Copyright © 2004 by the Genetics Society of America

PIF- and *Pong*-Like Transposable Elements: Distribution, Evolution and Relationship With *Tourist*-Like Miniature Inverted-Repeat Transposable Elements

Xiaoyu Zhang, Ning Jiang, Cédric Feschotte and Susan R. Wessler[1]

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

Manuscript received May 19, 2003
Accepted for publication October 30, 2003

ABSTRACT

Miniature inverted-repeat transposable elements (MITEs) are short, nonautonomous DNA elements that are widespread and abundant in plant genomes. Most of the hundreds of thousands of MITEs identified to date have been divided into two major groups on the basis of shared structural and sequence characteristics: *Tourist*-like and *Stowaway*-like. Since MITEs have no coding capacity, they must rely on transposases encoded by other elements. Two active transposons, the maize *P Instability Factor* (*PIF*) and the rice *Pong* element, have recently been implicated as sources of transposase for *Tourist*-like MITEs. Here we report that *PIF*- and *Pong*-like elements are widespread, diverse, and abundant in eukaryotes with hundreds of element-associated transposases found in a variety of plant, animal, and fungal genomes. The availability of virtually the entire rice genome sequence facilitated the identification of all the *PIF/Pong*-like elements in this organism and permitted a comprehensive analysis of their relationship with *Tourist*-like MITEs. Taken together, our results indicate that *PIF* and *Pong* are founding members of a large eukaryotic transposon superfamily and that members of this superfamily are responsible for the origin and amplification of *Tourist*-like MITEs.

TRANSPOSABLE elements (TEs), which are a major component of all characterized eukaryotic genomes, have been divided into two classes according to their transposition intermediate. Class 1 (RNA) elements transpose via an RNA intermediate and most either have long terminal repeats (LTR-retrotransposons) or terminate at one end with a poly(A) tract [long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)]. Class 2 (DNA) elements transpose via a DNA intermediate and usually have short terminal inverted repeats (TIRs). DNA elements can be further classified into families on the basis of the transposase (TPase) that catalyzes their movement. A TE family is composed of one or more TPase-encoding autonomous elements and up to several thousand nonautonomous elements that do not encode functional TPases but retain the *cis*-sequences necessary to be mobilized by the cognate TPase (for reviews see Capy *et al.* 1998; Feschotte *et al.* 2002a).

Miniature inverted-repeat transposable elements (MITEs) were first discovered in the grasses and later found in other flowering plants as well as in animal genomes (for review see Feschotte *et al.* 2002b). Structurally, MITEs are reminiscent of nonautonomous DNA elements, but their high copy number and intrafamily homogeneity in size and sequence distinguish them from most previously described nonautonomous elements (Wessler *et al.* 1995). Since MITEs lack coding sequences, their classification has been based on the sequence similarity of TIRs and target site duplication (TSD). Using these criteria, most of the tens of thousands of plant MITEs have been divided into two groups: *Tourist*-like (3-bp TSDs, usually TTA/TAA) and *Stowaway*-like (2-bp TSDs, usually TA) (Feschotte *et al.* 2002b).

Two distantly related families of active DNA transposons have recently been associated with *Tourist*-like MITEs. The maize *P Instability Factor* (*PIF*) and a *Tourist*-like MITE family called *miniature PIF* (*mPIF*) share identical TIRs, similar subterminal sequences, and a strong preference for insertion into the 9-bp palindrome CWCTTAGWG with duplication of the central TTA (Walker *et al.* 1997; Zhang *et al.* 2001). *PIF* contains two open reading frames (ORFs), one encoding a TPase, whereas the function of the other ORF is unknown (Zhang *et al.* 2001; Figure 1a). An even closer relationship was found in rice, where the 430-bp *Tourist*-like MITE called *mPing* was shown to be a deletion derivative of a 5.2-kb transposase-encoding element called *Ping* (Jiang *et al.* 2003; Kikuchi *et al.* 2003; Nakazaki *et al.* 2003). Several lines of evidence, however, led to the conclusion that a related element in the rice genome, called *Pong*, was the most likely source of TPase mobilizing *mPing* elements (Jiang *et al.* 2003). Indeed, *Pong* elements and *mPing* MITEs were found to be actively

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY362792–AY362819.

[1]*Corresponding author:* 4505 Miller Plant Science Bldg., University of Georgia, Athens, GA 30602. E-mail: sue@plantbio.uga.edu

transposing in the same cell culture line. Similar to *PIF*, *Pong* also contains two ORFs (ORF1 and ORF2). ORF1 may be involved in DNA binding as it includes a domain with weak similarity to the DNA-binding domain of *myb* transcription factors (Jiang *et al.* 2003; Figure 1b). ORF2 most likely encodes the TPase, as it contains an apparent DDE motif, a signature consisting of three acidic residues found in the catalytic domains of some eukaryotic and prokaryotic TPases (Rezsohazy *et al.* 1993; Mahillon and Chandler 1998; Jiang *et al.* 2003).

PIF and *Pong* elements share several features, including amino acid sequence conservation in the catalytic domain of ORF2, homology in their ORF1s, nucleotide sequence similarity in their TIRs, and identical TSDs (TTA; Jiang *et al.* 2003). The putative TPases of both *PIF* and *Pong* have a large number of homologs in GenBank that have been annotated as unknown/hypothetical proteins. Examination of several *PIF* TPase homologs from plants, nematodes, and a fungus showed that they resided in *PIF*-like transposable elements (Zhang *et al.* 2001). Furthermore, *PIF*-like and *Pong* TPases were shown to be distantly related to the TPases of bacterial insertion sequences of the IS*5* group (Kapitonov and Jurka 1999; Le *et al.* 2001; Zhang *et al.* 2001; Jiang *et al.* 2003). Thus, *PIF* and *Pong* are the founding members of what appears to be a new and widespread superfamily of DNA transposons called *PIF*/IS*5*. Significantly, some of these *PIF*-like elements were found to be associated with *Tourist*-like MITEs in their respective genomes (Le *et al.* 2001; Zhang *et al.* 2001; Aparicio *et al.* 2002; Feschotte *et al.* 2002b).

In this study we look at the distribution and evolution of the *PIF*/IS*5* superfamily of transposases and characterize their relationship with *Tourist*-like MITEs. To this end we conducted a systematic survey (database searches and PCR assays) of putative *PIF*- and *Pong*-like TPases in plants and animals. Phylogenetic analyses of >600 TPase fragments from 56 species define three major groups, each represented by multiple ancient and distinct lineages. The availability of virtually the entire sequence of rice (Goff *et al.* 2002; Yu *et al.* 2002) permitted the identification and characterization of all *PIF*- and *Pong*-like elements in a single genome. Furthermore, the association between rice *PIF*- and *Pong*-like elements and *Tourist*-like MITEs was explored by performing a genome-wide comparison of these elements. This represents the first comprehensive analysis of the origin of *Tourist*-like MITEs in any organism.

MATERIALS AND METHODS

PCR amplification of *PIF*-like TPases: Degenerate primers were derived from the regions encoding amino acid residues GALDGTH (D1F1, 5′-GGIGCHHTIGATGGHACWCA-3′; I, Inosine; H, A, C, or T; W, A, or T) and ELFNPRH (KR1, 5′-ATGICKMIRRTTRAACAAYTC-3′; K, G, or T; M, A, or C; R, A, or G; Y, C, or T; Figure 1a, positions indicated by arrows). PCR amplifications were performed with 10–100 ng of genomic DNA in 50-μl reactions. Cycling parameters were: 1 cycle at 94° for 3 min, 36 cycles at 94° for 30 sec, 50° for 30 sec, 72° for 1 min, and 1 cycle at 72° for 5 min. Forty microliters of the reaction was resolved on 1% agarose gels, and desired fragments were purified from agarose using the QIAquick gel extraction kit (QIAGEN, Valencia, CA) and cloned using the TOPO-TA cloning kit (Invitrogen, San Diego) according to manufacturers' instructions. Sequencing reactions were performed by the Molecular Genetics Instrumentation Facility of the University of Georgia. The sequences of 28 *PIF*-like TPase fragments described here were deposited in the GenBank database (accession nos. AY362792–AY362819).

Database searches and sequence and phylogenetic analyses: Database searches were performed with blast servers available from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov, databases nr, gss, est, and wgs_Anopheles) as well as the rice genomic database at The Institute for Genomic Research (http://tigrblast.tigr.org/euk-blast/index.cgi?project=osa1). Nucleotide sequences obtained from database searches and degenerate PCR (dPCR) amplifications were conceptually translated into amino acid sequences and aligned with CLUSTALW. Introns were predicted with Netgene2 [available at http://www.cbs.dtu.dk (Hebsgaard *et al.* 1996)]. Putative helix-turn-helix (HTH) motifs were predicted using the NPS@ program (available at http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_hth.html). Multiple sequence alignments used to generate the phylogenetic tree in Figure 2 were performed with the CLUSTALW server available at European Bioinformatics Institute (http://www.ebi.ac.uk/clustalw/) with default parameters. Multiple alignments used to generate other phylogenetic trees were performed using the MacVector program. Phylogenetic trees were generated on the basis of the neighbor-joining method, using PAUP* version 4.0b8 (Swofford 1999) with default parameters. Pictograms were generated at http://genes.mit.edu/pictogram.html.

RESULTS

Distribution and abundance of *PIF*/*Pong*-like TPases: *Identification of PIF/Pong-like TPases through database searches:* A systematic survey was carried out using the TPases of *PIF* and *Pong* as queries in tBlastn searches against several public databases. Significant similarity was detected in >1000 entries from a wide range of eukaryotic species, including 21 plants, 19 animals, and two fungi (listed in Table 1). Six hundred and seventy-three hits (574 were unique) with significant homology to the catalytic domains of *PIF* or *Pong* were selected for further analysis.

