Identification and Mapping of Expressed Genes, Simple Sequence Repeats and Transposable Elements in Centromeric Regions of Rice Chromosomes

Hiroshi Mizuno,1 Kazue Ito,2 Jianzhong Wu,1 Tsyoshi Tanaka,1 Hiroyuki Kanamori,2 Yuichi Katayose,1 Takuji Sasaki,1 and Takashi Matsumoto1,*

National Institute of Agrobiological Sciences, 1-2, Kannondai 2-chome, Tsukuba, Ibaraki 305-8602, Japan1 and Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, 446-1 Ippaizuka, Kamiyokoba, Tsukuba, Ibaraki 305-0854, Japan2

(Received 2 November 2006; revised 14 January 2007; published online 12 February 2007)

Abstract

The genomic sequences derived from rice centromeric regions were analyzed to facilitate the comprehensive understanding of the rice genome. A rice centromere-specific satellite sequence, RCS2/TrsD/CentO, was used to screen P1-derived artificial chromosome (PAC) and bacterial artificial chromosome (BAC) genomic libraries derived from Oryza sativa L. ssp. japonica cultivar Nipponbare. Physical maps of the centromeric regions were constructed by DNA fingerprinting methods and the aligned clones were analyzed by end sequencing. BLAST analysis revealed the composition of genes, centromeric satellites and other repetitive elements, such as RIRE7/CRR, RIRE8, Squiq, Anaconda, CACTA and miniature inverted-repeat transposable elements. Fiber-fluorescent in situ hybridization analysis also indicated the presence of distinct clusters of RCS2/TrsD/CentO satellite interspersed with other elements, instead of a long homogeneous region. Several expressed genes, sequences representative of ancestral organelar insertions, relatively long simple sequence repeats (SSRs), and sequences corresponding to 5S and 45S ribosomal RNA genes were also identified. Thirty-one gene sequences showed high-similarity to rice full-length cDNA sequences that had not been matched to the published rice genome sequence in silico. These results suggest the presence of expressed genes within and around the clusters of RCS2/TrsD/CentO satellites in unsequenced centromeric regions of the rice chromosomes.

Key words: genome; FL-cDNA; centromere; transposon

1. Introduction

The centromere of a chromosome plays a key role during mitosis and meiosis. The centromeric region contains species-specific satellite sequences and various kinds of transposable elements1–6 as well as genes.7 These highly repetitive sequences complicate many aspects of genomic sequencing—including cloning, mapping, chromosomal walking and assembling the fragments of DNA sequences in silico—so that genomic sequencing of the centromeric regions of higher eukaryotes can be difficult.

Therefore, the genes in centromeric regions could not be fully elucidated, even for the so-called completely sequenced genomes.8,9 Rice is considered to be a model monocot plant because of its relatively small genome size and its high-synteny with other cereal crops.10 As the first step in the functional characterization of the rice genome, the International Rice Genome Sequencing Project (IRGSP) completed a high-quality map-based sequencing of the genome of Oryza sativa L. ssp. japonica cv. Nipponbare.11–15 Even though it covers >95% of the estimated 390 Mb total genome sequence, the published rice genome sequence includes the centromeric regions of chromosomes 4, 5 and 8, and only portions of the...
centromeric regions of the 9 remaining chromosomes. The overall sequence data corresponds to only 13% of the full set of rice centromeric regions.

The rice centromeric region consists mainly of repeats of 155–165 bp RCS2/TrsD/CentO satellite DNA and the centromere-specific retroelement RIRE7/CRR.16–21 In addition to the repetitive elements, high-quality sequence data of the centromeres of chromosomes 4 and 8 revealed that genes are predicted to lie between and around the satellite sequences.16,17 and 12 of these genes are expressed in leaf and root tissues.18 Do other rice centromeric regions contain active genes? Although the rice Genome Research Program (RGP) attempted to construct sequence-ready PAC/BAC physical maps of the centromeric regions, interference with repetitive sequences prevented further chromosomal walking and subsequent genomic sequencing of the regions.11

Here we describe a comprehensive analysis that identifies and maps centromeric sequences derived from the 12 rice chromosomes. We reveal the occurrence of expressed genes, organellar insertions, relatively long simple sequence repeats (SSRs), and ribosomal RNA genes in rice centromeric regions. In addition, we describe the composition and distribution of centromeric repetitive sequences. Finally, we evaluate our PAC/BAC library and discuss strategies that might be used to complete the genomic sequencing of rice centromeric regions.

