DE-FG02-98ER62675 **Pieter J. de Jong**
"Library Resources for BAC End-Sequencing"

FINAL TECHNICAL REPORT

This "Final Technical Report" represents progress attained on this project at Roswell Park Cancer Institute prior to my relocation to the Children's Hospital Oakland Research Institute (CHORI). The work proposed in this project will continue under a new award number at CHORI.

Studies directed towards the specific aims outlined for this budget period include: 1) the expansion of the current human BAC library (RPCI-11); 2) development of a new TARBAC cloning vector and it's incorporation in the new libraries; and 3) Create a new ten-fold redundant library for the mouse genome using the new TARBAC vector. In addition, the distribution of all BAC and PAC libraries, high-density screening filters and individual positively screened clones continues to be a high priority in our lab. This responsibility is performed in a cost-recovery manner via BACPAC Resources.

Extensive in-depth characterization and expansion of the RPCI-11 Male Human BAC Library has been performed in our laboratory over the past budget period. The RPCI-11 library characteristics are summarized in the table below.

The RPCI-11 Human Male BAC Library

DNA source: anonymous male donor

Cloning vector: pBACe3.6 (segment-1-4) and pTARBAC1 (segment-5)

Segment	Cloning enzyme	Total clones	Number of plates	Non-insert clones	Insert size (kb)	Genomic representation
1	*Eco*RI	108,499	288	1.7%	164	5.8
2	*Eco*RI	109,496	288	0.5%	168	6.0
3	*Eco*RI	109,657	288	1.0%	181	6.7
4	*Eco*RI	109,382	288	1.0%	183	6.8
5	*Mbo*I	106,763	288	0.5%	195	6.9
Total		543,797	1,440	0.8%	178	32.2

Genomic Representation

The genomic representation of the library is 32-fold equivalents based upon an estimation of average insert size and the total number of clones. In order to confirm the genomic representation, the library was screened through colony hybridization using 46 different probes. To do this, overlapping probes were designed from randomly chosen chromosome 5, 19 and 21 markers. The hybridization positive clones were confirmed using restriction enzyme fingerprinting and Southern hybridization. The screening results are summarized in the table below. The segments 1-4 were constructed using *Eco*RI partially digested DNA, whereas the segment 5 was constructed using *Mbo*I partially digested DNA. Therefore, the screening results were divided into two parts that correspond to segment 1-4 and segment 5. A total of 1,275 clones and a total

DISCLAIMER

of 281 clones were screened using the 46 mixed probes as a result of the first step screening against segment 1-4 and segment 5, respectively. Out of these, 1,076 clones and 272 clones were assigned to each probe and confirmed to be part of 46 single-marker contigs. Identical 46 clones were positive using two independent markers, D5S2032, D5S423 from segment 1-4. Also, 7 clones from segment 5 were commonly positive using these probes. Two additional positive clones were observed using D5S2032 marker from segment 5. These markers are localized at 168.5 cM and 169.3 cM from the top of chromosome 5 linkage group, respectively. A physical distance of these markers might be very close. It might be conceivable that these positive clones might be isolated from chromosome duplicated regions, since the number of positive clones are almost twice as many as the expected genomic representation. Six out of 46 clones were randomly chosen and used as probes for the Fluorescence in situ hybridization (FISH). These results clearly showed that these clones were isolated from a single region on chromosome 5 under a resolution of FISH. Therefore, the determination of genome redundancy based on marker representation excludes the D5S423 marker in the calculation. The average genome redundancy of the library was determined to be 23.9 per marker from segment 1-4 and 6.0 per marker from segment 5. This result is closely consistent with the genome coverage estimated from the total number of clones and the average insert size.

Chromosome 5			Chromosome 19		
Markers	No. of Positives		Markers	No. of Positives	
	Segment 1-4	Segment 5		Segment 1-4	Segment 5
D5S417	27	11	D19S221	22	2
D5S406	26	9	D19S411	31	10
D5S635	26	7	D19S425	21	9
D5S676	28	7	D19S220	18	2
D5S1986	37	3	D19S422	14	8
D5S426	25	5	D19S219	21	3
D5S634	34	5	D19S412	22	5
D5S2076	21	5	D19S596	16	6
D5S407	24	13	D19S214	21	8
D5S491	25	8			
D5S2028	31	5	Chromosome 21		
D5S2029	25	10	Markers	No. of Positives	
D5S456	24	2	D21S1904	23	7
D5S505	24	2	D21S1911	28	3
D5S2065	20	10	D21S262	26	6
D5S657	27	4	D21S1252	13	2
D5S2017	23	7	D21S1893	27	10
D5S2090	31	5	D21S1260	22	3
D5S673	16	2			
D5S487	18	8			
D5S412	29	5			
D5S403	15	13			
D5S2066	17	4			
D5S2032	46	9			
D5S672	24	3			
D5S413	15	2			
D5S496	24	6			