One striking result from database searches was the abundance of sequences for *PIF*/*Pong*-like TPases in some organisms, especially plants. This is most apparent in genomes with large amounts of sequence information: for the catalytic domains alone, there were ~80 hits (*e*-value $< 10^{-23}$) in *Arabidopsis thaliana*, ~350 hits (*e*-value $< 10^{-10}$) in rice (*Oryza sativa* c.v. Nipponbare), and ~170 hits (*e*-value $< 10^{-30}$) in *Brassica oleracea* [~30% of its 600-Mb genome available for blast at TIGR (http://www.tigr.org/tdb/e2k1/bog1/)]. In animals, the number of sequences for *PIF*/*Pong*-like TPases varies from >100 in the African malaria mosquito (*Anopheles gambiae*) and ~300–400 in zebrafish (*Danio rerio*) to a

TABLE 1

Species with *PIF*- or *Pong*-like TPases

Plants	
Angiosperms	
Monocot	Asparagus, barley (Hv), bamboo (Aa), coix (Ca), *Dianella ensifolia*, Ehrharta (Ec), foxtail millet, *Gongora ilense* (Gi), Johnson grass (Sh), *Joinvillea ascendens* (Ja), maize (Zm)[a], oat (As), *Pharus latifolius* (Pl), red millet (Pm), rice (Os)[a], sorghum (Sb)[a], sugarcane (Shc), *Tripsacum pilosum* (Tp), teosinte (Zp, Zh), wheat (Ta)
Dicot	*Arabidopsis* (At)[a], *Brassica oleracea* (Bo), ice plant, lettuce, *Lotus japonicus* (Lj), Medicago (Mt), peppermint, potato, soybean, sweet leaf, sugar beet, tomato (Le)
Gymnosperms	Pine
Algae	*Physcomitrella patens*, *Porphyra yezoensis*
Animals	
Invertebrate	
Nematode	*Caenorhabditis briggsae*[a], *C. elegans*[a]
Insect	Drosophila, African malaria mosquito, silkworm
Echinoderm	Sea urchin
Ascidian	Sea squirt
Vertebrate	
Fish	Medaka fish, Takifugu[a], trout, zebrafish
Amphibian	Xenopus
Bird	Chicken
Mammal	Chimpanzee, cow, human, mouse, pig, rat
Fungi	*Filobasisiella neoformans*[a] (Fn), *Neurospora crassa*

For sequences used in generating the phylogenetic tree in Figure 3, initials of species names are shown.

[a] *PIF*-like TPases were identified in these species by previous studies (Kapitonov and Jurka 1999; Le *et al.* 2001; Zhang *et al.* 2001; Aparicio *et al.* 2002; Jiang *et al.* 2003).

few (<3) in *Drosophila melanogaster*, *Caenorhabditis elegans*, and human.

Nucleotide sequences of the 574 unique *PIF*- and *Pong*-like TPases were conceptually translated into amino acid sequences after removal of introns (see below) and judicious correction of frameshifts caused by small (1–2 bp) insertions or deletions. The resulting amino acid sequences were compared to detect conserved regions that might signify functional domains. Several blocks of highly conserved residues were identified for *PIF*-like TPases (Figure 1a). Block H corresponds to a predicted HTH domain that may be involved in DNA binding. Blocks N2, N3, and C1 most likely comprise the catalytic domain as they contain an apparent DDE motif with the three acidic residues centered in blocks N2, N3, and C1, respectively. A DDE motif is also present in *Pong*-like TPases (Figure 1b), but, unlike *PIF*, no HTH domain was predicted.

PIF/Pong-like TPases are usually adjacent to ORF1 homologs: tBlastn searches using the ORF1s of *PIF* and *Pong* as queries also yielded a large number of hits. When located on long contigs, these ORF1 homologs were usually found within 1–2 kb of *PIF*- or *Pong*-like TPases, indicating that each "pair" of ORF1 and TPase was encoded by the same element. In fact, when the termini of *PIF/Pong*-like elements were defined in *O. sativa* (see below), *A. thaliana*, and *A. gambiae* (X. Zhang, C. Feschotte, and S. R. Wessler, unpublished data), nearly all elements were found to encode both ORFs. ORF1s are significantly more divergent than the TPases. Two blocks of conserved residues were found in *PIF/Pong*-like ORF1s (Figure 1, blocks A and B), with the most conserved block (block A) centered in a ~100-amino-acid (aa) region that displays weak homology to the DNA-binding domains of *myb* transcription factors from some plants and animals (Jiang *et al.* 2003). *Pong*-like ORF1s contain an additional well-conserved block (Figure 1b, block C).

Additional PIF-like TPases from grasses: The majority of *PIF*-like sequences (~80%) were from only a few species (rice, *Arabidopsis*, *B. oleracea*, and *A. gambiae*) since this survey was limited by the availability of DNA sequences in databases. To better resolve the phylogeny of *PIF* elements, additional TPase sequences were isolated from species with established evolutionary relationships but limited sequence information. To this end, a dPCR procedure was employed to amplify *PIF*-like TPase fragments from selected grass species. Grasses were chosen for this analysis because their phylogeny is well characterized (Kellogg 2001) and they harbor the only known active *PIF* and *Pong* elements (Walker *et al.* 1997; Zhang *et al.* 2001; Jiang *et al.* 2003).

dPCR primers were derived from the conserved blocks N2 and C1 in *PIF*-like TPases (see Figure 1a for

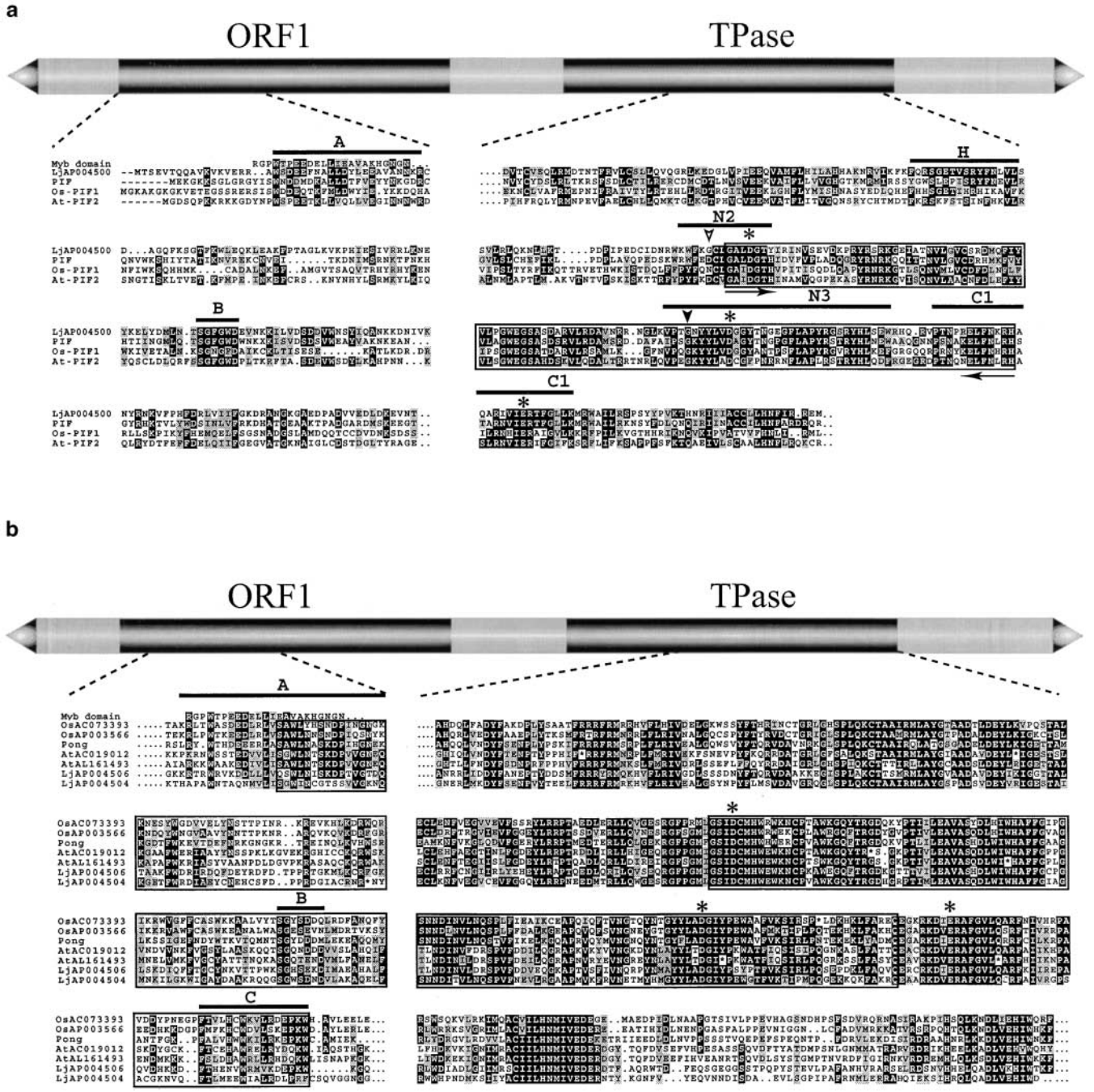


FIGURE 1.—Structure and conserved coding regions of *PIF*- and *Pong*-like elements. Triangles represent element TIRs. Coding regions are represented by dark gray boxes. Regions that are unrelated between different *PIF*- or *Pong*-like families are represented by light gray boxes. DDE motifs in TPases are indicated by asterisks, and open boxes localize regions that were used to deduce the phylogeny of *PIF*- or *Pong*-like elements (Figures 2–5 and 7). (a) *PIF*-like elements. Multiple alignments of the conserved regions of select *PIF*-like ORF1s and TPases are shown. Horizontal arrows indicate the position of dPCR primers, and the open and solid arrowheads denote the position of intron 1 and intron 2, respectively. Horizontal lines indicate the predicted HTH domain (H) and three blocks of conserved residues (N2, N3, and C1) that likely comprise the catalytic domain (see text). (b) *Pong*-like elements. Multiple alignments of the conserved regions from ORF1 and TPase of *Pong*-like elements are shown. A consensus sequence of the *Myb* domain is shown above ORF1 alignments for comparison. Horizontal lines in the ORF1 alignments indicate three blocks of conserved residues (A, B, and C). *Os-PIF1*, *At-PIF2*, and *Pong* were previously reported (KAPITONOV and JURKA 1999; ZHANG *et al.* 2001; JIANG *et al.* 2003). The *PIF* TPase shown in a is a full-length TPase isolated from maize (X. ZHANG and S. R. WESSLER, unpublished data). Other sequences were named according to the species initials (see Table 1), followed by their GenBank accession number.

positions and MATERIALS AND METHODS for sequences) and used to amplify an ~120-aa region from 20 grass species as well as several basal monocots (listed in Table 1). The amplified region included the majority of the catalytic domain in *PIF*-like TPases, extending from 3 aa upstream of the first Asp to 8 aa upstream of the Glu of the DDE motif (Figure 1a, boxed region). PCR products of the expected size (~360 bp) were successfully amplified from all 20 grasses tested and their close relative *Joinvillea*, as well as several Asparagales (*e.g.*, *Gongora ilense*; data not shown). In addition to the ~360-bp fragments, most species yielded larger PCR products (~450 bp) that, when sequenced, were found to contain an intron (see below).

Forty-five fragments from 15 species were sequenced; all were unique, indicating that there are multiple distinct TPases in each species and that only a small fraction had been sampled. No product was amplified from the more basal plants such as Zamia, Ginkgo, or Gnetum. Failure to amplify TPase fragments by dPCR from these species may be due to nucleotide variation in the primer recognition region or to the absence of *PIF*-like TPases.

