Published online April 2, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 6 ***2023–2030***
DOI: 10.1093/nar/gkh521

Structural features of the rice chromosome 4 centromere

Yu Zhang, Yuchen Huang, Lei Zhang, Ying Li, Tingting Lu, Yiqi Lu, Qi Feng, Qiang Zhao, Zhukuan Cheng[1], Yongbiao Xue[1], Rod A. Wing[2] and Bin Han*

National Center for Gene Research, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China, [1]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Road, Andingmenwai, Beijing 100101, China and [2]Department of Plant Sciences, Arizona Genomics Institute, The University of Arizona, Tucson, AZ 85721, USA

Received January 4, 2004; Revised February 12, 2004; Accepted March 10, 2004

ABSTRACT

A complete sequence of a chromosome centromere is necessary for fully understanding centromere function. We reported the sequence structures of the first complete rice chromosome centromere through sequencing a large insert bacterial artificial chromosome clone-based contig, which covered the rice chromosome 4 centromere. Complete sequencing of the 124-kb rice chromosome 4 centromere revealed that it consisted of 18 tracts of 379 tandemly arrayed repeats known as CentO and a total of 19 centromeric retroelements (CRs) but no unique sequences were detected. Four tracts, composed of 65 CentO repeats, were located in the opposite orientation, and 18 CentO tracts were flanked by 19 retroelements. The CRs were classified into four types, and the type I retroelements appeared to be more specific to rice centromeres. The preferential insert of the CRs among CentO repeats indicated that the centromere-specific retroelements may contribute to centromere expansion during evolution. The presence of three intact retrotransposons in the centromere suggests that they may be responsible for functional centromere initiation through a transcription-mediated mechanism.

INTRODUCTION

The centromere is essential for correct segregation of chromosomes in both mitotic and meiotic cells. Although centromere function is conserved in eukaryotes, centromeric sequences appear to be variable (1,2). It is believed that a complete sequence of a chromosome centromere is necessary for fully understanding centromere function. Centromere functions can be recapitulated by artificial chromosome constructs (3,4) or chromosome fragments (5,6), revealing the important role of specific centromere sequences.

Except for the centromere of the budding yeast *Saccharomyces cerevisiae*, which consists of only ~125 bp unique sequence (7,8), most eukaryotic species are composed of long highly repetitive DNA sequences. The currently available eukaryotic genome sequencing projects have provided the virtually complete physical maps and sequences of many species, including *Caenorhabditis elegans* (9), *Arabidopsis thaliana* (10), *Homo sapiens* (11,12) and *Oryza sativa* (13,14) in the last few years. However, as a highly heterochromatic part of the chromosome, the centromere is still left to be a big 'gap' to be sequenced. The highly repetitive DNA is more difficult to map, clone, sequence and assemble than the low copy number DNA. Though some detailed sequences about centromere regions have been analyzed, for example the satellite arrays of human centromeres (15) and the pericentromeric regions on *Arabidopsis* chromosomes (16), the sequences are not completely available yet and the genome sequencing project has left the big challenge of determining the primary sequence of a functional higher eukaryotic centromere. The full length of natural centromeres in most eukaryotes is at the megabase level generally. Such a long repeat DNA region is the most difficult barrier to cloning and sequencing it. The ordinary approach is to use different methods that isolate specific centromere regions from the rest of the genome. A successful example is the use of pulse field gel electrophoresis (PFGE) for isolation of the γ*1230* minichromosome derivative as a template for cloning and sequencing a significant proportion of the functional *Drosophila* centromere (17). Bacterial artificial chromosome (BAC) clones and similar clones with large insert size are also a good choice for isolation of centromere sequences from other genomic regions. However, this approach still has limitations for manipulation of a full centromere region such as those in human and *Arabidopsis*. Rice is an exception in the important model species because of its different centromere size, and the size of the satellite repeat is quantitatively variable among the 12 rice centromeres detected by the fluorescence intensities of the fluorescence *in situ* hybridization signals (18). Though some chromosomes have a similar centromere size to those in other species (>1 Mb), the centromeres of several chromosomes are surprisingly small and can be fully covered by BAC contigs constructed using the normal available technical approach. Thus, rice provides an opportunity to obtain the full

*To whom correspondence should be addressed. Tel: +86 21 64845260; Fax: +86 21 64825775; Email: bhan@ncgr.ac.cn

centromere sequence and the truly complete understanding of centromere sequence composition and organization.

Some DNA components of rice centromeres have been reported. The centers of rice centromeres are occupied by two kinds of repetitive elements: the 155-bp satellite repeat CentO and the centromere-specific retrotransposons (18). CentO satellites were found to be located exclusively in rice centromeres and were regarded as a key component of functional rice centromeres. The retroelements found in the rice centromere, such as RCS1 (19), RCB11 (20,21), RIRE7 (22,23), are mostly derived from *gypsy*-like retrotransposon family. With the rapid progress of the rice genome sequencing, complete sequence composition and structure of rice centromere become available now.

Here, the physical structure of the rice chromosome 4 centromere was determined directly by sequencing a large insert clone-based contig, which covered the centromere. The unique structures of the rice chromosome 4 centromere were identified. Some of the structures appear to be specific to the rice centromeres.

