

Anchoring of rice BAC clones to the rice genetic map *in silico*

Qiaoping Yuan, Feng Liang, Joseph Hsiao, Victoria Zismann, Maria-Ines Benito, John Quackenbush, Rod Wing¹ and Robin Buell*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA and ¹Clemson University Genomics Institute, 100 Jordan Hall, Clemson, SC 29634-5708, USA

Received May 1, 2000; Revised and Accepted July 27, 2000

ABSTRACT

A wealth of molecular resources have been developed for rice genomics, including dense genetic maps, expressed sequence tags (ESTs), yeast artificial chromosome maps, bacterial artificial chromosome (BAC) libraries and BAC end sequence databases. Integration of genetic and physical maps involves labor-intensive empirical experiments. To accelerate the integration of the bacterial clone resources with the genetic map for the International Rice Genome Sequencing Project, we cleaned and filtered the available EST and BAC end sequences for repetitive sequences and then searched all available rice genetic markers with our filtered databases. We identified 418 genetic markers that aligned with at least one BAC end sequence with >95% sequence identity, providing a set of large insert clones with an average separation of 1 Mb that can serve as nucleation points for the sequencing phase of the International Rice Genome Sequencing Project.

INTRODUCTION

Rice, *Oryza sativa*, is a member of the Gramineae family that includes wheat, barley, maize, sorghum, millet, sugarcane and oats. The estimated size of the haploid rice genome is significantly smaller than that of other cereal family members, 430 Mb as compared to 2500 Mb for maize, 4873 Mb for barley and 15 966 Mb for wheat (1). Because of its small genome size, and in recognition of its importance as the world's major food crop, rice has been developed as a model organism for the grasses and is currently the focus of an International Genome Sequencing effort using a bacterial artificial chromosome/P1 artificial chromosome (BAC/PAC)-based shotgun approach (<http://rgp.dna.affrc.go.jp/seqcollab.html>). Extensive molecular resources have been developed to assist in completion of the rice genome. These include a dense genetic map (2; <http://ars-genome.cornell.edu/rice>), a rice expressed sequence tag (EST) database (3,4), the TIGR Rice Gene Index (5; <http://www.tigr.org/tdb/tgi.html>), a yeast artificial chromosome (YAC) physical map (6; <http://rgp.dna.affrc.go.jp/publicdata/physicalmap99/yacall.html>), a P1 artificial chromosome

(PAC) physical map (<http://rgp.dna.affrc.go.jp/genomicdata/seqstrategy/seq-strategy.html>), two BAC libraries and over 80 000 BAC end sequences (http://www.genome.clemson.edu/projects/rice/rice_bac_end/index.html).

In a clone-by-clone sequencing strategy, such as that adopted for the International Rice Genome Sequencing Project, a series of anchored seed BAC and PAC clones are chosen as initial sequencing targets. Upon completion of the sequence of each clone, new, minimally overlapping clones are selected to extend the sequence. The initial selection of well-spaced, anchored seed clones, integrated with the genetic and physical maps, is crucial for the efficient completion of the project, particularly for directing and minimizing redundancy in the final closure phase.

The identification of an anchored set of seed clones is generally extremely labor-intensive, requiring the development of a validated set of genetic markers, hybridization to colony filters and the confirmation of selected clones by Southern hybridization or PCR amplification. As an alternative, we developed an approach in which more than 80 000 BAC end sequences were screened against the high-density genetic markers to identify and anchor BAC clones to the genetic map. In order to achieve this, we had to develop strategies to overcome a number of obstacles. First, both the BAC end sequences and the ESTs, which comprise the majority of the markers linked to the genetic map, are single-pass sequences. The relatively short lengths of these sequences and the errors inherent in such data require the development of stringent overlap criteria to assure unique, high confidence map assignments. Second, many ESTs contain stretches of poly(A) that can produce false hits to homopolymer stretches in the BAC end sequences. Third, the rice genome, like those of other higher eukaryotes, contains repetitive DNA sequence intermixed with coding sequence, which confounds interpretation of alignments between sequences.

