Bot. Bull. Acad. Sin. (2001) **42:** 159-166

Characterization of a 10 cM region of rice chromosome 5

Teh-Yuan Chow, Ya-Ting Chao, Su-Mei Liu, Hong-Pang Wu, Mu-Kuei Chu, Ching-San Chen, and Yue-Ie C. Hsing*

Institute of Botany, Academia Sinica, Taipei 115, Taiwan, Republic of China

(Received May 31, 2000; Accepted January 3, 2001)

Abstract. Rice is a model species for the cereals and a good candidate for genome sequencing due to its relatively small genome (430 Mb), dense physical and genetic maps, and good transgenic systems. As part of an international effort to decode the rice genome, a PAC clone localized at 10 cM of chromosome 5 is completely determined for its sequence using shotgun libraries of its two inserts, 2-kb and 5-kb in length. In total 2,998 sequencing reads were used for the assembly of the final sequence, covering 175,439 bp. This sequence may code for at least 28 putative proteins, as deduced from computational search for homology with other known coding sequences and EST, or predicted using GenScan package. Also present in this sequence are simple repeats, palindrome and retrotransposons. On the basis of these findings, the gene density in the gene-rich region of rice genome is about 6 kb/gene.

Keywords: Annotation; High-throughput genome sequencing; Repetitive sequences; Retrotransposons; Rice genome.

Abbreviations: BAC, Bacterial artificial chromosome; **EST,** Expressed sequence tags; **LTR,** Long terminal repeat; **NR,** Non-redundant database; **ORF,** Open reading frame; **PAC,** P1-derived artificial chromosome.

Introduction

As the first genome of the higher plants, the small mustard species *Arabidopsis thaliana* will soon be completely sequenced. The sequences of chromosome 2 and 4 were recently published (Lin et al., 1999; Mayer et al., 1999). Additional knowledge of the genomes of other plant species is desirable to understand how plant genes evolved and are organized and regulated.

Rice (*Oryza sativa*) has been chosen as the first crop to be sequenced by an international sequencing consortium, the IRGSP (International Rice Genome Sequencing Project, Sasaki and Burr, 2000) for the following reasons: (1) Rice is an important crop in the world, feeding about one half of the world's population; (2) Rice's genome size, 430 Mb, is the smallest among crops (Arumuganathan and Earle, 1991); (3) Rice linkage and physical maps have been established (e.g. Harushima et al., 1998), and over 40,000 expressed sequence tags (ESTs) have been reported (Yamamoto and Sasaki, 1997) and mostly mapped. A yeast artificial chromosome (YAC) library that has been fingerprinted and ordered with mapped markers currently covers 60% of the rice genome (Kurata et al., 1997). Several bacterial artificial chromosome (BAC) libraries and P1-derived artificial chromosome (PAC) libraries have also been described. (4) The transgenic technology for rice has been established, and rice has become the easiest of all cereal plants to transform genetically. (5) Rice shares a co-linear gene organization with other cereal grasses and thus a key to knowledge of the genomic organization of the other grasses (Gale and Devos, 1998).

*Corresponding author. Tel: 02-2789-9590 ext. 312; Fax: 02-2782-7954; E-mail: bohsing@ccvax.sinica.edu.tw

IRGSP, of which this lab is a member, adopts a map-based clone-by-clone shotgun strategy. Sheared bacterial artificial chromosome/P1-derived artificial chromosome (PAC) libraries are constructed from *Oryza sativa* ssp. *japonica* variety "Nipponbare" by labs in the States and Japan, respectively. BAC end-sequencing, fingerprinting and marker-aided PCR screening are used to make sequence-ready contigs. Using these libraries and information, each IRGSP member is for high-throughput sequencing and subsequent annotation of one or more of the twelve chromosomes. In this international effort, this lab in Taiwan works on the sequencing work of chromosome 5.

In this report, we describe the sequencing strategy and the annotation method used in the study. We also present the characterization of all the putative open reading frames and its repetitive sequence in P0699E04, a contig with a sequence localized around 10 cM of the short arm of rice chromosome 5.

Materials and Methods

Sequencing

P0699E04 is a PAC clone of the *Hind*III PAC library constructed by members of the Japan Rice Genome Research Program (RGP) using the genomic DNA of the japonica rice Nipponbare with the vector pCYPAC2. Its PAC DNA was sheared (1.6-2 kb and 4.5-5 kb), ligated to a pUC18 vector, and transformed into *Escherichia coli*. Sequencing reactions were performed using either BigDye Terminators, BigDye primers, or Dichlororhodamine Terminators (P.E. Biosystems) and were run on ABI377 sequencers (P.E. Biosystems). Shotgun clones were sequenced

to generate at least 9-10 fold coverage. In total, about 4000 reactions were carried out to generate the sequence of the whole PAC clone. These sequences were then assembled using the Phred/Phrap/Consed package (developed by P. Green, E. Ewing and D. Gordon at the University of Washington).

