Genome-Wide Analysis of *mariner*-Like Transposable Elements in Rice Reveals Complex Relationships With *Stowaway* Miniature Inverted Repeat Transposable Elements (MITEs)

Cédric Feschotte,[1] Lakshmi Swamy[2] and Susan R. Wessler

Departments of Plant Biology and Genetics, The University of Georgia, Athens, Georgia 30602

Manuscript received September 26, 2002
Accepted for publication November 11, 2002

ABSTRACT

Stowaway is a superfamily of miniature inverted repeat transposable elements (MITEs) that is widespread and abundant in plant genomes. Like other MITEs, however, its origin and mode of amplification are poorly understood. Several lines of evidence point to plant *mariner*-like elements (MLEs) as the autonomous partners of the nonautonomous *Stowaway* MITEs. To better understand this relationship, we have taken advantage of the nearly complete genome sequences of two rice subspecies to generate the first inventory of virtually all MLEs and *Stowaway* families coexisting in a single plant species. Thirty-four different MLEs were found to group into three major clades and 25 families. More than 22,000 *Stowaway* MITEs were identified and classified into 36 families. On the basis of detailed sequence comparisons, MLEs were confirmed to be the best candidate autonomous elements for *Stowaway* MITEs. Surprisingly, however, sequence similarity between MLE and *Stowaway* families was restricted to the terminal inverted repeats (TIRs) and, in a few cases, to adjacent subterminal sequences. These data suggest a model whereby most of the *Stowaway* MITEs in rice were cross-mobilized by MLE transposases encoded by distantly related elements.

Tc1/*mariner* is a diverse and widespread superfamily of eukaryotic class 2 transposable elements (reviewed in Capy *et al.* 1998; Plasterk *et al.* 1999; Plasterk and van Luenen 2002). One hallmark of the superfamily is insertion into the dinucleotide TA that is duplicated upon insertion and flanks the element as a target site duplication (TSD). Tc1/*mariner* elements are relatively short (1.2–3.5 kb) and are simple in structure with terminal inverted repeats (TIRs) and a single gene encoding the transposase. A common model for the transposition mechanism of Tc1/*mariner* elements has emerged from the functional study of a limited number of animal transposases (Plasterk and van Luenen 2002). The N-terminal region of Tc1/*mariner* transposases contains DNA-binding domain(s) that bind specifically to the TIRs (Plasterk *et al.* 1999; Lampe *et al.* 2001; Zhang *et al.* 2001). A C-terminal domain is characterized by an amino acid signature called the DDE/D motif consisting of two aspartic acid residues and a glutamic acid residue (or a third D). This motif is required for catalysis of both the DNA cleavage and the strand transfer steps of the "cut and paste" transposition reaction (reviewed in Hartl *et al.* 1997; Plasterk and van Luenen 2002).

[1]*Corresponding author:* Department of Plant Biology, University of Georgia, Athens, GA 30602.
E-mail: cedric@dogwood.botany.uga.edu

[2]*Present address:* Tri-Institutional MD/PhD Program, Weill Medical College, Cornell University, New York, NY 10021.

Tc1/*mariner* elements were recently found to be widespread in plants (reviewed in Feschotte *et al.* 2002a). The first reported plant members were *Soymar1*, a *mariner*-like element (MLE) from soybean (Jarvik and Lark 1998) and *Lemi1*, a *pogo*-like element from *Arabidopsis thaliana* (Feschotte and Mouchès 2000). Three additional rice MLEs were subsequently identified by database searches, but none were characterized further (Tarchini *et al.* 2000; Shao and Tu 2001; Turcotte *et al.* 2001; Feschotte and Wessler 2002). These five elements were used to derive plant-specific primers that successfully amplified MLE transposase genes in PCR assays with DNA from a wide spectrum of flowering plant genomes (Feschotte and Wessler 2002). For the majority of genomes assayed, multiple divergent lineages of transposases were amplified from single species.

Demonstration that MLEs are widespread and diverse in plants provided support for the hypothesis that MLEs are the autonomous elements responsible for the origin and spread of *Stowaway*, a large group of miniature inverted repeat transposable elements (MITEs; Bureau and Wessler 1994). MITEs are structurally reminiscent of class 2 nonautonomous elements with their small size (<600 bp), lack of coding capacity, and TIRs (reviewed in Feschotte *et al.* 2002b). However, their high copy number and structural homogeneity have served to distinguish them from most of the previously described class 2 elements (Wessler *et al.* 1995). MITEs were first discovered in plants, where they are now recognized as the predominant type of transposable element associ-

ated with the noncoding regions of plant genes. This is particularly evident in the cereals, including rice, maize, barley, and wheat (Bennetzen 2000; Feschotte *et al.* 2002a; Goff *et al.* 2002; Yu *et al.* 2002). Vast amounts of MITEs have also been discovered in many invertebrate and vertebrate genomes (reviewed in Feschotte *et al.* 2002b).

Most of the tens of thousands of MITEs in plant genomes have been divided into two groups on the basis of the similarity of their TIRs and TSDs: *Tourist*-like MITEs and *Stowaway*-like MITEs (Wessler *et al.* 1995; Feschotte *et al.* 2002b). That *Stowaway*-like MITEs and plant MLEs share similar terminal sequences (5′-CTC CCTCCRT-3′, where R stands for A or G) and target site preference (TA) strongly suggested that *Stowaway* MITEs were mobilized *in trans* by transposases encoded by MLEs (Turcotte *et al.* 2001; Feschotte *et al.* 2002b). A model was formulated that hypothesized that *Stowaway* elements originated by internal deletion(s) from a larger autonomous element (like previously described nonautonomous DNA elements) and were amplified to very high copy number by the transposase encoded by the autonomous element (Feschotte *et al.* 2002b). The diversity of *Stowaway* families observed in a single genome was explained by proposing that the families originated as deletion derivatives of distinct lineages of MLEs (Feschotte *et al.* 2002a,b). If this model is correct, one should encounter *Stowaway* families that have extensive sequence similarity (*i.e.*, not just in their termini) with MLEs present in the same genome. In addition, the diversity of *Stowaway* families should correspond with a similar diversity of MLEs in that same genome. Failure to match *Stowaway* families with MLEs would indicate that the model was incorrect or overly simplistic.

Comparison of all of the MLEs and *Stowaway* elements in a genome is possible only for Arabidopsis and rice for which entire genome sequences are available. Although remnants of MLE transposases are still recognizable in the sequence of *A. thaliana*, no full-length MLEs are identifiable (Shao and Tu 2001; Feschotte and Wessler 2002; C. Feschotte, unpublished data). Furthermore, *Stowaway* MITEs are relatively scarce in this species (at least in the sequenced ecotype), with <250 copies organized into fewer than five families (Le *et al.* 2000; C. Feschotte, unpublished data). In contrast, previous searches of a limited amount of rice genomic sequence identified numerous families of *Stowaway* MITEs and full-length MLEs (Bureau *et al.* 1996; Jiang and Wessler 2001; Shao and Tu 2001; Turcotte *et al.* 2001; Feschotte and Wessler 2002). For these reasons, the goal of this study was to characterize all MLE and *Stowaway* families in rice and determine the extent of sequence relatedness between these two groups.

A semiautomated computational approach was used to identify and compare MLEs and *Stowaway* MITEs in the two draft genome sequences of rice (Goff *et al.* 2002; Yu *et al.* 2002). In this way 34 MLEs were identified, with 22 considered full-length, as they contain a complete transposase coding region, TIRs, and TSD. Phylogenetic analysis and other criteria, such as the presence or absence of introns, led to their grouping into 25 distinct families falling into three major clades. In addition, up to 33,000 *Stowaway* MITEs were identified, with the high-copy-number elements grouping into at least 36 families. Surprisingly, none of the 25 MLE families could be associated by simple internal deletion with any of the *Stowaway* families. Instead, sequence similarity between *Stowaway* and MLE families was restricted to the TIRs and, in a few cases, to some adjacent subterminal sequence. These data have led us to conclude that most of the *Stowaway* MITEs in rice were probably cross-mobilized by MLE transposases encoded by distantly related elements.