In rice it is estimated that 50% of the rice genome is comprised of repetitive sequence (7). Experimental and computational genome analyses indicate rice repetitive sequences are found in tandemly repeated microsatellites (1–7 bp), longer and more complex minisatellite repeating units (up to 40 bp) and satellite DNAs with lengths of 140–360 bp. Mobile DNA sequences, such as transposons and retrotransposons, make up a high proportion of plant middle repetitive DNA. Retroelements are divided into mobile sequences with long

*To whom correspondence should be addressed. Tel: +1 301 838 3558; Fax: +1 301 838 0208; Email: rbuell@tigr.org

terminal repeats (LTRs) and non-LTR retrotransposons (LINEs, long interspersed nuclear elements) and the related SINEs (short interspersed nuclear elements). Plant genomes may also contain solo-LTRs, miniature inverted-repeat transposable elements (MITEs) and virus-like sequences. Analysis of rice centromeric sequences indicates that the centromere is a complex region with stretches of tandemly repeated sequences intermixed with middle repetitive elements. At least seven centromeric repetitive DNA families have been described in the rice centromere, six middle repetitive sequences (50–300 copies) and one tandem 168 bp repeat, RCS2, that is unique to rice centromeres (8). Rice telomeric DNA consists of conserved 7 bp repeats (TTTAGGG) (9,10). A final class of repetitive sequences found in all eukaryotic genomes is the 18S–5.8S–25S and 5S rRNA gene loci, clustered at a small number of sites, that encode the structural RNA components of ribosomes. All of these repetitive sequences can obscure the presence of real alignments between a marker and a BAC.

To address these problems, we devised a series of sequence filters and a screening process that has allowed us to generate high confidence links between the genetic markers and the BAC end sequences. With these improvements in the marker and BAC end sequence databases, we were able to anchor 418 markers to a collection of BAC clones. We were able to validate the robustness of our data by experimentally verifying the anchoring of candidate BACs on chromosome 10.

MATERIALS AND METHODS

Markers used in this study

The 2152 markers used in this study were obtained from the Rice Genome Program (RGP; <http://rgp.dna.affrc.go.jp/publicdata/geneticmap98/geneticmap98.html>) and the Cornell Rice Genes database (<http://ars-genome.cornell.edu/rice>) and are summarized in Table 1. The RGP markers (1474) are composed primarily of rice ESTs but also contain rice genomic sequences. The markers from the Rice Genes database (678) were derived from various *O.sativa* cDNA libraries and genomic sequences. Although both the RGP and the Cornell markers have been genetically mapped, the groups used different mapping populations and consequently the maps cannot be directly integrated. We also obtained 26 markers from oat, a related cereal species, which have been placed on the rice genetic map to test whether sufficient conservation of nucleotide sequence was present between these two cereal species such that BACs could be anchored to the rice genetic map using orthologous sequences.

Molecular methods

BAC clones were obtained from the Clemson University Genomics Institute and were grown in LB medium supplemented with chloramphenicol (11). BAC DNA was isolated using a standard alkaline lysis method (11,12). Yeast artificial chromosome (YAC) clones were grown in AHC medium and total DNA was isolated from YAC clones using methods described by Matallana *et al.* (13). Primers were designed to the BAC end sequences and used to amplify YAC and BAC DNA using cycling conditions of 94°C for 30 s, 55°C for 30 s and 72°C for 30 s with a total of 35 cycles (12). Products were fractionated on a 1% agarose gel (12).

Table 1. Source and nomenclature of markers used

Marker nomenclature ^a	Description	Number of markers ^b
RGP		
R	Root cDNA	442
S	Shoot cDNA	322
	S < 10 000 etiolated shoot	155
	S > 10 000 green shoot	167
C	Callus cDNA	420
G	Genomic DNA	184
Y	YAC clone	28
L	<i>NotI</i> linking clone	63
P	RAPDs	12
Other		3
Cornell		
RG	Genomic	165
RZ	Leaf cDNA	347
Others		166

^aMarker information and accompanying sequences were obtained from the RGP web site (2; <http://rgp.dna.affrc.go.jp/publicdata/geneticmap98/geneticmap98.html>), Rice Genes (<http://ars-genome.cornell.edu/cgi-bin/WebAce/webace?db=ricegenes>) or GenBank.

