Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice

Jianxin Ma and Jeffrey L. Bennetzen*

Department of Genetics, University of Georgia, Athens, GA 30602

Contributed by Jeffrey L. Bennetzen, November 11, 2005

Centromeres have many unusual biological properties, including kinetochore attachment and severe repression of local meiotic recombination. These properties are partly an outcome, partly a cause, of unusual DNA structure in the centromeric region. Although several plant and animal genomes have been sequenced, most centromere sequences have not been completed or analyzed in depth. To shed light on the unique organization, variability, and evolution of centromeric DNA, detailed analysis of a 1.97-Mb sequence that includes centromere 8 (CEN8) of *japonica* rice was undertaken. Thirty-three long-terminal repeat (LTR)-retrotransposon families (including 11 previously unknown) were identified in the CEN8 region, totaling 245 elements and fragments that account for 67% of the region. The ratio of solo LTRs to intact elements in the CEN8 region is ≈0.9:1, compared with ≈2.2:1 in noncentromeric regions of rice. However, the ratio of solo LTRs to intact elements in the core of the CEN8 region (≈2.5:1) is higher than in any other region investigated in rice, suggesting a hotspot for unequal recombination. Comparison of the CEN8 region of *japonica* and its orthologous segments from *indica* rice indicated that ≈15% of the intact retrotransposons and solo LTRs were inserted into CEN8 after the divergence of *japonica* and *indica* from a common ancestor, compared with ≈50% for previously studied euchromatic regions. Frequent DNA rearrangements were observed in the CEN8 region, including a 212-kb subregion that was found to be composed of three rearranged tandem repeats. Phylogenetic analysis also revealed recent segmental duplication and extensive rearrangement and reshuffling of the CentO satellite repeats.

CentO repeats | DNA evolution | nucleotide substitution | unequal conversion

The centromeres of eukaryotic chromosomes are essential for precise chromosome segregation during mitosis and meiosis. Although this function is conserved, the DNA content, organization, and complexity of centromeres vary considerably across different organisms (1–3). Budding yeast (*Saccharomyces cerevisiae*) centromeres consist of only ≈125 bp of unique sequence (4, 5). In contrast, centromeres from most multicellular eukaryotes, including *Arabidopsis* (6–8), rice (9–12), maize (13, 14), *Drosophila* (15), and human (16, 17) are much more complex. These large heterochromatic centromeres consist of large arrays of satellite repeats that are usually arranged in a tandem head-to-tail fashion, intermixed with additional repeats, including transposable elements. Although tandem repeats of some sort appear to be necessary for efficient centromere function in most eukaryotes, the direct effector of kinetochore formation is the assembly of an altered chromatin state in the centromere, associated with a unique H3 histone (CENH3) (18).

In all plant centromeres investigated, the sizes of satellite repeat units (also called monomers) are relatively consistent, ranging from ≈150 to ≈180 bp [e.g., 155 bp for rice CentO (9), 156 bp for maize CentC (13), and 180 bp for *Arabidopsis* pAL1 (19)]. However, the sequences of satellite monomers are quite different among species. The monomers of rice CentO and maize CentC, for instance, do not show significant sequence similarity (9). These differences can arise quite quickly, as shown by the lack of CentO repeat homology between two species within the genus *Oryza*, *Oryza sativa* and *Oryza brachyantha* (20). Moreover, the copy numbers of satellite repeats vary dramatically across species and even among different centromeres of a single organism or the same centromeres from different ecotypes or varieties (9, 21). This variation is partially responsible for tremendous differences in centromere size, as revealed by FISH (9, 14).

Another category of DNA found within plant centromeres is the long-terminal repeat (LTR) retrotransposons (1–3) that usually surround or intermingle with the satellite repeat arrays (9, 11, 12). A few LTR-retrotransposon families have been found to be centromere-enriched by FISH analyses and are called centromere-specific retrotransposons (CRs). These include the CRRs of rice (9, 10, 22–24) and the CRMs of maize (13). A recent survey of CRR distribution indicated that none of the CRR families or subfamilies previously described are completely specific to centromeres (25).

Because of the abundance of repetitive DNA, especially the large arrays of satellite repeats in centromeric regions, the sequencing and assembly of complete centromeres of higher eukaryotes is extremely challenging. Hence, it is not surprising that sequence gaps remain in all centromeres of *Arabidopsis* and human (26, 27). Recently, a 1.97-Mb region including rice CEN8, which contains the least satellite DNA among the 12 centromeres of rice (9), has been completely sequenced (11). Previous studies have described the dynamics of local sequence change across the rice genome, but these studies were completed before completion of any rice centromere (28–30). Because centromeres have such unusual structure and function, it will be interesting to investigate whether these properties influence the rates, mechanisms, or tolerated outcomes of local genome evolution. Here, we present detailed sequence analysis of the CEN8 region from *japonica* rice and its orthologous segments from *indica* rice, including structural analysis of LTR retrotransposons, phylogenetic analysis of CentO satellite repeats, and comparative analysis of genic and intergenic segments from these two subspecies. Our data indicate that the rates and outcomes of divergence differ substantially within different centromere-associated domains and are also quite different from the properties of sequence divergence in noncentromeric regions of rice.

