Theor Appl Genet (2003) 107:652–660
DOI 10.1007/s00122-003-1302-4

Tae-Jin Yang · Yeisoo Yu · Gyoungju Nah ·
Michael Atkins · Seunghee Lee · David A. Frisch ·
Rod A. Wing

Construction and utility of 10-kb libraries for efficient clone-gap closure for rice genome sequencing

Received: 18 November 2002 / Accepted: 27 January 2003 / Published online: 3 June 2003

Abstract Rice is an important crop and a model system for monocot genomics, and is a target for whole genome sequencing by the International Rice Genome Sequencing Project (IRGSP). The IRGSP is using a clone by clone approach to sequence rice based on minimum tiles of BAC or PAC clones. For chromosomes 10 and 3 we are using an integrated physical map based on two fingerprinted and end-sequenced BAC libraries to identifying a minimum tiling path of clones. In this study we constructed and tested two rice genomic libraries with an average insert size of 10 kb (10-kb library) to support the gap closure and finishing phases of the rice genome sequencing project. The *Hae*III library contains 166,752 clones covering approximately 4.6× rice genome equivalents with an average insert size of 10.5 kb. The *Sau*3AI library contains 138,960 clones covering 4.2× genome equivalents with an average insert size of 11.6 kb. Both libraries were gridded in duplicate onto 11 high-density filters in a 5 × 5 pattern to facilitate screening by hybridization. The libraries contain an unbiased coverage of the rice genome with less than 5% contamination by clones containing organelle DNA or no insert. An efficient method was developed, consisting of pooled overgo hybridization, the selection of 10-kb gap spanning clones using end sequences, transposon sequencing and utilization of in silico draft sequence, to close relatively small gaps between sequenced BAC clones. Using this method we were able to close a majority of the gaps (up to approximately 50 kb) identified during the finishing phase of chromosome-10 sequencing. This method represents a useful way to close clone gaps and thus to complete the entire rice genome.

Keywords Rice genome sequencing · 10-kb Library · Gap closure

Introduction

The complete DNA sequence of an organism provides a large amount of information for biological studies, not only in the sequenced organism but also in closely related taxa. In plants, the genome of the model plant *Arabidopsis thaliana* has been sequenced completely by an International collaboration (Arabidopsis Genome Initiative 2000). Rice (*Oryza sativa* L.), a model agronomic crop, is closely related to other cereals such as sorghum, corn and wheat. Information obtained from sequencing the rice genome will impact agricultural and biological understanding in rice as well as other cereal crops. The sequencing work is being done using *O. sativa japonica* var. Nipponbare under the auspices of the International Rice Genome Sequence Project (IRGSP), a consortium of research institutions from nine countries, including Brazil, Great Britain, China, France, India, Japan, Korea, Taiwan and the US. Individual chromosomes were allocated to each country for sequencing. In addition, three independent sequence drafts were announced: two drafts using the same japonica variety "Nipponbare", one clone-based sequencing (CBS) by Monsanto (Barry 2001; http:www.rice-research.org) and one whole-genome shotgun sequencing (WGS) by Syngenta (Goff et al. 2002). The third draft was a WGS approach using the indica cultivar "93-11" by the Beijing Genomics Institute of the Chinese Academy of Sciences (Yu et al. 2002).

Communicated by Q. Zhang

T.-J. Yang · Y. Yu · G. Nah · S. Lee · R. A. Wing (✉)
Arizona Genomics Institute, Department of Plant Sciences,
303 Forbes Building, University of Arizona,
Tucson, AZ 85721, USA
e-mail: rwing@genome.arizona.edu
Tel.: +1-520-626-9595
Fax: +1-520-621-7186

T.-J. Yang · Y. Yu · G. Nah · M. Atkins · S. Lee · D. A. Frisch ·
R. A. Wing
Clemson University Genomics Institute, 100 Jordan Hall,
Clemson, South Carolina 29634-5727, USA

D. A. Frisch
Genome Center of Wisconsin, 425 Henry Mall,
Madison Wisconsin 53706, USA

There are two approaches for sequencing large genomes: (1) whole-genome shotgun sequencing (WGS), and (2) clone-based shotgun sequencing (CBS). The WGS approach was applied to the human genome (Venter et al. 2001) and rice genome (Goff et al. 2002; Yu et al. 2002). More than 90% of the human genome is in large scaffold assemblies of 100-kb or more (Venter et al. 2001). However, there is a possibility of large-scale sequence misassemblies and gaps because of sequence repeats, uncloned or unclonable regions (Green 1997).

CBS involves obtaining a collection of large-insert clones, such as a bacterial artificial chromosome (BAC) library, covering a genome and performing shotgun sequencing on a minimum tile of clones. This approach was successfully used for sequencing of the *Arabidopsis* genome (Arabidopsis Genome Initiative 2000) and adopted for rice-genome sequencing by the IRGSP (Sasaki and Burr 2000; http://rgp.dna.affrc.go.jp). Individual assembly of each large insert clone eliminates the possibility of large-scale misassemblies and simplifies closing sequence gaps using small insert shotgun clones known to span the gaps. Some amount of overlap between adjacent BAC clones is indispensable to achieve a complete sequence. Therefore, clone validation is an important task to minimize overlap, and thus to reduce the cost and efforts for sequencing. Two approaches were applied for clone validation in CBS: (1) random clone sequencing (sequencing of random clones and walking from the clone), and (2) map-based sequencing (complete physical mapping and choosing seed BAC clones based on the map).

Random clone sequencing was suggested for sequencing of a genome without physical mapping (Batzoglou et al. 1999; Siegel et al. 1999; Roach et al. 2000; Wendl et al. 2001). The approach proceeds directly to sequence an initial collection of random clones without overlap between them, and then "walks" the genome by iteratively selecting minimally overlapping clones based on BAC-end sequence information, also called "sequence-tagged connectors" (STCs) (Venter et al.1996).