MATERIALS AND METHODS

Materials and physical map construction

Construction of a clone-based physical map of the rice chromosome 4 was described previously (24). The rice *Oryza sativa* ssp. *japonica* cv. Nipponbare clones used for constructing the physical map of the centromere contig were from four genomic libraries: two BAC libraries of OSJNBa and OSJNBb provided by Clemson University Genomics Institute (CUGI); one PAC library provided by Rice Genome Research Program (RGP); and some of BACs were provided by the Monsanto.

DNA sequencing and assembly

OSJNBb0062N22 and other clones were purified by caesium chloride-gradient. For a shotgun approach, sheared BAC DNA (2–3 kb) was ligated into a pBluescript vector and transformed into *Escherichia coli* DH5α. The shotgun subclones were sequenced from both ends by the dideoxy chain termination method using either BigDye Terminator Cycle sequencing V2.0 Ready Reaction (Applied Biosystems) or DYEnamic ET Dye Terminator Kit (MegaBACE; Amersham Pharmacia Biotech, Inc.). Most of the reactions were analyzed on ABI3730 sequencers and Megabace 1000 capillary sequencing machines. The shotgun sequences were assembled using the PHRED and PHRAP programs first (25) and primary assembly results were refined by careful manual checking to overcome the misalignments caused by repeats. Manual editing corrected many mis-assemblies caused by automatic assembly software such as excessive coverage in some regions and the separation of two ends sequence pairs from one subclone. The empirical result of PFGE and several restriction enzymes profiling were used to validate the length and the accuracy of the assembly. Sequence gaps were closed by using various dye-labeled terminator chemistries or by a combination approach of primer walking and PCR with oligonucleotides. Sequence regions of poor quality were re-sequenced from cloned plasmids. The nucleotide sequence of BAC OSJNBb0062N22 has been deposited in GenBank under the accession number BX890594.

Sequence analysis

DNA sequences similar to the BAC sequences were searched in the GenBank database and the TIGR Repeat Database (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/) using the BLASTN homology search software. Sequence alignments between different CentO monomers were performed and refined manually using GeneDoc (http://www.psc.edu/biomed/genedoc). The ages of full-length retroelements were measured by comparing their 5′ and 3′ long terminal repeat (LTR) sequences (26). Kimura-2 parameter distances between the two LTRs of individual elements were calculated using MEGA program (http://www.megasoftware.net/). The reported substitution rate of 6.5×10^{-9} per synonymous site per year for grasses (27) was used to estimate the ages of the elements.

RESULTS

Physical mapping of the chromosome 4 centromeric region

As part of an international effort to completely sequence the rice genome, we constructed a comprehensive clone-based physical map of chromosome 4 of *O.sativa* ssp. *japonica* Nipponbare, consisting of four contiguous BAC clones (contigs), through an integrated approach. Two large insert BAC libraries (OSJNBa and OSJNBb) and high-quality of DNA fingerprinting data allowed us to construct big contigs covering nearly all regions of the rice *O.sativa* ssp *japonica* Nipponbare genome, including regions of high repetitive DNA sequences (28). Contig3 (Fig. 1), the second largest contig of 12.9 Mb, fully covered the genetic region from 19.6 to 19.9 cM, where the centromere had been located by the genetic mapping (29,30). Using the key centromere component CentO satellite as a probe, we screened the BACs in Contig3 tiling path for sequencing, and found that only OSJNBb0062N22 and the overlap region between OSJNBb0062N22 and OSJNBa0032B23 contained the CentO satellites. The BAC clone OSJNBb0062N22 located at the position ~0.4 Mb in Contig3 and the gap between Contig2 and Contig3 was identified as a chloroplast genome insertion (our unpublished data). Thus, the chromosome 4 centromere core region was fully covered by the clone OSJNBb0062N22 in Contig3 (Fig. 1).

Sequencing and assembly of the centromere BAC clone

The centromeric BAC clone, OSJNBb0062N22, was sequenced by a random shotgun approach on both strands of subclones of ~2–3 kb and achieved a 10-fold coverage. The sequences were assembled by PHRED and PHARP first, and primary assembly result was refined with careful manual checking to overcome the misalignments caused by repeats. The assembled size of the BAC clone agreed with the size determined by PFGE. The validity of sequence assembly was further verified by *in silico* and empirical profiling with several restriction enzymes. The adjacent clones OSJNBb0026I12 and OSJNBa0032B23 were assembled separately and the large overlap regions showed identical sequence composition and organization. All of these evidences suggested that the quality and assembly of the sequence were of high accuracy and were reliable. The total

