The Plant Cell, Vol. 16, 967–976, April 2004, www.plantcell.org © 2004 American Society of Plant Biologists

Composition and Structure of the Centromeric Region of Rice Chromosome 8 [W]

Jianzhong Wu, Harumi Yamagata, Mika Hayashi-Tsugane, Saori Hijishita, Masaki Fujisawa, Michie Shibata, Yukiyo Ito, Mari Nakamura, Miyuki Sakaguchi, Rie Yoshihara, Harumi Kobayashi, Kazue Ito, Wataru Karasawa, Mayu Yamamoto, Shoko Saji, Satoshi Katagiri, Hiroyuki Kanamori, Nobukazu Namiki, Yuichi Katayose, Takashi Matsumoto, and Takuji Sasaki[1]

Rice Genome Research Program, National Institute of Agrobiological Sciences/Institute of the Society for Techno-Innovation of Agriculture, Forestry, and Fisheries, Tsukuba, Ibaraki 305-8602, Japan

Understanding the organization of eukaryotic centromeres has both fundamental and applied importance because of their roles in chromosome segregation, karyotypic stability, and artificial chromosome-based cloning and expression vectors. Using clone-by-clone sequencing methodology, we obtained the complete genomic sequence of the centromeric region of rice (*Oryza sativa*) chromosome 8. Analysis of 1.97 Mb of contiguous nucleotide sequence revealed three large clusters of CentO satellite repeats (68.5 kb of 155-bp repeats) and >220 transposable element (TE)–related sequences; together, these account for ~60% of this centromeric region. The 155-bp repeats were tandemly arrayed head to tail within the clusters, which had different orientations and were interrupted by TE-related sequences. The individual 155-bp CentO satellite repeats showed frequent transitions and transversions at eight nucleotide positions. The 40 TE elements with highly conserved sequences were mostly *gypsy*-type retrotransposons. Furthermore, 48 genes, showing high BLAST homology to known proteins or to rice full-length cDNAs, were predicted within the region; some were close to the CentO clusters. We then performed a genome-wide survey of the sequences and organization of CentO and RIRE7 families. Our study provides the complete sequence of a centromeric region from either plants or animals and likely will provide insight into the evolutionary and functional analysis of plant centromeres.

INTRODUCTION

In multicellular eukaryotes, the heterochromatic component known as the centromere plays an important role in both mitotic and meiotic nuclear divisions. The centromere functions in holding sister chromatids together during the early stages of mitosis and serves as the site for assembly of the kinetochore multiprotein complex, which binds to microtubules that mediate chromosome separation at later stages. Many studies have focused on repetitive sequences located at, or near, the centromeres and their protein binding regions to identify the sequences responsible for the centromeric activity of higher eukaryotes (for a review, see Cleveland et al., 2003). Except for that of budding yeast (*Saccharomyces cerevisiae*), the structure of eukaryotic centromeres contains various types of elements with repetitive sequences at the center, including satellite DNA, and contains retrotransposons and transposons in the flanking regions. Although centromeres play the same functional roles in different eukaryotes, they vary in size from several hundred kilobases to several megabases and even lack sequence homology among various organisms. Human centromeres are estimated to be 3 to 4 Mb long in each chromosome and are composed primarily of ~171-bp AT-rich repeats (α-satellites) that are tandemly arrayed in a head-to-tail arrangement (Waye and Willard, 1987; Choo et al., 1991; Schueler et al., 2001). In *Drosophila melanogaster*, the centromeres contain highly repeated satellite sequences, AATAT and TTCTC, interspersed with transposable elements (Sun et al., 2003). In higher plants, the centromere is estimated to range in size between 2 and 9 Mb and appears to be composed of tandemly arrayed satellites of 150 to 178 bp (Kaszas and Birchler, 1996; Ananiev et al., 1998; Copenhaver et al., 1999; Cheng et al., 2002; Hosouchi et al., 2002). Because of the highly heterochromatic structure of centromeres, completely cloning, sequencing, and assembling their genomic components have remained a significant challenge.

Rice (*Oryza sativa*) is one of the most important cereal crops and has been used as a model plant for genome sequencing (Sasaki and Burr, 2000). Genetic studies have led to the establishment of a high-density linkage map that contains >3200 loci on its 12 chromosomes (Harushima et al., 1998; http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html). Cytological studies have shown that the central region of rice centromeres is composed of 155-bp satellite repeats, amounting to ~7 Mb in the *japonica* cultivar Nipponbare (Dong et al., 1998; Cheng et al., 2002). In preparation for sequencing the entire rice genome, we have focused on constructing a physical map of each rice chromosome. Among the 12 rice chromosomes, chromosome 8

[1] To whom correspondence should be addressed. E-mail tsasaki@nias.affrc.go.jp; fax 81-29-838-2302.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Takuji Sasaki (tsasaki@nias.affrc.go.jp).
[W] Online version contains Web-only data.
Article, publication date, and citation information can be found at www.plantcell.org/cgi/doi/10.1105/tpc.019273.

contains the least amount of satellite DNA sequence—estimated to be only ~64 kb by fiber fluorescence in situ hybridization (FISH) (Cheng et al., 2002). We have successfully constructed a contig map covering the centromeric region of rice chromosome 8 using P1-derived artificial chromosomes (PACs) and BACs through PCR screening by DNA markers, fingerprinting, end-sequencing of candidate clones, as well as gap filling by end walking. This article reports the complete genomic sequence of the centromeric region of rice chromosome 8. Important features of and implications for the functional rice centromere are discussed.

RESULTS

Physical Map and Sequence Assembly

Previous studies of dosage analysis and genetic mapping of DNA markers showed that the centromeric position of rice chromosome 8 was located within a region spanning 3.5 centimorgan on the rice linkage map (Singh et al., 1996; Harushima et al., 1998). During construction and chromosomal mapping of yeast artificial chromosome contigs using rice EST and centromere-specific satellite sequences, the centromeric position of this chromosome was found to be confined to ~2 Mb (Wu et al., 2002). This region was located at 54.3 centimorgan on the linkage map and had DNA markers that exhibited no recombination among 186 F2 plants (http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html). The complete, sequence-ready PAC/BAC physical map that covers the genetically defined centromeric region (from marker C1374 to S21882S) of rice chromosome 8 is shown in Figure 1. Two Monsanto BAC clones and six Rice Genome Research Program (RGP) PAC or BAC clones were chosen, respectively, from the results of in silico mapping and PCR screening using 11 sequence tagged site/EST marker sequences (Wu et al., 2003). Seven BAC clones were selected from the Clemson University Genomics Institute (CUGI)-assembled BAC contigs. The remaining three clones were screened from the RGP BAC library based on chromosome walking. These 18 clones formed the tiling path of the entire region and were used for genomic sequencing.