D5S2030	12	7			
D5S1987	31	5			
D5S1991	26	6			

Actual genome redundancy of RPCI-11 library using 45 independent markers.

Total 45 markers (30 from chromosome 5, 9 from chromosome 19, and 6 from chromosome 21) have been applied to screen the RPCI-11 libraries. Overall genome redundancy has been estimated at 25.3-fold from the average insert size and the number of clones. Average genome redundancy of RPCI-11 was determined to be 23.9 by screening the library using the 45 markers.

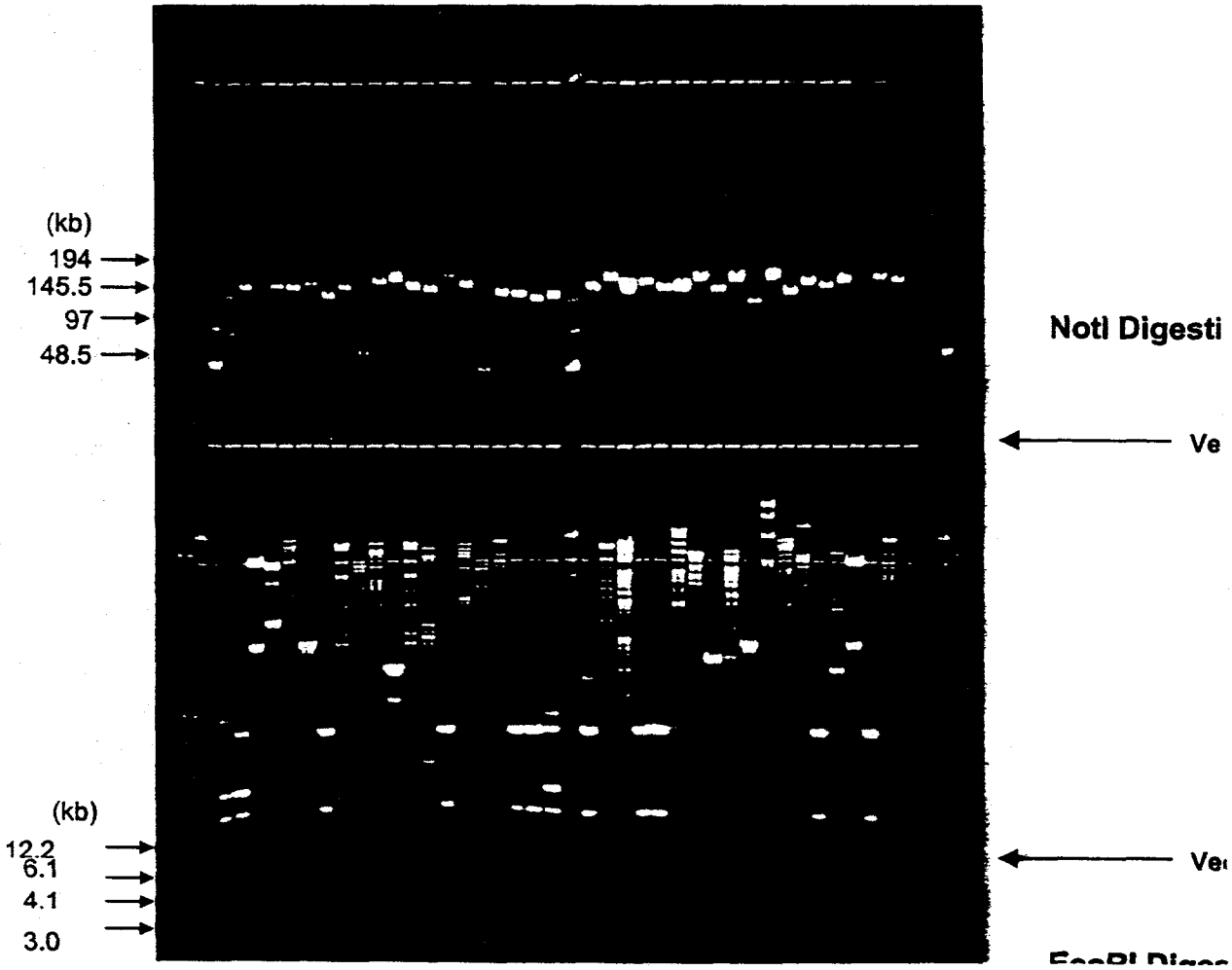
Randomness of *Eco*RI and *Mbo*I partial digestion