Map-based sequencing involves constructing a complete physical map by the assembly of fingerprinting data of BAC clones into contigs using FPC (Soderlund et al. 2000), and mapping the contigs onto chromosomes with molecular genetic markers, RFLP, SSR and STS. A minimum tiling path of seed BAC clones is then selected for sequencing.

Both WGS and CBS are based on high-throughput shotgun sequencing. One major task in genome sequencing is gap closure subsequent to the high-throughput sequencing phase in both WGS and CBS. There are three types of gaps in the draft genome sequence: (1) sequence gaps, gaps within unfinished sequenced clones; (2) clone gaps, gaps between sequenced-clone contigs, but within fingerprint clone contigs; (3) physical gaps, gaps between fingerprint clone contigs. Closing sequence gaps is relatively straight forward in the CBS approach, but is more complicated in the WGS approach. Gap closure in CBS is achieved by performing additional sequencing and finishing reactions on already identified clones using various methods such as primer walking, shatter libraries (McMurray et al. 1998) and transposons (Devine et al. 1997). Clone gaps and physical gaps are more difficult to evaluate directly, because the draft genome-sequence flanking many of the gaps is often not precisely aligned with the fingerprinted clones. Clone gaps are generally small and most are bridged by one or more individual BAC clones. The 'bridged' clone gaps can be sequenced using the bridging BAC clone. However, the use of BAC clones can be inefficient due to re-sequencing of large overlaps. For example, if a clone gap is 15 kb and the BAC insert is 150 kb, 135 kb or 90% involves redundant sequencing. Frohme et al. (2001) suggested an efficient way to make 'a direct gap-filling library'. The library was made by subtractive hybridization with 1,051 sequenced cosmid libraries against a small insert whole-genome library of *Xylella fastidiosa*. Over 50% of the subtracted library represented gap-specific sequence information comprising about 13 gaps (700–40,000 bp, totalling 208 kb) against 27 Mb of sequence. This method, however, is not feasible for the sequencing of large genomes. Batzoglow et al. (1999) and Wendl et al. (2001) suggested the use of a small insert genomic library to reduce the redundant sequence often accompanied with the use of BAC clones.

For rice, we constructed an integrated sequence-ready rice physical map using two BAC libraries (Chen et al. 2002). The physical map was used to facilitate the rice-genome sequencing consortium by providing a minimum tiling path to the IRGSP. To complete the physical map resource, we report here on the construction of two 10-kb insert libraries (10-kb library) with an 8.4-fold rice genome coverage and a useful resource to support the finishing phase of rice genome sequencing. The 10-kb libraries were successfully used for closing clone gaps. We made an efficient and a cost-effective protocol to close clone gaps using the 10-kb library: library screening by pooled overgo hybridization; clone selection of the 10-kb clones spanning each gap by blast-dbase and fingerprinting; and transposon-based full sequencing of the 10-kb clone. We were able to close large gaps (up to 46 kb) efficiently by using the Monsanto rice draft sequence.

Materials and methods

Plasmid vector preparation

Medium-copy plasmid vector (25 copies/cell), pCUGIblu21 (Yang et al. 2002), was chosen for 10-kb library construction. Plasmid DNA was isolated using the Qiagen plasmid midi kit (Qiagen, Calif.) according to instructions. Two restriction enzymes were used to prepare the vector: *Eco*RV for blunt-end ligation and *Bam*HI for *Sau*3AI cohesive-end ligation. The vector was prepared following the method of Luo et al. (2001) with slight modification. The plasmid DNA, 5 μg, was digested at 37 °C for 1 h with 10 U of *Eco*RV (New England Biolab, NEB) or *Bam*HI (NEB) in a total volume of 50 μl with 1× CIP buffer (NEB buffer III). The restriction enzyme was inactivated by incubation at 65 °C for 15 min. The digested DNA was de-phosphorylated for 1 h at 37 °C after adding calf intestinal (CIP) alkaline phosphatase: 1 U for the *Bam*HI sticky end and 5 U for the *Eco*RV blunt end in 10 μl of 1×

CIP buffer. Self-ligation was conducted to circularize the undephophorylated vector with 10 U of T4 ligase (Promega Company) at 16 °C overnight and the linearized pCUGIblu21 DNA was recovered from an agarose gel using the concert gel extraction system (Gibco BRL). Fifty nanograms of aliquots were stored at –20 °C. The quality of the prepared vector was evaluated by cloning *Sma*I- and *Bam*HI-digested lambda DNA for blunt-end and cohesive *Bam*HI-end ligations, respectively.

10-kb Library construction

Young leaves of 30-day old rice (*O. sativa* ssp. *japonica* cultivar Nipponbare) were collected and stored at –80 °C. High molecular DNA plugs were prepared by embedding nuclei in 0.5% low-melting agarose as described by Zhang et al. (1995). Washed DNA plugs were stored in TE buffer at 4 °C. Partial digestions and double-size selections were conducted as described by Budiman et al. (2000) and Luo et al. (2001) with modifications. A series of pilot partial digestions, with *Hae*III and *Sau*3AI separately, using different enzyme concentrations, were performed for 30 min at 37 °C. Finally, one plug was digested with 10 U of *Hae*III and 2.8 U of *Sau*3AI, respectively. Digestion was performed at 37 °C for 30 min and terminated by adding 10 μl of 0.5 M EDTA. The partially digested DNA fragments were separated on a 1% agarose CHEF (Bio-Rad, Calif.) gel with a 0.1–10 s linear ramp at 4 V/cm at 14 °C in 1× TAE buffer for 14 h. DNA fragments ranging from 8 to 20 kb were excised and the agarose pieces were placed upside down into a new 1% TAE agarose gel. The second CHEF electrophoresis was conducted under identical conditions, but for 16 h, to remove small trapped fragments and to concentrate the DNA. Insert DNA was recovered from the condensed gel fractions by elution using the concert gel-extraction system (Gibco BRL). For ligation, 25-ng of de-phosphorylated, linearized pCUGIblu21 and 100 ng of insert DNA were incubated with 3 U of T4 ligase (Promega Company) in 1× buffer. The ligation was stopped by incubation at 65 °C for 15 min and de-salted in 0.1 M glucose/1% agarose cones for 1 h on ice as described by Atrazhev and Elliott (1996). The ligation was transformed into electro-competent *Escherichia coli*, DH10B, by electro transformation under instructions (GibcoBRL). Transformants were selected on LB/X-gal/IPTG plates containing 50 mg/l of kanamycin. White colonies were picked and arrayed into 384-well microtiter plates (Genetix LTD, UK) using a Genetix Q-bot (Genetix LTD, UK). Libraries were stored at –80 °C.