MATERIALS AND METHODS

Semiautomated mining of full-length rice MLEs: A series of Perl scripts was written to automate the process of identifying and fetching full-length elements related to a particular transposase. In a first step, the transposase amino acid sequence is used as a query in a local WU-TBLASTN search (http://blast.wustl.edu) against a genomic database. The output file is parsed and the significant hits (in this study, *E* values $<10^{-5}$) are extracted from the database along with up to 10 kb of flanking DNA sequence. In a second step, the flanking sequences are searched for the possible ends of the elements using a subroutine called MATCH-TIR. This program scans the 5′ and 3′ flanking regions of each hit with a 16-mer sliding window for the presence of a consensus motif corresponding to the 5′ and 3′ ends of the element plus the expected target site duplications (user input). MATCH-TIR extracts 5′ and 3′ hits (sequences with >80% similarity to the motif) along with 50 nucleotides internal to the hits and produces pairwise alignments between 5′ extended hits and the reverse complement of 3′ extended hits. The alignments are inspected visually and the best matching pairs (usually fewer than four mismatches in the first 22 nucleotides) are considered as the TIRs of the element. In this study, the *Osmar1* transposase sequence was used as the query in a WU-TBLASTN search against two databases. The first database contained ~360 Mb of bacterial artificial chromosome (BAC)/P1-derived artificial chromosome (PAC) sequences from *Oryza sativa* ssp. *japonica* cv. Nipponbare (downloadable at http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/seqcollab-assign.pl). The second database contained ~430 Mb of contigs generated by whole-genome sequencing of *O. sativa* ssp. *indica* cv. 9311 (downloadable at http://btn.genomics.org.cn/rice/). The motif 5′-TACTCCCTCCAG-3′ and its reverse complement were used for MATCH-TIR searches of the 5′ and 3′ ends of rice MLEs, respectively. Other searches were performed using WU-BLASTN against the two databases described above and a third database containing the whole-genome shotgun assembly of *O. sativa* ssp. *japonica* cv. Nipponbare produced by Syngenta (390 Mb; http://www.tmri.org).

Compilation of *Stowaway* families and copy number determinations: Twenty-four *Stowaway* families analyzed in this study were previously published (from Stow-Os1-Os32; Bureau and Wessler 1994; Bureau *et al.* 1996; Jiang and Wessler 2001; Turcotte *et al.* 2001). These 24 families were compared to a collection of rice repeats identified *de novo* by the program RECON in ~30 Mb of BAC/PAC sequences of *O. sativa* ssp. *japonica* (Bao and Eddy 2002). RECON identified and com-

puted consensus sequences for the 24 previously recognized Stowaway families and for four newly identified families (from *Stow-Os1-Os37*; see Figure 3). An additional family (*Stow-Os38*) was identified through BLASTN searches with the subterminal regions of *Osmar4*. Jerzy Jurka and A. Drazkiewicz (Repbase; http://www.girinst.org) contributed 7 additional families (*Stow-Os42-Os52*). Copy numbers of *Stowaway* families were estimated for 360 Mb of BAC/PAC sequences from *O. sativa japonica* by two different methods and extrapolated to a genome size of 430 Mb. In the first method, the database was analyzed with RepeatMasker using the compilation of *Stowaway* consensus described above. Crude values obtained from this search were refined by dividing the number of hits by two for queries with TIRs >45 bp (these sequences will produce systematically two hits per position, one from each strand). To correct for multiple hits due to interfamily similarity, we combined families predicted to cross-hit (high level of similarity in their TIRs) and considered the highest value obtained for these families as the copy number of the combined families. This method gave values corresponding to the upper estimate in the range shown in Figure 3. The lower estimate was obtained by counting the number of hits produced in BLASTN and FASTA searches using each consensus as a query against the same database with default parameters. Each BAC was counted as a hit if it contained a sequence matching >50% of the length of the query with at least 85% similarity. This method gave a more conservative estimate partly because BAC/PAC sequences containing multiple family members produce a single hit.

Sequence and phylogenetic analyses: Rice MLEs were conceptually translated in the six reading frames with MacVector (http://www.accelrys.com/products/macvector/). Transposase open reading frames (ORFs) were assembled by removing introns predicted with >85% confidence by NetGene2 (http://www.cbs.dtu.dk) and/or FGENESH (http://genomic.sanger.ac.uk/gf/gf.html). When necessary, frameshifts were judiciously introduced according to nucleotide alignments of closely related elements. Putative initiation codons were predicted by NetStart (http://www.cbs.dtu.dk). The resulting transposase sequences were aligned with ClustalW using default parameters in MacVector 7.0. Phylogenetic trees were generated with PAUP* version 4.0b8 (http://paup.csit.fsu.edu/) using the neighbor-joining and maximum parsimony methods with default parameters and rooted with the distantly related *Soymar1* transposase from soybean. Sequence comparisons of *Osmar* and *Stowaway* elements were carried out using the LFASTA and BLAST2 servers available at http://www.infobiogen.fr.

RESULTS

Extracting MLEs from rice genomic sequence: Prior analyses of small fractions of the rice genome identified four MLEs and several *Stowaway* families (Bureau *et al.* 1996; Mao *et al.* 2000; Tarchini *et al.* 2000; Jiang and Wessler 2001; Shao and Tu 2001; Turcotte *et al.* 2001; Feschotte and Wessler 2002). The amount of rice sequences available in publicly accessible databases has increased dramatically since those studies and now encompasses nearly two complete genomes from two *O. sativa* subspecies: *japonica* (cv. Nipponbare) and *indica* (cv 93-11; see materials and methods for details of the databases used in this analysis). This vast resource has been exploited to identify, classify, and compare MLEs and *Stowaway* MITEs coexisting within the rice genome.

To compare MLE and *Stowaway* families, it was first necessary to obtain full-length MLEs including complete ORFs and flanking TIRs and TSDs. The strategy employed is detailed in materials and methods. The putative transposase sequence of *Osmar1* was used as the query in TBLASTN searches against two different databases. The first database contained ~360 Mb of BAC/PAC sequences generated from *O. sativa* ssp. *japonica* (cv. Nipponbare) by the International Rice Genome Sequencing Project (IRGSP). The second database was the draft genome sequence of *O. sativa* ssp. *indica* (cv. 9311) recently released by the Beijing Genomics Institute (BGI; ~420 Mb of shotgun sequence; Yu *et al.* 2002). After manual filtering of redundant hits, a total of 39 sequences with significant similarity to *Osmar1* transposase were identified (*E* values $<10^{-5}$). To define the ends of the corresponding MLEs, 5 kb flanking each hit was searched for TIRs similar to those of previously identified rice MLEs and *Stowaway* MITEs. Elements with perfect or near perfect TIRs of >20 bp and large ORF(s) encoding the putative transposase were extracted from the database along with 50 bp of flanking genomic sequence and used as queries in BLASTN searches against the IRGSP and BGI databases and the whole-genome shotgun assembly of *O. sativa japonica* cv. Nipponbare produced by Syngenta (Goff *et al.* 2002). These BLASTN searches enabled us to isolate incomplete and/or noncoding copies and determine whether elements isolated from *japonica* and *indica* were present at orthologous positions (for this study orthologous MLEs are considered as the same insertion event). A list of all identified MLEs, their accession numbers, and coordinates are available in a supplemental table (available at http://www.genetics.org/supplemental).

Twenty-two MLEs, ranging in size from 3167 to 11290 bp, were classified as full length because they contained ORF(s) corresponding to the transposase, had TIRs ranging from 20 to 36 bp (with fewer than four mismatches in most cases), and were flanked by a TA target site duplication (Figures 1 and 2). Searches with RepeatMasker and BLASTN revealed that other transposable elements had inserted into a few of the MLEs (Figure 1). For example, a 1795-bp *Mutator*-like element was found in *Osmar4* while *Osmar7* contained a 2708-bp insertion consisting of a *Tourist*-like MITE nested into a solo LTR from the retrotransposon *RIRE1*. By excluding secondary insertions in size determinations, full-length MLEs ranged from 3167 to 7072 bp. Ten additional MLEs appear to contain a full-length transposase gene and a substantial amount of subterminal sequence (see Figure 1). However, these elements were missing one or both termini due to either secondary mutations or rearrangements after insertion (such as large deletions or insertions) or gaps in the whole-genome sequence assembly from the BGI.