The ultimate usefulness of any BAC library greatly depends on how well it faithfully represents the genome. The clone collection should be isolated from the entire genome in a random fashion and retain the original genomic sequences without cloning artifacts, such as chimeras and rearrangements. The most conclusive method to assess these aspects of library quality would be by directly comparing the cloned DNA with the source genomic DNA. However, these procedures are not realistic and practical. We have designed experimentation to show clonal reliability of the BAC clones contained in the library. It is conceivable that independent cloning events can cause different cloning artifacts, these are manifested by clones in the library that contain and maintain these artifacts. Therefore we have designed an alternative strategy that compares the cloned DNA with the other cloned DNA derived from the same region. A 3.5-Mb STS-based PAC contig map was previously constructed in our laboratory using 20-fold genome equivalent human PAC libraries in Alzheimer disease (AD3) region on human chromosome 14q24.3 (Wu et al., unpublished data). A 1.5-Mb region containing 205 STS markers in which presenilin-1 gene exists has been chosen to characterize the RPCI-11 BAC library. We have reconstructed a 1.5-Mb BAC contig in this region. A total of 121 clones from segment 1-4 and 48 clones from segment 5 were localized in this region and all the clone ends were sequenced. Two different restriction enzymes (*Eco*RI for segment 1-4 and *Mbo*I for segment 5) were used for construction of the library. We compared each sequence to see the randomness of *Eco*RI and *Mbo*I partial digestion using a cross-match program. We could not find any identical end sequences among 48 clones isolated from segment 5. We examined randomness of the *Eco*RI partial digestion introducing the *Poisson* distribution ($P(X = k) = e^{-\mu}\mu^k/k!$). *Eco*RI recognizes the GAATTC sequence. The average base compositions of human genome for A, C, G, and T have been reported to be 0.288, 0.206, 0.213, and 0.293 respectively. The average length per *Eco*RI fragment in the human genome is calculated to be 3,200-bp. Therefore, a 1.5-Mb contig with 121 BACs from segment 1-4 can be assessed to contain 469 possible *Eco*RI cutting sites with 0.2580 BACs per *Eco*RI site. In these case, the *Poisson* variable (μ) is 0.2580. We found 25 cases that two different clones start from a same *Eco*RI site to a same direction. The expected frequency (P) corresponding to the two rare events of the same BAC end sequence in the same directions will be 0.02571351. Considering each BAC has two ends, the total number of pairs from both directions (left and right) of this case will be 24.1 (= 2 x 0.02571351 x 469) within the contig. We also found 3 cases that three different clones start from a same *Eco*RI site to a same direction. The expected frequency (P) corresponding to the three rare events of the same BAC end sequence in the same directions will be 0.00221136. Considering each BAC has two ends, the total number of pairs from both directions (left and right) of this case will be 2.1 (= 2 x 0.00221136 x 469) within the contig. These expected numbers are in close agreement with our results (24.1 vs. 25 in

two overlapping ends and 2.1 vs. 3 in three overlapping ends). Thus the BAC library appears to have been generated from random *Eco*RI sites.