Out of the 18 PAC or BAC clones, sequences of 16 clones were relatively easy to assemble. However, sequences of two BAC clones, OSJNBa0061E21 and B1052H09, that contained the CentO repeats were found to be severely misassembled. Incorrect repeat assemblies were detected by discriminating one or two base differences between the aligned repeat sequences with high quality values using CONSED. Subclones of each sequence subtype were subjected to sequencing by the transposon insertion/sequencing method. The resulting subclone sequences, typically 5 kb in length, were fixed as the continuous sequences and were reassembled with other shotgun sequences of the PAC/BAC clones (MK trace function of PHRAP). These processes were repeated until the assembled PAC/BAC sequences agreed with the experimental data. In total, 10 shotgun subclones were fully sequenced and were used for contig assembly of the CentO and other repetitive sequences. All sequence gaps were closed. Finished assemblies of the sequences were confirmed by comparing sizes of the fragments from restriction enzyme digests of each PAC/BAC clone with those of the virtual fragment sizes of the finished sequence. DNA gel blot hybridization of the digested DNAs of these two clones using the CentO sequence as a probe also confirmed the final assemblies. This enabled us to assemble the 18 PAC/BAC sequences into a single 1,972,546-bp contiguous DNA sequence (Figure 1). By comparing the overlapping sequences of adjacent clones, we estimated the overall accuracy of the finished sequence to be >99.99%. There were only 194 ambiguous nucleotides dispersed within the clusters of CentO satellite repeats. Clones and sequence information have been submitted to the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/).

Base Composition and Gene Content

Overall analysis of base composition in the obtained 1.97-Mb sequence revealed an average G+C content of 45.2%, which is slightly higher than that observed for the entire sequences of chromosomes 1, 4, and 10 (Feng et al., 2002; Sasaki et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003). We identified G+C- and A+T-rich regions (Figure 2A). The distribution patterns of most of the G+C-rich regions matched those of two types of retrotransposable elements, RIRE3 (G+C content, 51.2%) and hopi (G+C content, 62.0%) (Figures 2D and 3). Gene prediction for the 1.97-Mb sequence revealed the presence of 201 open reading frames (excluding those of DNA transposons and retrotransposons) with an average gene size of 1767 bp (Figure 3, see Supplemental Table 1 online). This average gene size is much smaller than those predicted for the complete sequences of chromosomes 1 (3.4 kb), 4 (2.8 kb), and 10 (2.6 kb). Among these predicted genes, we found 48 sequences (23%) highly similar to known proteins or sequences of rice full-length cDNA clones as shown in Supplemental Table 1 (Rice Full-Length cDNA Consortium, 2003). Two putative genes, TGF-beta receptor-interacting protein and defective chloroplasts and leaves protein chloroplast precursor, were located very near the centromere domain (only ~8 kb and 4 kb, respectively, away from the CentO clusters). The remaining 81% of the predicted genes encode hypothetical proteins. This high proportion might indicate the existence of as yet unknown genes specific to the centromere or the decreased reliability of prediction programs when applied to unusual (e.g., highly repetitive) sequences, such as those of centromeres.

Repetitive Sequences

On the basis of the sequence analysis with the RepeatMasker and BLASTN programs, 59% of the 1.97-Mb centromeric sequence was repetitive and mostly identical to retrotransposable elements (Figures 2B to 2E). There were very few miniature inverted repeat transposable elements or MITEs within the centromeric region (Figure 2E). The BAC clones OSJNBa0061E21 and B1052H09 located in the central part of the 1.97-Mb sequence contained many CentO repeats (Figure 1). We found three sequence clusters (34,589, 7616, and 26,292 bp