^bA total of 2152 rice markers with known map positions were selected from the public rice databases and were obtained from cDNA libraries, genomic libraries, RAPD markers or YAC clones.

Computational methods

To align the marker sequences with the BAC end sequences we used FLAST, a rapid sequence comparison program based on DDS (14). FLAST first concatenates the markers into a single query sequence with a non-alphanumeric spacing character separating the individual input sequences. This new query sequence is then searched against the BAC end sequence database using a hashing algorithm to identify high-scoring segment matches. High scoring hits are then extended in each direction until the sequence similarity score falls below a threshold or one of the separation characters is encountered. Segment pairs are then combined into chains, where adjacent elements in the chain can be derived from different reading frames or adjacent exons, making FLAST tolerant of frameshifts in EST-derived markers as well as introns in genomic sequence. As FLAST computes high-scoring segment pairs in a batch fashion, it runs several times faster than other sequence comparison programs such as BLASTN without sacrificing accuracy. FLAST runs under the Unix operating system and is available free of charge to academic and non-profit research organizations (see <http://www.tigr.org/softlab/> for additional information).

RESULTS

Extension of marker length using the TIGR rice gene index

The TIGR Gene Indices provide an analysis of the publicly available EST and gene sequence data in order to enumerate the genes and to provide likely consensus sequences for the

underlying transcripts (5). A total of 43 095 rice EST sequences were downloaded from dbEST and trimmed to remove vectors, poly(A/T) tails, adaptor sequences and contaminating bacterial sequences. A total of 2279 rice gene sequences were also included: 1804 transcripts (NP sequences) passed through Entrez from CDS and CDS-join features in GenBank records and 475 curated expressed transcript (ET) sequences from the TIGR EGAD database (<http://www.tigr.org/tdb/egad/egad.html>). These sequences were clustered by comparing all pairs using WU-BLAST (15; <http://blast.wustl.edu>) and collecting sequences with $\geq 95\%$ identity over regions ≥ 40 bp in length, with unmatched overhangs < 20 bp. The sequences comprising each cluster were assembled using CAP3 (16) to produce tentative consensus (TC) sequences. The TCs provide a high confidence consensus to represent each transcript that is generally longer than the individual ESTs that comprise it. A TC containing a known gene was assigned the function of that gene; TCs without assigned functions were searched using DPS (14) against a non-redundant protein database; high-scoring hits were assigned a putative function. The *O.sativa* Gene Index (OsGI, Release 3; <http://www.tigr.org/tdb/ogi>) contains a total of 20 336 unique rice sequences (either TCs, singleton ETs or singleton ESTs) reducing the redundancy in the rice EST database by 55%.

We then searched the RGP and Cornell rice marker data set against OsGI. We were able to identify 1540 markers (1185 RGP markers and 355 Cornell markers) that were represented by a TC. Through assignment of the markers to TCs, we were able to extend the average length of the mapped sequences by 70 bases, an average increase in length of 19.3%.

Cleaning and trimming the marker and BAC end sequence data sets

BAC end sequences were trimmed to remove low quality sequence regions using a 2% probability of error as a cutoff; contaminating vector sequences were also removed. From an initial set of 105 197 BAC end sequences, 83 014 were high quality sequences with an average clear range of 676 bases. Of these, 58 679 BAC end sequences were from the *HindIII* library and 24 335 sequences were from the *EcoRI* library. The TC and other marker sequences were also trimmed to remove low quality and homopolymer sequences. A recursive trimming process was implemented to remove low quality sequences using a cutoff criterion of < 1 unidentifiable nucleotide (N) every 10 nt. Poly(A/T), defined as > 5 A/T per 10 nt, were trimmed from the terminal segments of the sequences.

Construction of a rice repeat database and filtering of repetitive BAC end sequences

We searched rice sequences from GenBank for minisatellite sequences, mobile elements, rDNA, centromeric repeat sequences and telomeric repeat sequences and generated a curated Rice Repeat Database. This database can be accessed for BLAST searches through the TIGR Rice Genome Project web site (<http://www.tigr.org/tdb/rice>). BAC end sequences were searched against the Rice Repeat Database using FLAST and those containing high-scoring hits were eliminated from subsequent analysis. A total of 2688 BAC end sequences had $\geq 95\%$ identity with entries in the TIGR Rice Repeat Database. A majority of these were matches to transposon or transposon-like sequences; centromeric and telomeric repeats were the second most abundant. These results are summarized in Table 2.