10-kb Clone characterization

To estimate insert size and the distribution of clone sizes, DNA from a total of 768 clones, 384 from *Hae*III and 297 from the *Sau*3AI library, were prepared from 1.2 ml of 2× YT culture in 96-well format deep-well plates using an alkaline-lysis method. DNA was digested with 5 units (3 h at 25 °C) of *Swa*I (NEB), a rare cutting enzyme, and analyzed by normal horizontal electrophoresis (1.5 V/cm, 14 h) or by CHEF in 1% agarose gels (4 V/cm, 0.1–10 s switch time, 14 h run time, 14 °C).

10-kb Library screening

High-density colony filters were prepared using the Gentix Q-Bot. Clones were gridded in duplicate spots using a 5 × 5 array pattern on 22.5 × 22.5-cm Hybond N+ filters (Amersham). This gridding pattern allows 27,648 independent clones to be represented per filter. Three rice chloroplast probes, *psbA, rpoB* and *psaA*, which are PCR-amplified, were used to determine chloroplast genome contamination by hybridization against one filter of each library as described by Budiman et al. (2000). The probes were radioactively labelled using the Ambion random priming kit.

Overgo design and hybridization

Overgo probe design strategy was used to screen the 10-kb library and to select 10-kb clones spanning gaps between two fully or partially sequenced BAC clones. BLASTN from NCBI (http://ncbi.nlm.nih.gov) was used to determine the unique BAC terminal sequence flanking the gap. The script Overgo marker (http://genome.wustl.edu) was used for designing overgo primer pairs, two 24-mers that overlap by 8 bp resulting in a 40-bp probe. Radioactive labeling and hybridization was performed as described by Chen et al. (2002) with modifications. An overgo pooling strategy was adapted for dealing with several gaps at the same time (see Figs. 5 and 6). Each overgo was evaluated by hybridization to one filter to identify non-specific overgos. The unique overgos representing less than four hits per filter (0.8-fold genome coverage per filter) were pooled and used for hybridization against a whole set of 11 filters. Up to 13 radioactively labeled probes were pooled at one time.

End-sequence of 10-kb clones and homology search

The 10-kb clones selected by overgo hybridizations were picked into 96-well plates and end-sequenced using BigDye terminator chemistry v2.0 (ABI) with T3 and T7 vector primers and analyzed using the ABI3700 automatic DNA sequencer (ABI). Base-calling was performed automatically using phred, and vector sequences were removed by CROSS_MATCH (Ewing and Green 1998; Ewing et al. 1998). High quality, vector-trimmed, end-sequences (defined as those having >100 non-vector bases with a PHRED quality value >20) were kept in FASTA as a database. BLAST-DATABASE, "blast/db" of NCBI was used for homology searches with an individual BAC sequence flanking each gap against the database of the 10-kb end-sequence. Ten kilobase clones in which one or both end-sequences showed over 96% sequence identity were selected for future analysis.

Full sequencing of the 10-kb clones and sequence assembly

A transposon based sub-library construction system (TGS system F-700, Finnzymes) was used to construct a sub-library of a 10-kb clone according to the manufacturer's instruction. Two × 96 clones from each sub-library were sequenced using transposon specific primers, seqA and seqB, provided from the manufacturer. High quality, vector-trimmed sequences were thus used for sequence assembly of the 10-kb clone using phrap and consed (Gordon et al. 1998). The consensus sequence of the 10-kb clone was used for filling and closing the gaps by assembly with the BAC sequence.

The sequence data described in this paper was submitted to GenBank together with the north part of the BAC sequence under the accession numbers in Table 1.

Results

Characterization of 10-kb libraries

Two rice whole genome libraries with an average insert size of 10 kb (10-kb library) were constructed from *O. sativa* ssp. *japonica* cultivar Nipponbare to support the gap-closure and finishing-phases of the rice genome sequencing project. A series of plasmid vectors were constructed (Yang et al. 2002) and one, pCUGIblu21, was selected for construction of the libraries using *Hae*III and *Sau*3AI partially digested genomic DNA. A total of 384 and 297 clones were randomly picked from *Hae*III and *Sau*3AI, respectively, and their insert sizes were deter-

