INTEGRATING COMPUTATIONAL AND EXPERIMENTAL ANALYSES TO STUDY PLANT TRANSPOSABLE ELEMENTS

by

DAWN H HOLLIGAN

(Under the Direction of Susan R Wessler)

ABSTRACT

Transposable elements (TEs) are the single largest component of most eukaryotic genomes and account for more than half the DNA content in some of these organisms. This dissertation contains several studies designed to investigate the abundance and amplification of different TE types in plant genomes.

The research presented in the first half of this dissertation focuses on the identification and analysis of TEs in the model legume *Lotus japonicus* (*Lotus*) using a combined computational and experimental approach. *Lotus* belongs to the Leguminosae family, one of the largest plant families, containing several agronomically important species, such as soybean and garden pea. The availability of a significant amount (32.4Mb at the time of this study) of *Lotus* genome sequence has permitted, for the first time, comprehensive TE analysis in a legume species. While computer-assisted analysis facilitated a determination of TE abundance and diversity, the availability of complete BAC sequences permitted identification of full-length TEs, which facilitated the design of tools for genome wide experimental analysis.

The second half of this dissertation presents research aimed at understanding how MITEs are able to amplify to very high copy numbers in the host genome. MITEs are highly abundant in plants and animals, comprising > 100,000 copies in *Oryza sativa* (rice) and ~16% in *A. Aegypti (mosquito)*. The analysis described in Chapter 4 involves the use of a large-scale yeast assay to examine the functional relationship between the rice *Stowaway*-like MITEs and their putative transposase sources, *Osmars* (*Tc1/Mariner*-like). Results from these analyses provide insight into different TE superfamilies and how they interact within host genomes.

INDEX WORDS: Transposable elements, *Lotus japonicus*, TE abundance, TE diversity, recently amplified TEs, Sireviruses, Pack-MULEs, *Osmars*, *Stowaway*-like MITEs.

INTEGRATING COMPUTATIONAL AND EXPERIMENTAL ANALYSES TO STUDY PLANT TRANSPOSABLE ELEMENTS

by

DAWN H HOLLIGAN

B.S., University of Georgia, 2000

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

INTEGRATING COMPUTATIONAL AND EXPERIMENTAL ANALYSES TO STUDY PLANT TRANSPOSABLE ELEMENTS

by

DAWN H HOLLIGAN

Major Professor: Susan R Wessler

Committee: R. Kelly Dawe
Katrien Devos
Jeff Bennetzen
Xiaoyu Zhang

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

ACKNOWLEDGEMENTS

I offer my sincere thanks and gratitude to my major advisor Susan Wessler for all of her support, guidance and encouragement. I would also like to thank my committee members for all of their support and guidance during my graduate career. To the current and past members of the Wessler Lab, Xiaoyu Zhang, Ning Jiang, Cedric Feschotte, Ellen Pritham, Eunyoung Cho, Guojun Yang, Eleanor Kuntz, Yujun Han, Feng Zhang and Alex Nagel, a sincere thank you for all of your assistance and helpful conversations. I thank the Plant Biology Department staff, Susan Watkins and Carla Ingram, for their administrative assistance. Finally, I would like to thank my mother Pauline Michael, my aunts Phyllis Thomas and Denese Stubblefield, and my wonderful supportive and caring husband Alex Nagel for their never-ending support.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Transposable elements (TEs) are DNA fragments that can insert into new chromosomal locations in the genome, and in doing so, often duplicate in the process. First discovered by Barbara McClintock more than 50 years ago, TEs have since been found in a variety of organisms (McClintock 1948; McClintock 1949; Capy *et al.* 1998). Furthermore, TEs are the single largest component in most eukaryotic genomes accounting for ~ 50% of the DNA content in humans and >70 of the maize genome (SanMiguel and Bennetzen 1998; Lander *et al.* 2001). Eukaryotic TEs are generally divided into two classes according to whether their transposition intermediate is RNA (class 1) or DNA (class 2). Class 1 elements transpose via a copy and paste mechanism involving an RNA intermediate and as a result can increase their copy number due to transposition (Capy *et al.* 1998; Craig *et al.* 2002). In contrast, class 2 elements that transpose via a cut and paste mechanism are unable to increase their copy due to transposition. Instead class 2 elements rely on the host gap repair mechanism and replication machinery to increase their copy number. If the host repair the excision gap using a copy from the homologous chromosome or sister chromatids and/or the excised copy reinserts ahead of the replication fork, new copies are gained (Fedoroff 1989; Gloor *et al.* 1991; Nassif *et al.* 1994;

WILSON *et al.* 2003). Each TE class contains both coding and noncoding elements (also called autonomous and nonautonomous elements). Coding elements encode the proteins involved in the transposition reaction(s). In contrast, noncoding elements do not encode transposition-associated proteins but can be mobilized by coding elements because they retain the cis-sequences necessary for transposition (CRAIG *et al.* 2002; FESCHOTTE *et al.* 2002a).

Class 1 elements include retrotransposons with long terminal repeats (LTR-retrotransposons) and non-LTR retrotransposons, also called long interspersed nuclear elements (LINEs) and small interspersed nuclear elements (SINEs). All class 2 elements, except *Helitrons*, have short terminal inverted repeats (TIRs) and are grouped into superfamilies based on the similarities of their encoded transposes [e.g., *Tc1/mariner*, *hAT*, CACTA, *Mutator*-like elements (MULEs), and *PIF/Pong*], the enzyme that binds to the end of the TE and catalyzes both excision and reintegration. In addition to the superfamilies described below, Class 2 elements also include a special group of noncoding elements called miniature inverted repeat transposable elements (MITEs) that are present in very high copy numbers in several eukaryotic genomes (see literature review below) (CAPY *et al.* 1998; CRAIG *et al.* 2002; ZHANG and WESSLER 2004).

Based on TE analyses from several available genomic sequences, it is becoming clear that the relative abundance of both classes of TEs varies across eukaryotic genomes (KIDWELL 2002). For example, while the genome of vertebrates is dominated by class 1 TEs (largely non-LTR retrotransposons) with

only a single active class 2 element, plant genomes contain a wealth of class 1 and class 2 elements including a few active TE families (KOGA and HORI 2001; YAMAZAKI *et al.* 2001; JIANG *et al.* 2003). Although the TE abundance and, in some cases, activity have been reported for several plant species, analyzing the TE abundance across diverse taxa is still necessary for several reasons. Besides the reasons stated above, TE characterization across the plant kingdom will provide a more representative dataset for TE comparative studies between closely and distantly related species. Second, new genome analysis could potentially identify active elements that can be used to understand the transposition mechanism for the different superfamilies of TEs. Presently, with a variety of genomic sequences (shotgun reads and BACs) available and more to come, it is now feasible to choose the best organism for studying TEs.