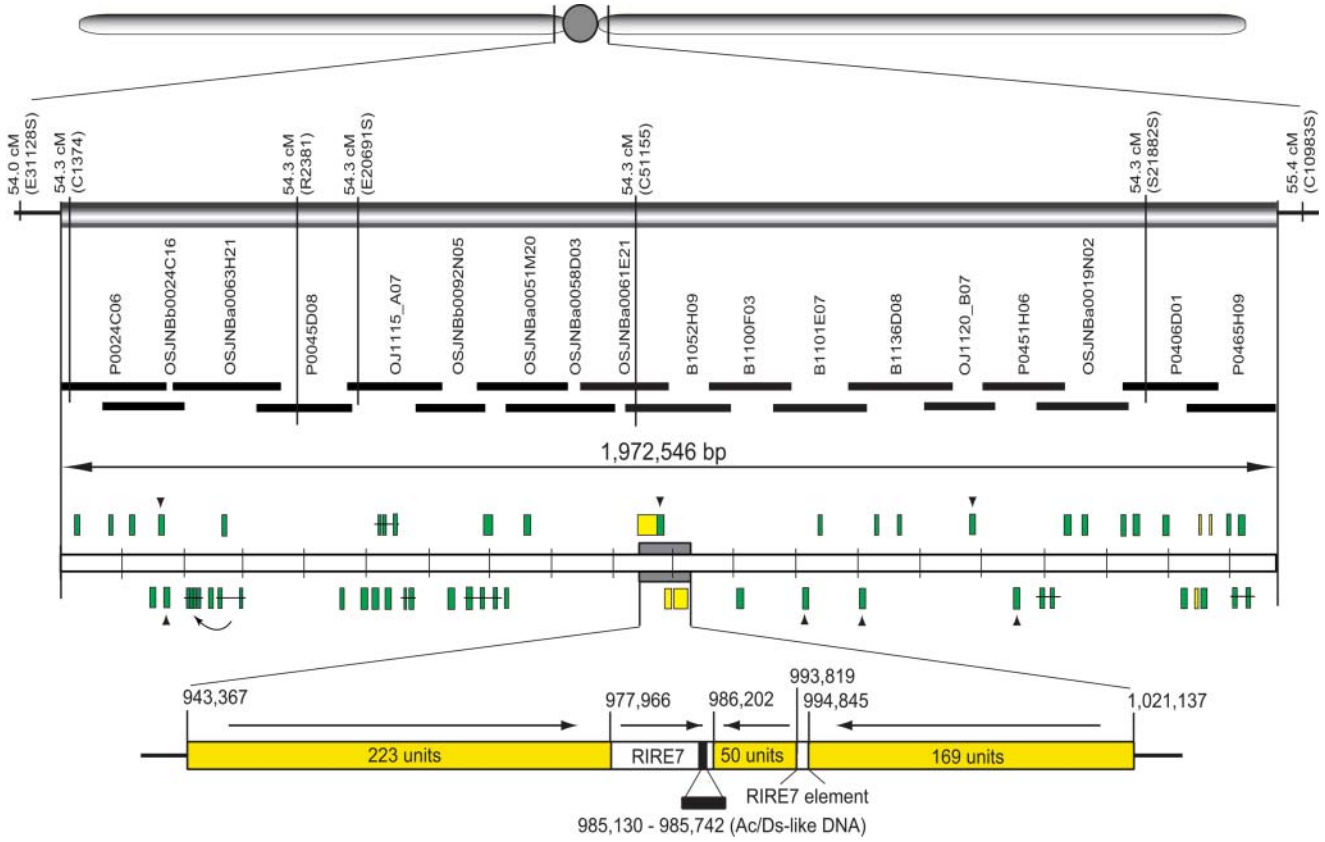

Figure 1. Genetic, Physical, and Sequence Maps of the Centromeric Region of Rice Chromosome 8.

Genetic map, PAC/BAC contigs, and sequence map are shown from the top to bottom. The green boxes above and below the sequence map represent transposable elements that showed homology to known sequences within GenBank (overall alignment coverage of >50% and sequence identity of >80%). Sequences of a transposable element–related element interrupted internally by DNA insertions are grouped with horizontal black lines and (if necessary) an additional arched arrow. The positions of the seven RIRE7-related elements that formed two groups are indicated by arrowheads. Yellow boxes represent the positions of CentO clusters on the central part of the sequenced region that are enlarged at the bottom; the number and orientation of the 155-bp satellite repeats are given. *Ac/Ds*, Activator/Dissociation; cM, centimorgan.

in length) composed of CentO satellite repeats within these two BAC clones (see Supplemental Figure 1 online), which overlapped each other. Our results coincide very well with the CentO amount and location pattern detected by fiber-FISH analysis of this chromosome (Cheng et al., 2002).

To characterize the rice repetitive sequences other than CentO satellites, we classified the transposable element (TE)-related sequences within the 1.97-Mb centromere sequence in light of our gene annotation results and findings from BLASTN searches against sequences in GenBank. We identified at least 224 fragments homologous to various TE sequences, including long terminal repeat (LTR) and polyprotein sequences (Table 1, see Supplemental Table 2 online). We also found 28 rice centromere-specific element1 (RCE1)-like elements, which are a specific component of the rice centromere (Nonomura and Kurata, 1999). Most of these elements were in an ~600-kb region within or near the CentO clusters (Figure 3). Among the identified TE sequences, 39 retrotransposons and a DNA transposon were full size or nearly so. Twenty-five of these TEs contained sequences >95% identical to the published data in GenBank, and 15 TEs had high sequence homology to the internal polyprotein region but pronounced divergence in their 5′ and/or 3′ regions, including the LTRs and 5′ untranslated region. The majority of these TE sequences within the chromosome 8 centromeric region are *gypsy*-type retrotransposons. The three retroelements we detected most frequently were those of the RIRE8, RIRE3, and RIRE7 families, a finding that supports previous results of an analysis of partial sequences from the centromeric region of rice chromosome 5 (Nonomura and Kurata, 2001; Kurata et al., 2002). Of the 40 identified TE sequences, six (three RIRE3s, one RIRE8-like, one hopi, and one HIPA) contained large inserted DNA segments at single or multiple positions (Figure 3). For instance, the hopi sequence contained four insertions, comprising full-size sequences of RIRE3 and RIRE8-like retrotransposons, which expanded the region to 66 kb (from nucleotide 205,082 to 270,964). Similar patterns of nested insertions of retroelements within repetitive sequences also have been reported to occur in maize (*Zea mays*), *Triticum monococcum* (wheat), and *Hordeum*

