DOI: 10.1534/genetics.106.062752

# The Transposable Element Landscape of the Model Legume *Lotus japonicus*

**Dawn Holligan,* Xiaoyu Zhang,*[,1] Ning Jiang,† Ellen J. Pritham*[,2] and Susan R. Wessler*[,3]**

**Department of Plant Biology, University of Georgia, Athens, Georgia 30602 and* †*Department of Horticulture, Michigan State University, East Lansing, Michigan 48824*

Manuscript received June 28, 2006
Accepted for publication September 18, 2006

ABSTRACT

The largest component of plant and animal genomes characterized to date is transposable elements (TEs). The availability of a significant amount of *Lotus japonicus* genome sequence has permitted for the first time a comprehensive study of the TE landscape in a legume species. Here we report the results of a combined computer-assisted and experimental analysis of the TEs in the 32.4 Mb of finished TAC clones. While computer-assisted analysis facilitated a determination of TE abundance and diversity, the availability of complete TAC sequences permitted identification of full-length TEs, which facilitated the design of tools for genomewide experimental analysis. In addition to containing all TE types found in previously characterized plant genomes, the TE component of *L. japonicus* contained several surprises. First, it is the second species (after *Oryza sativa*) found to be rich in Pack-MULEs, with >1000 elements that have captured and amplified gene fragments. In addition, we have identified what appears to be a legume-specific MULE family that was previously identified only in fungal species. Finally, the *L. japonicus* genome contains many hundreds, perhaps thousands of Sireviruses: *Ty1/copia*-like elements with an extra ORF. Significantly, several of the *L. japonicus* Sireviruses have recently amplified and may still be actively transposing.

TRANSPOSABLE elements (TEs) are the single largest output of genome sequencing projects, accounting for almost 50% of the human genome (Lander *et al.* 2001) and ~10 and 30% of the sequenced *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000) and *Oryza sativa* (rice) genomes, respectively (Turcotte *et al.* 2001; Goff *et al.* 2002; Jiang *et al.* 2004b). Partial sequencing of other genomes indicates that TEs account for >75% of the larger plant genomes, including maize and barley (SanMiguel *et al.* 1996; Kumar and Bennetzen 1999; Vicient *et al.* 1999). The fact that TEs represent a huge fraction of the genomes of multicellular organisms means that computer-assisted analyses of even partial genome sequencing projects can be informative. Such analyses have revealed significant features of genomewide TE content, including the composition of TE types and their evolutionary trajectory for some species in the Brassicaceae (Zhang and Wessler 2004).

Eukaryotic TEs are divided into two classes, according to whether their transposition intermediate is RNA (class 1) or DNA (class 2) (Capy *et al.* 1998; Craig *et al.* 2002). Each TE class contains coding and noncoding elements (also called autonomous and nonautonomous elements). Coding elements have complete or partial open reading frames (ORFs) that encode products involved in the transposition reaction. Noncoding elements do not encode transposition-associated proteins but can be mobilized because they retain the *cis*-sequences necessary for transposition. Integration of most TEs results in the duplication of a short genomic sequence (called target-site duplication, or TSD) at the site of insertion (Feschotte *et al.* 2002a).

Coding class 1 elements include retrotransposons with long terminal repeats (LTR retrotransposons) and non-LTR retrotransposons, also called long interspersed elements (LINEs). The most prevalent TE type in plant genomes, LTR retrotransposons, has been further classified as either *Ty1/copia*-like or *Ty3/gypsy*-like on the basis of the order of their encoded proteins that include reverse transcriptase (RT) and integrase (Kumar and Bennetzen 1999). All coding class 2 elements, except *Helitrons*, have short terminal inverted repeats (TIRs) and are grouped into superfamilies on the basis of the similarity of their encoded transposes [*e.g.*, *Tc1/mariner*, *hAT*, CACTA, *Mutator*-like elements (MULEs), and *PIF/Pong*], the enzyme that binds to the TE ends and catalyzes both excision and insertion (Capy *et al.* 1998; Craig *et al.* 2002; Zhang *et al.* 2004). Another class 2 TE

[1] *Present address:* Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA 90095.

[2] *Present address:* Department of Biology, University of Texas, Arlington, TX 76019.

[3] *Corresponding author:* Department of Plant Biology, 2502 Miller Plant Sciences Bldg., University of Georgia, Athens, GA 30602. E-mail: sue@plantbio.uga.edu

type, miniature-inverted repeat transposable elements (MITEs) are noncoding elements that are frequently associated with plant genes but are also found in animals including zebrafish and *Caenorhabditis elegans* (Feschotte *et al.* 2002b).

Given all that is known about TEs, why is it necessary to continue to characterize them in newly emerging sequence databases? First, TE content has been shown to vary dramatically across diverse taxa (Kidwell 2002). Because no two genomes are alike, each has a different story to tell. For example, vertebrates are dominated by class 1 TEs (largely non-LTR retrotransposons) and only fish are known to have active class 2 elements (Koga and Hori 2001). In contrast, plant genomes have a wealth of class 1 and class 2 elements including several active TE families (Yamazaki *et al.* 2001; Jiang *et al.* 2003). With a variety of genomes characterized and more to come, TE biologists can choose the best organism for studying a particular TE type. Second, comparative analysis of the TE content of related taxa provides data to build models of species divergence. For example, the >20-fold difference in genome size since the divergence of the grass clade over ∼70 million years ago can be explained in large part by the amplification of TEs, especially LTR retrotransposons (SanMiguel and Bennetzen 1998; Vicient *et al.* 1999; Jiang and Wessler 2001; Meyers *et al.* 2001). Third, from a practical point of view, knowledge of TE content is essential for correct genome annotation. This is especially important in light of recent findings that two plant TE types, Pack-MULEs and *Helitrons*, routinely capture and amplify gene fragments, thus further confounding genome annotation (Jiang *et al.* 2004a; Gupta *et al.* 2005; Lai *et al.* 2005).

The approaches used to identify TEs and the information that can be attained are greatly influenced by the characteristics of the sequence database. The availability of complete genome sequences for *O. sativa* and *A. thaliana* will ultimately permit the identification of full-length and fragmented TEs along with their chromosomal locations. In contrast, for a partial database consisting of short reads (such as *Brassica oleracea*), only coding TEs can be identified and their copy numbers can be approximated only by extrapolation to the whole genome (Zhang and Wessler 2004). The *Lotus japonicus* (*L. japonicus*) database, which is the focus of this study, represents a third type, where long transformation-competent bacterial artificial chromosome (TAC) sequences covering a significant fraction of the genome are available, but their chromosomal positions are yet to be determined.

*L. japonicus* belongs to the Fabaceae family, one of the largest plant families, with several agronomically important species (Young *et al.* 2003). It is an ideal model legume because of its small genome size (∼472 Mb), relatively short life cycle (2–3 months), and the ease of genetic manipulation (*e.g.*, ease of transformation, self compatible) (Jiang and Gresshoff 1997). For these reasons, a large-scale sequencing project was initiated for the *L. japonicus* accession Miyakojima (MG-20), and a subset of genomic sequence is now available (Sato *et al.* 2001; Nakamura *et al.* 2002; Asamizu *et al.* 2003; Kaneko *et al.* 2003; Kato *et al.* 2003). The available database represents ∼50% of the euchromatic (gene-rich) regions and includes ∼32.4 Mb of finished sequence (∼443 finished TACs) and 94 Mb of phase 1 sequence (Young *et al.* 2005).

Here we report the results of a combined computer-assisted and experimental analysis of the TEs in the 32.4 Mb of finished TACs (sequence available at time of study). Computer-assisted analysis provided information on the abundance (copy number), diversity (lineages), and temporal features of TE amplification for all major TE types. In addition to containing all TE types found in previously characterized plant genomes, *L. japonicus* is only the second species found to be rich in Pack-MULEs. As mentioned above, one reason for continuing to analyze TEs in genomes is to be able to identify unusual elements and/or previously described elements that may still be active. In this regard, our analysis has been particularly satisfying as we have identified what appears to be a legume-specific MULE family that was previously identified only in fungal species. In addition, we found lineages of *Ty1/copia*-like elements with an extra ORF that have recently amplified and may still be actively transposing.

## MATERIALS AND METHODS

**Plant material and DNA extraction:** Miyakojima (MG-20) and Gifu (B-129) ecotypes were obtained from the National Agricultural Research Center for the Hokkaido Region of Japan. Genomic DNA was extracted from leaves of 4-week-old seedlings from six individual plants each of MG-20 and B-129 and purified using the DNAeasy plant mini kit (QIAGEN, Chatsworth, CA).

**Transposon display:** Transposon display was carried out as described (Casa *et al.* 2000) with the following modifications. Element-specific primers were designed on the basis of the consensus subterminal sequences of CACTA, *Pong*, MULE, *copia*, and *gypsy*. Final annealing temperature for selective amplification was 56° with the $^{33}$P-labeled primer for all P2 primers except for *copia* P2, which was 45°. Primer sequences were the following: *Bfa*I+0, 5′-GACGATGAGTCCTGAGTAG-3′; *Bfa*1+T, 5′-GACGATGAGTCCTGAGTAGT-3′; CACTA P1, 5′-AAATGTTGTTGCGAAAAAGTCGCTG-3′; CACTA P2, 5′-CGCTGCGAATTAACTCATCTC-3′; *Pong* P1, 5′-CTTKAAGGCTCTCTCCAATG-3′; *Pong* P2, 5′-GGTCTTAGCAACTCCAG-3′; MULE P1, 5′-AAAGGAGATGGCGGACTTAGC-3′; MULE P2, 5′-AGATGGCGGACTTAGCAAAACAG-3′; *copia* P1, 5′-GAGAATAAATCTCCTAATACTG-3′; *copia* P2, 5′-CTCCTAATACTGAATATAATMTTC-3′; *gypsy* P1, 5′-GCAAAGCGTTTTCTCAAAAGGAC-3′; and *gypsy* P2, 5′-TCTAAACTTCCTTTAGTCGAAC-3′.