Phylogenetic analysis and classification of *Osmar* elements: As a first level of classification, MLEs were

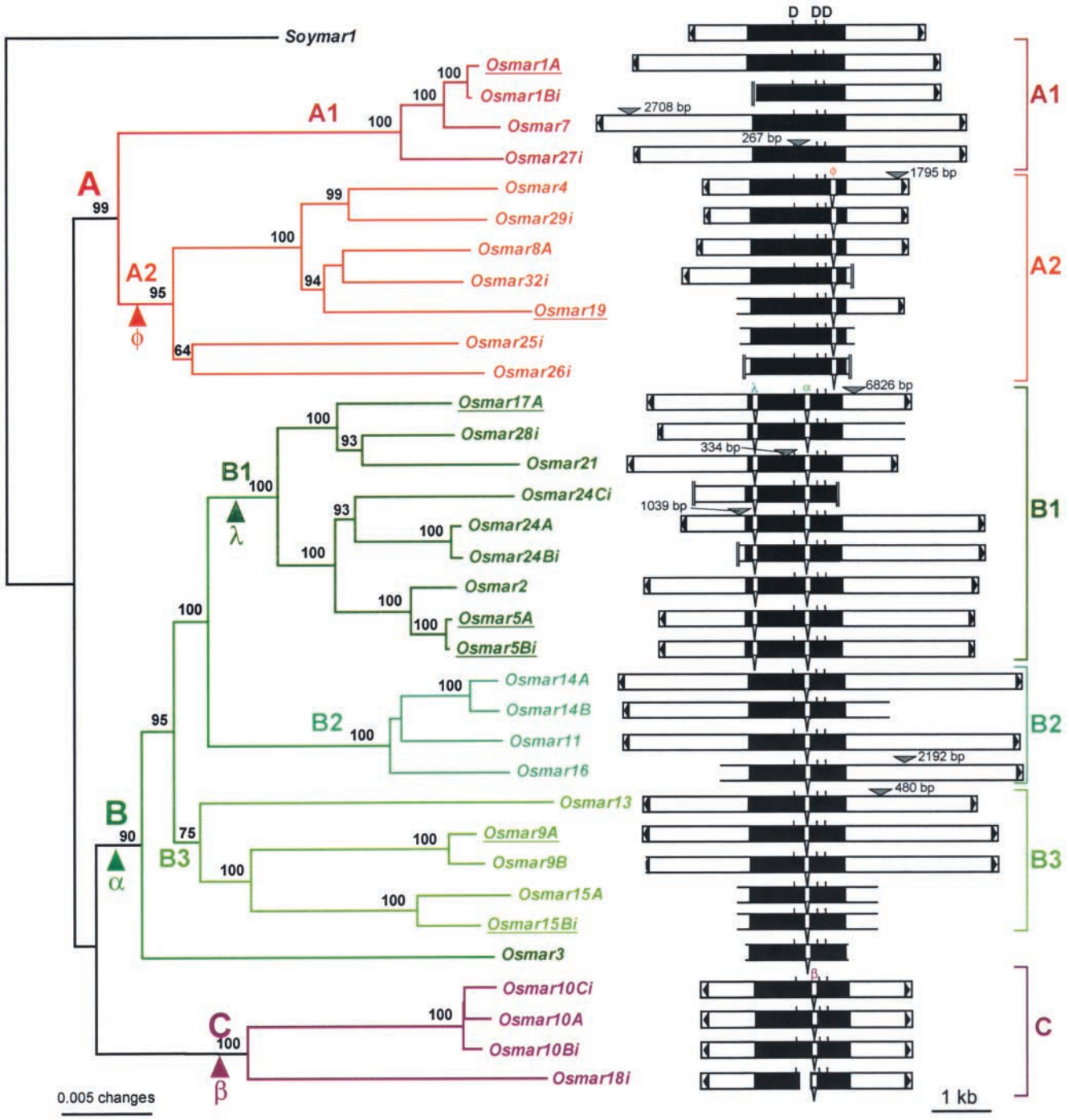

FIGURE 1.—Phylogenetic relationships and structures of rice MLEs. The neighbor-joining tree was generated from a multiple alignment of conceptually translated transposase sequences of 34 rice MLEs (*Osmar*s; see supplemental table for accession numbers at http://www.genetics.org/supplemental) and *Soymar1* from soybean (GenBank accession no. AF078934), which served as the outgroup. Bootstrap values >60 are shown as a percentage of 1000 replicates. Underlined names denote elements with potentially intact transposase genes. All *Osmar*s are from the subspecies *japonica*, except those followed by *i*, which are from the subspecies *indica*. Capital letters and different colors emphasize different lineages and sublineages of *Osmar*s. Arrowheads indicate the presence of a particular intron prior to the divergence of a lineage or a sublineage. The structure of the corresponding MLE is depicted on the right. Full-length elements are delimited by terminal inverted repeats (solid triangles). Other elements are incomplete due to secondary mutations or to an interruption in the *indica* contig sequence (this latter situation is shown by a double vertical bar). Transposase coding sequences are depicted as solid boxes and the position of the DD39D triad is shown. *Soymar1* and *Osmar*s of the sublineage A1 harbor an intronless transposase gene while other *Osmar*s are predicted to contain one or two introns (shown as an open triangle below the element). Introns occur at four different positions (α, β, ϕ, and λ), which are specific for a lineage or sublineage of transposase. Insertions of other repeats in *Osmar*s are shown as shaded triangles above the element along with the insertion size.

clade	element	size (bp)	TIR (bp)	TIR sequence
Clade A1	*Osmar1*	5259	26	5' CTCCCTCCGTTTCGTTTTGTTTGTCG
				3' CTCCCTCCGTTTCGTTTTGTTTGTCG
	Osmar7	6654*	26	5' TTCCCTCCGTTTCGTTTTATTTGTCG
				3' TACCCTCCGTTTCGTTTTATTTGACG
	Osmar27i	5745*	26	5' CTCCCTCCGTCCCGTTATGTTTGACG
				3' CTCCCTCCGTTCTGTTATGTTTGACG
				TTCG
Clade A2	*Osmar4*	3228*	29	5' CTCCCTCCATACTCATATTAGAAGTCGTT
				3' CTTCCTTCATACTCATATTAAAAGTCGTT
	Osmar29i	3167	29	5' CTCCCTCCATACCCACAAAACAAGTCGTT
				3' CTCCCTCCATACCCACAAAACAAGTCGTT
	Osmar8A	3387	29	5' CTCCCTCCATACTGGTAAAAAATGACGTT
				3' CTCCCTCCCTATAGAAAAAAAACGTCGTT
	Osmar32i	>3034	29	5' CTCCCTCCGTACTCGTAAAACATGTCGTT
	Osmar19	>2427	29	3' CTCCCTCCGTATCACAATTAGAAGATGTT
				ACTC
Clade B	*Osmar11*	6457	23	5′ CTCCCTCCGTCCCATAATATCTC
				3′ CTCCCTCCGTCCCACAATATCTC
	Osmar14A	7072	32	3′ CTCCCTCCGTCCCAGAAAGAAGGGATTCCTGG
				5' CTCCCTCCGTCCCAGAAAGAAGCGATTTCTGG
	Osmar14B	>4370	32	5' CTCCCTCCGTCCCAGAAAGAAGGGATTCCTGG
	Osmar14C	>930	31	5' CTCCCTCCGTCCCAGGAAGAAGGGATT-CTGG
	Osmar16	>5127	32	3' CTCCCTCCGTCCCAAAAAGAGACGAGTTCTGG
	Osmar5A	5195	29	5' CTCCCTCCGTCCCACAAAACATGACGTTT
				3' CTCCCTCCGTCCCACAAAACCTGCCGTTT
	Osmar2	5565	31	5' CTTCCTCCGTCCCAGTAAACGTGACGTTTTA
				3' CTTCCTCCGTCCCAGTAAACGTGTCGTTCTA
	Osmar24	5977	29	5' CTTCCTCCGTCCCAGAAAATATGTCGTTT
				3' CTTCCTCCGTCCCAGAAAAGTTGATGTTT
	Osmar24Bi	>4444	29	3' CTCCCTCCGTCCCACAAAACCTGCCGTTT
	Osmar21	4908*	23	5' CTCCCTCTATCCCACTATAGTGG
				3' CTCCTTCCGTCCCACTTTAGTGG
	Osmar28i	>4982	30	5' CTCCCTCCGTCCCAGTATAGTGGATGTCCT
	Osmar17A	4465*	30	5' CTCCCTCCGTCCCAGAAAGGAGGACGTTCT
				3' CTCCCTCCGTCCCAGAAAGGAGGACGTTCT
	Osmar13	5732	36	5′ CTCCCTCCGTCCCAGAATATAACAACTTTTAGCCTT
				3′ CTCCCTCCGTCCCAGAATACAACGAGCTTTAGCCTT
	Osmar9A	5819	28	5′ CTCCCTCCGTCCCACATTATATGGGACT
				3′ CTCCCTCCGTCCCATATTATATGGGATT
	Osmar9B	6906	20	5' CTCCCTCCGTCCCACATTGTCGGAGAAT
				3' CTCCCTCCGTCCCATATTATATGGGATT
				CCCA
Clade C	*Osmar10A*	3266	27	5′ CTCCCTCCGTTCCTTAATATAGGGCGT
				3′ CTCCCTCCGTTCCTTAATATAGGGCGT
	Osmar18i	3182#	25	5′ CTCCCTCCGTTCCTTAATATAGGGC
				3′ CTCCCTCCGTTCCTTAATATAGGGC
				TCCT

FIGURE 2.—Classification of *Osmar*s based on the TIR. *Osmar*s are classified into four clades (A1, A2, B, and C) on the basis of the sequence of a 4-bp motif in their TIR (boxed). This division is supported by the phylogenetic analysis (see Figure 1). The diagnostic consensus motifs are shown as white letters on a black background. An asterisk indicates that the size of *Osmar* was calculated after removal of nested TE insertions (see Figure 1). The sign > is used for the size of incomplete *Osmar*s, where only one terminus could be identified. The open triangle above the 5′ TIR of *Osmar8A* and the 3′ TIR of *Osmar27i* denotes an insertion of 5 and 2 bp, respectively, removed from the TIR sequence.

grouped into the same family when they shared >85% similarity over their entire length. Using these criteria, 25 different families of MLEs were distinguished (Figures 1 and 2). Consistent with the nomenclature introduced in animals, rice MLEs were designated *Osmar* (for ***O**. **s**ativa **mar**iner*) followed by the number of the family. Members of the same family were further designated by capital letters (for example, *Osmar1A* and *Osmar1B*; see Figure 1).