The research presented in the first half (Chapter 2 and Chapter 3) of this dissertation focuses on the TE analysis of the model legume *Lotus japonicus* (*Lotus*) using a combined computational and experimental approach. *Lotus* belongs to the Leguminosae family, one of the largest plant families, containing several agronomically important species, such as soybean and garden pea (JIANG and GRESSHOFF 1997). Besides providing the largest single source of vegetable protein in human diets and livestock feed, legumes form a symbiotic relationship with the Rhizobium family of bacteria and, as such, aid in massive production of biological nitrogen (JIANG and GRESSHOFF 1997; SATO *et al.* 2001). *Lotus* is an ideal model organism within the legume family because of its small genome size (~470), short-life cycle (2-3 months) and the ease of genetic

manipulation (JIANG and GRESSHOFF 1997). For these reasons, a large-scale sequencing project was initiated and 32.4Mb (6.9%) of genomic sequences were available at the time of this study (SATO *et al.* 2001; NAKAMURA *et al.* 2002; ASAMIZU *et al.* 2003; KANEKO *et al.* 2003; KATO *et al.* 2003; YOUNG *et al.* 2005). These available sequences have permitted, for the first time, a survey of TEs within the legume family. Using a computational approach, both coding and noncoding TEs representing all known TE superfamilies were identified in the available 32.4Mb *Lotus* sequence. A detailed description of the TE abundance and diversity of the different TE superfamilies (reviewed below), as well as new TE lineages and putative active elements, is described in Chapter 2. Chapter 3 describes the transposition analysis performed on two recently amplified elements (putative active candidates) discovered during Chapter 2 studies. This report provides for the first time a comprehensive survey of TEs in any legume genome.

The second half of this dissertation presents research aimed at understanding how MITEs are able to amplify to very high copy numbers in the host genome. MITEs are short noncoding DNA elements that are highly abundant in eukaryotic genomes (reviewed below) (FESCHOTTE *et al.* 2002b). Since being discovered more than 15 years ago, one of the main questions concerning MITEs has been, how they are able to amplify to such high copy numbers given the conservative (cut and paste) mechanism of transposition for DNA transposons. It has been proposed that MITEs might be able to use the transposase source necessary for transposition from related coding elements

(FESCHOTTE et al. 2002b; JIANG et al. 2004) or that MITEs possess special features in their sequence that provides an advantage for transposition. Based on these two hypotheses, the experiments described in Chapter 3 directly address how MITEs are able to amplify to high copy numbers. This analysis involved the use of a large-scale yeast excision assay system described previously, to examine the functional relationship between the rice *Stowaway*-like MITEs and their putative transposase sources *Osmars* (*Tc1/Mariner*-like).

Literature Review of Class 1 and Class 2 Elements

Class 1 Elements. LTR retrotransposons are the most prevalent TE type in plant genomes and are further divided into two groups, *Ty1/copia*-like or *Ty3/gypsy*-like based on the order of their encoded proteins that include reverse transcriptase (RT) and integrase (KUMAR and BENNETZEN 1999; BEAUREGARD *et al.* 2008). Structurally, LTR retrotransposons contain long terminal direct repeats that in most cases begin and end with a dinucleotide inverted repeat (5'-TG and 3'-CA, respectively), a 4-6 bp target site duplication, a polypurine tract (PPT) downstream of the 5' LTR and a primer binding site upstream of the 3' LTR (CRAIG *et al.* 2002). The internal regions of LTR retrotransposons contain two genes that are required for transposition of the element. The *GAG* gene encodes a capsid protein that is involved in the formation of virus-like intracellular particles that are the site of reverse transcription. The *POL* gene encodes protease (PR) which cleaves the *Pol* polyprotein, integrase (IN) involved in integration of the cDNA into the genome, reverse transcriptase (RT) which copies the RNA into

cDNA and RnaseH (RH) involved in degradation of the RNA template, etc. The main structural difference between *Ty1/copia*-like and *Ty3/gypsy*-like elements is the order of the proteins encoded by the *POL* gene (KUMAR and BENNETZEN 1999; CRAIG *et al.* 2002).

LTR retrotransposons with a third ORF. In addition to the *GAG* and *POL* genes, some LTR retrotransposons (including both *Ty1*/*copia*-like and *Ty3*/*gypsy*-like elements) have been identified in several plant species that contain a third ORF. The first elements of this type identified were the *Athila gypsy*-like element from *Arabidopsis* and the SIRE-1 *copia*-like elements from soybean (LATEN 1999; PETERSON-BURCH *et al.* 2000). LTR retrotransposons containing a third ORF are considered to be retrovirus-like (*env*-like) based on similarity of the third ORF with the envelope ORF of retroviruses. This similarity is primarily limited to the presence of such features as a transmembrane and a coil-coil domain. Recently, this group of LTR retrotransposons has been given the name *Sireviruses* to reflect the founding SIRE-1 element (from soybean) and their structural affinity with retroviruses (LATEN *et al.* 2003). The third ORF of different *Sireviruses* usually have very little sequence similarity (~20%). A recent molecular study revealed that the additional ORF of SIRE-1 uses stop-codon suppression to express their *env*-like protein. This is in contrast to most retroviruses that express their envelope protein from a spliced genomic mRNA (HAVECKER *et al.* 2005). However, the actual function of this extra ORF in the context of LTR retrotransposition is yet to be determined.

Non-LTR retrotransposons: LINEs and SINEs. LINEs are retrotransposons that do not contain flanking direct repeats. However, LINEs do contain a 5' UTR and a 3' UTR that ends with a poly (A) tail (CRAIG *et al.* 2002). The internal domain of LINEs contains two ORFs that are both required for transposition. ORF1 (*GAG*-like) encodes a nucleic acid binding protein involved in nucleic acid chaperone activity and ORF2 encodes a putative endonuclease protein (EN), reverse transcriptase (RT), both involved in retrotransposition, and a cysteine-histidine-rich domain (C) considered a putative nucleic acid binding domain (BOEKE 1997; CRAIG *et al.* 2002).

