Human BAC Ends

Shaying Zhao*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received September 2, 1999; Accepted September 16, 1999

ABSTRACT

The Human BAC Ends database includes all non-redundant human BAC end sequences (BESs) generated by The Institute for Genomic Research (TIGR), the University of Washington (UW) and California Institute of Technology (CalTech). It incorporates the available BAC mapping data from different genome centers and the annotation results of each end sequence for the contents of repeats, ESTs and STS markers. For each BAC end the database integrates the sequence, the phred quality scores, the map and the annotation, and provides links to sites of the library information, the reports of GenBank, dbGSS and GDB, and other relevant data. The database is freely accessible via the web and supports sequence or clone searches and anonymous FTP. The relevant sites and resources are described at http://www. tigr.org/tdb/humgen/bac_end_search/bac_end_intro.html

INTRODUCTION

BAC ends are the single-pass sequence reads from each end of BAC clones. The availability of BAC ends for ~20-fold coverage in human BAC libraries has had a significant impact on large scale genomic sequencing projects. The use of BAC ends for large scale sequencing follows the basic strategy of using a finished BAC sequence for searching against the BAC Ends database to identify the minimally overlapping clones which extend in each direction (1). In addition, the paired-ends can be used to validate, order and to join contigs. For example, the recent whole-genome shotgun sequencing strategy (2) will rely significantly on BAC ends as the primary scaffold onto which the end sequences from the smaller clones will be assembled. Since 1997 the US Department of Energy has been funding the Institute for Genomic Research (TIGR) and the University of Washington (UW) to end sequence human BAC clones on large scales (http://www.ornl.gov/meetings/bacpac/index.html). To date >750 000 sequences from >450 000 clones have been generated by both centers using human BAC libraries from the California Institute of Technology (CalTech) and Roswell Park Cancer Institute (RPCI). Meanwhile, other large scale mapping projects are also underway at several genome centers and a significant number of BACs have been mapped. Ung-Jin Kim's group at CalTech has hybridized >5000 BACs to chromosome 16 and 22. Barbara Trask's group at UW has FISH-mapped ~900 RPCI-11 clones to single locations. David Cox's group at Stanford Human Genome Center is conducting a large scale RH mapping with BAC end sequences. Marco Marra's group at the Washington University Genome Sequencing Center (WU GSC) is fingerprinting the RPCI-11 BAC library.

In order to better support the resource, TIGR has been collecting all BAC ends from both centers and annotating each sequence for repeats, ESTs, STSs and finished genomic sequences. The available map data have been gathering from various genome centers. The human BAC ends database at TIGR is daily updated to integrate the end sequences, the phred quality scores, the map and the annotation data for each BAC end. The database is free to the public via a web sequence similarity and clone search service, and via anonymous FTP.

DATA SOURCE

End sequences

The database has incorporated >450 000 UW ends, >303 000 TIGR ends and 3825 CalTech ends, providing a total of ~740 000 non-redundant sequences from ~470 000 clones (Table 1). With an average read-length of 477 bp for each sequence the database contains a total of 350 million nucleotides (12% genome). With 58% of clones having sequences from both ends the database has >270 000 paired-ends representing 11.5× coverage of the genome in BACs with paired-ends. Human BAC libraries from CalTech (libraries A, B, C and D, http://www.tree.caltech.edu/lib_status.html) and RPCI (RPCI-11 library, http://bacpac.med.buffalo.edu/11framehmale.htm) were used. The status of end sequencing from the CalTech libraries are summarized in Table 2.

