

Retrotransposon-Related DNA Sequences in the Centromeres of Grass Chromosomes

Joseph T. Miller, Fenggao Dong, Scott A. Jackson, Junqi Song and Jiming Jiang

Department of Horticulture, University of Wisconsin, Madison, Wisconsin 53706

Manuscript received June 5, 1998

Accepted for publication August 21, 1998

ABSTRACT

Several distinct DNA fragments were subcloned from a sorghum (*Sorghum bicolor*) bacterial artificial chromosome clone 13I16 that was derived from a centromere. Three fragments showed significant sequence identity to either Ty3/*gypsy*- or Ty1/*copi*a-like retrotransposons. Fluorescence *in situ* hybridization (FISH) analysis revealed that the Ty1/*copi*a-related DNA sequences are not specific to the centromeric regions. However, the Ty3/*gypsy*-related sequences were present exclusively in the centromeres of all sorghum chromosomes. FISH and gel-blot hybridization showed that these sequences are also conserved in the centromeric regions of all species within Gramineae. Thus, we report a new retrotransposon that is conserved in specific chromosomal regions of distantly related eukaryotic species. We propose that the Ty3/*gypsy*-like retrotransposons in the grass centromeres may be ancient insertions and are likely to have been amplified during centromere evolution. The possible role of centromeric retrotransposons in plant centromere function is discussed.

RETROTRANSPOSONS are mobile DNA elements which, like retroviruses, transpose through reverse transcription of an RNA intermediate. Retrotransposons have been characterized according to the yeast/*Drosophila* type elements as either Ty1/*copi*a class or Ty3/*gypsy* class, on the basis of both the order of their protein-coding domains found between the long terminal repeats (LTRs) and their sequence similarities (Xiong and Eickbush 1990). A major difference between the Ty1/*copi*a class and the Ty3/*gypsy* class of retrotransposons is the location of their integrase (IN) domain with respect to the reverse transcriptase (RT) domain. The Ty3/*gypsy*-like elements are like retroviruses in that they are arranged as 5'LTR-*gag*-protease-RT-RNaseH-IN-3'LTR (Figure 1). By contrast, Ty1/*copi*a-like elements are organized as 5'LTR-*gag*-protease-IN-RT-RNaseH-3'LTR (*gag* encoding the structural protein for the capsid).

Numerous retrotransposons have been discovered in plant species (reviewed by Bennetzen 1996). Recent work indicated that retrotransposon-related DNA sequences play a significant role in the organization and evolution of complex plant genomes (Wessler *et al.* 1995; Bennetzen and Kellogg 1997). It is estimated that at least 50% of the nuclear DNA of maize (*Zea mays*) is composed of different retrotransposons (SanMiguel *et al.* 1996). Plant retrotransposons have two major characteristics. First, most of them appear to be limited to a narrow range of related species, indicating

a rapid divergence of such elements during evolution (Fuerstenberg and Johns 1990; Joseph *et al.* 1990; Aledo *et al.* 1995; Brandes *et al.* 1997). Second, based on limited information from cytological analysis, retrotransposons may not be uniformly distributed in plant genomes and many are underrepresented in the centromeric regions (Moore *et al.* 1991; Aledo *et al.* 1995; Brandes *et al.* 1997).

A 90-kb sorghum bacterial artificial chromosome (BAC) clone, 13I16, was derived from a centromere (Jiang *et al.* 1996b). A number of distinct DNA fragments were subcloned from this BAC. Some of these DNA fragments showed significant DNA and amino acid sequence similarities to either Ty1/*copi*a- or Ty3/*gypsy*-like retrotransposons. Here we report the organization and distribution of these sequences in the centromeres of chromosomes from grass species. The potential role of these sequences in centromere function is discussed.

MATERIALS AND METHODS

Materials: A number of species in the grass family Gramineae were used to analyze for the presence of the retrotransposon-related centromeric DNA sequences, including three species from the Bambusoideae subfamily [rice (*Oryza sativa*), bamboo (*Bambusa vulgaris*), and *Pharus* sp.], three species from the Panicoideae subfamily [sorghum, maize, and sugarcane (*Saccharum officinarum*)], seven species from the Pooideae subfamily [barley (*Hordeum vulgare*), *Agropyron intermedium*, *Brachypodium sylvaticum*, oat (*Avena sativa*), rye (*Secale cereale*), wheat (*Triticum aestivum*), and *Aegilops squarrosa*]. Non-Gramineae species included were two other monocots (*Juncus effusus* and *Cyperus alternifolius*) and a dicot species *Arabidopsis thaliana*.

DNA isolation and gel-blot hybridization: Five grams of leaf

Corresponding author: Jiming Jiang, Department of Horticulture, University of Wisconsin-Madison, 1575 Linden Dr., Madison, WI 53706. E-mail: jjjiang1@facstaff.wisc.edu

tissue were ground in liquid nitrogen. The resulting powder was mixed with 6× CTAB (hexadecyltrimethylammonium bromide) and incubated for 1 hr at 60°. An equal volume of chloroform-isoamyl alcohol (24:1) was then added and the contents were gently mixed. The mixture was centrifuged for 10 min at 10,000 rpm and the resultant supernatant was filtered through miracloth and precipitated in an equal volume of cold isopropanol. The DNA was pelleted by centrifuging for 5 min at 10,000 rpm. The pellet was washed with 70% ethanol, dried, and resuspended in Tris-EDTA buffer.

Plant genomic DNA was digested with restriction enzymes, electrophoresed on 1% agarose gels, and transferred to Gene-scan nef-988 membrane. Prehybridization and hybridization were performed at 65° in 5× SSC, 0.5% SDS, 0.02 M NaPO₄ (pH 6.5), 2 mM EDTA, 10 mM Tris (pH 7.4), and 0.02% denatured salmon sperm DNA. Probes were labeled with ³²P and hybridized for 24 hr. Posthybridization washes were performed at either a low-stringency condition (0.5× SSC, 1% SDS at 65°) or a high-stringency condition (0.1× SSC, 1% SDS at 65°).

Sequence analysis: Cycle sequencing reactions were performed using a Sequencing Ready Reaction Kit (Applied Biosystems, Inc., Foster City, CA) and a Perkin-Elmer thermocycler (model 2400; Norwalk, CT) with the following cycling conditions: 95° incubation for 3 min followed by 25 cycles of 95° for 15 sec, 50° for 20 sec, and 60° for 4 min, followed by 72° for 10 min. The reaction products were precipitated with ethanol, dried, and analyzed on an ABI Automated DNA Sequencer (model 373; Columbia, MD). DNA sequences were edited with SeqEd software v1.0.3 and aligned with the Pileup program of the GCG Wisconsin Package v9.1. Homology searches were made against sequences in the nucleic acid database of GenBank using BLASTN. Translated amino acid sequences were compared to the Swissprot protein database using BLASTX and to the translated GenBank sequences using TBLASTX.

