DOE BAC End Sequencing Project Summary

We have successfully completed the project and achieved all goals within the project period as summarized below.

Human BAC End Sequencing We have generated over 300,000 end sequences from >186,000 human BAC clones from CalTech BAC libraries and the RPCI-11 library, with an average read length of 460 bp and average Q20 bases of 394 bp (Zhao *et al.*, 2000). Over 60% of the clones have BAC end sequences (BESs) from both ends representing >5X coverage of the genome by the paired-end clones (Table 1). Compared to other centers, TIGR's BESs were found to be of the highest quality (see Figure 1). These sequences have provided the basic substrates to build the human genome assembly scaffold (Venter *et al.*, 2001, International Human Genome Sequencing Consortium, 2001)

Table 1 TIGR Large Scale BAC End Sequencing Projects

Species	Total BESs[a]	Total Clones[b]	%Pairs[c]	Success Rate[d] (%)	Read Length[e] (bp)	Q20 Bases[f] (bp)	Clone Tracking Accuracy[g]
Human	300,000	186,000	60	67	460	394	>91%

Table 1. Legend:

[a] The total number of BAC end sequences (BESs) generated by the project.

[b] Clones with at least one BES.

[c] % of clones with both BESs (pairs).

[d] % of sequencing attempts that yielded reads with an overall error rate of <2.5%, an edited read length of >100 bp, and free of *E. coli* and vector sequences.

[e] Read length after sequences were trimmed.

[f] Average number of bases with phred QV ≥ 20 per sequence, after sequences were trimmed.

[g] % of BESs that were associated with the correct clone identifiers.

DOE Patent Clearance Granted
Mark P Dvorscak 9/27/01
Mark P. Dvorscak Date
(630) 252-2393
E-mail: mark.dvorscak@ch.doe.gov
Office of Intellectual Property Law
DOE Chicago Operations Office

DISCLAIMER

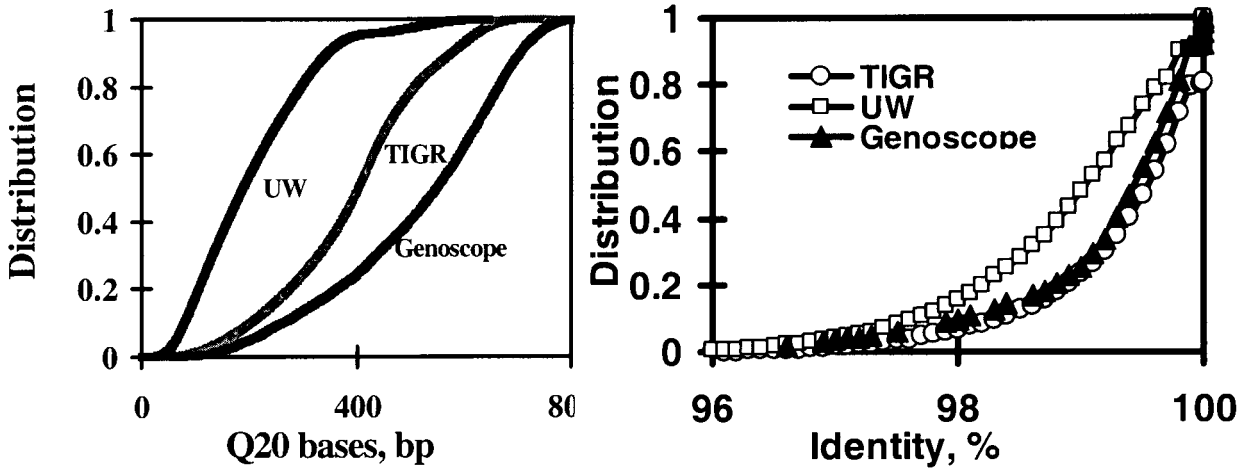

Figure 1a. **Q20 length distributions of raw traces of human BAC end sequences (BESs).** The base calling program, phred, was run on 328,084 chromatogram files from UW (Mahairas *et al.* 1999), 282,825 files from TIGR and 7,586 files from Genoscope. From phred output quality files, the quality score of each base was examined. The number of bases with a score of ≥ 20 (Q20 length) was counted for each trace and distributions of Q20 lengths were plotted.