2. Materials and Methods

2.1. PCR screening

PAC and BAC libraries were constructed from genomic DNA derived from the rice cultivar Nipponbare (JP 229579 in NIAS GenBank, Oryza sativa L. ssp. japonica) that was generated by the Rice Genome Research Program. PCR screening was performed as described previously.22 We designed four sets of primers and selected the set that most efficiently amplified the RCS2/CentO/TrsD array from the DNA pools. We confirmed that we could use this primer set to obtain clones that had already been mapped by chromosomal walking.11,22 DNAs from 192 PAC clones randomly selected from the library and 545 RCS2/CentO/TrsD-containing PAC/BAC clones were isolated using standard alkaline lysis procedure from 2 ml overnight cultures in Luria–Bertani (LB) medium containing kanamycin (40 μg/ml) or chloramphenicol (12.5 μg/ml). Insert sizes of these clones were determined by pulsed-field gel electrophoresis (PFGE). PAC/BAC DNAs were digested by NotI and separated in 1% agarose gels in 0.5× TBE at 6 V/cm with a 5–40 s pulse time for 14 h.

2.2. BAC/PAC end sequencing

The BAC and PAC clones were grown for 22 h with shaking. DNA of each BAC/PAC clone was prepared in a 96-well plate using a liquid-handling robot (Quadra 3, Tomtec, Hamden, CT, USA). Bacterial cells were harvested by centrifugation and lysed according to the standard alkaline lysis protocol. The insert DNA was purified using isopropanol precipitation, dissolved in deionized H2O and sheared into ~1 kb fragments by sonication (Sonifier 450, Branson, Danbury, CT, USA). Sequencing was performed using the Big Dye terminator reaction mix (ABI, Foster City, CA, USA) on ABI3700 capillary sequencers (ABI). Repetitive elements were identified by BLAST analysis using the TIGR repetitive sequences library (http://www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml). Sequence similarity search was performed against the rice full-length complementary DNA (FL-cDNA) sequences23,24 including previously unmapped sequences (identity <95%, total coverage <90%).

2.3. DNA fingerprinting

Agarose-gel-based fingerprinting methodology25 was used to generate HindIII fingerprints of PAC/BAC clones. PAC/BAC DNAs were isolated using a REAL kit (Qiagen, Hilden, Germany) and then digested with HindIII. The restriction digestion fragments were electrophoresed in 1.2% agarose gels. The gels were then stained with SYBR Green I (Invitrogen, Carlsbad, CA, USA). Positions of DNA bands on the gel images were identified using the program Image25 and checked manually. DNA fragments derived from vector DNA (one fragment from pBeloBAC11 or four fragments from pCYPAC2) were removed manually. Fingerprints derived from PAC/BAC clones in pericentromeres were integrated to determine the chromosomal location of each contig. The initial fingerprint assembly was performed automatically using fpc V4.725–27 at a Sulston score cut-off of E−12 and a tolerance of 2. To correct errors, the contigs were edited manually at cutoff of E−8 to E−13 and a tolerance of 2.

2.4. Genetic mapping

Genetic mapping was performed according to previously published methods26 with minor modifications. In brief, after a BLAST search against the TIGR rice repeat database and the RGP genomic sequence database, unique sequences were selected. Rice molecular linkage map using 186 F2 plants derived from a single cross between the japonica cultivar Nipponbare and the indica cultivar Kasalath was used.