Alpha-satellite content analysis of the RPCI-11 Human BAC Library

It is well established that the human genome contains repetitive sequence elements based on the 340 bp EcoRI dimeric satellite sequence or EcoRI periodicity of 680 bp. These repeat units are commonly referred to as alpha-satellites. Since *Eco*RI and *Eco*RI methylase were used to perform the partial digestion of the source human DNA in the construction of Segments 1-4 of the RPCI-11 library, it is conceivable that the periodic regions of *Eco*RI sites may cause preferential cloning events, thus permitting over representation of these regions in the library. Prior to picking all of the transformed clones, a total of 18,432 clones were picked into 48 384-well plates that corresponds to plate number 1 through 48 in the actual RPCI-11 library. These clones were gridded onto nylon membranes as low-density (6 plates per 22 x 22 cm membrane) and high density (48 plates per 22 x 22 cm membrane) screening filters. We identified alpha satellite positive clones by colony hybridization using an alpha-satellite specific PCR product as a probe. Two hundred-nine BAC clones (1.1%) showed positive signals out of 18,432 clones. The result indicates that the library does not over-represent the centromeric regions, since the human centromeric region is present at 10% level in the human genome. Out of these positives, 37 clones were analyzed by fingerprinting and pulsed-field gel electrophoresis after digestion with *Eco*RI and *Not*I restriction enzyme, respectively (see figure below). We could not find a BAC clone containing the 340 bp *Eco*RI repeat elements in the 37 clones. However, we found other alpha-satellite units containing different size of the periodic regions of *Eco*RI sites. Eighteen out of 37 clones contained the periodic regions of EcoRI sites that were observed as condensed bands in specific sizes after digestion with *Eco*RI (see figure below). The other 19 clones showed random *Eco*RI digestion pattern. Fingerprinting analysis of 10 single colonies from each clone revealed that 22 out of 37 clones were inherently unstable (data not shown).



Pulsed-field electrophoretic analysis of alpha-satellite repeat elements in the RPCI-11 Human BAC Library

Chimerism and clone rearrangements in the RPCI-11 Library

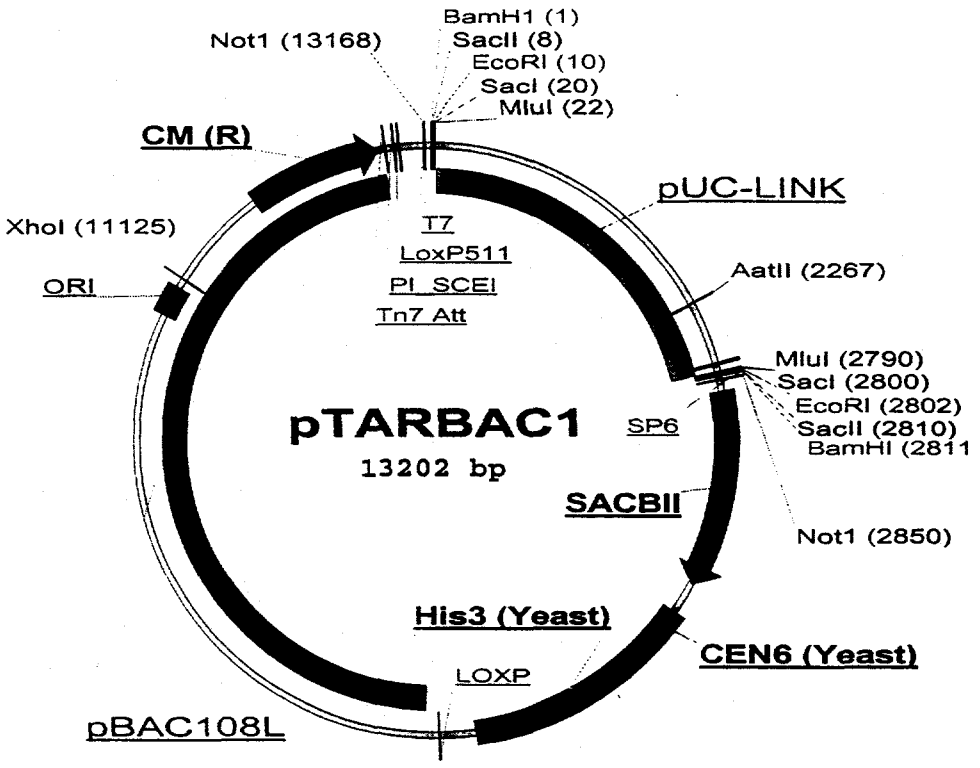
Unlike the YAC cloning system where homologous recombination in the yeast cells results in chimeric clones, co-ligation of two (or more) insert DNA molecules presents the most probable cause of chimerism in the BAC cloning system. Data accumulated through the use of BAC clones for high throughput physical mapping and genomic sequencing projects has indicated little evidence for the presence of chimeric clones in the RPCI-11 Human BAC Library. The use of FISH analysis to detect chimeric BAC clones is less useful due to the fact that the low level of chimerism is difficult to determine in relation to the "background" of apparent chimeric signals due to hybridization to duplicated chromosomal regions. It is also very difficult to detect chimeric clones composed predominately of one large DNA fragment fused to a smaller DNA molecule by FISH due to lacking or low signal corresponding to the small molecular weight portion of the chimera. We have designed a more sensitive way to identify the chimeric clones. We have constructed the 1.5 Mb chromosome 14 contig and examined whether the both ends of BAC clones are localized in the other contiguous clone members. Overgo probes were designed from the BAC end sequences to generate new markers. A total of 121 overgos derived from BAC ends were successfully localized in the contig without any inconsistency. Both ends from 52 clones were mapped back to the other member of contiguous clones thus these clones are not chimeric. The Washington University Genome Sequencing Center has been sequencing the 3.5-Mb PAC contig. All the end sequences were searched against the genomic sequence. An additional 34 clones were conclusively shown to be non-chimeric.