**TE insertion polymorphism:** To test whether the polymorphisms on transposon display gels were due to TE insertion or restriction site polymorphism, gel bands were excised, reamplified, and cloned as described (Casa *et al.* 2000). Sequences of cloned fragments were determined by the Molecular

Genetics Instrumentation Facility (University of Georgia). PCR was then performed with primer pairs designed to amplify regions containing the flanking sequence and the element, to verify whether the insertion site was indeed polymorphic. Primer sequences are available upon request.

**Database search strategies:** The available 32.4 Mb of *L. japonicus* genome sequence was downloaded from GenBank at the NCBI database (http://www.ncbi.nlm.nih.gov) and from the *L. japonicus* database (http://www.kazusa.or.jp/lotus). The regions of the *L. japonicus* sequenced were enriched for genes (genomic clones containing ESTs and cDNAs) and as such are not necessarily representative of the entire genome (Sato *et al.* 2001). The following procedure was used to identify TE coding sequences from *L. japonicus.* For each TE type (*e.g.*, CACTA, MULE, *copia*), a consensus sequence based on the most conserved coding region of the previously described *A. thaliana* elements (Zhang and Wessler 2004) was used as a query in TBLASTN searches against the *L. japonicus* sequences. Full-length class 2 elements and LTR retrotransposons were identified using NCBI-BLAST 2 SEQUENCES (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi) on the NCBI server, where 10 kb upstream and downstream of the coding region of each homolog was blasted against itself to define the TIR or the LTR. Comparison of two or more closely related elements served to identify the ends of LINEs on the basis of 5′ homology and poly(A) tail. Complete elements of all types were verified manually by identification of the target-site duplication. To identify noncoding TEs, full-length elements were combined into a repeat library file and used as input for RepeatMasker analysis (version 07/07/2001, using default parameters; http://repeatmasker.genome.washington.edu/RM/RepeatMasker.html) of the *L. japonicus* sequences. RepeatMasker was reiterated until no new TEs were identified. A comprehensive repeat library was then generated and all the *L. japonicus* TEs were used to mask the 32.4 Mb of sequences. The remaining unmasked portion was then subjected to RECON (version 1.03) (Bao and Eddy 2002) to identify novel or undetected TEs missed by RepeatMasker. TE sequences for *L. japonicus* are available upon request and will be available shortly at The Institute for Genomic Research (TIGR) plant repeat database (http://www.tigr.org/tdb/e2k1/plant.repeats/). Transmembrane domain predictions for ORFX were performed using TMpred (http://www.ch.embnet.org/software/TMPRED_form.html), PHDhtm (Rost *et al.* 1995), and TMHMM v.2.0 (http://www.cbs.dtu.dk/services/TMHMM/). Motifs were searched using MotifScan and InterProScan (http://www.expasy.ch/prosite/). Multicoil (Wolf *et al.* 1997) and Paircoil2 (McDonnell *et al.* 2006) (http://multicoil.lcs.mit.edu/cgi-bin/multicoil; http://paircoil2.csail.mit.edu/, respectively) were used to detect coil-coil domains. All ORFs were identified using the ORF finder (http://www.ncbi.nlm.nih.gov).

***Helitron* identification:** *Helitrons* were identified by using as query the two distinct protein regions representing the rolling-circle motif and domain 5 of the helicase of previously described *A. thaliana Helitrons* in TBLASTN searches of the *L. japonicus* sequences. A list of contigs with significant hits (*e*-value $<10^{-4}$) to both domains within a 10-kb neighborhood was compiled and these sequences plus 10 kb upstream and downstream of the outer coordinates of each domain were extracted and translated in all six reading frames. Comparisons to known *Helitron* proteins were used to demarcate the beginning and end of element-encoded proteins within that fragment. Where possible, the 5′ and 3′ terminal regions of the elements were more precisely determined either by comparisons of closely related sequences or by the presence of key structural hallmarks (AT insertion site, 5′-TC, and 3′-CTRR and the 15- to 20-nucleotide palindrome close to the 3′-end).

**Phylogenetic analysis:** Sequences of each TE type were used to generate multiple alignments and resolved into lineages by generating phylogenetic trees. Multiple sequence alignment was performed by CLUSTALW (http://www.ebi.ac.uk/clustalw) with default parameters for each TE type. Phylogenetic trees were generated on the basis of the neighbor-joining method (Saitou and Nei 1987) using PAUP* version 4.0b8 (Swofford 1999) with default parameters. Bootstrap values were calculated for each tree from 250 replicates.

**TE copy number estimation:** The copy number for each TE type in the 32.4 Mb of *L. japonicus* sequences was calculated on the basis of the elements obtained from TBLASTN searches and the RepeatMasker analysis described above.

**Pack-MULE:** Two approaches were employed to search for MULE–TIRs. For the first approach, the catalytic domain (most conserved region) of *Mutator* transposases from *A. thaliana* (Zhang and Wessler 2004) was used to search the *L. japonicus* sequences and the sequences flanking the transposase were examined for the presence of TIRs and target-site duplications. For the second approach, sequences of 50 randomly chosen TACs from the 443 TAC sequences were screened for the presence of inverted repeats by FINDMITE (Tu 2001). The recovered inverted repeats, which included both MULE–TIRs and other inverted repeats, were manually examined for features of MULE–TIRs (>40 bp, 7- to 10-bp target-site duplication). Furthermore, if a newly recovered TIR was <80% similar to a known TIR, it was defined as a new TIR family. The resulting MULE–TIR sequences were used to mask the 32.4-Mb *L. japonicus* genomic sequence with RepeatMasker (version 07/07/2001, using default parameters; http:/ftp.genome.washington.edu/RM/webrepeatmaskerhelp.html). RepeatMasker output files contain annotations of all sequences that matched MULE–TIRs as well as their position in the input genomic sequence and were the basis for identification of Pack-MULEs. Specifically, Pack-MULEs were identified by using the following search criteria: (1) TIRs should be separated by <10 kb; (2) sequences between the TIRs should be >100 bp; (3) TIRs should be in inverted orientation with terminal sequences pointing outward (as in previously described MULEs); (4) sequences between the TIRs must be highly similar ($E < 10^{-9}$) to nontransposase and nonhypothetical proteins in GenBank or to the genes in the *L. japonicus* gene index (provided by TIGR at http://www.tigr.org); and (5) the TIRs must be flanked by a recognizable target-site duplication of 7–10 bp. For individual Pack-MULEs, if the TIRs of two elements (with different target-site duplications) can be aligned (BLASTN, $E = 1e^{-10}$) and if >50% of the sequence between the TIRs can be aligned (BLASTN, $E = 1e^{-10}$), then the two elements are defined as copies that arose from the same element (presumably during transposition).

## RESULTS

**TE abundance:** The strategies and procedures used to identify TEs in *L. japonicus* are detailed in materials and methods. Briefly, consensus sequences based on the most conserved coding regions of previously described plant TEs were used as queries in TBLASTN searches of *L. japonicus* genomic DNA. Next, 10 kb flanking these coding regions were searched for element termini and target-site duplications. TE sequences identified in this way were used as input for RepeatMasker to find noncoding versions of these TEs based on nucleotide similarity. The RepeatMasker output was used as input for a second round of RepeatMasker and

**TABLE 1**

**Transposable elements in the 32.4 Mb of *L. japonicus* sequences**

| TE type | Copy number | Coding/noncoding | DNA amount (Mb) | % TE composition in the 32.4 Mb |
|---|---|---|---|---|
| Class 1 | | | | |
| *copia*-like | 309 | 212/97 | 1.50 | 4.6 |
| *gypsy*-like | 245 | 191/54 | 1.46 | 4.5 |
| LINEs | 124 | 124/0 | 0.40 | 1.2 |
| Total class 1 | 678 | 527/151 | 3.36 | 10.4 |
| Class 2 | | | | |
| *PIF/Pong* | 384 | 29/355 | 0.47 | 1.5 |
| CACTA | 24 | 23/1 | 0.14 | 0.4 |
| MULEs | 1,140 | 75/1065 | 1.56 | 4.8 |
| *hAT*s | 118 | 47/71 | 0.12 | 0.4 |
| *Helitrons* | 27 | 21/26 | 0.14 | 0.4 |
| *Mariner* | 1 | 1/0 | 0.002 | 0.006 |
| MITEs | 370[a] | 0/370 | 0.20 | 0.62 |
| Total class 2 | 2,064 | 196/1868 | 2.63 | 8.1 |
| TE fragments[b] | 9,810 | — | 3.98 | 12.3 |
| Total TEs | 12,552 | 723/4038 | 9.97 | 30.8 |

[a] Of these, 213 were *Tourist*-like and 157 were *Stowaway*-like.

[b] Elements lacking one or both ends. See Table 2 for distribution of each TE type.

this process was reiterated until no new TE sequences were identified. Finally, RECON (Bao and Eddy 2002) was used to identify any novel TEs (those not related to previously described elements or to coding elements in *L. japonicus*). RECON is a program for *de novo* identification of repeats based solely on their repetitive nature. The RECON analysis, however, did not detect any novel repeats, indicating that the vast majority of TEs had been identified. The results of these analyses are summarized in Table 1.