A phylogenetic analysis of transposase sequences was carried out to resolve evolutionary relationships among rice MLE families. Conceptual translation and multiple alignments of 34 *Osmar* transposases revealed that most (27/34) are corrupted by substitutions and small insertions/deletions (indels) that introduced premature stop codons in the protein sequence. However, several *Osmar*s had intact ORFs and may encode active transposase (names underlined in Figure 1). After removal of predicted introns (see below) and, where necessary, introduction of frameshifts to restored ORFs, putative full-length *Osmar* transposases were found to range in size from 432 to 505 residues with pairwise amino acid identities that varied from 36% (*Osmar13 vs. Osmar26i*) to 99% (*Osmar5A vs. Osmar5Bi*). The most conserved region is a central domain of ~150 residues that is roughly delimited by the DD39D motif (see multiple alignment provided as supplemental data at http://www.genetics.org/supplemental). This motif is characteristic of plant MLE transposases (SHAO and TU 2001; FESCHOTTE and WESSLER 2002) and is found intact in 30 out of 34 putative *Osmar* proteins. Phylogenetic trees were generated from a multiple alignment of 34 *Osmar* transposases and the *Soymar1* transposase using the neighbor-joining and maximum parsimony methods. Both methods produced trees with very similar topology that defined three major clades of *Osmar* transposase (A, B, and C in Figure 1). Clade A (10 families, 11 sequences) and clade B (13 families, 19 sequences) were more abundant and diverse than clade C (2 families, 4 sequences) and can be further divided into subclades with strong bootstrap values (A1, A2, etc.; see Figure 1).

An analysis of the positions of predicted introns in

Osmar transposase genes provides additional support for the phylogenetic groupings. Four different introns (called α, β, ϕ, and λ) were associated with *Osmar* transposases and result in genes with zero, one, or two introns. When the distribution of these introns was superimposed on the transposase phylogeny, each type of intron was found to be specific to a clade or to a subclade of transposases (Figure 1). That is, intron-α was restricted to clade B, intron-β to clade C, intron-ϕ to subclade A2, and intron-λ to subclade B1.

The phylogenetic organization of *Osmar*s is also supported by a comparison of their TIRs. For all *Osmar* elements, the first 10 bp of the TIRs are well conserved and match the consensus 5′-CTCCCTCCRT-3′ (Figure 2). Adjacent to this motif is a 4-bp sequence that serves to define a subset of *Osmar*s. There is a striking correspondence between these groupings and those defined by the phylogenetic groupings of transposases (compare groups in Figure 2 and phylogeny in Figure 1). Indeed, all *Osmar*s in subclade A1 have a TTCG motif in their TIRs while *Osmar*s in subclade A2 display a consensus ACTC motif. *Osmar*s clustered in clade B are characterized by a CCCA motif and those falling in clade C are characterized by TCCT. That each motif is diagnostic of an *Osmar* transposase clade (or subclade) suggests coevolution between transposase and TIR sequences.

Classification of *Stowaway* MITEs and sequence relationship with *Osmar*s: Although numerous *Stowaway* families were previously identified in rice, analysis of repeats was limited to a small fraction of the genomic sequence (<50 Mb; Mao *et al.* 2000; Tarchini *et al.* 2000; Jiang and Wessler 2001; Turcotte *et al.* 2001). For this reason, a more comprehensive search for *Stowaway* MITEs was undertaken in ~360 Mb of BAC/PAC sequence from the IRGSP.

Searches were carried out with BLASTN and RepeatMasker using a collection of previously characterized *Stowaway* elements (Jiang and Wessler 2001; Repbase Update, http://www.girinst.org) in addition to elements identified *de novo* by RECON (Bao and Eddy 2002; N. Jiang, Z. Bao, S. Eddy and S. R. Wessler, unpublished data). Depending on the stringency of these searches (see materials and methods), a total of 22,000–33,000 *Stowaway* elements are estimated to populate the Nipponbare genome. Sequence comparisons led to the grouping of most of these elements into 36 high-copy-number families (Figure 3). As with *Osmar* families, members of the same *Stowaway* family share at least 85% similarity over their entire length. *Stowaway* families are represented by consensus sequences that range in size from 96 to 312 bp with TIRs of 21 to 94 bp (Figure 3; consensus sequences were deposited in Repbase Update, http://www.girinst.org).

With the *Osmar* and *Stowaway* elements organized into families, it was of interest to determine whether any correspondence existed that would indicate a clear-cut relationship between autonomous (*Osmar*) and nonautonomous (*Stowaway*) elements. Two complementary approaches were used to compare the sequences in the terminal regions of *Osmar*s with *Stowaway* families. First, each rice MLE was used as a query in BLASTN searches against the three rice genomic databases (IRGSP, BGI, and Syngenta). These searches revealed that *Osmar*s were associated with few, if any, deletion derivatives (see supplemental table at http://www.genetics.org/supplemental and Figure 4). Furthermore, these deleted copies were heterogeneous in size (Figure 4) and were usually larger than *Stowaway* elements (280 bp–~2 kb *vs.* 94–350 bp for *Stowaway* consensus). Although a few MLE families, such as *Osmar10*, include a small homogeneous group of short deletion derivatives (Figure 4), none of the derivatives have attained the high copy number that is a hallmark of MITE families.

In a second approach, each *Stowaway* consensus was used as a query in BLASTN searches against a database containing all full-length rice MLEs. These searches revealed that when significant sequence similarity existed, it was restricted to the terminal nucleotides (usually <50 bp; see example of *Osmar1* and *Stow-Os6*, Figure 5). The most extensive matches were found between *Osmar4* and *Stow-Os10b*, *Osmar13* and *Stow-Os16*, and *Osmar11* and *Stow-Os49* (Figure 5).

Most of the *Stowaway* families can be assigned to one of the four major *Osmar* clades on the basis of similarities in their TIRs. In fact, 34 out of 36 *Stowaway* consensus sequences display one of the four TIR motifs diagnostic of *Osmar* clades (compare Figures 2 and 3). For example, group B of *Stowaway* and *Osmar* are characterized by the same CCCA motif in the TIRs. Guided by these groupings, we generated consensus TIR sequences for each major clade of *Osmar* and *Stowaway* in the form of pictograms (Figure 6). Comparison of the pictograms further revealed the similarities in the TIRs of corresponding clades of *Osmar* and *Stowaway*.

DISCUSSION

Several lines of evidence point to plant MLEs as the autonomous partners of *Stowaway* MITEs. In this study, we have taken the next step in testing this hypothesis by generating an inventory of virtually all MLEs and *Stowaway* families coexisting in a single genome and analyzing in detail their sequence relationships.

The first whole-genome picture of plant MLEs: A total of 39 MLE transposases and 22 potentially full-length MLEs were identified from the genomes of the two rice subspecies. On the basis of the phylogenetic analysis of transposases, the intron/exon structure of the transposase gene, and a comparison of terminal sequences, rice MLEs could be divided into 25 families that group into three major clades (Figures 1 and 2). These clades correspond to the three lineages of MLE transposases that were recently isolated by PCR using plant-specific MLE primers and genomic DNA from a wide range of plant species (Feschotte and Wessler 2002; results not shown). A conclusion of this prior study was that three