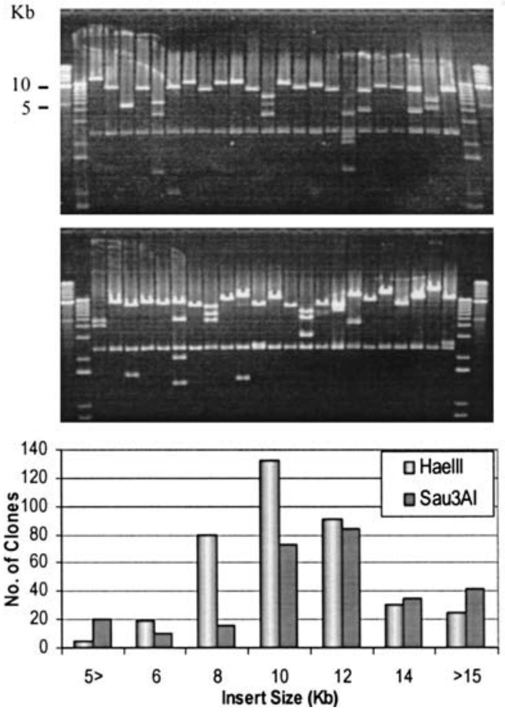

Fig. 1 Analysis of 10-kb clones randomly selected from two rice 10-kb libraries, *Hae*III library (**A**) and *Sau*3AI (**B**), and insert distribution (**C**). A total of 23 clones from each library was digested with the *Swa*I restriction enzyme. Ethidium bromide-stained CHEF gels (0.1–10 s switch time, 4 V/cm for 14 h) show insert DNA around the common 2.4-kbp pCUGIblu21 vector. The two lanes outside are molecular-weight markers, the 5 kb and 1 kb ladders (Gibco BRL) for the first and the second from outside, respectively

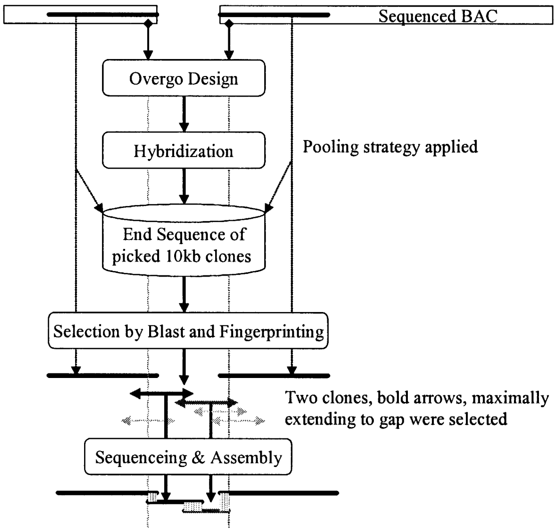

Fig. 2 Strategy for gap closure using a 10-kb library. Two overgos designed from the distal ends of BAC sequence flanking gaps were used to screen the 10-kb library. All the selected 10-kb clones were end-sequenced and blasted with BAC sequence flanking gaps. One or two 10-kb clones having a minimum overlap with the BAC sequence and extending maximally to the gap were chosen for full sequencing and thus for closing the gap

mined by digestion with *Swa*I followed by electrophoresis (Fig. 1A and B). The *Hae*III library (OSJNPb000, http://www.genome.clemson.edu) contains 166,752 clones with an average insert size of 10.6 ± 2.44 kb, and the *Sau*3AI library (OSJNPc000) contains 138,960 clones with an average insert size of 11.6 ± 2.76 kb. The clones containing inserts greater than 8 kb and less than 5 kb were 93.7% and 1.5%, respectively, in the *Hae*III library, and 89.2% and 5%, respectively, in the *Sau*3AI library (Fig. 1C). The coverage for the *Hae*III and *Sau*3AI libraries is estimated at 4.4 and 4.0 haploid genome equivalents respectively, thus resulting in an 8.4-fold equivalent for both, based on the genome size of 400 Mb (Chen et al. 2002; Ohmido et al. 2000). The combined libraries were archived in 792 × 384-well micro-titer plates and gridded in duplicate onto 11 high-density nylon filters in a 5 × 5 pattern. We obtained an average of ten hits from more than 50 overgo probe hybridizations with a range from 6 to 40 hits, supporting our estimation regarding genome coverage (8.4-fold). Approximately 200 hits were obtained by hybridization with the telomere repeat $(TTTAGGG)_6$. Less than 5% of the clones from each library are considered as contaminants such as those containing organelle DNA or small inserts below 2 kb. Around 3% of the libraries represented the chloroplast genome based on hybridization with three chloroplast genes.

Clone gap closure using the 10-kb library

A total of 202 BAC clones were sequenced for chromosome (Yu et al. 2003). Although most of the sequenced BAC clones were assumed to align contiguously with a slight overlap with adjacent BAC clones, unexpected clone gaps were found after sequence assembly. Most of the gaps were spanned by several individual BAC clones; however, it would be expensive and time-consuming to perform shotgun sequencing of a BAC clone containing a 50-kb insert for closure of a small gap. A small insert library such as the 10-kb library is an alternative to reduce the amount of redundant sequence, thereby reducing the finishing cost. PCR might be applied if the gap size is small and accurate. However, it is sometimes difficult to estimate the gap size especially if the gap is from a complicated region of the genome.

We devised a walking strategy to close most clone gaps using the 10-kb libraries as shown in Fig. 2. The strategy consists of three steps, (1) library screening; (2) clone selection; and (3) sequence assembly: (1) library screening, we screened the 10-kb library using the overgo probe designed from the BAC sequence flanking the gap; (2) clone selection, we tried to identify one or two 10-kb clones for each gap based on blast analysis of the flanking BAC sequence against a local data base of end sequences and subsequent fingerprinting of the selected 10-kb clones; clones overlapping with the flanking BAC sequence with more than 96% sequence identity and

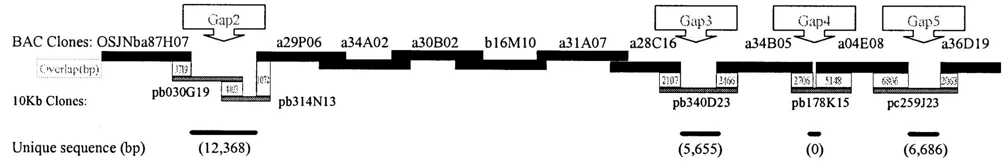

Fig. 3 Representation of successful gap closures using 10-kb clones for a medium and three small gaps among ten contiguous sequence BAC clones on chromosome 10