TIGR. TIGR has been end sequencing BACs from CalTech libraries A, B, C and D as well as RPCI-11 (3). For the CalTech D library, TIGR has sequenced some of plates 2003–2063 and most of plates 2163–2387 from the *Hin*dIII segment, and plates 2501–2657 from the *Eco*RI segment. For the RPCI-11 library, TIGR has sequenced most of plates 1–490 from segment 1 and 2. TIGR has generated 303 939 non-redundant sequences from 185 656 clones, 63.7% of which have paired-ends. Assuming an average insert size of 120 kb for CalTech clones and 165 kb for RPCI-11 segment 1 and 2 clones, the coverage by the paired-ends clones on the genome is 5.6×. TIGR's ends were processed by the same vector- and quality-trimming routines as plasmid sequence reads (3) and the average trimmed reads are 465 bp indicating a total of 142 Mb (4.7% genome). The quality assessment and sequence analyses indicate that TIGR's ends are of high quality. The average Q20 length is 394 bp before trimming and 396 bp after trimming. The end sequences match finished genomic sequences with an average identity of 98%. Table 3 summarizes TIGR ends.

*Tel: +1 301 838 3532; Fax: +1 301 838 0208; Email: szhao@tigr.org

Table 1. The BAC ends database summary

	BESs[a]	ReadLn[b]	%Genome[c]	Clones[d]	Pairs[e]	%Pair[f]	Cov[g]	SingleEnd BACs[h]
Total	742 913	477	11.8	470 619	272 294	57.9	11.5×	198 325; 42%
CalTech	387 569	452	5.8	233 226	154 340	66.2	4.9×	78 883; 34%
RPCI-11	355 344	505	6.0	237 393	117 954	49.7	6.5×	119 442; 50%

[a] Total non-redundant BAC end sequences (BESs).
[b] Read length after vector- and quality-trimming.
[c] The sequence coverage of BAC ends of the genome.
[d] Total clones with either one end or both ends.
[e] Clones with sequences from both ends (paired-ends clones).
[f] The percentage of paired-ends clones.
[g] The paired-ends clone coverage of the genome.
[h] BACs with only one end.

Table 2. CalTech human BAC libraries end sequencing status

Library	Plates	BESs	ReadLn	Clones	Pairs	%Pair
A		1751	422	1009	741	73.4
B		1616	385	899	717	79.8
C		23 860	378	15 457	8403	54.4
D1	2001–2423	180 905	454	107 536	73 369	68.2
D2	2501–2671	52 276	465	32 321	19 955	61.7
D2	3000–3253	125 645	458	74 499	51 146	68.9
D2	>3253	1460	516	1456	4	0.3

UW. The High Throughput Sequencing Center at UW (http://www.htsc.washington.edu) has been end sequencing BACs from CalTech C, D and RPCI-11. UW has sequenced most of plates 2001–2278 from the CalTech D *Hin*dIII segment and most of plates 3000–3253 from the *Eco*RI segment, and most of plates 577–1152 from segments 3 and 4 of RPCI-11. A total of 458 801 sequences submitted to GenBank by UW were ftp’d to TIGR and were vector cleaned. Excluding 16 834 ends overlapping with TIGR a total of 439 012 non-redundant ends from 283 822 clones were incorporated into the database. The average reads are 485 bp with a total of 209 Mb (7.1% genome). Clones with paired-ends are 54.6% indicating a 6.3× coverage on the genome, assuming an average insert size of 65 kb for CalTech D plates 3000+ clones and 120 kb for the rest of the CalTech clones, and 170 kb for RPCI-11 clones from segments 3 and 4. Table 4 summarizes UW ends.

CalTech. CalTech has sequenced 2038 ends from libraries C and D, and 1787 ends from A and B with 44% paired-ends. The ends are chromosome 16 specific and have been incorporated into the database.

Table 3. TIGR BAC ends summary

	BESs	Q20Ln[a]	ReadLn	%Genome	Clones	Pairs	%Pair	Insert	Cov
Total	303 939	394	465	4.7	185 656	118 283	63.7		5.6×
CalTech	143 793	391	443	2.1	87 737	56 056	63.9	120 kb	2.2×
RPCI-11	160 146	397	485	2.6	97 919	62 227	63.5	165 kb	3.4×

[a] The numbers of bases with phred quality score ≥20 before vector- and quality-trimming.