2.5. Fluorescence in situ hybridization (FISH)

Fiber-FISH analysis was performed according to previously published methods29 with minor modifications. PCR fragments amplified from centromere specific repeats or retrotransposon sequences in the centromeric
region of chromosome 8 were used as DNA probes. DNA fragments of ~5 kb were amplified by PCR and then mixed to make a DNA probe for retrotransposons. Probes were labelled with digoxigenin- or biotin-labelled dUTP and detected with a fluorescein isothiocyanate-conjugated anti-digoxigenin antibody or Cy3-conjugated avidin. All fluorescent images were captured using a fluorescence microscope (BX51, Olympus, Tokyo, Japan) with a charge-coupled device camera (CoolSNAP HQ, Roper Scientific, Tucson, AZ, USA).

3. Results and Discussion

3.1. Screening of PAC/BAC clones that cover the 12 centromeric regions of rice

In this work we sought to comprehensively screen PAC/BAC clones from the 12 rice centromeric regions. Because all rice centromeres contain centromere-specific tandem repeats, RCS2/TrsD/CentO satellite sequences,19 we screened the two rice genomic libraries using RCS2/TrsD/CentO as a probe. We obtained a total of 545 clones—398 PAC clones and 147 BAC clones, which correspond to 0.57 and 0.30% of the total clones in the PAC and BAC libraries, respectively (Table 1).

In addition to the RCS2/TrsD/CentO satellite sequence, other rice centromere-specific motifs include the RCH1, RCH2, RCH3 and RCS1 sequences.19 Though we screened both the PAC and BAC libraries with these sequences, only four positive clones in total were obtained (data not shown). Because these retrotransposon-derived centromeric sequences may demonstrate high-diversity,17,20 they seem to be unsuitable as probes for overall screening of rice centromere-specific clones.

3.2. Composition and structure of repetitive sequences in rice centromeric regions

We analyzed the sequence of both ends for a total of 545 rice centromere-specific clones (Supplementary Table 1). BLASTN analysis using various databases revealed the composition of centromeric sequences. The analysis identified the RCS2/TrsD/CentO satellite sequence and various repetitive elements, including RIRE7/CRR, RIRE8, Squiq, miniature inverted-repeat transposable elements (MITEs), Anaconda and CACTA, in rice centromeric regions (Table 2). The RCS2/TrsD/CentO satellite accounted for 34.7% of the total region sequenced in this study, and RIRE7/CRR, a centromere-specific Ty3/gypsy type retroelement of rice,20,21 accounted for 14.9% (Table 2). Thus, these 2 centromere-specific repetitive elements comprised 49.6% of the sequenced region. The MITE family30,31 accounted for 12.7% of the total region sequenced (Table 2). MITEs have a pronounced bias toward euchromatic regions11,13,14 and are considered to be associated with genes.32,33 This propensity suggests that genes might be present near the MITEs in rice centromeric regions.

A cytogenetic approach was applied to analyze the structure of these repetitive sequences. Fiber-FISH analysis using RCS2/TrsD/CentO and three repetitive sequences we described in Table 2 as probes indicated that RCS2/TrsD/CentO sequences are arranged in discrete clusters, instead of being distributed uniformly throughout the centromeric region of rice (Fig. 1A). This result is somewhat contrary to the previous finding from pachytene FISH analysis,21 but recent sequence analysis from rice chromosomes 416 and 817,18 supports our finding. The repetitive elements RIRE3, RIRE7/CRR

Table 1. Centromeric PAC/BAC clones

|        | Total no. of clones | No. of centromeric clones | %   |
|--------|---------------------|---------------------------|-----|
| PAC    | 69 276              | 398                       | 0.57|
| BAC    | 48 960              | 147                       | 0.30|
| Total  | 118 236             | 545                       | 0.46|

MITIE, miniature inverted-repeat transposable element; SINE, short interspersed nuclear element; FL-cDNA, full length cDNA.
and RIRE8 seem to be dispersed throughout centromeric regions, with no obvious clustering (typical images, Fig. 1B–D). Fiber-FISH analysis illustrates the complex structure of rice centromeric regions and supports previous estimates regarding their composition in Table 2.