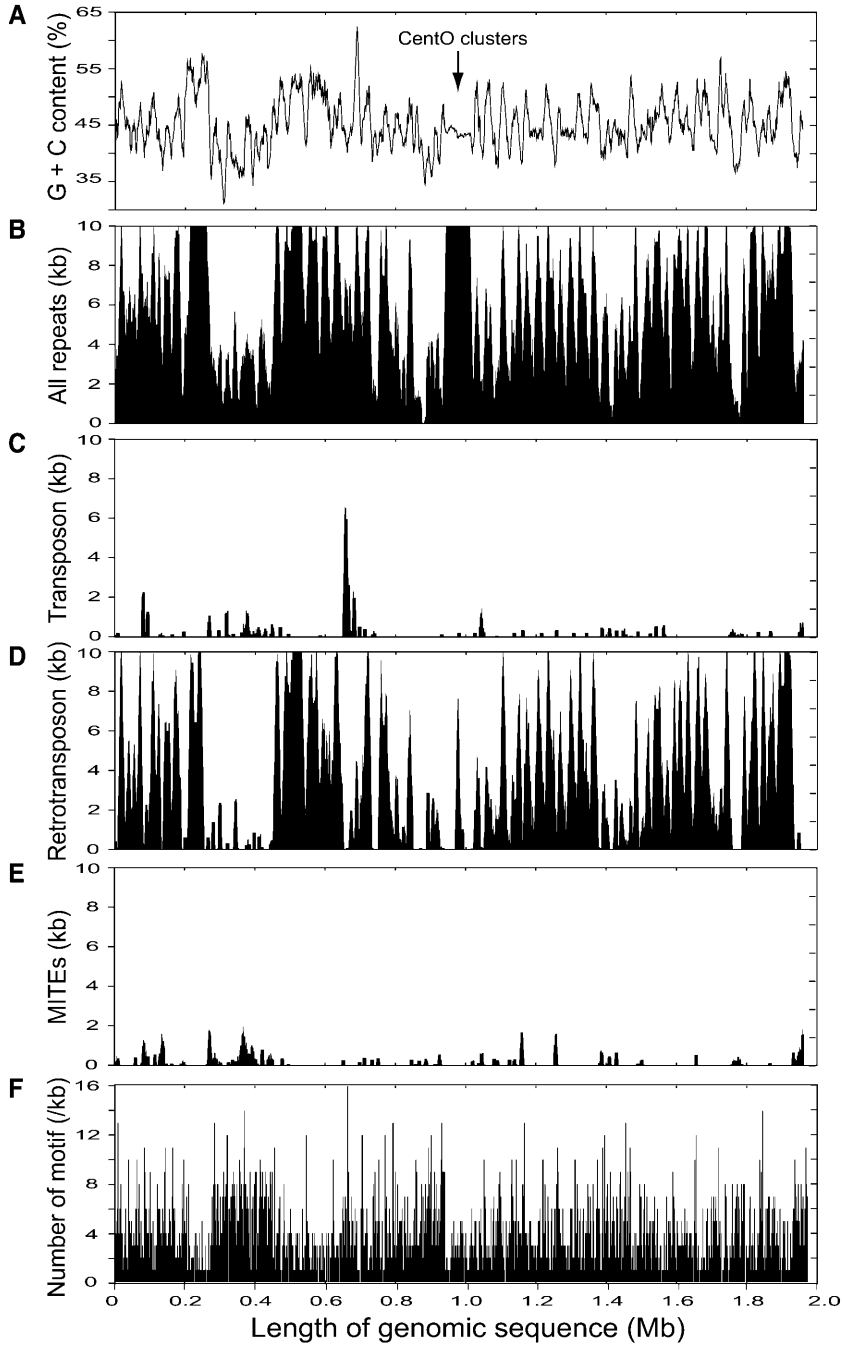

Figure 2. Distribution of G+C Fractions, Repetitive Sequences, and DNA Motifs along the Centromeric Region.

The horizontal axis represents the 1.97-Mb contiguous sequence from the short to long arm. The vertical axis shows the G+C content and frequency of various repeat sequences obtained by analyzing the sequence every 10 kb using a 1-kb sliding window and the occurrence of the DNA motif A(/T)AT(/C)ATT within every 1 kb.

(A) G+C content.

(B) All repeat sequences.

(C) DNA transposons.

(D) Retrotransposons.

(E) Miniature inverted repeat transposable elements (MITEs).

(F) A(/T)AT(/C)ATT motifs.

Figure 3. Annotation Map of the 1.97-Mb Centromeric Region of Chromosome 8.

White boxes represent the predicted genes. Green and gray boxes indicate the TEs. The names of TEs or TE-like sequences identical to those in GenBank as described in the text are shown above the corresponding sequences (also see Supplemental Tables 1 and 2 online for details). Arrows represent LTR sequences and their orientations. Arched arrows indicate that the TE sequence is interrupted by DNA insertions. Red bars indicate elements containing sequences similar to the rice centromere-specific element RCE1 (accession number AB013613). Yellow boxes represent CentO clusters; the red arrows above show the orientations of the 155-bp satellite repeats. The positions of five genetic markers without recombination are indicated by their names in orange.

vulgare (barley) (SanMiguel et al., 1996; Shirasu et al., 2000; Wicker et al., 2001).

DISCUSSION

Learning about the composition and structure of centromeres can be greatly helpful for understanding the mechanisms needed for their functional roles. Although several centromeric regions from Arabidopsis (*Arabidopsis thaliana*) have been sequenced, these sequences contain many gaps (Arabidopsis Genome Initiative, 2000; Kumekawa et al., 2000, 2001; Hosouchi et al., 2002). In this article, we report the successful sequencing and analysis of an entire and genetically defined rice centromeric region that contained ~1.97 Mb of contiguous nucleotide sequence. Despite the presence of abundant repetitive sequences, we predicted 201 opening reading frames (mostly encoding hypothetical proteins) within the region, revealing an average gene density of 1 gene per 10 kb. Although biochemical demonstrations to define which part of the sequences on the above region is the functional centromere were not done in this experiment, our upfront results will aid future research into biology as well as genomic evolution of the centromeres in plants.

Organization and Sequence Conservation of Rice CentO Satellites

Satellite repeats associated with centromeres have been reported in several plant species (Cheng et al., 2002; Jiang et al., 2003; Nagaki et al., 2003). As shown in Figure 1, three large clusters of rice CentO satellite sequence were present within the genetically defined centromeric region of chromosome 8. Extensive analysis of the CentO sequences revealed that the first cluster (nucleotide 943,367 to 977,965) comprised 223 units of

Table 1. Repetitive Sequences Detected from the Centromeric Region of Rice Chromosome 8

Repeats	Copy No.	Type	Note
CentO	458	Centromeric satellite DNA	155-bp satellites
Total TE-related elements	224	LTRs, polyprotein, CRRs	Some truncated
Individual TEs in a full or nearly full-length			
HIPA	1	*Rim2*, DNA transposon	With insertions
Hopi	3	*Gypsy*	One with insertions, one with divergence in 5′ and 3′ regions
Houba	1	*Copia*	
RIRE2	2	*Gypsy*	One with divergence in 3′ region
RIRE3	12	*Gypsy*	Three with insertions
RIRE7	7	*Gypsy*	Four with divergence in 5′ and 3′ regions
RIRE8	14	*Gypsy*	Nine with divergence in 5′ and/or 3′ regions

the rice 155-bp satellite repeat (CentO), including six copies with 12 bp of tandem duplication occurring at the same position and three copies with 55-, 56-, and 71-bp deletions at different positions (see Supplemental Figure 1 online). All of the units were uninterrupted and tandemly arrayed in a head-to-tail arrangement with the orientation of 5′ to 3′. The second cluster (from nucleotide 986,202 to 993,818) contained 50 units of the 155-bp satellite repeat, including 12 copies with deletions, nine of which were 10-bp deletions that occurred at the same position and three (7-, 14-, and 78-bp deletions) at different positions. The third cluster (nucleotide 994,845 to 1,021,137) contained 169 units of the 155-bp satellite repeat, including seven copies with tandem duplications. Of these, six were 12-bp duplications that occurred at the same position; the remaining one was 34 bp long and occurred at a different position. In addition, one copy had a 56-bp deletion. Like those for the first cluster, these satellite repeats were uninterrupted and tandemly arrayed in a head-to-tail arrangement, but they were opposite in orientation (3′ to 5′). Judging from the sequences at the breakpoint between the second and third clusters, we found that they were originally one continuous fragment and that a short DNA segment was inserted at some point. Interestingly, this breakpoint corresponds to the position where the 12-bp tandem duplication takes place frequently within the 155-bp repetitive sequence. This finding indicates the presence of sequence rearrangement even within the highly conserved CentO satellite. It also is very interesting that the total number of satellite DNAs in the second and third clusters is almost the same as that in the first cluster.

