

A rapid and accurate strategy for rice contig map construction by combination of fingerprinting and hybridization

Guofan Hong*

National Center for Gene Research, Chinese Academy of Sciences, Caobao Road 500, Shanghai 200233, China

Key words: rice, genome, contig map

Abstract

A rapid and accurate strategy for rice contig map construction was described. Rice BAC library with average insert of 120 kb in length was used as building materials in contig mapping. The contigs of varied lengths ranging from 500 kb to several megabases with sufficient redundancy to ensure the accuracy of the joining between individual BACs were formed by fingerprinting. The contigs were then assigned to and ordered along the chromosomes by various molecular markers through their hybridization against the whole rice genomic library. The accuracy of clone overlaps in contig was further confirmed by the existence in contigs of well fit stacks of marker-lodged clones. He contigs thus obtained covered nearly the rice genome.

Introduction

Though the strategy described in this paper was applied to the construction of rice contigs, it could be used for the contig map construction of other genomes.

Nearly half of the world population, who are largely living in the developing countries, feed on rice. Rice, *Oryza sativa*, the smallest genome of any monocots known (4.3×10^8 bp in length) [2] has now become a model plant among the cereals for molecular genetic studies.

If genes can be tagged with closely linked molecular markers, then sections of the genome that are close to the desired gene can be selected from a genomic library with the markers themselves as probes. If the markers are very close and a library of very large inserts, such as YACs, is available, the gene of interest might be on the same clone as the marker. This is unlikely, however, because it is difficult to obtain markers that are close enough. One will have to 'walk along' the chromosome by selecting a clone adjacent to the marker clone, then the next adjacent clone, and so on, until the gene is reached. These tedious, costly and time-

consuming procedures will be rendered superfluous when a contig map of high resolution is available.

In order to detect overlaps between clones a procedure for comparing clones with one another will be required. Overlaps can simply be detected by hybridization. However, it is not in itself a satisfactory criterion for matching, because dispersed repeats can generate numerous false-positives. Like the rice genome, which contains about 50% repeats, many others have repeats to various degrees. The fingerprinting procedure, which is based on principles entirely different from those of hybridization [6, 17, 18] is therefore a good choice for generating rice contigs.

Contig construction by fingerprinting requires a representative genomic library. There are quite a few vectors available for the construction of genomic libraries, such as cosmid, BAC and YAC. With the rice genome of ca. 4.3×10^8 bp in length, the insertion capacity of around 4.5×10^4 bp of cosmid seemed to be too small. Though YACs with large insertion capacity of up to 10^6 bp were widely used in constructing many genome libraries including human [5] and rice genomes [14], they suffer from high rates of chimerism and thus are unsuitable for genomic sequencing [4, 13, 15]. With its average insertion capacity of around 1.2×10^5 bp, BAC has a number of advantages over

* Also research professor of Shanghai Institute of Biochemistry, Chinese Academy of Sciences.

the YAC. With their very low copy number (mostly single-copy), BACs have very low rates of chimerism with a high degree of structure stability in the host cell, even after 100 generations of serial growth [16]. That the cloning with BAC is as easy as with ordinary *E. coli* vectors greatly facilitates the construction of libraries of large inserts. Furthermore, much easier manipulation of BAC clone DNAs than of YACs made it possible to save huge amounts of time and money, associated with the mapping and analysis of complex genomes such as rice's genome. The much shorter average insert of BAC libraries made the BAC contigs obtained have many times higher resolution than that made of YACs, which significantly simplifies the time-consuming and tedious procedures involved in identifying, by map-based cloning, genes of interest. Since the contigs consist of an ordered library of BAC inserts, it provides a scaffold for large-scale DNA sequencing of the genome. A genetically stable and representative BAC library of the rice genome was constructed earlier [19], which has an average insert of 120 kb.