Annotation

Annotation involved both DNA and protein database searches and gene prediction programs. The DNA sequences were searched against the non-redundant database using BLASTX (Altschul et al., 1997) and searched against the EST database using BLASTN (Altschul et al., 1997). Gene predictions were made by GenScan (Burge and Karlin, 1997; Burge and Karlin, 1998) trained for *Arabidopsis* or maize. These were confirmed with EST database whenever possible. Predicted protein sequences were searched against a non-redundant amino-acid database using BLASTP (Altschul et al., 1997). Output from the gene finding and signal detection programs was displayed using the Genotator viewer (Harris, 1996). Genes encoding tRNAs were predicted by tRNAscan-SE (Rivas and Eddy, 1999).

Analysis of DNA Sequences and the Putative Open Reading Frames

The repetitive DNA sequences were identified using Miropeat (Parsons, 1995) or RepeatMasker (A.F.A. Smith and P. Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html).

Each of the putative open reading frames were analyzed or scanned by several kinds of software, which included: BLOCKS — looking for the most highly conserved regions in groups of proteins documented in the Prosite Database (Henikoff et al., 1999); ChloroP — presenting the cleavage site scores for chloroplast protein (Emanuelsson et al., 1999); Peptidestructure — prediction of protein hydrophilicity and glycosylation sites; Pfam — multiple alignments match the majority of proteins (Bateman et al.,1999); PRINTS — a protein motif fingerprint database (Attwood and Beck, 1994); ProDom -- a database of protein domain families (Corpet et al., 1998); PROSITE Pattern — patterns defined in the PROSITE Dictionary of Protein Sites and Patterns (Hofmann et al., 1999); PROSITE Profile — detection of distantly related proteins (Eddy, 1998; Gribskov et al., 1987); PSORT — prediction of protein localization sites in cells (Nakai and Kanehisa, 1992); TMHMM — prediction of transmembrane helices in proteins (Sonnhammer et al., 1998).

Results

Sequence Assembly and its Quality

A total of about 4000 sequencing reads, including 3000 for the 2 kb clones and 1000 for the 5 kb clones, were collected and subjected to Phred/Phrap/Consed for assembly. The final assembly used 2998 reads, or 2,455,250 bp, with an average coverage about 14 fold. The insert length for the P0699E04 is 175,439 bp, and its Phred/Phrap scores are listed in Table 1. All the information for low quality bases is also displayed on our web site (http://biometrics.sinica.edu.tw/genome/index_e.htm). There are several mapping markers in the sequence, and the orientation of the six markers was C50503, E50988, S21107, E50955, R830, and C53640, arranged with the telomere at the 5' end and centromere at the 3' end. R830 is the RFLP marker and all the others are cDNA markers. According to the sequence we got, the orientation should be C50503, E50988, S21107, E50955, R830 and C53640. The complete sequence of P0699E04 and its annotated information bear the accession number AP001111.

Gene Identification and Structure

There are 28 genes in the 175,439 bp P0699E04 insert as predicted through computational search by the packages indicated in Materials and Methods. These genes were named ORF 1-28 as a working nomenclature. The average length of these genes is 2,978 bp. The main features, localization and similarity search results of each gene are described in Table 2 and Figure 1. Out of these 28 ORF, two ABC transporter genes are present tandem in this fragment, but in reverse direction. A Ca^{++}-ATPase, several DNA-binding proteins, and many other membrane proteins are also present. There was no tRNA gene in this 175 kb fragment, according to the analysis by tRNAscan-SE.

Table 1. The sequence quality of the P0699E04 DNA by Phred/Phrap analysis.

Quality score	Number of bases	Cumulative bases	Cumulative frequency
> 90	145883	145883	0.8315
80-90	11527	157410	0.8972
70-80	5678	163088	0.9296
60-70	6449	169537	0.9664
50-60	4263	173800	0.9907
40-50	1240	175040	0.9977
30-40	297	175337	0.9994
20-30	70	175407	0.9998
10-20	8	175415	0.9999
0-10	24	175439	1.0000

Table 2. The characterization of the 28 ORF found in P0699E04.