Clade	Family	Copy number	Size (bp)	TIR (bp)	Consensus TIR sequence
Clade A1	*Stow-Os1*	3,500-4,500	150	75	CTTCCTCCGTTTCACAATGTAAGTCATTCTAGyATTTCCCACATT…
	Stow-Os25		96	31	CTTCCTYCGTTTCACAATGCAAGACTTTCTAG
	Stow-Os50		170	85	CTTCCTCCGTTTCATATTATAAGACTTTCTAGCATTGCCCATATT…
	Stow-Os5	2,000-3,000	258	94	CTMCCTCCGTTTCATATTATAAGWCGTTTTGACTTTKGTCAAAGT…
	Stow-Os6		251	76	CTCCCTCCRTTTCATATTATAAGTCGTTTGActtTTTTWTTCTAG…
					TTCA
Clade A2	*Stow-Os32*	1,500-2000	269	21	CTCCCTCCRKRYTSATAATAC
	Stow-Os45		262	40	CTCCCTCCGGNYTGATAATACTTGTCGTTTTRGACAAGGG
	Stow-Os10		259	37	CTCCCTCCGTAYTYATAAWAAWWGTCGTTTTGGACAA
	Stow-Os38		270	36	CTCCCTCCGTACTCGTAAAGGAAGTCGTTTAGGACA
					AYTC
Clade B	*Stow-Os8*	1,200-2000	257	32	CTCCCTCCGTCCCAAAATATAAGYATTTTTAG
	Stow-Os23		100	48	CTMCCTCCGTCCCAAAATATAASMAYTTTTRGCTATGAATCTRGA…
	Stow-Os18	500-700	216	30	CTWCCTCCGTCYCAAAATRTARSTATTTYT
	Stow-Os24	1,000-2,000	246	33	CTMCCTCYGTCCCAAAATAATTGTATTTCTAGG
	Stow-Os30	2,000-3,500	149	69	CTCCCTCCGTCYCAAAATATAAGARATTTTRAYGRGATGTGAYAT…
	Stow-Os15		150	65	CTCCCTCCGTCYCAWAATATAASAACCTATGTACTGGATGTATGT…
	Stow-Os26		113	49	CTCCCTCCGTCCCAAAATATAGCAACCTAGAAYYGGATGGGACAT…
	Stow-Os46		150	70	CTCCCTCCGTCCCAAAATATAAGGGATTTTGGATGGATGTGACAT…
	Stow-Os21	3,000-4,500	240	78	CTCCCTCCGTCYCAAAATATWTGACGCTRTTRAYTTTTCTRTTTW…
	Stow-Os3		240	73	CTCCCTCCGTYTCWAAATRTTTGACRCCGTTGACTTTTTACTAAA…
	Stow-Os13	500-1,000	250	34	CTCCCTCCATCCYAAAATATAAGGCAYAACCACT
	Stow-Os28	500-1,000	257	24	CTCCCTCYATCCCATAATATAAGG
	Stow-Os34	400-700	256	21	CTCCCTCCGTCCCAAATTATA
	Stow-Os9	2,500-3,000	239	66	CTCCCTCCGTCCMAAAAAAAATAGWCAAACCTYTGGTTTCCGTGT…
	Stow-Os52		236	79	CTCCCTCCGTCCCATTTTAAGTGCAACCATGAGTTTYCGTGTCCA…
	Stow-Os51		242	27	CTMCCTCCGTCCCAAAATAAGTGYAG
	Stow-Os36	250-400	268	30	CTCCCTCCTTCCCAAATTGATCATCATATA
	Stow-Os37	150-200	312	24	CTCCCTCCATCYCARTTTAATCAT
	Stow-Os16	1,200-2,000	225	28	CTCCCTCCGTCCCAAAATATAASAAYTTTTAG
	Stow-Os35	200-400	241	33	CTCCCTCCGTCYCCAAAAAAAMTCMAYTYCTAG
	Stow-Os49		237	25	CTCCCTCCGTCCCAMAAWAAATCAA
	Stow-Os20	100-150	277	30	CTCCCTCCGTCCCATAAAAATTGAATTTCT
	Stow-Os42	200-250	238	30	CTMCCTCCGTYCCAWAATAAGTTTATTTTT
					CCCA
Clade C	*Stow-Os12*	350-500	312	29	CTCCCTCCATTCCAAATTGATCTACATAT
	Stow-Os29	50-100	244	33	CTCMCTCTGTTCYYAAATATAAGCATTTCTAGG
					TCCH
Unclassified	*Stow-Os14*	250-350	259	30	CTCCCTCCATCCACAAAAGTTAKACATATT
	Stow-Os11	600-800	292	31	CTYCCTCCATCTAYTTTTGATAGTCATATTT
Total		21950-33050			

FIGURE 3.—Classification of *Stowaway*s based on the TIR. A consensus sequence was derived for each high-copy-number *Stowaway* family. Sources and methods of collecting these sequences are detailed in MATERIALS AND METHODS. *Stowaway* families were classified into four major clades on the basis of the same 4-bp motifs used to classify *Osmar*s (see Figure 2). Copy numbers were estimated using two methods, producing a lower and upper estimate for each family or a group of closely related families (see MATERIALS AND METHODS).

major lineages of MLE transposases had diversified prior to the divergence of the Poaceae family (~70 MYA) and have been maintained in the genomes of most extant grass species (FESCHOTTE and WESSLER 2002). The fact that the three MLE clades identified in rice correspond to these three major lineages indicates that no other, more divergent, lineages are in rice. Thus, the rice genome is representative of the diversity of MLE transposases in the grasses and as such should serve as a suitable model for the evolutionary analysis of plant MLEs.

Despite the ancient origin of the three MLE lineages, all appear to include families recently active in rice. This is reflected by the high level of sequence similarity among members of several *Osmar* families and the presence of copies with intact coding capacity (see Figure 1 and examples in Figure 4). On the basis of these criteria, one of the most recently active MLE family is *Osmar5*: the three full-length members are >99.5% identical to each other and harbor intact transposase ORFs. It is therefore possible that one or more active MLEs may still reside in the rice genome.

Full-length *Osmar*s are heterogeneous in size, ranging from 3.2 to 7.1 kb, and there is also extensive size varia-

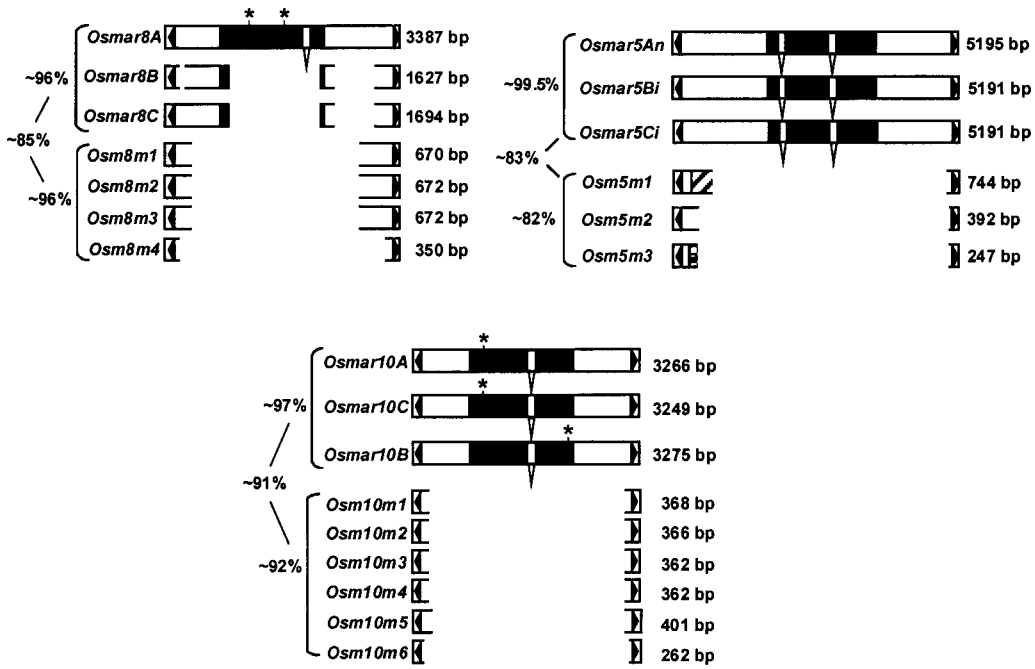

FIGURE 4.—Structure of some *Osmar* families. Shown is the structure of *Osmar* families representative of the three major clades of rice MLEs (clade A, *Osmar8*; clade B, *Osmar5*; and clade C, *Osmar10*). Each family includes members of variable size that either contain or do not contain transposase sequences (solid boxes). Average percentages of pairwise similarity within and between the two subsets of family members are shown. Hatched and dotted fragments in *Osm5m1* and *Osm5m2* represent portions that are unrelated to each other and to other *Osmar5* sequences. Otherwise, smaller members resemble internal deletion derivatives of the larger members. Asterisks indicate the positions of premature stop codons caused by nucleotide substitution or small indels in the transposase sequence.

tion within *Osmar* families (see Figures 1, 2, and 4). Full-length *Osmar*s harbor a single gene corresponding to the putative transposase, which generally occupies a central position in the element (but see Figure 2 for the few exceptions) and has a similar size among rice MLEs. Thus, most of the size variation among *Osmar*s is due to the variable length of the subterminal regions. These regions do not display any obvious structural features, such as direct or inverted motifs, like those of some other plant DNA transposons, including hAT or CACTA superfamily members (KUNZE and WEIL 2002).