SINE elements are nonautonomous non-LTR retrotransposons. The 5' halves of several SINEs resemble tRNA genes and for this reason, SINEs have been proposed to arise from tRNAs (DEININGER 1989; CRAIG *et al.* 2002). However the 3' end of SINEs is similar to the 3' end of LINEs with both ending with a poly (A) tract or A or T rich sequences. This similarity suggests that SINEs probably utilize proteins encoded by LINEs for retrotransposition (OKADA 1991). LINE and SINE elements are usually the predominant type of class 1 elements found in mammalian genomes (SMIT 1996).

Class 2 Elements. As mentioned in the introduction, all class 2 elements, except *Helitrons*, have short terminal inverted repeats (TIRs) and are grouped into superfamilies based on the similarities of their encoded transposases.

Tc1*/*Mariners. The *Tc1*/*Mariner*-like superfamily was initially discovered in *C.elegans* (*Tc1*) and *Drosophila* (*Mariner*) and is now known to be a diverse

group of elements found in the genomes of most higher eukaryotes (FESCHOTTE and WESSLER 2002) at relatively low copy numbers (<100 copies). *Tc1*/*Mariner*-like elements contain terminal inverted repeats (~20–30 bp) that vary across species and a conserved target duplication of TA. The internal coding sequence contains a single ORF, which encodes the transposase protein that is required for transposition. This transposase contains a DDE or DDD motif that is found in most eukaryotic transposases and integrases (CRAIG *et al.* 2002). Several *Tc1*/*Mariner*-like elements are currently active in the genomes of diverse taxa including *Tc1* and *Tc3* from *C.elegans*, *Minos* and *Mos1* from flies, and *Impala* and *Fot1* from fungus (*F. oxysporum*) (EMMONS *et al.* 1983; COLLINS *et al.* 1989; DABOUSSI *et al.* 1992; FRANZ *et al.* 1994; LANGIN *et al.* 1995; DAWSON and FINNEGAN 2003). However, no *Tc1*/*Mariner*-like elements have been shown to be active in plants, although the rice *Tc1*/*Mariner*-like element *Osmar5* has been demonstrated to be active in a yeast transposition assay (YANG *et al.* 2006). *Tc1*/*Mariner*-like elements are usually grouped into three major clades and have a wide phylogenetic distribution unrelated to phylogenetic relationships of the hosts. This has led to widespread speculation that horizontal transfer may play a role in the spread of this group of elements (CRAIG *et al.* 2002; FESCHOTTE and WESSLER 2002).

hATs. First identified in maize, the *hAT* superfamily is widespread in eukaryotes and contains several currently active members. *hAT* elements contain the active founding members *Ac* (coding, autonomous) and *Ds*

(noncoding, nonautonomous) from maize (BANKS *et al.* 1985; CRAIG *et al.* 2002). *hATs* consist of terminal inverted repeats that are flanked by an 8 bp target site duplication and a transposase protein. The TIR and parts of the subterminal region have been reported to be required for transposition (WEIL and KUNZE 2000). Well-studied and currently active members of the *hAT* superfamily include maize *Ac/Ds*, *hobo* from *Drosophila melanogaster, Tam3* from *Antirrhinum majus*, *Dart/ndart* from *Oryza sativa*, *Arabidopsis thaliana Tag1* and *Tag2,* and *dTph* from *Petunia hybrida* (MCCLINTOCK 1956; DORING and STARLINGER 1984; LIM 1988; MARTIN *et al.* 1989; TSAY *et al.* 1993; HENK *et al.* 1999; FUJINO *et al.* 2005). The *hAT* superfamily was named after the representative *hobo*, *Ac* and *Tam3* elements. All characterized *hAT*-like elements fall into three distinct but closely related clades suggesting that they share common ancestry (RUBIN *et al.* 2001; CRAIG *et al.* 2002).

CACTA. Elements belonging to the CACTA superfamily contain a short (~13 bp) terminal inverted repeat with a signature of 5'-CACTA-3' at the most terminal end and they are flanked by a 3 bp target site duplication (KUNZE and WEIL 2002). However, unlike *hAT* elements, CACTA elements encode two proteins (TNP1-like and TNP2), both involved in transposition and derived from alternative splicing of the same mRNA precursor (TRENTMANN *et al.* 1993; CRAIG *et al.* 2002). Some members of this superfamily include the active founding member *En/Spm* of maize, *Cs1* of sorghum, *Tam1* and *Tam2* elements of *Antirrhinum* and *Rim2/Hipa* elements of rice (BONAS *et al.* 1984; UPADHYAYA *et al.*

1985; GIERL and SAEDLER 1989; CHOPRA *et al.* 1999; WANG *et al.* 2003). Furthermore, CACTA-like elements in plants phylogenetically group into two major clades (CRAIG *et al.* 2002; ZHANG and WESSLER 2004).

PIF/Harbinger. The maize *PIF and Arabidopsis Harbinger elements* belong to a recently described superfamily that has since been found in the genomes of most characterized higher eukaryotes (KAPITONOV and JURKA 1999; CRAIG *et al.* 2002; ZHANG *et al.* 2004). Most recently, the rice *Ping* and *Pong*-like elements have been added to this superfamily based on sequence and structure similarity (JIANG *et al.* 2003; ZHANG *et al.* 2004). Structurally, these elements contain terminal inverted repeats (~30 bp), target site duplications (TAA/TTA) and are usually 4-5 kb (ZHANG *et al.* 2004). Unlike most DNA transposons, *PIF and Pong*-like elements contain two open reading frames (ORFs 1 and 2). ORF1 contains a domain that shares weak similarity to the DNA binding region of *Myb* transcription factors, and as such is predicted to be involved in DNA binding of transposon ends (JIANG *et al.* 2003; ZHANG *et al.* 2004). ORF2 encodes the transposase and contains the catalytic DDE motif. Both ORFs are required for transposition (YANG *et al.* 2007). Chapter 3 of this dissertation focuses on two members of this superfamily, the *Lotus Pong*-like elements.