Phylogeny of *PIF*-like and *Pong*-like TPases: *Three major clusters of PIF- and Pong-like TPases:* The TPase fragments identified by database mining and those isolated by dPCR were pooled and their evolutionary relationships examined. A multiple alignment was constructed from the 45 dPCR products and 574 unique database hits and used to generate an unrooted phylogenetic tree (Figure 2). The majority of the sequences clustered into three groups: the plant *PIF*-like group, the plant *Pong*-like group, and the animal group. In addition, the five fungal sequences clustered into two small, species-specific groups.

Clustering of the two plant groups was supported by bootstrap values as well as by several features that were shared within each group but not between groups. First, the spacing (*i.e.*, numbers of residues) between the second Asp and the Glu of the DDE motif differed between *PIF*-like and *Pong*-like groups but was consistent within each group. *PIF*-like TPases exhibit DD47E or DD48E spacing whereas *Pong*-like TPases exhibit an invariant DD35E spacing. Second, the TIRs of *PIF*- and *Pong*-like elements contain sequence motifs that are highly conserved within each group but distinct between the two groups (see below). On the basis of the comparison of TPase sequences, the animal group was related equally to both plant groups.

Phylogeny of plant PIF- and Pong-like TPases: Phylogenetic relationships among plant *PIF*- and *Pong*-like elements were determined by analyzing a subset of 99 sequences (63 *PIF*-like, 36 *Pong*-like) that were selected to represent the different lineages within each group. A CLUSTALW multiple alignment was constructed from the catalytic domains of these sequences and used to generate a phylogenetic tree (Figure 3). Both plant groups are monophyletic and heterogeneous. In each group, amino acid identity between sequences from distantly related species can be higher than that between two sequences from the same or from a closely related species, suggesting the presence of multiple ancient lineages of both *PIF*- and *Pong*-like elements.

The plant *PIF*-like group is composed of four major lineages (A–D). Lineage A is the largest and most complex with members from both monocots and dicots. It can be further divided into five sublineages (A1–A5). A1 includes five grass subfamilies (Panicoideae, Ehrhartoideae, Bambusoideae, Pooideae, and the ancestral Pharoideae), indicating that this sublineage was present before the diversification of the grasses ~70 MYA. Although only two grass subfamilies (Panicoideae and Ehrhartoideae) contributed sequences to A2, this lineage may be even more ancient than A1 as it is also found in the orchid *G. ilense* (order Asparagales). A3 and A4 are each found in a single dicot family (A3 in Brassicaceae and A4 in Fabaceae). A5 is present in both monocots and dicots and includes the only known active *PIF*-like element, the maize *PIF*. B and C are two small lineages from dicots, both restricted to the Brassicaceae family. Lineage D is another monocot-specific lineage found in four grass subfamilies (Panicoideae, Ehrhartoideae, Bambusoideae, and Pharoideae).

Pong-like TPases clustered into three major lineages (O–Q). Lineage O included two sublineages, the dicot-specific O1 and the monocot-specific O2. Lineage P is dicot specific, suggesting that it emerged in dicots after their separation from monocots. P could also be divided into two sublineages (P1 and P2). P1 was found in only the Brassicaceae family and included the majority of *Pong*-like TPases from *A. thaliana* (71%) and nearly all from *B. oleracea* (137 of 139). The P2 sublineage is probably older than P1 as it is also present in the Fabaceae family. Most TPase sequences in lineages Q were from *O. sativa*. However, the presence of one sequence from *Zea mays* and two from *Lotus japonicus* in lineage Q suggests that it is also an ancient lineage.

Introns in plant PIF-like elements: Although the original maize *PIF* element lacks introns (ZHANG *et al.* 2001), many plant *PIF*-like TPases contain one or two introns in their catalytic domains. The boundaries of these introns (donor/acceptor sites) were predicted with very high confidence (90–100%), and the coding sequences were restored (compared to intronless TPases) after their removal. Introns in *PIF*-like TPases can be classified into two classes on the basis of their position (Figure 1a, intron 1 and intron 2). Intron 1 is located 6 aa upstream of the first Asp residue of the DDE motif and intron 2 is located 6 aa upstream of the second Asp residue. Both introns are short (83 bp on average) and A/T rich (71% on average), with little conservation in length or sequence either within or among species. Significantly, the intron number and position from *PIF*-like TPases were consistent with the lineage designations. Two introns were present in three sublineages of A (A1, A3,

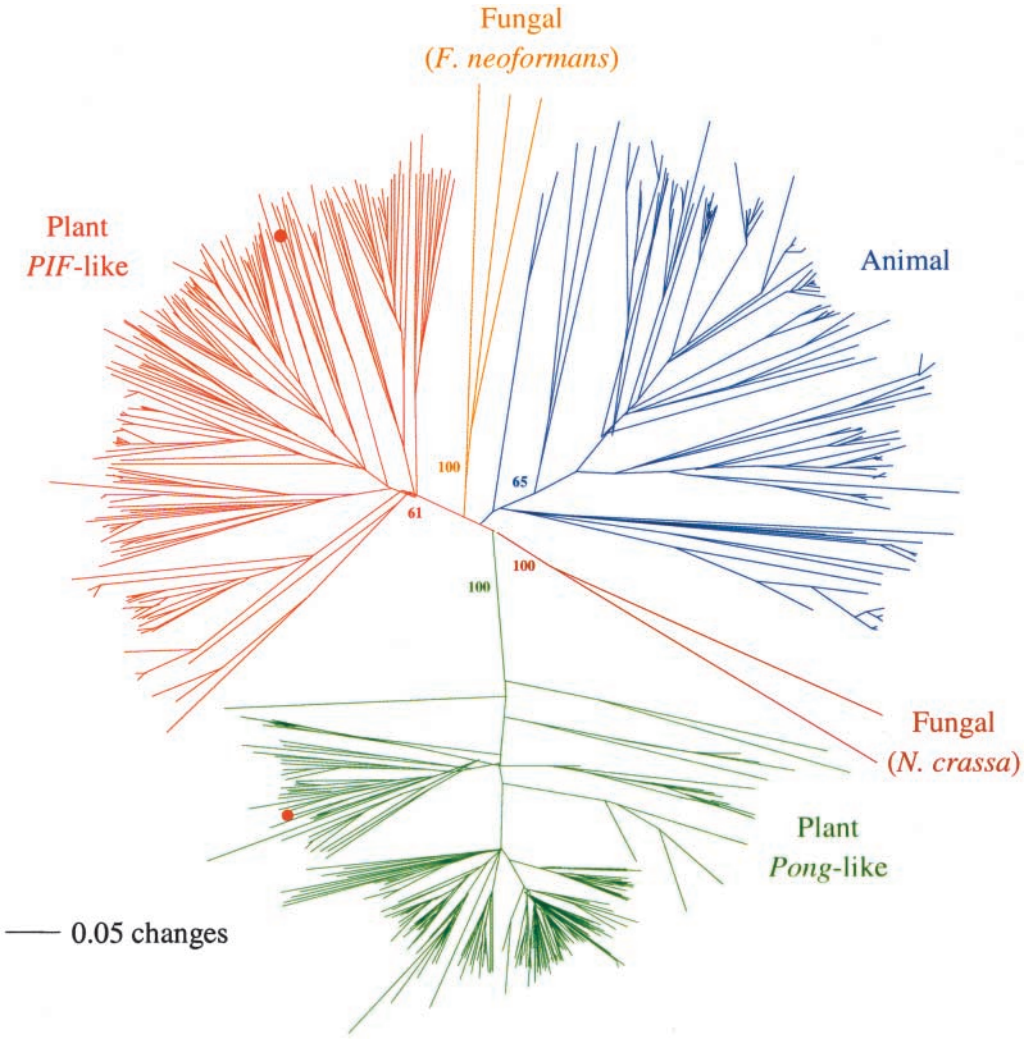


FIGURE 2.—Phylogeny of *PIF* and *Pong*-like TPases. The unrooted tree was generated using the neighbor-joining method from a CLUSTALW multiple alignment of the catalytic domain from 619 *PIF*- or *Pong*-like TPases identified by database searches or isolated by dPCR (see text). The maize *PIF* TPase and rice *Pong* TPase are represented by the solid red circle in their respective group. Bootstrap values were calculated from 500 replicates.

and A4), only intron 1 was present in A2, only intron 2 was present in lineages of D, and no introns were found in lineages B and C. Sublineage A5 was an exception: some TPases did not contain any introns while others contained two.

Two models have been proposed to explain the diversity of introns associated with related coding sequences: the "intron-early" model (loss of introns from an intron-rich ancestor; GILBERT *et al.* 1997) or the "intron-late" model (addition of introns to an intron-less ancestor; LOGSDON 1998). It is unlikely that the intron-late model explains the distribution of *PIF* introns as it would require multiple and independent intron acquisitions at identical positions. The intron-early explanation is more parsimonious since the data can be most easily interpreted by hypothesizing the existence of an ancestral *PIF* TPase with both introns and that multiple independent loss events occurred during evolution (Figure 3). According to this model, both introns were retained in the ancestor of lineage A, but intron 2 was lost in sublineage A4. Intron 1 was lost in the common ancestor of lineages B–D so that none of these lineages contains intron 1. Intron 2 was subsequently lost in the common ancestor of lineages B and C. The predicted stepwise loss of introns from *PIF*-like TPase genes contrasts with the plant *mariner*-like TPases, where the data are more consistent with the acquisition of introns during evolution (FESCHOTTE *et al.* 2002a,b).

TPase/ORF1 arrangements: Several different arrangements of TPase and ORF1 were observed for *PIF*- and *Pong*-like elements. All elements within a lineage or sublineage exhibit the same organization. Specifically, TPase and ORF1 in *PIF*-like elements are transcribed toward the same direction ("tail-to-head") but are organized in two different patterns, with the TPase gene located upstream of ORF1 in lineage A but downstream of ORF1 in lineages B, C, and D. Three different arrangements were found for *Pong*-like elements. TPase and ORF1 were organized in a "head-to-head" alignment for O1, a "tail-to-tail" alignment for O2, and a tail-to-head alignment for P and Q with TPase located downstream of ORF1 (see Figure 3).