Materials and Methods

Identification of LTR Retrotransposons. Structural analyses and sequence homology comparisons were used to identify LTR retrotransposons in the CEN8 region of *japonica* rice (c.v. Nipponbare) (11). Intact LTR retrotransposons were identified by using LTR-STRUC, an LTR-retrotransposon mining program (28), and by methods previously described (30, 31). Solo LTRs and truncated elements were identified by sequence homology searches against a rice LTR-retrotransposon database that was developed by collecting known LTR retrotransposons and by scanning the 371-Mb Nipponbare genome sequence generated by the International Rice

Conflict of interest statement: No conflicts declared.

Abbreviations: LTR, long terminal repeat; CRR, centromere-specific retrotransposons of rice; mya, million years ago.

*To whom correspondence should be addressed. E-mail: maize@uga.edu.

Genome Sequencing Project (Build 3.0 pseudomolecules, http://rgp.dna.affrc.go.jp/IRGSP/Build3/build3.html). The structures of all LTR retrotransposons identified were confirmed by manual inspection. New LTR-retrotransposon families were defined by the criteria previously described (31).

Sequence Alignments and Comparisons. Targeted query sequences in the CEN8 region of *japonica* cultivar Nipponbare were used in BLASTN and CROSS_MATCH searches against the assembled whole-genome shotgun sequence for *indica* cultivar 93-11 to identify orthologous segments. The orthologous relationship between *japonica* and *indica* sequences was assumed when a query sequence from *japonica* exhibited only one match to assembled 93-11 genome sequence data.

Homologous sequences were aligned by using CLUSTALX (32) and were manually inspected. Synonymous and nonsynonymous substitution distances for the orthologous genes in the CEN8 regions and the euchromatic regions of *japonica* and *indica* (31), and sequence divergences of the two LTRs of single retrotransposons, were calculated as described (31).

A triplication in the CEN8 region was identified by BLAST2 (33) and CROSS_MATCH (University of Washington, Seattle). A segmental duplication of CentO satellite arrays was detected by phylogenetic analysis of all CentO satellite monomers in CEN8.

Results

Analysis of the LTR Retrotransposons in the CEN8 Region. As previously argued (30, 34), the abundance, broad distribution, and known structures of LTR retrotransposons make them particularly well suited surrogates for the study of local genome evolution. The initial focus of this study was to accurately identify LTR retrotransposons and their structures in the rice CEN8 region. This was challenging, given that LTR retrotransposons are highly enriched in CEN8 and are preferentially arranged in a nested pattern (11), and that numerous LTR retrotransposons in rice have undergone various inter-/intraelement rearrangements, primarily through unequal homologous recombination and illegitimate recombination (30, 34). To ensure a detailed characterization of LTR retrotransposons in CEN8, we have optimized the method that was previously developed by our laboratory for identification of LTR retrotransposons (31, 35, 36). A large database of rice LTR retrotransposons was first generated by scanning the whole-genome sequence of rice (c.v. Nipponbare, Build 3.0 pseudomolecules, http://rgp.dna.affrc.go.jp/IRGSP/Build3/build3.html) by using LTR-STRUC (28) and by collecting all rice LTR retrotransposons previously reported [refs. 25, 31, and 37; The Institute for Genomic Research (TIGR), Rockville, MD]. Subsequently, the LTR retrotransposons in CEN8 were identified by a combination of homology searches against this rice LTR-retrotransposon database or on the basis of their canonical structural characteristics (38), and each predicted element was confirmed by manual inspection.

Using the methods described above, we identified 245 LTR retrotransposons in the CEN8 region. These were comprised of 65 intact elements and 61 solo LTRs that are flanked by standard target-site duplications (TSDs), two intact elements and four solo LTRs lacking TSDs, and 113 truncated elements >500 bp (see Table 1, which is published as supporting information on the PNAS web site). Smaller fragments without any recognized PBS (primer binding site), polypurine tract, or TSD were not further investigated.

LTR sequences from LTR retrotransposons identified in this study were compared with each other in a pair-wise manner and compared with LTR sequences extracted from previously reported LTR retrotransposons by CROSS_MATCH and BLASTN. On the basis of LTR sequence homology, 22 known LTR-retrotransposon families and 11 previously unknown LTR-retrotransposon families were identified. These 11 families account for 24 LTR retrotransposons in the CEN8 region (see Table 2, which is published as supporting information on the PNAS web site). The copy numbers of these 33 families vary from 1 to 25 in the CEN8 region, but none of these families is completely unique to either the CEN8 region or the other 11 centromeric regions of rice (data not shown). This observation deviates from the previous description of CRRs revealed by FISH analyses (9) but does parallel a recent survey of distribution of CRRs by sequence analysis (25), indicating that the CRRs are enriched in the centromeric regions of rice, but that rare copies are found elsewhere. Overall, CRRs make up 19% of the core region and 16% of the entire CEN8 region (see Table 2).

A Low Rate of Unequal Recombination for LTR Retrotransposons in the CEN8 Region. A high frequency of solo LTRs, derived from unequal homologous recombination between two LTRs of single retrotransposons (36), has been previously noted in the rice genome (10, 30, 31). In this study, we found that solo LTRs are considerably less abundant on average in the CEN8 region than in most of the characterized genome. The overall ratio of solo LTRs to intact elements in the CEN8 region (0.94:1) is significantly lower ($P < 0.05$) than the 2.2:1 ratio seen in euchromatic regions (31) and also significantly lower ($P < 0.05$) than the 1.6:1 ratio seen in the rice genomic sequences available in GenBank on August 21, 2002, which was primarily composed of noncentromeric regions (30) (see Table 3, which is published as supporting information on the PNAS web site). These data suggest that unequal intraelement recombination has been considerably suppressed in the centromeric region in contrast to noncentromeric regions.