ORF	Protein ID	Exon	a.a	kD	pI	Putative characteristics	Predict localization	EST	NR
1	BAA90492.1	3	275	29	8.9	Homeodomain, bZIP protein	Nuclei	yes	Arabidopsis transcription factors, etc.
2	BAA90493.1	3	88	10	12.4	LEA IV protein; nuclear targeting signals	Nuclei	n.d.	n.d.
3	BAA90494.1	2	278	30	11.2	4.7 Transmembrane domains	Plasma membrane	yes	Arabidopsis hopothetical protein
4	BAA90495.1	2	177	18	5.6	Hydrophobic	ER membrane	n.d.	n.d.
5	BAA90496.1	2	113	13	12.5			n.d.	n.d.
6	BAA90497.1	2	170	17	8.4	With repeats		yes	n.d.
7	BAA90498.1	2	141	15	12.3	With repeats		yes	n.d.
8	BAA90499.1	2	91	10	11	Hydrophilic		n.d.	n.d.
9	BAA90500.1	3	227	26	4.7		Motichondria	yes	n.d.
10	BAA90501.1	3	174	19	4.7	Zinc finger protein	Nuclei	n.d.	Arabiopsis RING protein
11	BAA90502.1	3	895	101	4.7	With signal peptide, glycoprotein		n.d.	Arabidopsis hypothetical protein
12	BAA90503.1	2	236	26	11.6	Myc-type helix-loop-helis domain	Nuclei	n.d.	n.d.
13	BAA90504.1	2	95	10	12.3	LEA IV protein		yes	n.d.
14	BAA90505.1	3	133	14	6.3	HMG DNA binding domain	Nuclei	n.d.	n.d.
15	BAA90506.1	3	266	29	4.3		Cytoplasm	n.d.	Similar to retrotransposon RIRE2 orf5
16	BAA90507.1	2	654	71	9.5	5 Transmembrane domains		yes	Arabidopsis ABC transporter
17	BAA90508.1	2	705	76	9.5	6 Transmembrance domains, P-loop		yes	Arabidopsis ABC transporter
18	BAA90509.1	4	522	55	11.2	With signal peptide, glycoprotein		yes	n.d.
19	BAA90510.2	8	1055	114	5.7	8 Transmembrane domains	Plasma membrane	yes	Rice Calcium ATPase
20	BAA90511.1	6	1018	110	5.7	With nuclear targeting signals	Nuclei	n.d.	Rice unknown protein
21	BAA90512.1	8	314	34	6.6	With signal peptide, hydrophobic		n.d.	n.d.
22	BAA90513.1	2	170	19	10.5	With nuclear targeting signals	Nuclei	n.d.	n.d.
23	BAA90514.1	1	292	31	10			yes	n.d.
24	BAA90515.1	4	334	37	9.7	4 Transmembrane domains		yes	n.d.
25	BAA90516.1	7	363	41	6.8			yes	n.d.
26	BAA90517.1	9	273	30	5	With signal peptide, glycoprotein		yes	n.d.
27	BAA90518.1	7	299	32	7.3	Hydrophobic at the C-terminus		n.d.	Arabidopsis LRR-like protein
28	BAA90519.1	11	449	50	8.7	C2H2 type Zinc finger protein	Nuclei	yes	Human transcription factor TFIIIA