Of the 32.4 Mb of *L. japonicus* sequence searched in this way, ~10 Mb is derived from TEs. Approximately 6 Mb of the 10 Mb were complete elements (both termini identified) and the remaining 4 Mb were TE fragments (one or both ends missing) (Tables 1 and 2). The 6 Mb of complete elements included 525 coding elements (with relatively intact coding sequences) and ~2000 noncoding elements (with no significant coding capacity but with ends related to a coding element). Of the ~2000 noncoding elements, 370 are MITEs (213 *Tourist*-like and 157 S*towaway*-like; Table 1). For the other noncoding elements, significant sequence identity with the corresponding coding TE could not be detected between the element TIRs. For example, coding and noncoding *PIF/Pong*-like elements share sequence homology only in the first 9–13 bp of the TIR region.

*Helitrons* were identified by using as query in TBLASTN searches the rolling-circle motif and the most conserved domain of the helicase (domain 5) (see materials and methods). Twenty-one elements were identified containing both domains (Table 1).

**TE diversity:** The identification of thousands of TEs in *L. japonicus* provided the raw material to address questions about TE diversity and temporal aspects of TE amplification. Such issues include whether *L. japonicus* harbors most of the previously identified TE lineages, whether new lineages have evolved, and whether certain TE lineages have recently amplified. As a first step in addressing these questions, we generated phylogenetic trees for all TE types (Figures 1A, 3A, and 6A; supplemental Figures S1A, S2A, S3, and S4 at http://www.genetics.org/supplemental/). In each case, the most conserved coding sequences of *L. japonicus* TEs as well as those representing all lineages previously identified

**TABLE 2**

**Distribution of transposable element fragments**

| TE type | Copy number | DNA amount (Mb) |
|---|---|---|
| Class 1 | | |
| *copia*-like | 831 | 0.69 |
| *gypsy*-like | 524 | 0.66 |
| LINEs | 584 | 0.45 |
| Class 2 | | |
| *PIF/Pong* | 1132 | 0.29 |
| CACTA | 285 | 0.29 |
| MULEs[a] | 5503 | 1.23 |
| *hAT*s | 400 | 0.13 |
| *Helitrons* | 551 | 0.24 |
| Total TEs | 9810 | 3.98 |

Transposable element fragments are those lacking one or both ends.

[a] Some of the MULEs lacking both ends might be non-TIR MULEs.

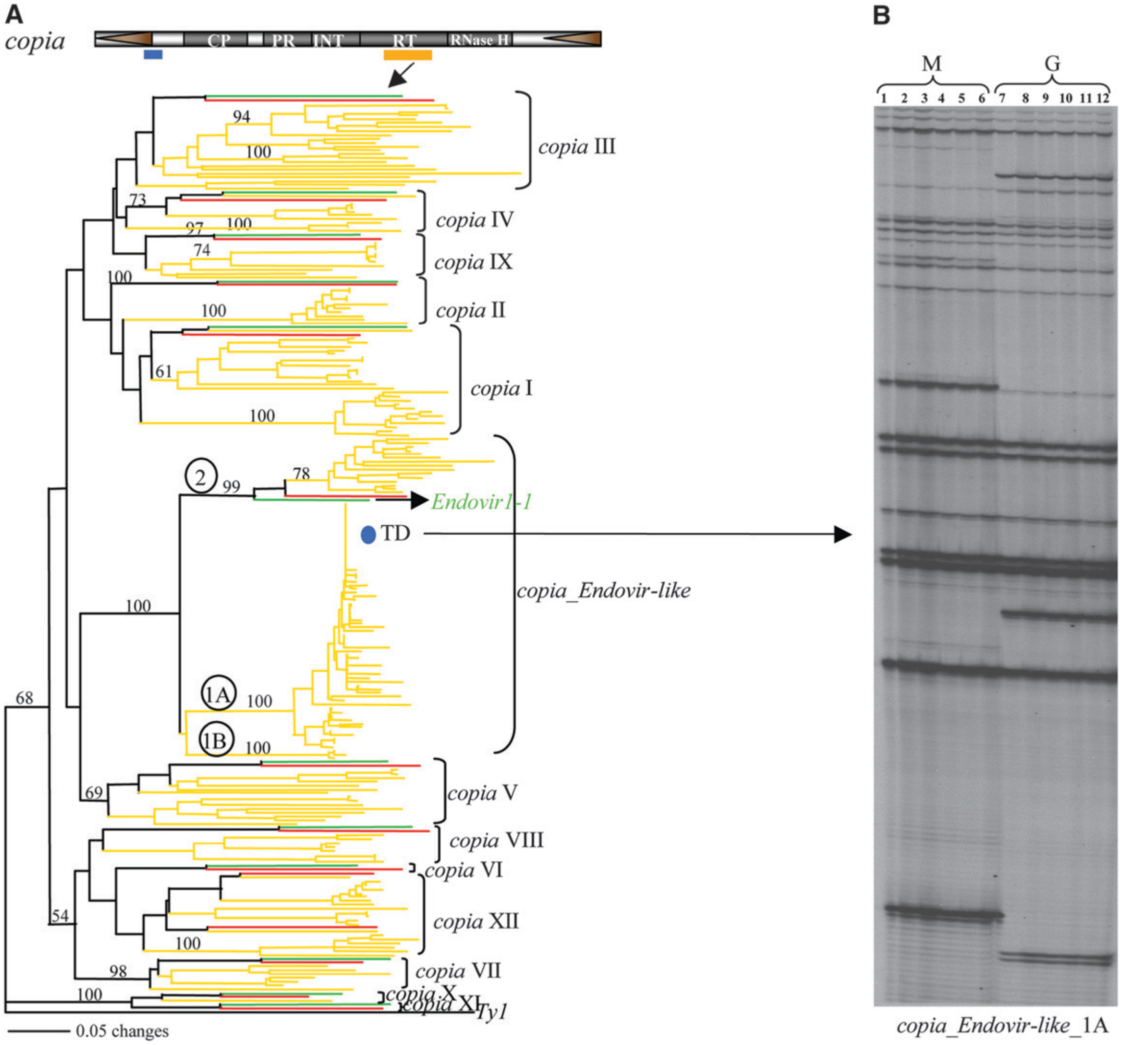

FIGURE 1.—Phylogeny and transposon display of *copia*-like elements. (A) The phylogenetic tree was generated using the RT domain (orange bar) of elements from *L. japonicus* (yellow) and a representative of each lineage from *A. thaliana* (green) and *B. oleracea* (red) and rooted with the corresponding RT from the yeast *Ty1* element. This tree, and all trees in subsequent figures, was generated using the neighbor-joining method and bootstrap values were calculated from 250 replicates. The 13 bracketed *copia*-like lineages are those identified in previous studies, and the three sublineages of *copia_Endovir*-like are labeled (2, 1A, 1B) and discussed in the text. The blue bar indicates the region in the LTR used to generate primers for transposon display analysis from the element noted with a blue circle. (B) Transposon display analysis of one of the *copia_Endovir*-like sublineages. Sublineage-specific primers were designed and PCR analysis was performed with these primers and with a *Bfa*1+T primer and resolved on a 6% polyacrylamide gel. Lanes 1–6, genomic DNAs from individual (siblings) plants from Miyakojima (M); lanes 7–12, individual plants from Gifu (G).

from the Brassicaceae (*A. thaliana* and *B. oleracea*) (ZHANG and WESSLER 2004) were compared by multiple alignments using CLUSTALW and used in neighbor-joining tree construction (SAITOU and NEI 1987).

Most of the major TE lineages from the Brassicaceae were also in *L. japonicus* (Figures 1A, 3A, and 6A; supplemental Figures S1A, S2A, S3, and S4 at http://www.genetics.org/supplemental/), and several new TE lineages were found in the limited amount of *L. japonicus* sequence. For example, the Brassicaceae and *L. japonicus* share *MuDR*-like and *Jittery*-like MULEs, *PIF*, and *Pong*-like elements, clades A and B of CACTA elements, *Tag1*-like, *Tag2*-like, and *Tip100* of *hAT* elements (Figures 3A and 6A; supplemental Figures S1A and S3 at http://www.genetics.org/supplemental/), and all 13 lineages of *copia*-like elements (Figure 1A). In contrast, several *gypsy*-like and LINE lineages in *L. japonicus* are not in the Brassicaceae. Of 23 Brassicaceae *gypsy*-like lineages, five (*Gimli*, *Gloin*, *Tft*, *Meriadoc*, and *Athila*, supplemental Figure S2A at http://www.genetics.org/supplemental/) were in the *L. japonicus* database, and the three remaining *L. japonicus gypsy*-like lineages were not in the Brassicaceae (*Lj_gypsy_1A*, *Lj_gypsy_2A*, and *Lj_gypsy_3A*; supplemental Figure S2A at http://www.genetics.org/supplemental/). Similarly, 11 of the 12 Brassicaceae LINE lineages (including the entire clade I) were not in this subset of the *L. japonicus* sequence; the majority of the *L. japonicus* LINEs (87of 114 copies) were grouped into three lineages (*Lj*_LIII, *Lj*_IV, and II-L; supplemental Figure S4 at http://www.genetics.org/supplemental/) that were not reported in the Brassicaceae. Given the limited amount of *L. japonicus* genomic sequence analyzed, we were surprised to identify several major TE lineages that were not previously described. In addition to the examples cited above, there was a large lineage of *copia*-like elements with an additional conserved ORF (*copia_Endovir*-like, Figure 1A) and a group of MULEs that are related to the fungal *Hop* element (Figures 3A and 4). Finally, several class 1 and class 2 TE families contain nearly identical members, suggesting recent or ongoing transposition. The most significant results from this analysis are considered in more detail below.