### 3.3. Identification of expressed genes, SSRs and organellar insertions in rice centromeric regions

To identify expressed genes in rice centromeric regions, we conducted a sequence similarity search between the end sequences of the screened PAC/BAC clones and rice full-length cDNA (FL-cDNA) sequences in silico. BLASTN analysis indicated that 31 end sequences had high-similarity (E value < E−05) to rice FL-cDNA sequences with no match to published rice genome sequences in silico (Table 3). These results suggest that genes are present and expressed in unsequenced centromeric regions of the rice chromosomes. Whereas the putative proteins encoded by 11 FL-cDNA sequences were similar to various types of known proteins, such as ribose-phosphate diphosphokinase, urease and transcription factors.

---

**Figure 1.** Structure of the centromeric region revealed by FISH analysis of extended DNA fibers (fiber-FISH). Typical fluorescent images from DNA fibers probed with RCS2/TrsD/CentO (green, A–E) and various repetitive sequences (red)-RIRE3 (B), RIRE7/CRR (C), RIRE8 (D), and 5S rDNA (E) are shown. Scale bar, 5 μm.
factors, those encoded by the remaining FL-cDNA sequences had no pronounced similarity to any known protein (Table 3). We searched genes in Arabidopsis thaliana, which had sequence similarity with rice centromeric genes described in Table 3 by using MIPS (Munich Information center for Protein Sequences). Sixteen rice genes had sequence similarity with those in A.thaliana but 15 rice genes had no similarity (data not shown). These 16 genes were not located in the centromeric regions of A.thaliana, indicating that the chromosomal location of these genes is different between rice and A.thaliana.

We identified SSRs in rice centromeric regions. Among the 47 distinctive motif families of rice SSRs, only the AT motif had been found in centromeres (Table 2). It is interesting that although the AT motif in centromeric region was repeated more than 200 times (Supplementary Table 1) only 2 loci with >200 AT repeats had previously been identified among the 18 828 SSR loci in the rice genome, suggesting that AT repeats might be relatively long in centromeric regions for the rice genome overall.

The sequences we identified also include data suggestive of insertion of ancestral organellar DNA in the rice

| Unmapped rice FL-cDNA Accession no. | PAC/BAC end sequence | Score | E-value | Name of encoded protein | Score | E-value |
|-------------------------------------|----------------------|-------|---------|-------------------------|-------|---------|
| AK071067                            | P0613F05_R           | 827   | 0       | Hypothetical protein    | 255   | 4.00E-67 |
| AK067951                            | B1048E04_F           | 581   | 1.00E-167| Ribose-phosphate diphosphokinase | 505 | 1.00E-142 |
| AK069198                            | P0672A03_F           | 492   | 1.00E-140| Unknown                |       |         |
| AK063425                            | P0639D05_F           | 254   | 4.00E-68 | Zinc finger protein     | 110   | 4.00E-25 |
| AK121415                            | P0511A04_F           | 250   | 5.00E-67 | Hypothetical protein    | 462   | 1.00E-129|
| AK121105                            | P0483C05_F           | 248   | 5.00E-66 | Cleavage and polyadenylation specificity factor | 186 | 8.00E-46 |
| AK070768                            | P0700F03_F           | 246   | 1.00E-65 | Unknown                |       |         |
| AK120382                            | B1001D06_R           | 236   | 7.00E-63 | Unknown                |       |         |
| AB075476                            | P0498E11_R           | 204   | 3.00E-53 | Urease                 | 1216  | 0       |
| AK069176                            | P0498E11_R           | 204   | 4.00E-53 | Urease                 | 1216  | 0       |
| AK073065                            | P0639D05_F           | 200   | 4.00E-52 | Zinc finger protein     | 74.7  | 1.00E-12 |
| AK120053                            | P0700F03_F           | 196   | 7.00E-51 | Chitinase              | 587   | 1.00E-167|
| AK107762                            | P0495C03_F           | 186   | 3.00E-48 | Unknown                |       |         |
| AK111153                            | P0577E01_F           | 161   | 5.00E-40 | Unknown                |       |         |
| AK109830                            | P0622D06_F           | 145   | 2.00E-35 | Protein F3F19.15 of A.thaliana | 407 | 1.00E-113|
| AK106598                            | P0577E01_F           | 143   | 9.00E-35 | Unknown                |       |         |
| AK105573                            | P0672D12_F           | 111   | 2.00E-25 | Unknown                |       |         |
| AK107728                            | P0630F09_F           | 103   | 3.00E-23 | Unknown                |       |         |
| AK101043                            | P0495C03_R           | 82    | 3.00E-16 | Unknown                |       |         |
| AK108608                            | P0571C08_R           | 80    | 6.00E-16 | Unknown                |       |         |
| AK070410                            | P0708D04_F           | 76    | 2.00E-14 | Unknown                |       |         |
| AK106650                            | P0495C03_R           | 74    | 6.00E-14 | F-box family protein   | 68    | 4.00E-12 |
| AK064426                            | P0038H11_R           | 66    | 2.00E-11 | Unknown                |       |         |
| AK065444                            | B1121D11_F           | 64    | 5.00E-11 | Unknown                |       |         |
| AK063480                            | P0504C07_F           | 60    | 7.00E-10 | Unknown                |       |         |
| D21313                              | P0559C02_F           | 56    | 1.00E-09 | Wilm's tumor suppressor | 337 | 8.00E-90 |
| AK105461                            | P0550C11_F           | 56    | 2.00E-09 | Unknown                |       |         |
| AK066246                            | P0705D08_F           | 56    | 2.00E-08 | Unknown                |       |         |
| AK121903                            | P0672D12_F           | 54    | 1.00E-07 | Unknown                |       |         |
| AK109863                            | B1116G02_F           | 52    | 2.00E-07 | Unknown                |       |         |
| AK062276                            | P0495C03_R           | 50    | 1.00E-06 | F-box family protein   | 65    | 7.00E-11 |