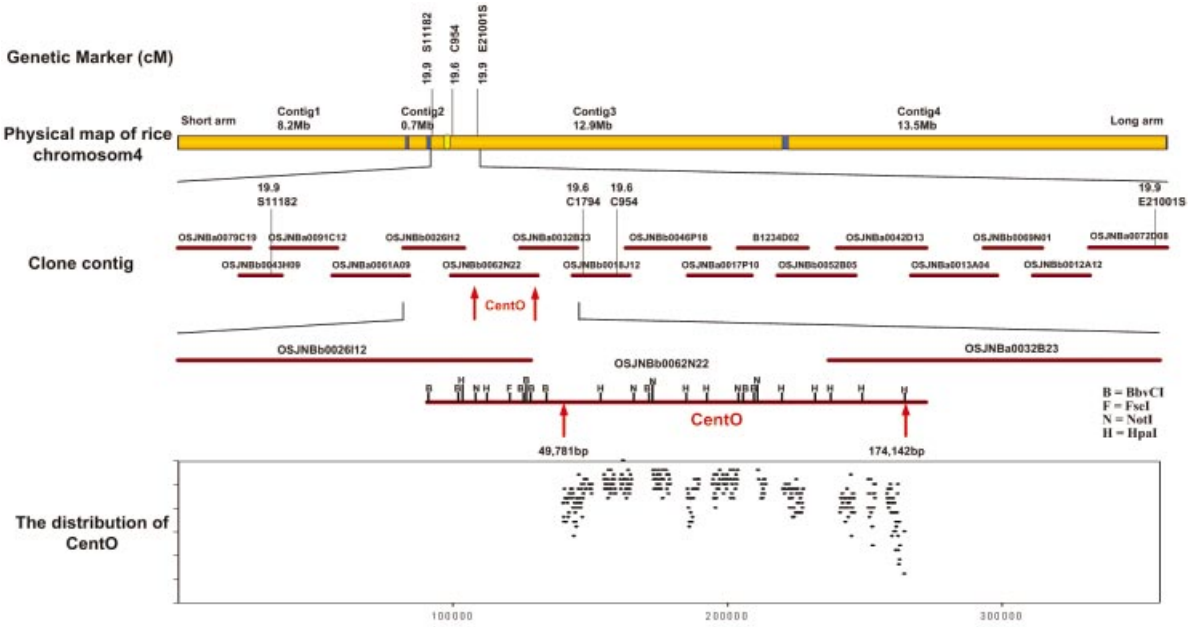

Figure 1. Map of the centromere region of rice chromosome 4. Four contigs, which covered the entire chromosome 4, are indicated in orange and as described. A part of tiling path of BAC clones of Contig3, which covered the whole centromeric region (yellow) was shown. It was also indicated that the genetic distance between markers S11182 (19.9 cM) and E21001S (19.9 cM) corresponds to the physical distance of 1200–1300 kb (24,29). The restriction enzyme sites of the BAC OSJNBb0062N22 of 181 kb, which contain all CentO repeats were indicated by the black vertical short-lines and described as following: B = BbvCI, F = FseI, N = NotI, H = HpaI. The CentO region was centralized on the BAC clone OSJNBb0062N22 from 19 781 to 174 142 bp, indicated by two red vertical arrows. The detailed distribution of CentO satellite repeats was shown in a square box. Each dot represented one satellite monomer.

length of BAC OSJNBb0062N22 was 181 586 bp and fully covered the CentO repeats ranging from 49 781 to 174 142 bp, thus defining the core region of the centromere.

Sequence contents of the centromere core region and CentO satellite repeats

Overall, the 124 kb centromere core region consisted of two kinds of repetitive elements: the 155/165 bp CentO satellite repeats and retroelements (Fig. 2). The centromeric retroelements (CRs), positioned among the tandem CentO monomer arrays, divided the CentO arrays into 18 tracts. CentO tracts were separated by the retroelements on both sides.

Eighteen CentO tracts (designated CentO-4A to CentO-4R) were found to be dispersed in the centromere core region (Fig. 2). The length of CentO tracts ranged widely from 477 to 8571 bp (Table 1). The total length of all CentO repeats was 58 865 bp, representing 47% of the centromere core region. Unexpectedly, the directions of the 18 tracts were not the same and four tracts, which located in the internal part of the core region (CentO-4H, I, J and N) were found in the opposite orientation, and tracts CentO-4H, I and J were highly homologous to CentO-4E and F but in the opposite orientation. The orientations of CentO monomers in each tract were the same.

The chromosome 4 centromere had 379 copies of CentO satellite repeat in the 18 tracts. According to their lengths and structures, these CentO monomers could be classified into three subgroups: 155 bp CentO (154 copies), 165 bp CentO (161 copies) and incomplete CentO (64 copies). The 165-bp CentO had a 10-nt duplication (TATTGGCATA, see Fig. 3), compared with the 155-bp monomer. These two types of CentO repeats were found in all 18 tracts and it seemed that 165 bp CentO monomers were inclined to appear towards the short arm telomere and 155 bp ones were biased towards the other end. A lot of incomplete CentO monomers were also found in the array. They existed as internal deletions, fragments lacking the 5′ or 3′ ends and short internal fragments. The deletion usually happened in a different position in different monomers but some incomplete monomers shared the same deletion patterns. Figure 3 showed a monomer named as CentO-4A23, with the internal 47 bp deletion repeated nine times. About half of incomplete monomers existed at the edge of the tracts. Some of these truncated monomers at ends of adjacent tracts could be merged into one complete monomer by removing the inserted retroelement, indicating that the CentO arrays were usually disrupted by CRs. Interestingly, the retroelements seemed preferentially inserted into two target sites of 85 bp and 125–128 bp of the 155 bp CentO repeat through monitoring nine insertion events (Fig. 3).