A Hotspot for Unequal Homologous Recombination in the Core of CEN8. Because unequal homologous recombination rates of LTR retrotransposons in the CEN8 region and euchromatic regions of rice are significantly different, we wondered whether the rates of recombination also vary across the CEN8 region. Hence, the 1.97-Mb CEN8 region was dissected into 10 adjacent nonoverlapping subregions (each of the first nine subregions is 200 kb, and the last subregion is 173 kb; Fig. 1), and we subsequently investigated the distribution of solo LTRs and intact LTR retrotransposons in these subregions. We found that the ratio of solo LTRs to intact elements in the core of the CEN8 region is an average of approximately three times higher than in other subregions (Fig. 1). This core region at the center of the CEN8 centromere is included within the CENH3-binding domain that defines the chromosome segregation property of CEN8 (10) but is a distinct subdomain within that region (Fig. 1).

One could propose that this apparent high rate of unequal recombination is a property of the particular families of LTR retrotransposons in the core of the CEN8 region. There are 14 LTR retrotransposons (solo LTRs and intact elements) harbored in the core subregion, and these belong to 12 different families (see Table 2). The distribution and structures of all 91 retrotransposons in the CEN8 region that belong to these 12 families were analyzed. The average ratio of solo LTRs to intact elements of these 12 families in the CEN8 region was found to be 0.90:1 (see Table 2), similar to the value calculated for all retrotransposons in the CEN8 region (0.94:1) (see Table 3). Thus, these families do not show any overall bias toward solo LTR accumulation, suggesting that the hotspot of intraelement unequal recombination is specific to the core subregion.

Comparison of LTR Retrotransposons in the CEN8 Regions of *japonica* and *indica* Rice. To determine the timing and lineage specificities of the dramatic accumulation of LTR retrotransposons in CEN8, we conducted sequence comparison between the genomes of two rice subspecies, *japonica* and *indica*, targeting LTR-retrotransposon insertion sites. For each LTR-retrotransposon insertion in CEN8 of *japonica* cultivar Nipponbare, two unique 500-bp sequences, each composed of 200 bp of one retrotransposon terminal sequence and 300 bp of flanking DNA, were extracted and used to search against

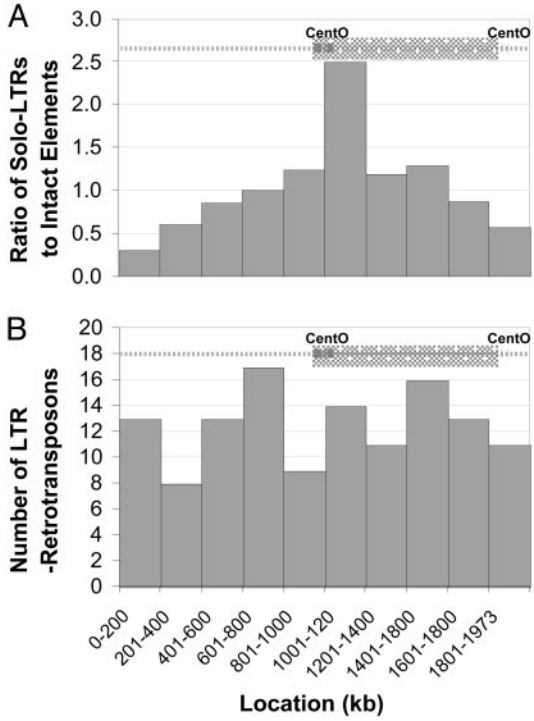

Fig. 1. Variation in the ratios of solo LTRs to intact LTR retrotransposons across the CEN8 region. The CEN8 region was dissected into 10 contiguous subregions, as shown on the *x* axis. Bars indicate ratios of solo LTRs to intact LTR retrotransposons (*A*), and total numbers of solo LTRs plus intact LTR retrotransposons (*B*) in respective subregions.

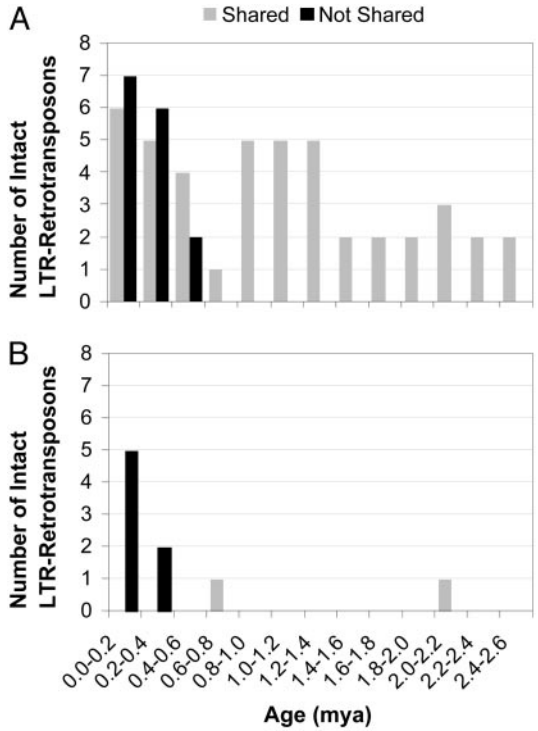

Fig. 2. Ages of LTR retrotransposons in the CEN8 region (*A*) and previously analyzed euchromatic regions (31) (*B*) of Nipponbare rice. Ages were estimated from LTR sequence divergence by using a substitution rate of 1.3×10^{-8} mutatuons per site per year (31). Gray bars show elements shared by Nipponbare (*japonica*) and 93-11 (*indica*); black bars show elements unique to Nipponbare.