Table 1 The clone gap list and summary of the gap closure using the 10-kb clones

Gap no.	BAC clones flanking gap	Estimated size (kb)	Actual size (bp)	10-kb Clone	Insert (kb)	Genebank acc. no.
1	OSJNba0079B05 OSJNba0049K09	29.0	39,382	OSJNpc115P09 OSJNpc187H07 OSJNpc297A03 OSJNpb230J05	10.2 12.2 14.0 10.0	AC116601
2	OSJNba0087H07 OSJNba0029P06	4.5	12,368	OSJNpb030G19 OSJNpb314N13	13.4 8.2	AC099733
3	OSJNba0034B05 OSJNba0004E08	27.0	5,655	OSJNpb340D23	10.3	AC122145
4	OSJNba0004E08 OSJNba0036D19	4.5	0	OSJNpb178K15	14.0	AC091724
5	OSJNba0036D19 OSJNba0050E08	50.0	6,686	OSJNpc259J23	15.0	AC116601
6	OSJNbb0015J03 OSJNbb0081F12	40.5	46,104	OSJNpc072A13 OSJNpc200H08 OSJNpb148E04 OSJNpb196I08 OSJNpb406G03	11.0 15.0 13.2 11.4 10.5	AC131375
7	OSJNba0050M22 OSJNba0035F15	3.0	0	OSJNpb192I17	9.6	AC090483
8	OSJNba0011L09 OSJNbb0058B20	4.5	1,058	OSJNpc015A22	10.1	AC122144

extending into the gap were chosen; an average of three clones (range 1–8) for each target was selected at this step; based on fingerprint and insert size evaluation, we selected one or two clones extending maximally into each gap; (3) sequence assembly, we fully sequenced the 10-kb clone using the transposon-insertion method; most of the 10-kb clones were successfully assembled as one contig by sequencing of the 192 transposon insertions; the consensus sequence of the 10-kb clone was combined and re-assembled with the BAC sequence to close the gap, resulting in the contiguous sequence with sequenced BACs; if the gap is larger than can be spanned by two 10-kb clones, a second round of screening was applied from the end of the 10-kb sequence.

Small- and medium-size gap closure

One medium-size gap (Gap 2 in Fig. 3) and three small gaps (Gap 3, 4 and 5 in Fig. 3) were found between ten contiguous sequenced BACs aligned on the physical map (Chen et al. 2002). For gap 3, we identified a few 10-kb clones, including OSJNpb340D23, having both ends sharing a redundant sequence with the adjacent BAC clones, OSJNba0034B05 and OSJNba0004E08. They spanned the gap even though the estimated gap size was 27 kb (Table 1). The gap between OSJNba0034B05 (171,162 bp) and OSJNba0004E08 (188,163 bp) was closed by 5,655 bp of unique sequence from a 10-kb clone, OSJNpb340D23 (10,228 bp). Another gap, gap 4 between a04E08 and a36D19 (144,197 bp), turned out to be a pseudo-gap having a total of 6 bp overlap. Gap 5 between OSJNba0036D19 and OSJNba0050E08 (155,878 bp) was closed by 6,686 bp of unique sequence from a 10-kb clone, OSJNpc259J23 (14,555 bp), with a total of 8.8 kb of redundant sequence. Two other gaps, gap 7 and 8 in Table 1, were closed by one 10-kb clone for each gap by an identical method.

One medium size gap (gap 2 in Fig. 3) was successfully closed by the sequence assembly of two 10-kb clones. The gap size was estimated as less than 4.5 kb based on BAC fingerprints (Table 1). However, we could not find any 10-kb clones representing overgos from both OSJNBa0087H07 and OSJNBa0029P06, even though we obtained four hits from OSJNBa0087H07 and 22 hits from OSJNBa0029P06. Blast/db showed an end of one 10-kb clone, OSJNpb314N13, having sequence identity with OSJNba0087H07, and several 10-kb clones, including OSJNpb314N13, having sequence identity only with OSJNba0029P06. Fingerprinting by digestion with *Dra*I revealed common bands between them, OSJNpb030G19 and OSJNpb314N13 representing an overlap between two clones and thus spanning the gap. These two clones span 12,368 bp of unique sequence between OSJNBa0087H07 (156,654 bp) and OSJNBa0029P06 (151,533 bp).

Large gap closure

Two large gaps, gap 6 and 1 in Table 1, were closed using 10-kb clones with the aid of the Monsanto draft sequence as in Fig. 4. The draft sequence was used as a template for designing overgos from inside of the clone gaps by an in silico sequence comparison. We positioned the sequenced Monsanto BAC clones into our physical map by combining their fingerprint data into our map. We found BAC clones spanning large gaps such as Monsanto BAC clone OJ1004C03 for gap 1 and OJ1001D01 for gap 6. By combining the Monsanto shotgun sequence trace files (up to 5× equivalents) of OJ1001D01 and two independent consensus sequences of our BACs, OSJNbb0015J03 and OSJNbb0081F12, one Monsanto sequence contig was extending 6.8 kb from OSJNbb0015J03 and another 6-kb contig was assumed to be floating inside the gap (Fig. 4). Based on this information, even if the quality of the sequenced contig is below the standard, we were able to design three overgos based on the Monsanto sequence contigs (arrows no. 1, 2 and 3) and one overgo from OSJNbb0081F12 (arrow no. 4). By selection based on blast/db, we selected four 10-kb clones at the first round of selection: two 10-kb clones, OSJNpc072A13 and OSJNpb406G03, for extending from both BAC ends; two 10-kb clones, OSJNpb148E04 and OSJNpb196I08, extending to the opposite direction from the Monsanto 6-kb contig inside the gap. Three 10-kb clones were assembled together, resulting in a 33-kb extension from BAC clone, OSJNbb0081F12, and one 10-kb clone, OSJNpc072A13, extending 14 kb from BAC clone, OSJNbb0015J03. Even though the gap was not closed by the four 10-kb clones, we found two other small contigs extending into the remaining gap after re-assembling the sequence of four 10-kb clones and the Monsanto sequence trace files. At the 2nd round, two overgos (arrows no. 5 and 6), were used to select a 10-kb clone, OSJNpc200H08, spanning the remaining gap. The selected 10-kb clones spanned a gap totaling 46,104 bp. The other gap, gap 1 in Table 1, was closed with 39,382 bp of unique sequence in a similar manner using four 10-kb clones.