***Copia*-like elements with an additional ORF:** Among class 1 elements, *copia*-like elements are the most abundant in the available *L. japonicus* sequence (Table 1). The most numerous lineage, *copia_Endovir*-like, includes ~40% of all *copia*-like elements and can be further divided into three sublineages (Figure 1A). Of these, sublineage 2 is more closely related to a small group of *A. thaliana* elements (called *Endovir1-1*) that contain an extra ORF (Laten 1999; Peterson-Burch *et al.* 2000). Like *Endovir1-1*, all *L. japonicus* elements in this lineage contain an extra ORF (called ORF3) located between *pol* and the 3′ LTR. As such, the *L. japonicus* elements are members of the newly named Sireviruses, a group of *Ty1/copia* elements like *Endovir1-1* that often have an extra ORF. This name comes from the founding *SIRE-1* element of soybean (Havecker *et al.* 2005).

The structural features of a typical member of each sublineage are shown in Figure 2. Note that ORF3 from each sublineage varies in length (~630–950 aa) and that there is >75% amino acid sequence identity within each sublineage. However, intersublineage sequence similarity is <20%. Our survey identified 82 elements, 64 from sublineage 1A (40 full-length with LTR and target-site duplication defined), 3 from sublineage 1B (all full length), and 15 from sublineage 2 (9 full length). For the majority of full-length copies, ORF3 is intact. The most recently amplified elements are in sublineage 1A where ORF3 is intact for 34 of the 40 full-length elements (no frameshifts and/or stop codons). Most strikingly, 10 elements are nearly identical along their entire length of ~12 kb. Of these 10 elements, 2 are identical, 4 are 99% identical (~200 mismatches), and 4 are ~97–98% identical (<500 mismatches). For sublineage 2, the lineage most closely related to *Endovir1-1*, 6 of the 9 full-length ORF3 copies are intact.

For all *L. japonicus* families with a third ORF, interfamily sequence similarity did not extend beyond the reverse transcriptase domain and significant similarity could not be detected when the third ORF from different sublineages was compared. Prior studies have referred to the extra ORF encoded by plant LTR retrotransposons as envelope-like (Havecker *et al.* 2004, 2005) because several have transmembrane and coil-coil domains like the *env* genes of retroviruses (Laten *et al.* 1998). To search for these and other domains in the *L. japonicus* elements, consensus sequences were generated for each sublineage and the predicted ORF was screened for transmembrane and coil-coil domains using appropriate software (see materials and methods). A transmembrane domain was detected for the representative ORF3 from sublineages 1A and 1B (19 aa and 11 aa, respectively) but not for sublineage 2. In addition, a coil-coil domain was detected for the representative ORF3 from sublineage 1A but not from sublineages 1B and 2.

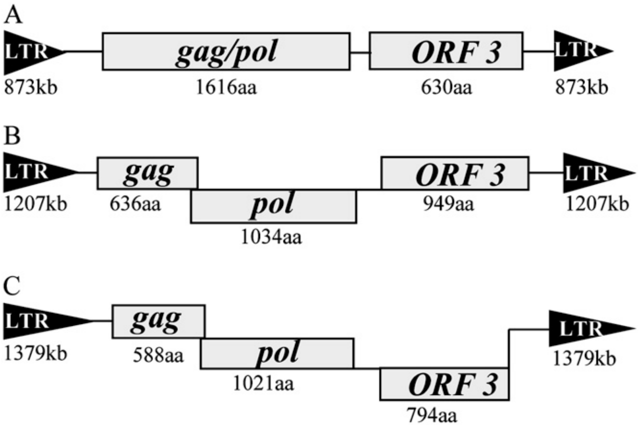

Figure 2.—Structural organization of a representative *copia_Endovir*-like element from the three sublineages. (A) *Lj_copia_Endovir*-like_2, (B) *Lj_copia_Endovir*-like_1A, and (C) *Lj_copia_Endovir*-like_1B. For each sublineage, shaded boxes represent the different ORFs and solid arrowheads represent the LTRs. The length of the average ORF (in aa) and LTR (in base pairs) is shown. For B and C, the *gag* and *pol* ORFs overlap and a frameshift is presumed to occur to generate both proteins.

**A legume-specific lineage of MULEs:** MULEs represent a diverse family of TEs that are widespread in plants and are in some fungal species (Yu *et al.* 2000; Chalvet *et al.* 2003). While previous studies identified two major groups of plant MULEs (*MuDR* and *Jittery*-like) (Lisch 2002; Xu *et al.* 2004), we were surprised to find a third lineage in *L. japonicus* (Figure 3A). In addition to *MuDR*-like (27 elements) and *Jittery*-like (24 elements) MULEs, the third lineage is most closely related to a small group of recently discovered fungal MULEs called *Hop* from *Fusarium oxysporum* (Chalvet *et al.* 2003) and to a MULE element from *Magnaporthe grisea* than to any other plant element. Elements from this lineage (named herein *Hop*-like) are 3–9 kb in length, contain ~40-bp TIRs, and have 9-bp target-site duplications. Of the 25 elements, 18 are full length, with TIR identity ranging from 80 to 100%. Furthermore, at least 7 elements share >90% sequence homology over a 2- to 4-kb region. Overall,

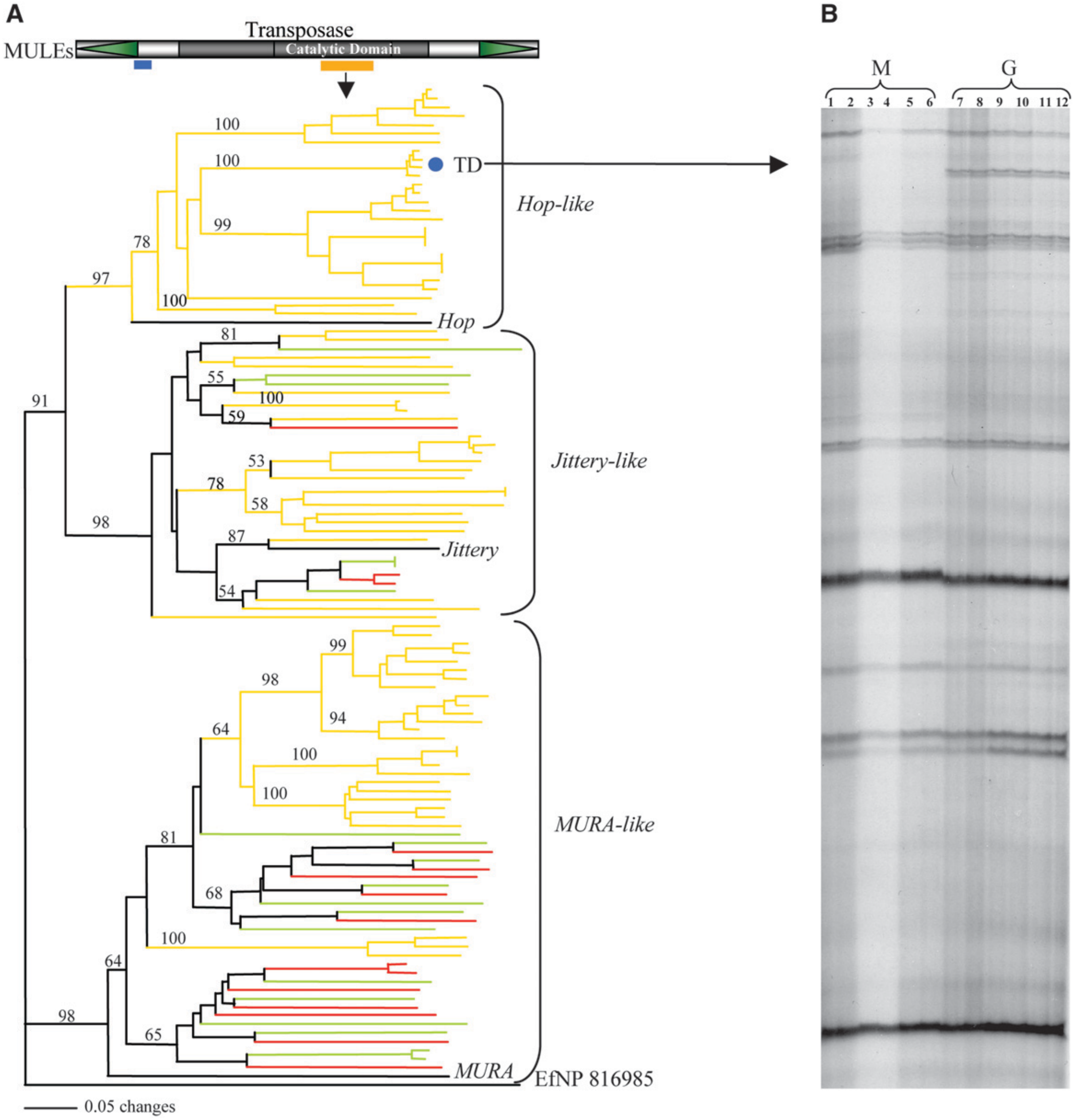

Figure 3.—Phylogeny and transposon display of MULE elements. (A) The phylogenetic tree was generated using the catalytic domain (orange bar) of the transposase of elements from *L. japonicus* (yellow) and a representative of each lineage from *A. thaliana* (green) and *B. oleracea* (red) and rooted with the transposase from a fungus. The three bracketed lineages include two previously identified (*MURA* and *Jittery*) and the new *Hop*-like lineage. The solid blue circle indicates the elements and the blue bar indicates the region used to generate primers for transposon display (TD) analysis. (B) Transposon display analysis of *L. japonicus* *Hop*-like elements. Sublineage-specific primers were designed and PCR analysis was performed with these primers and with a *Bfa*1+T primer and resolved on a 6% polyacrylamide gel. Lanes 1–12 are the same as Figure 1B.

this lineage contains the most recently amplified MULEs in the *L. japonicus* genome.