Efforts to understand the nature and specification of the centromere have demonstrated that this central element for ensuring inheritance is itself epigenetically determined (Charlesworth et al., 1994; Cleveland et al., 2003; Hall et al., 2003). In human, the α-satellite DNA in centromeres is known to interact with the CENP-A protein to form a binding site for the multiple microtubules that attach to metazoan centromeres (Vafa and Sullivan, 1997). Rice CentO satellite repeats have known to be located within the chromosome regions at which the kinetochore is formed and the spindle fibers attach (Cheng et al., 2002). We then examined the DNA sequences of the 442 units of the 155-bp satellite repeat obtained in this study. We found that, except for 49 units that contained sequences with a low quality, the 155-bp DNA sequences (accuracy >99.99%) among the 393 units were highly conserved, with an identity of 91 to 99% relative to a relevant consensus sequence (Figure 4A). Among the base changes, six were transversions (AT and CG) and two were transitions (TC) (Figure 4B). Because transitions are thought to be greatly more frequent than transversions, the preceding finding is surprising. It is noteworthy that five transversions and a transition occurred in the 5′ region of the satellite repeat that contained AT-rich motifs, similar to the conserved centromere DNA element of CDEII that most likely interacts with Cse4p and Mif2 (the homologs of CENP-A and CENP-C) in budding yeast (Cottarel et al., 1989; Meluh and Koshland, 1995; Meluh et al., 1998). Another transversion, in the 3′ region, corresponded to one of the six nucleotides of CACGTG, which has been demonstrated in the budding yeast as the palindromic core sequence of CDEI and as such is important for the in vivo protein binding of centromeres (Clarke, 1990). In addition, the 3′ region of the 1.97-Mb sequence contained two small clusters of CentO repeats, each of which had only eight copies of the 155-bp satellite (Figure 3).

Organization and Sequence Divergence of TEs

In general, the detected TE or TE-like sequences in this study were concentrated within five large clusters surrounding the CentO clusters to form a somewhat symmetrical structure for the centromere core (Figure 2D). It now becomes clear that the three large clusters of CentO satellite repeats are separated by two retrotransposon fragments derived from RIRE7. The interval between the first and second clusters was 8235 bp (nucleotide 977,966 to 986,201) and corresponds fully to the published RIRE7 sequence in GenBank except for a 612-bp insertion in the 3′ LTR; this insertion is similar to the Activator/Dissociation-like-transposon sequence (Figure 1). The interval between the second and third clusters was only 1025 bp (nucleotide 993,819 to 994,844) and showed 84% identity with the polyprotein-encoding sequence of RIRE7 in its 3′ end region. On the basis of the sequence and FISH analysis of a BAC clone derived from a centromeric region of *O. sativa* subsp *indica* cv IR-BB21,

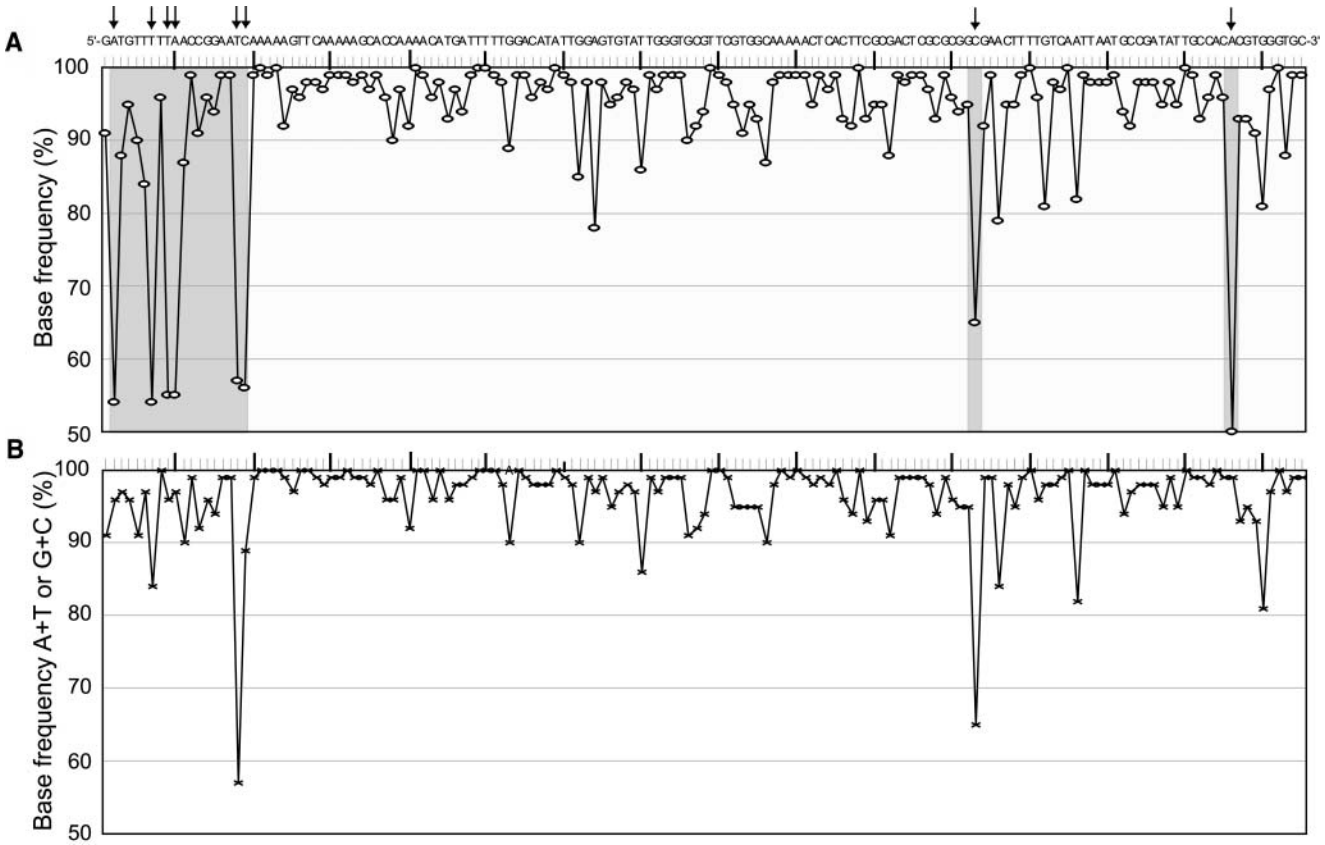

Figure 4. Sequence Variations among Each CentO Satellite Repeat.

(A) The most frequent bases (consensus shown above the graph) detected among the 393 units of 155-bp satellite repeats that showed >99.99% sequence accuracy.