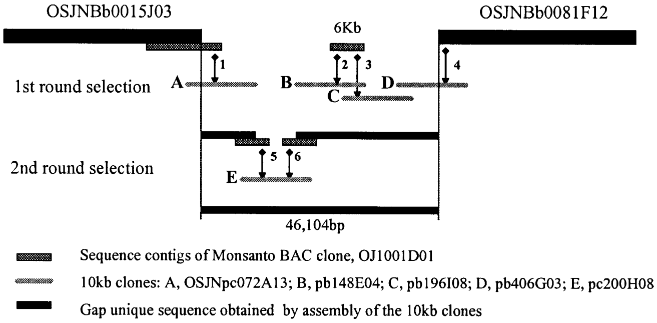
OSJNBb0015J03
6Kb
OSJNBb0081F12
1
2
3
4
1st round selection
A
B
C
D
2nd round selection
5
6
E
46,104bp
Sequence contigs of Monsanto BAC clone, OJ1001D01
10kb clones: A, OSJNpc072A13; B, pb148E04; C, pb196I08; D, pb406G03; E, pc200H08
Gap unique sequence obtained by assembly of the 10kb clones

Fig. 4 Representation of a successful gap closure using five 10-kb clones and Monsanto sequence draft for a big clone gap (46,104 bp) between two sequenced BAC clones in rice chromosome 10

Discussion

10-kb Library and the 5 × 5 patterned high-density filter

Most BAC libraries are gridded on nylon filters (22.5 × 22.5 cm) using a 4 × 4 pattern. They represent different genome coverage, depending on the genome size and average insert size of the BAC library. Approximately 25-fold rice genome coverage is represented by five filters using a 4 × 4 pattern, resulting in a 5× coverage per filter (Chen et al. 2002); a 15-fold tomato genome coverage by seven filters, 2× per filter (Budiman et al. 2000); a 8-fold cotton genome coverage by eight filters, 1× per filter (Tomkins et al. 2001); and a 6-fold barley genome coverage by 17 filters, 0.3× per filter (Yu et al. 2000). We required more than 320,000 clones carrying a 10-kb insert to represent 8× rice genome equivalents, resulting in 17 filters based on a 4 × 4 pattern (18,432 clones/filter), but only 11 filters based on a 5 × 5 pattern (27,648 clones/filter), a 33% reduction. Therefore, we used a 5 × 5 pattern on filters for the 10-kb libraries and obtained clear hybridization data such as Fig. 5. The 10-kb library is a valuable resource not only for gap closure, but can also be used to isolate single genes for functional or comparative genomics. It is more straight-forward to isolate a complete gene structure from a 10-kb clone than a 150-kb BAC clone because the size of most rice genes is less than 10 kb (Feng et al. 2002; Goff et al. 2002; Sasaki et al. 2002; Yu et al. 2002; Yu et al. 2003).

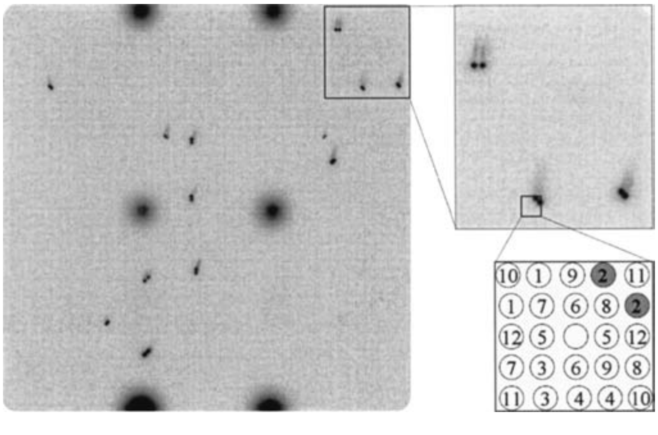

Fig. 5 Hybridization image of one 5 × 5 patterned filter of a 10-kb library using pooled Overgo probes. Colonies were double-spotted in high density with the Genetix Q-bot on a 22.5 × 22.5-cm filter in a 5 × 5 pattern for the over-sized figures. Each filter contains 27,648 colonies

Selection method for gap closure

We relied on overgo hybridization for screening the 10-kb library. Overall, we have obtained 20 clones from each overgo hybridization; among all the overgos designed from 14 clone gaps, 70% were unique (less than 20 hits), 20% were unique with unspecific hybridization (40 to 70 hits) and the other 10% were unspecific recognizing repeat sequences, which were discarded. To select one ideal clone from among an average of 20 putative clones screened by overgo hybridization, we compared several methods: subtraction hybridization, PCR, blast-dbase and fingerprinting.