As a majority of the genetic markers for rice are derived from ESTs rather than random genomic DNA segments, it is unlikely that a substantial fraction contain repetitive DNA (which is typically associated with non-coding regions). However, it is possible that the rice marker set could contain additional repetitive sequences that we had not previously curated. Therefore, we searched our repetitive sequence-depleted BAC end sequence database with our set of cleaned markers to identify additional repetitive sequences within the rice BAC end data set. If two or more BAC end sequences hit a single marker, these were considered candidate repetitive sequences. A total of 183 BAC end sequences were identified; these were searched against GenBank to further curate the nature of repeat. If the sequence aligned with a class of sequences known to be repetitive, we added that sequence to the TIGR Rice Repeat Database. After these two phases of repeat filtering, the final BAC end data set contained 80 143 sequences. The search method we employed may not provide an exhaustive identification of repeat sequences within the BAC end sequence database. However, this significant reduction in representation of repetitive sequences, in conjunction with the use of high stringency cutoff criteria in subsequent alignments, will reduce the occurrence of false associations in our alignments between the markers and BAC end sequences.

Alignment of the cleaned markers with the non-repetitive rice BAC end data set

Using FLAST, we searched the filtered BAC end sequence data set with the cleaned marker sequences and where possible, the corresponding TC. We searched the original markers (without the corresponding TCs) using a stringent cutoff of $\geq 95\%$ with a minimum of 78 bases of overlap and identified

Table 2. Number of BAC end sequences with matches to the TIGR Rice Repeat Database

BAC library	Number of BAC end sequences with matches to TIGR Rice Repeat Database ^a			
	Telomere/centromere	Transposon/transposon-like	rDNA	Others
<i>HindIII</i>	468 (0.80%)	1040 (1.77%)	27 (0.05%)	110 (0.19%)
<i>EcoRI</i>	320 (1.31%)	371 (1.52%)	266 (1.09%)	269 (1.10%)
Total	788 (0.95%)	1411 (1.70%)	293 (0.35%)	379 (0.46%)

^aAlignments from the BAC end database to the curated Rice Repeat Database were scored as a match if there was $> 95\%$ identity over a minimum of 78 bases. The representation of the matches within the *EcoRI* and the *HindIII* BAC libraries is also presented.

328 markers (out of 2152 total) that aligned with at least one BAC end sequence (Table 3). We were able to further increase the number of candidate anchored BACs by searching the BAC end sequence database with TCs for the mapped markers. This identified an additional 90 markers that aligned with at least one BAC end sequence, an overall increase of 27.4%, allowing a total of 418 mapped markers to anchored BAC clones. A complete listing of these results is available (<http://www.tigr.org/tdb/rice/mappedbacends/>). Although our lower limit for candidate alignments was $\geq 95\%$ identity over 78 nt, our alignments were much more robust. Alignments between the marker sequences and the candidate BAC end sequences were, on average, 98.58% identical over 215.9 bases. The average identity between the alignments of the TC sequences and their corresponding BAC end sequences was slightly better, 98.95% over 265.9 bases, reflecting the greater sequence length and fidelity of the TC assemblies. On average there were 1.4 BAC end sequences per marker, with a maximum of seven BAC end sequences aligning with one marker. No BAC end sequences were identified using the non-rice marker sequences.

Table 3. Alignment of the rice markers with the filtered rice BAC end sequence data set

Query	Number of markers that aligned with a BAC end sequence $\geq 95\%$ identity ^a
Rice markers	
Markers without a corresponding TC	69
Markers with a TC ^b	
Search using marker sequence	259
Search using TC	349
Non-rice markers	0
Total	418

^aThe markers were searched against the rice BAC end sequence data set that had been filtered to remove known rice repeat sequences. Using a cutoff of $\geq 95\%$ identity, candidate BACs were identified for 418 rice markers. No candidate BACs were identified for markers derived from oats.

^bBoth the TC and the underlying marker sequence were used to search the rice BAC end data set.