Figure 1b. **The match identity distributions of human BESs to human finished sequences**. Human finished sequences (2,078 contigs ranging from 2 kb to 12 Mb) were searched against a repeat-masked BES database consisting of sequences from TIGR, UW and Genoscope. The search was done using BLASTN and FASTA. Matches with identity ≥ 90% and match length ≥ 0.75 BES length were selected.

Process Improvements During the course of the large scale BAC end sequencing project, we have been conducting extensive R&D work to continuously improve our protocols. This was demonstrated by several facts: 1) the sequencing success rate was increased from 67% to 85%, 2) the average read length was increased from 470 bp to above 650 bp and the average phred Q20 length was increased from 400 bp to over 570 bp, and 3) the clone tracking accuracy was increased from >91% to >98%. In addition, a fully automated BAC purification protocol has been developed and implemented into the production pipeline (Figure 2).

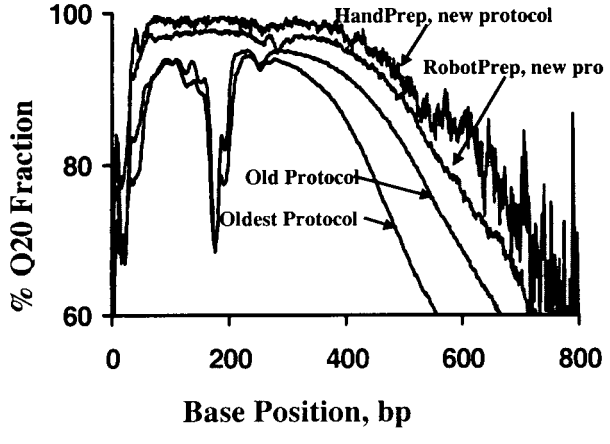

Figure 2. Process Improvement. The phred quality score was examined for each base of each BES sequenced with different protocols. The percentage of the Q20 base at each position was calculated and plotted. New protocol yielded higher quality sequences, as indicated by both the longer reads and the disappearance of a quality dip in the region of bases 160-200. Also the robot prep had a similar sequence quality as the hand prep.

Cost Reduction from human to rat We have improved our BAC end sequencing protocols. As a result, the sequencing success rate was increased from 67% to above 84%, the average read length was increased from 470 bp to 650 bp, and the average Q20 bases were increased from 400 bp to 570 bp. This alone decreases the cost by reducing the total number of attempted sequences by 20% per successful sequence generated, 27% per base generated and 30% per Q20 base generated. The sequencing was performed on ABI capillary sequencers instead of ABI 377 machines, which eliminated lane mis-tracking and increased the clone tracking accuracy by at least 7% (therefore 7% more sequences were useful). The primary base calling was performed by the software phred and the quality scores were further adjusted by the software TraceTuner (Paracel) that was specifically trained for our 3700 traces. As a result, the base call accuracy and the trimmed read length were increased. For the sequencing process itself, we decreased the amount of Big Dye mix from the original 10 μl to the current 1-3 μl. We have successfully developed and implemented a robotic BAC DNA isolation method, which decreased the labor and the downstream requirement for the amount of Big Dye mix. In addition, we have automated most steps in the pipeline. As a result, the total cost for paired BAC end sequencing decreased from $14-15/clone to $5-6/clone.