***PIF*- and *Pong*-like elements in rice:** The availability of virtually the entire genomic sequence of *O. sativa* (GOFF

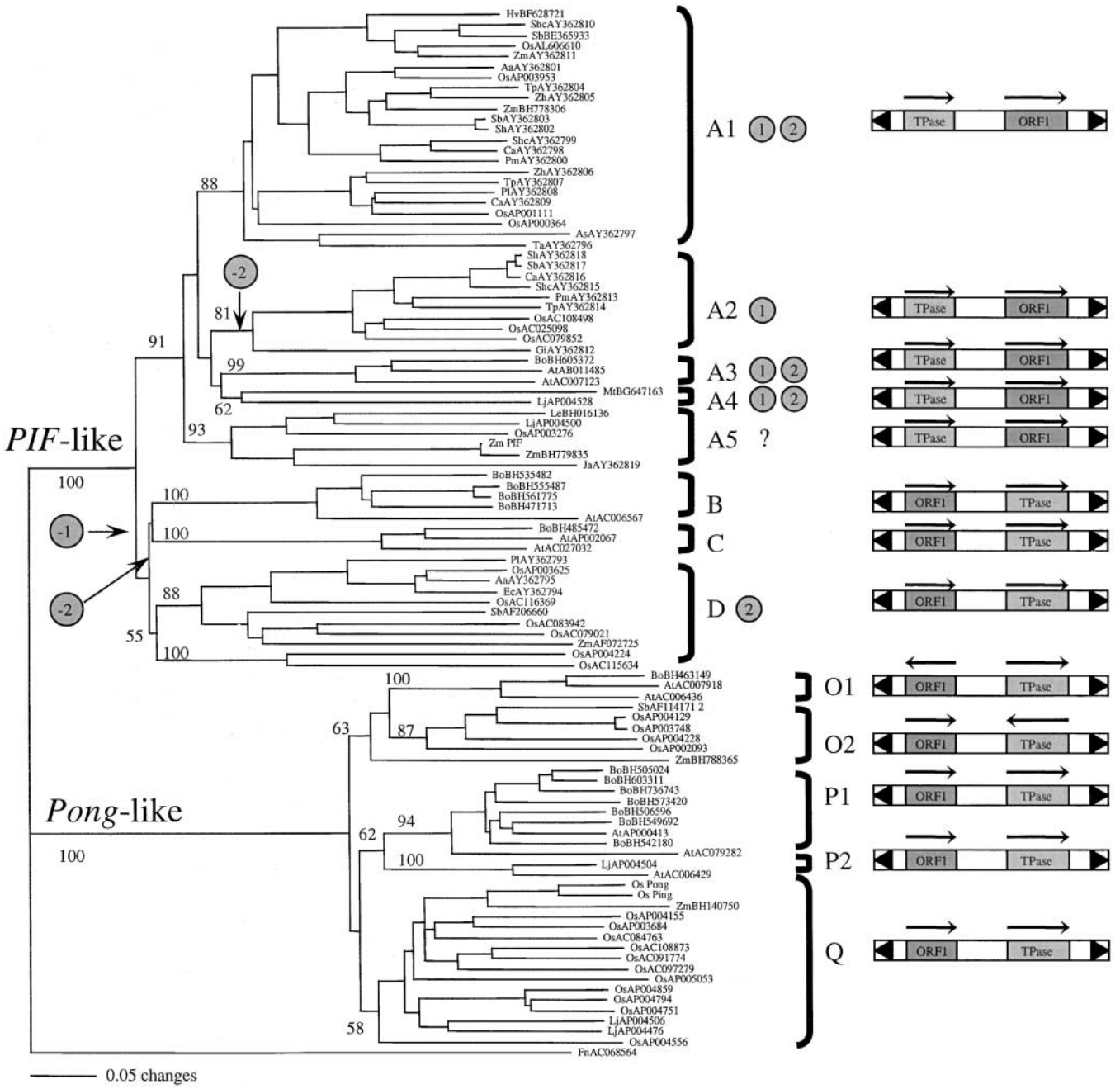


FIGURE 3.—Phylogeny of plant *PIF*- and *Pong*-like TPases. The phylogenetic tree was generated using the neighbor-joining method from a CLUSTALW multiple alignment of 99 catalytic domains of *PIF*- and *Pong*-like TPases and rooted with the catalytic domain of a *PIF*-like TPase from *F. neoformans*. Presence of intron 1 and/or intron 2 in a certain *PIF*-like lineage is shown as a "1" or "2" in a gray circle. The intron content of lineage A5 is variable (see text). Loss of intron 1 and/or intron 2 from *PIF*-like TPases (a "−1" or "−2" in a gray circle) is indicated by arrows. The relative organizations of ORF1/TPase for *PIF*/*Pong*-like element lineages are shown (arrows indicate direction of transcription). Sequences are named according to the species initial (see Table 1) followed by the GenBank accession number. Bootstrap values were calculated from 1000 replicates.

et al. 2002; YU *et al.* 2002) made it possible to conduct a comprehensive analysis of the relationships between *PIF*- and *Pong*-like elements and *Tourist*-like MITEs. To do this, *PIF*- and *Pong*-like TPase sequences were first identified by computer-assisted analysis and then the sequences flanking these hits were searched to define full-length *PIF*- and *Pong*-like elements.

PIF and Pong TPases: tBlastn searches using as queries the TPases of *PIF* and *Pong* led to the identification of 205 and 145 hits (*e*-value < -10), respectively, from the TIGR rice database (ssp. *japonica*, cv. Nipponbare). Duplicate hits located on overlapping regions of bacterial artificial chromosomes were excluded as were severely truncated TPases (containing <50% of the com-

plete coding region). The remaining 116 *PIF*-like TPases and 80 *Pong*-like TPases were relatively full-length and contained the entire catalytic domain. After removal of introns and correction of frameshifts caused by small insertion/deletions (1–2 bp), full-length *PIF*-like TPases were found to range in size from 392 to 432 aa, while full-length *Pong*-like TPases were from 416 to 549 aa.

The evolutionary relationships of rice *PIF*- and *Pong*-like TPases were determined by generating phylogenetic trees from CLUSTALW multiple alignments of their catalytic domains (Figures 4 and 5). Two major lineages for *PIF*-like TPases that correspond to lineages A (including sublineages A1, A2, and A5) and D were resolved as shown in Figure 3. Correlation between intron content of a *PIF*-like TPase and the TPase lineage is similar to that described in the broader plant survey (Figure 3) except for two additional intron loss events: *OsPIF3* lost intron 1 and *OsPIF18* lost intron 2. A tail-to-head alignment for TPase and ORF1 was found for all *OsPIF* elements (TPase located upstream of ORF1 in lineage A but downstream of ORF1 in lineage D) with one exception: the two ORFs in *OsPIF6* are organized in a tail-to-tail alignment, possibly due to a recent rearrangement. *Pong*-like TPases also clustered into two major groups, corresponding to the sublineage O2 and lineage Q in Figure 3. The ORF1/TPase alignment in lineage Q was found to be tail-to-head while that in lineage O2 is head-to-head.

Characterizing full-length elements: PIF- and *Pong*-like TPases were grouped into families on the basis of TPase sequence identities, with members of the same family being >90% identical. In this way, 27 *PIF*-like families and 26 *Pong*-like families (including the *Pong* family) were defined. These families were designated *OsPIF* (for *O. sativa PIF*) and *OsPong* (for *O. sativa Pong*), followed by the number of the family. Elements of the same family were further designated with a letter (*e.g.*, *OsPIF1a* and *OsPIF1b*, see Figures 4 and 5).

The identification of complete *OsPIF* and *OsPong* elements was complicated by the fact that interfamily comparisons indicated that sequence similarity was restricted to the known ORFs. For this reason, full-length elements were identified by comparison of sequences flanking the TPases within the same family where high sequence similarity extended into sequences flanking the ORFs. Sequences marking the boundary of similarity between elements of the same family were then searched for TIRs related to those of *PIF* or *Pong* and the flanking 3-bp TSDs characteristic of *PIF* and *Pong* elements (TTA/TAA). In this way, TIRs of 21 of the 27 *OsPIF* families (71 elements) and 20 of the 26 *OsPong* families (61 elements) were identified (see supplemental data at http://www.genetics.org/supplemental/ for accession numbers and positions).

The TIRs of *OsPIFs* were of variable length, ranging from 10 bp (*OsPIF4*) to 45 bp (*OsPIF20*). In contrast, *OsPong* TIRs were more uniform: all were 14–18 bp long except for one family (represented by a single-element *OsPong5*, 66-bp TIRs). Comparison of the TIRs of *OsPIFs* and *OsPongs* showed similarities (most began with 5′-GGSC-3′, where S represents G or C) as well as differences (the fifth nucleotide was usually A in *OsPongs* but was rarely an A in *OsPIFs*; Figure 6, a and b). The inner TIRs contained *PIF*-specific and *Pong*-specific motifs [*OsPIF*, 5′-TGTTTGGTT-3′ (positions 6–14); *OsPong*, 5′-STMCAA-3′ (positions 7–12), where M stands for A or C].

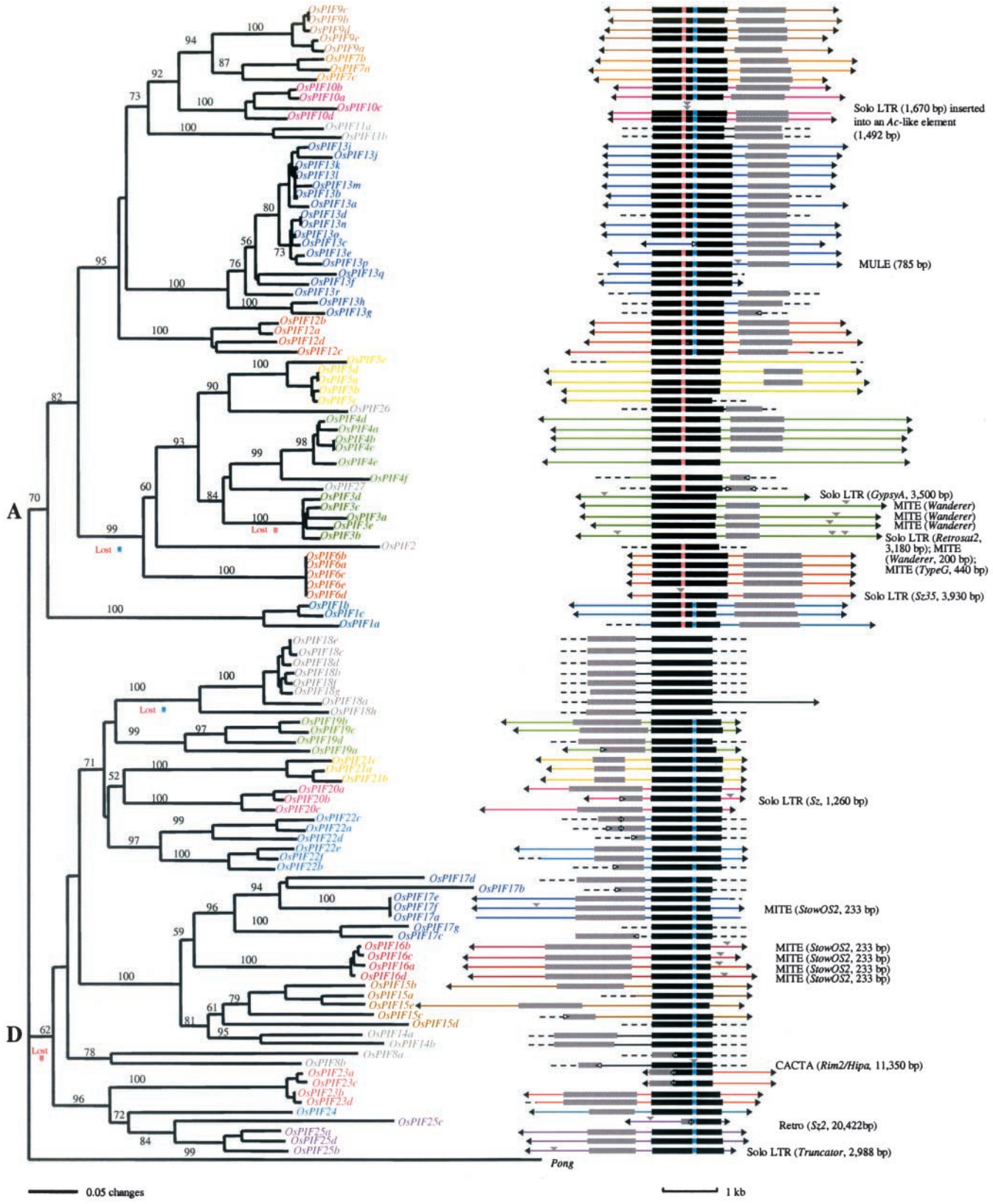
Full-length *OsPIFs* ranged from 2305 bp (*OsPIF23a*) to 22,169 bp (*OsPIF25c*) and *OsPongs* ranged from 2612 bp (*OsPong18d*) to 18,753 bp (*OsPong15a*). Most of this variation was due to the insertion of other TEs. These secondary TE insertions were located by searching full-length elements with RepeatMasker and Blastn. Sixteen *OsPIFs* and 24 *OsPongs* were found to contain a variety of TEs insertions (see Figures 4 and 5 for their positions and identities), including other DNA elements [*Ac*-like, *Mutator*-like (MULEs), and CACTA-like], MITEs (*Tourist*-like and *Stowaway*-like), LTR retroelements, solo LTRs (*Copia*-like and *Gypsy*-like), non-LTR retroelements (LINEs), and, in one case, a *Helitron* element. In a few instances, members of *OsPIF* and *OsPong* families (*e.g.*, *OsPIF3*, *OsPIF16*, and *OsPong18*) harbored the same MITE insertion at the same position, indicating that the MITE insertion did not prevent further transposition of these elements. When TE insertions were excluded, the length of most *OsPIFs* (50 of 71) and *OsPongs* (45 of 61) was found to be in the range of 4–6 kb.