Although all CentO monomers were well conserved in length, identical repeats among the CentO arrays appeared to be rare. Except some conserved nucleotide, the polymorphisms were dispersed along the whole 155 bp consensus satellite sequence. Most polymorphism sites were only variations in one monomer while some have variations in many monomers. The identities between different monomers were mostly from 90 to 98%, but the divergences of the monomers in the CentO tracts near the edge of the centromere core region were more apparent than others. Some of them were only <85% identical to the consensus sequence. In maize and *Arabidopsis*, chromatin immunoprecipitation studies have

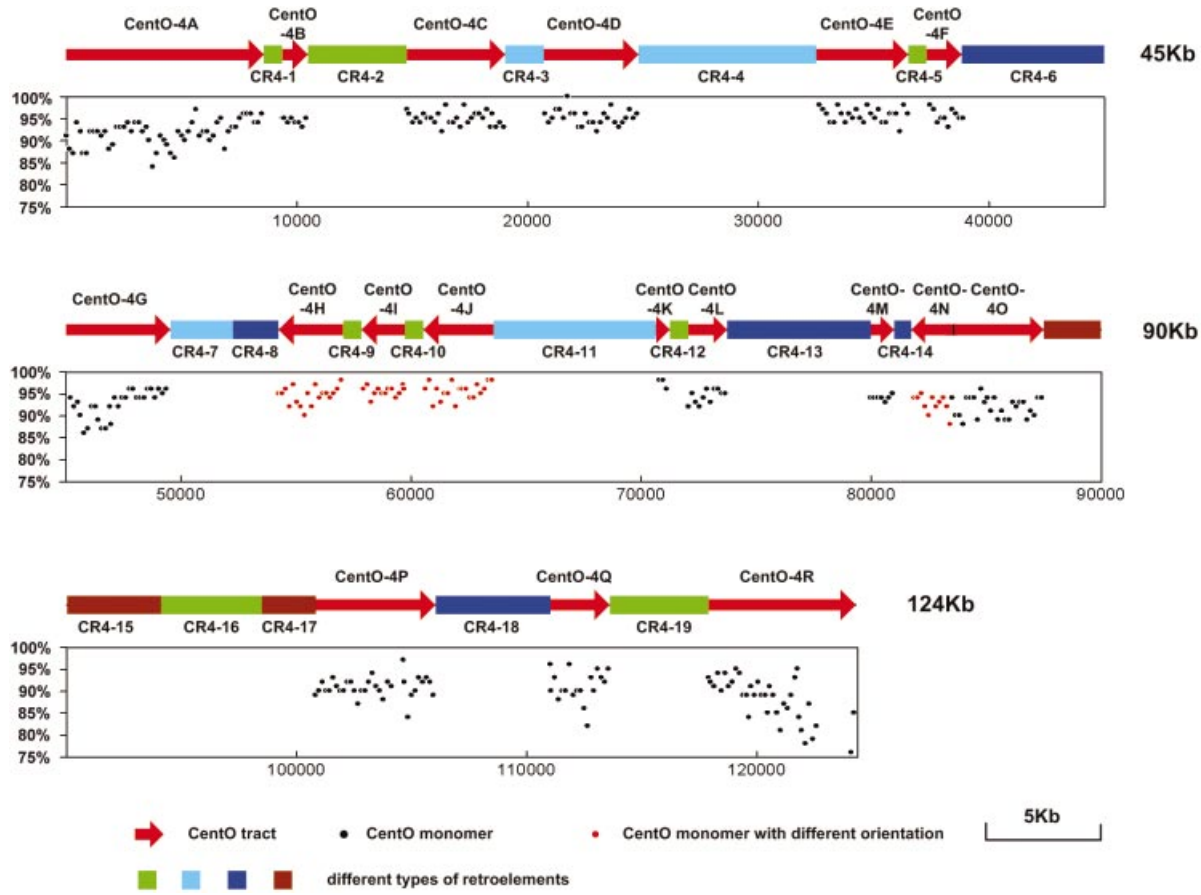

Figure 2. The complete organization of the tandemly repeated arrays of CentO sequences and CRs in the core region of rice chromosome 4 centromere. The horizontal red arrows represented the length and orientation of the 18 CentO tracts. The detail monomer's identity and orientation were shown by the black and red dots. The *y* axis showed different identity between these monomers. The CentO tracts were separated by 19 retroelements of four types. Four type retroelements were marked by four different colors, individually.

Table 1. Detailed information of the 18 CentO tracts of rice chromosome 4 centromere

Tract	Position	Length	Orientation	Copy number	155 bp CentO	165 bp CentO	Incomplete CentO	Identity
CentO-4A	1–8571	8571	+	55	14	35	6	84–97%
CentO-4B	9366–10 486	1121	+	8	3	3	2	93–95%
CentO-4C	14 778–19 078	4301	+	28	9	16	3	92–98%
CentO-4D	20 753–24 877	4125	+	26	11	13	2	92–100%
CentO-4E	32 620–36 620	4001	+	26	8	15	3	92–98%
CentO-4F	37 411–38 948	1538	+	11	4	5	2	93–98%
CentO-4G	45 200–49 506	4307	+	28	7	17	4	86–96%
CentO-4H	54 200–57 022	2823	–	19	7	9	3	90–98%
CentO-4I	57 815–59 712	1898	–	13	4	6	3	93–97%
CentO-4J	60 507–63 570	3064	–	20	4	12	4	92–98%
CentO-4K	70 771–71 247	477	+	4	1	1	2	98%
CentO-4L	72 077–73 739	1663	+	11	3	6	2	92–96%
CentO-4M	79 992–81 005	1014	+	7	3	3	1	93–94%
CentO-4N	81 749–83 514	1766	–	13	6	2	5	90–95%
CentO-4O	835 34–87 515	3982	+	26	21	3	2	88–96%
CentO-4P	100 834–106 044	5211	+	31	22	4	5	84–94%
CentO-4Q	111 043–113 628	2586	+	18	9	5	4	82–96%
CentO-4R	117 925–124 341	6417	+	35	18	6	11	78–95%
Total	1–124 341	58 865		379	154	161	64	

shown that only a portion of the centromeric satellites are involved in the centromeric function (31,32). It is still unclear whether this is related to the variable divergence found between different satellites.