To determine whether *Hop*-like MULEs are also present in other plant genomes, additional TBLASTN searches were performed using as query a consensus sequence derived from the catalytic domains of the 25 *Hop*-like elements in *L. japonicus* against the NCBI NR (nonredundant), GSS (genome survey sequences), and HTGS (high-throughput genomic sequence) databases. These searches identified 130 additional elements, all from legumes (including soybean, chickpea, and medicago) (Figure 4). Surprisingly, not a single *Hop*-like element was detected from any nonlegume plant species, including the completely sequenced genomes of *A. thaliana*, *O. sativa*, and *Populus trichocarpa*. Furthermore, a phylogenetic tree generated from the legume *Hop*-like MULEs suggested that elements from each legume formed species-specific monophyletic sublineages (Figure 4).

**Pack-MULEs:** Pack-MULE is the name given to MULEs that have captured fragments of host genes. These elements were first discovered in maize (Talbert and Chandler 1988) and were found to be abundant in the rice genome where >3000 were identified (Jiang *et al.*

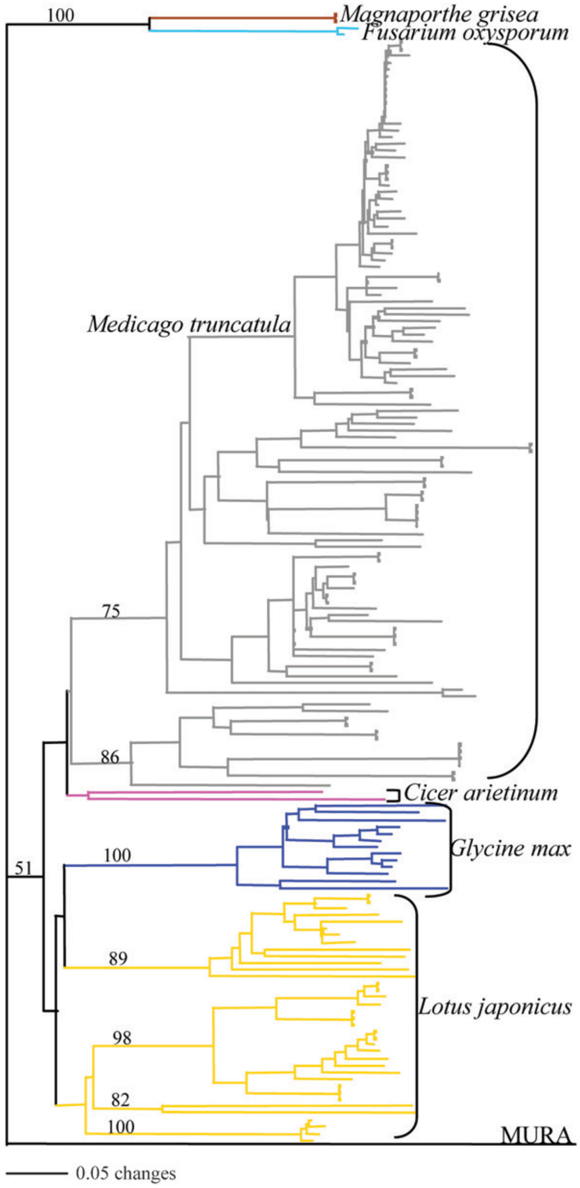

Figure 4.—Legume- and fungal-specific MULE lineage. Phylogenetic tree of legume and fungal MULEs. The phylogenetic tree was generated using the catalytic domain of *Hop*-like MULEs from *L. japonicus* and from the indicated legume and fungal species.

2004a). The identification of only a few Pack-MULEs in the genome of *A. thaliana* (5 in 17 Mb of genomic sequence, or 36 genomewide) (Yu *et al.* 2000; Hoen *et al.* 2006) suggested that gene capture by MULEs might occur frequently only in the grasses. To determine whether Pack-MULEs are abundant in *L. japonicus*, a search of the RepeatMasker output files was conducted using the same parameters as those employed previously in rice (see materials and methods). A total of 160 Pack-MULEs were identified, including 73 (46%) with two or more copies in the available genomic sequence. The amplification of these elements was likely due to transposition rather than to large-scale genome duplication because each copy has a unique target-site duplication. Furthermore, for 23 of these 73 amplified elements, the copies had sequence similarity of 99% or higher. Like the rice Pack-MULEs, the protein hits of the *L. japonicus* Pack-MULEs include a variety of functional domains such as kinases, transcription factors, and transporters (supplemental Table 1 at http://www.genetics.org/supplemental/). To assess whether any of the 160 Pack-MULEs were expressed, their sequences were used as queries to search the *L. japonicus* EST database (http://www.kazusa.or.jp/en/plant/lotus/EST). Nine elements (6%) had exact matches (supplemental Table 2 at http://www.genetics.org/supplemental/), indicating that some of the captured gene fragments are transcribed.

In a prior study, availability of the entire rice genomic sequence facilitated the identification of most of the rice genes whose sequences were captured by Pack-MULEs (Jiang *et al.* 2004a). To investigate the origin of the sequences captured by *L. japonicus* Pack-MULEs, the internal regions of the 160 Pack-MULEs were used to query all the *L. japonicus* sequences in GenBank (a total of 122.2 Mb, including unfinished TACs). Among the 160 elements, 71 (44%) had one or more significant homologs (BLASTN $E < 10^{-10}$) that were not flanked by MULE–TIRs. Of the 71 Pack-MULEs with identified genomic homologs, 15 (9% of total Pack-MULEs) contain sequences from two or more loci (Figure 5; supplemental Table 3 at http://www.genetics.org/supplemental/). As shown in Figure 5, two highly similar chimeric Pack-MULEs (from chromosomes 1 and 3) contain sequences from three genomic loci (chromosomes 1, 2, and an unknown locus) and one of the deduced ORFs contains sequences from all three loci (Figure 5).

**Recently amplified elements and TE insertion polymorphism:** With the exception of LINEs, all other major TE types have lineages with highly similar members (Figures 1A, 3A, and 6A; supplemental Figures 1A and 2A at http://www.genetics.org/supplemental/). For example, two families of *Pong*-like elements have members that share ∼98% nucleotide sequence similarity and both ORF1 and ORF2 are intact (not interrupted by stop codons) (Figure 6A, *Lj_Pong3A* and *Lj_Pong1A*). In addition, ∼58% (48/82) of the *copia_Endovir*-like elements have a LTR identity of ∼98% with only one to three mismatches over the entire length of the ∼0.8- to 1.2-kb LTR, and ∼20 share 99% identity overall. If these elements have in fact amplified recently, their insertion sites may be polymorphic in different *L. japonicus* ecotypes. Polymorphic sites not only would provide evidence for recent element activity, but also could be developed into molecular markers.

To assess the extent of insertion-site polymorphism, we performed a modification of the AFLP technique called transposon display (see materials and methods). To this end, PCR primers were designed from the

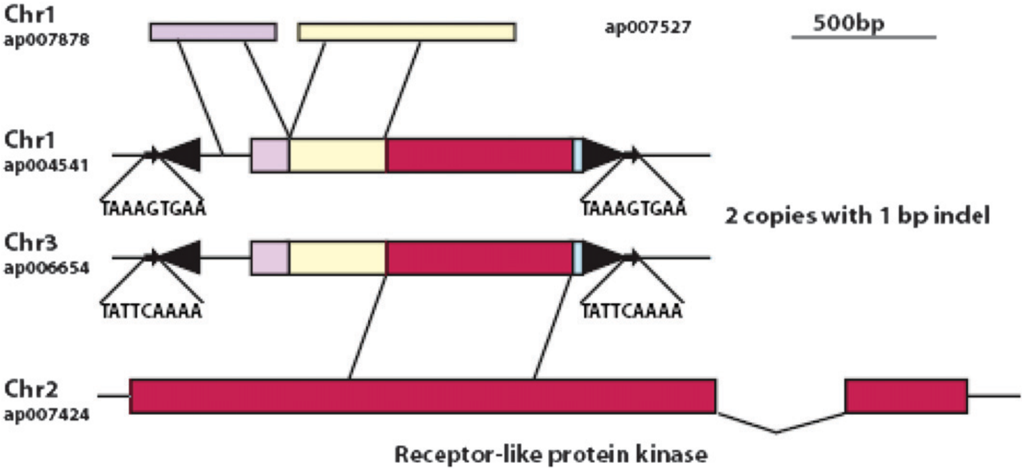

FIGURE 5.—Structure and genomic origin of gene fragments in a chimeric Pack-MULE. Pack-MULE–TIRs are shown as black arrowheads and black horizontal arrows indicate target-site duplications. Homologous regions are connected with solid lines and the GenBank accession number of the TAC sequences where the Pack-MULE or the genomic copy was found is indicated. The chromosomal location of TAC AP007527 is unknown. For the Pack-MULE and the receptor-like protein kinase gene, exons are depicted as colored boxes and introns as the lines connecting exons. The light blue box represents part of an exon where the sequence is of unknown origin. The identity of the two other genomic copies (accession nos. AP007878 and AP007527) is not known and are depicted as narrow boxes.

subterminal sequences of specific lineages of *Pong*-like, CACTA, MULE, *copia*-like, and *gypsy*-like elements that harbored multiple highly similar members (lineages and regions for primer design are indicated in Figures 1A, 3A, and 6A; supplemental Figures 1A and 2A at http://www.genetics.org/supplemental/) and were used to amplify genomic DNA from six individual (sibling) plants from each of two ecotypes: Miyakojima (MG-20, the sequenced genome) and Gifu (B-129). Results of the transposon display analysis for each TE type are shown in Figures 1B, 3B, and 6B and in supplemental Figures 1B and 2B at http://www.genetics.org/supplemental/. For each TE lineage tested, polymorphic bands were seen between the two ecotypes but not among the six individuals of a single ecotype. Overall, the TE insertion polymorphism between the two ecotypes ranged from ~15% (CACTA elements) to 48% (*Pong*-like elements), which is about four times higher than the ~4% polymorphism observed in previous studies using traditional AFLP methods (JIANG and GRESSHOFF 1997; KAWAGUCHI *et al.* 2001).

