Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species

Hye-Ran Lee[†], Wenli Zhang[‡], Tim Langdon[§], Weiwei Jin[†], Huihuang Yan[†], Zhukuan Cheng[‡], and Jiming Jiang[†¶]

[†]Department of Horticulture, University of Wisconsin–Madison, Madison, WI 53706; [‡]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China; and [§]Institute of Grassland and Environmental Research, Plas Gogerddan, Aberystwyth SY23 3EB, United Kingdom

Edited by Susan R. Wessler, University of Georgia, Athens, GA, and approved June 14, 2005 (received for review May 10, 2005)

The functional centromeres of rice (*Oryza sativa*, AA genome) chromosomes contain two key DNA components: the CRR centromeric retrotransposons and a 155-bp satellite repeat, CentO. However, several wild *Oryza* species lack the CentO repeat. We developed a chromatin immunoprecipitation-based technique to clone DNA fragments derived from chromatin containing the centromeric histone H3 variant CenH3. Chromatin immunoprecipitation cloning was carried out in the CentO-less species *Oryza rhizomatis* (CC genome) and *Oryza brachyantha* (FF genome). Three previously uncharacterized genome-specific satellite repeats, CentO-C1, CentO-C2, and CentO-F, were discovered in the centromeres of these two species. An 80-bp DNA region was found to be conserved in CentO-C1, CentO, and centromeric satellite repeats from maize and pearl millet, species which diverged from rice many millions of years ago. In contrast, the CentO-F repeat shows no sequence similarity to other centromeric repeats but has almost completely replaced other centromeric sequences in *O. brachyantha*, including the CRR-related sequences that normally constitute a significant fraction of the centromeric DNA in grass species.

centromere | centromeric histone H3 | rice | satellite repeat

Chromosomes in most eukaryotic species contain a single centromere that serves as the site for kinetochore formation and sister chromatid cohesion. Several proteins associated with the centromere/kinetochore complex are highly conserved (1, 2). In particular, a centromere-specific histone H3 variant, referred to as CenH3, appears to be a universal marker for centromeric chromatin, with homologues having been characterized in species, including yeast (Cse4p), insects (CID), vertebrates (CENP-A), and plants (CenH3) (3). Despite the conservation of the centromere function and several centromeric proteins in distantly related species, the DNA sequences associated with the centromeres show little or no conservation, which has been the most interesting enigma in centromere biology.

Human centromeres are the most intensively studied centromeres among multicellular eukaryotes. The main DNA component in human centromeres is the ≈171-bp α satellite repeat (4). Subfamilies of α satellite repeats have diverged significantly not only among the primates but also between chromosomes within the same species (4). Some subfamilies contain a 17-bp motif, called the CENP-B box, which is recognized by CENP-B, a highly conserved centromeric protein in humans and mice (5, 6). Mutations in the CENP-B box within α satellite repeats have been found to alter their capacity to organize a functional centromere (7, 8). Motifs similar to the CENP-B box were reported in the centromeric repeats of various eukaryotes, including insects (9, 10), rodents (11), birds (12), and plants (13, 14). These results suggest that evolution of centromeric satellite repeats may be linked with centromere function/evolution.

Recently, it has been found that some centromeric proteins contain regions that are undergoing rapid adaptive evolution (15–18). It has been proposed that these proteins serve as adaptors that match highly variable centromeric DNA to the well conserved centromeric protein machinery (17), and that their evolution is driven by selection to minimize the consequences of centromeric satellite changes, which may be inherently destabilizing for the genome (19). An intriguing question is whether centromeric satellites are amplified and eliminated solely by the same mechanisms for all tandemly repeated sequences (20, 21) or the evolution of the centromeric satellite is constrained by their interaction with centromeric proteins, as suggested by the presence of the CENP-B box-similar motifs in distantly related species.

Rice and its wild relatives are potentially ideal models to study the coevolution of centromeric DNA and proteins. The centromeres of rice have been well characterized (22–24), and their functional components have been identified as arrays of the CentO satellite and dispersed copies of CRR retrotransposons, members of a well conserved family found in the centromeres of many grass species (25). The CentO satellite repeat shares sequence similarity with the centromeric satellite repeat CentC of maize (23). Some *Oryza* species have been reported to be missing CentO (26), indicating major changes in centromeric DNA composition. To study the evolution of centromeric DNA in *Oryza* species, we developed a chromatin immunoprecipitation (ChIP)-based method to recover DNA from functional centromeres of divergent *Oryza* species. Several previously uncharacterized centromeric satellite repeats were identified, and their distributions were surveyed. Our results reveal rapid evolutionary patterns of centromeric DNA among a group of closely related higher eukaryotic species.

Materials and Methods

Materials. A total of 17 different *Oryza* species containing different genomes were used in the study (Table 1). Two species, *O. rhizomatis* (PI 105440) and *O. brachyantha* (PI 105171), were used in immunoassaying and ChIP cloning. Four species were used in fluorescence *in situ* hybridization (FISH) analysis, including *O. rhizomatis* (105659), *O. brachyantha* (w1057), *O. punctata* (w1564), and *O. latifolia* (w0019). Plasmid pRCS2 (22) was used as a FISH probe to detect the CentO satellite repeat. Six different plasmids derived from different parts of CRR elements (22) were mixed and used as a FISH probe to detect CRR elements.

ChIP Cloning. Nuclei were prepared from young leaf tissue according to published protocols in ref. 27 without cross-linking treatment. ChIP experiments by using the rice anti-CenH3 antibody were conducted as described in ref. 24. Immunoprecipitated DNA was extracted with phenol/chloroform and precipitated with ethanol in the presence of Pellet Paint coprecipitant (Novagen). The extracted

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: ChIP, chromatin immunoprecipitation.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ058468–DQ058600).

See Commentary on page 11573.

[¶]To whom correspondence should be addressed. E-mail: jjiang1@wisc.edu.