Table 3. BLAST analysis of genes at rice centromeres
nuclear genome. P0698F05, F is similar to the DNA sequence that encodes mitochondrial NADH dehydrogenase subunit 1 (Supplementary Table 1). This result suggests that mitochondrial insertions occur even in centromeric regions and similar organellar insertions account for 0.18–0.19% of the published nuclear genome of rice.11

3.4. Mapping of PAC/BAC sequence data, genes and repetitive sequences

We constructed clone-based physical maps of the 12 rice centromeric regions. A total of 545 centromere-specific clones and 42 selected clones (position markers, Supplementary Table 2) were subjected to DNA fingerprinting analysis to make contigs. Contigs were mapped to a specific centromeric region if they contained marker clones. The locations of four newly mapped PAC/BAC sequences corresponded to genetically identify centromeric regions of the 12 rice chromosomes. In total, 13 contigs and 4 clones were mapped genetically or in silico (Supplementary Figure 1), corresponding to 2.98 Mb of the 12 rice centromeric regions. In addition, 2 contigs totaling 360 kb in size have not been anchored yet, and 2 FL-cDNA sequences (AK067951, AK063425) belonged to unmapped contigs (Table 3, Supplementary Figure 1). Our findings suggest that these contigs might exist as gene-containing islands in centromeric regions.

After we mapped the 13 contigs, we then mapped the genes and transposable elements described in section 3.2. Mapping the 13 contigs allowed us to map the genes and transposable elements described earlier to their respective chromosomes in silico so there was no need to map each gene or motif individually (Table 4). In addition to the repetitive elements, the 5S rDNA genes were mapped on chromosome 11 (Table 4). It was previously reported that the 5S rDNA cluster lies close to the centromeric region on the short arm of chromosome 11.34 We confirmed the presence of the single long cluster of 5S rDNA near the RCS2/TrsD/CentO cluster by using fiber-FISH and found that the gap between these 2 clusters is <50 kb (Fig. 1E). The 45S rDNA consisting of 17S, 5.8S and 25S rDNA coding units was identified on chromosome 1 (Table 4). However, fiber-FISH and mitotic FISH both failed to identify the 45S rDNA cluster near the RCS2/TrsD/CentO satellite (data not shown); the 45S rDNA cluster previously had been detected at the end of the short arm of chromosome 9.35 Our finding suggests that 45S rDNA sequences are not clustered but are dispersed or fragmented in the centromeric region.