To amplify a centromeric DNA fragment from different grass species, two primers were designed based on an ~220-bp sequence that is conserved between sorghum clone pSau3A9 and rice clone pRCS1 (see results). Forward and reverse primers used were 5′GATTTGAAGCCATATTTGGG3′ and 5′GGTCCTCTCCATCATTCCT3′, respectively. The DNA fragments were amplified by polymerase chain reaction (PCR), ligated to pGEM-T vectors (Promega Inc., Madison, WI), transformed into *Escherichia coli* strain XL2, and sequenced.

Fluorescence *in situ* hybridization: Detailed procedure for chromosome preparation and fluorescence *in situ* hybridization (FISH) analysis was described previously (Jiang *et al.* 1996a). The DNA probes were labeled by biotin-11-dUTP and detected by an FITC-conjugated anti-biotin antibody (Vector Laboratories, Burlingame, CA). Chromosomes were counterstained with propidium iodide. The formamide in the hybridization mixture was 50 and 30% in regular and low stringency hybridizations, respectively. Detection and analysis of FISH signals were accomplished using an Olympus (Melville, NY) BX60 microscope with an external Olympus BH2-RFL-T3 epifluorescence source and 60× Olympus PlanApo and 100× Olympus UPlanFl lenses. Images were captured with a SenSys CCD (charge coupled device) camera (Photometrics, Tucson, AZ) coupled to a Macintosh computer. Gray scale images were captured individually and merged using IPLab Spectrum v3.1 software.

RESULTS

A 90-kb sorghum BAC clone, 13I16, was derived from a sorghum centromere (Jiang *et al.* 1996b). Several

distinctive DNA fragments were subcloned from this BAC. Two subclones, pHind22 and pSau3A9, showed significant DNA and amino acid sequence similarities to Ty3/*gypsy*-like retrotransposons. Another subclone, pHind12, had sequence similarity to Ty1/*copia*-like retrotransposons.

The centromeric DNA sequences related to Ty3/*gypsy* retrotransposons are conserved in the centromeres of grass species: *Sorghum clone pHind22:* Clone pHind22 (GenBank accession number AF078901) contains a 510-bp *Hind*III fragment. Significant DNA sequence identity was found between pHind22 and several Ty3/*gypsy* retrotransposons. Nucleotides 27–223 in pHind22 had 57% sequence identity to the *Skipper* element in *Dictyostelium discoideum* (AF017040). Likewise, bases 26–208 had 55% sequence identity to the Tf1 and Tf2 elements of fission yeast (*Schizosaccharomyces pombe*; Weaver *et al.* 1993). The entire pHind22 sequence was aligned with the Tf2 element and sequence similarity was found throughout the 510-bp fragment. On the basis of amino acid sequence similarity with the Tf2 element, it can be deduced that the pHind22 sequence is homologous to a portion of the integrase coding sequence of the Ty3/*gypsy* retrotransposons (Figure 1). Putative amino acid sequence of bases 27–223 in pHind22 showed significant similarity to 25 different Ty3/*gypsy* retrotransposons reported in diverse organisms, including *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *S. pombe*, *Caenorhabditis elegans*, *Z. mays*, *Brassica napus*, *Lilium henryi*, and *A. thaliana* (Figure 2).

Gel-blot hybridizations under high-stringency conditions showed that probe pHind22 hybridized to the genomic DNA from all the grass species analyzed but not to the DNA from *A. thaliana*, a dicot species, nor to monocot species outside of the grass family (Figure 3A). Strong FISH signals were detected in all of the sorghum centromeres (Figure 4A). FISH analysis also revealed that the hybridization signals were specific to the centromeric regions of both A and supernumerary B chromosomes from other grass species (Figure 4, B–D). Unambiguous signals outside the centromeric regions were not observed in any of the species analyzed although undetectable noncentromeric signals cannot be excluded. In several species, the FISH signals were restricted to the primary constriction of metaphase chromosomes (Figure 4, B and C).

Sorghum clone pSau3A9: Clone pSau3A9 contains a 745-bp *Sau*3AI fragment. Like pHind22, the pSau3A9 sequence is specific to the centromeric regions and is conserved in distantly related grass species (Jiang *et al.* 1996b). Probe pSau3A9 also hybridized to the genomic DNA from all the grass species analyzed under high stringency conditions in gel-blot hybridization (Figure 3B). Sequences homologous to the Sau3A9 family were not detected in the GenBank database before the previous paper (Jiang *et al.* 1996b) was published. However, a recent deposit of a partial sequence from a Ty3/*gypsy* retrotransposon of maize (1572 bp in length, AF030633)

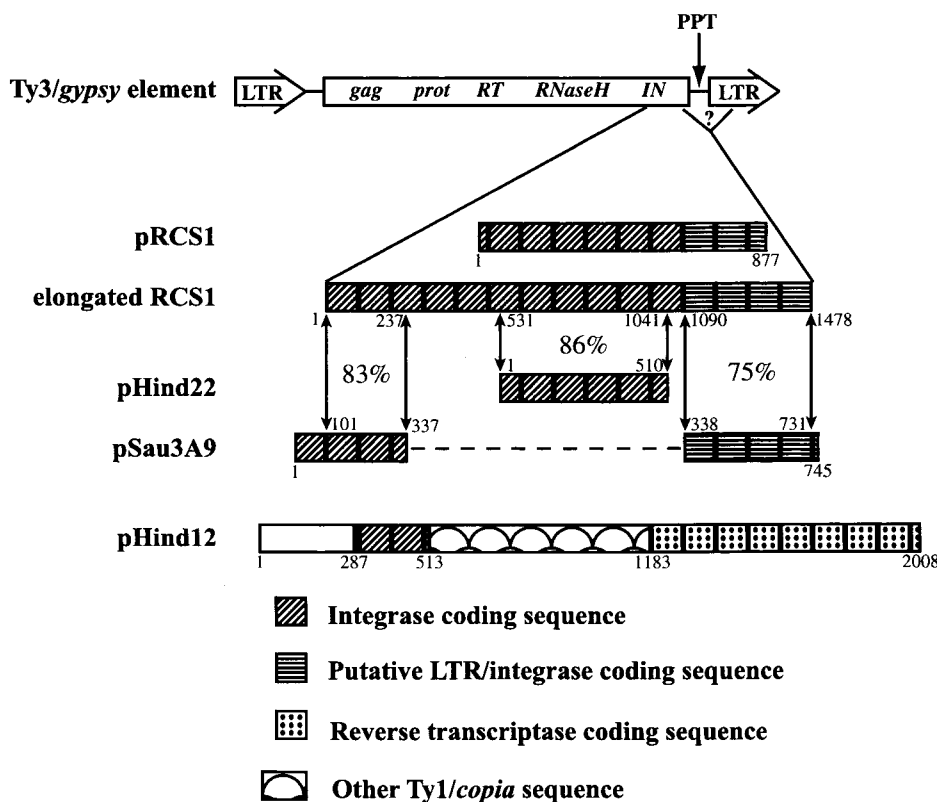


Figure 1.—Schematic diagrams of sorghum and rice clones isolated from centromeric regions. The diagram of Ty3/*gypsy* element is modified from Bennetzen (1996; prot, protease; RT, reverse transcriptase; IN, integrase). Vertical arrows indicate the alignment and similarity between sequences. Nucleotide positions for arrows and sequence ends are marked. Dashed line in clone pSau3A9 indicates the gap that is needed to align it with elongated RCS1. Clone pHind22 aligns to elongated RCS1 in a portion of this gap. Bases 1090–1478 of elongated RCS1 are putative LTR/integrase coding sequence and their position corresponding to the Ty3/*gypsy* element is not known.