Table 1. *Oryza* species used in molecular and cytological analyses of centromeric DNA

Species		Chromosome no.	Genomes	Accession no.	Source
1	*O. sativa* spp. *japonica*	24	AA	Nipponbare	—
2	*O. sativa* spp. *indica*	24	AA	Teging	USDA
3	*O. glaberrima*	24	AA	PI 596842	USDA
4	*O. rufipogon*	24	AA	PI 590417	USDA
5	*O. nivara*	24	AA	PI 590404	USDA
6	*O. meridinalis*	24	AA	ACC 103317	USDA
7	*O. minuta*	48	BBCC	ACC 101386	IRRI
8	*O. punctata*	48	BBCC	w1564*	China
9	*O. officinalis*	24	CC	ACC 105088	IRRI
10	*O. eichingeri*	24	CC	ACC 105163	IRRI
11	*O. rhizomatis*	24	CC	PI 105440 105659*	USDA China
12	*O. alta*	48	CCDD	ACC 105143	IRRI
13	*O. grandiglumis*	48	CCDD	ACC 101405	IRRI
14	*O. latifolia*	48	CCDD	PI 269727 w0019*	USDA China
15	*O. australiensis*	24	EE	PI 101410	USDA
16	*O. brachyantha*	24	FF	PI 105171 w1057*	USDA China
17	*O. granulata*	24	GG	ACC 102118	IRRI
18	*O. meyeriana*	24	GG	ACC 106474	IRRI

IRRI, International Rice Research Institute; USDA, U.S. Department of Agriculture.

*These accessions were used for FISH analysis.

DNA was resuspended in 10 mM Tris/1 mM EDTA, pH 8.0, supplemented with 10 μg/ml RNase A. Precipitated DNA was purified by using the QIAquick PCR purification kit (Qiagen, Valencia, CA) and treated with T4 DNA polymerase at 12°C for 20 min. A-overhangs were added by incubation with TaqDNA polymerase at 72°C for 20 min. Modified DNA was cloned into the pCR 2.1-TOPO vector (Invitrogen). Recombinant clones were transferred to 384-well microtiter plates (Nalge Nunc, Rochester, NY) containing 30 μl of LB freezing buffer. Filter preparation and hybridization were according to published protocols in ref. 28. Immunoprecipitated DNAs were labeled with ^{32}P and purified by using Sepadex G50 columns. Hybridization was carried out at 42°C in ULTRAhyb hybridization buffer (Ambion, Austin, TX). Sequencing was performed by the DNA sequencing facility of the Biotechnology Center at the University of Wisconsin-Madison.

Sequence Analyses. Tandemly repeated elements were identified by using DOTTER (29), other elements were identified by BLAST search against GenBank and the TIGR repeat database (www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml). Specific searches were made by using DOTTER and BLAST against data sets of rice LTR retrotransposons (30) and CRR elements (31), and with PATTERNHUNTER (32) against rice chromosome assemblies, the PLANTSAT database (http://w3lamc.umbr.cas.cz/PlantSat) and GenBank entries whose description identified them as potential centromeric satellites (based on presence of text terms such as "centromer*," "heterochromat*," and "tandem"). The seed model for PATTERNHUNTER was 1110000000011100000000111. Satellite repeat sequences were aligned by MUSCLE (33) and refined manually, with consensus sequences viewed by using the WEBLOGO program (34). All clusters of four or more CentO monomers were extracted from the current TIGR rice pseudomolecule assemblies and aligned with CentC. A total of 811 monomers were then used to generate a consensus for the conserved 80-bp region; 10 of these monomers were chosen as representative of the intraspecific divergence based on a preliminary neighbor-joining analysis for use in tree reconstruction. GenBank accession nos. U63974–U63992 and X86001 were used to represent the CentO sequences from the CCDD genome species *O. alta* and *O. grandiglumis*. *Pennisetum* satellite sequences were obtained from the PLANTSAT database. CentC and CentO-C1 consensus were generated from the subsets of GenBank entries and ChIP-cloned plasmid sequences, respectively, which span the region of similarity. A consensus was also generated for GenBank entries of CentC-similar sequences from *Tripsacum dactyloides*. The phylogeny of aligned sequences was analyzed by the neighbor-joining method with the Jukes–Cantor model as implemented by MEGA3 (35). Similar trees were also reconstructed by using maximum likelihood methods as implemented by PHYLIP (http://evolution.genetics.washington.edu/phylip.html) and MRBAYES (http://mrbayes.net) (data not shown).

Southern Blot and Slot Blot Hybridization. Approximately 5 μg of genomic DNA from different *Oryza* species were digested with selected restriction endonucleases, separated by running on 1% agarose gels, and blotted onto Hybond N+ membranes. DNA fragments corresponding to the LTRs, UTR, and *gag* regions of three CRR subfamilies, CRR1, CRR2, and noaCRR1 (31), were amplified and mixed as a single probe. In estimation of the copy numbers of the satellite repeats, dilutions of rice genomic DNA (5×10^2 to 10^5 copies of the haploid genome) and dilutions of cloned repeats as a control (10^5 to 10^{10} copies) were quantitatively slot-blotted on Hybond N+ membranes. The amount of blotted DNAs were calculated based on genome or insert sizes.

FISH and Chromosomal Immunoassay. FISH and immunoassays on chromosomes were performed according to published protocols (36, 37).

Results

The CentO Satellite Is Not Present in Diploid Wild Rice Species with CC, FF, or GG Genomes. We previously demonstrated that the centromeres of rice chromosomes contain two major components, a 155-bp centromere-specific satellite repeat CentO (22, 23) and CRR retroelements from the widely distributed grass centromeric retrotransposon family (23, 31). Both CentO and CRR sequences are highly enriched in chromatin associated with the rice centromeric histone H3 variant CenH3 (24).

The rice genus, *Oryza*, comprises 23 species (38). These species contain AA, BB, CC, BBCC, CCDD, EE, FF, GG, and HHJJ genomes, respectively (39). A recent PCR-based survey indicated that the CentO repeat is not present in several wild rice species (26). We used Southern blot hybridization to survey CentO distribution in 17 different *Oryza* species (Table 1). Strong hybridization signals were observed in species with AA, BBCC, and CCDD genomes (Fig. 1*A*). Weak signals were observed in the EE-genome species *O. australiensis* and one of the CCDD-genome species, *O. latifolia*. Hybridization signals were not detected in six different species with CC, FF, or GG genomes (Fig. 1*A*), suggesting that in these species the CentO repeats have either diverged significantly or been replaced by unrelated sequences. We conducted immunoassays on the somatic metaphase chromosomes of *O. rhizomatis* (CC) and *O. brachyantha* (FF) by using a rice anti-CenH3 antibody (24). Strong signals were observed on the sister kinetochores of all chromosomes in both species (Fig. 1 *B* and *C*), suggesting that the CenH3 proteins in *O. rhizomatis*, *O. brachyantha*, and cultivated rice are well conserved despite the lack of similarity in centromeric DNA.