Contig construction is essentially a process of image analysis of restriction enzyme fingerprint autoradiograms. Genomic clones are processed by restriction digestion and isotopic labelling to produce mixtures of radioactive DNA fragments. The resulting mixtures are separated on DNA sequencing gels with groups of such samples being interspersed with standard markers of various sizes. To obtain reliable fingerprints, the restriction digestion must come to completion. Digestion of individual rice BAC clones with a combination of *HindIII/HaeIII* for generating fingerprint autoradiograms has proven to be very satisfactory in our lab. For obtaining reproducible fingerprints, the reaction conditions should be optimized to be optimal for clones of different base composition, and the enzymes and reagents used in the reaction must be of high quality.

Before starting large scale image analysis, we did a series of tests to learn whether the fingerprints produced from such rice genomic BAC clones were adequate for clone mapping. About four hundred clones selected at random were analyzed for this purpose. The results obtained showed that the bands generated were distributed across the length of the lane and the average number of bands generated per clone was sufficient for fingerprint analysis.

When restriction enzyme fingerprint autoradiograms have been done, the computer does the job. The Contig9 genome mapping package, which was kindly offered by Dr Alan Coulson of Sanger Center, Cam-

bridge, UK, is currently run on a Sun sparystation 10 with a SunOS 4.0.5 operating system and an SGI indigo 2 with an IRIX 5.2 system. The package comprises data entry, matching, assembly, statistical analysis, etc. The data are entered by scanning the autoradiography films and interpreted by an image-processing system. Several graphics workstations connected to the local network are used for viewing, editing and analyzing the processed data.

At the beginning of the project, a typical film was chosen and edited by using the program IMAGE contained in Contig9, then a list of standard marker band positions from one marker lane on this film was placed to the file standard1.marker.dat. Every time this file was used to normalize the sample band positions.

In our system a Sharp JX-610 transparent scanner, which can scan films up to 304 mm wide and 431 mm length, is used for data entry. Scanning is controlled by the program PhotoStyler 1.1a, running on a 486/33 personal computer. It generated a greyscale image in TIFF format. The image data is then transferred to a SGI indigo 2 workstation through the net. To view and edit a film, activate the image-analysis software, called IMAGE, in the SGI workstation. The program IMAGE will analyze the scanned fingerprint autoradiograms, find the lanes, detect the sample bands on lanes, correct gel distortion by aligning known standard marker lanes and normalize their positions with respect to the marker bands. At each stage IMAGE displays all the results graphically and offers hand editing tools for correcting errors so we can manually delete artificial bands and interpolate bands which are not marked. Accepted bands are automatically digitized, normalized and written to the database.

To progressively assemble clones into overlapping groups, two programs MAPSUB and CONTIG (contained in Contig9) are used. The MAPSUB program takes all the text in the file Map.in, matches every clone of the new subset against the clones of the second subset with preset values of main tolerance and probability cutoff. The Map.in file was shown below.

- rice name of project
- SET1 start of first set (usually a newly scanned batch)
- & get clone set, followed by two clone names, marking beginning and end of set
- first clone name (of new subset)
- last clone name (of new subset)
- SET2 start of second set (usually the entire database)
- & get clone set,
- first clone name (of entire database)
- last clone name (of entire database)
- 7 main tolerance set
- 10 difmap tolerance set
- 0.1E-3 match probability cutoff set
- 0.1E-10 analytical probability cutoff set

Clone matching involves counting the number of bands in one clone that occupy the same position, within a preset tolerance, as bands on another clone. The results are written to a text file Map.out, which can be printed out as in Table I.

The signs in the above file were explained as follows. C/n denotes the clone name; No./b, the number of bands; ct./n, the contig name; No./mt, number of matches; mt/c, the matching clone; b/mt, the bands of matching clone; n/ct, the name of contig containing matching clone; pr, probability of match being coincidental; s/d, shorter distance in bands from end of matching clone to the end of its contig; x, number with d denotes the bands unmatched; number without d denotes number of bands in incoming clone not found in top matching clone, or number of such different bands found in subsequent matching clones.