Single colony isolation was performed from all the clones in the contig. BAC DNA was isolated from five independent colonies of each clone and analyzed by pulsed-field gel electrophoresis after digestion with *Not*I. Unlike the alpha satellite clones, we could not observe major rearrangements comparing the insert size from each colony. Another 168 STS markers that were mostly derived from PAC clone ends were localized in the contig. Based on approximate contig size of 1.5 Mb, the average marker distances are calculated to be 5.2-kb. Since most of the markers are derived from clone-ends, it was possible to determine a definite linear order for all markers using the SEGMAP program. This permitted the establishment of consistent colony hybridization or PCR patterns. In other words, if a BAC or PAC clone is positive for two distant markers, then it should also be hybridization positive for all markers internal to the distant markers. A clone deletion would be indicated by a negative hybridization or PCR result for one or more internal markers. Using all of the clones and all of the 289 markers in colony hybridization and PCR, we have not found a single inconsistent result. This indicates the absence of large deletions spanning more than one marker interval. However, these screening results do not exclude deletions much smaller than the average 5kb marker intervals. In addition to this, fingerprinting analysis has been applied 169 clones within the contig to detect rearrangements. The fingerprinting is based on the *Eco*RI restriction fragment patterns from duplicate single-colony isolates for each clone. Clonal rearrangements can be detected as differences in the duplicate fingerprints resulting from clonal heterogeneity. Rearrangements can also be detected by comparing different clones from the same contig for consistent fragment patterns. Thirteen small rearrangements were detected in the contigs as heterogeneity between duplicate sub colonies. Five additional fingerprint inconsistencies were found within single clones by

comparing them to all their corresponding overlapping clones. All of the fingerprint inconsistencies affected only one or two fragments per clone. In summary, 18 clones out of 169 were found to have rearrangements due to alterations to a single genomic fragment during or after the cloning process.

pTARBAC Vector Development

To make the process of Transformation Associated Recombination (TAR) rescue of genomic DNA more universally applicable, we have taken a new strategy that creates large-insert BAC/YAC shuttle libraries which provides unlimited resources for TAR cloning. Several hybrid BAC/YAC vectors (pTARBAC1, see figure below, pTARBAC2, and pTARBAC4) were constructed by addition of a yeast centromere (CEN6) and a yeast-selectable marker (his3) to our earlier BAC vector, pBACe3.6 . Combination of the presence of ARS sequences in the insert and the CEN6 element and His3 selection marker in the BAC vector provides the capability of transferring the BACs into yeast and propagating them as yeast artificial chromosomes (YACs). However, by treating the BACs with a restriction enzyme (e.g. *Eco*RI) which lacks corresponding sites in the TARBAC vector, most of the insert sequences in the BACs can be deleted, including the contained ARS sequences. Such *Eco*RI-deleted BAC clones have (unique) genomic sequences at the two ends of a hybrid BAC/YAC vector which can be used as "hooks" for homologous recombination. Hence, most of the BAC clones in a TARBAC library can be used to generate TAR-rescue vectors to (re-)clone genomic segments from different haplotypes or from related species. We have tested the whole system to re-isolate the deleted sequences by co-transformation of deleted BAC DNA with genomic Trypanosome DNA. The less-complex genomes of unicellular eukaryotes have been used to model future work with mammalian TARBAC libraries. Most (35 out of 40) of the Trypanosome BACs, with 107 kb average inserts, transform yeast at high efficiency, indicating the presence of ARS elements in most of the genomic insert fragments. Most of the insert sequences in the Trypanosome BACs can be deleted by treating the BACs with a restriction enzyme that lacks corresponding sites in the TARBAC vector. Such deleted BAC clones are functionally similar to the previous TAR-rescue vectors because they have (unique) genomic sequences at the ends of a hybrid BAC/YAC vector. We have successfully explored the re-isolation of the deleted sequences by co-transformation of deleted BAC DNA with genomic Trypanosome DNA. The fidelity of TAR-cloning for the specific TbSir2 gene locus turned out to be 60% identity for a 50 kb, and 30% for a 100 kb genomic fragment. TAR-cloning of the large genomic fragments from randomly selected BAC clones revealed 17% identity for a 150 kb fragment, and 8% for a 180 kb fragment. The sequencing results prove that homologous recombination can take place at many potential sites between a "TAR-cloning" vector and genomic fragments. Our approach provides a new strategy to identify targeted genomic region and can greatly facilitate a large-scale analysis for positional cloning.