showed significant identity to the pSau3A9 sequence. The 337 bp on the 5' end of pSau3A9 has 67% sequence identity to this maize retrotransposon. The deduced amino acid sequence from the same 337 nucleotides showed significant similarity to the sequences from the Tf2 element of *S. pombe* and several other Ty3/*gypsy* retrotransposons in *Drosophila*, flour beetle, and *C. elegans*. On the basis of amino acid sequence similarity with the Tf2 element (Weaver *et al.* 1993), it can be deduced that this 337-bp sequence is homologous to part of the integrase coding sequence of Ty3/*gypsy* retrotransposons (Figure 1).

Nucleotides 338–745 of pSau3A9 had no relationship to any retrotransposons based on searches in both GenBank and Swissprot databases. It is not known if this

fragment is part of the integrase coding region or part of the 3' LTR sequence of the retrotransposon. Probes pHind22 and pSau3A9 hybridized the same DNA fragments from various grass species in gel-blot hybridization (Figure 3), indicating that these two sequences were derived from the same retrotransposon. All retrotransposons contain a polypurine tract that is found immediately before the 3' LTR (Figure 1). This string of 10–18 purines acts as a priming site during reverse transcription of the element. Two purine-rich regions were found at the beginning of this DNA fragment (Figure 5), but it is not known if these regions represent the polypurine tract of this retrotransposon.

Rice clone pRCS1: A rice BAC clone (17p22) derived from a rice centromere was identified by screening a

	Homologous region to elongated RCS1			Homologous region to pHind22			GenBank #	
	GLYMPLFVPT	VPWEDLSMDF	VLGLPRTKRG	RDSIFVVVDR	FSRMAHFIPC	HPQTDGQTEV	VNRTLSTMLR	This study ^a
1	GLLLPLPIPD	QAWQVLSLDF	ISGLP-TSRR	FNCILVVVK	FSRYAHFLAM	HPQSDGQTEW	VNQCVEAYLR	U69258
2	GLLQPLEVSE	CKWEHLMDF	IICFPLSKRC	HDSIWWVDR	FTRSAHFIP	HPQTDG*SER	TIQILEDMLR	X13886
3	WYVSELEIPE	WKWDHLMDF	VTGFPMTFRN	KDAVWVVDR	LTRSAHFLLV	HPQTDGQSER	TIQTLLEDMLR	X99804
4	GLLQPLEIAE	GRWLDLSMDF	VTGLPPTSNN	LNMLVVVDR	FSRRAHFAT	HPQTDGQSER	TIQTLNRLLR	M23367
5	CPLOEILPSE	RPWESLSMDF	ITALPESG-G	YNALFVVVDR	FSRMAILLVPC	RPQTDGQTER	TNQTVEKLLR	L10324
6	Not available					HPQTLGTER	SHRTFNEYVR	S68526

Figure 2.—Alignment of the integrase of Ty3/*gypsy*-like retrotransposons with deduced amino acid sequences from portions of the DNA sequences of elongated RCS1 and pHind22. The DNA sequences corresponding to the two regions in this figure are nucleotides 237–386 of elongated RCS1 and 23–82 of pHind22, respectively. Shaded background indicates that the amino acid is found in all or all but one of the sequences. – indicates a gap and * is a stop codon. ^a GenBank accession numbers for pHind22 and elongated RCS1 are AF078901 and AF078903, respectively. 1, the *Reina* element from *Z. mays*; 2, the *del* element from *L. henryi*; 3, the *Tna1* element from *B. napus*; 4, the Ty3-2 element from *S. cerevisiae*; 5, the Tf2 element from *S. pombe*; 6, the *mdg1* element from *D. melanogaster*.

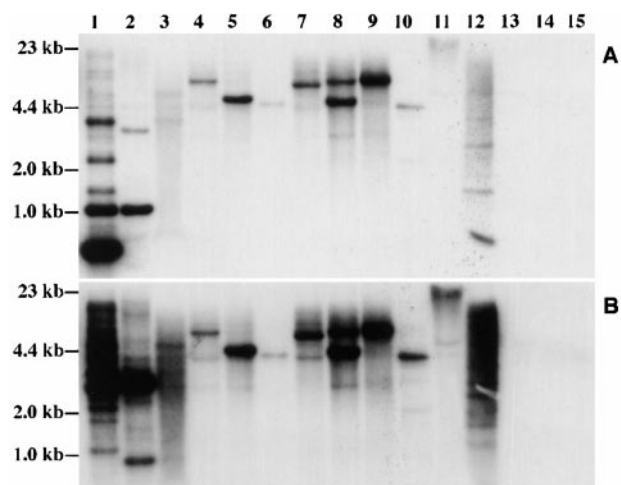


Figure 3.—Conservation of the pHind22 and pSau3A9 sequences in grass species. Genomic DNA from sorghum (lane 1), maize (lane 2), sugarcane (lane 3), *Ag. intermedium* (lane 4), barley (lane 5), oat (lane 6), rye (lane 7), wheat (lane 8), *Ae. squarrosa* (lane 9), rice (lane 10), bamboo (lane 11), *Pharus* sp. (lane 12), *J. effusus* (lane 13), *C. alternifolius* (lane 14), and *A. thaliana* (lane 15) were digested with *Hind*III and probed with pHind22 (A) and pSau3A9 (B), respectively. A stringent wash condition ($0.1\times$ SSC at 65°) was used in the gel-blot hybridization. Signals were detected in all lanes (1–12) of grass species but not in lanes of non-Gramineae species. The two probes hybridized to the same major bands in several lanes.

rice BAC library (Wang *et al.* 1995) using pSau3A9 as a probe (Dong *et al.* 1998). A subclone pRCS1, which hybridized to probe pSau3A9, was isolated from BAC 17p22 (Dong *et al.* 1998). Clone pRCS1 contains an 877-bp *Sau*3AI fragment. Nucleotides 71–580 of pRCS1 had 86% sequence identity to the 510-bp pHind22 sequence (Figure 1). Like pHind22, the 618 nucleotides on the 5' end of pRCS1 had significant sequence identity to the integrase coding sequence of the Ty3/*gypsy* retrotransposons.

