Principal Investigator: Nierman, William C., Ph.D.

FINAL PROGRESS REPORT

GRANT NO. DE-FC02-97ER62500
"Construction of a Genome-Wide, Highly Characterized Clone Resource for Genome Sequencing"
between
U.S. Department of Energy
and
The Institute for Genomic Research

Libraries constructed in Bacterial Artificial Chromosome (BAC) vectors have become the standard clone sets in high throughput genomic sequencing projects primarily because of their high stability (Shizuya *et al.*, 1992). End sequences from BACs provide highly specific sequence markers in large-scale sequencing projects. A genome sequencing approach (Venter *et al.*, 1996) has been described in which a clone contig is extended by selecting the minimally overlapping clones in each direction by searching the finished BAC sequence against a BAC end sequence (BES) database, so that the BAC clone with the smallest overlap can be chosen to extend the contig. Since BAC clones (average insert size 150 kb) are sufficiently large to traverse most tandem arrays of homology units and repeats, end sequences are very useful in genome assembly and chromosome walking. The recent whole-genome shotgun sequencing announcement (Venter *et al.*, 1998) will rely on BESs as the primary scaffold onto which the end sequences from the smaller clones will be assembled. End sequences have been utilized extensively to confirm, join and order existing contigs (for example, http://compbio.ornl.gov/tools/channel/index.html).

The US Department of Energy (DOE) has funded both The Institute of Genomic Research (TIGR) and the University of Washington (UW) for a large scale human BAC end sequencing from Sep. 15, 1997 to Nov. 15, 1999. The goal of this project is to generate 600,000 BESs from 300,000 clones, which represents 15X clone coverage and 10% sequence coverage of the genome, providing one sequence marker every 5 kb. We have generated and deposited in GenBank 304,106 BAC end sequences (BESs) at TIGR, reaching our goal on June 1, 1999, five month ahead of the schedule. Additional 460,000 has been generated at UW (Mahairas, *et al.*, 1999). Sequence data from both TIGR and UW are available in GenBank and are also available for ftp and search through our website at http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_search.html (Zhao, 2000).

At TIGR, the human BAC end sequencing and trimming were performed as described (Kelley *et al.*, 1999) with an overall sequencing success rate of 65%. CalTech human BAC libraries A, B, C and D as well as Roswell Park Cancer Institute's library RPCI-11 were used. To date, we have generated >300,000 end sequences from >186,000 human BAC clones with an average read length ~460 bp for a total of 141 Mb covering ~4.7% of the genome. Over sixty percent of the clones have BAC end sequences (BESs) from both ends representing over five-fold coverage of the genome by the paired-end clones. The average phred Q20 length is ~400 bp. This high accuracy makes our BESs match the human finished sequences with an average identity of 99% and a match length of 450 bp, and a frequency of one match per 12.8 kb contig sequence. Our sample tracking has

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

ensured a clone tracking accuracy of >90%, which gives researchers a high confidence in 1) retrieving the right clones from the BAC libraries based on the sequence matches; and 2) building a minimum tiling path of sequence-ready clones across the genome and genome assembly scaffolds.

Additional sequencing performed at UW has resulted in over 700,000 total end-sequences from BAC clones. We have conducted quality assessment and sequence analyses on BESs from both UW and TIGR ends (Zhao *et al*, in press). All our analyses indicate that BESs from CalTech libraries and RPCI-11 have similar properties. The average number of high-quality bases (phred score >=20) of the raw traces is 394 bp for TIGR and 202 bp for UW. Because of the higher phred quality scores, more TIGR BESs hit human genomic sequences, STSs, ESTs and proteins with higher percent identity. BESs from TIGR and UW match finished human genomic sequences with an average identity of 99% over an average of 400 bp, indicating BESs from both labs are sufficiently accurate for use in large-scale sequencing projects. Although the nominal pairing rate (clones sequenced from both ends) is about 60% overall, there appears to be a substantial number of end-sequences attributed to an incorrect clone. This raises questions as to the ability to use the BES resource for contig extension and for confirming the large-scale assembly of sequenced regions. The effective clone coverage was 4.4X, based on both ends of a clone matching one of 82 large contigs from GenBank that are at least 600 kb in length. This suggests that it will be necessary to rely on end sequences from two independent BAC clones to verify contig extension or genome assembly. The BES database contains a large number of sequences from CalTech D library plates >=3000; these clones have a much smaller predicted insert size than previously realized. These clones may be useful for filling smaller gaps between previously sequenced regions. Taken together, these results demonstrate that the BES data are not uniform in quality. It is therefore important for users of the BES information to be aware of the differences in average clone insert size, variability in accuracy, and potential for misnaming of end-sequences.

We have extensively analyzed BESs for the contents of STSs, ESTs, protein coding regions and repeats. We have found that 0.26%, 2.7% and 0.4% BESs contain STSs, ESTs and protein coding regions, respectively. It has been estimated that 3% of the human genome encodes genes. Our EST analyses indicate that about 1% of the BES bases match ESTs, implying that substantial additional coding regions of the genome remain to be identified. In addition, both EST and protein analyses identified some BACs potentially encoding genes of significant research interest. The STS analyses as well as other analyses allow us to putatively assign >10,000 BACs to unique chromosome locations, providing important information for both genome projects and projects involving particular genes. Between 55% and 60% of BESs contain known genome-wide repeats. The percentage of base pairs in repeats (34%) is slightly smaller than previously reported (Smit, 1996) with most of the difference attributable to a lower abundance of LINE elements. BESs are quite unique since <0.06% of repeat-masked BESs match more than fifty other BESs, indicating few novel genome-wide repeats. At least 60% of the BACs have >=100 bp contiguous unique sequences from both ends. Those BESs are most useful for the genome assembly project and we expect that this resource will prove valuable in many areas of genome research.

References:

Kelley, J.M., Field, C.E., Craven, M.B., Bocskai, D., Kim, U., Rounsley, S.D and Adams, M.D. (1999) High Throughput Direct End Sequencing of BAC Clones. *Nucleic Acids Res.*, 27, 1539-1546.

Mahairas Gregory G. James C. Wallace, Kim Smith, Steven Swartzell, Ted Holzman, Andrew Keller, Ron Shaker, Jeff Furlong, Janet Young, Shaying Zhao, Mark D. Adams, and Leroy Hood (1999) Sequence Tagged Connectors: A Sequence Approach to Mapping and Scanning the Human Genome. *PNAS* 96, 9739-44.

Shizuya, H., Birren, B., Kim U., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M. I. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl.Acad. Sci. USA*, 89, 8794-8797.

Smit, A.F., (1996) The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev*, 6, 743-748.

Venter, J. C., Adams, M. D, Sutton, G.G., Kerlavage, A. R., Smith, H. O. and Hunkapiller, M. (1998) Shotgun sequencing of the human genome. *Science*, 280, 1540-1542.

Venter, J. C., Smith, H. O. and Hood, L. (1996) A New Strategy for Genome Sequencing. *Nature* , 381, 364-366.

Zhao, S. (2000) Human BAC Ends. *Nucleic Acids Res.* 28,129-132.

Zhao, S, Joel Malek, Gregory Mahairas, Lily Fu, William Nierman, J. Craig Venter, and Mark D. Adams (2000) Human BAC Ends Quality Assessment and Sequence Analyses. *Genomics*,, in press.