We encountered difficulty in constructing a complete physical map using PAC/BAC clones because the clones only minimally extended the contigs that had been constructed by chromosomal walking. Overall, 71% (427 of 598) of the PAC/BAC clones have not been mapped because they contain too few DNA fragments to facilitate assigning into contigs by DNA fingerprinting methods (Supplementary Figure 2). The distribution of insert length indicates that 47.2% (188 of 398) of centromeric PAC clones have inserts <60 kb (Fig. 2), compared with 13.6% in the total library.36 Moreover 10.3% (41 of 398) of PAC clones seem to have no insert (data not shown). This finding suggests that inserts containing repetitive sequences are easily deleted from PAC clones, as we also experienced when we constructed genomic libraries in YAC vector (unpublished data). Thus it may be necessary to use more efficient cloning vectors to obtain genomic clones containing the overall centromeric region of rice.

3.5. Toward the complete identification of genes in centromeric regions

In this study, we obtained PAC/BAC clones and constructed a partial physical map of unsequenced rice centromeric regions (Supplementary Figure 1), which could be useful for analysis of the centromeres by clone-by-clone genomic sequencing and subsequent

| Table 4. Chromosomal location of repetitive sequences in rice centromeric regions |
|---------------------------------|----------------------|
| Sequence name                  | Chromosome no.       |
| 45S rDNA (18S-5.8S-25S)        | 1                    |
| 5S rRNA                        | 11                   |
| (AT)n                          | Unknown              |
| Ac/Ds                          | 4                    |
| Anaconda                       | 1, 5, 12             |
| CACTA                          | 1, 8                 |
| CACTG                          | 6, 8                 |
| Castaway                       | 1, 5, 8              |
| Ditto                          | 4, 5, 6              |
| Explorer                       | 1                    |
| Gaijin                         | 6                    |
| Karma                          | 2                    |
| Mariner                        | 2                    |
| Pangrangja                     | 8                    |
| pSINE1                         | 4, 12                |
| Retrosat1                      | 10                   |
| Rim2                           | 8                    |
| RIRE7/CRR                      | 1, 2, 4, 5, 6, 7, 8, 10, 11, 12 |
| RIRE8                          | 1, 4, 8              |
| RIRE9                          | 11                   |
| Squiq                          | 1, 2, 5              |
| Tourist                        | 1, 8                 |
| Wonderer                       | 10                   |
| Wuji                           | 1                    |
| Youren                         | 4, 8                 |
prediction of genes. Comparison between the end sequences of the PAC/BAC clones and sequences of rice FL-cDNAs (Table 3) suggested that expressed genes exist within and around the RCS2/TrsD/CentO clusters in the centromeric regions. Previous studies have also identified new genes in the centromeres of chromosomes 4, 8, 17, and 18, and our current findings suggest that other rice centromeric regions contain many more expressed genes. An overall comparison between the published rice genome sequence and total rice FL-cDNAs was conducted recently and showed that 2257 of 28,469 FL-cDNA sequences do not match the rice genome, suggesting that the unsequenced regions, especially centromeric regions, contain many genes. The centromeric PAC/BAC clones we have identified may be used as probe for DNA microarrays comprising the entire set of FL-cDNAs in order to provide further insights into expressed genes in rice centromeric regions. In our strategy, however, we detected genes with evidence of expression, and we did not consider genes without transcriptional evidence. In addition, genes silenced by DNA methylation may also exist due to the flanking repetitive sequences/transposons in the centromeric region.