In contrast to rice MLEs, there is a remarkable conservation in the size of full-length MLEs described from a wide range of metazoan species. The dozens of elements described from species as diverse as planarians, hydra, nematodes, insects, or humans vary in size from only 1.2 to 1.4 kb, despite extreme variation in sequence (ROBERTSON *et al.* 1998). Furthermore, MLE families seem to be mainly represented by full-length copies in these species (ROBERTSON *et al.* 1998). One notable exception is *Hsmar1* in humans, which is present in 200 full-length copies (1.3 kb), but is responsible for the spread of >2000 80-bp MITEs (MORGAN 1995). In rice, there is an overwhelming copy number excess of *Stowaway* MITEs over *Osmar*s (22,000–33,000 *vs.* <40). Together, these differences may reflect differences in the *cis*- and *trans*-requirements of animal and plant MLE transposases and/or their evolutionary dynamic.

A comprehensive inventory of *Stowaway* MITEs in rice: A comprehensive collection of *Stowaway* families was obtained by combining data gathered from previous studies with those generated *de novo* by the program RECON for ~30 Mb of rice sequences (see MATERIALS AND METHODS). Searches of ~360 Mb of Nipponbare BAC/PAC sequences with this collection indicate that this genome contains from 22,000 to 33,000 *Stowaway* elements that group into 36 families (Figure 3). These values are in the range of those reported in previous studies (JIANG and WESSLER 2001; TURCOTTE *et al.* 2001; GOFF *et al.* 2002; YU *et al.* 2002) and confirm that *Stowaway* is one of the most abundant classes of interspersed repeats in rice, contributing up to ~2% of the total genomic DNA.

Like most previously described MITE families, *Stowaway* families are characterized by relatively high numbers of copies (for class 2 transposons) and a remarkable conservation in size (standard deviation from consensus size is typically <2% per family; data not shown). There are, however, large variations in copy number among families, ranging from several dozen to a few thousand (Figure 3). Interestingly, the most expansive families are also those with the longest TIRs (*e.g.*, *Stow-Os1*, *Stow-Os23*, and *Os-Stow46*; see Figure 3). It is tempting to speculate that the palindromic structure of these *Stowaway*s may have contributed to their success.

The complex relationship of *Osmar* and *Stowaway* elements: Multiple alignments and phylogenetic analyses of hundreds of family members show that most *Stowaway* families are made of multiple subfamilies of variable age (not shown). This phylogenetic structure indicates

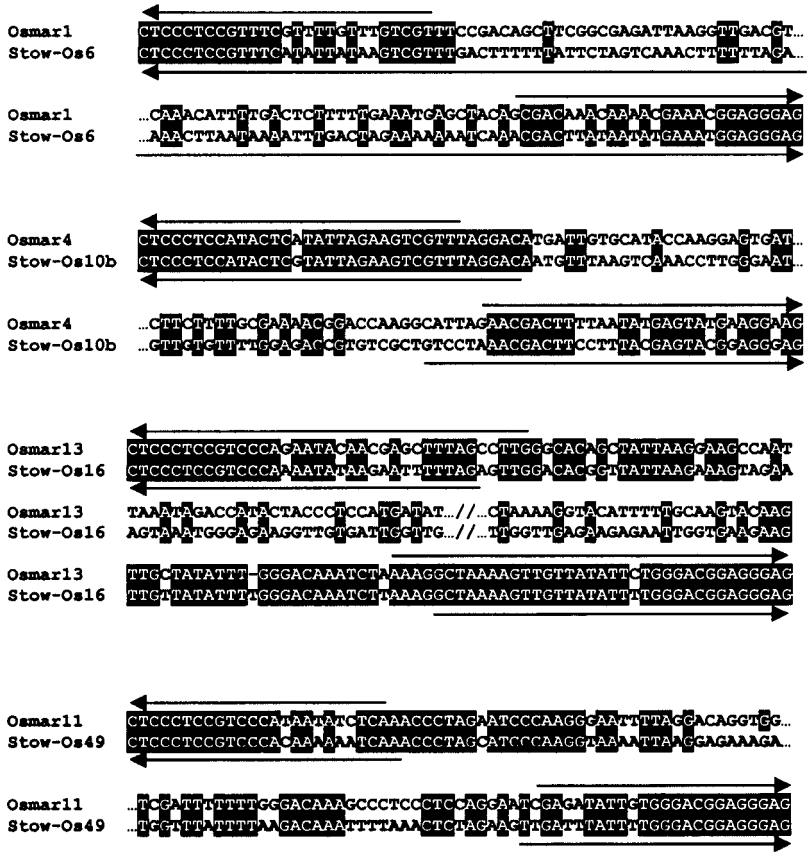

FIGURE 5.—Selected examples of pairwise sequence comparisons between the terminal sequences of *Osmar* and *Stowaway* elements. Examples of *Osmar-Stowaway* matches were selected to illustrate the range and extent of sequence similarity that can be found between the two groups. Significant similarities are usually restricted to the TIRs (*Osmar1 vs. Stow-Os6* and *Osmar4 vs. Stow-Os10b*) but, in a few cases, can be extended to the subterminal regions (*Osmar13 vs. Stow-Os16* and *Osmar11 vs. Stow-Os49*). TIRs are shown as arrows above *Osmar* sequences and below *Stowaway*.

that there have been multiple waves of amplification of a limited number of *Stowaway* progenitors. Having identified and characterized *Stowaway* and MLE families, we are now in a position to address two key questions: What are the enzymatic sources responsible for the bursts of *Stowaway* transposition? How do *Stowaway* progenitors originate?

Osmars as the transposase sources for Stowaway MITEs: Comparison of *Osmar* and *Stowaway* sequences shows that similarity is primarily restricted to the first 20–30 bp of the elements (Figures 5 and 6). In some pairwise comparisons, significant similarities could be extended to the subterminal regions, associating a given *Stowaway* family with an *Osmar* family (see Figure 5). However, the level of similarity in these comparisons (<85%) is below the value of a typical intrafamily relationship. Nevertheless, these are the closest matches that can be established in the rice genome between a high-copy-number *Stowaway* family and an element encoding a transposase. Therefore, *Osmar* elements are the best candidates as the autonomous partners of *Stowaway* families.

Our comparative analysis of the TIRs of *Osmar* and *Stowaway* provides further evidence for a functional relationship between these two groups of transposons. We showed that *Osmar* and *Stowaway* families could be placed into corresponding groups on the basis of characteristic motifs in their TIRs (Figures 2, 3, and 6). In turn, each of these motifs was found to be diagnostic of a distinct group of MLE transposase. Thus, similarities of *Osmar* and *Stowaway* in TIRs were used to connect almost every *Stowaway* family with one of four distinct clades of *Osmar* transposase (see Figure 6). Coevolution of TIR and transposase sequences is expected because transposase molecules recognize and bind specifically to the TIRs during the transposition reaction of most class 2 transposons, including Tc1/*mariner* elements (LAMPE *et al.* 2001; ZHANG *et al.* 2001; PLASTERK and VAN LUENEN 2002). Hence, changes in the transposase sequences are likely to be accompanied by changes in the TIR sequences and vice versa (LAMPE *et al.* 2001; NAUMANN and REZNIKOFF 2002). The conservation of TIR sequences for different element families should thus reflect the use of the same or a very similar source of transposase. For these reasons, we believe that the correspondence of *Osmar* and *Stowaway* TIRs is functionally significant and supports the notion that different *Stowaway* families have amplified by using distinct MLE transposases (FESCHOTTE *et al.* 2002a).

Origin of Stowaway MITEs: Although evidence for a functional relationship between *Stowaway* MITEs and *Osmar* transposases is accumulating, there were very few cases of clear-cut sequence relationship between *Stowaway* and *Osmar* elements (*i.e.*, where the MITE resembles an internal deletion derivative of the larger element; see examples in Figure 4). In fact, such cases are restricted to *Stowaway* elements that have not amplified

to high copy numbers and represent a negligible fraction of the 22,000–33,000 *Stowaway*s present in rice. Thus, the origin of high-copy-number *Stowaway* families remains enigmatic.

One possible explanation for this situation is the differential retention of *Stowaway* and *Osmar* elements in the rice genome over evolutionary time. Assuming that the loss of transposons is primarily a stochastic process (Hartl *et al.* 1997), MITEs may simply have a greater chance to persist because they outnumber their autonomous partners. MITEs may also have a selective advantage over *Osmar*s, because their insertions are less likely than those of larger elements to be deleterious. While differential retention may explain some *Osmar-Stowaway* situations, it cannot explain all of them. Among the dozens of *Osmar* and *Stowaway* families described in this study, one would have expected to find at least a few cases of direct association. After all, clear-cut relationships between large MITE families and full-length Tc1/*mariner* transposons were previously found in the Arabidopsis and human genomes (*e.g.*, Morgan 1995; Smit and Riggs 1996; Feschotte and Mouchès 2000; reviewed in Feschotte *et al.* 2002b).