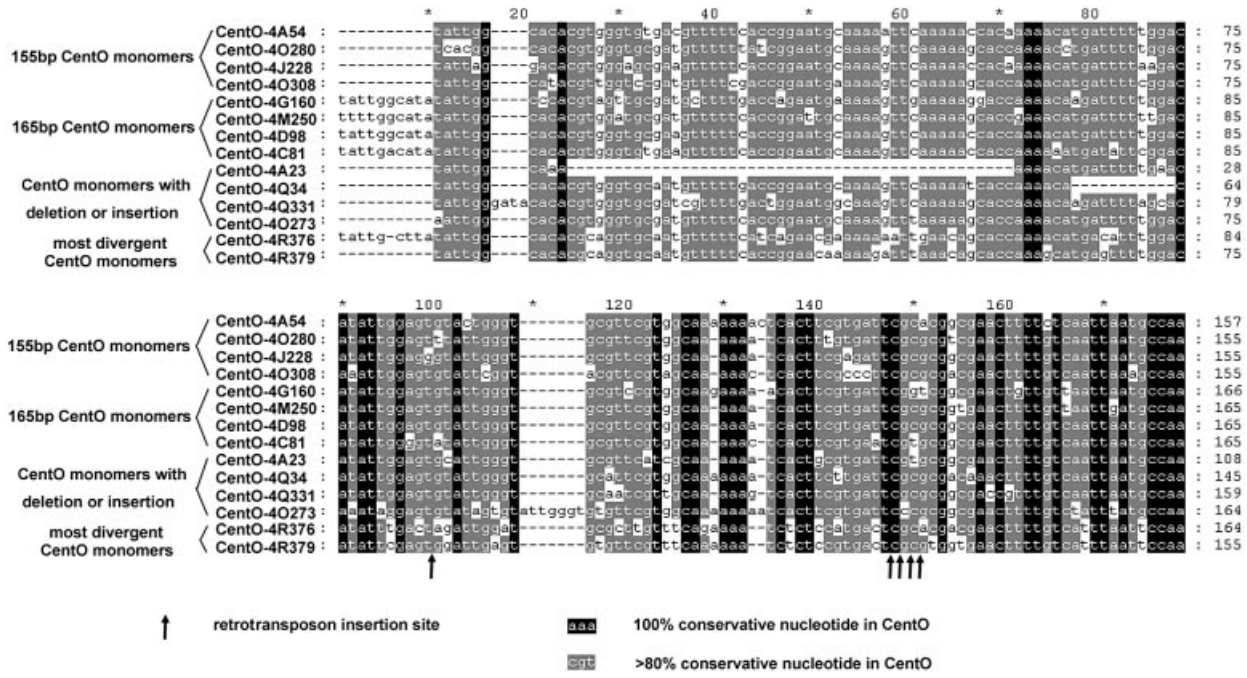

Figure 3. Sequence comparison between different rice chromosome 4 CentO monomers. From top to bottom: four typical 155-bp monomers, four typical 165-bp monomers, four monomers with several nucleotide indels, two monomers at the edge of centromere core region with the most variations. Sequence conservation between different monomers was indicated by background shading. Dark shading represented 100% conservation and light shading >80%. The preferential insertion sites of CR were indicated by the vertical arrows.

Rice CRs

Nineteen retroelements, within the 18 CentO tracts, were divided into four types depending on their structure and sequence homology (Fig. 4).

Type I retroelements belonged to a novel rice retrotransposon family. Eight retroelements of this type, including five solo LTRs, two intact retrotransposons and one incomplete copy, were found in the centromere core region. In this 124 kb core region, we detected only solo LTRs that belong to this type. Solo LTRs are thought to arise from an intra-element recombination between paired LTRs and there would be few solo LTRs in the regions where recombination rates are low. It is interesting to note that there were still some solo LTRs within the centromere since recombination in centromeric regions is known to be highly suppressed. The solo LTRs in centromere suggested that there was no significant correlation between the frequency of solo LTR formation and regional recombination rates, and that different types of solo LTRs may be formed by several different processes. Except the solo LTRs, CR4-2 and CR4-16 were two intact elements with the coding capacity of a *gag* protein. Despite having structural features of LTR retrotransposons, CR4-19 near the edge of the core region diverged extensively and lacked the full ORF that codes for the polyprotein, suggesting that this element was likely non-autonomous.

Type II and type III retroelements were typical Ty3-*gypsy* class retrotransposons with a relatively large 5′ UTR and a polyprotein reading frame that overlapped the downstream LTR; furthermore, type III represented the *gypsy*-type retrotransposon RIRE7 identified previously with the character of preferential insertion into the tandem repeat sequence (CentO) in rice centromere regions (22). These two types were closely related because their coding sequences shared ~80% similarity despite different LTRs and 5′ UTR regions. They were mostly fragmented and truncated retroelements except for a single intact type II retrotransposon, CR4-4.

The type IV element was actually a kind of repetitive sequence with no homology to any known repeats in rice collected by the TIGR Repeat Database, but it was found to be dispersed along the whole chromosome 4. The only copy of type IV repeat in chromosome 4 centromere region was divided into two parts (CR4-15 and CR4-17) by an intact type I retrotransposon (CR4-16).