RPCI-24 male C57BL/6J Mouse BAC Library

We have completed the construction of the RPCI-24 male C57BL/6J mouse BAC library and plan to have this clone collection in distribution by late June – early July 2000. The library was constructed from isolated genomic kidney DNA that was partially digested with MboI. This will allow for a different representation of genomic fragments than that present in the RPCI-23 C57BL/6J female mouse BAC library which was constructed using EcoRI-digested DNA. The library will be divided into 2 equal segments of 288 (384 well) plate, each representing approximately 5-6 fold coverage of the mouse genome. We anticipate that the library will be comprised of clones that contain recombinant inserts that average 155-160 Kbp. Less that 5% of the clones will be non-recombinant. This library will be extensively characterized in the next year as to it's fidelity in representing the mouse genome by mapping a set of know mouse markers to high density filters containing all the clones in the library.

Our plans are to replicate copies of this library and first distribute to the laboratories of Marco Marra in Vancouver, BC, Canada and to TIGR for the mouse clone fingerprinting and end-sequencing efforts current underway.

Resource Distribution

Distribution of the current human and mouse BAC libraries has continued during this budget year. We have endeavored to improve usage of the human and the mouse BAC libraries by providing screening resources on a cost recovery basis. All reimbursements are deposited in a

academic account in the domain of Health Research Inc. (a non-profit New York State organization responsible for grant accounting at New York State Institutes, including Roswell Park Cancer Institute) and are used to pay for supplies used in library replication, high density filter production, equipment maintenance, minor new equipment and labor. This arrangement has trtansfered to the Children's Hospital of Oakland Research Institute (CHORI) in late January, 2000. The same cost recovery mechanism is employed at CHORI with all funds generated used to support the process. The two main requirements attached to users of the libraries are: 1) Maintain and use the library/clone nomenclature in all publications and databases. 2) Secondary library distribution is not permitted without explicit approval, but individual clones of clone collections can be distributed as long as the clone names and library origins are acknowledged and remain permanently attached with the clones. Distribution of library resources occurs in three forms: 1) distribution of library culture plate arrays, 2) distribution of high density hybridization filters, and 3) distribution of individual "positive" clones. We have developed a WWW page (http://www.chori.org/bacpac) to inform all users (and potential users) on the availability and reimbursement costs of our library resources. In addition, there is information on the library construction, characterization, vector maps and sequences, as well as, our DNA prep method used for the clones from the various libraries offered.

Distribution of library culture plate arrays: By taking care of the library plate distribution from our laboratory, we have been able to facilitate rapid transfer of the library copies and maintain optimal library quality. We have been able to replicate library copies within one to four weeks after receipt of the request. To date we have replicated 44 copies of RPCI-11 Segments 1&2 and have distributed these to 36 research centers and commercial resource center throughout the world. 30 replicas of the RPCI-11 Segments 3&4 have been generated and we have distributed these to 25 research centers. Six replicas of RPCI-13 Segments 1-4 have been made and three of these have been distributed to the major human genome sequencing centers. Due to the complexities of relocating our laboratory from RPCI to CHORI, we have found it necessary to temporarily suspend the replication of all libraries until the time that we have completed the relocation. We anticipate replicating both RPCI-14 and RPCI-15 soon after our relocation to CHORI and make these libraries immediately available to those researchers requesting them.