ChIP Cloning of Centromeric Sequences in CC- and FF-Genome Species. To investigate the evolution of centromeric DNA in wild rice species we developed a technique to clone sequences from immunoprecipitated DNA (see *Materials and Methods* for details). Briefly, ChIP with the rice anti-CenH3 antibody is carried out by using nuclei isolated from leaf tissue of the target species. DNA fragments associated with the immunoprecipitated complexes are

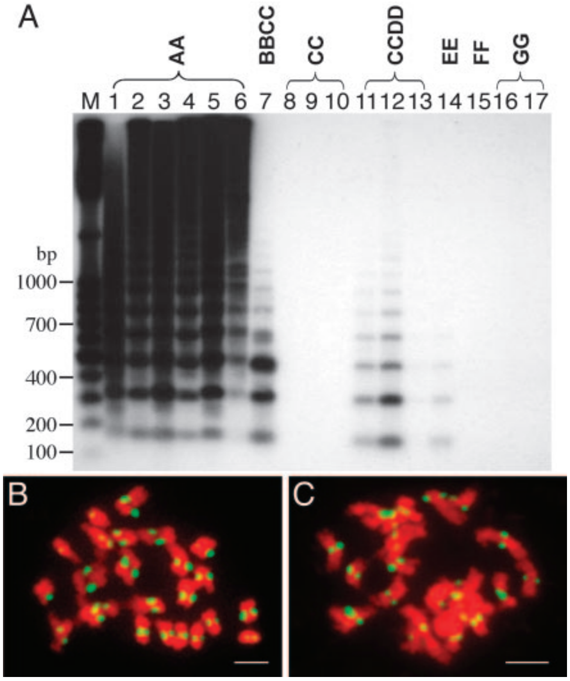

Fig. 1. Presence of the CentO repeat and CenH3 protein in different *Oryza* species. (*A*) Southern blot hybridization of the CentO repeat to TaqI-digested genomic DNA from lanes: 1, *O. sativa* (spp. *japonica*); 2, *O. sativa* (spp. *indica*); 3, *O. glaberrima*; 4, *O. rufipogon*; 5, *O. nivara*; 6, *O. meridinalis*; 7, *O. minuta*; 8, *O. officinalis*; 9, *O. eichingeri*; 10, *O. rhizomatis*; 11, *O. alta*; 12, *O. grandiglumis*; 13, *O. latifolia*; 14, *O. australiensis*; 15, *O. brachyantha*; 16, *O. granulata*; 17, *O. meyeriana*. The genome assignment of each species is indicated at the top of each lane. Lane M is a 100-bp DNA ladder. Note that no hybridization signals were detected in CC-, FF- and GG-genome species. Weak hybridization was observed in lane 13 after a longer exposure of the film. (*B*) Immunoassaying on somatic metaphase chromosomes of *O. rhizomatis* by using a rice anti-CenH3 antibody. (*C*) Immunoassaying on somatic metaphase chromosomes of *O. brachyantha* by using a rice anti-CenH3 antibody. (Scale bars: 5 μm.)

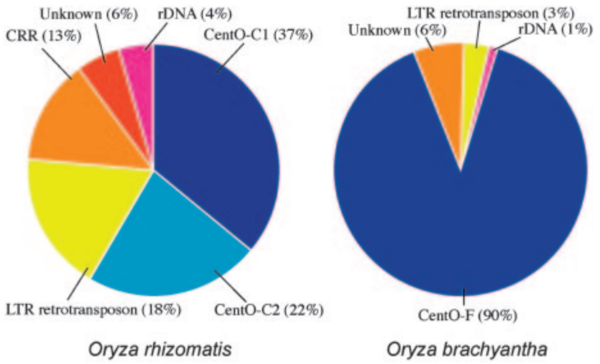

Fig. 2. The proportions of centromeric DNA elements isolated by ChIP cloning in *O. rhizomatis* and *O. brachyantha*. Note that the ChIP cloning libraries may contain various type of DNA sequences, and only the preselected high-copy elements are included in the diagrams.

extracted and cloned to create a plasmid library, which is then screened for centromeric repeats. We developed ChIP cloning libraries from *O. rhizomatis* (CC) and *O. brachyantha* (FF), which do not contain the CentO repeat (Fig. 1*A*). The libraries consisted of 1,152 clones for *O. rhizomatis* and 1,536 clones for *O. brachyantha*. We randomly selected 16 plasmid clones from each library for insert size estimations. The *O. rhizomatis* clones average 303 bp, ranging from 20 to 1,200 bp. The O. *brachyantha* clones average 269 bp, ranging from 50 to 800 bp.

Satellite Repeats Are the Main Centromeric DNA Elements in both *O. rhizomatis* and *O. brachyantha*. We screened the plasmid libraries by using immunoprecipitated DNA from the same species as probes. This screening method should identify high-copy centromeric sequences because these sequences are significantly enriched in the probes. Approximately 18% of the *O. rhizomatis* clones, and 33% of the *O. brachyantha* clones showed medium to strong hybridization signals. We then randomly picked 96 clones with medium to strong hybridizations from each species for sequencing.