The pRCS1 and pSau3A9 sequences were aligned and the 259 nucleotides on the 3' end of pRCS1 (bases 619–877) had 80% sequence identity to a central portion (bases 338–602) of the pSau3A9 sequence, which is a putative coding sequence for the integrase or a putative LTR sequence. The conservation of this DNA fragment between rice and sorghum suggests that it is likely a part of the same centromeric retrotransposon. To analyze the degree of conservation of this fragment in other grass species, two primers were designed for PCR amplification (see materials and methods; Figure 5). A single band around 220 bp was amplified from the genomic DNA of six species analyzed, including sorghum, rice, maize, bamboo, wheat, and *B. sylvaticum*, but it was not amplified from barley. Because the pSau3A9 and pRCS1 sequences are found in every centromere, the amplified products are a mixture of paralogous sequences. One PCR fragment from rice and sorghum and two fragments from each of the other four

species were cloned and sequenced. The 10 PCR fragments ranged from 214 bp to 226 bp and shared at least 60% sequence similarity with each other (Figure 5). The pSau3A9 and pRCS1 sequences and the 10 PCR fragments can be divided into three groups based on the degree of sequence similarity (Figure 5). The two fragments from any one species were not always located within the same group. The sequence similarities within the three groups were 71–87, 92, and 99%, respectively (Figure 5). Sequence data confirmed that this putative LTR/integrase coding sequence is highly conserved among grass species.

Elongated RCS1: We have recently sequenced several hundred M13 clones derived from the rice BAC 17p22 (J. Jiang, unpublished results) and identified additional 473 bp flanking the 5' end of pRCS1 and 128 bp on the 3' end of pRCS1. These flanking sequences extended the pRCS1 to 1478 bp and this contig was named as elongated RCS1 (AF078903) (Figure 1). On the basis of the amino acid sequence similarity of this contig to the Tf2 element of *S. pombe* (Weaver *et al.* 1993), it can be deduced that the 1090 bp on the 5' of the elongated RCS1 is homologous to a portion of the integrase coding sequence of Ty3/*gypsy* retrotransposons (Figures 1 and 2). Two sorghum clones, pHind22 and pSau3A9, had sequence similarity to different portions of this 1090-bp sequence from rice (Figure 1). A gap of 853 nucleotides must be inserted into pSau3A9 in order to align it with the elongated RCS1. The pHind22 sequence had 86% identity to the sequence within this gap (Figure 1).

The centromeric DNA sequence related to Ty1/*cop*ia retrotransposons is not specific to centromeric regions: Another sorghum subclone derived from BAC 13I16, pHind12 (AF078902), contains a 2008-bp *Hind*III fragment (Figure 1). Sequence analysis showed that the 287-bp sequence on the 5' end of pHind12 was homologous to the previously isolated repetitive sequence pSau3A10. The pSau3A10 sequence is a tandem repeat and is located in the centromeres of sorghum and closely related species (Miller *et al.* 1998).

DNA and amino acid sequence analysis revealed that the 1721 bp on the 3' end of pHind12 (bases 288–2008) had significant sequence similarity to several Ty1/*cop*ia retrotransposons, including the PREM-2 element of maize (Turich *et al.* 1996), the ToRTL1 element of tomato (Daraselvia *et al.* 1996), and the *cop*ia element from *Drosophila* (Mount and Rubin 1985). The pHind12 sequence from bases 288 to 2008 was aligned with the homologous sequences in the PREM-2 element of maize (bases 4798–6644) and the *cop*ia element of *Drosophila* (bases 2082–3993), and the overall sequence identities were 61 and 56%, respectively.

Based on amino acid sequence similarity with the *cop*ia element of *Drosophila* (Mount and Rubin 1985), nucleotides 288–513 of pHind12 may code for the N-terminal portion of an integrase protein (Figure 1). Nucleotides 1184–2008 of pHind12 may code for a por-