Because the observed polymorphic bands could be the result of sequence variation at the restriction site (*Bfa*1) in the genomic DNA, a PCR-based analysis was performed with primers derived from sequences of cloned transposon display bands to verify polymorphic insertions (see MATERIALS AND METHODS). Of the cloned sequences, half (7 of 14) were demonstrated to be actual polymorphic insertion sites (TE present at a locus in all individuals of only one ecotype). The nature of the remaining polymorphic bands (7 of 14) could not be verified due in part to PCR artifacts caused by the repetitive nature of the amplified (TE) sequence. All monomorphic bands tested represented insertions in both ecotypes.

## DISCUSSION

The availability of a significant amount of *L. japonicus* genome sequence has permitted for the first time a comprehensive study of the TE landscape in a legume species. To analyze a database consisting of finished or nearly finished TAC sequences representing ~7% of the genome, we devised computational strategies to identify and characterize coding, noncoding, and novel TEs. While computer-assisted analysis facilitated a determination of TE abundance (copy number) and diversity (TE lineages), the availability of complete TAC sequences permitted identification of full-length TEs. These sequences in turn facilitated the design of tools for genomewide experimental analysis.

**The TE landscape:** As mentioned in MATERIALS AND METHODS, the available *L. japonicus* sequence analyzed in this study was from the gene-rich regions of the genome. However, despite this limited data set, the abundance of class 1 *vs.* class 2 elements and coding *vs.* noncoding elements in *L. japonicus* is similar to what has been observed in the complete genome sequences of *O. sativa* and *A. thaliana*. In the genome sequence analyzed, coding class 1 elements of *L. japonicus* were more abundant than noncoding elements (~500 coding *vs.* ~150 noncoding) and all of the noncoding elements were solo LTRs derived from the LTR retrotransposons. For DNA elements, noncoding elements significantly outnumbered coding elements (~1800 noncoding *vs.* ~200 coding), except for CACTA elements and *Helitrons* (Table 1). Overall, class 2 elements are numerically more abundant than class 1 elements (~2000 DNA *vs.* ~600 RNA); however, class 1 elements account for a larger fraction of DNA in the 32.4 Mb of *L. japonicus* sequence (~10 Mb of class 1 *vs.* 8 Mb of class 2). This reflects the fact that most class 1 elements on average are larger than class 2 elements.

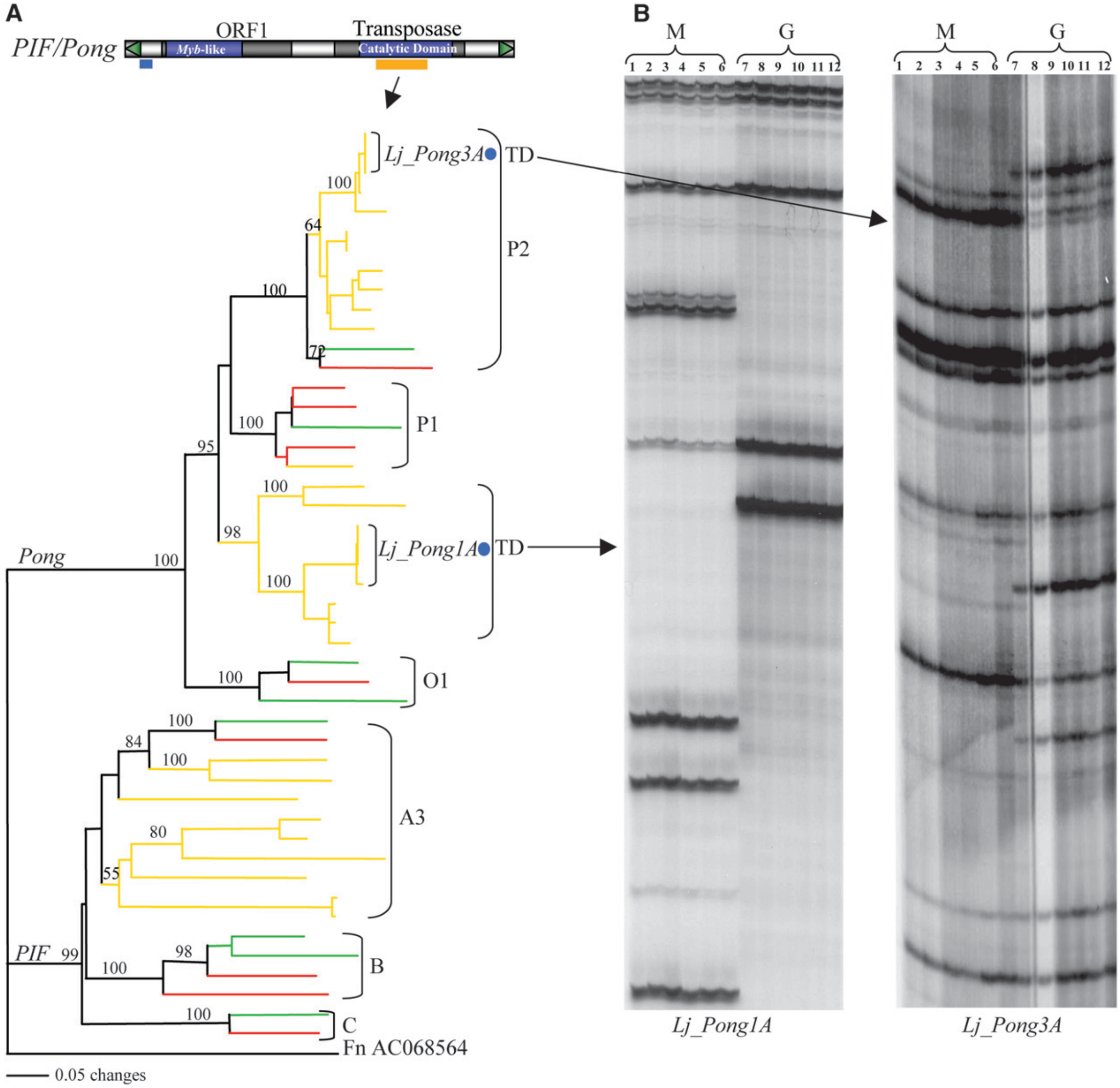

FIGURE 6.—Phylogeny and transposon display of *PIF/Pong*-like elements. (A) The phylogenetic tree was generated as described in previous figures. Lineages identified in previous studies are indicated next to the brackets. The solid blue circle indicates the elements used to generate sublineage-specific primers for transposon display analysis. (B) Transposon display using primers derived from two distantly related *Pong* lineages. See Figure 1B for details and DNA analyzed.

It is important to note that because this subset of *L. japonicus* sequences was derived from gene-rich regions in the genome, the observed TE abundance can be biased. For example, in *O. sativa*, RNA elements (especially LTR retrotransposons) appear to be more concentrated in gene-poor regions (including heterochromatic DNA) while DNA elements appear to be predominantly near genes (INTERNATIONAL RICE GENOME SEQUENCING PROJECT 2005). As such, TE abundances reported in this study probably underestimate class1 elements and overestimate class 2 elements. Despite these caveats, the TE component of *L. japonicus* contained a few surprises and some unusual elements. These are discussed in more detail below.

***Copia*-like elements containing an extra ORF:** The most abundant and recently amplified LTR retrotransposons in *L. japonicus* are 10- to 12-kb *copia*-like elements containing a third ORF (Figure 1A). Elements with similar features have been described in several plant species and have been given the name Sirevirus to reflect the founding *SIRE*-1 element (from soybean) and their structural affinity with retroviruses (HAVECKER *et al.* 2005). The third ORF of Sireviruses is diverse and of apparent independent origin. However, several resemble the envelope ORF of retroviruses in that they encode transmembrane and coil-coil domains (WRIGHT and VOYTAS 1998; LATEN 1999; PETERSON-BURCH *et al.* 2000; HAVECKER *et al.* 2005). A transmembrane domain was detected in the third ORF from elements in sublineages 1A and 1B, and a coil-coil domain was detected in elements from sublineage 1A. However, the third ORF of elements in sublineage 2 revealed no obvious domains.

While the *L. japonicus* elements add to the mystery of Sireviruses, they also offer a potentially valuable experimental system to address many of the outstanding questions. To date, there is no experimental evidence supporting a retrovirus lifestyle for any Sirevirus nor is there any evidence that its so-called *env*-like genes encode proteins that can mediate viral infection or cell-to-cell transmission. An alternative explanation for the presence of an additional ORF is that these *copia*-like elements, like the newly described plant Pack-MULE and *Helitron* elements, are able to capture and amplify plant gene fragments. However, this scenario seems unlikely because there is no evidence of significant homology between the numerous extra ORFs of Sireviruses, including the *L. japonicus* elements, and any known plant gene, ORF, or cDNA.

The question of function could be addressed by observing a Sirevirus that is active, that is, one that is capable of retrotransposition. Our analysis suggests that some of the *L. japonicus* Sireviruses may still be active. With <10% of the genome sequence, we identified 34 full-length Sireviruses in sublineage 1A with no interrupting stop codons. In fact, 10 of these 34 elements are virtually identical, indicating that they have integrated very recently.