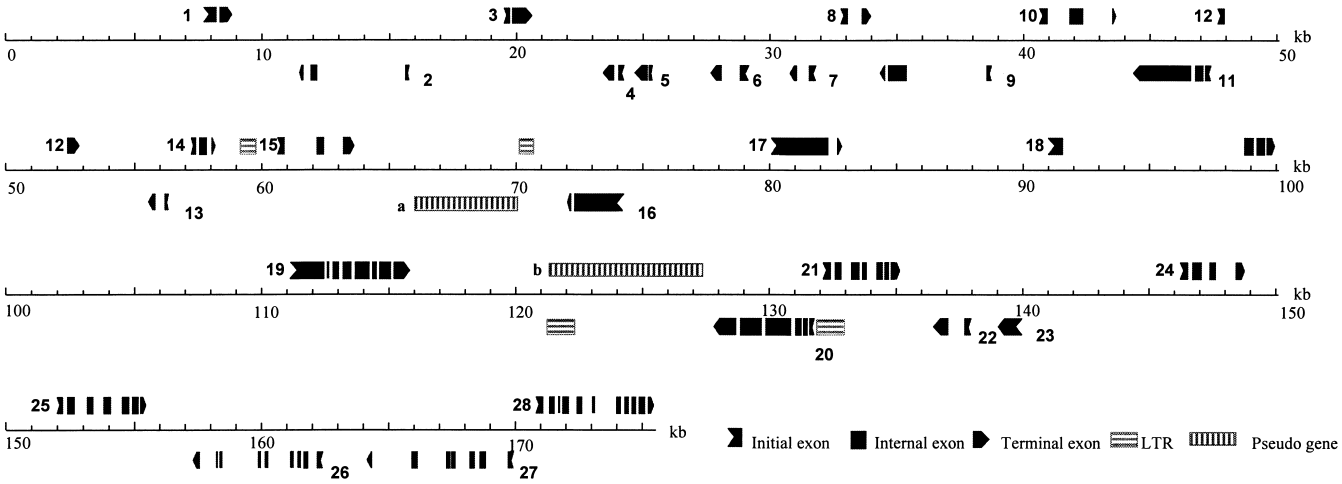

Figure 1. Organization of genes on P0699E04. The display shows the directions and exon-intron structure of the annotated genes. The pseudogenes and LTR are also illustrated.

Analysis of Intergenic Regions

Altogether, the 28 predicated genes, including exons and introns, account for about 50.93% of the 175 kb contig. In other words, about half of the P0699E04 sequence is intergenic regions. The overall GC content of the contig is 44.26%, with an average content of 61.89% in exons and an average content of 39.94% in other region (introns plus intergenic region).

Several kinds of repetitive sequences are present in the intergenic regions, including retrotransposons, short repeated motifs of mono-, di-, tri-, tetra- or pantanucleotides, direct or invert repeats, and palindromes. To search for these repetitive sequences present in the 175 kb contig, two packages were used, as indicated in Materials and Methods.

The RepeatMasker was used to screen for different classes of simple repeats present in the sequences. There are many AT-rich, CT-rich, GA-rich, GC-rich regions, with lengths ranging from 20 bp to about 100 bp. For instance, the sequence from bp 2 to bp 27 is an AT-rich one, with the sequence of AAAATTTTATTTATAATATTTATTAT. The sequence from bp 1701 to bp 1784 is another AT-rich one, with the sequence of TAAATTTATTATAAAAATA TTTTTAATTATTAATTTAATAAACTTAATTTGGTAA TATAAAATATTACTATATTTGTATATAAA. There are also many simple repeats like (G)n, (TA)n, (TC)n, (CGG)n, (CCG)n, (TCC)n, (CCA)n, (CAG)n, (CGA)n, (TCG)n, (TTAA)n, (TTTTC)n, (CCGGG)n where n ranged from 3 to 60. These are typical microsatellite motifs. Generally they were found upstream from the 5'-UTR of several genes and sometimes in the ORF of specific genes. For instance, the sequence from bp 10562 to bp 10581 is a (TA) n repeat, with the sequence of TATATATATATATATATATA, and is localized at the intergenic region. The sequence from bp 52771 to bp 52790 is a (G)n repeat, with the sequence of GGGGGGGGGGGGGGGGGGG, and is part of ORF12.

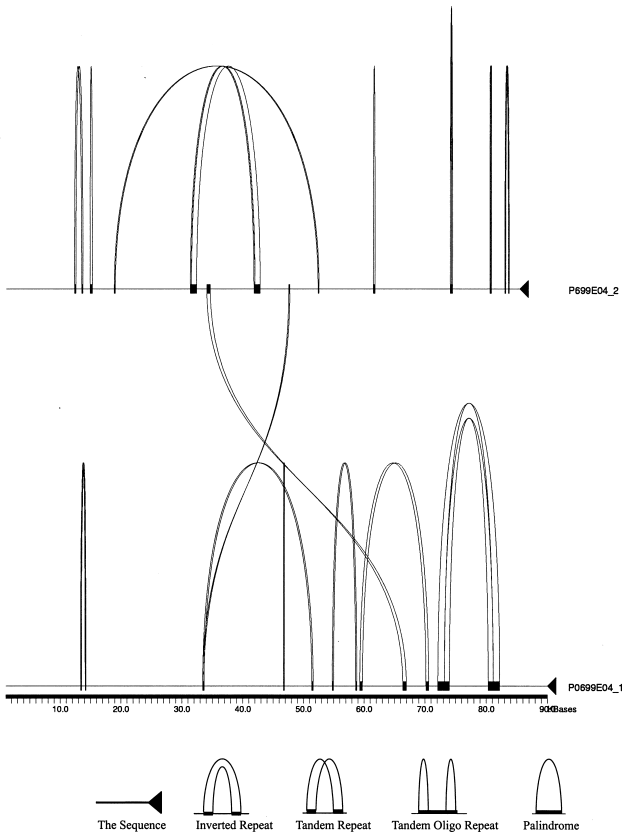

Figure 2. The repeats present in the P0699E04 sequence analyzed by Miropeats. The threshold value is set to 100. The lower section includes the first base to 90,000 bp, and the upper section includes 90,001 bp to 175,439 bp. Keys to graphics are illustrated as above.