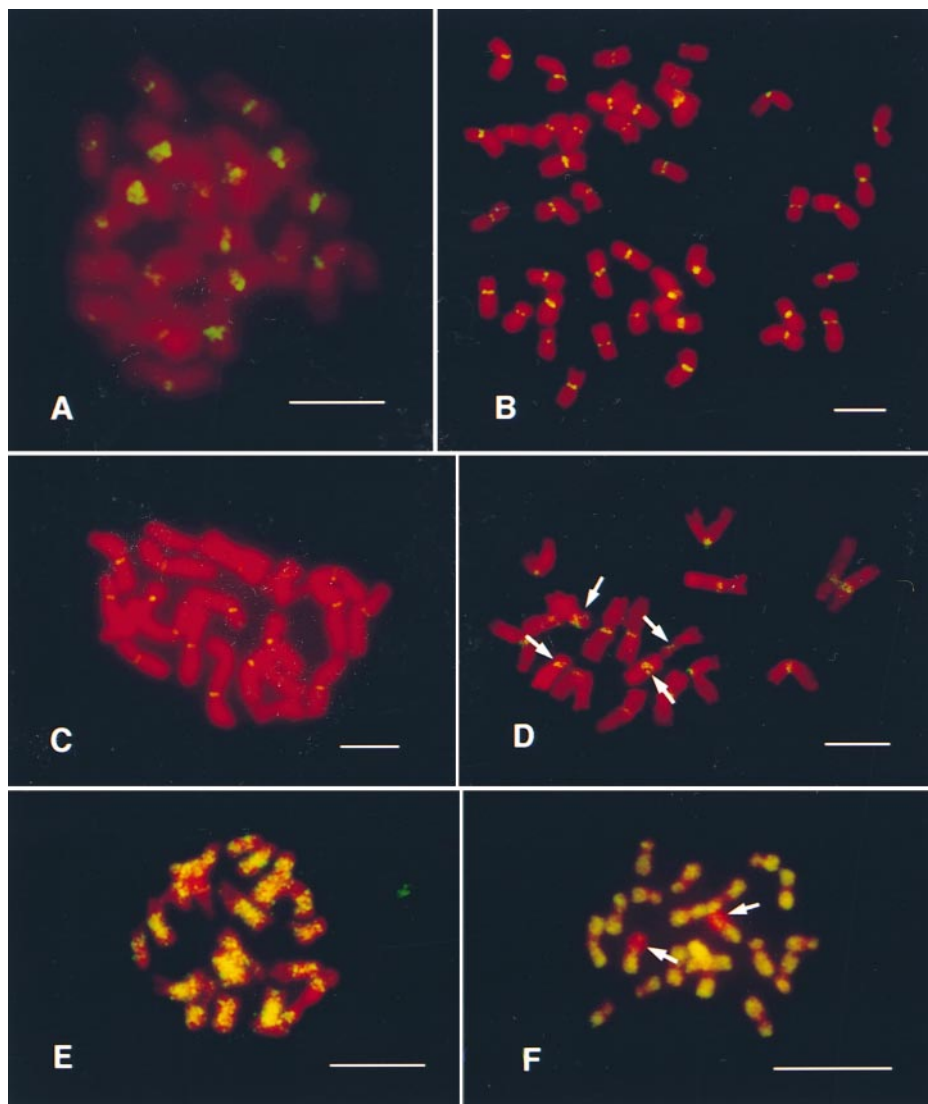


Figure 4.—Chromosomal locations of the retrotransposon-related DNA sequences. DNA probes were detected with a FITC-conjugated anti-biotin antibody (green color). Chromosomes were counterstained with propidium iodide (red color). (A–D) FISH analysis of probe pHind22 on metaphase chromosomes of (A) sorghum, (B) wheat, (C) maize, and (D) rye. Strong signals were observed only in centromeric regions or exclusively in primary constrictions. FISH signals were also present in the centromeric regions of the supernumerary B chromosomes (arrows) of (D) rye. (E) Sorghum metaphase chromosomes were hybridized with a 563-bp fragment (bases 1445–2008) derived from probe pHind12. This DNA fragment hybridized throughout the sorghum chromosomes with enriched signals in the proximal regions. (F) Maize chromosomes were hybridized with the 1.3-kb LTR sequence of the PREM-2 element (Turich *et al.* 1996). Strong FISH signals can be observed on the entire length of all chromosomes. The intensity of the FISH signals was significantly reduced in the NOR (arrows) and in the centromeric regions. (A, C, E, and F) Bar, 5 μ m. (B and D) Bar, 10 μ m.

tion of the reverse transcriptase (Figure 1). Nucleotides 514–1183 had 55% sequence identity to a DNA fragment in the *copia* element of *Drosophila*. This DNA fragment of *copia* separates the coding regions of the integrase and the reverse transcriptase (Mount and Rubin 1985).

The 563 nucleotides (bases 1445–2008) corresponding to part of the reverse transcriptase coding sequence in pHind12 were amplified by PCR and used as a probe for both gel-blot hybridization and FISH analysis. Under low stringency conditions this fragment hybridized to the genomic DNA of many but not all the grass species from three different subfamilies of Gramineae (Figure 6A). Signals were detected only in species within the Panicoideae subfamily, including sorghum, maize and sugarcane, under high stringency conditions (Figure 6B). FISH signals were dispersed throughout the sorghum chromosomes (Figure 4E), indicating this fragment is not specific to the centromeric regions. Since pHind12 has high sequence similarity to the maize PREM-2 element, a clone containing the 1.35-kb LTR

sequence of the maize PREM-2 element (Turich *et al.* 1996) was also used for FISH analysis. Strong signals were observed along the entire length of all maize chromosomes, but the signal distribution was underrepresented in the centromeric and nucleolus organizing regions (NORs) (Figure 4F). FISH signals were not detected on sorghum chromosomes using this probe. The FISH results demonstrated that the distribution patterns of pHind12 and the LTR sequence of PREM-2 on sorghum and maize chromosomes are similar to most previously reported plant retrotransposons (see discussion).

DISCUSSION

Rearrangements of retrotransposon sequences in sorghum BAC 13I16: DNA sequences related to both Ty3/*gypsy* and Ty1/*copia* retrotransposons were identified in the sorghum centromeric BAC clone 13I16. The sequence information suggests that these retrotranspo-

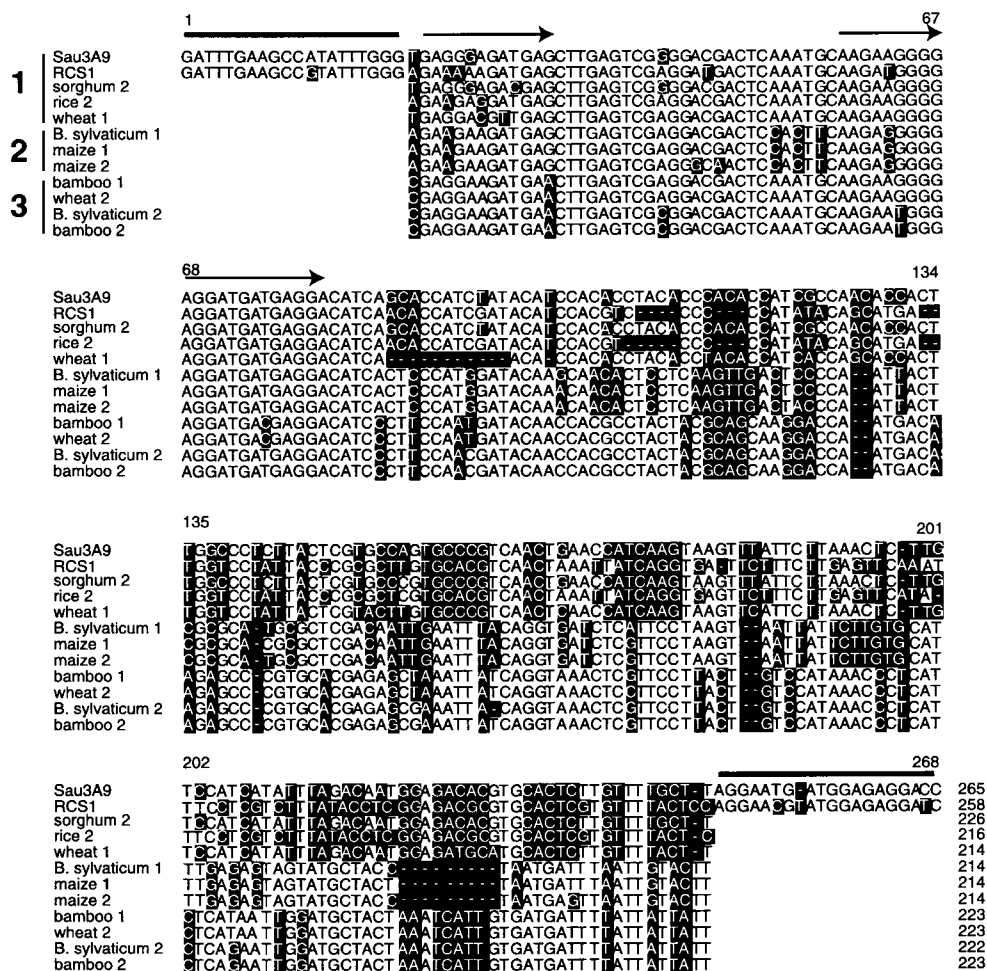


Figure 5.—Aligned sequences of an ~220-bp centromeric DNA fragment amplified from six grass species. The 12 sequences can be divided into three groups based on the degree of sequence similarity. Group 1 includes 2 sequences from sorghum, 2 from rice and 1 from wheat; group 2 includes 2 from maize and 1 from *B. sylvaticum*; group 3 includes 2 from bamboo, 1 from wheat and 1 from *B. sylvaticum*. Sequence similarities within the three groups are 71–87, 92, and 99%, respectively. Unshaded regions indicate consensus nucleotides. Arrows point to the putative polypurine tracts. Bars above sequences indicate primers.