***Mutator*-like elements.** *Mutator*-like elements (MULEs) are related to the founding maize *MuDR* except that they lack the *MudrB* protein present in all *MuDR* elements (ROBERTSON 1981; CRAIG *et al.* 2002) (ROBERTSON 1981; LISCH

2002). *Mutator*-like elements contain long terminal inverted repeats (~200 bp) that are flanked by a 8-10 bp TA rich target site duplication. The transposase protein *MudrA* is required for transposition and is encoded by a single ORF. Structurally, *Mutator*-like elements in numerous plant species are more similar to the maize *Mutator*-like elements (e.g., *Jittery*) than to *MuDR* (LISCH 2002; XU *et al.* 2004). Furthermore, unlike the typical non-coding DNA elements that are homologous deletion derivates of coding elements, noncoding *Mutator*-like elements often share only TIR similarity with the coding elements and instead a variety of sequences are present between the TIRs including host gene fragments (YU *et al.* 2000) (see Pack-MULEs below).

Pack-MULEs. Pack-MULE is the name given to *Mutator*-like elements that have captured fragments of host genes (YU *et al.* 2000; JIANG *et al.* 2004a). These elements contain TIRs and TSD similar to typical *Mutator*-like elements. Many Pack-MULEs contain gene fragments from multiple chromosomal loci that are fused to form new open reading frames (JIANG *et al.* 2004a). Pack-MULEs were first discovered in maize and later were found to be abundant in the rice genome, where over 3000 copies were identified (TALBERT and CHANDLER 1988; JIANG *et al.* 2004a). Even the small genome of *Arabidopsis* contains a few Pack-MULEs (36 copies), suggesting that gene capture by MULEs might be a widespread occurrence in plant genomes and might contribute to the creation of new genes. With the availability of a vast number of genomic sequences, and the fact that *Mutator*-like elements are widespread in eukaryotes, a more

comprehensive representation of Pack-MULEs across genomes is needed to determine the range of this phenomenon.

Helitrons. *Helitrons* are a new superfamily of DNA transposons containing both coding and non-coding copies (FESCHOTTE and WESSLER 2001; KAPITONOV and JURKA 2001). Structurally, they lack TIRs and TSD but have short conserved terminal motifs, 5'TC and a 3'CTRR (KAPITONOV and JURKA 2001). The internal region of coding *Helitrons* contains ORFs that share similarity with DNA helicases and to the replicator initiator protein of rolling circle plasmids (KAPITONOV and JURKA 2001). In addition a 16-20 bp palindromic structure is located a few base pairs upstream of the 3'CTRR (KAPITONOV and JURKA 2001). Unlike typical DNA transposons, *Helitrons* have been hypothesized to move via a rolling circle transposition mechanism (KAPITONOV and JURKA 2007). Initially discovered computationally in *Arabidopsis*, rice and *C.elegans*, *Helitrons* have been identified in most of the available eukaryotic genomic sequences (FESCHOTTE and WESSLER 2001; POULTER 2003; ZHOU *et al.* 2006; HOLLISTER and GAUT 2007; PRITHAM and FESCHOTTE 2007). Most *Helitrons* are noncoding elements because they lack the proteins required for transposition and can be very small in size (~200 bp, *C.elegans*)(KAPITONOV and JURKA 2007). However, many maize *Helitrons*, like Pack-MULEs, contain gene fragments that have been captured from more than one chromosomal locus (GUPTA *et al.* 2005; LAI *et al.* 2005; MORGANTE *et al.* 2005). A more recent study reports the presence of *Helitrons* carrying cellular genes in *Arabidopsis*, rice and bats (*M.*

lucifugus) suggesting that these types of elements are not limited to maize (RAY *et al.* 2008; SWEREDOSKI *et al.* 2008).

Miniature inverted repeat transposable elements (MITEs). MITEs are a unique group of noncoding DNA transposons that are characterized by their high copy number and small size (FESCHOTTE *et al.* 2002b). MITEs were initially discovered in maize and since then have been found in most eukaryotes (thousands per family) (BUREAU and WESSLER 1994; CRAIG *et al.* 2002). For example, MITEs account for ~16% of the mosquito genome (*A. aegypti*) (NENE *et al.* 2007). Like other class 2 elements, MITEs contain terminal inverted repeats of variable length and they generate a target site duplication upon insertion (FESCHOTTE *et al.* 2002a). MITEs preferentially accumulate near or within genes (introns and UTRs) (MAO *et al.* 2000; ZHANG *et al.* 2000). This preference, along with their abundance, suggests that MITEs may play an important role in the evolution of genes (FESCHOTTE *et al.* 2002a) (OKI *et al.* 2008).

MITEs are classified into two groups based on their sequences: *Tourist* and *Stowaway*. *Tourist-like* MITEs usually contain ~15 bp TIRs flanked by a TAA or TTA target site duplication, and *Stowaway*-like MITEs contain ~30 bp TIRs flanked by a TA target site duplication. While most MITEs do not share any sequence similarity with coding elements besides a few bases in the TIRs (FESCHOTTE *et al.* 2002b), some MITEs such as the active rice *mPing* element are direct deletion derivatives of a coding elements (JIANG *et al.* 2003). Based on TIR and TSD similarities, the *PIF/Harbinger* superfamily of elements is

considered the likely transposase source for *Tourist*-like MITEs (JURKA and KAPITONOV 2001; JIANG *et al.* 2004b; ZHANG *et al.* 2004). This association has been experimentally demonstrated with the mobilization of the rice *Tourist*-like MITE *mPing* by *Ping* and *Pong* elements. For *Stowaway*-like MITEs, the likely transposase source based on structural similarity (TIR and TSD) is the *Tc1*/*Mariner-like* superfamily (FESCHOTTE *et al.* 2002b; JIANG *et al.* 2004b). Furthermore, a previous biochemical study demonstrated the ability of the putative transposase encoded by the rice *Tc1*/*Mariner*-like element *Osmar* to bind to distantly related *Stowaways* (FESCHOTTE *et al.* 2002b; FESCHOTTE *et al.* 2005). One chapter of this dissertation is focused on the functional relationship between *Tc1/Mariner*-like elements and *Stowaway*-like MITEs in rice.

TE Abundance and Activity

TEs are highly abundant in eukaryotes, accounting for almost 50% of the human and mosquito (*A. aegypti*) genomes, and ~15 - 20% and ~30% of the sequenced *Arabidopsis thaliana* and *Oryza sativa* (rice) genomes, respectively (LANDER *et al.* 2001; TURCOTTE *et al.* 2001; NENE *et al.* 2007; LIU and BENNETZEN 2008). In the larger plant genomes including maize and barley, TEs (mostly LTR retrotransposons) account for more than 75% of their genomic content (KUMAR and BENNETZEN 1999; VICIENT *et al.* 1999). Studies in maize and *Oryza australiensis* revealed that the amplification of LTR-retrotransposons is largely responsible for genome size doublings that can occur in less than 5 million years (BENNETZEN *et al.* 1998; PIEGU *et al.* 2006). Even in single-celled organisms such

as *Trichomonas* (*T. vaginalis*), TEs (mostly DNA transposons) account for ~65% of the genome (CARLTON *et al.* 2007). Although, in plants DNA transposons do not generally contribute significantly to genome expansion, some families are present in very high copy numbers. For example, MITEs can account for significant fractions of some eukaryotic genomes (~6% of rice and 16% of mosquito) (NENE *et al.* 2007; OKI *et al.* 2008).