Analysis of all 192 sequences indicated that satellite repeats are the most common elements in the data sets for both species (Tables 2 and 3, which are published as supporting information on the PNAS web site). Two classes of tandem repeats were found in the data set of *O. rhizomatis*. The first, CentO-C1, has a monomeric unit size of 126 bp and accounts for 37% of the sequences in the data set. The second, CentO-C2, has a unit size of 366 bp and accounts for 22% of the sequences (Fig. 2). A single class of tandem repeat, CentO-F, with a 154-bp monomer size was found in *O. brachyantha*. This repeat accounts for 90% of the sequences in the data set. We confirmed that each of the repeats is present at a high copy number in their genome of origin by using slot blot hybridizations (data not shown). Assuming that *O. rhizomatis* has the same genome size as *O. officinalis* (CC), which is 1,100 Mb (40), *O. rhizomatis* contains ≈10^4 and 10^5

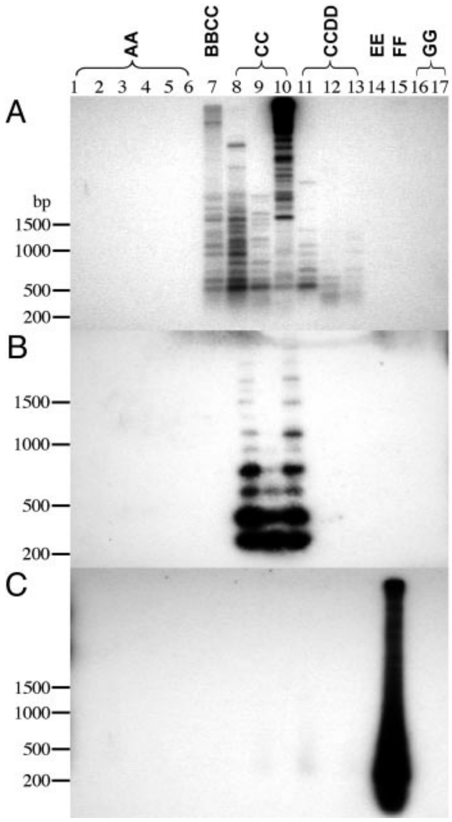

Fig. 3. Distribution of the centromeric satellite repeats CentO-C1, CentO-C2, and CentO-F in different *Oryza* species. (*A*) HaeIII-digested genomic DNAs were probed with the CentO-C1 repeat. (*B*) TaqI-digested genomic DNAs were probed with the CentO-C2 repeat. (*C*) *Tru*9-digested genomic DNAs were probed with the CentO-F repeat. Lanes: 1, *O. sativa* (spp. *japonica*); 2, *O. sativa* (spp. *indica*); 3, *O. glaberrima*; 4, *O. rufipogon*; 5, *O. nivara*; 6, *O. meridinalis*; 7, *O. minuta*; 8, *O. officinalis*; 9, *O. eichingeri*; 10, *O. rhizomatis*; 11, *O. alta*; 12, *O. grandiglumis*; 13, *O. latifolia*; 14, *O. australiensis*; 15, *O. brachyantha*; 16, *O. granulata*; 17, *O. meyeriana*. The genome assignment of each species is indicated at the top of each lane.

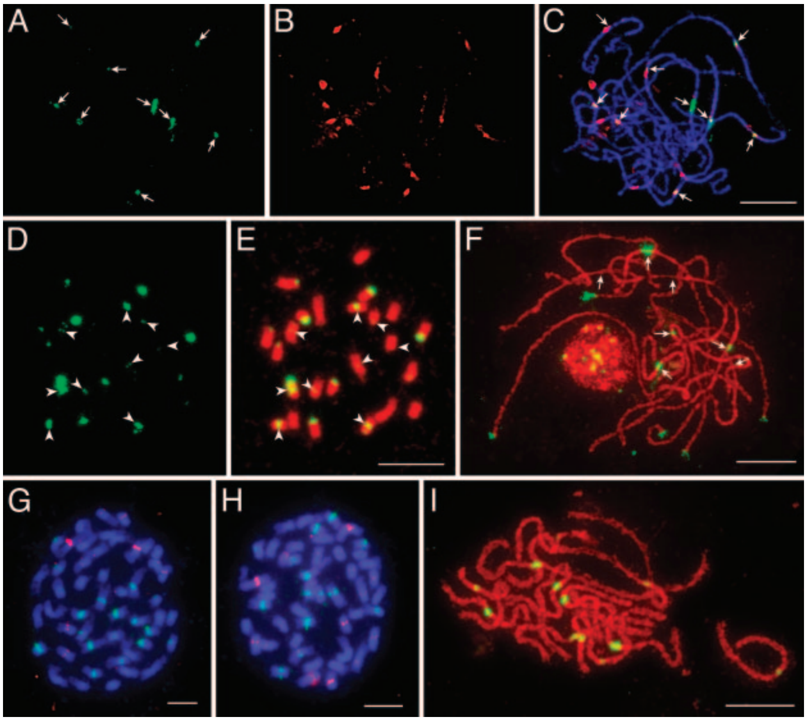

Fig. 4. FISH mapping of centromeric repeats. (*A*) FISH signals derived from satellite repeat CentO-C1. Arrows point to nine unambiguous hybridization sites. (*B*) FISH signals derived from the rice CRR probe (23). (*C*) The FISH signals from CentO-C1 and CRR are merged with the pachytene chromosomes of *O. rhizomatis*. (*D*) FISH signals derived from satellite repeat CentO-C2. Arrowheads point to hybridization sites at the centromeres. (*E*) The FISH signals from CentO-C2 are merged with the somatic metaphase chromosomes of *O. rhizomatis*. (*F*) FISH mapping of CentO-C2 to the pachytene chromosomes of *O. rhizomatis*. Arrows point to unambiguous centromeric signals. (*G*) FISH mapping of CentO-C1 (red signals) and CentO (green signals) to the somatic metaphase chromosomes of *O. punctata* (BBCC). (*H*) FISH mapping of CentO-C1 (red signals) and CentO (green signals) to the somatic metaphase chromosomes of *O. latifolia* (CCDD). (*I*) FISH mapping of the CentO-F satellite repeat to the pachytene chromosomes of *O. brachyantha*. Chromosomes were counterstained with DAPI in all images and pseudocolored in either blue or red. (Scale bars: 5 μm in *G* and *H*; 10 μm in *C*, *E*, *F*, and *I*.)

copies (per haploid genome) of the CentO-C1 and CentO-C2 repeats, respectively. Similarly, the copy number of CentO-F in *O. brachyantha* was estimated to be 2×10^4 per haploid genome. For comparison, detailed FISH measurements of CentO tracts in *O. sativa* indicate that it has a copy number of 4 to 5×10^4 (23).