sions are not intact elements and therefore are most likely inactive. The coding region for the integrase in pHind12 was flanked at the 5' end by a sorghum-specific tandem repeat, pSau3A10 (Miller *et al.* 1998) rather than by other parts of the retrotransposon. Sequence alignment of the elongated RCS1 from rice with the sorghum clones pHind22 and pSau3A9 indicated the Ty3/*gypsy*-like element in sorghum BAC 13I16 has also been rearranged (Figure 1). The sequence data suggested that part of the coding region for the integrase in the pSau3A9 sequence has been deleted. However, the rearrangement of the Ty1/*cop* and Ty3/*gypsy*-like elements in sorghum centromeres is not conclusive. One alternative explanation is that the insert of BAC 13I16 has been significantly rearranged, resulting in a deletion in pSau3A9 and a possible sequence translocation in pHind12. Although BAC clones are more stable than yeast artificial chromosomes, the highly repetitive nature of the insert of BAC clone 13I16 may cause such rearrangements.

Distribution of retrotransposon sequences derived from sorghum and rice centromeres: Many transposable elements do not appear to be randomly distributed. For example, *in situ* hybridization analysis in *Drosophila* revealed that the heterochromatic regions accumulate

significantly more transposable elements than the euchromatic regions (Charlesworth *et al.* 1994; Csink and McDonald 1995). To date no general distribution pattern of retrotransposons has been found in plant species. The *BIS1* element of barley and the ZLRS element of maize are distributed along chromosome arms, but are reduced or missing from heterochromatic centromeres, telomeres, and NORs (Moore *et al.* 1991; Aledo *et al.* 1995). Brandes *et al.* (1997) studied the genomic organization of Ty1/*cop* elements in ferns, gymnosperms, and angiosperms. Using degenerate PCR primers for the reverse transcriptase coding region they amplified the Ty1/*cop* sequences from 12 plant species for FISH analyses. Dispersed hybridization throughout the chromosomes was found in most species, but reduced hybridization was detected in NOR and centromeric regions. However, in *A. thaliana* and *Cicer arietinum* the Ty1/*cop*-like elements were clustered in paracentromeric heterochromatin. The *Athila* element in *A. thaliana* is also concentrated in paracentromeric heterochromatin. Defective *Athila* elements were found flanking the major 180-bp centromeric satellite DNA (pAL1) in *A. thaliana* (Pélissier *et al.* 1996).

FISH analysis demonstrated that pHind12, a Ty1/*cop*-related DNA sequence isolated from a sorghum

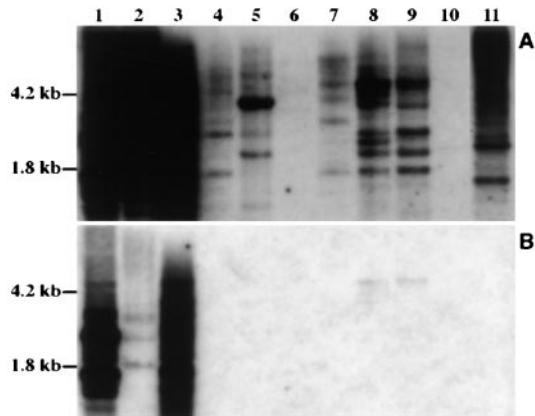


Figure 6.—Gel-blot hybridization of a 563-bp fragment of pHind12 (bases 1445–2008), which is homologous to the coding region for the reverse transcriptase of the Ty1/ *copia* retrotransposon, to *Hind*III-digested genomic DNA from sorghum (lane 1), maize (lane 2), sugarcane (lane 3), *Ag. intermedium* (lane 4), barley (lane 5), oat (lane 6), rye (lane 7), wheat (lane 8), *Ae. squarrosa* (lane 9), rice (lane 10), and *Pharus* sp. (lane 11). Posthybridization washes were performed at (A) low ($0.5\times$ SSC at 65°) and (B) high ($0.1\times$ SSC at 65°) stringency conditions, respectively. Hybridization signals were not detected in oat (lane 6) and rice (lane 10) under low stringency conditions and were detected only in the species within the Panicoideae subfamily, including sorghum, maize, and sugarcane, under high stringency conditions.

centromere, distributed throughout the sorghum chromosomes, a pattern similar to that of many previously reported plant retrotransposons. However, the Ty3/*gypsy*-related DNA sequences, including pHind22, pSau3A9, and pRCS1, had a strikingly restricted distribution pattern to the centromeric regions (Figure 4, A–D; Jiang *et al.* 1996b). Although undetectable FISH signals may exist on other chromosomal regions, it can be concluded that more than 95% of the fluorescent signals were concentrated in the centromeric regions. In several species, the FISH signals can be located within the primary constriction of metaphase chromosomes (Figure 4, B and C). One possible explanation for this centromere-specific distribution is that the copies in the centromeric regions are clustered and thus generate strong FISH signals, whereas the copies outside of the centromeres are too dispersed to be detected. All the available results argue against this hypothesis: (1) gel-blot hybridization of sorghum and rice genomic DNA digested with numerous restriction enzymes suggests that the Ty3/*gypsy*-related centromeric sequences are not tandem repeats, but are dispersed in the centromeric regions; (2) clustered Fiber-FISH signals, characteristic of tandem repeats, were not observed when pHind22, pSau3A9, and pRCS1 were used as probes (S. A. Jackson and J. Jiang, unpublished results); (3) there are only one to two copies of the pHind22 and pSau3A9 sequences in BAC 13I16, indicating that these sequences are not clustered. Therefore, the centromeric Ty3/*gypsy* element described in this report represents a

new type of Ty3/*gypsy* retrotransposon that is conserved exclusively in a specific chromosomal region of distantly related eukaryotic species.

We propose two possible mechanisms for the centromere-restricted distribution pattern. First, it is possible that the retrotransposon identified in the present study preferentially transposed into the centromeres. An example of such region-specific transposition is the telomere-specific retrotransposons reported in *Drosophila* (Mason and Biessmann 1995). Transposition specific to the pericentromeric regions has also been discovered in humans even though the mechanism of such transpositions is still poorly understood (Eichler *et al.* 1996, 1997; Regnier *et al.* 1997). Second, the retrotransposons that transposed into the rice and sorghum centromeres may have been amplified by an unknown mechanism. A recent report from a mammalian species supports the possibility of this model (O'Neill *et al.* 1998). This research discovered a dramatic amplification of retrotransposon-related DNA sequences in the centromeric regions of macropodid chromosomes. The amplification was caused by undermethylation induced by interspecific hybridization (O'Neill *et al.* 1998).