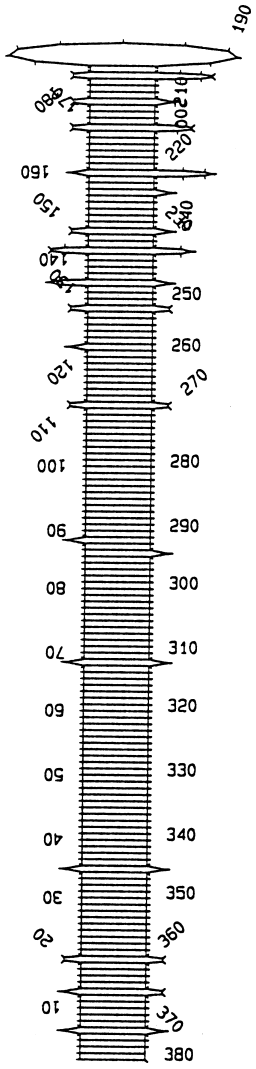

Figure 3. The stem-loop structure of the palindrome 104,882 bp-105,262 bp.

Miropeats (Parsons, 1995) was used to search for the presence of direct repeats, invert repeats and palindromes present in sequences. Figure 2 illustrates that many of these repeats are indeed present in the 175 kb conitg, including two palindromes, three invert repeats, and many direct repeats. The length and type of these repeats, and the occurrence of these repeats in other PAC/BAC genome sequences available now, are listed in Table 2. The stem-loop structure of one of the palindromes is indicated in Figure 3. The length of these repeats ranged from about 135 bp to 1,000 bp. Some of these repeats occur very often in the rice genome while some are very rare.

Two pairs of the direct repeats are long terminal repeats (LTR) of retrotransposons, the first pair are 59272-59714 and 70145-70587, and the others are 121315-122341 and 131769-132797. According to the results of BLASTX, polyproteins, the characteristics of retrotransposons, are in the regions between the two terminal repeats in each pair. The distances between the two terminal repeats of each set, about 11 kb, also fitted well with the average length of retrotransposon. Thus, two retrotransposons were in the 175 kb contig of P0699E04. The one in the 5' side (59 kb to 90 kb) contained the orf5 of a gypsy-type retrotransposon and a pseudo polyprotein, with many stop codons in its reading frame, between the LTRs. The one in the 3' side (121 kb to 132 kb) contained a pseudo polyprotein and an unknown protein which shared similarity with another unknown protein present in rice genome. The length of the LTRs of the two retrotransposons was also quite different: the former one had the LTR of 443 bp in length, and the latter had one of 1027 bp. The two LTR sets had no sequence homology.

Discussion

Sequence Quality

The IRGSP has adopted the standards of the Human Genome Project, which sets a standard of less than one base-pair error in 10,000 bp. Although this level of accuracy is difficult to verify, it is achievable through a combination of high-quality shotgun sequence reads, at least seven-fold redundancy, and the insistence that 97% of all bases are sequenced on both strands or that two sequencing chemistries are used. Thus, the requirement set by IRGSP is that "ninety percent of the bases should have a Phred score higher than 70, and ninety-nine percent should have one higher than 40" (Sasaki and Burr, 2000). Table 1 illustrates the sequence quality of P0699E04. The two Phred scores for this clone are 96.64% and 99.94%, indicating the sequence quality meets the standard demanded by IRGSP.

Gene Identification and Density

Of the 28 genes identified by the prediction programs in the 175,439 bp contig, only one lacks intron while the others contain from one to ten. The gene density is about 6.3 kb per gene, very similar to those obtained from other rice BAC/PAC clones sequenced. For rice plant, the genome size is 430 Mb, and the estimated number of rice genes is about 30,000, thus the average gene density should be about 14.5 kb per gene. Our work has used BAC/PAC clones pulled out by ES and RFLP markers, thus a gene-rich region may have been selected. This phenomenon has also been pointed out in the sequencing projects of Arabidopsis chromosome 2 and chromosome 4 (Lin et al., 1999; Mayer et al., 1999)

For the 28 predicted genes in P0699E04, nine of them are similar to sequences in EST databases; five of them are similar to sequences in NR databases; six of them are similar to sequences in both EST and NR databases; and eight of them do not have any hit in the database. These numbers are also very similar to those of *Arabidopsis* genomic clones (Lin et al., 1999; Mayer et al., 1999). However, many studies indicate that the annotation of long stretches of genomic sequences generally results in many mistakes. For instance, the 60 kb around the *Arabidopsis thaliana*

Table 3. The presence of the simple repeats in 52 rice BAC/PAC clones available in the database[a]. BLASTN comparison was used. The subject fragment >75% of query fragment in length with >80% sequence identity would be counted.