We tried to reduce the number of putative clones by subtractive hybridization. We designed two overgos, one from a 1-kb region proximal to the gap and another from a region 5 kb distal to the gap, and hybridized individual overgos against all 11 filters. By comparison of the two hybridization data, we classified the 10-kb clones into three groups: 1 kb-unique; 1 kb and 5 kb in common; and 5 kb-unique. Fingerprints showed similarity within a group but no similarity between groups, supporting our expectation. We preferred to select one clone spanning the gap from the 1 k-unique group rather than the other groups. To validate a clone, we performed PCR-amplification and blast-dbase similarity searches. The PCR-amplification was conducted using one of the vector primers, T3 or T7, and one overgo primer. The smaller PCR product represents less overlap because the size represents the overlap length. However the PCR result was not credible because it did not correlate with the result of blast/db. We could not find any clones having sequence identity from six clones in the 1 kb-unique class, on the other hand, we found five out of eight clones in the 1 kb and 5 kb common class. Only one clone correlated with the PCR and blast results. Another clone showed homology with both sides of BAC clones flanking the gap, indicating it spanned the gap. It closed the gap successfully even though there are 6,806 bp of redundant sequence (gap 5 in Fig. 3). The whole sequence of one clone in the 1 kb-unique group showed no relation with the gap. Therefore, we concluded that most of the 1 kb-unique clones are from other chromosomal regions which have similarity with the 40 bp overgo sequence. The hybridization data supported the assumption; the signal of the 1 k-unique hits are weaker than the 1 kb and 5 kb common hits. The false PCR products, therefore, might be a result of using the overgo primer for PCR amplification. The blast-dbase is thought to be the most-reliable method for selecting the ideal 10-kb clone by excluding the false positive based on small nucleotide sequence homology, i.e., 40 bp and 24 bp by overgo hybridization and PCR, respectively.

Overgo-pooling strategy for hybridization

The overgo-pooling for hybridization was considered because we could allocate and select putative 10-kb clones based on the blast-dbase method as mentioned previously. All the overgos designed from several gaps were evaluated individually by hybridization with one filter. The overgos providing less than four hits at a filter (around an 0.8-fold genome-equivalent per filter) were considered as a unique probe and pooled without replication for screening. We obtained clear hybridization hits, such as shown in Fig. 5, using up to 13 pooled overgo primer pairs. Figure 6 shows the pooling strategy. After screening by pooling, all clones detected were sequenced from both ends. We could assign clones to each gap and select the clones extending into the gap by blast-dbase with each of the BAC sequences flanking the gap. Around 20% were selected and considered as putative clones for each gap. By fingerprinting and sizing the insert of the clones selected by blast-dbase, we tried to select one or two clones for each gap. Most gaps, except gap nos. 1 and 6 in Table 1, were closed by this strategy. The pooling strategy allowed us to reduce hybridization experiments by more than 70% and deal with many gaps simultaneously.

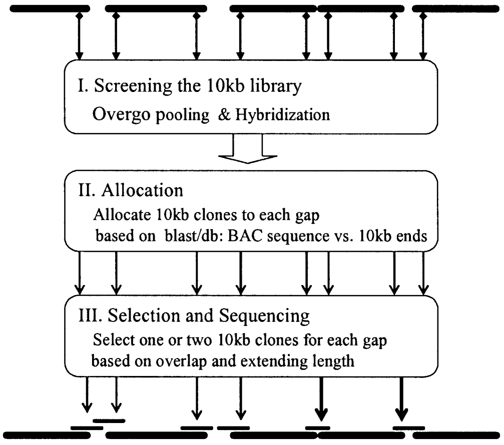
I. Screening the 10kb library
Overgo pooling & Hybridization
II. Allocation
Allocate 10kb clones to each gap
based on blast/db: BAC sequence vs. 10kb ends
III. Selection and Sequencing
Select one or two 10kb clones for each gap
based on overlap and extending length

Fig. 6 The Overgo pooling strategy for hybridization. A maximum of 13 unique overgos were pooled for hybridization against 11 10-kb filters (8.8× genome coverage). Each of the 10-kb clones were re-allocated into each gap based on blast with the BAC sequence flanking each gap, and one or two 10-kb clones were finally selected for closure of each gap

Gap closure strategy using the 10-kb library

A number of methods can be used for clone gap closure. It is possible to sequence a small clone gap by primer walking; this is, however, an inefficient and time-consuming process. Relatively large gaps up to approximately 30 kb might be closed by sequencing PCR amplification products when a BAC clone spanning the gap is available and the gap size is precisely estimated. The PCR amplification technique was applied at the finishing phase of clone-based genome sequencing (CBS) projects, such as *Arabidopsis* (Arabidopsis Genome Initiative 2000). Even though the PCR technique is convenient and useful, PCR has some defects for the generation of sequencing templates. In vitro mutation by base-pair substitution or deletion results in inaccuracies of the genome sequence. McMurray et al. (1998) reported that the PCR reaction could skip a 400-bp fragment between two repeat regions, resulting in loss of the sequence.

The IRGSP adopted CBS for rice whole-genome sequencing because it gives a more accurate sequence than WGS. Chromosome 10 (Yu et al. 2003), chromosome 1 (Japan rice genome sequenceing, http://rgp.dna.affrc.go.jp, Sasaki et al. 2002) and chromosome 4 (Feng et al. 2002) were completely sequenced based on this method. Most of the clone gaps encountered on the rice chromosome-10 sequencing project were not expected. They are much smaller than a single BAC and are bridged by one or more individual BACs. However, it is very difficult to estimate the gap size because the sequenced BACs flanking many of the gaps are often not precisely aligned with the fingerprinted clones on the FPC map. There are big differences, in many cases, between the estimated gap size vs the real gap size: 4.5 kb vs 12.5 kb in gap 1 and 27 kb vs 56 kb in gap 3 (Table 1), supporting our strategy.

There is another benefit of using the 10-kb library. The Monsanto draft sequence was successfully used for closure of a large clone gap (as shown in Table 1 and Fig. 4). The Monsanto draft is available to IRGSP members for in silico sequence comparison with their own sequencing progress. Abundant sequence information might be provided and the 10-kb library successfully used to confirm the in silico sequence extending into the gap and to close the remaining gap. The Syngenta draft sequence is another valuable resource for finishing the rice genome sequence. Even though all the sequence contigs were not allocated onto each chromosome, the sequence drafts could be applied in the effort of gap closure. If the draft sequences are successfully incorporated into the gap-closure procedure by IRGSP members, the rice whole-genome sequencing will be completed in advance. We expect the gap-closure strategy using the 10-kb library to contribute to completing the rice genome sequence.