Southern blot hybridization was used to survey the distribution of the three centromeric repeats in a wide range of *Oryza* species (Fig. 3). The CentO-C1 repeat was detected in all species containing CC genomes, including BBCC and CCDD tetraploids (Fig. 3*A*). The CentO-C2 repeat was detected only in diploid species with CC genomes but not in BBCC or CCDD species (Fig. 3*B*). The CentO-F repeat was detected only in the FF genome species *O. brachyantha* (Fig. 3*C*).

FISH analysis showed that the CentO-C1 repeat is exclusively located at the centromeres of *O. rhizomatis* chromosomes (Fig. 4 *A–C*), and was present in 9 of the 12 chromosomes. The sizes and intensities of the FISH signals varied greatly between chromosomes, suggesting that their centromeres contain different amount of the CentO-C1 repeat. Surprisingly, the CentO-C2 repeat was detected in both centromeric and telomeric regions of several *O. rhizomatis* chromosomes (Fig. 4 *D–F*). Some of the FISH signals were very weak, and the number of detectable FISH signals varied in different metaphase spreads. We counted up to 15 FISH sites at the telomeric regions and 9 sites at the centromeric regions. We also conducted FISH of CentO-C1 on somatic metaphase chromosomes of *O. punctata* (BBCC) and *O. latifolia* (CCDD). CentO-C1 hybridization was observed at the centromeres of less than half of the chromosomes in both species. A CentO probe hybridized to more centromeres than the CentO-C1 probe in both species (Fig. 4 *G* and *H*).

FISH analysis showed that the CentO-F repeat is exclusively located at the centromeres of *O. brachyantha* chromosomes (Fig. 4*I*). The CentO-F repeat was detected in the centromeres of all 12 *O. brachyantha* chromosomes with significantly varied size and intensities of the FISH signals, a pattern which is highly similar to that of CentO when hybridized to *O. sativa* chromosomes (23).

The CentO-C1 Repeat Share Sequence Similarities with CentO and CentC Repeats. CentO was previously reported to share sequence similarity with the 156-bp satellite CentC (23), which is found within the functional centromeres of maize chromosomes (37, 41–43). Comparison with this CentO/CentC alignment revealed no matches for CentO-C2 or CentO-F, but significant similarities were found to CentO-C1 elements over a 80-bp region (Fig. 5 *A* and *B*). This region contains a number of short adenine tracts spaced at regular intervals of ≈10 bp, an arrangement that occurs frequently in satellite DNA and may confer advantageous structural properties (44). These tracts are particularly well conserved within each family, as illustrated by the logo for CentO-C1 (Fig. 5*A*) and were found to underlie the predicted sites of maximum curvature for monomers of each family, as calculated by the BEND.IT server (45). The similarity of the repeat families may therefore reflect a common structure.

Surprisingly, phylogenetic analyses by using a variety of alignments and models consistently place CentO-C1 on a lineage predicted to have diverged from CentO and CentC at around the time that these two last shared a common ancestor (Fig. 5*C*). A possible explanation for this result is that strong selection has driven the assembly of similar sequences from independent elements (convergent evolution), giving the spurious appearance of a common origin for CentO-C1 and CentO/CentC. If this selection were a general feature of plant centromeres or heterochromatin, other sequences with similar structural potential might be expected to be found in relevant databases. We used the sensitive PATTERNHUNTER software (32) to search collections of centromeric and satellite database entries (*Materials and Methods*). Ten-base pair-spaced seed models were used to detect sequences with similar structural potential. From >10,000 entries, a single family of sequences was found that had significant similarity to the 80-bp CentO/CentO-C1/CentO alignment. This family was the ≈150-bp centromeric satellites from pearl millet (*Pennisetum glaucum*) (46, 47), a species related to rice and maize (Fig. 5*B*). Given this relationship, it appears most likely that the *Pennisetum* satellites and CentO-C1 are indeed derived from a common ancestor of CentO and CentC, and that the CentO-C1/CentO divergence predates the rice/maize split. The common size of each subfamily suggests that their ancestor was also ≈150-bp long. The sequences outside of the 80-bp conserved region are highly variable both between and within