Table 4. UW BAC ends summary

	Plates	BESs	ReadLn	%Genome	Clones	Pairs	%Pair	Insert	Cov
Total		439 012	485	7.1	283 822	155 190	54.6		6.3×
RPCI-11		195 241	521	3.4	133 680	61 561	46	170 kb	3.5×
CIT-C		16 237	349	0.2	11 260	4977	44	120 kb	0.2×
CIT-D1	2001–2423	100 422	470	1.6	62 916	37 506	60	120 kb	1.5×
CIT-D2	≥3000	125 645	458	1.9	74 499	51 146	69	65 kb	1.1×
CIT-D2	≥4000	1451	517						

Mapping data

Hybridization. The chromosome 16 (7812 ends from 4839 clones) and chromosome 22 (781 ends from 458 clones) specific database contains the end sequences from BACs hybridized to chromosome 16 and chromosome 22 by Ung-Jin Kim's group at CalTech (http://informa.bio.caltech.edu/idx_www_tree.html).

FISH. About 3000 RPCI-11 clones will be FISH-mapped by Barbara Trask's group at UW using TIGR end sequencing DNA templates (http://fishfarm.biotech.washington.edu/BACResource/). The FISH data of 898 BACs mapped to single locations are currently available in the database.

RH mapping. A total of 29 120 TIGR end sequences from 25 906 clones have been sent to David Cox's group at Stanford Human Genome Center for RH mapping (http://www-shgc.stanford.edu/Mapping/index.html). The RH data will be incorporated into the database once available.

Fingerprints. The fingerprints of >200 000 RPCI-11 clones from plates 1–818 by Marco Marra's group at WU GSC are available at http://genome.wustl.edu/gsc/human/human_database.shtml , which is a good complementary resource to end sequences. The entire RPCI-11 library will be fingerprinted.

Sequence annotation

ESTs. About 3% of BAC ends matched ESTs from the TIGR EST database (~296 602 sequences) with an average identity of 99% and an average match length of 185 bp. Similar results were observed when comparing BAC ends to the human UniGene database from NCBI.

Protein coding regions. About 0.4% BAC end sequences contained protein coding regions. The matched genes are involved in signal transduction/communication (16%), cell defence (7.5%), gene expression (35%), cell cycle (6%), structure and motility (5%) and metabolism (9%). About 3% match candidate disease genes such as tumor suppressors or oncogenes. A substantial number (19%) are zinc fingers. Most of the matches are to genes from *Homo sapiens* (50%), *Mus musculus* (14%) and *Rattus norvegicus* (7%). BAC end sequences cover ~12% genome and should be a rich resource for gene finding.

STS markers. About 0.3% BAC ends matched STSs of dbSTS database (55 950 sequences) from GenBank with an average identity of 99% and an average match length of 246 bp. Similar results were observed using program ePCR (4). Putative map locations were assigned to ~1000 clones based on the results.

Finished sequence. BAC ends matched finished sequences with an average identity of 98% and match length of 431 bp. A total of 89 paired-ends were found to match the 3.9 Mb chromosome 6 contig Hs6_1643 indicating a 3.3× effective coverage of the genome by the paired-ends and with six gaps of 3–75 kb (see Supplementary Material, Fig. 1). Putative map locations were assigned to >5000 BACs whose paired-ends matched the same sequence with a reasonable insert size.

Repeats. About 58% of BAC ends and 34% of bases were repeat-masked by RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). The categories of repeats identified were similar to what was described by Smit (5). About 60% of BACs with paired-ends and 75% of BACs with single-end have ≥100 bp contiguous unique sequences. These ends are the most useful in the genome assembly projects. Comparing the repeat-masked BAC ends with each other indicated that >95% of sequences were unique. Further studies are underway to find out if the non-unique sequences are due to gene family, novel repeats or other reasons.