Although TEs make up a significant fraction of most plant genomes, only a few have been shown to be active under normal conditions such as the maize *Ac/Ds* and *Mutator* elements and the rice *Ping* and *mPing* elements (MCCLINTOCK 1956; ALLEMAN and FREELING 1986; JIANG *et al.* 2003). This is probably due to the fact that most TEs are full of inactivating mutations. Intact TEs are further controlled (silenced) by host regulatory mechanism(s) (FESCHOTTE and PRITHAM 2007). For example in *Arabidopsis*, most TEs are silenced by DNA methylation, but can be reactivated in *Arabidopsis* methylation mutants (ZHANG *et al.* 2006; ZHANG 2008). In some cases, treating plants with drugs that result in decreased DNA methylation can also reactivate TEs. For example, *hAT* elements in rice (*Dart*) and *medaka* fish (*Tol2*) are transpositionally activated by 5-azacytidine (decreased methylation) treatments (IIDA *et al.* 2006; TSUGANE *et al.* 2006). The maize *Ac* element is also mobilized in transgenic *Arabidopsis* lines after treatment with 5-azacytidine (SCORTECCI *et al.* 1997). Furthermore, under abiotic or biotic stresses, many transposons, in addition to those mentioned above, are either transcriptionally and/or transpositionally activated. The LTR retrotransposons *Tos17* in rice and *Tto1* in tobacco are both activated in tissue

culture lines (HIROCHIKA 1995; HIROCHIKA *et al.* 2000). In addition, several TEs can be activated when introduced into heterologous hosts. For example, the rice *Tc1/Mariner*-like element *Osmar5* is active in yeast and *Tto1* from tobacco is active in both *Arabidopsis* and rice (HIROCHIKA *et al.* 1996; YANG *et al.* 2006).

Strategies to Identify TEs Computationally

The identification of TEs can be quite challenging, depending on the type of TE family and the quality of the genomic sequence being studied. Within the last few years, a variety of computer programs have been developed to identify TEs. Some of these programs identify TEs by using known TEs as queries (e.g. Repeatmasker and MITE Analysis Kit), while others (e.g. RECON and FINDMITE) are based on either identifying repeated sequences in a genome or on structural criteria (e.g. LTRs, TIRs, TSDs) (TU 2001; BAO and EDDY 2002) (YANG and HALL 2003). For a comprehensive TE analysis, no one program is sufficient to identify all TEs in a genome. For example, although, RepeatMasker can efficiently identify known TEs, it is unable to identify novel TEs. In contrast, programs such as RECON that can identify novel TEs require further TE annotation of the repeat output files to determine the TE type identified (Jiang, N., personal communication; (BAO and EDDY 2002)). Therefore, using a combination of these programs will most likely provide the most accurate information on TE content in a genome.

However, regardless of the approach, the TE information generated is also limited by the database used for the analysis. Genomic databases such as

whole genome shotgun reads or BAC end sequences (~ 600 bp reads) are only useful for identifying MITEs and conserved protein regions (catalytic domain and reverse transcriptase) in coding TEs because these sequences are usually less than 600 bp (ZHANG and WESSLER 2004). In contrast, assembled genome sequences and whole BACs (~100,000 bp) are the most useful in that both MITEs and large (> 600 bp up to 15 kb) full-length TEs can be identified (INITIATIVE 2000; LE *et al.* 2000). The identification of full-length TEs that includes defined terminal sequences is crucial for genome-wide experimental analysis. Chapter 2 of this dissertation presents a study on whole BACs from the *Lotus japonicus* genome and illustrates the importance of having whole BAC sequences for TE studies.

References

ALLEMAN, M., and M. FREELING, 1986 The Mu transposable elements of maize: evidence for transposition and copy number regulation during development. Genetics **112:** 107-119.

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 Research **10:** 115-122.

BANKS, J., J. KINGSBURY, V. RABOY, J. W. SCHIEFELBEIN, O. NELSON *et al.*, 1985 The *Ac* and *Spm* controlling element families in maize. Cold Spring Harbor Symposium on Quantitative Biology **50:** 307-311.

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

BEAUREGARD, A., M. J. CURCIO and M. BELFORT, 2008 The Take and Give Between Retrotransposable Elements and Their Hosts. Annu Rev Genet. **42**.

BENNETZEN, J. L., P. SANMIGUEL, M. CHEN, A. TIKHONOV, M. FRANCKI *et al.*, 1998 Grass genomes. PNAS **95:** 1975-1978.

BOEKE, J. D., 1997 LINEs and Alus--the polyA connection [news; comment]. Nat Genet **16:** 6-7.

BONAS, U., H. SOMMER, B. J. HARRISON and H. SAEDLER, 1984 The transposable element *Tam1* of *Antirrhinum majus* is 17-kb long. Mol.Gen.Genet. **194:** 138-143.

BUREAU, T. E., and S. R. WESSLER, 1994 Mobile inverted-repeat elements of the *Tourist* family are associated with genes of many cereal grasses. Proc. Natl. Acad. Sci. USA **91:** 1411-1415.

CAPY, P., C. BAZIN, D. HIGUET and T. LANGIN, 1998 *Dynamics and evolution of transposable elements*. Springer-Verlag, Austin, Texas.

CARLTON, J. M., R. P. HIRT, J. C. SILVA, A. L. DELCHER, M. SCHATZ *et al.*, 2007 Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. Science **315:** 207-212.

CHOPRA, S., V. BRENDEL, J. ZHANG, J. D. AXTELL and T. PETERSON, 1999 Molecular characterization of a mutable pigmentation phenotype and

isolation of the first active transposable element from Sorghum bicolor. Proc Natl Acad Sci U S A **96:** 15330-15335.

COLLINS, J., E. FORBES and P. ANDERSON, 1989 The Tc3 family of transposable genetic elements in Caenorhabditis elegans. Genetics **121:** 47-55.