It is important to note that our discussion is limited to the centromeric regions represented by the PAC/BAC clones we obtained. Although we tried to analyze rice centromeric sequences comprehensively, the sequences we obtained in this study may not completely cover the entire rice centromeric region. The percentage of RCS2/TrsD/CentO in the rice genome data is ~1.8% (6.9 of 390 Mb) according to an estimate based on the signal intensity of RCS2/TrsD/CentO satellites in FISH analysis but only 0.46% in the PAC/BAC libraries used here (Table 1). This suggests that the PAC/BAC clones obtained is insufficient to cover the overall rice centromeric region. In addition, we sequenced only the ends of the clones. Because the average read length was 536 bp after trimming of the vector sequence, a total of 0.58 Mb (including redundancy) was sequenced, corresponding to 8.4% (0.58 of 6.9 Mb) of the estimated length of the 12 rice centromeric regions.

Complete genomic sequencing of the remaining centromeric regions, which are composed mainly of repetitive elements, clearly presents a considerable challenge. We consider that almost all remaining regions, typically long clusters of RCS2/TrsD/CentO satellites (Figure 1), would not be clonable because they contain few restriction enzyme sites and because the clones would be unstable in *Escherichia coli* (Fig. 2). Therefore, PAC/BAC clone-based sequencing and other essentially different methods for genome analysis, such as fiber-FISH, is likely needed to be used in concert to achieve a strategy that will be effective in completely characterizing the 12 rice centromeric regions.

**Acknowledgements:** The authors thank Dr B. A. Antonio for critical reading of the manuscript. This work was supported by Grants-in-Aid for Scientific Research for Young Scientists (B) 17710165 from the Ministry of Education, Culture, Sports, Science and Technology of Japan to H.M., and grants GD-2007 from the Ministry of Agriculture, Forestry and Fisheries of Japan.

**Supplementary data:** Supplementary data are available online at www.dnares.oxfordjournals.org.

**REFERENCES**

1. Henikoff, S. and Dalal, Y. 2005, Centromeric chromatin: what makes it unique? *Curr. Opin. Genet. Dev.*, 15, 177–184.
2. Hall, A. E., Keith, K. C., Hall, S. E., Copenhaver, G. P., and Preuss, D. 2004. The rapidly evolving field of plant centromeres, *Curr. Opin. Plant Biol.*, 7, 108–114.

3. Houben, A. and Schubert, I. 2003. DNA and proteins of plant centromeres, *Curr. Opin. Plant Biol.*, 6, 551–560.

4. Lamb, J. C. and Birchler, J. A. 2003. The role of DNA sequence in centromere formation, *Genome Biol.*, 4, 214.

5. Dawe, R. K. 2005. Centromere renewal and replacement in the plant kingdom, *Proc. Natl Acad. Sci. USA*, 102, 11573–11574.

6. Feschotte, C., Jiang, N., and Wessler, S. R. 2002. Plant transposable elements: where genetics meets genomics, *Nat. Rev. Genet.*, 3, 329–341.

7. Cooke, H. J. 2004. Silence of the centromeres—not, *Trends Biotechnol.*, 22, 319–321.

8. Kumeekawa, N., Hosouchi, T., Tsuruoka, H., and Kotani, H. 2000. The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 5, *DNA Res.*, 7, 315–321.

9. Kumeekawa, N., et al. 2001. The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 4, *DNA Res.*, 8, 285–290.

10. Devos, K. M. 2005. Updating the ‘crop circle’, *Curr. Opin. Plant Biol.*, 8, 155–162.

11. International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome, *Nature*, 436, 793–800.

12. Sasaki, T., Matsumoto, T., Yamamoto, K., et al. 2002. The genome sequence and structure of rice chromosome 1, *Nature*, 420, 312–316.

13. Feng, Q., Zhang, Y., Hao, P., et al. 2002. Sequence and analysis of rice chromosome 4, *Nature*, 420, 316–320.

14. Rice Chromosome 10, Sequencing Consortium. (2003). In-depth view of structure, activity, and evolution of rice chromosome 10. *Science*, 300, 1566–1569.

15. Rice Chromosomes 11 and 12 Sequencing Consortia. (2003). The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications, *BMC Biol.*, 3, 1–18.

16. Zhang, Y., Huang, Y., Zhang, L., et al. 2004. Structural features of the rice chromosome 4 centromere, *Nucleic Acids Res.*, 32, 2023–2030.