For the future analysis of these elements, the value of having *L. japonicus* TAC contigs cannot be overestimated. By aligning four to six full-length elements, we were able to design element-specific primers for transposon display analysis of progeny from two *L. japonicus* ecotypes (Figures 1, 2, and 6; supplemental Figures S1 and S2 at http://www.genetics.org/supplemental/). While our preliminary analysis did not reveal any new integration events, the ability of transposon display to visualize hundreds of Sireviruses in the genome will be a powerful way to screen for retrotransposition of this family in a variety of strains and crosses and in plants subjected to stresses known to activate retrotransposons in other plant species.

**Pack-MULEs:** Previous studies demonstrated that Pack-MULEs are abundant in at least two grasses, rice, and maize, but not in *A. thaliana* (Yu *et al.* 2000; Jiang *et al.* 2004a; Hoen *et al.* 2006). The limited amplification of Pack-MULEs observed in *A. thaliana* could have been a consequence of its streamlined genome where few TEs have amplified significantly. Alternatively, the dearth of Pack-MULEs in *A. thaliana* may reflect a paucity of these elements in the genomes of dicotyledonous plants. Analysis of the Pack-MULEs in *L. japonicus* provided an opportunity to investigate their abundance in the genome of another dicotyledenous plant. Similar to what was observed in *O. sativa*, Pack-MULEs are abundant in *L. japonicus*; we identified 160 Pack-MULEs in the available *L. japonicus* sequence. If the sequences used in this study are representative of the rest of the genome, we estimate that there would be 2300 Pack-MULEs in the genome (160 × 472 Mb/32.4 Mb).

Of the 160 characterized Pack-MULEs, all have amplified by transposition (as deduced by the detection of target-site duplications), and 6% of the captured genes (or gene fragments) are transcribed. Like the Pack-MULEs in *O. sativa*, where one-fifth of the Pack-MULEs are chimeric, 21% of the Pack-MULEs (those with identifiable genomic copies, or 9% of all Pack-MULEs) in *L. japonicus* carry sequences from multiple loci. As such, the existence of numerous *L. japonicus* Pack-MULEs indicates that gene fragment acquisition and amplification by Pack-MULEs is a significant phenomenon in the shuffling of gene segments in many and diverse plant taxa. One of the critical issues regarding Pack-MULEs is whether some of the captured gene fragments could possibly evolve into real coding regions or whether all are pseudogenes. Due to the limited genomic resources (the unavailability of a sequenced cDNA library and gene annotation), a systematic evaluation of the gene structure and potential ORFs for Pack-MULEs in *L. japonicus* was not possible at the time of analysis. Such a thorough analysis of Pack-MULE origins will have to wait for the availability of more *L. japonicus* genome sequence as well as the availability of large cDNA collections and genome annotation.

Although the Pack-MULEs of *L. japonicus* and *O. sativa* share many features, our analysis indicates that their amplification in *L. japonicus* has been more recent. For example, 23 of the 73 amplified Pack-MULEs (32%) in *L. japonicus* have another copy with a sequence similarity of 99% or higher. This compares with only 2 of 73 amplified rice Pack-MULEs with 99% or higher sequence similarity. Thus, like the Sireviruses, many of the Pack-MULEs of *L. japonicus* have recently amplified and should prove to be a valuable resource in understanding the mechanism of gene fragment acquisition.

**A new MULE lineage:** Virtually all MULEs identified to date in plants belong to one of two groups: *MUDR* and *Jittery*-like (Yu *et al.* 2000; Lisch 2002; Xu *et al.* 2004). It was therefore surprising to find a third group of MULEs in *L. japonicus* that is more closely related to the fungal element *Hop* (Chalvet *et al.* 2003) than to *MUDR* or *Jittery*-like MULEs (Figure 3A). Several features of this *Hop*-like MULE lineage suggest that they may have arisen during the emergence of the legume family. First, they are present in all legumes examined but absent from nonlegume species (Figure 4). Second, each of the elements from each legume species forms a monophyletic group, which contrasts with other plant TE lineages where elements in monophyletic groups come from a variety of taxa. These features, coupled with the fact that a fungus such as *F. oxysporum* is a pathogen of legumes (Altier and Groth 2005), lend

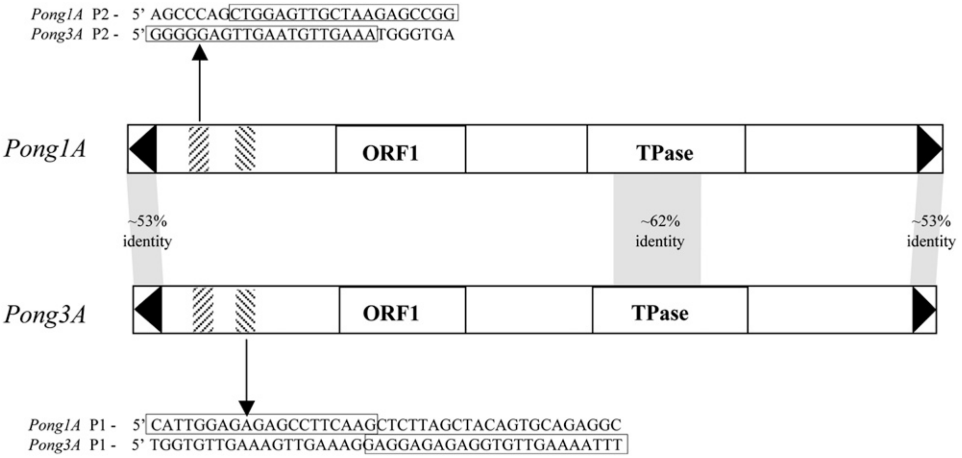
Pong1A P2 - 5' AGCCCAGCTGGAGTTGCTAAGAGCCGG
Pong3A P2 - 5' GGGGGAGTTGAATGTTGAAATGGGTGA
Pong1A
ORF1
TPase
~53% identity
~62% identity
~53% identity
Pong3A
ORF1
TPase
Pong1A P1 - 5' CATTGGAGAGAGCCTTCAAGCTCTTAGCTACAGTGCAGAGGC
Pong3A P1 - 5' TGGTGTTGAAAGTTGAAAGGAGGAGAGAGGTGTTGAAAATTT

FIGURE 7.—Structure of *Pong*1A and 3A showing regions used for transposon display primer design. The ORF1 and the transposase (TPase) regions are indicated for each *Pong*. Solid arrows represent terminal inverted repeats. Shaded areas share sequence homology between the two *Pongs*. Diagonal lines indicate the corresponding regions and sequences between the *Pongs* that were used for primer design.

themselves to an intriguing scenario whereby the *Hop*-like MULE elements originated from an ancient horizontal transfer event between fungus and legumes. This event may have occurred in the ancestor of today's legumes and the monophyletic groups of *Hop*-like elements in legume genomes may be the result of independent amplification and diversification in each derivative species.

**A dearth of MITEs:** Fewer than 400 MITEs belonging to the two MITE superfamilies *Tourist* and *Stowaway* were identified in the available *L. japonicus* sequence (∼200 *Tourist*-like, ∼150 *Stowaway*-like; Table 1). This value is significantly lower, even when extrapolated to the whole genome, than the ∼90,000 MITEs reported in *O. sativa* (Turcotte *et al.* 2001; Juretic *et al.* 2004; International Rice Genome Sequencing Project 2005). To identify noncoding elements like MITEs, we first characterized full-length coding elements and used their ends to query the *L. japonicus* sequence. For example, all of the *Tourist*-like MITEs were identified by similarity searches to full-length *PIF/Pong*-like elements. However, *Stowaway* MITEs could not be searched in this way because no full-length *L. japonicus* elements with *Stowaway* TIRs (Tc1/*mariner* elements) were recovered. Instead, *Stowaway* MITEs were identified by BLAST searches using previously characterized *Stowaway*-like TIRs. In addition, because our analysis was based on sequence similarity searches with coding elements, it was possible, even likely, that we missed many MITEs that shared no sequence similarity with the available queries. To address this limitation, an additional search was employed using RECON, a program that allows for *de novo* identification of repetitive sequences. Because the RECON output contained no additional MITEs, we conclude that MITEs are not as common in *L. japonicus* as they are in other plant genomes, especially in the grasses. As discussed below, the relatively small number of *Tourist* elements cannot be explained by an absence of their cognate *PIF/Pong* coding elements, which are well represented in the *L. japonicus* genome.

**Development of tools for experimental analysis:** As discussed above, the analysis of TEs in a genome database is greatly facilitated by the availability of finished TACs. Without these long contigs, full-length members of both class 1 and class 2 elements are often unrecognizable because the terminal and subterminal regions of most TEs share very limited sequence identity even when sublineages in the same TE family are compared. This point is nicely illustrated by comparing members of two full-length *Pong* families that have recently amplified in the *L. japonicus* genome (Figures 6A and 7). It should be obvious from this comparison that full-length elements could not be retrieved from a genome database made up of short sequence reads. While coding regions of most TEs, including *Pong*, can be identified on the basis of homology to previously derived conserved catalytic domains, it is virtually impossible to retrieve their respective TIRs and subterminal regions (or LTRs for class 1 elements) from a database of short sequence fragments (Figure 7). In this study, primers for transposon display analysis were derived from the alignment of the terminal sequences of elements of interest and served a crucial role as tools for whole-genome experimental analysis.