Removal of repetitive sequences in the BAC end sequences was essential for successful interpretation of the data. When the marker sequences were used to search against the unfiltered BAC end sequences, we identified an average of 3.3 BAC end sequences per marker, with one marker identifying 131 BAC end sequences. Thus, without filtering the BAC end data set for repetitive sequences, the probability of a positive alignment being due to repetitive sequences conserved throughout the genome is much greater.

Experimental verification of candidate BACs

To provide empirical evidence that our filtering and alignment tools are robust, we selected markers from chromosome 10 to validate our *in silico* alignments. Due to the availability of partial data sets, we used two complementary experimental

approaches. First, the two BAC libraries used in this study have been fingerprinted and overlapping clones can be clustered into contigs based on shared restriction fragment patterns (17; http://www.genome.clemson.edu/tools/contig_viewer/index.html). As more than 66 BACs have been anchored to chromosome 10 as part of the International Rice Genome Sequencing Project (<http://www.tigr.org/tdb/rice>, <http://www.genome.clemson.edu/cgi-bin/status.pl>), we can verify whether the candidate BACs group into the same fingerprint contigs as BACs that have been validated and selected for sequencing on chromosome 10. In our second approach, we verified the physical map location of the candidate BACs through PCR amplification of BAC end sequences on the YAC clones that comprise a minimal tile for chromosome 10 (6; <http://rgp.dna.affrc.go.jp/publicdata/physicalmap99/YACall.html>). Both of these analyses have constraints, including gaps in the YAC map, deletions and chimeras in the YAC clones, absence of a BAC in the fingerprint database and fixed assembly parameters within the fingerprint contigs. However, it was apparent that the cleaning and filtering techniques used in this study provide a robust method to identify anchored BAC clones.

Our analyses identified 20 BACs anchored by 13 markers from chromosome 10. We examined 11 BACs corresponding to 10 markers using the techniques described above and were able to experimentally verify nine BACs corresponding to eight markers (Table 4). For two other chromosome 10 markers (RZ400, R1933), we could not use either experimental approach to verify the *in silico* anchoring as neither a YAC map position nor an anchored sequence map position was available. However, for the BACs identified for the Cornell marker RZ400, clustering in the fingerprint contigs was observed. Three of the seven RZ400 candidate BACs were in contig 1108, another three BACs were in contig 781 and the remaining BAC was not present in the fingerprint database. Thus, although we could not anchor the candidate BAC clones for RZ400 to chromosome 10, the clustering of the BACs into similar fingerprint contigs is consistent with the *in silico* analyses that these BACs share common features.

DISCUSSION

Using a combination of computational tools, we were able to identify BAC clones anchored to the rice genetic map from the available marker and BAC end sequence data sets. We were able to address the low quality nature of the EST and BAC end sequences and remove the lower quality portions within these sequences using stringent cutoff parameters. We were able to enhance the marker sequences by identifying the corresponding TC within the TIGR Rice Gene Index, increasing the average length of the markers from 362 to 432 bases, an increase of 19.3%. One complication of larger eukaryotic genomes is the presence of repetitive sequences that can confound alignments between sequences. To address this problem, we created a Rice Repeat Database and used this database to remove BAC end sequences that contained repeats. From searches with the cleaned, trimmed and extended marker set against the repeat-depleted BAC end database, we were able to identify BAC end sequences corresponding to 418 mapped markers. Experimental verification of these alignments using markers from chromosome 10 revealed our computational tools and alignments to be robust.

Table 4. Summary of experimental evidence validating alignment of BAC end sequences with chromosome 10 markers

Marker ^a	cM	Candidate BAC	Confirmation with ^b	
			FPC	YAC
C913A	11	OSJNBa0016A20	+	NA, 3
C8	15.9	OSJNBa0074B12	-	+
C148	17.5	OSJNBb0052C09	-	NA, 4
R1629	22.7	OSJNBb0021C03	+	+
S1786B	24.3	OSJNBb0005A06	NA, 1	NA, 4
Y1053R	34.6	OSJNBa0077L01	+	+
G37	44.8	OSJNBb0040J03	-	+
S14155	58.1	OSJNBa0026L12	+	+
		OSJNBa0047A22	+	NA, 3
C488	58.4	OSJNBb0001N17	NA, 2	+
RZ583	71.8	OSJNBb0067B07	NA, 2	+