Instead of differential retention, we propose two, not mutually exclusive, alternative hypotheses. First, some *Stowaway* families may not be derived from *Osmar*, but may originate *de novo* following the fortuitous association and recognition of TIRs flanking unrelated segments of DNA. The creation of a new DNA transposon by capture of flanking sequence has been reported for the *P* element in Drosophila (Tsubota and Huong 1991) and a similar scenario was proposed for the origin of *Ds1* elements in maize (MacRae and Clegg 1992). Support for this hypothesis comes from the fact that 13 high-copy-number *Stowaway* families are characterized by elements with long TIRs (48–94 bp; Figure 3), whereas there are no *Osmar*s with TIRs longer than 36 bp (Figure 2). Long arrays of palindromic DNA are frequently encountered in eukaryotic genomes (*e.g.*, Cavalier-Smith 1974; Deininger and Schmid 1976) and may provide the raw material for the *de novo* origin of some MITE families.

De novo origins are unlikely for other *Stowaway* families that have extended regions of similarity with coexisting MLEs (Figure 5). These *Stowaway* families may have originated by internal deletion of *Osmar*s, but amplification to higher copy numbers could be a secondary event mediated by a transposase encoded by a distantly related element (see Figure 7). In this model, the origin and amplification of MITEs are considered as two different steps that may be separated by a long period of time. The more time elapsed between these two steps, the more difficult it will be to recognize the filiation between a MITE family and an autonomous element.

MITE amplification via cross-mobilization: Regardless of the origin of MITEs (*de novo* or ancient deletion derivatives), our results suggest that cross-mobilization is one

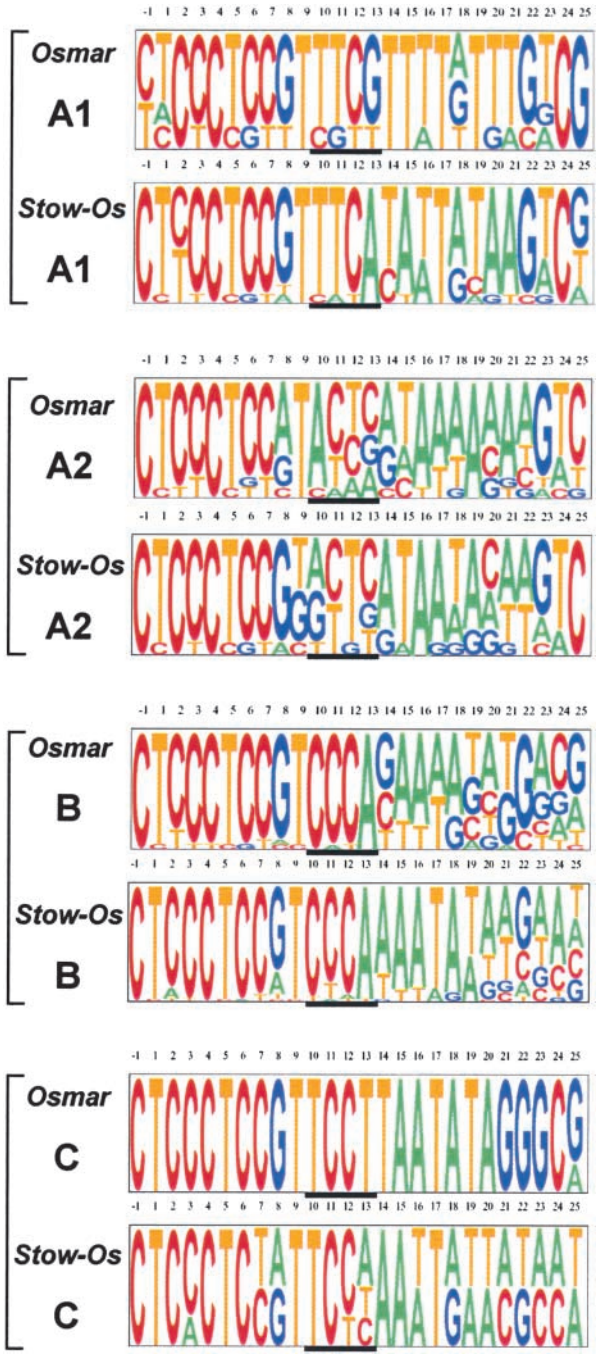

Figure 6.—Comparison of *Osmar* and *Stowaway* TIRs. *Osmar* and *Stowaway* can be classified into corresponding clades (A1, A2, B, and C) on the basis of the presence of a 4-bp diagnostic motif in their TIR sequences (see Figures 2 and 3). To further illustrate the TIR similarity in corresponding clades of *Osmar* and *Stowaway*, a pictogram was constructed using the first 25 nucleotides and the reverse complement of the last 25 nucleotides of all clade members (see Figures 2 and 3). In this representation, the size of the letter is proportional to its frequency at a given position. A thick black line underscores the clade-specific 4-bp motif used to classify *Osmar* and *Stowaway* families. Pictograms were generated at http://genes.mit.edu/pictogram.html, using default parameters.

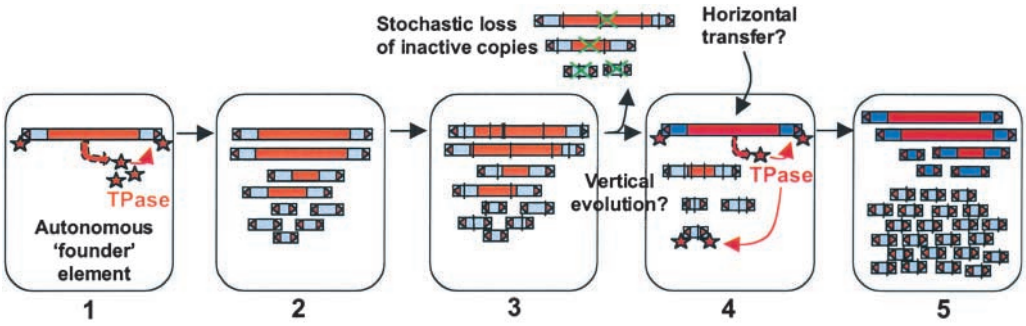

FIGURE 7.—Model for the amplification of *Stowaway* MITEs. The first three steps of this model are based on the life cycle of MLEs proposed by HARTL *et al.* (1997). An autonomous element, newly introduced in the genome of a species by either vertical inheritance or horizontal transmission, first transposes at relatively high frequency (step 1). This may lead to a rapid increase in copy number if the double-strand gap left after excision is repaired using a locus containing the same transposon (for example, the homologous chromosome). Many newly synthesized transposons are internally deleted versions of the autonomous copy because of frequent interruption and/or slippage during gap repair (step 2). Copy numbers may increase until host defense mechanisms (homology-dependent silencing) and auto-regulatory processes (transposase titration, overproduction inhibition, etc.) act to repress transposition and stabilize copy numbers (HARTL *et al.* 1997; OKAMOTO and HIROCHIKA 2001; HANNON 2002; PLASTERK and VAN LUENEN 2002). Over time, both active and inactive copies are progressively degraded by point mutations (vertical inactivation) and are stochastically lost or fossilized in the genome (step 3). By chance, some of the decayed elements might preserve (or evolve *de novo*) sequences recognized by the transposase of a newly introduced autonomous element (step 4). The new autonomous element might be introduced from another species by horizontal transfer or genetic introgression, but it may also emerge "vertically" by diversifying evolution of a previously inactivated full-length element (as discussed by LAMPE *et al.* 2001). The newly expressed transposase will thus be able to mobilize its own family members and distantly related MITEs (step 4). We propose that sequence divergence between the MITE and its autonomous partner may favor the propagation of MITEs and allow their amplification to high copy numbers (step 5; see text for details).

of the major mechanisms operating in the rice genome to amplify MITEs to high copy numbers. There are previous examples of cross-mobilization of short DNA transposons by distantly related autonomous elements. In maize, *Ds1* elements (~400 bp) have only the 5′ terminal 13 bp and the 3′ terminal 26 bp in common with *Ac* elements, but they can be mobilized by the *Ac* transposase (*e.g.*, SHEN *et al.* 1998). In *Caenorhabditis elegans*, the nonautonomous Tc7 elements are mobilized *in vivo* and *in vitro* by the Tc1 transposase, even though Tc7 and Tc1 share only their 36 terminal nucleotides (REZSOHAZY *et al.* 1997). Interestingly, Tc7, like many other MITE-like families in this species, has no parental autonomous copies recognizable in the *C. elegans* genome. For example, the MITE families *CeleTc2*, *Cele11*, and *Cele12* all have Tc2-like TIRs while their internal sequences have little similarity to each other or to other Tc2 sequences (OOSUMI *et al.* 1996). To explain this and related instances, the authors hypothesized that a single Tc family can cross-mobilize a variety of highly divergent sequences (OOSUMI *et al.* 1996). The cross-mobilization model also gains support from the recent discovery that another rice MITE family, *mPing*, is co-mobilized in cell culture with a closely, but not directly related, autonomous *Pong* element (JIANG *et al.* 2003). Interestingly, *mPing* elements are *Tourist*-like MITEs, the other principal MITE group in plants. Together with our study of the *Osmar-Stowaway* relationships in rice, these data converge toward a model where cross-mobilization plays a major role in the amplification of MITEs.