The types I, II and III retroelements were all typical LTR-retrotransposons and had similar primer binding sites (PBS) complementary to the 3′ end of initiator methionyl tRNA and polypurine tracts (PPT). Through sequence homology searches with the genetically anchored publicly available rice genome sequence, we found that these retroelements were primarily located in the centromeric or pericentromeric heterochromatin regions. However, additional copies of these elements could be found elsewhere across all rice chromosomes. The retroelements were not as unique to the centromere as CentO, the other centromere component.

Based on sequence divergence between 5′/3′ LTRs (Kimura parameter distances) and an estimate of the nucleotide substitution rate for grasses (6.5×10^{-9} substitutions per synonymous site per year) (27), we measured the insertion time of the three retrotransposons (CR4-2, CR4-16 and CR4-4) to be 0.88, 0.19 and 1.63 million years, respectively. The ages of other retroelements were difficult to measure because of the significant sequence degeneracy and the loss of the pair of complete LTRs. These data suggested that type I

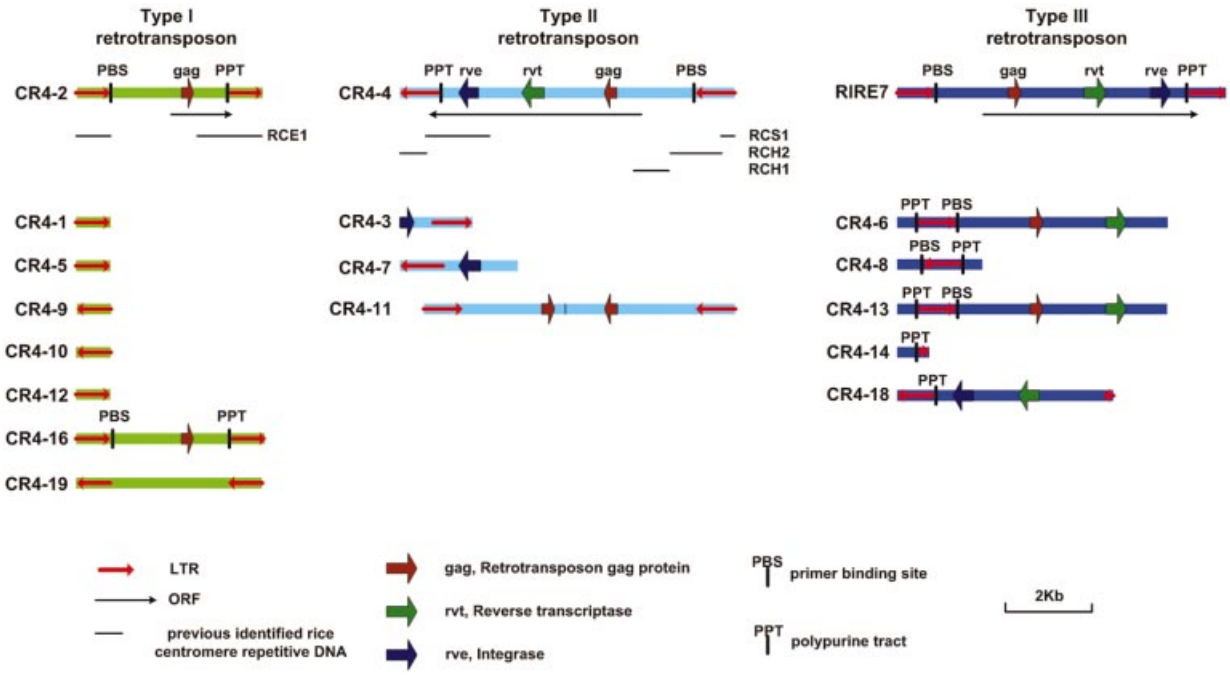

Figure 4. Detailed structural features of three types of CRs. CR4-2, CR4-16 and CR4-11 were intact retroelements. Others were fragments of retroelements. Types I, II and III retroelements were indicated in green, light blue and dark blue, respectively. It was also indicated that the previous identified rice centromeric repetitive DNA fragments (RCE1, RCS1, RCH2, RCH1) were found to correspond to type I or II retroelements.

retroelements transposed more recently than type II retroelements. Homology searches also showed that the type I retroelements only exist in rice, whereas types II and III retroelements have homologs in other cereals. This suggested that type I retroelements have transposed into the centromere more recently than other elements and may be more specific to rice centromeres.

We identified three intact centromeric retrotransposons (CR4-2, CR4-4 and CR4-16) in the chromosome 4 centromere (Fig. 4), while no intact retrotransposon in the centromere had been found in rice in the previous research (21). These intact retrotransposons, which included the LTRs, PBSs and PPTs immediately internal to the LTR and the retrotransposon reading frames, belonged to types I and II retroelements separately, indicating that they had the possibilities to be actively transcribed. The presence of the full intact retrotransposons provided evidence that the CR elements flanking the satellite DNA were capable of initiating transcription, which is thought to be relative to the functional centromere initiation by a transcription-mediated mechanism (33).