Evolution of retrotransposon sequences in sorghum and rice centromeres: Most plant retrotransposons appear to be limited to a narrow range of related species or a single genus based on gel-blot hybridization experiments (Bennetzen 1996). The ZLRS element of maize hybridized to all *Zea* species but not to those in its sister genera *Tripsacum* or *Saccharum* (Aledo *et al.* 1995). The *Bs1* element of maize hybridized only to *Zea* and *Tripsacum* species under low stringency conditions (Fuerstenberg and Johns 1990). Under a high stringency the *BIS1* element from barley hybridized to wheat and rye but not to oat (Moore *et al.* 1991). The *del* element in *L. henryi* was found in most but not all *Lilium* species (Joseph *et al.* 1990). The DNA sequence coding for the reverse transcriptase of a Ty1/ *copia* element in *Pennisetum glaucum* hybridized, under low stringency conditions, to *Pennisetum*, *Setaria*, and barley but not to other grasses such as wheat and rye (Brandes *et al.* 1997). In all these reports the retroelements did not hybridize to the genomic DNA of species outside a genus or a tribe.

Surprisingly, the Ty3/*gypsy*-related DNA sequences identified in the sorghum centromeres were detected in a much wider range of plant species than all previously reported retrotransposons. Positive gel-blot hybridization signals were detected in grass species across the three examined subfamilies of the Gramineae when pSau3A9 and pHind22 were used as probes (Figure 3). There are two possible explanations for this rare conservation. First, the centromeric Ty3/*gypsy* retrotransposons may represent ancient transpositions and were amplified possibly before the divergence of the grass species. Mutation and other modifications of these centromeric Ty3/*gypsy* sequences have accumulated at

a much slower pace than retrotransposons located outside the centromeres, resulting in the high conservation within the centromeric regions. Second, the centromeric Ty3/*gypsy* sequences might be associated with centromere function and functional constraints result in the high conservation (see below).

Transposable elements and centromere function: In *S. cerevisiae*, a 125-bp DNA sequence encodes all the information needed for full centromere function (Clarke 1990). The centromeres of chromosomes from other eukaryotic species, including *S. pombe*, *D. melanogaster*, humans, and plants, encompass many kilobases or even megabases of DNA (Clarke 1990; Willard 1990; Murphy and Karpen 1995; Kaszas and Birchler 1996; Round *et al.* 1997; Miller *et al.* 1998). DNA sequences responsible for centromere function in these species have not been fully defined. Recently, the centromere of a minichromosome from *D. melanogaster* has been located within a 420-kb region (Sun *et al.* 1997). This region is composed of satellite DNA and single, complete transposable elements. The fine scale restriction maps of the transposable elements indicated that they were nearly identical to previously published elements, suggesting that these centromeric elements are recent insertions, or that they are ancient insertions conserved due to selective/functional constraints (Sun *et al.* 1997). Since the transposable elements identified in the centromere of the minichromosome are neither unique to the centromeres nor present in all centromeres, it is not known whether these elements play a direct role in centromere function.

A relationship between transposable elements and centromere structure has also been proposed in mammalian species (Kipling and Warburton 1997). A highly conserved centromere-associated protein, CENP-B, is a common feature of mammalian centromeres. Binding sites for CENP-B, called "CENP-B boxes," are present in the otherwise unrelated centromeric satellite DNA sequences identified in various mammalian species (Masumoto *et al.* 1989; Kipling *et al.* 1995; Kipling and Warburton 1997), suggesting a role for CENP-B in centromere function. Extensive sequence similarity was found between CENP-B and the transposase protein encoded by the *pogo* superfamily of transposable elements. CENP-B is proposed to be involved in promoting recombination in the centromeric regions (Kipling and Warburton 1997). In this hypothesis, CENP-B, or a transposable element, facilitates the evolution and maintenance of centromeric DNA sequences, rather than playing a direct role in centromere function.

We have identified a highly conserved Ty3/*gypsy*-like retrotransposon in the centromeres of grass species. In several aspects this Ty3/*gypsy*-like retrotransposon is different from the transposable elements found in the centromere of the *D. melanogaster* minichromosome. First, preliminary sequence data suggest that the Ty3/*gypsy*-like retrotransposons in sorghum centromeres are

not intact elements, while the transposable elements identified in *Drosophila* are all complete elements (Sun *et al.* 1997). Second, the Ty3/*gypsy*-like retrotransposons are specific to centromeres and are present in every centromere, while the transposable elements identified in *Drosophila* are neither unique to the centromeres nor present in all centromeres (Sun *et al.* 1997). Third, the centromere-specific Ty3/*gypsy*-like retrotransposons are remarkably conserved in the centromeres of distantly related plant species. The grass species diverged from a common ancestor about 60 to 100 mya (Martin *et al.* 1989; Wolfe *et al.* 1989), and there are no reports of repetitive DNA elements conserved in specific chromosomal regions among all the grass species, except the telomeric DNA sequences. We demonstrated that the centromere-specific Ty3/*gypsy*-like retrotransposons are also present in the centromeres of supernumerary B chromosomes from rye and maize (Figure 4D for pHind22; for pSau3A9 see Jiang *et al.* 1996b). These special characteristics of the centromere-specific retrotransposons in grasses led to a speculation that these sequences might be part of the functional centromeres (Jiang *et al.* 1996b). It will be a major challenge to test whether such sequences have any direct roles in centromere function.

We used a DNA sequence located in the Tf2 element of *S. pombe*, which has sequence similarity to pHind22, as a query against the GenBank databases. This sequence was found to have 76% identity to 165 nucleotides located in the central core sequence of centromere 2 in *S. pombe* (J. T. Miller and J. Jiang, unpublished observation). The central core sequences and its flanking repeat K are the critical parts of the functional centromeres of *S. pombe* chromosomes (Baum *et al.* 1994). The present sequence comparison shows that part of the central core sequence in *S. pombe* may also be derived from a Ty3/*gypsy*-like retrotransposon.

We thank Dr. R. A. Wing of Clemson University and Dr. P. C. Ronald of the University of California-Davis for providing the sorghum and rice BAC clones and Dr. J. P. Mascarenhas of the University of Albany for the LTR probe of the maize PREM-2 element. We are grateful to Drs. S. R. Wessler, M. J. Havey, and T. C. Osborn for critical reading of the manuscript. This research is supported by Hatch Funds (142-3935, 142-D395) and Funds 135-0534 and 135-0528 from the Graduate School of the University of Wisconsin-Madison to J.J.

LITERATURE CITED

- Aledo, R., R. Raz, A. Monfort, C. M. Vicent, P. Puigdomènech *et al.*, 1995 Chromosome localization and characterization of a family of long interspersed repetitive DNA elements from the genus *Zea*. *Theor. Appl. Genet.* **90**: 1094-1100.
- Baum, M., V. K. Ngan and L. Clarke, 1994 The centromeric K-type repeat and the central core are together sufficient to establish a functional *Schizosaccharomyces pombe* centromere. *Mol. Biol. Cell* **5**: 747-761.
- Bennetzen, J. L., 1996 The contribution of retroelements to plant genome organization, function and evolution. *Trends Microbiol.* **4**: 347-353.