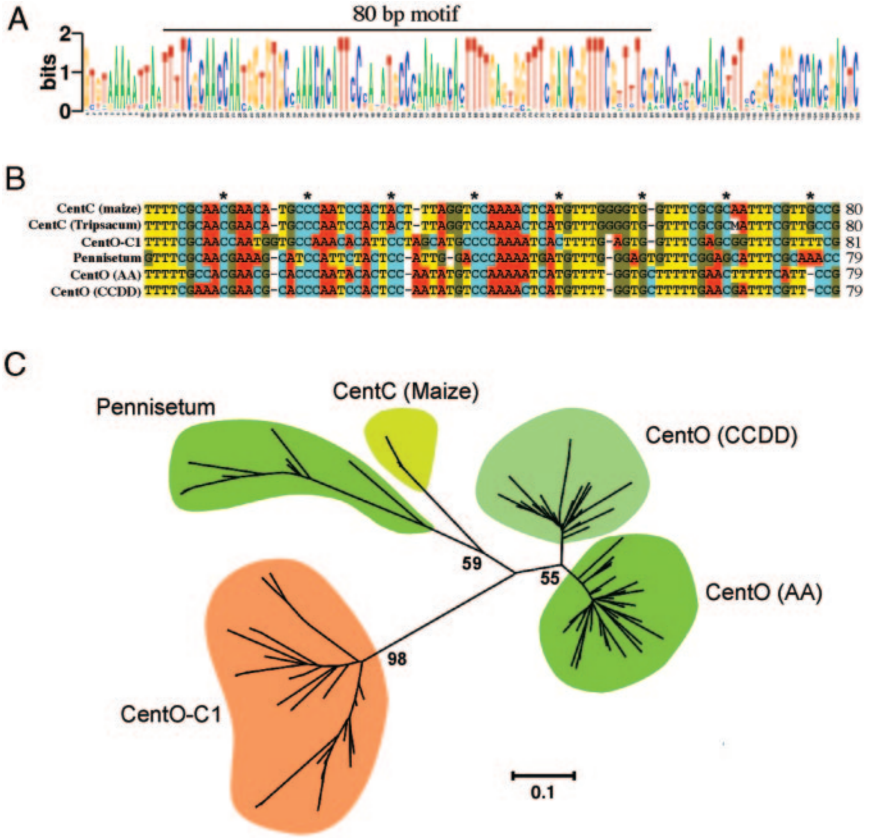
A
80 bp motif
bits
2
1
0
B
CentC (maize) TTTTCGCAACGAACA-TGCCCAATCCACTACT-TTAGGTCCAAAACTCATGTTTGGGGTG-GTTTCGCGCAATTTCGTTGCCG 80
CentC (Tripsacum) TTTTCGCAACGAACA-TGCCCAATCCACTACT-TTAGGTCCAAAACTCATGTTTGGGGTG-GTTTCGCGCMATTTCGTTGCCG 80
CentO-C1 TTTTCGCAACCAATGGTGCCAAACACATTCCTAGCATGCCCCAAAATCACTTTTG-AGTG-GTTTCGAGCGGTTTCGTTTTCG 81
Pennisetum GTTTCGCAACGAAAG-CATCCATTCTACTCC-ATTG-GACCCAAAATGATGTTTG-GGAGTGTTTCGGAGCATTTCGCAAACC 79
CentO (AA) TTTTTGCCACGAACG-CACCCAATACACTCC-AATATGTCCAAAAATCATGTTTT-GGTGCTTTTTGAACTTTTTCATT-CCG 79
CentO (CCDD) TTTTCGAAACGAACG-CACCCAATCCACTCC-AATATGTCCAAAACTCATGTTTT-GGTGCTTTTTGAACGATTTCGTT-CCG 79
C
Pennisetum
CentC (Maize)
CentO (CCDD)
CentO (AA)
CentO-C1
59
55
98
0.1

Fig. 5. Evolution of the CentO-C1 and its related centromeric satellite repeats. (*A*) Consensus sequences of CentO-C1 repeat of *O. rhizomatis* displayed as a sequence logo. The line indicates the location of the 80-bp region that is conserved among the CentO-C1, CentO, and CentC repeats. (*B*) Alignment of the 80-bp region from the consensus sequences of the CentO-C1, CentO (from *O. sativa* with AA genomes and *O. alta* with CCDD genomes), CentC (from *Z. mays* and *T. Tripsacum* genomes), and the centromeric repeat from *P. glaucum*. (*C*) Neighbor-joining tree of the 80-bp region from representative monomers of CentO-C1, CentO, CentC, and the *Pennisetum* subfamilies. The tree was constructed by using MEGA3 (Jukes–Cantor model, γ parameter of 2 for rate variation between sites, gaps included in the calculation). Numbers at branch points represent bootstrap values for 500 replications, and the scale bar represents estimated substitutions per site.

subfamilies, more consistent with a relative lack of selective constraint than the acquisition of species-specific functions/structures.

CentO-C1 Sequences Are Found in the Cultivated Rice Genome. Although CentO-C1 is currently present at a high copy number only in species with a CC genome, the phylogenetic analysis suggests that it was present in the original *Oryza* ancestor. Searches of the cultivated rice (AA) chromosome assemblies identified a small number of CentO-C1-like sequences. Two are found in tandem on chromosome 2 within 160 bp of a CentO-like sequence at 6381552–6381755 bp (TIGR 3 release, http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml). The three monomers are found at equivalent positions in *japonica* and *indica* genomes but are only ≈93% identical (excluding deletions), suggesting a relatively ancient origin (≈5 million years based on the revised intergenic mutation rate, ref. 48). Another tandem CentO-C1 repeat is found adjacent to a CentO cluster near the centromere of *japonica* chromosome 5 at 12337759–12338069 bp, and a partial tandem is found at an equivalent position on both *japonica* and *indica* chromosome 1 (29859493–29859584 bp in *japonica*). Analyses unambiguously group these sequences with CentO-C1 and not with CentO, although they appear to represent a separate lineage within CentO-C1 (data not shown). Given that they are dispersed on a number of chromosomes, they may represent relics of ancestral arrays rather than subsequent introgressions.

Retroelements in the Centromeres of *O. rhizomatis* and *O. brachyantha* Chromosomes. Besides the three previously uncharacterized centromeric satellite repeats, retroelements are the most dominant sequences in the rest of the data sets in both species. In *O. rhizomatis*, the CentO-C1 and CentO-C2 satellite repeats account for 59% of the sequences in the data set (Fig. 2). The remaining sequences include 13% CRR-related DNA fragments and 18% other types of LTR retrotransposons that show homology to the Osr25, Osr30, Osr34, and Osr40 families (30), all of which belong to the Ty3/*gypsy* class (Table 2). In *O. brachyantha*, the CentO-F repeat accounts for 90% of the sequences. The remaining sequences include 3% LTR retrotransposon-related DNA fragments (Fig. 2 and Table 3). Surprisingly, no CRR-related sequences were detected from the data set. FISH analysis by using the CRR probe from rice (23) did not reveal discrete signals in the centromeric region of *O. brachyantha* pachytene chromosomes (data not shown). Southern blot hybridizations confirmed that *O. brachyantha* has few CRR elements, if any. Nine DNA fragments corresponding to LTR, UTR, and *gag*-coding regions from the CRR1, CRR2, and noaCRR1 subfamilies were amplified and mixed as a single probe. Genomic DNA from *O. brachyantha* showed the weakest hybridization to this mixed probe of 16 *Oryza* species tested (data not shown).

Discussion

Although satellite repeats are the most dominant DNA components in the centromeres of most higher eukaryotic species (49, 50), some plant species do not appear to contain major satellite repeats in the centromeres (51, 52). Here, we have demonstrated that satellite repeats make up the bulk of CenH3-associated nucleosomes in *Oryza* species that lack CentO sequences. The percentages of the CentO-C1, CentO-C2, and CentO-F satellite repeats in the ChIP cloning data sets (Fig. 2) and the FISH hybridization patterns (Fig. 4) show that these repeats are the most dominant centromeric DNA components in the diploid *Oryza* species containing the CC and FF genomes.