Comparison with the rice chromosome 8 centromere-related sequences

Rice chromosome 8 has also been identified to be with limited amount of CentO repeats (18). We searched a rice chromosome 8 centromere-related sequence of B1052H09 (AP006480) that was completely sequenced and assembled from GenBank. Comparative analysis revealed that the CentO spanned region of chromosome 8 centromere was 78 kb and the length of CentO repeats were 68 kb. Chromosome 4 has the most heterochromatic region in the rice genome. About one-third of the chromosome including the entire short arm and the pericentric region of the long arm is highly heterochromatinized (34). Though chromosomes 4 and 8 have similar sizes of CentO repeats (59 and 68 kb), more retroelements inserted into the CentO repeats of chromosome 4 centromere core region. In addition, though the CR elements of rice chromosome 8 centromere were much less than that of chromosome 4, an intact RIRE7 retrotransposon (type III retroelement) was found in the chromosome 8 centromeric region.

DISCUSSION

The centromere is the most essential element of chromosomes for the faithful segregation and inheritance of genetic information in higher eukaryotic species. However, centromeric sequences in different species appear to be variable, although their function is highly conserved (1,2). There seems no doubt that achieving a complete sequence of a chromosome centromere will contribute to establishing the sequences required for the function and a better understanding of the function of the centromere. Unfortunately, so far, the sequences of natural centromeres are not completely available yet, although some detailed sequences about centromere regions have been studied (15,16). The sequence structure of the centromere of rice chromosome 4 reported here is, to our knowledge, the first complete centromere sequence, at least in rice.

DNA sequences associated with centromeric regions have been reported in numerous species including some plant species. The centromeres of higher eukaryotic species are mainly composed of satellite repeats and other repetitive elements. In rice chromosome 4, tandemly arrayed repeats

CentO and 19 CRs constituted the centromere core region. The 155 bp CentO repeats had the similar monomer length with the centromeric satellite repeats in other species, such as alpha satellite in human (15), pAL1 in *Arabidopsis* (16) and CentC in maize (35). The conservative repeat lengths found for most centromeric satellites are thought to be corresponding to the nucleosomal unit lengths (1). Retroelements have been reported to be conserved components of cereal centromeres whereas the retrotransposon family shows no such clear localization in *Arabidopsis* genome (21). Previous researches have identified a highly conserved Ty3/*gypsy*-like retrotransposon family in the centromeres of maize (35), sorghum (36), barley (37), rice (18,22,23) and many other cereals (21).

Several centromeric repetitive DNA elements have been identified to play important roles in interactions with the centromeric-specific histone H3 variant (31,32). And the complete composition of centromeric repeats will provide a platform for identification of centromere function and minimal sequences that provide centromere function. The preferential insertion of the CR elements among CentO repeats indicates that the centromere-specific retroelements have contributed to centromere expansion during evolution. Among the CentO tracts, four tracts of 65 CentO repeats were located in an opposite orientation. Our study also reveals that type I retroelements have transposed into the centromere more recently than other elements and may be more specific to rice centromeres.

The structure of the complete rice chromosome 4 centromere suggests that a certain number of tandemly arrayed repeats and at least one intact retrotransposon element might be necessary for maintaining the full centromere function in higher plants through a transcription-mediated mechanism. Recent studies in fission yeast suggest that either, or both, tandem repeats and LTR retrotransposons may play a role in the heterochromatinization of centromeric DNA through an RNA interference (RNAi) related mechanism (38–40). The heterochromatin that coats centromeric repeats is required for the assembly of an active centromere. As in fission yeast, the centromeric heterochromatins of most eukaryotes are frequently made up of tandem arrays of simple satellites interspersed with other repetitive elements (e.g. retrotransposons). Those tandem repeats and repetitive elements would be subject to triggering chromatin modification by the same mechanism as long as one strand is transcribed and the components of the RNAi machinery are recruited.

Rice provides an almost perfect model to obtain the full centromere sequence, for several rice chromosomes contain a limited amount of the satellite repeats. Beside the centromere of chromosome 4, the centromere of chromosome 8 is also small enough to be covered by constructed BAC contigs and be sequenced. We compared the only two available centromere sequences in rice till now, the complete sequence in chromosome 4 and the centromere-related sequence in chromosome 8 obtained from GenBank, to detect the differentiation of centromere sequence between chromosomes in the same species. The two centromeres comprised almost the same amount of CentO repeats, whereas the number of retroelements varied, which indicated that the similarity between the centromere sequences might be associated with their function.

Since there are no centromeric regions that have been completely sequenced in higher plants, the full structural information of the centromere of rice chromosome 4 may shed some light on the functional dissection of a centromere and identification of minimal sequences that provide the centromere function.

ACKNOWLEDGEMENTS

This work was supported by the grants from the Ministry of Sciences and Technology (2002AA2Z1003 and 2003AA222091), the Chinese Academy of Sciences, the Shanghai Municipal Commission of Sciences and Technology (038019315) and the National Natural Science Foundation of China (30221002 and 30325014).