CRAIG, N. L., R. CRAIGIE, M. GELLERT and A. M. LAMBOWITZ, 2002 *Mobile DNA II*. American Society for Microbiology Press, Washington, D.C.

DABOUSSI, M.-J., T. LANGIN and Y. BRYGOO, 1992 Fot1, a new family of fungal transposable elements. Mol. Gen. Genet. **232:** 12-16.

DAWSON, A., and D. J. FINNEGAN, 2003 Excision of the Drosophila mariner transposon Mos1. Cell **11:** 225-235.

DEININGER, P. L., 1989 SINEs: Short Interspersed Repeated DNA Elements in higher eukaryotes, pp. 619-636 in *Mobile DNA*, edited by D. E. BERG and M. M. HOWE. American Society for Microbiology, Washington, DC.

DORING, H. P., and P. STARLINGER, 1984 Barbara McClintock's controlling elements: now at the DNA level. Cell **39:** 253-260.

EMMONS, S. W., L. YESNER, K. S. RUAN and D. KATZENBERG, 1983 Evidence for a transposon in Caenorhabditis elegans. Cell **32:** 55-65.

FEDOROFF, N., 1989 Maize transposable elements, pp. 375-411 in *Mobile DNA*, edited by D. E. BERG and M. M. HOWE. American Society for Microbiology, Washicngton, 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., M. T. OSTERLUND, R. PEELER and S. R. WESSLER, 2005 DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. Nucleic Acids Res. **33:** 2153-2165.

FESCHOTTE, C., and E. J. PRITHAM, 2007 DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. **41:** 331-368.

FESCHOTTE, C., and S. R. WESSLER, 2001 Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci U S A **98:** 8923-8924.

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., 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.

FRANZ, G., T. G. LOUKERIS, G. DIALEKTAKI, C. R. THOMPSON and C. SAVAKIS, 1994 Mobile *Minos* elements from *Drosophila hydei* encode a two-exon transposase with similarity to the paired DNA-binding domain. Proc. Natl. Acad. Sci. USA **91:** 4746-4750.

FUJINO, K., H. SEKIGUCHI and T. KIGUCHI, 2005 Identification of an active transposon in intact rice plants. Mol Genet Genomics. **273:** 150-157.

GIERL, A., and H. SAEDLER, 1989 The En/Spm transposable element of Zea mays. Plant Mol Biol **13:** 261-266.

GLOOR, G. B., N. A. NASSIF, D. M. JOHNSONSCHLITZ, C. R. PRESTON and W. R. ENGELS, 1991 Targeted gene replacement in Drosophila via P element-induced gap repair. Science **253:** 1110-1117.

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, 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 Physiology **139:** 857-868.

HENK, A. D., R. F. WARREN and R. W. INNES, 1999 A new Ac-like transposon of Arabidopsis is associated with a deletion of the RPS5 disease resistance gene. Genetics **151:** 1581-1589.

HIROCHIKA, H., 1995 Activation of plant retrotransposons by stress, pp. in press in *Modification of Gene Expression and Non-Mendelian Inheritance*, edited by K. OONO. National Institute of Agrobiological Resources, Japan.

HIROCHIKA, H., H. OKAMOTO and T. KAKUTANI, 2000 Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. Plant Cell **12:** 357-369.

HIROCHIKA, H., H. OTSUKI, M. YOSHIKAWA, Y. OTSUKI, K. SUGIMOTO *et al.*, 1996 Autonomous transposition of the tobacco retrotransposon *Tto1* in rice. The Plant Cell **8:** 725-734.

HOLLISTER, J. D., and B. S. GAUT, 2007 Population and evolutionary dynamics of Helitron transposable elements in Arabidopsis thaliana. Mol Biol Evol. **24:** 2515-2524.

IIDA, A., A. SHIMADA, A. SHIMA, N. TAKAMATSU, H. HORI *et al.*, 2006 Targeted reduction of the DNA methylation level with 5-azacytidine promotes excision of the medaka fish Tol2 transposable element. Genet Res. **87:** 187-193.

INITIATIVE, A. G., 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

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., 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., 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). Current Opinion in Plant Biology **7:** 115-119.

JIANG, Q., and P. M. GRESSHOFF, 1997 Classical and molecular genetics of the model legume *Lotus japonicus*. Molecular Plant Microbe Interactions **10:** 59-68.

JURKA, J., and V. V. KAPITONOV, 2001 PIFs meet Tourists and Harbingers: A superfamily reunion. Proc Natl Acad Sci U S A **98:** 12315-12316.

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 Research **10:** 27-33.

KAPITONOV, V. V., and J. JURKA, 1999 Molecular paleontology of transposable elements from Arabidopsis thaliana. Genetica **107:** 27-37.

KAPITONOV, V. V., and J. JURKA, 2001 Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. USA **98:** 8714-8719.

KAPITONOV, V. V., and J. JURKA, 2007 Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. **23:** 521-529.

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 Research **10:** 277-285.

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 & Genetic Systems **76:** 1-8.

KUMAR, A., and J. L. BENNETZEN, 1999 Plant retrotransposons. Annu Rev Genet **33:** 479-532.

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.

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 U S A. **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.

LANGIN, T., P. CAPY and M. J. DABOUSSI, 1995 The transposable element impala, a fungal member of the Tc1-mariner superfamily. Mol. Gen. Genet **246:** 19-28.

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

LATEN, H. M., E. R. HAVECKER, L. M. FARMER and D. F. VOYTAS, 2003 SIRE1, an Endogenous Retrovirus Family from *Glycine max*, Is Highly Homogeneous and Evolutionarily Young. Molecular Biology and Evolution **20:** 1222-1230.

LE, Q. H., S. WRIGHT, Z. YU and T. BUREAU, 2000 Transposon diversity in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **97:** 7376-7381.

LIM, J. K., 1988 Intrachromosomal rearrangements mediated by *hobo* transposons in *Drosophila melanogaster*. pnas **85:** 9153-9157.

LISCH, D., 2002 *Mutator* transposons. Trends in Plant Science **7:** 497-504.

LIU, R., and J. L. BENNETZEN, 2008 Enchilada redux: how complete is your genome sequence? New Phytol **179:** 249-250.

MAO, L., T. C. WOOD, Y. YU, M. A. BUDIMAN, J. TOMKINS *et al.*, 2000 Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res. **10:** 982-990.