Fragment position	Fragment length (bp)	Repeat type	Clone number (fragment number)				
			Chrom. 1	Chrom. 4	Chrom. 5	Chrom. 6	Chrom. 10
13347-13481	135	Direct repeat	–	–	1 (2)[b]	–	–
33406-33591	186	Direct repeat	14 (22)	1 (1)	1 (2)	6 (6)	3 (3)
54755-54954	220	Direct repeat	–	–	1 (2)	–	–
59272-59714	443	LTR of retrotrans.	2 (4)	–	1 (2)	–	2 (4)
73170-73982	803	Direct repeat	–	–	1 (2)	–	–
102318-102545	228	Direct repeat	17 (38)	2 (3)	1 (2)	8 (19)	8 (14)
104882-105262	381	Palindrome	–	–	1 (1)	–	–
108856-109019	164	Direct repeat	5 (8)	–	1 (2)	3 (3)	1 (1)
121315-122341	1027	LTR of retrotrans.	4 (8)	–	1 (2)	4 (8)	–
124026-124597	572	Invert repeat	3 (3)	–	1 (2)	2 (2)	–
137518-137703	186	Invert repeat	14 (26)	1 (1)	1 (2)	6 (7)	3 (3)
142355-142518	164	Direct repeat	7 (10)	1 (1)	1 (2)	4 (4)	1 (1)
151458-151743	286	Palindrome	10 (12)	1 (2)	1 (1)	3 (5)	3 (5)
164068-164254	187	Invert repeat	3 (4)	–	1 (2)	–	–
172973-173108	136	Direct repeat	–	–	1 (2)	–	–

[a]There were 52 rice genome BAC/PAC sequences available in the database at the end of May, 2000.
[b]BAC/PAC clone number (fragment number).

AtEm1 locus on chromosome 3, analyzed by the commonly used prediction programs specifically trained for *Arabidopsis*, i.e. GENSCAN, GeneFinder and GeneMark, failed to locate any of the exons of the ten genes in that region (Comella et al., 1999). Thus, the exon status of the predicted genes in our study should be viewed with reservation until the putative gene product can be identified.

Simple Repeats

As Table 2 shows, many simple repeats are present in the sequences of P0699E04. Several of the direct repeats occur only rarely, e.g. those within the 13347-13881 fragment. However, several of the direct repeats occur very often, e.g. those within the 102318-102545 fragment. Similarly, of the two palindromes, one does not have a match in the database, while the other one appears frequently in other genome in the rice database. Our observation is consistent with reports on *Arabidopsis* chromosome 2 and 4, which showed that the occurrence of simple repeats is region specific (Mayer et al., 1999).

Retrotransposons in Rice Genome

Retrotransposons, a group of mobile elements that transpose via the RNA intermediate, are important components of the eukaryotic genomes (Boeke and Corces, 1989). The structure of retrotransposons resembles that of integrated retroviruses, with long terminal repeats (LTRs), andan internal domain encoding a group-specific antigen, and a polyprotein (Pol). The Pol region has conserved domains characteristic of protease, reverse transcriptase, integrase, and RNase H genes, and this region is present in both retrotransposons of P0699E04.

Many clones containing retrotransposons are currently in the annotated rice BAC/PAC database. About three-fourths of them are LTR-retrotransposons, as those present in P0699E04. Many of the rice retrotransposons, such as the two described here, are pseudogenes because they are interrupted by stop codons or frameshifts. Thus, the rice genome indeed contains many retrotransposons, although many are inactive.

Acknowledgements. The authors are grateful to RGP, Japan for providing the rice PAC clones. We also express our deep thanks to Dr. P.C. Huang for a critical reading of the manuscript and to Academia Sinica Computing Center for technical support. This research is supported by grants from the National Science Council, Council of Agriculture and Academia Sinica, Taiwan to T.Y.C, H.P.W., C.S.C. and Y.I.H. This work is the effort of all members in the Academia Sinica Plant Genome Center.

Literature Cited

Altschul, S.F., T.L. Madden, A.A. Axhaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389-3402.

Arumuganathan, K. and E.D. Earle. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **3:** 208-218.

Attwood, T.K. and M.E. Beck. 1994. PRINTS—a protein motif fingerprint database. Protein Engin. **7:** 841-848.