REFERENCES

1. Henikoff,S., Ahmad,K. and Malik,H.S. (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293**, 1098–1102.
2. Sullivan,B.A., Blower,M.D. and Karpen,G.H. (2001) Determining centromere identity: cyclical stories and forking paths. *Nature Rev. Genet.*, **2**, 584–596.
3. Harrington,J.J., Van,Bokkelen,G., Mays,R.M., Gustashaw,K. and Willard,H.F. (1997) Formation of *de novo* centromeres and construction of first-generation human artificial microchromosomes. *Nature Genet.*, **15**, 345–355.
4. Ikeno,M., Grimes,B., Okazaki,T., Nakano,M., Saitoh,K., Hoshino,H., McGill,N.I., Cooke,H. and Masumoto,H. (1998) Construction of YAC-based mammalian artificial chromosomes. *Nat. Biotechnol.*, **16**, 431–439.
5. Murphy,T.D. and Karpen,G.H. (1995) Localization of centromere function in a Drosophila minichromosome. *Cell*, **82**, 599–609.
6. Kaszas,E. and Birchler,J.A. (1996) Misdivision analysis of centromere structure in maize. *EMBO J.*, **15**, 5246–5255.
7. Clarke,L. (1990) Centromeres of budding and fission yeast. *Trends Genet.*, **6**, 150–154.
8. Clarke,L. (1998) Centromeres: proteins, protein complexes, and repeated domains at centromeres of simple eukaryotes. *Curr. Opin. Genet. Dev.*, **8**, 212–218.
9. The *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **11**, 2012–2018.
10. The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
11. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
12. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
13. Sasaki,T., Matsumoto,T., Yamamoto,K., Sakata,K., Baba,T., Katayose,Y., Wu,J., Niimura,Y., Cheng,Z., Nagamura,Y. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.
14. Feng,Q., Zhang,Y., Hao,P., Wang,S., Fu,G., Huang,Y., Li,Y., Zhu,J., Liu,Y., Hu,X. *et al.* (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420**, 316–320.
15. Schueler,M.G., Higgins,A.W., Rudd,M.K., Gustashaw,K. and Willard,H.F. (2001) Genomic and genetic definition of a functional human centromere. *Science*, **294**, 109–115.
16. Copenhaver,G.P., Nickel,K., Kuromori,T., Benito,M.I., Kaul,S., Lin,X., Bevan,M., Murphy,G., Harris,B., Parnell,L.D. *et al.* (1999) Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science*, **286**, 2468–2474.
17. 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.
18. 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.
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. Nomomura,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.
21. Langdon,T., Seago,C., Mende,M., Leggett,M., Thomas,H., Forster,J.W., Thomas,H., Jones,R.N. and Jenkins,G. (2000) Retrotransposon evolution in diverse plant genomes. *Genetics*, **156**, 313–325.
22. Kumekawa,N., Ohmido,N., Fukui,K., Ohtsubo,E. and Ohtsubo,H. (2001) A new *gypsy*-type retrotransposon, RIRE7: preferential insertion into the tandom repeat sequence TrsD in pericentromeric heterochromatin regions of rice chromosomes. *Mol. Genet. Gemomics*, **265**, 480–488.
23. Nonomura,K.I. and Kurata,N. (2001) The centromere composition of multiple repetitive sequences on rice chromosome 5. *Chromosoma*, **110**, 284–291.
24. Zhao,Q., Zhang,Y., Cheng,Z., Chen,M., Wang,S., Feng,Q., Huang,Y., Li,Y., Tang,Y., Zhou,B. *et al.* (2002) A fine physical map of the rice chromosome 4. *Genome Res.*, **12**, 817–823.
25. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
26. SanMiguel,P., Gaut,B.S., Tikhonov,A., Nakajima,Y. and Bennetzen,J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nature Genet.*, **20**, 43–45.
27. Gaut,B.S., Morton,B.R., McCaig,B.C. and Clegg,M.T. (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene rbcL. *Proc. Natl Acad. Sci. USA*, **93**, 10274–10279.
28. Chen,M., Presting,G., Barbazuk,W.B., Goicoechea,J.L., Blackmon,B., Fang,G., Kim,H., Frisch,D., Yu,Y., Sun,S. *et al.* (2002) An integrated physical and genetic map of the rice genome. *Plant Cell*, **14**, 537–545.
29. Harushima,Y., Yano,M., Shomura,A., Sato,M., Shimano,T., Kuboki,Y., Yamamoto,T., Lin,S.Y., Antonio,B.A., Parco,A. *et al.* (1998). A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics*, **148**, 479–494.
30. Wu,J., Maehara,T., Shimokawa,T., Yamamoto,S., Harada,C., Takazaki,Y., Ono,N., Mukai,Y., Koike,K., Yazaki,J. *et al*. (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell*, **14**, 525–535.
31. Nagaki,K., Talbert,P.B., Zhong,C.X., Dawe,R.K., Henikoff,S. and Jiang,J. (2003) Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics*, **163**, 1221–1225.
32. 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.
33. 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.
34. Cheng,Z., Buell,C.R., Wing,R.A., Gu,M. and Jiang,J. (2001) Toward a cytological characterization of the rice genome. *Genome Res.*, **11**, 2133–2141.
35. 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.
36. Miller,J.T., Dong,F., Jackson,S.A., Song,J. and Jiang,J. (1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. *Genetics*, **150**, 1615–1623.
37. Hudakova,S., Michalek,W., Presting,G.G., ten Hoopen,R., dos Santos,K., Jasencakova,Z. and Schubert,I. (2001) Sequence organization of barley centromeres. *Nucleic Acids Res.*, **29**, 5029–5035.
38. Volpe,T., Schramke,V., Hamilton,G.L., White,S.A., Teng,G., Martienssen,R.A. and Allshire,R.C. (2003) RNA interference is required for normal centromere function in fission yeast. *Chromosome Res.*, **11**, 137–146.
39. Martienssen,R.A. (2003) Maintenance of heterochromatin by RNA interference of tandem repeats. *Nature Genet.*, **35**, 213–214.
40. Schramke,V. and Allshire,R. (2003) Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science*, **301**, 1069–1074.