the assembled whole-genome shotgun sequence (WGS) generated from *indica* cultivar 93-11 (39, 40). An insertion of an LTR retrotransposon was considered to be shared between *japonica* and *indica* when two unique sequences/regions were found in assembled *indica* WGS that perfectly matched the two query sequences targeting that insertion (see Fig. 4, which is published as supporting information on the PNAS web site). Alternatively, an insertion was judged to be unique in *japonica* when a unique sequence/region was found in assembled *indica* WGS that perfectly matched the combined sequences that are composed of two query sequences without the LTR-retrotransposon terminal sequences (see Fig. 4).

The insertion sites of 126 intact elements and solo LTRs identified in *japonica* were investigated by *japonica* and *indica* comparison. The results indicated that 100 (45 intact elements and 55 solo LTRs) were inserted into the common ancestor of *japonica* variety Nipponbare and *indica* variety 93-11, whereas 18 (15 intact elements and 3 solo LTRs) were inserted into a Nipponbare ancestor after its divergence from a common ancestor with 93-11 (see Table 4, which is published as supporting information on the PNAS web site). It remains unclear whether the other eight elements are shared by *japonica* and *indica*, because there were no sequences found in the *indica* database that match corresponding query sequences flanking these elements. In contrast to the 1.1-Mb noncentromeric orthologous regions from *japonica* (c.v. Nipponbare) and *indica* [c.v. either GLA4 (31) or 93-11], in which >50% of the retrotransposons were unique to *japonica* (see Table 4), the CEN8 region exhibits a significantly ($P < 0.01$, Fisher's exact test) higher percentage (85%) of insertions shared between *japonica* and *indica* (see Table 4). This suggests that the majority of LTR retrotransposons in the CEN8 region accumulated before the divergence of *japonica* and *indica* from a common ancestor.

Estimated Insertion Times of LTR Retrotransposons in the CEN8 Region. LTR divergence was used as a tool to date insertions of LTR retrotransposons (30, 41). The data demonstrated that the majority of the intact LTR retrotransposons in the CEN8 region (Fig. 2 and see Table 5 and Fig. 5, which are published as supporting information on the PNAS web site) were calculated to be slightly older [average insertion dates of 0.8 million years ago (mya)] than those harbored in the 1.1 Mb of euchromatic DNA (average insertion dates of 0.59 mya) that was previously analyzed (31). However, 27 dated LTR retrotransposons in the CENH3 domain (10) were calculated to be significantly older (average insertion dates of 1.06 mya) than those in the euchromatin regions or the flanking pericentromeric heterochromatin (see Fig. 5).

Eleven of 45 intact LTR retrotransposons shared between *japonica* and *indica*, expected to be inserted into rice before the divergence of *japonica* and *indica*, were calculated to be more recent insertions than 0.44 mya [the estimated divergence time of the *indica* and *japonica* varieties studied (31)] (Fig. 2 and data not shown). These results suggest that the LTR retrotransposons in the CEN8 region diverge at a slower rate than in the euchromatic regions (31), where all of the retrotransposons shared between *japonica* and *indica* were found to be older than 0.69 mya (Fig. 2*B*). Hence, the ages of LTR retrotransposons in the CEN8 region could be underestimated if the divergence of LTR sequences is occurring at a slower rate. Alternatively, the CEN8 region analyzed may have been more recently shared by a common Nipponbare and 93-11 ancestor than was the 1.1 Mb of euchromatic DNA previously characterized.

Low Levels of Genic Sequence Divergence in the CEN8 Region. To further address the genomic sequence divergence in the centromeric region, synonymous and nonsynonymous nucleotide substitutions were analyzed in the predicted coding regions of 20 genes in the Nipponbare (*japonica*) CEN8 region and their homologues identified in the 93-11 (*indica*) genome by the method of Nei–Gojobori (42) by using the Jukes–Cantor cor-