Bateman A., E. Birney, R. Durbin, S.R. Eddy, R.D. Finn, and E.L. Sonnhammer. 1999. Pfam 3.1: 1313 multiple alignments match the majority of proteins. Nucleic Acids Res. **27:** 260-262.

Boeke, J.D. and V.G. Corces. 1989. Transcription and reverse transcription of retrotransposons. Annu. Rev. Microbiol. **43:** 403-434.

Burge, C.B. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268:** 78-94.

Burge, C.B. and S. Karlin. 1998. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. **8:** 346-354.

Comella, P., H. Wu, M. Laudie, C. Berger, R. Cooke, M. Delseny, and F. Grellet. 1999. Fine sequence analysis of 60 kb around the *Arabidopsis thalana AtEm1* locus on chromosome III. Plant Mol. Biol. **41:** 687-700.

Corpet, F., J. Gouzy, and D. Kahn. 1998. The ProDom database of protein domain families. Nucleic Acids Res. **26:** 323-326.

Eddy, S.R. 1998. Profile hidden Markov models. Bioinformatics **14:** 755-763.

Emanuelsson, O., H. Nielsen, and G. von Heijne. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage site. Protein Sci. **8:** 978-984.

Gale, M.D. and K.M. Devos. 1998. Comparative genetics in the grasses. Proc. Natl. Acad. Sci. USA **95:** 1971-1974.

Gribskov, M., A.D. McLachlan, and D. Eisenberg. 1987. Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. USA **84:** 4355-4358.

Harris, N.L. 1996. Genotator: A workbench for sequence annotation. Genome Res. **7:** 754-762.

Harushima, Y., M. Yano, A. Shomura, M. Sato, T. Shimano, Y. Kuboki, T. Yamamoto, S.Y. Lin, B.A. Antonio, and A. Parco. 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population. Genetics **148:** 479-494.

Henikoff, S., J.G. Henikoff, and S. Pietrokovski. 1999. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics **15:** 471-479.

Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. Nucleic Acids Res. **27:** 215-219.

Kurata, N., Y. Umehara, H. Tanoue, and T. Sasaki. 1997. Physical mapping of the rice genome with YAC clones. Plant Mol. Biol. **35:** 101-113.

Lin, X., S. Kaul, S. Rounsley, T.P. Shea, M Benito, C.D. Town, C.Y. Fujii, T. Mason, C.L. Bowman, M. Barnstead, T.V. Feldblyum, C.R. Buell, K.A. Ketchum, J. Lee, C.M. Ronning, H.L. Koo, K.S. Moffat, L.A. Cronin, M. Shen, G. Pai, S. van Aken, L. Umayam, L.J. Tallon, J.E. Gill, M. D. Adams, A.J. Carrera, T.H. Creasy, H.M. Goodman, C. R. Somerville, G.P. Copenhaver, D. Preuss, W.C. Nierman, O. White, J.A. Eisen, S.L. Salzberg, C.M. Fraser, and J.C. Venter. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402:** 761-768.