^aAll of the markers and their corresponding map position, with the exception of RZ583, were obtained from the RGP (<http://rgp.dna.affrc.go.jp/publicdata/geneticmap98/geneticmap98.html>). Marker RZ583 and its map position were obtained from Rice Genes (<http://ars-genome.cornell.edu/cgi-bin/WebAce/webace?db=ricegenes>). The cM positions between the RGP and the Cornell maps are not directly comparable.

^bThe alignment of the candidate BACs on the physical and sequence maps for chromosome 10 was verified through comparison of the fingerprint contig data with BACs anchored to the sequencing map and/or through PCR amplification of YAC clones as described in Materials and Methods. +, the experimental data was consistent with the *in silico* analyses; -, the experimental data did not support the placement of the candidate BAC on chromosome 10; NA, not available. Numbers following NA: 1, the marker was not anchored on current sequence map; 2, the fingerprint data was not available for this clone; 3, the experiment was not performed; 4, the primers designed to the BAC end sequence failed to amplify the correct product from the BAC clone using the amplification conditions specified in Materials and Methods.

The 80 143 BAC end sequences used in our searches comprise 54.2 Mb and represent ~11.8% of the 431 Mb rice genome (1). If the library, and the end sequences derived from it, are representative of the rest of the genome, there should be about a 12% chance of identifying any particular randomly selected sequence-based marker within the end sequence database. The 2152 previously mapped independent markers considered in our analysis spanned a total sequence length of 0.78 Mb; the total sequence length matched in the 418 markers we were able to successfully map to BACs *in silico* was 0.09 Mb or 11.5% of the total marker length. This correlation between genomic coverage and representation of the markers in the BAC end data set is consistent with our experimental results and suggests that the sequence filtering and screening protocol we developed is robust.

The rice genome has been reported to be composed of ~50% repetitive sequences. Our computational analyses identified only 3.5% of the rice BAC end data set as containing repetitive sequences. There are several reasons for this apparent discrepancy. First, we searched the rice BAC end sequence database for repeat sequences using a curated set of known rice sequences (215 sequences in total) which is not a comprehensive catalog of rice repeats. For example, simple repeats such as dinucleotide and trinucleotide repeats are not comprehensively

represented in our repeat database and as a consequence would not be identified in any alignment search of the rice BAC end sequences. Second, we used a high stringency cutoff of $\geq 95\%$ identity to highlight repetitive sequences within the BAC end dataset. We can increase by 2-fold the number of BAC end sequences defined as repetitive by reducing the cutoff from ≥ 95 to 90% identity (data not shown). Thus, as a consequence of this high stringency, only identical or nearly identical members of a repeat family were identified. Third, we used 78 nt as the minimum length in our alignments with the Rice Repeat Database, thus excluding the detection of more simple repeat sequences within the data set and partial repeats in the BAC end sequences. To comprehensively identify repeats in a genome, an unbiased search of repeated nucleotides must be performed using alternative computational programs such as MUMmer (18). Indeed, preliminary analyses of the rice BAC end sequences using the MUMmer program are consistent with the rice genome containing ~50% repetitive sequences (S.Salzberg, unpublished data).

Coupled with the limitations we employed in our computational approach to identify repetitive sequences, the choice of restriction enzyme used to generate a BAC library can influence the types of highly conserved repeat classes that are represented in the BAC end sequences derived from the library. For example, the *EcoRI* BAC library has >20-fold higher representation of rDNA repeats, 5.5-fold higher representation of other repeats and a 1.6-fold higher representation of centromeric/telomeric repeats than the *HindIII* library. The representation of *HindIII* and *EcoRI* restriction sites in these repeat classes provides an explanation for the large difference in the abundance of rDNA between the libraries. The 25S ribosomal DNA sequence (19) contains an *EcoRI* site yet no *HindIII* site and 265 *EcoRI* BAC end sequences, but no *HindIII* BAC ends, aligned ($\geq 95\%$) with this sequence. Likewise, entire classes of centromeric and telomeric or other tandem repeats may not be represented in the BAC end sequence database if the repeats do not contain the restriction enzyme site used to construct the library. For example, a search of the available BAC end sequences with the conserved (TTTAGGG)_n telomere repeat sequence did not reveal any BACs that contain this sequence.