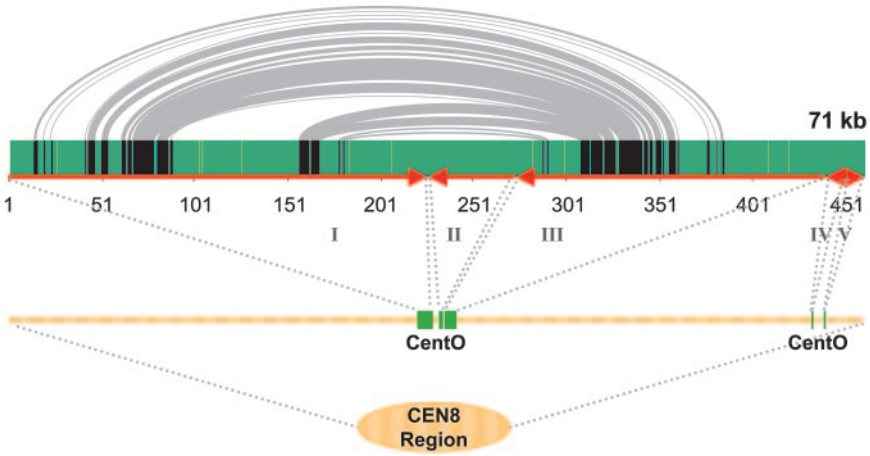

Fig. 3. Segmental duplication of CentO blocks and reshuffling of CentO satellite repeats. Satellite repeats from five CentO blocks (I, II, III, IV, and V) in the CEN8 region are represented by green, black, and yellow vertical lines, The most identical pairs of satellite repeats were often adjacent, but many (indicated by black lines) were most closely related to repeats that were located at a significant distance. These distant pairs are connected by the gray curved lines. The yellow lines indicate satellite repeats that share a 12-bp duplication. The red arrows indicate the orientation of CentO blocks.

rection. Each of these 20 predicted genes was chosen because it had a unique copy in both *japonica* and *indica* genomes (data not shown); therefore, it is likely they represent 20 orthologous loci in the CEN8 regions of *japonica* and *indica* (see Table 6, which is published as supporting information on the PNAS web site). Similarly, we reanalyzed the 24 genes previously investigated in the 1.1-Mb euchromatic regions of Nipponbare (31) and their orthologues in the 93-11 genome (see Table 6).

The 20 pairs of genes in the CEN8 region that we investigated exhibit variable synonymous and nonsynonymous substitution distances, ranging from 0 to 0.0186 and from 0 to 0.0064, with average distances of 0.0025 and 0.0012, respectively (see Table 6). In contrast, the 24 orthologous genes previously investigated in the 1.1-Mb euchromatic regions exhibit average synonymous and nonsynonymous substitution distances of 0.0057 and 0.0016 (see Table 6). The average nonsynonymous substitution distances calculated by using these two sets of genes are almost identical, probably due to similar levels of filtration applied by natural selection. However, the average synonymous substitution distances in the CEN8 region are significantly lower than observed in euchromatic regions. This observation, together with the analysis of LTR-retrotransposon sequence divergence, suggests that the CEN8 region exhibits nucleotide sequence divergence at an ≈1.6- to 2.2-fold lower rate than euchromatic regions, or that the CEN8 region was more recently shared by a common ancestor of the studied *japonica* and *indica* varieties.

Indels and Point Mutations. Because synonymous (gene) and possibly neutral (LTR-retrotransposon) rates of base pair substitution appeared to be lower in centromeric regions, it is important to determine whether the frequency of indel generation was also unusual in the CEN8 region. Previous studies have shown that rice genomic sequences are rapidly removed by small deletions that, in the absence of natural selection to retain the sequence, will remove all nuclear DNA with a half life of <6 million years (30, 31). For the introns of the 20 pairs of orthologous CEN8 region genes described above, it was found that the ratio of indels to point mutations was 0.4 (data not shown), whereas the same ratio (0.4) was observed in the 24 euchromatic genes previously studied (31). Hence, this lower sequence divergence rate or more recent common ancestry was manifest in both classes of sequence variation.

Segmental Duplication in the CEN8 Region. A 212-kb subregion, composed of three tandem triplicated segments "a" (96 kb), "b" (90 kb), and "c" (26 kb), was identified in the CEN8 region (see Fig. 6, which is published as supporting information on the PNAS web site). Segments a and b share two intact LTR retrotransposons and a solo LTR, whereas segments a and c share a 1.6-kb indel that is absent in segment b. Insertions of the two shared intact LTR retrotransposons, *Osr30* and *Osr31*, were dated to 1.35 and 0.27 mya, respectively, demonstrating that the duplication of segment a and b occurred after the insertion of these LTR retrotransposons. Because the overall pair-wise sequence similarities among segments a, b, and c in the shared region are high (99.0–99.3%), it is likely that the two duplications occurred successively in similar time frames. A residue of *Osr33* was detected at the end of segment c, indicating that at least one deletion event occurred after the segmental duplication, thereby removing a large portion of *Osr33* from segment c (see Fig. 6).

A previous FISH study revealed dramatic variation in the amount of CentO satellite DNA between different chromosomes and between the corresponding chromosomes of different varieties of rice (9). Most CentO satellite repeats in the CEN8 region share 91–99% sequence identity with their consensus sequence (11). These observations suggest that many recent amplifications and rearrangements of CentO satellite repeats have occurred in rice centromeric regions.

In an attempt to shed light on the processes that give rise to the dramatic changes in copy numbers of CentO satellite repeats between different chromosomes (9), phylogenetic analysis was performed on all 460 CentO satellite repeats that are clustered in five blocks within the CEN8 region (11) by using the MEGA program (43) (Fig. 3). Based on the neighbor-joining phylogenetic trees obtained (see Fig. 7, which is published as supporting information on the PNAS web site), we identified 48 pairs of highly identical monomers (98–99% sequence similarity) that are dispersed in the two largest CentO blocks (I and III). Interestingly, the orders of these two sets of 48 monomers were considerably conserved in their corresponding CentO blocks, and these blocks are arranged in opposite orientation (Fig. 3), suggesting a recent segmental duplication event that drove the amplification of CentO satellite repeats.