Our results extend our earlier report that the centromeric repeats of rice and maize share sequence similarity (23). The origin of the CentO/CentC family clearly precedes the divergence of the various *Oryza* genomes, and closely related sequences are still present at high copy numbers in both *O. sativa* and *O. australiensis* (EE), a species relatively divergent from rice (39) (Fig. 1*A*). However, in both the distantly related FF genome and the more recently diverged CC genome, canonical CentO sequences appear to have disappeared. In the FF genome, a single repeat family has replaced both CentO and the bulk of CRR sequences. In the CC genome species, the CentO repeat is replaced by an ancient sister lineage, CentO-C1, as well as another repeat, CentO-C2.

It is striking that CentO homologies can be readily detected among distantly related grass species, and that the best conserved nucleotides are also those predicted to play the most important role in determining the element's structure. Models of this structure (MODEL.IT server, http://hydra.icgeb.trieste.it/~kristian/dna/model_it.html) suggest that the DNA backbone forms a relatively compact solenoid when consensus sequences are used; single nucleotide changes in some A-tracts of the 80-bp region are sufficient to relax the predicted structure, whereas changes outside this region have little effect. Conserved features are also similar to a pattern found to be common in a range of satellites (44), where two 50–60-bp "bending elements" are separated by a 20- to 30-bp region of low curvature. In CentO and relatives, the variable sequences between 80-bp regions typically contain a short GC-rich motif which does not bend in models. CentO and CentO-C1 therefore appear to represent highly evolved satellites that can tolerate only limited sequence change before key structural determinants are lost.

If this model is correct, why and how would the well adapted CentO family be replaced in functional centromeres? One possibility is that replacement satellites initially evolve in regions outside of the functional centromeres and then invade centromeric domains when they have become coadapted with the necessary chromatin components (49). This hypothesis may explain the hybridization of CentO-C2 to subtelomeric and centromeric domains (Fig. 4*E*). Based on the FISH patterns, it appears that there are more copies of the CentO-C2 repeat in the telomeric regions compared with the centromeric regions. Thus, the CentO-C2 repeat likely originated in the subtelomeric regions and was later recruited into the centromeres. It is tempting to speculate that relics of CentO-C1, because they appear to have survived in the *O. sativa* genome, have been revived and reamplified in the CC-genome species, as proposed in the "library" hypothesis (53). In this case, then, any adaptive interaction between centromeric DNA and proteins does not appear to have been driven by the host's need to restrain an "aggressive" satellite.

In contrast, CentO-F may fit the pattern of an element responsible for "centromere drive" (19). It has almost completely replaced the CentO/CentO-C1 family and its amplification in *O. brachyantha* also coincides with the elimination of CRR-related sequences. The CRR elements appear to maintain their presence in the centromeric DNA of a wide range of grass species by using targeted retrotransposition (54), and they are likely to be sensitive to increased rates of satellite array homogenization that eliminate old elements faster than they may be replaced by retrotransposition. CentO-F might therefore be predicted either to have some property that increases its propensity for recombination and, hence, turnover, or to be strongly favored at meiosis so that newly amplified arrays are rapidly fixed in the population.

Both CentO and CentO-C1 are present in tetraploid species *O. punctata* (BBCC) and *O. latifolia* (CCDD). These two repeats are separated in different centromeres (Fig. 4 *G* and H). Interestingly, almost half of the chromosomes in both species contain few or no copies of either of these two centromeric repeats. This result indicates that the centromeric DNA may have undergone rapid and dynamic changes after the formation of the tetraploids. It would be interesting to know whether new centromeric satellite repeats have emerged in these tetraploid species. When the maize chromosome carrying the *CenH3* gene is transferred into the genetic background of an oat, the maize *CenH3* gene is silenced and the maize centromeres adapt the oat CenH3 (37). It will be interesting to know whether only one or both *CenH3* genes from the two parental genomes are expressed in the tetraploid rice species and whether inactivation of a *CenH3* gene has impacted on the elimination of a centromeric satellite from the same parental genome. Such studies will shed more light on adaptive evolution between DNA and proteins in the centromeres.

We thank Dr. Scott Jackson for his valuable comments on the manuscript. This research was supported by Grant FG02-01ER15266 from the Department of Energy and Grant 9975827 from the National Science Foundation (to J.J.). Z.C. is supported by Grant 2002AA225011 from the Chinese State High-Tech Program and Grants 30325008 and 30428019 from the National Natural Science Foundation of China. T.L. is supported by the Biotechnology and Biological Sciences Research Council.