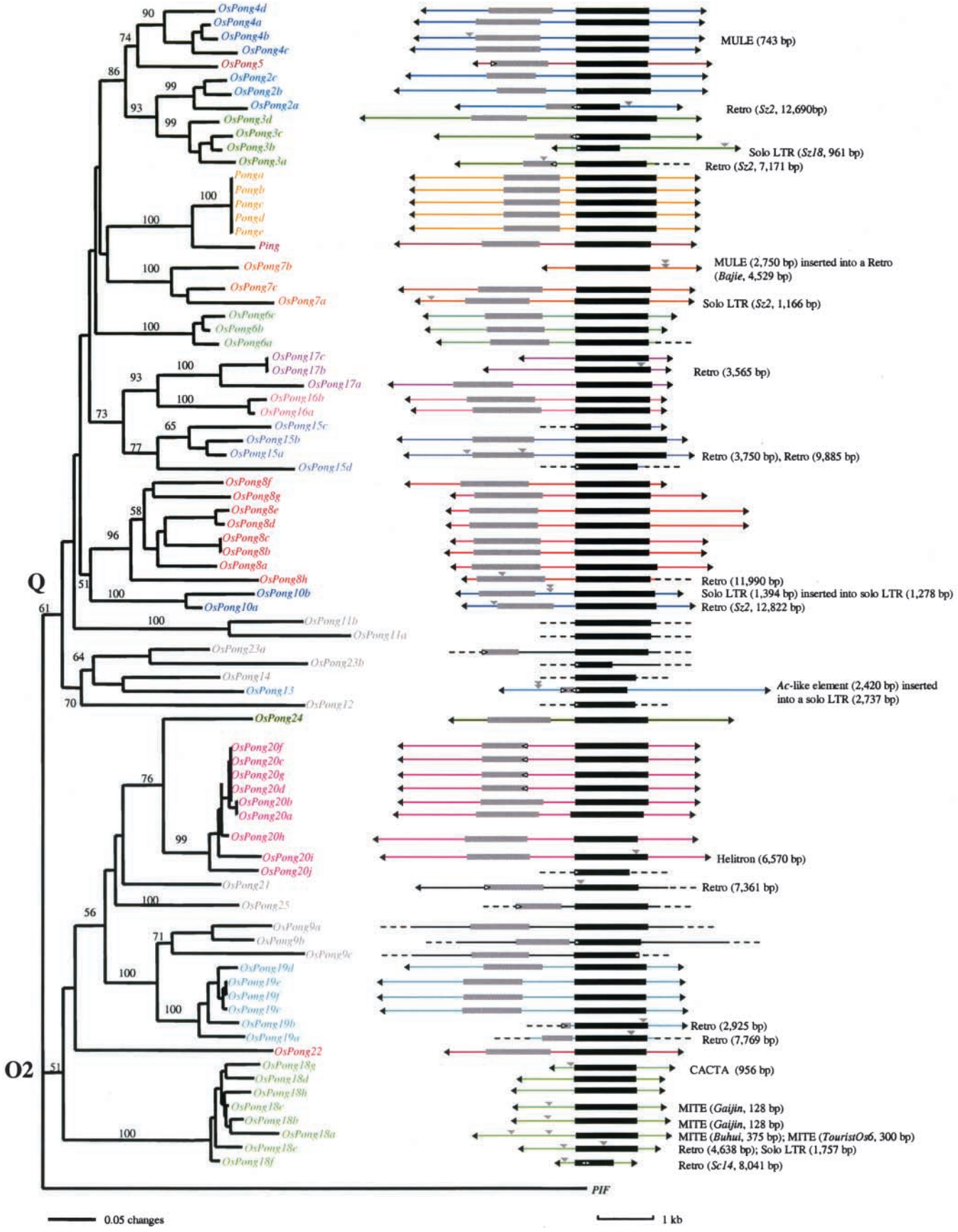
In several instances, *OsPIF* and *OsPong* families include elements that are nearly identical, suggesting that they transposed recently and may still be capable of further transposition. For example, *OsPIF6* includes five complete elements (~4.1 kb) located on four different chromosomes (chromosomes 3, 7, 9, and 10) that are, on average, ~99.6% identical over their entire length. In addition, their coding sequences are not interrupted

Figure 4.—Phylogeny of *OsPIF* TPases and the structure of the encoding element. The neighbor-joining tree was constructed from a CLUSTALW multiple alignment of the catalytic domains (boxed region in Figure 1a) of 116 *OsPIF* TPases and rooted with the catalytic domain of *OsPong*. Bootstrap values were calculated from 1000 replicates. The structure of *OsPIF* elements is depicted at the right. Black triangles represent element TIRs, colored lines represent noncoding regions, gray boxes represent ORF1s, black boxes represent TPase genes, and open triangles indicate truncation in TPase genes. The positions of intron 1 (pink box) and intron 2 (blue box) are shown. The positions of insertions by other TEs are indicated by gray triangles above the elements (the identity and length of these insertions are described to the right). Dashed lines represent missing regions from elements that are incomplete because of gaps in genomic sequences or rearrangements after insertion (such as deletions or large insertions). The length of elements as well as ORF1 and TPase genes is drawn to scale. *OsPIF3d* and *OsPIF7b* were previously reported as *Os-PIF1* and *Os-PIF2*, respectively (Zhang *et al.* 2001).

by stop codons or frameshifts. Similarly, *OsPong20* includes four elements (~5.3 kb) that are on average >99.7% identical.

In contrast, interfamily sequence conservation, even between closely related families, was restricted to coding regions and TIRs. For example, the nucleotide sequences of *OsPIF9* and *OsPIF7* were 60% identical in their TPases (~1.2 kb) and 55% identical in their ORF1s (~1 kb), but these two families did not share additional sequence similarity aside from their TIRs. Similarly, the




OsPong4d
OsPong4a
OsPong4b
OsPong4c
OsPong5
OsPong2c
OsPong2b
OsPong2a
OsPong3d
OsPong3c
OsPong3b
OsPong3a
Ponga
Pongb
Pongc
Pongd
Ponge
Ping
OsPong7b
OsPong7c
OsPong7a
OsPong6c
OsPong6b
OsPong6a
OsPong17c
OsPong17b
OsPong17a
OsPong16b
OsPong16a
OsPong15c
OsPong15b
OsPong15a
OsPong15d
OsPong8f
OsPong8g
OsPong8e
OsPong8d
OsPong8c
OsPong8b
OsPong8a
OsPong8h
OsPong10b
OsPong10a
OsPong11b
OsPong11a
OsPong23a
OsPong23b
OsPong14
OsPong13
OsPong12
OsPong24
OsPong20f
OsPong20c
OsPong20g
OsPong20d
OsPong20b
OsPong20a
OsPong20h
OsPong20i
OsPong20j
OsPong21
OsPong25
OsPong9a
OsPong9b
OsPong9c
OsPong19d
OsPong19e
OsPong19f
OsPong19c
OsPong19b
OsPong19a
OsPong22
OsPong18g
OsPong18d
OsPong18h
OsPong18c
OsPong18b
OsPong18a
OsPong18e
OsPong18f
PIF
Q
O2
MULE (743 bp)
Retro (*Sz2*, 12,690bp)
Solo LTR (*Sz18*, 961 bp)
Retro (*Sz2*, 7,171 bp)
MULE (2,750 bp) inserted into a Retro (*Bajie*, 4,529 bp)
Solo LTR (*Sz2*, 1,166 bp)
Retro (3,565 bp)
Retro (3,750 bp), Retro (9,885 bp)
Retro (11,990 bp)
Solo LTR (1,394 bp) inserted into solo LTR (1,278 bp)
Retro (*Sz2*, 12,822 bp)
Ac-like element (2,420 bp) inserted into a solo LTR (2,737 bp)
Helitron (6,570 bp)
Retro (7,361 bp)
Retro (2,925 bp)
Retro (7,769 bp)
CACTA (956 bp)
MITE (*Gaijin*, 128 bp)
MITE (*Gaijin*, 128 bp)
MITE (*Buhui*, 375 bp); MITE (*TouristOs6*, 300 bp)
Retro (4,638 bp); Solo LTR (1,757 bp)
Retro (*Sc14*, 8,041 bp)
0.05 changes
1 kb