Reshuffling of CentO Satellite Repeats in the CEN8 Region. In addition to the segmental duplication described above, an apparent insertion or deletion (indel) of a cluster of 69 contiguous monomers was found by aligning two duplicated segments (Fig. 3). Some small indels composed of one or a few monomers were also found by comparing the two duplicated segments. Moreover, 11 CentO

satellite monomers were found to share a duplication of 12 bp of DNA, indicating these monomers share a common origin. However, none of these 11 monomers are adjacent to each other. These observations suggest that amplification and reshuffling of CentO satellite repeats have occurred quite often during CentO evolution.

Discussion

The initial establishment of a centromere, as evidenced by studies of neocentromere formation, does not require long stretches of alphoid satellite repeats like CentO of rice, but it does require attainment of a heterochromatic state featuring the specification of CENH3 (the centromere-specific H3 histone) within the neocentromere nucleosomes. However, stable and highly efficient centromere function in chromosome segregation in most animals and plants is associated with satellite repeats interspersed with other repetitive elements like LTR retrotransposons (1–3). Hence, the functional status of a centromere may dramatically influence the accumulation and subsequent divergence of centromere-associated DNA sequences. Sequences with different affinities for kinetochore components may compete, providing a foundation for the theory that meiotic competition between centromeres in the female gametophyte could enhance the rate of centromeric sequence variation (3, 18). To see what unusual effects these functional constraints might have upon centromere sequence divergence, an appropriate first step is to determine the nature and rate of centromere-associated sequence variation.

Previous analyses of the centromeric region of rice chromosome 8 have provided valuable information regarding the content and composition of centromeric DNA in a higher plant species (10, 11). However, the repeat structures in this region were not analyzed in depth, probably due to the difficulties that have been commonly met by most genome sequencing groups in annotation of large sets of complex genomic sequences (44). In this study, 11 previously unknown LTR-retrotransposon families (including 23 elements) were identified, and solo LTRs, truncated LTR-retrotransposon elements, and elements that are arranged in nested patterns (35) were characterized. Combined with the analyses of sequence divergence and rearrangement, this study provides a comprehensive description of a completely sequenced centromeric region in a plant genome.

Recent whole-genome sequencing projects in plants have revealed the dramatic accumulation of LTR retrotransposons in centromeric and pericentromeric regions (11, 26). As evidenced by the 36 intact LTR retrotransposons and solo LTRs found in the CEN8 region in this study, the percentage of LTR-retrotransposon DNA in a genome is considerably underestimated by genome-sequencing projects that concentrate on euchromatic regions. In total, LTR retrotransposons constitute at least 67% of the DNA in the CEN8 region. Assuming that 18–22% of the rice genome is composed of LTR retrotransposons (30, 37), LTR retrotransposons are at least 3- or 4-fold enriched in centromeric region compared with most noncentromeric regions. Although the evolutionary mechanisms behind preferential insertion and/or retention bias for LTR retrotransposons in centromeric regions are still poorly understood, the nonrandom distribution of LTR retrotransposons should at least partly reflect the action of purifying selection against the deleterious effects of LTR-retrotransposon insertion into genes (45).

Given that homologous recombination during meiosis is highly repressed or completely inhibited in all rice centromeres (46, 47), it is not surprising that a low relative abundance of solo LTRs was observed in the CEN8 region compared with previously investigated euchromatic regions (31). A suppression of homologous recombination would be expected to also inhibit unequal recombination events. The observation that LTR retrotransposons in the CEN8 region are older on average than in noncentromeric regions suggests that the other events that remove LTR-retrotransposon sequences from the genome, primarily small deletions caused by illegitimate (i.e., nonhomologous) recombination (30, 36), may also be suppressed, allowing a longer time for intact elements to persist and for solo LTRs to accumulate.

The discovery of a hotspot for solo LTR accumulation in the core of the CEN8 region was unexpected. Previous research has shown that the core of the CEN8 region is part of the kinetochore region that is enriched in CENH3 (10). The kinetochore region of CEN8 was also found to harbor active genes (10). Perhaps the kinetochore and possible other genetic functions of this core subregion have created an environment favoring a higher frequency of homologous recombination than in adjacent subregions. If this is true, then these equal and unequal homologous recombinations must result primarily in noncrossover conversion events, because recombinational mapping indicates few to no meiotic chromosomal exchanges in the entire CEN8 region (46, 47).

The degree of LTR sequence identity has been used to estimate the time of integration of LTR retrotransposons. This dating method is based on the observation that the nucleotide sequences of two LTRs of a single LTR retrotransposon are nearly always identical upon integration (41). However, a possible methodological problem regarding this dating strategy would be that LTR retrotransposons in different genomic regions of an organism could diverge at different rates. Comparative analysis of a set of genic and intergenic orthologous regions of Nipponbare (*japonica*) and GLA4 (*indica*) found that LTR retrotransposons exhibit at least a 2-fold higher rate of single-nucleotide substitution than observed in the synonymous codon positions in the average cereal gene (31). This higher divergence rate is predicted to be primarily a result of the greater transition rate of 5-methyl cytosine compared with cytosine, because most LTR retrotransposons are extensively 5-methylated at cytosine residues in all examined cereal species (48).