At the next step of analysis, the program CONTIG (an interactive database management) was invoked. It provides automatic assembly options which are useful for placing clones in its most likely position within contigs. In our case, we depend on human judgement to decide precise extents of overlap when contigs are being extended or joined. For positioning a new incoming clone (for example y8803, see above), the clone name w18670 (this clone has the best match with the new clone) was entered and the corresponding contig 372 was displayed on the screen, and then the new clone name y8803 was entered with its contig number

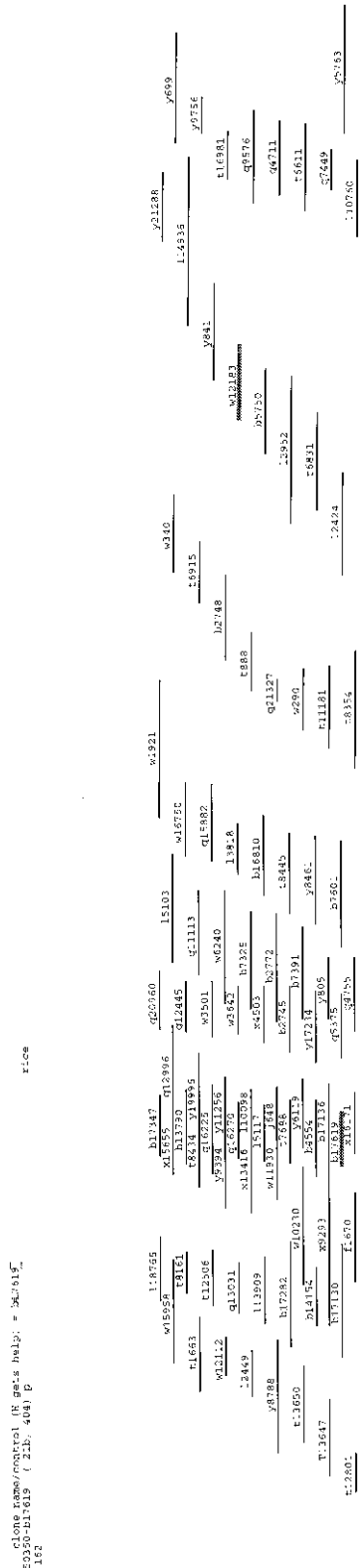


Figure 1. One of the contigs obtained. The horizontal lines denote the BAC inserts, the lengths of which were determined by the number of bands per clone displayed on fingerprint autoradiograms. This contig contained 92 clones and was of 2481 kb in length. Therefore, on average, every clone contributed 27 kb in length to the contig. Because the average insert of the library was 120 kb in length, the average number of redundant clones in this contig was 4.5.

Table 1. Data showing clone matching

| c/n | No./b | Ct./n | No./mt | mt/c | b/mt | n/ct | pr. | S/d | × |
|-------------|-------|-------|------------|--------------|-------|-------|---------|-----|-----|
| 20510-y8801 | (24b, | 0) | | | | | | | |
| | | | 15 matches | 80500-q1062 | (20b, | 1029) | 6.1e-12 | 18 | 9d |
| | | | 13 matches | 70280-w15442 | (18b, | 1029) | 3.6e-10 | 19 | 2 |
| | | | 15 matches | 60090-13845 | (25b, | 1029) | 6.1e-10 | 19 | 5 |
| | | | 13 matches | 10360-t14497 | (20b, | 1029) | 2.7e-09 | 22 | 3 |
| | | | 13 matches | 10581-t6864 | (20b, | 1029) | 2.7e-09 | 22 | 4 |
| 20510-y8803 | (43b, | 0) | | | | | | | |
| | | | 23 matches | 70040-w18670 | (32b, | 372) | 7.8e-12 | 27 | 20d |
| | | | 25 matches | 10320-t17519 | (37b, | 372) | 8.5e-12 | 14 | 4 |
| | | | 18 matches | 50350-b17593 | (27b, | 372) | 1.0e-08 | 16 | 5 |
| | | | 18 matches | 50350-b17595 | (31b, | 372) | 2.3e-07 | 15 | 6 |
| | | | 16 matches | 50342-b17582 | (27b, | 372) | 7.4e-07 | 0 | 3 |
| | | | 11 matches | 70420-w2578 | (19b, | 0) | 5.7e-05 | 0 | 6 |
| | | | 15 matches | 60200-t1297 | (33b, | 817) | 1.1e-04 | 18 | 5 |

0 being changed into contig number 372. The POS option was used to position the new incoming clone. The exact extent of its overlap with the existing clones was verified by examination of the MAPSUB's output file Map.out or visual inspection of the films, and then the position of the active clone could be adjusted by using the graphics cursor.