Mayer, K., C. SchoLler, R. Wambutt, G. Murphy, G. Volckaert, T. Pohl, A. DuSterhoFt, W. Stiekema, K.-D. Entian, N. Terryn, B. Harris, W. Ansorge, P. Brandt, L. Grivell, M. Rieger, M. Weichselgartner, V. De Simone, B. Obermaier, R. Mache, M. Muller, M. Kreis, M. Delseny, P. Puigdomenech, M. Watson, T. Schmidtheini, B. Reichert, D. Portatelle, M. Perez-Alonso, M. Boutry, I. Bancroft, P. Vos, J. Hoheisel, W. Zimmermann, H. Wedler, P. Ridley, S.-A. Langham, B. Mccullagh, L. Bilham, J. Robben, J. VanDer Schueren, Doggett, S. Hall, M. Kay, N. Lennard, K. Mclay, R. Mayes, A. Pettett, M.-A. Rajandream, M. Lyne, V. Benes, S. Rechmann, D. Borkova, H. BloCker, M. Scharfe, M. Grimm, T.-H. LuHnert, S. Dose, M. De Haan, A. Maarse, M. SchoFer, S. Muller-Auer, C. Gabel, M. Fuchs, B. Fartmann, K. Granderath, D. Dauner, A. Herzl, S. Neumann, A. Argiriou, D. Vitale, R. Liguori, E. Piravandi, O. Massenet, F. Quigley, G. Clabauld, A. McNdlein, R. Felber, S. Schnabl, R. Hiller, W. Schmidt, A. Lecharny, S. Aubourg, F. Grymonprez, Y.-J. Chuang, F. Vandenbussche, M. Braeken, I. Weltjens, M. Voet, I. Bastiaens, R. Aert, E. Defoor, T. Weitzenegger, G. Bothe, U. Ramsperger, H. Hilbert, M. Braun, E. Holzer, A. Brandt, S. Peters, M. Van Staveren, W. Dirkse, P. Mooijman, R. Klein, Lankhorst, M. Rose, J. Hauf, P. KoTter, S. Berneiser, S. Hempel, M. Feldpausch, S. Lamberth, H. Van Den Daele, A. De Keyser, C. Buysshaert, J. Gielen, R. Villarroel, R. De Clercq, M. Van Montagu, J. Rogers, A. Cronin, M. Quail, S. Bray-Allen, L. Clark, J. Chefdor, C. Cooke, A. Berger, E. Montfort, T. Casacuberta, N. Gibbons, N. Weber, M. Vandenbol, M. Bargues, J. Terol, A. Torres, A. Perez-Perez, B. Purnelle, E. Bent, S. Johnson, D. Tacon, T. Jesse, L. Heijnen, S. Schwarz, P. Scholler, S. Heber, P. Francs, C. Bielke, D. Frishman, D. Haase, K. Lemcke, H. W. Mewes, S. Stocker, P. Zaccaria, M. Bevan, R.K. Wilson, M. De La Bastide, K. Habermann, L. Parnell, N. Dedhia, L. Gnoj, K. Schutz, E. Huang, L. Spiegel, M. Sehkon, J. Murray, P. Sheet, M. Cordes, J. Abu-Threideh, T. Stoneking, J. Kalicki, T. Graves, G. Harmon, J. Edwards, P. Latreille, L. Courtney, J. Cloud, A. Abbott, K. Scott, D. Johnson, P. Minx, D. Bentley, B. Fulton, N. Miller, T. Greco, K. Kemp, J. Kramer, L. Fulton, E. Mardis, M. Dante, K. Pepin, L. Hillier, J. Nelson, J. Spieth, E. Ryan, S. Andrews, C. Geisel, D. Layman, H. Du, J. Ali, A. Shohdy, A. Hasegawa, A. Hameed, M. Lodhi, A. Johnson, E. Chen, M. Marra, R. Martienssen, and W.R. Mccombie. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana.* Nature **402:** 769-777.

Nakai, K. and M. Kanehisa. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics **14:** 897-911.

Parsons, J.D. 1995. Miropeats: graphical DNA sequence comparisons. Comput. Applic. Biosci. **11:** 615-619.

Rivas, E. and S.R. Eddy. 1999. A dynamic programming algorithm for RNA structure including pseudoknots. J. Mol. Biol. **285:** 2053-2068.

Sasaki, T. and B. Burr. 2000. International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. Curr. Op. Pl. Biol. **3:** 138-141.

Sonnhammer, E.L.L., G. von Heijne, and A. Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *In* J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff and C. Sensen (eds.), Proc. of Sixth Int. Conf. on Intelligent Systems for Molecular Biology, Menlo Park, CA: AAAI Press, pp. 175-182.

Yamamoto, K. and T. Sasaki. 1997. Large-scale EST sequencing in rice. Plant Mol. Biol. **35:** 135-144.

水稻第五條染色體 10 cM 片段的定性

周德源　趙雅婷　劉淑美　鄔宏潘　朱木貴　陳慶三　邢禹依

中央研究院植物研究所

水稻是世界上最重要的糧食作物之一。又由於它具有較小的基因組（430 百萬鹼基對），具備很密的物理輿圖及遺傳輿圖，以及轉殖系統已被建立，水稻成為最適合於進行基因組定序分析的作物。台灣參加了世界合作團隊、進行解讀水稻的遺傳密碼。水稻有十二條染色體，我們所參與的是第五條染色體的大量定序與序列分析。在第五條染色體短臂約 10 cM 處有一個 PAC 殖系 —P0699E04，本文敘述此殖系被定序與定性的情形。我們以亂槍法製備了 2 kb 與 5 kb 的基因庫，並進行了約 4000 個定序反應，其中的四分之三的序列被用以進行序列組合，得到了 P0699E04 的全長為 175,439 鹼基對。解讀的結果指出它有 28 個可能的基因，有些與已知的蛋白質或 EST 序列類似，而有些則為全新的蛋白質。另外、我們也研究了包括一般重複序列、反轉錄跳躍子等重複 DNA 序列。

關鍵詞：序列解讀；大量基因組定序；重複序列；水稻基因組。