Several lines of evidence indicate that different classes of DNA undergo very different rates of sequence evolution in the CEN8 region compared with euchromatic regions. The large number of intact LTR retrotransposons that are shared by Nipponbare (*japonica*) and 93-11 (*indica*) suggest one process that removes intact LTR retrotransposons, illegitimate recombination (31, 36), is suppressed, whereas a second process (unequal homologous recombination) may actually be enhanced. This apparent enhancement of unequal homologous recombination appears to be limited to a small region within the CENH3 domain (10), whereas the low level of sequence deletion is found throughout the centromeric and pericentromeric regions. Overall, a slower LTR-retrotransposon removal process, perhaps combined with an insertion preference for the chromatin state(s) found in centromeric regions, could explain the accumulation of transposons that is a consistent and distinctive feature of most centromeres.

The dating of LTR-retrotransposon insertion via analysis of LTR divergence (38) uncovered 11 apparent shared insertions that are predicted to have occurred after the divergence of the shared Nipponbare and 93-11 ancestor. Although genes in the CEN8 region show a similar degree of nonsynonymous sequence divergence, as do euchromatic-region genes, their degree of synonymous divergence is significantly lower (0.0057 vs. 0.0025). Many of these observations could be explained by a high rate of equal and unequal homologous recombination that frequently leads to conversion of polymorphisms (sequence homogenization) but rarely yields crossover events. A low frequency of crossovers might be beneficial, because it would minimize the rate of variation in centromere structure caused by unequal exchanges. As it is, centromeres are highly variable, but this diversity could be much greater if rates of unequal crossovers between the numerous repeats were not suppressed.

In this study, at least 85% of the LTR retrotransposons in the CEN8 region were found to be shared between the two *japonica* and *indica* varieties investigated. By contrast, only 50% of LTR

retrotransposons were shared in analyzed euchromatic regions of rice (see Table 4) between these same two *indica* and *japonica* genomes. Recently, an LTR-retrotransposon-rich region in rice, flanked by two clusters of genes including the *Orp* locus, was found to share ≈50% of its LTR retrotransposons between *japonica* and *indica* (49). This suggests that the presence of a cluster of repeated DNA, as in the CEN8 and *Orp* regions, does not explain an accumulation of older (i.e., shared) LTR retrotransposons. Moreover, an unusual abundance of solo LTRs is also not a feature associated with most repeat-rich regions. This is a property of the CEN8 core but not of the repetitive regions that flank the CEN8 core nor of the repeat cluster near the *Orp* locus.

Despite an overall lower frequency of homologous exchange and a lower apparent rate of nucleotide sequence divergence, many recent DNA rearrangements have dramatically reshaped the CEN8 region. In addition to the accumulation of LTR retrotransposons, segmental duplications and large indels have been a major force responsible for size expansion of the CEN8 region. Numerous ancient and recent segmental duplications have been documented in several plant and animal species (26, 27, 50–52), but none of these events were demonstrated in centromeric regions before this study.

It is especially interesting that the segmental duplication of a cluster of CentO satellite repeats was revealed by phylogenetic analysis. Subsequently, many indels and reshuffling of a single monomer or multiple contiguous monomers were identified based on alignments of satellite repeats or duplicated CentO segments. Because of their tremendous redundancy, centromeric satellite repeats can be homogenized by unequal conversion, whereas variation in copy number and arrangement can be caused by unequal exchange (53). In this scenario, recent events will obscure more ancient rearrangements. Hence, it is likely that many more reshufflings of CentO satellite repeats have occurred than were deciphered in this study.

We thank Drs. Jiming Jiang and Kiyotaka Nagaki (University of Wisconsin, Madison) for providing CCR sequences before they were released by GenBank (accession nos. AY827956–AY828189), Renyi Liu for assistance in sequence analysis, and Drs. Kelly Dawe and Steve Henikoff for comments on the manuscript. This research was supported by National Science Foundation Plant Genome Program Grant 9975618.