Random and Targeted Primate BAC End Sequencing Collaborating with Evan Eichler and Vladimer Larionov, we have sequenced over 1000 primate clones isolated from human duplicated or hyper-variable regions. These sequences have been useful in finishing the human genome by identifying assembly errors and facilitating gap closure (Kouprina *et al.* 2003, Liu *et al.* 2003). In addition, we have sequenced a number of random clones from several primates (Table 2) and a genetic distance from human has been calculated for each species with the data (Figure 3)

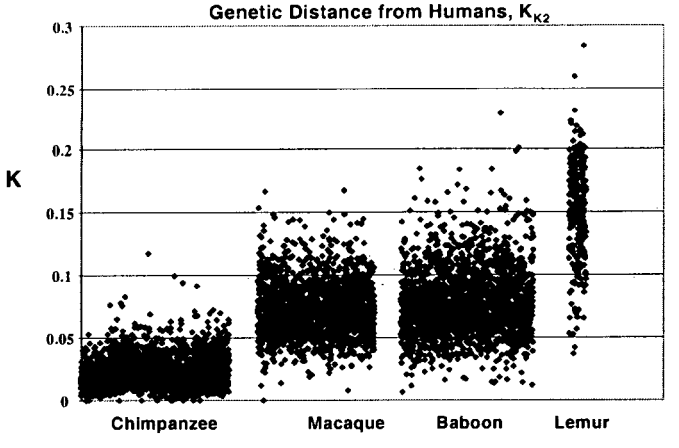

Figure 3. Estimated rate of mutation for primates (1.2-1.4 X 10 -9 changes/bp/year). The data were obtained based on the sequence identities to human genome sequences of 2000 random chimp BES pairs, 1300 rhesus macaque BES pairs, 1800 baboon BES pairs and 1000 lemur BES pairs that were placed onto the human genome with right orientation and insert size between 1-400kb.

Table 2. Primate BAC end sequencing

Species	Libraries	Paired End Clones	Read Length (bp)	Q20 Length (bp)
Chimp	RP43	3101	493	410
Baboon	RP41	3586	519	438
Rhesus	CH250	4844	552	463
Marmoset	CH259	2500	740	669
Lemur	LBNL	6792	590	503

Protocadherin Gene Cluster We have placed lemur paired-ends to the human genome and found that the average distance between the two ends were 220kb, which was higher than the experimentally determined average insert size of 175kb. To further confirm the data, we experimentally determined 50 BACs that have been placed on the human genome and found this was indeed the case. This would indicate that the lemur genome is ~15-20% smaller than the human genome. We are interested in the reasons behind.

We sequenced and annotated a lemur BAC with an actual insert size of 180kb and with its ends spanning >300kb regions on human chr5 that contains protocadherin alpha A1, A2 and A3, beta 1-15, ornithine transporter 2 as well as TAF7 RNA polymerase II TATA box binding protein. We found that the lemur clone contained protocadherin A1 and A3 but no A2, procadherin beta 1-7 and 14-15 with 8 and 13, 9 and 10, 11 and 12 each merged into one gene, as well as ornithine transporter 2 and TAF7. The gene order was identical to the human. We are now preparing a manuscript to report this and are sequencing more clones from more primate species to get a completely picture of evolution for this important gene cluster.

Data availability We have submitted all BAC end sequences generated to GenBank. In addition, all sequence data and quality files are available on the TIGR website.

Refereence

Ge Liu, NISC Comparative Sequencing Program, Shaying Zhao, Jeffrey A. Bailey, S. Cenk Sahinalp, Can Alkan, Eray Tuzun, Eric D. Green, and Evan E. Eichler (2003) Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome. *Genome Research* 13, 358-368.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature 409, 860-921.*

Natalay Kouprina, Sun-Hee Leem, Greg Solomon, Albert Ly, Maxim Koriabine, John Otstot, Eugene Pak, Amalia Dutra, Shaying Zhao, J. Carl Barrett & Vladimir Larionov (2003) Segments missing from the draft human genome sequence can be isolated by transformation-associated recombination cloning in yeast. *EMBO* 4, 257-62.

Venter J. C., Maker D. Adams, Eugene W. Myers, *et al.* (2001) The Sequence of the Human Genome. *Science 291, 1304-1351.*

Zhao, S., Joel Malek, Gregory Mahairas, Lily Fu, William Nierman, J. Craig Venter, lsand Mark D. Adams. (2000b) Human BAC Ends Quality Assessment and Sequence Analyses. *Genomics 63 (3), 321-332.*