1. Houben, A. & Schubert, I. (2003) *Curr. Opin. Plant Biol.* **6,** 554–560.
2. Amor, D. J., Kalitsis, P., Sumer, H. & Choo, K. H. A. (2004) *Trends Cell Biol.* **14,** 359–368.
3. Henikoff, S., Ahmad, K. & Malik, H. S. (2001) *Science* **293,** 1098–1102.
4. Willard, H. F. (1998) *Curr. Opin. Genet. Dev.* **8,** 219–225.
5. Earnshaw, W. C. & Rothfield, N. (1985) *Chromosoma* **91,** 313–321.
6. Masumoto, H., Masukata, H., Muro, Y., Nozaki, N. & Okazaki, T. (1989) *J. Cell Biol.* **109,** 1963–1973.
7. Ohzeki, J., Nakano, M., Okada, T. & Masumoto, H. (2002) *J. Cell Biol.* **159,** 765–775.
8. Basu, J., Stromberg, G., Compitello, G., Willard, H. F. & Bokkelen, G. V. (2005) *Nucleic Acids Res.* **33,** 587–596.
9. Lopez, C. C. & Edstrom, J. E. (1998) *Nucleic Acids Res.* **26,** 4168–4172.
10. Mravinac, B., Plohl, M. & Ugarkovic, D. (2004) *Gene* **332,** 169–177.
11. Stitou, S., de la Guardia, R. D., Jimenez, R. & Burgos, M. (1999) *Exp. Cell Res.* **250,** 381–386.
12. Solovei, I. V., Joffe, B. I., Gaginskaya, E. R. & Macgregor, H. C. (1996) *Chromosome Res.* **4,** 588–603.
13. Nagaki, K., Tsujimoto, H. & Sasakuma, T. (1998) *Chromosome Res.* **6,** 295–302.
14. Gindullis, F., Desel, C., Galasso, I. & Schmidt, T. (2001) *Genome Res.* **11,** 253–265.
15. Malik, H. S. & Henikoff, S. (2001) *Genetics* **157,** 1293–1298.
16. Talbert, P. B., Masuelli, R., Tyagi, A. P., Comai, L. & Henikoff, S. (2002) *Plant Cell* **14,** 1053–1066.
17. Cooper, J. L. & Henikoff, S. (2004) *Mol. Biol. Evol.* **21,** 1712–1718.
18. Talbert, P. B., Bryson, T. D. & Henikoff, S. (2004) *J. Biol.* **3,** 18.
19. Malik, H. S. & Henikoff, S. (2002) *Curr. Opin. Genet. Dev.* **12,** 711–718.
20. Smith, G. P. (1976) *Science* **191,** 528–535.
21. Charlesworth, B., Sniegowski, P. & Stephan, W. (1994) *Nature* **371,** 215–220.
22. Dong, F., Miller, J. T., Jackson, S. A., Wang, G.-L., Ronald, P. C. & Jiang, J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 8135–8140.
23. Cheng, Z. K., Dong, F., Langdon, T., Ouyang, S., Buell, C. B., Gu, M. H., Blattner, F. R. & Jiang, J. (2002) *Plant Cell* **14,** 1691–1704.
24. Nagaki, K., Cheng, Z. K., Ouyang, S., Talbert, P. B., Kim, M., Jones, K. M., Henikoff, S., Buell, C. R. & Jiang, J. (2004) *Nat. Genet.* **36,** 138–145.
25. Miller, J. T., Dong, F., Jackson, S. A., Song, J. & Jiang, J. (1998) *Genetics* **150,** 1615–1623.
26. Hass, B. L., Pires, J. C., Porter, R., Phillips, R. L. & Jackson, S. A. (2003) *Theor. Appl. Genet.* **107,** 773–782.
27. Wang, H., Tang, W., Zhu, C. & Perry, S. E. A. (2002) *Plant J.* **32,** 831–843.
28. Nizetic, D., Drmanac, R. & Lehrach, H. (1991) *Nucleic Acids Res.* **19,** 182.
29. Sonnhammer, E. L. L. & Durbin, R. (1995) *Gene* **167,** GC1–GC10.
30. Gao, L. H., McCarthy, E. M., Ganko, E. W. & McDonald, J. F. (2004) *BMC Genomics* **5,** 18.
31. Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C. R., Cheng, Z. & Jiang, J. (2005) *Mol. Biol. Evol.* **22,** 845–855.
32. Ma, B., Tromp, J. & Li, M. (2002) *Bioinformatics* **18,** 440–445.
33. Edgar, R. C. (2004) *Nucleic Acids Res.* **32,** 1792–1797.
34. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004) *Genome Res.* **14,** 1188–1190.
35. Kumar, S., Tamura, K. & Nei, M. (2004) *Brief. Bioinformatics* **5,** 150–163.
36. Jiang, J., Gill, B. S., Wang, G. L., Ronald, P. C. & Ward, D. C. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 4487–4491.
37. Jin, W. W., Melo, J. R., Nagaki, K., Talbert, P. B., Henikoff, S., Dawe, R. K. & Jiang, J. (2004) *Plant Cell* **16,** 571–581.
38. Khush, G. S. (1997) *Plant Mol. Biol.* **35,** 25–34.
39. Ge, S., Sang, T., Lu, B.-R. & Hong, D.-Y. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 14400–14405.
40. Martinez, C. P., Arumuganathan, K., Kikuchi, H. & Earle, E. D. (1994) *Jpn. J. Genet.* **69,** 513–523.
41. Ananiev, E. V., Phillips, R. L. & Rines, H. W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 13073–13078.
42. Zhong, C. X., Marshall, J. B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J. A., Jiang, J. M. & Dawe, R. K. (2002) *Plant Cell* **14,** 2825–2836.
43. Jin, W. W., Lamb, J. C., Vega, J. M., Dawe, R. K., Birchler, J. A. & Jiang, J. (2005) *Plant Cell* **17,** 1412–1423.
44. Fitzgerald, D. J., Dryden, G. L., Bronson, E. C., Williams, J. S. & Anderson, J. N. (1994) *J. Biol. Chem.* **269,** 21303–21314.
45. Vlahovicek, K., Kaján, L. & Pongor, S. (2003) *Nucleic Acids Res.* **31,** 3686–3687.
46. Ingham, L. D., Hanna, W. W., Baier, J. W. & Hannah, L. C. (1993) *Mol. Gen. Genet.* **238,** 350–356.
47. Kamm, A., Schmidt, T. & Heslop-Harrison, J. S. (1994) *Mol. Gen. Genet.* **244,** 420–425.
48. Ma, J. & Bennetzen, J. L. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 12404–12410.
49. Csink, A. K. & Henikoff, S. (1998) *Trends Genet.* **14,** 200–204.
50. Jiang, J., Birchler, J. B., Parrott, W. A. & Dawe, R. K. (2003) *Trends Plant Sci.* **8,** 570–575.
51. Houben, A., Brandes, A., Pich, U., Manteuffel, R. & Schubert, I. (1996) *Theor. Appl. Genet.* **93,** 477–484.
52. Schubert, I. (1998) *Trends Genet.* **14,** 385–386.
53. Salser, W., Bowen, S., Browne, D., El Adli, F., Fedoroff, N., Fry, K., Heindell, H., Paddock, G., Poon, R., Wallace, B. & Whitcome, P. (1976) *Fed. Proc.* **35,** 23–35.
54. Langdon, T., Seago, C., Mende, M., Leggett, M., Thomas, H., Forster, J. W., Thomas, H., Jones, R. N. & Jenkins, G. (2000) *Genetics* **156,** 313–325.