How could cross-mobilization contribute to MITE amplification? Recent studies have shown that the activity of many transposable element families is repressed by epigenetic mechanisms that act at the transcriptional or post-transcriptional level to repress the expression of the transposon gene product (OKAMOTO and HIROCHIKA 2001; FESCHOTTE *et al.* 2002a; HANNON 2002; PLASTERK and VAN LUENEN 2002). All of these mechanisms are based on recognition of nucleic acid sequence homology and triggered by multiple copies of the target sequence. As a result, there is usually an inverse correlation between the copy number of a transposable element family, the expression levels of their gene products, and/or the transpositional activity of the family (*e.g.*, CHANDLER and WALBOT 1986; CHABOISSIER *et al.* 1998; HIROCHIKA *et al.* 2000). We speculate that a MITE lacking extensive sequence homology with an active autonomous element, but retaining the short *cis*-sequences (TIRs) recognized by the corresponding transposase, may be able to multiply without triggering homology-dependant mechanisms of transposon silencing (see model in Figure 7). This would ensure the maintenance of a high level of transposase expression, which would allow MITE families to quickly spread and attain high copy numbers.

We are grateful to N. Jiang and Z. Bao for their help in the mining and analysis of *Stowaway* families and for sharing unpublished information. We also thank J. Jurka for providing access to data on *Stowaway* families prior to their publication in Repbase. We thank N. Jiang, M. Osterlund, E. Pritham, and X. Zhang for critical reading of the manuscript and helpful discussions. This work was supported by grants from the National Science Foundation Plant Genome Initiative, the National Institutes of Health, and the University of Georgia Research Foundation to S.R.W.

LITERATURE CITED

BAO, Z., and S. R. EDDY, 2002 Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. **12:** 1269–1276.

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

Bureau, T. E., and S. R. Wessler, 1994 *Stowaway*: a new family of inverted-repeat elements associated with genes of both monocotyledonous and dicotyledonous plants. Plant Cell **6:** 907–916.

Bureau, T. E., P. C. Ronald and S. R. Wessler, 1996 A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc. Natl. Acad. Sci. USA **93:** 8524–8529.

Capy, P., C. Bazin, D. Higuet and T. Langin, 1998 *Dynamics and Evolution of Transposable Elements.* Springer-Verlag, Austin, TX.

Cavalier-Smith, T., 1974 Long palindromes in eukaryotic DNA. Nature **262:** 255–256.

Chaboissier, M. C., A. Bucheton and D. J. Finnegan, 1998 Copy number control of a transposable element, the I factor, a LINE-like element in Drosophila. Proc. Natl. Acad. Sci. USA **95:** 11781–11785.

Chandler, V. L., and V. Walbot, 1986 DNA modification of a maize transposable element correlates with loss of activity. Proc. Natl. Acad. Sci. USA **83:** 1767–1771.

Deininger, P. L., and C. W. Schmid, 1976 An electron microscope study of the DNA sequence organization of the human genome. J. Mol. Biol. **106:** 773–790.

Feschotte, C., and C. Mouchès, 2000 Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. Mol. Biol. Evol. **17:** 730–737.

Feschotte, C., and S. R. Wessler, 2002 *Mariner*-like transposases are widespread and diverse in flowering plants. Proc. Natl. Acad. Sci. USA **99:** 280–285.

Feschotte, C., N. Jiang and S. R. Wessler, 2002a Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. **3:** 329–341.

Feschotte, C., X. Zhang and S. Wessler, 2002b Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons, pp. 1147–1158 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, DC.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. Wang *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92–100.

Hannon, G. J., 2002 RNA interference. Nature **418:** 244–251.

Hartl, D. L., A. R. Lohe and E. R. Lozovskaya, 1997 Modern thoughts on an ancyent *marinere*: function, evolution, regulation. Annu. Rev. Genet. **31:** 337–358.

Hirochika, H., H. Okamoto and T. Kakutani, 2000 Silencing of retrotransposons in Arabidopsis and reactivation by the ddm1 mutation. Plant Cell **12:** 357–369.

Jarvik, T., and K. G. Lark, 1998 Characterization of *Soymar*1, a *mariner* element in soybean. Genetics **149:** 1569–1574.

Jiang, N., and S. R. Wessler, 2001 Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. Plant Cell **13:** 2553–2564.

Jiang, N., Z. Bao, X. Zhang, H. Hirochika, S. R. Eddy *et al.*, 2003 An active DNA transposon family in rice. Nature **421:** 163–167.

Kunze, R., and C. F. Weil, 2002 The hAT and CACTA superfamilies of plant transposons, pp. 565–610 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, DC.

Lampe, D. J., K. K. Walden and H. M. Robertson, 2001 Loss of transposase-DNA interaction may underlie the divergence of mariner family transposable elements and the ability of more than one mariner to occupy the same genome. Mol. Biol. Evol. **18:** 954–961.

Le, Q. H., S. Wright, Z. Yu and T. Bureau, 2000 Transposon diversity in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **97:** 7376–7381.

MacRae, A. F., and M. T. Clegg, 1992 Evolution of *Ac* and *Ds1* elements in select grasses (Poaceae). Genetica **86:** 55–66.

Mao, L., T. C. Wood, Y. Yu, M. A. Budiman, J. Tomkins *et al.*, 2000 Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res. **10:** 982–990.

Morgan, G. T., 1995 Identification in the human genome of mobile elements spread by DNA-mediated transposition. J. Mol. Biol. **254:** 1–5.

Naumann, T. A., and W. S. Reznikoff, 2002 Tn5 transposase with an altered specificity for transposon ends. J. Bacteriol. **184:** 233–240.

Okamoto, H., and H. Hirochika, 2001 Silencing of transposable elements in plants. Trends Plant. Sci. **6:** 527–534.

Oosumi, T., B. Garlick and W. R. Belknap, 1996 Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. J. Mol. Evol. **43:** 11–18.

Plasterk, R. H. A., and H. G. van Luenen, 2002 The Tc1/*mariner* family of transposable elements, pp. 519–532 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, DC.

Plasterk, R. H. A., Z. Izsvák and Z. Ivics, 1999 Resident aliens: the Tc1/*mariner* superfamily of transposable elements. Trends Genet. **15:** 326–332.

Rezsohazy, R., H. G. A. M. van Luenen, R. M. Durbin and R. H. A. Plasterk, 1997 Tc7, a Tc1-hitch hiking transposon in *Caenorhabditis elegans*. Nucleic Acids Res. **25:** 4048–4054.

Robertson, H. M., F. N. Soto-Adames, K. O. Walden, R. M. Avancini and D. J. Lampe, 1998 The mariner transposons of animals: horizontally jumping genes, pp. 268–284 in *Horizontal Gene Transfer*, edited by M. Syvanen and C. I. Kido. Chapman & Hall, London.

Shao, H., and Z. Tu, 2001 Expanding the diversity of the *IS630-Tc1-mariner* superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics **159:** 1103–1115.

Shen, W. H., C. Ramos and B. Hohn, 1998 Excision of *Ds1* from the genome of maize streak virus in response to different transposase-encoding genes. Plant Mol. Biol. **36:** 387–392.

Smit, A. F. A., and A. D. Riggs, 1996 *Tiggers* and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA **93:** 1443–1448.

Tarchini, R., P. Biddle, R. Wineland, S. Tingey and A. Rafalski, 2000 The complete sequence of 340 kb of DNA around the rice adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. Plant Cell **12:** 381–391.

Tsubota, S. I., and D. V. Huong, 1991 Capture of flanking DNA by a P element in *Drosophila melanogaster*: creation of a transposable element. Proc. Natl. Acad. Sci. USA **88:** 693–697.

Turcotte, K., S. Srinivasan and T. Bureau, 2001 Survey of transposable elements from rice genomic sequences. Plant J. **25:** 169–179.

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.

Yu, J., S. Hu, J. Wang, G. K. Wong, S. Li *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92.

Zhang, L., A. Dawson and D. J. Finnegan, 2001 DNA-binding activity and subunit interaction of the mariner transposase. Nucleic Acids. Res. **29:** 3566–3575.

Communicating editor: M. J. Simmons