Distribution of high density hybridization filters: We are providing users with library screening filters that have been produced in our laboratory on the "Q-Bot" and the "BioGrid" robotic gridding systems. The clones are gridded onto the filters in a 4x4 clone duplicate array such that a positive clone will appear as two distinct spots in one of eight possible "vector" patterns on the autorad image of the probed filter. Each filter contains over 36,000 clones (>18,000 individual clones). We have found that the end users of these filters have no difficulty interpreting the data from this type of filter grid and locating the positive clones in the original library plates. These filters are provided on a cost recovery basis of $125.00 per filter. To date, over 20,000 of the high-density filters have been produced and distributed to laboratories worldwide.

Distribution of individual positive clones: During the past two years we have begun to distribute individual clones to small scale users who do not have direct access to the library plates. The clones are provided on a cost recovery basis of $12.50 per clone. Users have obtained high-density hybridization filters from our lab and perform the screening in their own laboratories. Upon reading the screening results, they contact us with the locations of their positives clones and we pick, culture and ship these clones to them either as glycerol stocks (US labs) or LB stabs (overseas shipment). We have been able to ship these clones within two to four days after receipt of the request. To date, we have supplied over 10,000 RPCI BAC clones directly to the end user. We are finding that this is an effective method to increase the use of our libraries while providing access to laboratories that due to budget constraints, would not be able to obtain the complete clone collection.)

Significance

The construction, characterization, and distribution of human BAC libraries have been, and remains, of critical importance for the successful completion of the Human Genome Project. Our efforts have concentrated in generating the highest quality libraries possible by developing improved cloning methodologies and incorporating these methodologies into our production-scale laboratory procedures. Implicit in our scientific mandate is the verification that the BAC libraries that we generate faithfully represent the genome from which the DNA was obtained. To this end, we have undertaken an extensive characterization of the RPCI-11 Human BAC Library to determine any potential failings of the current BAC cloning methodologies and to delineate genomic regions lacking in coverage, and to develop strategies to address these areas in future libraries. This work will be essential in order to provide sequencing template for "difficult" regions required for completing the human genome.

The distribution of these libraries resources in a timely, and more importantly, in a high quality fashion to the scientific community is a crucial function of our laboratory. BACPAC Resources functions as the distribution arm of our laboratory by incorporating a cost-recovery mechanism that serves the scientific community.

The development of improved cloning technologies in our laboratory and the publication of these methods has enabled other laboratories throughout the world to initiate BAC library construction for the genomes of many other organisms. This facilitates the gene discovery process for the genome community in general.

The pilot work on Trypanosoma Transformation Associated Recombination should set the stage for the use of our latest human and mouse BAC library clones (RPCI-14, RPCI-15, and RPCI-24) as DNA sources that will enable the end user to "re-clone" the same genomic DNA segments from different haplotypes and related species.

Publications

1. Osoegawa, K., Woon, P.-Y., Zhao, B., Frengen, E., Tateno, M., Catanese, J. J., and de Jong, P. J. An Improved Approach for Construction of Bacterial Artificial Chromosome Libraries. *Genomics* 52, 1-8 (1998)
2. Frengen, E., Weichenhan, D., Zhao, B., Osoegawa, K., van Geel, M., and de Jong, P. J. A modular positive selection bacterial artificial chromosome vector with multiple cloning sites. *Genomics* 58: 250-253 (1999)
3. Li, R., Mignot, E., Faraco, J., Kodatani, H., Catanese, J., Zhao, B., Lin, X., Hinton, L., Ostrander, E.A., Patterson, D.F. and de Jong, P.J. (1999). Construction and characterization of an eight fold redundant dog genomic bacterial artificial chromosome library. Genomics (in press).
4. Osoegawa, K., Frenegen, E., Ioannou, P. A., and de Jong, P. J. Construction of Bacterial Artificial Chromosome (BAC/PAC) Libraries. In *Current Protocols in Human Genetics* (N. C. Dracopoli, J. L. Haines, B. R. Korf, D. T. Moir, C. C. Morton, C.E. Seidman, J.G. Seidman, and D.R. Smith, Eds), pp 5.15.1-5.15.31. John Wiley & Sons, New York. (1999)
5. Osoegawa, K., Tateno, M., Woon, P.-Y., Frengen, E., Mammoser, A. G., Catanese, J. J., Hayashizaki, Y.,and de Jong, P. J. Bacterial Artificial Chromosome Libraries for Mouse Sequencing and Functional Analysis. *Genome Research*, In press.