1. Jiang, J., Birchler, J. A., Parrott, W. A. & Dawe, R. K. (2003) *Trends Plant Sci.* 8, 570–575.
2. Lamb, J. C., Theuri, J. & Birchler, J. A. (2004) *Genome Biol.* **5,** 239.
3. Henikoff, S. & Dalal, Y. (2005) *Curr. Opin. Genet. Dev.* **15,** 177–184.
4. Clarke, L. (1990) *Trends Genet.* **6,** 150–154.
5. Clarke, L. (1998) *Curr. Opin. Genet. Dev.* **8,** 212–218.
6. Heslop-Harrison, J. S., Murata, M., Ogura, Y., Schwarzacher, T. & Motoyoshi, F. (1999) *Plant Cell* **11,** 31–42.
7. 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) *Science* **286,** 2468–2474.
8. Kumekawa, N., Hosouchi, T., Tsuruoka, H. & Kotani, H. (2001) *DNA Res.* **8,** 285–290.
9. Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C. R., Gu, M., Blattner, F. R. & Jiang, J. (2002) *Plant Cell* **14,** 1691–1704.
10. Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P. B., Kim, M., Jones, K. M., Henikoff, S., Buell, C. R. & Jiang, J. (2004) *Nat. Genet.* **36,** 138–145.
11. Wu, J., Yamagata, H., Hayashi-Tsugane, M., Hijishita, S., Fujisawa, M., Shibata, M., Ito, Y., Nakamura, M., Sakaguchi, M., Yoshihara, R., *et al.* (2004) *Plant Cell* **16,** 967–976.
12. Zhang, Y., Huang, Y., Zhang, L., Li, Y., Lu, T., Lu, Y., Feng, Q., Zhao, Q., Cheng, Z., Xue, Y., *et al.* (2004) *Nucleic Acids Res.* **32,** 2023–2030.
13. Ananiev, E. V., Phillips, R. L. & Rines, H. W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 13073–13078.
14. Jin, W., Melo, J. R., Nagaki, K., Talbert, P. B., Henikoff, S., Dawe, R. K. & Jiang, J. (2004) *Plant Cell* **16,** 571–581.
15. Sun, X., Le, H. D., Wahlstrom, J. M. & Karpen, G. H. (2003) *Genome Res.* **13,** 182–194.
16. Rudd, M. K., Schueler, M. G. & Willard, H. F. (2003) *Cold Spring Harbor Symp. Quant. Biol.* **68,** 141–149.
17. Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. (2001) *Science* **294,** 109–115.
18. Talbert, P. B., Masuelli, R., Tyagi, A. P., Comai, L. & Henikoff, S. (2002) *Plant Cell* **14,** 1053–1066.
19. Martinez-Zapater, J. M., Estelle, M. A. & Somerville, C. R. (1986) *Mol. Gen. Genet.* **204,** 417–423.
20. Lee, H. R., Zhang, W., Langdon, T., Jin, W., Yan, H., Cheng, Z. & Jiang, J. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 11793–11798.
21. Hall, S. E., Kettler, G. & Preuss, D. (2003) *Genome Res.* **13,** 195–205.
22. Miller, J. T., Dong, F., Jackson, S. A., Song, J. & Jiang, J. (1998) *Genetics* **150,** 1615–1623.
23. Presting, G. G., Malysheva, L., Fuchs, J. & Schubert, I. (1998) *Plant J.* **16,** 721–728.
24. Langdon, T., Seago, C., Mende, M., Leggett, M., Thomas, H., Forster, J. W., Jones, R. N. & Jenkins, G. (2000) *Genetics* **156,** 313–325.
25. Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C. R., Cheng, Z. & Jiang, J. (2005) *Mol. Biol. Evol.* **22,** 845–855.
26. *The Arabidopsis Genome Initiative* (2000) *Nature* **408,** 796–815.
27. 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) *Science* **291,** 1304–1351.
28. McCarthy, E. M. & McDonald, J. F. (2003) *Bioinformatics* **19,** 362–367.
29. Han, B. & Xue, Y. (2003) *Curr. Opin. Plant Biol.* **7,** 134–138.
30. Ma, J., Devos, K. M. & Bennetzen, J. L. (2004) *Genome Res.* **14,** 860–869.
31. Ma, J. & Bennetzen, J. L. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 12404–12410.
32. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997) *Nucleic Acids Res.* **25,** 4876–4882.
33. Tatusova, T. A. & Madden, T. L. (1999) *FEMS Microbiol. Lett.* **174,** 247–250.
34. Bennetzen, J. L., Ma, J. & Devos, K. M. (2005) *Ann. Bot.* **95,** 127–132.
35. SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake Berhan, A., Springer, P. S., Edwards, K. J., Avramova, Z. & Bennetzen, J. L. (1996) *Science* **274,** 765–768.
36. Devos, K. M., Brown, J. K. & Bennetzen, J. L. (2002) *Genome Res.* **12,** 1075–1079.
37. McCarthy, E. M., Liu, J., Gao, L. Z. & McDonald, J. F. (2002) *Genome Biol.* **3,** RESEARCH0053.
38. Kumar, A. & Bennetzen, J. B. (1999) *Annu. Rev. Genet.* **33,** 479–532.
39. Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., *et al.* (2002) *Science* **296,** 79–92.
40. Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., Wei, S., Fu, J., Chen, Y., Ren, X., *et al.* (2004) *Nucleic Acids Res.*
41. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. (1998) *Nat. Genet.* **20,** 43–45.
42. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3,** 418–426.
43. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) *Bioinformatics* **17,** 1244–1245.
44. Bennetzen, J. L., Coleman, C., Liu, R., Ma, J. & Ramakrishna, W. (2004) *Curr. Opin. Plant Biol.* **7,** 732–736.
45. Pereira, V. (2004) *Genome Biol.* **5,** R79.
46. 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) *Plant Cell* **14,** 537–545.
47. Wu, J., Mizuno, H., Hayashi-Tsugane, M., Ito, Y., Chiden, Y., Fujisawa, M., Katagiri, S., Saji, S., Yoshiki, S., *et al.* (2003) *Plant J.* **36,** 720–730.
48. Martienssen, R. A. & Richards, E. J. (1995) *Curr. Opin. Genet. Dev.* **5,** 234–242.
49. Ma, J., SanMiguel, P., Lai, J., Messing, J. & Bennetzen, J. L. (2005) *Genetics* **170,** 1209–1220.
50. International Human Genome Sequencing Consortium (2001) *Nature* **309,** 860–921.
51. She, X., Jiang, Z., Clark, R. A., Liu, G., Cheng, Z., Tuzun, E., Church, D. M., Sutton, G., Halpern, A. L. & Eichler, E. E. (2004) *Nature* **431,** 927–930.
52. Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., *et al.* (2005) *PLoS Biol.* **3,** e38.
53. Smith, G. P. (1976) *Science* **191,** 528–535.