: II. Error probabilities. *Genome Research* 8:186-194.