Acknowledgements This work was supported by grants from the U.S. Department of Agriculture CREES, the National Science Foundation, and the Department of Energy and postdoctoral fellowship program from Korea Science and Engineering Foundation (KOSEF). Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Department of Agriculture, the National Science Foundation or the Department of Energy. We thank Monsanto for providing access to their draft sequence data.

References

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Atrazhev AM, Elliott JF (1996) Simplified de-salting of ligation reactions immediately prior to electroporation into *E. coli*. Biotechniques 21:1024

Barry G (2001) The use of the Monsanto draft rice genome sequence in research. Plant Physiol 125:1164–1165

Batzoglou S, Berger B, Mesirov J, Lander ES (1999) Sequencing a genome by walking with clone-end sequences: a mathematical analysis. Genome Res 9:1163–1174

Budiman MA, Mao L, Wood T, Wing RA (2000) A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. Genome Res 10:129–136

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, et al. (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:1–10

Devine SE, Chissoe SL, Eby Y, Wilson RK, Boeke JD (1997) A transposon-based strategy for sequencing repetitive DNA in eukaryotic genomes. Genome Res 7:551–563

Ewing B, Green P (1998) Base-calling of automated sequencer traces using *Phred*II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. Genome Res 8:175–185

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X, et al. (2002) Sequence and analysis of rice chromosome 4. Nature 420:316–320

Frohme M, Camargo AA, Czink C, Matsukuma AY, Simpson AJG, Hoheisel JD, Verjovski-Almeida S (2001) Directed gap closure in large-scale sequencing projects. Genome Res 11:901–903

Goff SG, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Green P (1997) Against a whole-genome shotgun. Genome Res 7:410–417

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–941

Luo M, Wang YH, Frisch D, Joobeur T, Wing RA, Ralph AD (2001) Melon BAC library construction using improved methods and identification of clones linked to the locus conferring resistance to melon Fusarium Wilt (*Fom*-2). Genome 44:154–162

McMurray AA, Sulston, JE, Quail MA (1998) Short-insert libraries as a method of problem solving in genome sequencing. Genome Res 8:562–566

Ohmido N, Kijima K, Akiyama Y, de Jong JH, Fukui K (2000) Quantification of total genomic DNA and selected repetitive sequences reveals concurrent changes in different DNA families in indica and jopanica rice. Mol Gen Genet 263:388–394

Roach JC, Thorsson V, Siegel AF (2000) Parking strategies for genome sequencing. Genome Res 10:1020–1030

Sasaki T, Burr B (2000) International rice genome sequencing project: the effort to completely sequence the rice genome. Curr Opin Plant Biol 3:138–141

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, et al. (2002) The genome sequence and structure of rice chromosome 1. Nature 420:312–316

Siegel AF, Trask B, Roach JD, Mahairas GG, Hood L, van den Engh G (1999) Analysis of sequence-tagged-connector strategies for DNA sequencing. Genome Res 9:297–307

Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers and FPC V4.7. Genome Res 10:1772–1787

Tomkins JP, Peterson DG, Yang TJ, Main D, Wilkins TA, Paterson AH, Wing RA (2001) Development of genomic resources for cotton (*Gosypium hirsutuum*): BAC library construction, preliminary STC analysis, and identification of clones associated with fiber development. Mol Breed 8:255–261

Venter JD, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381:364–366

Venter JD, Adams MD, Myers EW, Li PW, Mural RJ, Sutton SS, Smith HO, Yandell M, Evans CA, Holt RA, et al. (2001) The sequence of the human genome. Science 291:1304–1351

Wendl MC, Marra MA, Hillier LW, Chinwalla AT, Wilson RK, Waterston RH (2001) Theories and applications for sequencing randomly selected clones. Genome Res 11:274–280

Yang TJ, Yu Y, Frisch D, Wing RA (2002) Two series of plasmid vectors for the shotgun library. Plant and Animal Genome Conference X, p 24

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–91

Yu Y, Tomkins JP, Waugh R, Frisch D, Kudrna D, Kleinhofs A, Brueggeman RS, Muehlbauer GJ, Wise RP, Wing RA (2000) A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. Theor Appl Genet 101:1093–1099

Yu Y, Rambo T, Currie J, Sasaki C, Kim HR, Collura K, Thompson S, Simmons J, Yang TJ, Park GN, Patel AJ, Thurmond S, Henry D, Oates R, Palmer M, Pries G, Gibson J, Anderson H, Paradkar M, Crane L, Dale J, Carver MB, Wood T, Frisch D, Engler F, Soderlund C, Palmer LE, Tetylman L, Nascimento L, Bastide M de la, Spiegel L, Ware D, O'Shaughnessy A, Dike S, Dedhia N, Preston R, Huang E, Ferraro K, Kuit K, Miller B, Zutavern T, Katzenberger F, Muller S, Balija V, Martienssen RA, Stein L, Minx P, Johnson D, Cordum H, Mardis E, Cheng Z, Jiang J, Wilson R, McCombie WR, Wing RA, Yuan Q, Ouyang S, Liu J, Jones KM, Gansberger K, Moffat K, Hill J, Tsitrin T, Overton L, Bera J, Kim M, Jin S, Tallon L, Ciecko A, Pai G, Aken SV, Utterback T, Reidmuller S, Bormann J, Feldblyum T, Hsiao J, Zismann V, Blunt S, Vazeilles A de, Shaffer T, Koo H, Suh B, Yang Q, Haas B, Peterson J, Pertea M, Volfovsky N, Worman J, White O, Salzberg SL, Fraser CV, Buell CR, Messing J, Song R, Fuks G, Llaca V, Kovchak S, Young S, Bowers JE, Paterson AH, Johns MA, Mao L, Pan H, Dean RA (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. Science (in press)

Zhang HB, Zhao X, Ding X, Paterson AH, Wing RA (1995) Preparation of megabase-size DNA from plant nuclei. Plant J 7:175–184