DATABASE ACCESS

Sequence similarity search

TIGR provides the public with a free service for sequence similarity searches of BAC ends (http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_search.html). Users can submit a query sequence of up to 500 kb to search the database. The search is performed in two steps. The first step involves BLASTN to query against the database with liberal criteria to select the BAC ends to be used in the second step. In the second phase search, a mini-database is constructed from all sequences resulting from the BLASTN search which may potentially match the query sequence. FASTA (6) is then used to compare the query sequence to this mini-database. Options are provided to select the unmasked or repeat-masked BAC ends in each step. The database choices are the whole dataset, the monthly incremental dataset, and the chromosome 16 and chromosome 22 specific dataset. The results are presented graphically indicating (i) minimally overlapped clones in each direction as the candidates to be sequenced next, and (ii) paired-ends matches to validate the contigs. The sequence alignments are also shown. Each matched end is hyperlinked to an annotation report described below. An Email option is provided and the results are viewed using the URL provided within 4 days. An example of the output is shown in the Supplementary Material accompanying this paper**.**

To make the search service more powerful, a parallel virtual machine linux cluster with 5 units of four 450 MHz pentium III processors, 18 GB storage space, 1 GB RM is dedicated for the BAC ends search. Software tools are under development to: (i) reduce the false positive and false negative hits by varying the repeat-masking level of BAC ends; (ii) identify the minimally overlapping clones by combining end sequences and fingerprints from WU GSC; (iii) validate contigs of any size users send; and (iv) order and join contigs for multiple query sequences.

Clone search

Each BAC end's sequence, phred scores, map and annotation are available through the clone search at http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_search.html#clone . The clone name consists of plate#, row letter and column#. Furthermore, to make it unique in the database, CalTech library A clones start with 'A-' and RPCI-11 clones start with 'R-'. For example, R-3J8 is the BAC from RPCI-11, plate 3, row J and column 8.

The statistics indicate that the database is accessed >2000 times per week through the search service by users worldwide.

Annotation reports

The annotation reports (http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_anno.html) integrate the unmasked- and repeat-masked sequences, the phred quality scores, the map and the annotation results for repeats, ESTs and STSs for each BAC end. Links are provided to sites of the BAC library, the reports of GenBank, dbGSS and GDB as well as other relevant information.

FTP BAC ends

The database is free to the public via anonymous FTP at ftp://ftp.tigr.org/pub/data/h_sapiens/bac_end_sequences . Multi-FASTA files of unmasked sequence, repeat-masked sequences and headers-only are available. The FASTA headers include the BAC name, the primer, sequence read length, library, GenBank accession no., dbGSS#, GDB ID, the source institute and sequence name internal to the institute. The whole dataset (hbends) as well as monthly (hbends_month) and weekly (hbends_week) incremental datasets are provided.

SUPPLEMENTARY MATERIAL

Supplementary material for this paper is available at NAR Online and consists of the following data:

- Relevant URLs
- Effective coverage by paired-ends
- An example of sequence similarity output
- An example of the annotation reports

ACKNOWLEDGEMENTS

I wish to thank Norman Doggett for his critical comments and editing on the manuscript. I am grateful for the BAC end sequences provided by TIGR, UW HTSC and CalTech, for the mapping work provided by the groups of Ung-Jin Kim, Barbara Trask, David Cox and Marco Marra, for the database support provided by Lily Fu, Michael Heaney and Michael Holmes and to Mark Adams for his guidance and encouragement. This work is funded by the US Department of Energy.

REFERENCES

1. Venter,J.C., Smith,H.O. and Hood,L. (1996) *Nature*, **381**, 364–366.
2. Venter,J.C., Adams,M.D, Sutton,G.G., Kerlavage,A.R., Smith,H.O. and Hunkapiller,M. (1998) *Science*, **280**, 1540–1542.
3. Kelley,J.M., Field,C.E., Craven,M.B., Bocskai,D., Kim,U., Rounsley,S.D. and Adams,M.D. (1999) *Nucleic Acids Res.*, **27**, 1539–1546.
4. Schuler,G.D (1998) *Trends Biotechnol.*, **16**, 456–459.
5. Smit,A.F. (1996) *Curr. Opin. Genet. Dev.*, **6**, 743–748.
6. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.