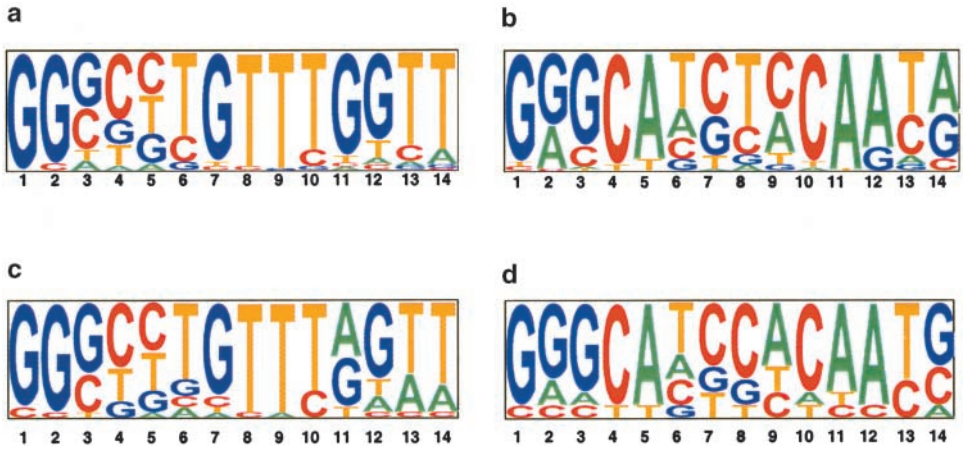


FIGURE 6.—Comparison of the TIRs of rice *PIF*-like elements, *Pong*-like elements, and *Tourist*-like MITEs, shown in the form of pictograms. The terminal 14 bases from both ends are compared. The nucleotide A is shown in green, C in red, G in blue, and T in yellow. Numbers indicate the position of a nucleotide from the element end (*e.g.*, "1" indicates the terminal nucleotide). The height of each nucleotide represents the relative frequency of that nucleotide at that position. (a) Pictogram of the TIRs from 21 *OsPIF* families (71 elements). (b) Pictogram of the TIRs from 20 *OsPong* families (61 elements). (c) Pictogram of the TIRs from 20 previously described rice *Tourist*-like MITE families [*Tourist_Ia*, *Ib*, *Ic*, *III* (*Gaijin*), *IV* (*Castaway*), *V* (*Wanderer*), *VII*, *VIII*, *IX*, *XI*, *XII*, *XIV*, *XV*, *XVI*, *Type C*, *Buhui*, *Casin*, *Centre*, *Stone*, and *Susu*]. (d) Pictogram of the TIRs from eight previously described rice *Tourist*-like MITE families (*Tourist_VI*, *Helia*, *Qiqi*, *ID-2*, *ID-3*, *ID-4*, *Lier*, *Stola*, and *Youren*; BUREAU *et al.* 1996; TARCHINI *et al.* 2000; JIANG and WESSLER 2001; TURCOTTE *et al.* 2001).

OsPong2 and *OsPong3* families share 81 and 85% nucleotide identity in their ORF1 (~900 bp) and TPase (~1.3 kb) coding regions, respectively, but have completely diverged noncoding regions.

Coevolution of ORF1 and TPase in OsPongs: All 21 *OsPIF* families and 19 of the 20 *OsPong* families with defined termini encoded both ORF1 and TPase. The only *OsPong* family with defined termini that does not harbor an ORF1 is *OsPong18*, where all eight elements contain only TPase. Absence of ORF1 in these elements is likely due to internal deletion for two reasons. First, the length of these elements is unusually short, ranging from 1365 to 2745 bp. Second, Blastn searches using *OsPong18* elements as queries identified several additional family members that do not encode TPase but contain coding sequence for ORF1 (*e.g.*, AP003799; 96,105–99,317). Thus, all *OsPong* families with defined ends encode both ORF1 and TPase.

As mentioned above, ORF1 and TPase in *OsPIFs* are arranged in three different alignments and those in *OsPongs* have two different alignments. Interelement recombination has been shown to be a significant force in the evolution of mobile elements (*e.g.*, ADEY *et al.* 1994; MCCLURE 1996; JORDAN and MCDONALD 1998; LERAT *et al.* 1999). The presence of two ORFs whose organization varies in different *PIF/Pong*-like elements prompted us to investigate whether interelement recombination had contributed to the evolution of these elements. To address this question, the phylogenies of ORF1 and the TPase of *OsPIF* and *OsPong* elements were compared to determine whether ORFs in the same element were coevolving or whether discrepancies existed that might suggest independent evolution.

Two phylogenetic trees were generated for *OsPong* elements, one based on an ~110-aa region in their ORF1s including all three conserved blocks (Figure 1b, boxed region) and one based on the catalytic domains of their TPases (Figure 7). Comparison of the two trees showed that the phylogenies determined from the two coding sequences were consistent. These results indicate that interelement recombination had probably not occurred in *OsPongs*. Similarly, the phylogeny of *OsPIF* ORF1s was found to be consistent with that of the TPases, albeit with less bootstrap support (data not shown).

Insertion sites of OsPIFs and OsPongs: Prior studies have shown that some plant DNA transposons, such as members of the *Ac/Ds* and *Mutator* families, have a preference for insertion into single-copy regions of the genome (CHEN *et al.* 1987; CRESSE *et al.* 1995; DIETRICH *et al.* 2002). Similarly, in a recent study it was shown that the majority of new *Pong* insertions (9 of 10) were into single-copy sequences of the rice genome (JIANG *et al.* 2003). To test whether this is also true for *OsPIFs*

FIGURE 5.—Phylogeny of *OsPong* TPases and the structure of the encoding element. The neighbor-joining tree on the left was constructed from a CLUSTALW multiple alignment of the catalytic domains (boxed region in Figure 1c) of 80 *OsPong* TPases and rooted with the catalytic domain of the maize *PIF*. Bootstrap values were calculated from 1000 replicates. Structure of *OsPong* elements is shown at the right. Black triangles represent element TIRs, colored lines represent noncoding regions, gray boxes represent ORF1s, black boxes represent ORF2s, and open triangles indicate truncations in ORF1s or ORF2s. The positions of insertions by other TEs are indicated by gray triangles above the elements (the identity and length of these insertions are described to the right). Dashed lines represent missing regions from elements that are incomplete because of gaps in genomic sequences or rearrangements after insertion (such as deletions or large insertions). The length of elements as well as ORF1 and TPase genes is drawn to scale.

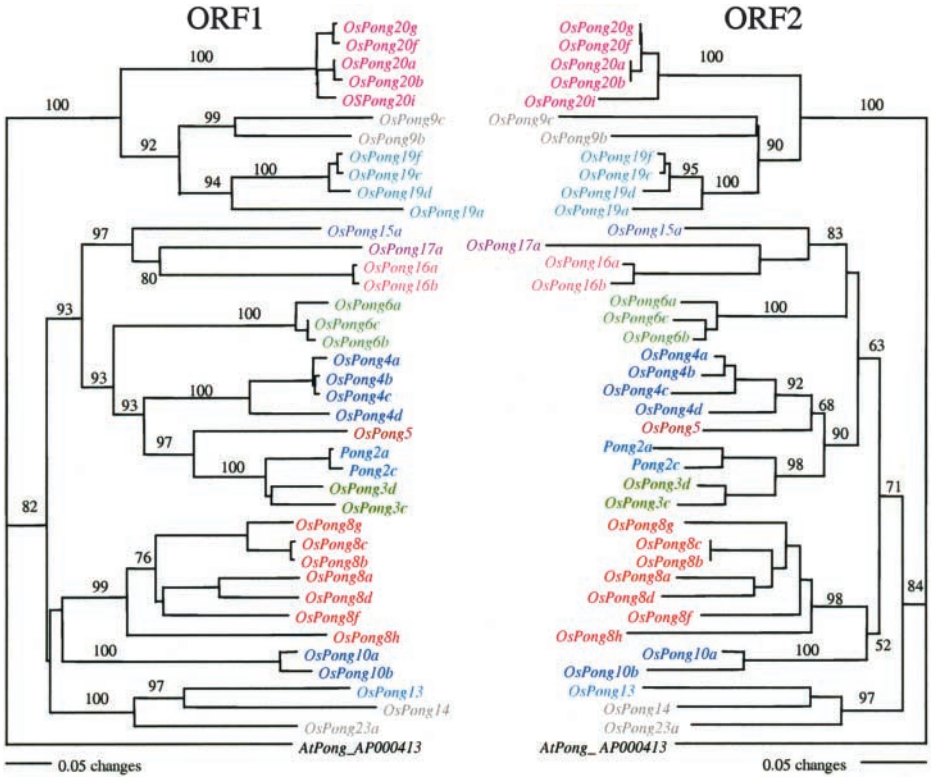


FIGURE 7.—Phylogenetic trees of *OsPong* ORF1 and *OsPong* ORF2. Trees were generated using the neighbor-joining method from CLUSTALW multiple alignments of the conserved regions in *OsPong* ORF1 (boxed in Figure 1b) and the catalytic domain of ORF2 (boxed in Figure 1b), respectively, and rooted with the corresponding regions of ORF1 and ORF2 from an Arabidopsis *Pong*-like element *AtPong_AP000413* (accession no. AP000413; ORF1, 34,047–34,412; ORF2, 32,091–32,453). Where multiple *OsPong* elements are identical in these regions, only one is shown. The name and color of *OsPongs* are the same as in Figure 6. Bootstrap values were calculated from 1000 replicates.

and other *OsPongs*, the immediate flanking sequences (100 bp from each end) of all elements with defined termini were searched using RepeatMasker and Blastn. Of the 132 *OsPIFs* and *OsPongs* examined, 109 (82.6%) were in single-copy sequences, 13 (9.8%) were in other DNA elements, 6 (5.5%) were in retroelements, and 4 (3.0%) were in unknown repeats.

Relationships between OsPIFs, OsPongs, and Tourist-like MITEs: Identification of full-length *OsPIF* and *OsPong* elements permitted the first genome-wide analysis of the relationship between these TPase-encoding elements and *Tourist*-like MITEs. Sequence identities between *OsPIFs* and *OsPongs* and *Tourist*-like MITEs were determined in two ways. First, we investigated whether *Tourist*-like MITE families could be associated with *OsPIFs* or *OsPongs* on the basis of the sequences of their TIRs. To do this, the TIRs of 31 published rice *Tourist*-like MITE families were examined (Bureau *et al.* 1996; Tarchini *et al.* 2000; Jiang and Wessler 2001; Turcotte *et al.* 2001). The TIRs of the majority of MITE families (28 of 31) were of two types: one (20 families) was strikingly similar to the TIRs of *OsPIFs* (Figure 6c) and the other (9 families) was more similar to *OsPongs* (Figure 6d). Second, more extensive sequence similarity between individual *OsPIF* or *OsPong* families and *Tourist*-like MITEs was examined. Terminal sequences (200 bp from each end) from the *OsPIF* and *OsPong* families with defined ends were used as queries to search a rice repetitive sequence database (N. Jiang, Z. Bao, S. R. Eddy and S. R. Wessler, unpublished data). Significant nucleotide similarity that extended beyond the TIRs was taken as evidence that an *OsPIF* or *OsPong* family was associated with a *Tourist*-like MITE family. Nine of the 21 *Tourist*-like MITE families with *OsPIF*-like TIRs were found to be associated with *OsPIF* families, while 3 of the 9 *Tourist*-like MITE families with *OsPong*-like TIRs were found to be associated with *OsPong* families.