- Bennetzen, J. L., and E. K. Kellogg, 1997 Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**: 1509–1514.
- Brandes, A., J. S. Heslop-Harrison, A. Kamm, S. Kubis, R. L. Dou-drick *et al.*, 1997 Comparative analysis of the chromosomal and genomic organization of Ty1- *copia*-like retrotransposons in pteridophytes, gymnosperms and angiosperms. *Plant Mol. Biol.* **33**: 11–21.
- Charlesworth, B., P. Jarne and S. Assimakopoulos, 1994 The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. III. Element abundances in heterochromatin. *Genet. Res.* **64**: 183–197.
- Clarke, L., 1990 Centromeres of the budding and fission yeasts. *Trends Genet.* **6**: 150–154.
- Csank, A. K., and J. F. McDonald, 1995 Analysis of *copia* sequence variation within and between *Drosophila* species. *Mol. Biol. Evol.* **12**: 83–93.
- Daraselia, N. D., S. Tarchevskaya and J. O. Narita, 1996 The promoter for tomato 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase has unusual regulatory elements that direct high level expression. *Plant Physiol.* **112**: 727–733.
- Dong, F., J. T. Miller, S. A. Jackson, G.-L. Wang, P. C. Ronald *et al.*, 1998 Rice (*Oryza sativa*) centromeric regions consist of complex DNA. *Proc. Natl. Acad. Sci. USA* **95**: 8135–8140.
- Eichler, E. E., F. Fu, Y. Shen, R. Antonacci, V. Jurecic *et al.*, 1996 Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.* **5**: 899–912.
- Eichler, E. E., M. L. Budarf, M. Rocchi, L. L. Deaven, N. A. Doggett *et al.*, 1997 Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.* **6**: 991–1002.
- Fuerstenberg, S. I., and M. A. Johns, 1990 Distribution of *BsI* retrotransposons in *Zea* and related genera. *Theor. Appl. Genet.* **80**: 680–686.
- Jiang, J., S. H. Hulbert, B. S. Gill and D. C. Ward, 1996a Interphase fluorescence *in situ* hybridization mapping: a physical mapping strategy for plant species with large complex genome. *Mol. Gen. Genet.* **252**: 497–502.
- Jiang, J., S. Nasuda, F. Dong, C. W. Scherrer, S.-S. Woo *et al.*, 1996b A conserved repetitive DNA element located in the centromeres of cereal chromosomes. *Proc. Natl. Acad. Sci. USA* **93**: 14210–14213.
- Joseph, J. L., J. W. Sentry and D. R. Smyth, 1990 Interspecies distribution of abundant DNA sequences in *Lilium*. *J. Mol. Evol.* **30**: 146–154.
- Kaszas, E., and J. A. Birchler, 1996 Misdivision analysis of centromere structure in maize. *EMBO J.* **15**: 5246–5255.
- Kipling, D., and P. E. Warburton, 1997 Centromeres, CENP-B and *Tigger* too. *Trends Genet.* **13**: 141–145.
- Kipling, D., A. R. Mitchell, H. Masumoto, H. E. Wilson, L. Nicol *et al.*, 1995 CENP-B binds a novel centromeric sequence in the Asian mouse *Mus caroli*. *Mol. Cell. Biol.* **15**: 4009–4020.
- Martin, W., A. Gierl and H. Saedler, 1989 Molecular evidence for pre-Cretaceous angiosperm origins. *Nature* **339**: 46–48.
- Mason, J. M., and H. Biessmann, 1995 The unusual telomeres of *Drosophila*. *Trends Genet.* **11**: 58–62.
- Masumoto, H., H. Masukata, Y. Muro, N. Nozaki and T. Okazaki, 1989 A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *J. Cell Biol.* **109**: 1963–1973.
- Miller, J. T., S. A. Jackson, S. Nasuda, B. S. Gill, R. A. Wing *et al.*, 1998 Cloning and characterization of a centromere-specific repetitive DNA element from *Sorghum bicolor*. *Theor. Appl. Genet.* **96**: 832–839.
- Moore, G., W. Cheung, T. Schwarzacher and R. Flavell, 1991 BIS 1, a major component of the cereal genome and a tool for studying genomic organization. *Genomics* **10**: 469–476.
- Mount, S. M., and G. M. Rubin, 1985 Complete nucleotide sequence of the *Drosophila* transposable element *copia*: homology between *copia* and retroviral proteins. *Mol. Cell. Biol.* **5**: 1630–1638.
- Murphy, T. D., and G. H. Karpen, 1995 Localization of centromere function in a *Drosophila* minichromosome. *Cell* **82**: 599–609.
- O'Neill, R. J. W., M. J. O'Neill and J. A. M. Graves, 1998 Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**: 68–72.
- Pélissier, T., S. Tutois, S. Tourmente, J. M. Deragon and G. Picard, 1996 DNA regions flanking the major *Arabidopsis thaliana* satellite are principally enriched in *Athila* retroelement sequences. *Genetica* **97**: 141–151.
- Regnier, V., M. Meddeb, G. Lecointre, F. Richard, A. Duverger *et al.*, 1997 Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum. Mol. Genet.* **6**: 9–16.
- Round, E. K., S. K. Flowers and E. J. Richards, 1997 *Arabidopsis thaliana* centromeres regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7**: 1045–1053.
- SanMiguel, P., A. Tikhonov, Y. Jin, N. Motchoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**: 765–768.
- Sun, X., J. Wahlstrom and G. Karpen, 1997 Molecular structure of a functional *Drosophila* centromere. *Cell* **91**: 1007–1019.
- Turich, M. P., A. Bokhari-Riza, D. A. Hamilton, C. He, W. Messier *et al.*, 1996 PREM-2, a *copia*-type retroelement in maize is expressed preferentially in early microspores. *Sex. Plant Reprod.* **9**: 65–74.
- Wang, G.-L., T. E. Holsten, W.-Y. Song, H.-P. Wang and P. C. Ronald, 1995 Construction of a rice bacterial artificial chromosome library and identification of clones linked to *Xa21* disease resistance locus. *Plant J.* **7**: 525–533.
- Weaver, D. C., G. V. Shpakavski, E. Caputo, H. L. Levin and J. D. Boeke, 1993 Sequence analysis of closely related retrotransposon families from fission yeast. *Gene* **131**: 135–139.
- Wessler, S. R., T. E. Bureau and S. E. White, 1995 LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5**: 814–821.
- Willard, H. F., 1990 Centromeres of mammalian chromosomes. *Trends Genet.* **6**: 410–416.
- Wolfe, K. H., M. Gouy, Y.-W. Yang, P. M. Sharp and W.-H. Li, 1989 Date of monocot-dicot divergence estimated from chloroplast sequence data. *Proc. Natl. Acad. Sci. USA* **86**: 6201–6205.
- Xiong, Y., and T. H. Eichbush, 1990 Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**: 3353–3363.

Communicating editor: J. A. Birchler