(B) The base frequency including the transversion sequence. The eight alternative bases comprising six transversions (AT and GC) and two transitions (TC) are indicated by arrows and are found within the regions with gray shading in **(A)**.

rice CentO repeats colocalize with the centromeric retrotransposon of rice (CRR) sequences RCS1, RCH1, RCH2, and RCH3 (Dong et al., 1998; Cheng et al., 2002). Similar results were obtained from maize centromeres in which centromeric retrotransposon in maize (CRM) sequences (homologous to rice CRR elements) with the CentC repeat are integral to centromere function by specifically interacting with the centromeric histone H3 (Zhong et al., 2002).

Compared with those of the *indica* rice IR-BB21 and maize, the centromeric region of chromosome 8 of the *japonica* cv Nipponbare manifested some differences in regard to sequence and organization patterns. Three elements (nucleotides 977,966 to 986,201, 1,208,944 to 1,216,562, and 1,300,951 to 1,308,568) showed highly conserved sequences with the RIRE7 encoding both LTRs and the internal polyprotein (overall identity, 99%). These sequences revealed strong homology only to the RCS1 element, one of the four CRR sequences, but weak homology to the RCH3 sequence. The three described RIRE7 sequences had an additional unique feature in that they all were located in regions with a G+C content of 43 to 45% (Figure 2A). We found another group of four elements whose sequences were highly conserved among each other, with an overall identity of >95% after ignoring some short inserted sequences. These elements were located near the two distal ends (nucleotides 155,094 to 162,712, 173,626 to 181,082, 1,484,577 to 1,492,205, and 1,549,994 to 1,558,460) of the centromeric region, and their overall identity with RIRE7 was only 83 to 85%, with alignment coverage of 52 to 55%. Noteworthy differences between these two groups of elements were mostly limited to the sequences encoding the two LTRs and the 5′ untranslated region of the polyprotein gene. Results of sequence comparison based on BLASTN analysis clearly confirmed the greater correspondence of the second group to CRR (four sequences of RCS1, RCH1, RCH2, and RCH3) and CRM than to the RIRE7 sequence (Figures 5A to 5C). Sequence similarity analysis indicated that full-size CRM elements in maize centromeres also can be divided into two groups that were suggested to be transposed 1.22 million years ago (Langdon et al., 2000; Nagaki et al., 2003).

Within the 1.97-Mb region, we also found 17 RIRE8-like sequences, which formed two groups in light of sequence differences, similar to the situation for the RIRE7-related elements (Table 1, see Supplemental Table 2 online). It seems difficult

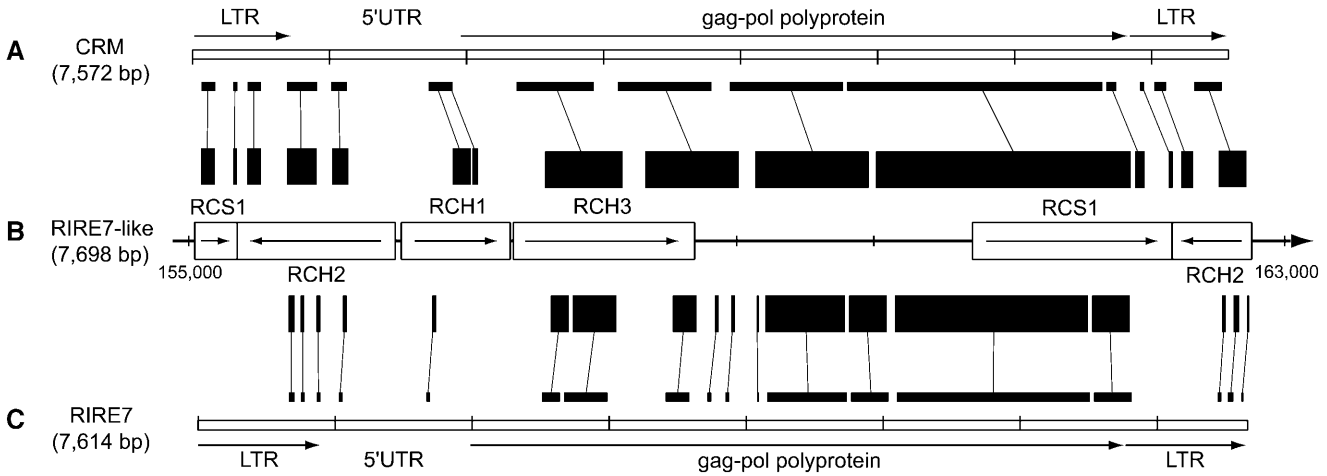

Figure 5. Sequence Comparison of a RIRE7-Related Element Found in the Centromeric Region of Chromosome 8 with GenBank Data.

The sequence alignment between each element on the basis of BLASTN analysis is shown with closed boxes. Sequence identities for the aligned segments ranged from 80 to 100%. Sequences identical to CRR sequences (accession numbers AF078903 for RCS1, AF058903 for RCH1, AF058904 for RCH2, and AF058905 for RCH3) within the RIRE7-related elements are represented with open boxes.

(A) The maize retrotransposon CRM (accession number AY129008).

(B) The RIRE7-related element from nucleotide 155,040 to 162,737 of this study.

(C) The rice retrotransposon RIRE7 (accession numbers AB033234 for LTR regions and AB033235 for the internal gag-pol polyprotein region).

at the moment to make a conclusion on the evolutionary relationships between the two groups of sequences in both the RIRE7 and RIRE8 families. Because both groups of RIRE7 elements are highly conserved not only within the rice centromere but also among the rice and other cereals, such as maize and *Sorghum bicolor* (sorghum), they may have derived from a common ancestor sequence and then been subjected to different evolutionary rates. We suppose the CRR sequence is one member of the RIRE7 family that originally arose before the divergence of Poaceae species ~55 million years ago (Kellogg, 2001).

For the remaining TE-related sequences identified in the centromeric region, pronounced rearrangements (insertions and deletions) were present, suggesting that these elements might have transposed much earlier and may have been largely involved in the Poaceae genome evolution (Bennetzen, 2000; Heslop-Harrison, 2002). In addition, by searching for DNA motifs within the 1.97-Mb region, we found 5988 sites for the sequence A(/T)AT(/C)ATT, 119 sites for AATAAAT(/C)AAA, and 712 sites for TTA(/T)TT(A)TTT(A)TT. These sites were distributed almost evenly throughout the region, in which the A(/T)AT(/C)ATT motif, the core sequence of the Topo-II box, which is an essential component of the nuclear matrix attachment regions in *D. melanogaster* and other organisms (for a review, see Gasser and Laemmli, 1987), showed an average of 3.0 sites per kb with a maximum of 17 sites per kb (Figure 2F). Whether these sequences play roles in the function of rice centromeres remains undetermined. Because accurate chromosome segregation requires that sister chromatids remain associated with the spindle, it will be important to know whether and how these sequences, together with the CentO satellites, are essential for centromere identity, stability, and complete function in rice.