In some cases a MITE family was clearly identified as a deletion derivative of a particular *OsPIF* or *OsPong* family (Figure 8). For example, the high copy number *Castaway* family (~3000 copies; Bureau *et al.* 1996; Jiang and Wessler 2001) appears to be derived by a simple deletion from the *OsPIF6* family. The apparent deletion breakpoints in *OsPIF6* occur at a 4-bp direct repeat (TTCC, underlined in Figure 8a) that is present as only a single copy in *Castaway*. Similar relationships between several other *OsPIF* families and *Tourist*-like MITE families are shown in Figure 8, b–d. In contrast, although sequence similarities between *OsPong* and *Tourist*-like MITEs were detected, they were not as extensive as those seen with *OsPIF* (see Figure 8, e and f).

DISCUSSION

***PIF*- and *Pong*-like elements are widespread and abundant:** Here we present the first comprehensive analysis of *PIF*- and *Pong*-like elements in eukaryotes. Prior studies reported that the TPases of *PIF* and *Pong* were distantly related to the bacterial IS*5* TPases and noted the presence of *PIF*-like elements in several plant (rice, sorghum, and Arabidopsis), animal (*C. elegans*, *C. briggsae*, and fugu fish) and fungal (*Filobasisiella neoformans*)

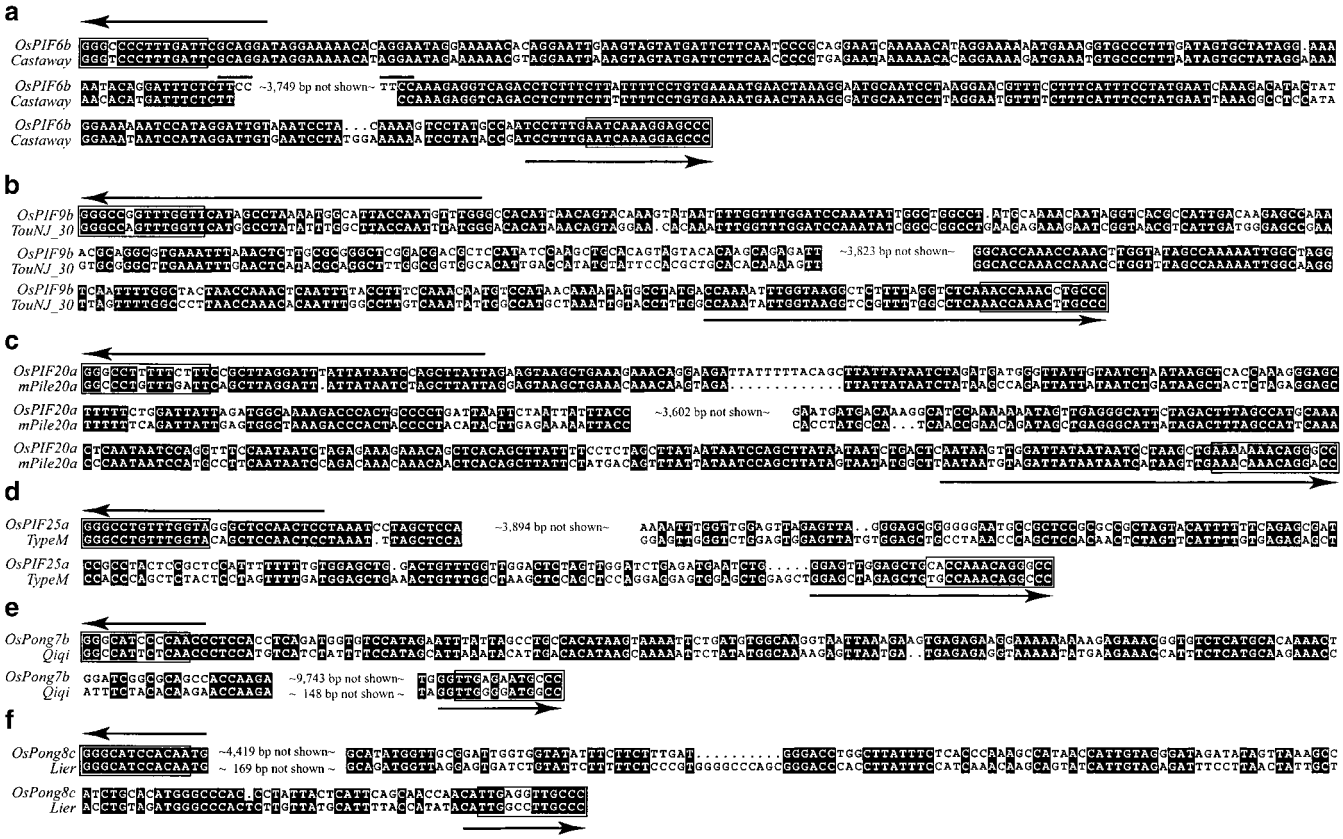


FIGURE 8.—Sequence similarity between *OsPIF*- (a–d) or *OsPong*-like elements (e and f) and *Tourist*-like MITEs in rice. Each alignment represents an example of an *OsPIF* or *OsPong* family that shares significant sequence similarity with a *Tourist*-like MITE. *TouNJ-30* and *mPile10* were identified in this study, while other MITEs were previously reported. Arrows indicate element TIRs, open boxes indicate the regions used to generate the pictograms in Figure 6, and horizontal lines denote the location of direct repeats flanking the deletion breakpoints (discussed in text).

genomes (KAPITONOV and JURKA 1999; LE *et al.* 2001; ZHANG *et al.* 2001; APARICIO *et al.* 2002; JIANG *et al.* 2003). In this study, >600 *PIF*- and *Pong*-like TPases were identified or isolated from 35 plants, 19 animals, and two fungi. Phylogenetic analyses of these TPases defined three major groups, each represented by multiple distinct lineages. Taken together, the *PIF*/IS*5* superfamily of TPases is ancient and its members are widespread. To date, the only other TPase superfamily with such a broad distribution in eukaryotes and prokaryotes is IS630/*Tc1*/*mariner* (FESCHOTTE and WESSLER 2002; ROBERTSON 2002).

PIF- and *Pong*-like elements are especially abundant in plants, including both monocotyledons and dicotyledons. Large numbers of *PIF*- and *Pong*-like TPases were detected in three plants with relatively small genomes: ~80 copies in Arabidopsis (130 Mb); ~350 copies in rice (450 Mb), and >1000 copies in *B. oleracea* (extrapolated to ~600 Mb). Although significant sequence is not yet available for plants with large genomes, such as maize (2500 Mb) and barley (5000 Mb), the degenerate PCR assay indicates that these genomes also harbor multiple and diverse lineages (Figure 3). Given that amplification of transposable elements is largely responsible for the huge differences in plant genome size (BENNETZEN 2002; FESCHOTTE *et al.* 2002a), it is reasonable to assume that even larger families of *PIF* and *Pong* will be found once these genomes are sequenced.

ORF1 of *PIF*/*Pong*-like elements: The maize *PIF* and rice *Pong* elements encode two ORFs (ORF1 and ORF2), of which ORF2 is most likely the TPase while the function of ORF1 is unknown. Database searches revealed a large number of homologs for both ORFs and, where found, they were usually in pairs with an ORF1 homolog located within ~1–2 kb of an ORF2 homolog. All *OsPIF* and *OsPong* families with defined termini also encoded both ORFs. The amino acid sequence similarity between *PIF*- and *Pong*-like ORF1s in blocks A and B (Figure 1) suggests a monophyletic origin, and the presence of this ORF in virtually all *PIF*/*Pong*-like lineages suggests that it is necessary for the active transposition of these elements.

A requirement for a protein other than the transposase is unusual for a eukaryotic transposon, having been described previously only for members of the CACTA superfamily (KUNZE and WEIL 2002). Although the au-

tonomous *Mutator* element *MuDR* from maize encodes two proteins (MURA and MURB), it is so far the only *Mutator* element shown to encode more than a single ORF among hundreds of MULEs examined (Yu *et al.* 2000; Singer *et al.* 2001; Lisch 2002). For CACTA-like elements, multiple proteins are encoded by alternatively spliced transcripts (*e.g.*, TNPA and TNPD of the maize *En/Spm* element; Kunze and Weil 2002). In contrast, our data indicate that ORF1 and ORF2 are separate transcription units. First, each ORF has a promoter that was predicted with high confidence by computer programs (data not shown). Second, elements harboring four distinct alignments of ORFs 1 and 2 were detected in plant genomes: head-to-tail (ORF2 either upstream or downstream of ORF1), head-to-head, and tail-to-tail (Figure 3). The fact that ORFs 1 and 2 would be transcribed from opposite strands in the head-to-head or tail-to-tail arrangements rules out the possibility that alternatively spliced transcripts are involved.

Several features of ORF1 provide clues to its possible function(s). Weak similarity between the most conserved region in ORF1 (Figure 1, block A) and the *myb* DNA-binding domain of some plant and animal transcription factors suggests that ORF1 may encode a DNA-binding protein (Jiang *et al.* 2003). One can envision a model whereby the product of ORF1 binds to the ends of *Pong*-like elements and recruits ORF2 by protein-protein interactions. Alternatively, products of ORFs 1 and 2 may form a heterodimer that binds to the element ends. If the products of ORFs 1 and 2 interact, it is reasonable to expect that they have been coevolving, a feature consistent with the data presented in this study (Figure 7). That is, the phylogenies of ORF1 and ORF2 are very consistent and no interfamily rearrangement was found (Figure 7). The requirements of *PIF* and *Pong* transposition as well as the possible interaction between ORF1 and ORF2 are under investigation.

***OsPIF* and *OsPong* elements:** This study identified 116 *OsPIF* and 80 *OsPong* TPases representing all of the lineages of *PIF* and *Pong* TPases detected in monocot genomes. As such, rice is a suitable model to study the evolution of *PIF*- and *Pong*-like elements in plants as well as their relationship with *Tourist*-like MITEs.

OsPIF and *OsPong* elements were grouped into 27 and 26 families, respectively, on the basis of sequence identity of their coding regions. These groupings received additional support when it was determined that elements of the same family share extensive sequence similarity in noncoding regions. Several *OsPIF* and *OsPong* families (such as *OsPIF4*, *-5*, *-6*, *-9*, *-12*, *-13*, *-23* and *OsPong8*, *-17*, *-19*, *-20*) include members that are nearly identical. Furthermore, each family includes at least one putative autonomous member whose coding region is not interrupted by a stop codon or a frameshift mutation. These features are indicative of recent and perhaps ongoing activity of multiple *OsPIF/OsPong* families.