One of the more labor-intensive parts of initiating a BAC-based sequencing project for an entire genome is the anchoring of BAC clones to the genetic map. We have demonstrated that our cleaning and filtering tools are sufficiently robust to identify candidate BACs for this purpose. Although these BACs will require further verification prior to initiating sequencing, they are an important resource for laboratories participating in the sequencing of the rice genome. These data also represent a resource for rice biologists who are positionally cloning genes of interest in rice. The large insert BAC clones anchored to the genetic map not only provide an immediate substrate for further analyses, but they also present a resource for construction of a high-resolution map in the region of interest.

ACKNOWLEDGEMENTS

The marker and YAC clones for chromosome 10 were a kind gift from Dr Takuji Sasaki of the Rice Genome Program and the Japanese Ministry of Agriculture, Forestry and Fisheries

(MAFF) Genome Research Program. Funding for the work was provided in part by a grant by the US Department of Agriculture (99-35317-8275), the National Science Foundation (DBI998282) and the US Department of Energy (DE-FG02-99ER20357). The fingerprint data for the BAC clones was funded by Novartis Agribusiness Biotechnology Research Inc.

REFERENCES

1. Arumuganathan, K. and Earle, E.D. (1991) *Plant Mol. Biol. Rep.*, **9**, 208–219.
2. Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., Yamamoto, T., Lin, S.Y., Antonio, B.A., Parco, A. *et al.* (1998) *Genetics*, **148**, 479–494.
3. Kurata, N., Nagamura, Y., Yamamoto, K., Harushima, Y., Sue, N., Wu, J., Antonio, B.A., Shomura, A., Shimizu, T., Lin, S.Y. *et al.* (1994) *Nature Genet.*, **8**, 365–372.
4. Yamamoto, K. and Sasaki, T. (1997) *Plant Mol. Biol.*, **35**, 135–144.
5. Quackenbush, J., Liang, F., Holt, I., Perte, G. and Upton, J. (2000) *Nucleic Acids Res.*, **28**, 141–145.
6. Umehara, Y. and Inagaki, A. (1995) *Mol. Breeding*, **1**, 79–89.
7. Deshpande, V.G. and Ranjekar, P.K. (1980) *Hoppe Seylers Z. Physiol. Chem.*, **361**, 1223–1233.
8. Dong, F., Miller, J.T., Jackson, S.A., Wang, G.-L., Ronald, P.C. and Jiang, J. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 8135–8140.
9. Ohmido, N. and Fukui, K. (1997) *Plant Mol. Biol.*, **35**, 963–968.
10. Wu, T., Wang, Y. and Wu, R. (1994) *Plant Mol. Biol.*, **26**, 363–375.
11. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
12. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Siedman, J.G., Smith, J.A., Struhl, K., Albright, L.M., Coen, D.M., Varki, A. and Janssen, K. (1994) *Current Protocols in Molecular Biology*. John Wiley and Sons, NY.
13. Matallana, E., Bell, C.J., Dunn, P.J., Lu, M. and Ecker, J.E. (1992) In Koncz, C., Chua, N. and Schell, J. (eds), *Methods in Arabidopsis Research*. World Scientific, Singapore, pp. 144–169.
14. Huang, X., Adams, M.D., Zhou, H. and Kertavage, A.R. (1997) *Genomics*, **46**, 37–45.
15. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
16. Huang, X. and Madan, A. (1999) *Genome Res.*, **9**, 868–877.
17. Soderlund, C., Longden, I. and Mott, R. (1997) *Comput. Appl. Biosci.*, **13**, 523–535.
18. Delcher, A., Kasif, S., Fleischmann, R.D., Peterson, J., White, O. and Salzberg, S.L. (1999) *Nucleic Acids Res.*, **27**, 2369–2376.
19. Prestle, J., Schoenfelder, M., Adam, G. and Mundry, K.W. (1985) *Gene.*, **37**, 255–259.