Distribution of CentO and RIRE7 Sequences on Other Rice Chromosomes

Neither CentO satellite repeats nor RIRE7 elements described above could be found on the other genomic regions of the chromosome 8 in which 95% of its sequence has been completed (T. Sasaki, unpublished data). To date, the International Rice Genome Sequencing Project (IRGSP) has already published 358 Mb of nonoverlapping nucleotide sequence of the rice genome, including that of chromosome 8, in public databases (http://rgp.dna.affrc.go.jp/IRGSP/). We then investigated the sequence and organization of CentO repeats and the two groups of RIRE7 elements in the published rice sequence. Using the 155-bp consensus sequence from the CentO repeats as the query for BLASTN analysis revealed the presence of ~440 kb of these satellite sequences in the other chromosomes (corresponding to >2800 copies of the satellite repeat). Except for only five copies that were located in telomeric regions, all of these CentO copies occurred in centromeric regions and showed distribution patterns and orientations similar to those in chromosome 8 (Wu et al., 2003). Using the full-size RIRE7 (7614 bp, including both LTRs and the internal polyprotein sequences) in GenBank and the RIRE7-related sequences in the centromere of chromosome 8 (7698 bp from nucleotide 155,040 to 162,737, which contain the four CRR elements) as queries for BLASTN analysis against the IRGSP rice sequences, we found 37 copies of the two described elements, 13 corresponding entirely to RIRE7 and 24 to the RIRE7-related sequences (alignment coverage >55%, overall identity >93%). All of these sequences were located within the recombination-suppressed regions of each centromere, except for one copy that had highly conserved sequences with the second group of RIRE7 elements

but was obviously located within the long-arm region of chromosome 4. From the sequences available for the centromeric region of chromosome 1, we found that four (one RIRE7 and three RIRE7-related) of the six copies colocalized with the CentO satellite repeats (Wu et al., 2003). Therefore, we suggest that not only sequences of the RIRE7-related (CRR) elements studied thus far but also the retrotransposon RIRE7 itself could be components essential to the functions of rice centromeres.

We have deduced the compositional and structural features of the centromere of rice chromosome 8 of the *japonica* rice Nipponbare in light of its complete genomic sequence. The next challenge is to extensively investigate in vivo the roles of the centromere-specific sequences, especially the CentO satellite repeats and retroelements of the RIRE7 and RIRE8 families, in both karyotypic stability and chromosome segregation. We believe that our findings will be very useful in studying not only the complicated mechanisms of centromere functions but also the evolutionary history of the centromere among plant genomes.

METHODS

Mapping and Sequencing

PAC/BAC clones were mapped to the centromeric region of rice chromosome 8 using a method similar to that already described (Wu et al., 2003). We used sequences of rice DNA markers to conduct the in silico mapping of Monsanto (St. Louis, MO) BAC clones and PCR screening of RGP PAC/BAC libraries. Gaps on the physical map were closed by screening and choosing the bridge clones from either CUGI-assembled BAC contigs or RGP PAC/BAC libraries based on the sequence-tagged connector or chromosomal walking methods.

We sequenced the PAC and BAC clones using a shotgun approach. Two shotgun libraries with different insert sizes (average 2 kb and 5 kb) for each clone were constructed by cloning the genomic fragments into the *Sma*I site of pUC18. Both ends of 960 2-kb subclones and 960 5-kb subclones were sequenced to achieve 10-fold coverage (Sasaki et al., 2002). Sequence reactions were performed using dye terminator chemistry (BigDye Terminator version 3.1; Applied Biosystems, Foster City, CA), and the sequences were analyzed on capillary sequencers (ABI 3700; Applied Biosystems). Each PAC or BAC clone was assembled from the shotgun sequences by PHRED/PHRAP software (Ewing and Green, 1998). Assembled sequence contigs were viewed and edited by CONSED (Gordon et al., 1998). Sequence gaps between the assembled contigs were generally filled by sequencing the bridge clones with custom primers. Regions with low quality scores were improved by resequencing with custom primers or by alternative chemistries. A transposon insertion/sequencing system (Genome Priming System GPS-1; New England Biolabs, Beverly, MA) was used for complete sequencing of the subclones in the problematic regions that contained repeats, such as the RCS2 satellite DNA.

Gene Prediction and Classification of Repetitive Sequences

Gene prediction for the 1.97-Mb sequence was conducted using our annotation system (http://rgp.dna.affrc.go.jp/genomicdata/AnnSystem.html) in a similar way as for the analysis of the chromosome 1 sequence (Sasaki et al., 2002). Homology searches of the rice full-length cDNA sequences against the predicted genes were performed using BLASTN (release 2.2.5) (Altschul et al., 1997; Rice Full-Length cDNA Consortium, 2003). The total amount of repetitive sequences was measured using the Institute for Genomic Research Rice Repeat Database for BLAST searches (http://www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml) and RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). Furthermore, locations of the centromere-specific sequence RCE1 were investigated (Nonomura and Kurata, 1999). Classification of rice TEs within the centromere region was conducted in light of the results of BLAST analysis of sequences against the data in GenBank, which include >40 TE elements identified from rice and other cereal crops (last update, May 2003).

Accession Numbers

The accession numbers of the 18 PAC or BAC clones we sequenced are AP005305, AP005843, AP005819, AP004458, AP004041, AP005540, AP005162, AP005166, AP005693, AP006480, AP006481, AP006482, AP005832, AP004228, AP005500, AP005407, AP005498, and AP004561. A contiguous sequence of the 1.97-Mb centromere region also can be downloaded at http://rgp.dna.affrc.go.jp/Publicdata.html, and its annotation result can be viewed at http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/status.pl.

Sequence data from this article have been deposited with the EMBL/GenBank data libraries under accession numbers AB013613, AF078903, AF058903, AF058904, AF058905, AY129008, AB033234, AB033235, AP005305, AP005843, AP005819, AP004458, AP004041, AP005540, AP005162, AP005166, AP005693, AP006480, AP006481, AP006482, AP005832, AP004228, AP005500, AP005407, AP005498, and AP004561.

ACKNOWLEDGMENTS

We thank the Monsanto Company for providing rice BAC clones and draft sequences and Rod Wing (Arizona Genomic Institute and Computational Laboratory, Tucson, AZ) for providing their rice Nipponbare BAC libraries and the fingerprint data. We thank Frances A. Burr (Brookhaven National Laboratory, Upton, NY) for critical reading and editing of the manuscript. We also thank Masahiro Nakagahra and Kyozo Eguchi for advice and encouragement and all RGP members for their support during this work. This study was supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Rice Genome Project, GS-1101, 1103, and GS-1201).