MARTIN, C., A. PRESCOTT, C. LISTER and S. MACKAY, 1989 Activity of the transposon Tam3 in *Antirrhinum* and tobacco: Possible role of DNA methylation. European Molecular Biology Organization Journal **8:** 997-1004.

MCCLINTOCK, B., 1948 Mutable loci in maize. Carnegie Institution of Washington Yearbook **47:** 155-169.

MCCLINTOCK, B., 1949 Mutable loci in maize. Carnegie Institution of Washington Yearbook **48:** 142-154.

MCCLINTOCK, B., 1956 Controlling elements and the gene. Cold Spring Harbor Symposium on Quantitative Biology **21:** 197-216.

MORGANTE, M., S. BRUNNER, G. PEA, K. FENGLER, A. ZUCCOLO *et al.*, 2005 Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet **9:** 997-1002.

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 Research **9:** 63-70.

NASSIF, N., J. PENNEY, S. PAL, W. R. ENGELS and G. B. GLOOR, 1994 Efficient copying of nonhomologous sequences from ectopic sites via P- element-induced gap repair. Mol Cell Biol **14:** 1613-1625.

NENE, V., J. R. WORTMAN, D. LAWSON, B. HAAS and E. AL., 2007 Genome sequence of Aedes aegypti, a major arbovirus vector. Science **316:** 1703-1704.

OKADA, N., 1991 SINEs: short interspersed repeated elements of the eukaryotic genome. Trends in ecology and evolution **6:** 358-361.

OKI, N., K. YANO, Y. OKUMOTO, T. TSUKIYAMA, M. TERAISHI *et al.*, 2008 A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica. Genes Genet Syst **83:** 321-329.

PETERSON-BURCH, B. D., D. A. WRIGHT, H. M. LATEN and D. F. VOYTAS, 2000 Retroviruses in plants? Trends Genet **16:** 151-152.

PIEGU, B., R. GUYOT, N. PICAULT, A. ROULIN, A. SANIYAL *et al.*, 2006 Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. **16:** 1262-1269.

POULTER, R. T. M., T J D GOODWIN AND M I BUTLER, 2003 Vertebrate helentrons and other novel Helitrons Gene **313:** 201-212.

PRITHAM, E. J., and C. FESCHOTTE, 2007 Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. Proc Natl Acad Sci U S A. **104:** 1895-1900.

RAY, D. A., C. FESCHOTTE, H. J. PAGAN, J. D. SMITH, E. J. PRITHAM *et al.*, 2008 Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res. **18:** 717-728.

ROBERTSON, D. S., 1981 Mutator Activity in Maize: Timing of Its Activation in Ontogeny. Science **213:** 1515-1517.

RUBIN, E., G. LITHWICK and A. A. LEVY, 2001 Structure and evolution of the hAT transposon superfamily. Genetics **158:** 949-957.

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.

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 Research **8:** 311-318.

SCORTECCI, K. C., Y. DESSAUX, A. PETIT and M. A. VAN SLUYS, 1997 Somatic excision of the Ac transposable element in transgenic Arabidopsis thaliana after 5-azacytidine treatment. Plant Cell Physiol. **38:** 336-343.

SMIT, A. F. A., 1996 The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. **6:** 743-748.

SWEREDOSKI, M., L. DEROSE-WILSON and B. S. GAUT, 2008 A comparative computational analysis of nonautonomous helitron elements between maize and rice. BMC Genomics **9:** 467.

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.

TRENTMANN, S. M., H. SAEDLER and A. GIERL, 1993 The transposable element En/Spm-encoded TNPA protein contains a DNA binding and a dimerization domain. Mol Gen Genet **238:** 201-208.

TSAY, Y. F., M. J. FRANK, T. PAGE, C. DEAN and N. M. CRAWFORD, 1993 Identification of a mobile endogenous transposon in *Arabidopsis thaliana*. Science **260:** 342-344.

TSUGANE, K., M. MAEKAWA, K. TAKAGI, H. TAKAHARA, Q. QIAN *et al.*, 2006 An active DNA transposon nDart causing leaf variegation and mutable dwarfism and its related elements in rice. Plant J **45:** 46-57.

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.

UPADHYAYA, K. C., H. SOMMER, E. KREBBERS and H. SAEDLER, 1985 The paramutagenic line *niv-44* has a 5-kb insert, Tam2, in the chalcone synthase gene of *Antirrhinum majus*. Mol.Gen.Genet. **199:** 201.

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.

WANG, G. D., P. F. TIAN, Z. K. CHENG, G. WU, J. M. JIANG *et al.*, 2003 Genomic characterization of Rim2/Hipa elements reveals a CACTA-like transposon

superfamily with unique features in the rice genome. Mol Genet Genomics. **270:** 234-242.

WEIL, C. F., and R. KUNZE, 2000 Transposition of maize Ac/Ds transposable elements in the yeast saccharomyces cerevisiae [In Process Citation]. Nat Genet **26:** 187-190.

WILSON, R., J. ORSETTI, A. D. KLOCKO, C. ALUVIHARE, E. PECKHAM *et al.*, 2003 Post-integration behavior of a *Mos1* mariner gene vector in *Aedes aegypti.* Insect Biochem Mol Biol. **33:** 853-863.

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.

YANG, G., and T. C. HALL, 2003 MAK, a computational tool kit for automated MITE analysis. Nucleic Acids Res **31:** 3659-3665.

YANG, G., C. F. WEIL and S. R. WESSLER, 2006 A rice Tc1/Mariner-like element transposes in yeast. Plant Cell **18:** 2469-2478.

YANG, G., F. ZHANG, C. N. HANCOCK and S. R. WESSLER, 2007 Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana. Proc Natl Acad Sci U S A. **104:** 10962-10967.

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 Physiology **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, Q., J. ARBUCKLE and S. R. WESSLER, 2000 Recent, extensive and preferential insertion of members of the miniature inverted-repeat transposable element family *Heartbreaker (Hbr)* into genic regions of maize. Proc. Natl. Acad. Sci. USA **97:** 1160-1165.

ZHANG, X., 2008 The Epigenetic Landscape of Plants. Science **320:** 489-492.

ZHANG, X., N. JIANG, C. FESCHOTTE and S. R. WESSLER, 2004 Distribution and evolution of *PIF*- and *Pong*-like transposons and their relationships with *Tourist*-like MITEs. Genetics **166:** 971-986.

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 U S A **101:** 5589-5594.

ZHANG, X., J. YAZAKI, A. SUNDARESAN, S. COKUS, S. W.-L. CHAN *et al.*, 2006 Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. Cell **126:** 1189-1201.