**Concluding remarks:** As mentioned in the Introduction, all plant and animal genomes characterized to date have a distinctive TE composition with respect to element types and their evolutionary trajectory. The analysis of the TEs in <10% of the *L. japonicus* genome allows us to approximate its TE landscape. Furthermore, the availability of complete BAC sequences covering ∼25% of *Medicago truncatula* (another model legume) will provide an unprecedented opportunity to study the TE relationship between these two closely related dicots relative to their distant relative *A. thaliana*. More importantly, the high quality of the *L. japonicus* sequence, in the form of hundreds of finished TACs, has facilitated our identification of novel elements, including recently amplified Sireviruses and Pack-MULEs and the legume-specific *Hop*-like MULES. Experimental strategies based on the use of PCR primers developed from full-length members of these and other *L. japonicus* elements promise to enrich our understanding of TEs and how they interact with host genomes.

We thank Sachiko Isobe from the National Agricultural Research Center for the Hokkaido Region for providing the *L. japonicus* seeds (Gifu–B129 and Miyakojima–MG20) and Cedric Feschotte for help with identification of the *Stowaway* MITEs. This work was supported by a grant from the National Science Foundation to S.R.W. and by a postdoctoral research fellowship from the National Science Foundation 0107590 to E.J.P.

## LITERATURE CITED

Altier, N., and J. Groth, 2005 Characterization of aggressiveness and vegetative compatibility diversity of *Fusarium oxysporum* associated with crown and root rot of birdsfoot trefoil. Lotus Newsl. **35:** 59–74.

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815.

Asamizu, E. T., S. Kato, Y. Sato, Y. Nakamura, K. A. Kaneko *et al.*, 2003 Structural analysis of a *Lotus japonicus* genome. IV. Sequence features and mapping of seventy-three TAC clones which cover the 7.5 Mb regions of the genome. DNA Res. **10:** 115–122.

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

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

Casa, A. M., C. Brouwer, A. Nagel, L. Wang, Q. Zhang *et al.*, 2000 The MITE family heartbreaker (*Hbr*): molecular markers in maize. Proc. Natl. Acad. Sci. USA **97:** 10083–10089.

Chalvet, F., C. Grimaldi, F. Kaper, T. Langin and M J. Daboussi, 2003 *Hop*, an active *mutator*-like element in the genome of the fungus *Fusarium oxysporum.* Mol. Biol. Evol. **20:** 1362–1375.

Craig, N. L., R. Craigie, M. Gellert and A. M. Lambowitz, 2002 *Mobile DNA II.* American Society for Microbiology, Washington, DC.

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.

Gupta, S., A. Gallavotti, G. A. Stryker, R. J. Schmidt and S. K. Lal, 2005 A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol. Biol. **57:** 115–127.

Havecker, E. R., X. Gao and D. F. Voytas, 2004 The diversity of LTR retrotransposons. Genome Biol. **5:** 225.

Havecker, E. R., X. Gao and D. F. Voytas, 2005 The Sireviruses, a plant-specific lineage of the *Ty1/copia* retrotransposons, interact with a family of proteins related to dynein light chain 8. Plant Physiol. **139:** 857–868.

Hoen, D. R., K. C. Park, N. Elrouby, Z. Yu, N. Mohabir *et al.*, 2006 Transposon-mediated expansion and diversification of a family of ULP-like genes. Mol. Biol. Evol. **23:** 1254–1268.

International Rice Genome Sequencing Project, 2005 The map-based sequence of the rice genome. Nature **436:** 793–800.

Jiang, Q., and P. M. Gresshoff, 1997 Classical and molecular genetics of the model legume *Lotus japonicus.* Mol. Plant Microbe Interact. **10:** 59–68.

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, S. R. McCouch, S. R. Eddy *et al.*, 2003 An active DNA transposon in rice. Nature **421:** 163–167.

Jiang, N., Z. Bao, X. Zhang, S. R. Eddy and S. R. Wessler, 2004a Pack-MULE transposable elements mediate genome evolution in plants. Nature **431:** 567–573.

Jiang, N., C. Feschotte, X. Y. Zhang and S. R. Wessler, 2004b Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr. Opin. Plant Biol. **7:** 115–119.

Juretic, N., T. E. Bureau and R. M. Bruskiewich, 2004 Transposable element annotation of the rice genome. Bioinformatics **20:** 155–160.

Kaneko, T., E. Asamizu, T. Kato, S. Sato, Y. Nakamura *et al.*, 2003 Structural analysis of a *Lotus japonicus* genome. III. Sequence features and mapping of sixty-two TAC clones which cover the 6.7 Mb regions of the genome. DNA Res. **10:** 27–33.

Kato, T., S. Sato, Y. Nakamura, T. Kaneko, E. Asamizu *et al.*, 2003 Structural analysis of a *Lotus japonicus* genome. V. Sequence features and mapping of sixty-four TAC clones which cover the 6.4 Mb regions of the genome. DNA Res. **10:** 277–285.

Kawaguchi, M., T. Motomura, H. Imaizumi-Anraku, S. Akao and S. Kawasaki, 2001 Providing the basis for genomics in *Lotus japonicus*: the accessions Miyakojima and Gifu are appropriate crossing partners for genetic analyses. Mol. Genet. Genomics **266:** 157–166.

Kidwell, M. G., 2002 Transposable elements and the evolution of genome size in eukaryotes. Genetica **115:** 49–63.

Koga, A., and H. Hori, 2001 The *Tol2* transposable element of the medaka fish: an active DNA-based element naturally occurring in a vertebrate genome. Genes Genet. Syst. **76:** 1–8.

Kumar, A., and J. L. Bennetzen, 1999 Plant retrotransposons. Annu. Rev. Genet. **33:** 479–532.

Lai, J., Y. Li, J. Messing and H. K. Dooner, 2005 Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. Proc. Natl. Acad. Sci. USA **102:** 9068–9073.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

Laten, H. M., 1999 Phylogenetic evidence for *Ty1/copia*-like endogenous retroviruses in plant genomes. Genetica **107:** 87–93.

Laten, H. M., A. Majumdar and E. A. Gaucher, 1998 SIRE-1, a *copia/Ty1*-like retroelement from soybean, encodes a retroviral *envelope*-like protein. Proc. Natl. Acad. Sci. USA **95:** 6897–6902.

Lisch, D., 2002 *Mutator* transposons. Trends Plant Sci. **7:** 497–504.

McDonnell, A. V., T. Jiang, A. E. Keating and B. Berger, 2006 Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics **22:** 356–358.

Meyers, B. C., S. V. Tingey and M. Morgante, 2001 Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. **11:** 1660–1676.

Nakamura, Y., T. Kaneko, E. Asamizu, T. Kato, S. Sato *et al.*, 2002 Structural analysis of a *Lotus japonicus* genome. II. Sequence features and mapping of sixty-five TAC clones which cover the 6.5 Mb regions of the genome. DNA Res. **9:** 63–70.

Peterson-Burch, B. D., D. A. Wright, H. M. Laten and D. F. Voytas, 2000 Retroviruses in plants? Trends Genet. **16:** 151–152.

Rost, B., R. Casadio, P. Fariselli and C. Sander, 1995 Prediction of helical transmembrane segments at 95% accuracy. Protein Sci. **4:** 521–533.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SanMiguel, P., and J. L. Bennetzen, 1998 Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann. Bot. **81:** 37–44.

SanMiguel, P., A. Tikhonov, Y.-K. Jin, N. Motchoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.

Sato, S., T. Kaneko, Y. Nakamura, E. Asamizu, T. Kato *et al.*, 2001 Structural analysis of a *Lotus japonicus* genome. I. Sequence features and mapping of fifty-six TAC clones which cover the 5.4 Mb regions of the genome. DNA Res. **8:** 311–318.

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

Talbert, L. E., and V. L. Chandler, 1988 Characterization of a highly conserved sequence related to *mutator* transposable elements in maize. Mol. Biol. Evol. **5:** 519–529.

Tu, Z., 2001 Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. Proc. Natl. Acad. Sci. USA **98:** 1699–1704.

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

Vicient, C. M., A. Suoniemi, K. Anamthawat-Jonsson, J. Tanskanen, A. Beharav *et al.*, 1999 Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. Plant Cell **11:** 1769–1784.

Wolf, E., P. S. Kim and B. Berger, 1997 MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci. **6:** 1179–1189.

Wright, D. A., and D. F. Voytas, 1998 Potential retroviruses in plants: *Tat1* is related to a group of *Arabidopsis thaliana Ty3/gypsy* retrotransposons that encode *envelope*-like proteins. Genetics **149:** 703–715.

Xu, Z., X. Yan, S. Maurais, H. Fu, D. G. O'Brien *et al.*, 2004 *Jittery*, a *mutator* distant relative with a paradoxical mobile behavior: excision without reinsertion. Plant Cell **16:** 1105–1114.

Yamazaki, M., H. Tsugawa, A. Miyao, M. Yano, J. Wu *et al.*, 2001 The rice retrotransposon Tos17 prefers low-copy-number sequences as integration targets. Mol. Genet. Genomics **265:** 336–344.

Young, N. D., J. Mudge and T. H. Ellis, 2003 Legume genomes: more than peas in a pod. Curr. Opin. Plant Biol. **6:** 199–204.

Young, N. D., S. B. Cannon, S. Shusei, K. Dongjin, D. R. Cook *et al.*, 2005 Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol. **137:** 1174–1181.

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, X., and S. R. Wessler, 2004 Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. Proc. Natl. Acad. Sci. USA **101:** 5589–5594.

Zhang, X., N. Jiang, C. Feschotte and S. R. Wessler, 2004 *PIF*- and *Pong*-like transposable elements: distribution, evolution and relationship with *Tourist*-like miniature inverted-repeat transposable elements. Genetics **166:** 971–986.

Communicating editor: D. Voytas