***OsPIF*, *OsPong*, and *Tourist*-like MITEs:** In previous studies, *PIF* and *Pong* elements were isolated as the TPase sources for two families of *Tourist*-like MITEs, *mPIF* and *mPing*, respectively (Zhang *et al.* 2001; Jiang *et al.* 2003). In this study, many other *PIF*- and *Pong*-like elements were identified, including all the families in rice. Characterization of these elements permitted a comprehensive analysis of the relationships between these TPase-encoding elements and *Tourist*-like MITEs. The major conclusion from this analysis is that most *Tourist*-like MITE families are related to either *PIF*- or *Pong*-like elements solely on the basis of a comparison of their TIRs. Of the 31 previously described *Tourist*-like MITE families in rice, the TIRs of 20 were found to be more closely related to the consensus *OsPIF* TIR while the TIRs of 9 were more closely related to the consensus *OsPong* TIR (Figure 6).

Attempts to associate individual *Tourist*-like MITE families with specific *OsPIF* or *OsPong* families uncovered many clear-cut relationships (Figure 8). For example, the MITE family *Castaway* (~3000 copies) was found to be derived from the *OsPIF6* family by internal deletion and subsequent amplification. Relationships between *Tourist*-like MITEs and *OsPong* families are less apparent as sequence similarity is limited to the subterminal regions (as shown in Figure 8, e and f). However, detection of sequence identity between subterminal regions of *OsPong* families and *Tourist*-MITE families, albeit limited, is significant in light of the fact that even closely related *OsPong* families display no sequence identity in their subterminal regions.

In summary, the characterization of virtually all full-length *PIF*- and *Pong*-like elements in the rice genome has permitted a determination of the extent of their relatedness with most of the 60,000 *Tourist*-like MITEs residing in this genome. Our data indicate that many *Tourist*-like MITEs originated from *OsPIF* and *OsPong* elements by internal deletion and subsequent amplification. However, 16 of the 28 *Tourist*-like MITEs examined in this study were not clearly associated with *OsPIF/OsPong* families. It is possible that their cognate *OsPIF/OsPong* families were lost from the genome. Such a scenario is not difficult to imagine considering *OsPIF/OsPong* elements are present at much lower copy number (several per family) than are *Tourist*-like MITEs (hundreds or thousands per family). Alternatively, some *Tourist*-like MITE families may have originated by chance events, where, for example, a pair of nearby inverted repeats (and other *cis* requirements, if any) were mobilized fortuitously by an endogenous *PIF*- or *Pong*-like TPase and subsequently amplified to high copy numbers.

The rice genome harbors >90,000 MITEs: 60,000 *Tourist*-like MITEs and 30,000 *Stowaway*-like MITEs (Feschotte *et al.* 2003; N. Jiang and S. R. Wessler, unpublished data). Whereas elements of the *PIF/Pong/IS5* superfamily (*OsPIF* and *OsPong*) appear to be re-

sponsible for the origin and amplification of *Tourist*-like MITEs, rice elements related to the *Tc1/mariner* superfamily (called *Osmars*) have been associated with *Stowaway*-like MITEs (Feschotte *et al.* 2003). Given that plant genomes are known to harbor many other DNA transposon families (including CACTA-like, *hAT*-like, and MULEs), one obvious question is, Why are most MITE families associated with only these two superfamilies? A key factor may be the *cis* requirements for transposition. Transposition of MULEs requires long TIRs (>200 bp; Benito and Walbot 1997), while that of CACTA-like and *hAT*-like elements requires both TIRs and subterminal repetitive motifs (reviewed in Kunze and Weil 2002). Several studies have shown that, for these elements, multiple TPase-binding sites reside in the subterminal repeats. In contrast, the binding sites for animal *mariner* TPases seem to be restricted to the short element TIRs (~28–31 bp; Auge-Gouillou *et al.* 2001; Lampe *et al.* 2001; Zhang *et al.* 2001). Indeed, artificial transposons containing just the *mariner* TIRs have been successfully mobilized by the cognate *mariner* TPase, *in vitro* and *in vivo* in bacteria (*e.g.*, Lampe *et al.* 1999, 2001; Tosi and Beverley 2000). Although the *cis* requirements for transposition of *PIF*- and *Pong*-like elements have yet to be determined, these elements do not contain any apparent repetitive motifs in subterminal regions, suggesting that their TPase-binding sites may also reside in their short TIRs. Thus, the minimal *cis* requirements for transposition by members of the *mariner*-like and *PIF/Pong*-like families may significantly enhance the probability of generating shorter elements by deletion from larger elements or by chance, which could be mobilized by the TPases encoded by these families.

Concluding remarks: The *PIF* and *Pong* elements are founding members of a very large and dynamic superfamily of class 2 elements that are widespread in flowering plants. The impact of these elements is significant, as *PIF*- and *Pong*-like families are capable of expansion through the amplification and diversification of both autonomous and nonautonomous members, including very-high-copy-number MITEs. Furthermore, with a demonstrated preference for insertion into genic regions, *PIF*- and *Pong*-like elements and their associated *Tourist*-like MITEs appear to be a major force generating genetic diversity and influencing the evolution of plants.

This work was supported by grants from the National Science Foundation and National Institutes of Health to S.R.W.

LITERATURE CITED

Adey, N. B., S. A. Schichman, D. K. Graham, S. N. Peterson, M. H. Edgell *et al.*, 1994 Rodent l1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. Mol. Biol. Evol. **11:** 778–789.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J. M. Chia *et al.*, 2002 Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science **297:** 1301–1310.

Auge-Gouillou, C., M. H. Hamelin, M. V. Demattei, M. Periquet and Y. Bigot, 2001 The wild-type conformation of the *Mos-1* inverted terminal repeats is suboptimal for transposition in bacteria. Mol. Genet. Genomics **265:** 51–57.

Benito, M. I., and V. Walbot, 1997 Characterization of the maize *Mutator* transposable element MURA transposase as a DNA-binding protein. Mol. Cell. Biol. **17:** 5165–5175.

Bennetzen, J. L., 2002 Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29–36.

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.

Chen, J., I. Greenblatt and S. Dellaporta, 1987 Transposition of *Ac* from the *P* locus of maize into unreplicated chromosomal sites. Genetics **117:** 109–116.

Cresse, A. D., S. H. Hulbert, W. E. Brown, J. R. Lucas and J. L. Bennetzen, 1995 *Mu1*-related transposable elements of maize preferentially insert into low copy number DNA. Genetics **140:** 315–324.

Dietrich, C. R., F. Cui, M. L. Packila, J. Li, D. A. Ashlock *et al.*, 2002 Maize *Mu* transposons are targeted to the 5′ untranslated region of the *gl8* gene and sequences flanking *Mu* target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics **160:** 697–716.

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.

Feschotte, C., L. Swamy and S. R. Wessler, 2003 Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with *Stowaway* MITEs. Genetics **163:** 747–758.

Gilbert, W., S. J. de Souza and M. Long, 1997 Origin of genes. Proc. Natl. Acad. Sci. USA **94:** 7698–7703.

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.

Hebsgaard, S. M., P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze *et al.*, 1996 Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. Nucleic Acids Res. **24:** 3439–3452.

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.

Jordan, I. K., and J. F. McDonald, 1998 Evidence for the role of recombination in the regulatory evolution of *Saccharomyces cerevisiae Ty* elements. J. Mol. Evol. **47:** 14–20.

Kapitonov, V. V., and J. Jurka, 1999 Molecular paleontology of transposable elements from *Arabidopsis thaliana*. Genetica **107:** 27–37.

Kellogg, E. A., 2001 Evolutionary history of the grasses. Plant Physiol. **125:** 1198–1205.

Kikuchi, K., K. Terauchi, M. Wada and H. Y. Hirano, 2003 The plant MITE *mPing* is mobilized in anther culture. Nature **421:** 167–170.

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., B. J. Akerley, E. J. Rubin, J. J. Mekalanos and H. M. Robertson, 1999 Hyperactive transposase mutants of the *Himar1 mariner* transposon. Proc. Natl. Acad. Sci. USA **96:** 11428–11433.

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., K. Turcotte and T. Bureau, 2001 *Tc8*, a *Tourist*-like transposon in *Caenorhabditis elegans*. Genetics **158:** 1081–1088.

Lerat, E., F. Brunet, C. Bazin and P. Capy, 1999 Is the evolution of transposable elements modular? Genetica **107:** 15–25.

Lisch, D., 2002 *Mutator* transposons. Trends Plant Sci. **7:** 498–504.

Logsdon, Jr., J. M., 1998 The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. **8:** 637–648.

Mahillon, J., and M. Chandler, 1998 Insertion sequences. Microbiol. Mol. Biol. Rev. **62:** 725–774.

McClure, M. A., 1996 The complexities of viral genome analysis: the primate lentiviruses. Curr. Opin. Genet. Dev. **6:** 749–756.

Nakazaki, T., Y. Okumoto, A. Horibata, S. Yamahira, M. Teraishi *et al.*, 2003 Mobilization of a transposon in the rice genome. Nature **421:** 170–172.

Rezsohazy, R., B. Hallet, J. Delcour and J. Mahillon, 1993 The IS4 family of insertion sequences: evidence for a conserved transposase motif. Mol. Microbiol. **9:** 1283–1295.

Robertson, H. M., 2002 Evolution of DNA transposons, pp. 1093–1110 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, DC.

Singer, T., C. Yordan and R. A. Martienssen, 2001 Robertson's *Mutator* transposons in *A. thaliana* are regulated by the chromatin-remodeling gene *Decrease in DNA Methylation* (*DDM1*). Genes Dev. **15:** 591–602.

Swofford, D. L., 1999 *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods*. Sinauer, Sunderland, MA.

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.

Tosi, L. R., and S. M. Beverley, 2000 *cis* and *trans* factors affecting *Mos1 mariner* evolution and transposition in vitro, and its potential for functional genomics. Nucleic Acids Res. **28:** 784–790.

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

Walker, E. L., W. B. Eggleston, D. Demopulos, J. Kermicle and S. L. Dellaporta, 1997 Insertions of a novel class of transposable elements with a strong target site preference at the *r* locus of maize. Genetics **146:** 681–693.

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.

Yu, Z., S. I. Wright and T. E. Bureau, 2000 *Mutator*-like elements in *Arabidopsis thaliana*. Structure, diversity and evolution. Genetics **156:** 2019–2031.

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.

Zhang, X., C. Feschotte, Q. Zhang, N. Jiang, W. B. Eggleston *et al.*, 2001 *P Instability Factor*: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. Proc. Natl. Acad. Sci. USA **98:** 12572–12577.

Communicating editor: M. J. Simmons