Received November 25, 2003; accepted February 1, 2004.

REFERENCES

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

Ananiev, E.V., Phillips, R.L., and Rines, H.W. (1998). Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. Proc. Natl. Acad. Sci. USA **95,** 13073–13078.

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

Bennetzen, J.L. (2000). Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. **42,** 251–269.

Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371,** 215–220.

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R., and Jiang, J. (2002). Functional rice centromeres are

marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell **14,** 1691–1704.

Choo, K.H., Vissel, B., Nagy, A., Earle, E., and Kalitsis, P. (1991). A survey of the genomic distribution of alpha-satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. Nucleic Acids Res. **19,** 1179–1182.

Clarke, L. (1990). Centromeres of budding and fission yeasts. Trends Genet. **6,** 150–154.

Cleveland, D.W., Mao, Y., and Sullivan, K.F. (2003). Centromeres and kinetochores: From epigenetics to mitotic checkpoint signaling. Cell **112,** 407–421.

Copenhaver, G.P., et al. (1999). Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science **286,** 2468–2474.

Cottarel, G., Shero, J.H., Hieter, P., and Hegemann, J.H. (1989). A 125-base-pair CEN6 DNA fragment is sufficient for complete meiotic and mitotic centromere functions in *Saccharomyces cerevisiae*. Mol. Cell. Biol. **9,** 3342–3349.

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.

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

Feng, Q., et al. (2002). Sequence and analysis of rice chromosome 4. Nature **420,** 316–320.

Gasser, S.M., and Laemmli, U.K. (1987). A glimpse at chromosomal order. Trends Genet. **6,** 16–22.

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

Hall, S.E., Kettler, G., and Preuss, D. (2003). Centromere satellites from *Arabidopsis* populations: Maintenance of conserved and variable domains. Genome Res. **13,** 195–205.

Harushima, Y., et al. (1998). A high-density rice genetic linkage map with 2,275 markers using a single F_2 population. Genetics **148,** 479–494.

Heslop-Harrison, J.S. (2002). Comparative genome organization in plants: From sequence and markers to chromatin and chromosomes. Plant Cell **12,** 617–635.

Hosouchi, T., Kumekawa, N., Tsuruoka, H., and Kotani, H. (2002). Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. DNA Res. **9,** 117–121.

Jiang, J., Birchler, J.A., Parrott, W.A., and Dawe, R.K. (2003). A molecular view of plant centromeres. Trends Plant Sci. **8,** 570–575.

Kaszas, E., and Birchler, J.A. (1996). Misdivision analysis of centromere structure in maize. EMBO J. **15,** 5246–5255.

Kellogg, E.A. (2001). Evolutionary history of the grasses. Plant Physiol. **125,** 1198–1205.

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

Kumekawa, N., Hosouchi, T., Tsuruoka, H., and Kotani, H. (2001). The size and sequence organization of the centromeric region of *Arabidopisis thaliana* chromosome 4. DNA Res. **8,** 285–290.

Kurata, N., Nonomura, K.-I., and Harushima, Y. (2002). Rice genome organization: The centromere and genome interactions. Ann. Bot. **90,** 427–435.

Langdon, T., Seago, C., Mende, M., Leggett, M., Thomas, H., Forster, J.W., Jones, R.N., and Jenkins, G. (2000). Retrotransposon evolution in diverse plant genomes. Genetics **156,** 313–325.

Meluh, P.B., and Koshland, D. (1995). Evidence that the MIF2 gene of *Saccharomyces cerevisiae* encodes a centromere protein with homology to the mammalian centromere protein CENP-C. Mol. Biol. Cell **6,** 793–807.

Meluh, P.B., Yang, P., Glowczewski, L., Koshland, D., and Smith, M.M. (1998). Cse4pis a component of the core centromere of *Saccharomyces cerevisiae*. Cell **94,** 607–613.

Nagaki, K., Song, J., Stupar, R.M., Parokonny, A.S., Yuan, Q., Ouyang, S., Liu, J., Hsiao, J., Jones, K.M., Dawe, R.K., Buell, C.R., and Jiang, J. (2003). Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics **163,** 759–770.

Nonomura, K.-I., and Kurata, N. (1999). Organization of the 1.9-kb repeat unit RCE1 in the centromeric region of rice chromosomes. Mol. Gen. Genet. **261,** 1–10.

Nonomura, K.-I., and Kurata, N. (2001). The centromere composition of multiple repetitive sequences on rice chromosome 5. Chromosoma **110,** 284–291.

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

Rice Full-Length cDNA Consortium (2003). Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. Science **301,** 376–379.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science **274,** 765–768.

Sasaki, T., and 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., et al. (2002). The genome sequence and structure of rice chromosome 1. Nature **420,** 312–316.

Schueler, M.G., Higgins, A.W., Rudd, N.K., Gustashaw, K., and Willard, H.F. (2001). Genomic and genetic definition of a functional human centromere. Science **294,** 109–115.

Shirasu, K., Schulman, A.H., Lahaye, T., and Schulze-Lefert, P. (2000). A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. **10,** 908–915.

Singh, K., Ishii, T., Parco, A., Huang, N., Brar, D.S., and Khush, G.S. (1996). Centromere mapping and orientation of the molecular linkage map of rice. Proc. Natl. Acad. Sci. USA **93,** 6163–6168.

Sun, X., Le, H.D., Wahlstrom, J.M., and Karpen, G.H. (2003). Sequence analysis of a functional *Drosophila* centromere. Genome Res. **13,** 182–194.

Vafa, O., and Sullivan, K.F. (1997). Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. Curr. Biol. **7,** 897–900.

Waye, J.S., and Willard, H.F. (1987). Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: A survey of alphoid sequences from different human chromosomes. Nucleic Acids Res. **15,** 7549–7569.

Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., and Keller, B. (2001). Analysis of a contiguous 211 kb sequence in diploid wheat (*T. monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J. **26,** 307–316.

Wu, J., et al. (2002). A comprehensive rice transcript map containing 6591 expressed sequence tag sites. Plant Cell **14,** 525–535.

Wu, J., et al. (2003). Physical maps and recombination frequency of 6 rice chromosomes. Plant J. **36,** 720–730.

Zhong, C.X., Marshall, J.B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J.A., Jiang, J., and Dawe, R.K. (2002). Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell **14,** 2825–2836.