The contig construction by fingerprinting is a random process in nature, whereby all the clone fingerprints generated from the representative library must be compared with one another for creating clone matching, which will naturally lead to the clone redundancy in contigs. Redundancy is necessary for the correct clone overlap. Though over-redundancy exists in some contigs this is unavoidable. The BAC library used in our project consisted of 22 000 clones, which were of about 7 equivalents of the rice genome. Interestingly, the average number of redundant clones in contigs was quite in agreement with the degree of clone redundancy in the library. About 600 contigs with various lengths ranging from 500 kb to several megabases were obtained, the total length of which was nearly the size of the rice genome.

Figure 1 shows one of the contigs obtained. It can be seen in the figure that there was an ideal array of clones for clone overlapping on the right hand of the figure, though on the left hand as many as 16 redundant clones existed. The entire length of the contig was 2481 kb.

It was observed that with variation in distance between loci and between molecular markers, their order were conserved in the genomes of different spe-

cies. Seven years ago Gale's group found that the maps for each of the three diploid ancestors of hexaploid wheat, *Triticum urartu*, *Aegilops squarrosa* and the still unidentified donor of the B genome, are actually remarkably similar [3]. Later in 1993 Gale's group further revealed that the same loci in wheat, barley and rye were arranged in exactly the same order along large stretches of their chromosomes [9]. Comparative mapping of wheat, barley and rye chromosomes has revealed conserved linkage and identical genetic order among groups of DNA markers [8–11]. The order of molecular markers on rice chromosome segments has also been conserved in the genomes of wheat [14] and maize [1]. More recently, using various approaches Dunford *et al.* [12] have assessed the fine-scale genetic and physical colinearity between rice and *Triticeae*. They showed that the order of tightly linked DNA sequences separated by 1.6 cM or 1×10^6 bp in rice is conserved on the much large chromosomes of the *Triticeae*. Zhang (Q. F. Zhang, Huazhong Agricultural University, China, personal communication) has recently mapped nearly two hundred randomly selected molecular markers derived from Nipponbare, an *O. sativa japonica* variety, to the chromosomes of Ai Jiao Nang Te (a *O. sativa indica* variety) and revealed the identical order of the markers for the two rice varieties.

The contigs were assigned to particular chromosomes and were ordered by various molecular markers through their hybridization against the whole genomic library by robot Biomek-2000 (Beckman). Hybridiza-

tion was here used at the same time as an independent and powerful tool to further confirm the accuracy in contigs of the clone overlaps, which have been built up by fingerprinting. Because the contigs have already marked with genetic markers, the contig map thus constructed was integrated with the genetic map, from which the genetic markers derived.

The strategy described in this paper was rapid. A contig map of the rice genome was constructed within less than 2 years' time. The gaps between contigs may be caused by the lack of gap-bridging clones in the library or by the high threshold values set in the mapping program, which make it impossible to identify the gap-bridging clones that exist in the library.

Large DNA sequencing of individual rice chromosomes could be initiated based on the contig map by selecting and sequencing the minimal overlapped clones. Gaps may be filled up during the course of sequencing by, for example, clone walking through hybridization. If molecular markers of different rice varieties were used to assign and order the contigs, slight variation in the order of contigs was possible, which could be corrected in the gap filling procedures as well.

Acknowledgements

This work was completed by the research team of National Center for Gene Research, Chinese Academy of Sciences, and was supported by the funds from State Commission of Science and Technology, the Chinese Academy of Sciences and the commission of Science and Technology of Shanghai. G.F.H. was recipient of Biotechnology Career Fellowship of Rockefeller Foundation.