17. Wu, J., Yamagata, H., Hayashi-Tsugane, M., et al. 2004. Composition and structure of the centromeric region of rice chromosome 8, *Plant Cell*, 16, 967–976.

18. Nagaki, K., Cheng, Z., Ouyang, S., et al. 2004. Sequencing of a rice centromere uncovers active genes, *Nat. Genet.*, 36, 138–145.

19. Dong, F., Miller, J. T., Jackson, S. A., Wang, G. L., Ronald, P. C., and Jiang, J. 1998. Rice (*Oryza sativa*) centromeric regions consist of complex DNA, *Proc. Natl Acad. Sci. USA*, 95, 8135–8140.

20. Kumeekawa, N., Ohmirio, N., Fukui, K., Ohtsubo, E., and Ohtsubo, H. 2001. A new gypsy-type retrotransposon, RIREf, preferential insertion into the tandem repeat sequence TrsD in pericentromeric heterochromatin regions of rice chromosomes, *Mol. Genet. Genomics*, 265, 480–488.

21. Cheng, Z., Dong, F., Langdon, T., et al. 2002. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon, *Plant Cell*, 14, 1691–1704.

22. Wu, J., Mizuno, H., Hayashi-Tsugane, M., et al. 2003. Physical maps and recombination frequency of six rice chromosomes, *Plant J.*, 36, 720–730.

23. Kikuchi, S., Satoh, K., Nagata, T., et al. 2003. Collection, mapping, and annotation of over 28 000 cDNA clones from japonica rice, *Science*, 301, 376–379.

24. The Rice Annotation Project (2007). Curated Genome Annotation of *Oryza sativa* ssp. *japonica* and Comparative Genome Analysis with *Arabidopsis thaliana*, *Genome Res.*, 17, 175–183.

25. Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T., and Coulson, A. 1988. Software for genome mapping by fingerprinting techniques, *Comput. Appl. Biosci.*, 4, 125–132.

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

27. Ness, S. R., Terpstra, W., Krzywinski, M., Marra, M. A., and Jones, S. J. 2002. Assembly of fingerprint contigs: parallelized FPC, *Bioinformatics*, 18, 484–485.

28. Harushima, Y., Yano, M., Shomura, A., et al. 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population, *Genetics*, 148, 479–494.

29. Mizuno, H., Wu, J., Kanamori, H., et al. 2006. Sequencing and characterization of telomere and subtelomere regions on rice chromosomes 1S, 2L, 6L, 7S, 7L and 8S, *Plant J.*, 46, 206–217.

30. Bureau, T. E. and Wessler, S. R. 1992. Tourist: a large family of small inverted repeat elements frequently associated with maize genes, *Plant Cell*, 4, 1283–1294.

31. Bureau, T. E. and Wessler, S. R. 1994. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants, *Plant Cell*, 6, 907–916.

32. Bureau, T. E. and Wessler, S. R. 1994. Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses, *Proc. Natl Acad. Sci. USA*, 91, 1411–1415.

33. Miao, L., Wood, T. C., Yu, Y., et al. 2000. Rice transposable elements: a survey of 73 000 sequence-tagged-connectors, *Genome Res.*, 10, 982–990.

34. Kamisugi, Y., Nakayama, S., Nakajima, R., Ohtsubo, H., Ohtsubo, E., and Fukui, K. 1994. Physical mapping of the 5S ribosomal RNA genes on rice chromosome 11, *Mol. Gen. Genet.*, 245, 133–138.

35. Shishido, R., Sano, Y., and Fukui, K. 2000. Ribosomal DNAs: an exception to the conservation of gene order in rice genomes, *Mol. Gen. Genet.*, 263, 586–591.

36. Baba, T., Katagiri, S., Tanoue, H., et al. 2000, Construction and characterization of rice genomic libraries: PAC library of *Japonica* Variety, Nipponbare and BAC library of *Indica* Variety, Kasalath, *Bull. Natl Inst. Agrobiol. Resour.*, 14, 41–52.