References

1. Ahn S, Tanksley SD: Comparative linkage maps of the rice and maize genomes. *Proc Natl Acad Sci USA* 90: 7980–7984 (1993).
2. Arumuganathan K, Earle ED: Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* 9: 208–218 (1991).
3. Chao S, Sharp PJ, Worland AJ, Warham EJ, Koebner RMD, Gale MD: RFLP-based genetic maps of the wheat homoeologous group 7 chromosomes. *Theor Appl Genet* 78: 495–504 (1989).
4. Chumakov IM, Rigault P, Le GI, Bellanne CC, Billault A *et al.*: A YAC contig map of the human genome. *Nature* 377 (suppl. 6547): 175–297 (1995).
5. Cohen D, Chumakov I, Weissenbach JA: A first generation map of the human genome. *Nature* 366: 698–701 (1993).
6. Coulson A, Sulston J: Genome mapping by restriction fingerprinting. In: Davies KE (ed) *Genome Analysis: A Practical Approach*, pp. 19–39. IRL Press, Oxford, UK (1988).
7. Coulson A, Kozono Y, Lutterbach B, Shownkeen R, Sulston J, Waterston R: YACs and the *C. elegans* genome. *BioEssays* 13: 413–417 (1991).
8. Devos KM, Atkinson MD, Chinoy CM, Liu C, Gal MD: RFLP-based genetic map of the homoeologous group 3 chromosomes of wheat and rye. *Theor Appl Genet* 83: 931–939 (1992).
9. Devos KM, Atkinson MD, Chinoy CM, Francis HAI, Harcourt RL, Koebner RMD, Liu CJ, Masojc P, Xie DX, Gale MD: Chromosomal rearrangements in the rye genome relative to that of wheat. *Theor Appl Genet* 85: 673–680 (1993).
10. Devos KM, Gale MD: Extended genetic maps of the homoeologous group 3 chromosomes of wheat, rye and barley. *Theor Appl Genet* 85: 641–652 (1993).
11. Devos KM, Millan T, Gale MD: Comparative RFLP maps of the homoeologous group-2 chromosomes of wheat, rye and barley. *Theor Appl Genet* 85: 784–792 (1993).
12. Dunford RP, Kurata N, Laurie DA, Monery TA, Minobe Y, Moore G: Conservation of fine-scale DNA marker order in the genomes of rice and the Triticeae. *Nucl Acids Res* 23: 14 (1995).
13. Haldi M, Perrot V, Saumier M, Desai T, Cohen D, Cherif D., Ward D, Lander ES: Large human YACs constructed in a *rad52* strain show a reduced rate of chimerism. *Genomics* 24: 478–484 (1994).
14. Kurata N, Moore G, Nagamura Y, Foote T, Yano M, Minobe Y, Gale M: Conservation of genome structure between rice and wheat. *Bio/technology* 12: 276–278 (1994).
15. Selleri L, Eubanks JH, Giovanninin M, Hermanson GG, Romo A, Diabali M, Maurer S, Mcelligott DL, Smith MW, Evans GA: Detection and characterization of 'chimeric' yeast artificial chromosome clones by fluorescent *in situ* suppression hybridization. *Genomics* 14: 536 (1992).
16. Shizuya H, Birren B., Kin UJ, Mancino V, Slepak T, Tachiir Y, Simon M: Cloning and stable maintenance of 300-kilo-base-pair fragment of human DNA in *E. Coli* using an f-factor-based vector. *Proc Natl Acad Sci USA* 89: 8794 (1992).
17. Sulston J, Mallett F, Durbin R, Horsnell T.: Image analysis of restriction enzyme fingerprint autoradiograms. *CABIOS* 5: 101 (1989).
18. Sulston J, Mallett F, Staden R, Durbin R, Horsnell T, Coulson A: Software for genome mapping by fingerprinting techniques. *CABIOS* 4: 125 (1988).
19. Tao QZ, Zhao HY, Qiu LF, Hong GF: Construction of a full bacterial artificial chromosome (BAC) library of *Oryza sativa* genome. *Cell Res* 4: 127–133 (1994).