ZHOU, Q., A. FROSCHAUER, C. SCHULTHEIS, C. SCHMIDT, G. P. BIENERT *et al.*, 2006 Helitron Transposons on the Sex Chromosomes of the Platyfish Xiphophorus maculatus and Their Evolution in Animal Genomes. Zebrafish **3:** 39-52.

CHAPTER 2

THE TRANSPOSABLE ELEMENT LANDSCAPE OF THE MODEL LEGUME *LOTUS JAPONICUS* [1]

---

[1] Holligan, D., Zhang, X., Jiang, N., Pritham, E. J. and S.R. Wessler. 2006. Genetics. 174: 2215 – 2228. Reprinted here with permission of publisher

Abstract

The largest component of plant and animal genomes characterized to date is transposable elements. 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 32.4Mb 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 genome wide 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 more than one thousand 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.

Introduction

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 approximately 10% and 30% of the sequenced *Arabidopsis thaliana* (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 over 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 genome-wide 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 almost all TEs except *Helitrons* results in the duplication of a short genomic sequence (called target site duplication, 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 based largely on 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 based on the similarity of their encoded transposases [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 type, miniature-inverted repeat transposable elements (MITEs) are noncoding elements that are frequently associated with plant genes but are also found in animals including zebra fish and *C. 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 chose the best organism to study 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 over 20-fold difference in genome size since the divergence of the grass clade over 50 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.* 2005a; LAI *et al.* 2005).

The approaches used to identify TEs and the information that can be attained is 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 *B. oleracea*), only coding TEs can be identified and their copy numbers can only be approximated 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 TAC (transformation competent bacterial artificial chromosome) 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 particular 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 Hokkaido Region, Japan. Genomic DNA was extracted from leaves of four-week-old seedlings from six individual plants each of MG-20 and B-129 and purified using the DNAeasy plant mini kit (Qiagen).

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 *copia* P2, which was 45°. Primer sequences were: *Bfa*I+0: 5'-GACGATGAGTCCTGAGTAG-3', *Bfa*1+T: 5'-GACGATGAGTC CTGAGTAGT-3',

CACTA P1: 5'-AAATGTTGTTGCGAAAAAGTCGCTG-3', CACTA P2: 5'-CGCTGCGAATTAACTCATCTC-3', *Pong* P1: 5'-CTTKAAG GCTCTCTCCAATG-3', *Pong* P2: 5'-GGTCTTAGCAACTCCAG-3', MULE P1: 5'-A AAGGAGATGGCGGACTTAGC-3', MULE P2: 5'-AGATGGCGGACTTAGCAAA ACAG-3', *copia* P1: 5'-GAGAATAAATCTCCTAATACTG-3', *copia* P2: 5'-CTCCT AATACTGAATATAATMTTC-3', *gypsy* P1: 5'-GCAAAGCGTTTTCTCAAAA GGAC-3', *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 primers pairs designed to amplify regions containing the flanking sequence and the element, to verify if the insertion site was indeed polymorphic. Primers 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 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 terminal inverted repeat (TIR) or the long terminal repeat (LTR). Comparison of two or more closely related elements served to identify the ends of LINEs based on the 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 identity 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 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 through comparisons of closely related sequences or by the presence of key structural hallmarks (AT insertion site, 5' -TC and 3' –CTRR and 15-20 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 based on 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 based on 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

(longer than 40 bp, 7-10 bp target site duplication). Furthermore, if a newly recovered TIR was less than 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 less than 10 kb; (2) sequences between the TIRs should be longer than 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 non-transposase and non-hypothetical proteins in GenBank or to the genes in the *L. japonicus* gene index [provided by the Institute of Genome Research (TIGR) (http://www.tigr.org)]; (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 more than 50% of the sequence between the TIRs could be aligned (BLASTN, $E= 1e^{-10}$), then the two elements were 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 the 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 (TSDs). TE sequences identified in this way were used as input for RepeatMasker in order 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 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 element 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 2.1.

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 2.1, 2.2). The 6 Mb of complete elements

Table 2.1. Transposable elements in the 32.4 Mb of *L. japonicus* sequences

TE type	Copy number	Coding/Noncoding	DNA amount (Mb)	Percentage (%)
Class I				
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 I	678	527/151	3.36	10.4
Class II				
PIF/Pong	384	29/355	0.47	1.5
CACTA	24	23/1	0.14	0.4
MULEs	1140	75/1065	1.56	4.8
hATs	118	47/71	0.12	0.4
Helitrons	27	21/6	0.14	0.4
Mariner	1	1/0	0.002	0.006
MITEs	370*	0/370	0.20	0.62
Total Class II	2064	196/1868	2.63	8.1
TE fragments**	9810	-	3.98	12.3
Total TEs	12552	723/4038	9.97	30.8

* 213 *Tourist*-like and 157 *Stowaway*-like

** elements lacking one or both ends. See Table 2.2 for distribution of each TE type

MULEs - *Mutator*-like elements

MITEs - Miniature inverted repeat transposable elements

LINEs - Long interspersed nuclear elements

Table 2.2. Distribution of transposable element fragments**

TE type	Copy number	DNA amount (Mb)
Class I		
copia-like	831	0.69
gypsy-like	524	0.66
LINEs	584	0.45
Class II		
PIF/Pong	1132	0.29
CACTA	285	0.29
MULEs†	5503	1.23
hATs	400	0.13
Helitrons	551	0.24
Total TEs	9810	3.98

** elements lacking one or both ends

† some of the MULEs lacking both ends are probably non-TIR MULEs

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 [(miniature inverted repeat transposable elements) (213 *Tourist*-like and 157 S*towaway*-like), Table 2.1]. 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 to 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 2.1).

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 2.1 (A), 2.2 (A), 2.3 (A), 2.4 (A), 2.5 (A), 2.6, 2.7). In each case, the most conserved coding sequences of *L. japonicus* TEs as well as those representing all lineages previously identified 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 2.1 (A), 2.2 (A), 2.3 (A), 2.4 (A), 2.5 (A), 2.6, 2.7), 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 hAT* elements (Figures 2.2 (A), 2.3 (A), 2.4 (A), and 2.6), and all 13 lineages of *copia*-like elements [Figure 2.1 (A)]

.

Figure 2.1. Phylogeny and transposon display of *copia*-like elements. A) The phylogenetic tree was generated using the reverse transcriptase (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).

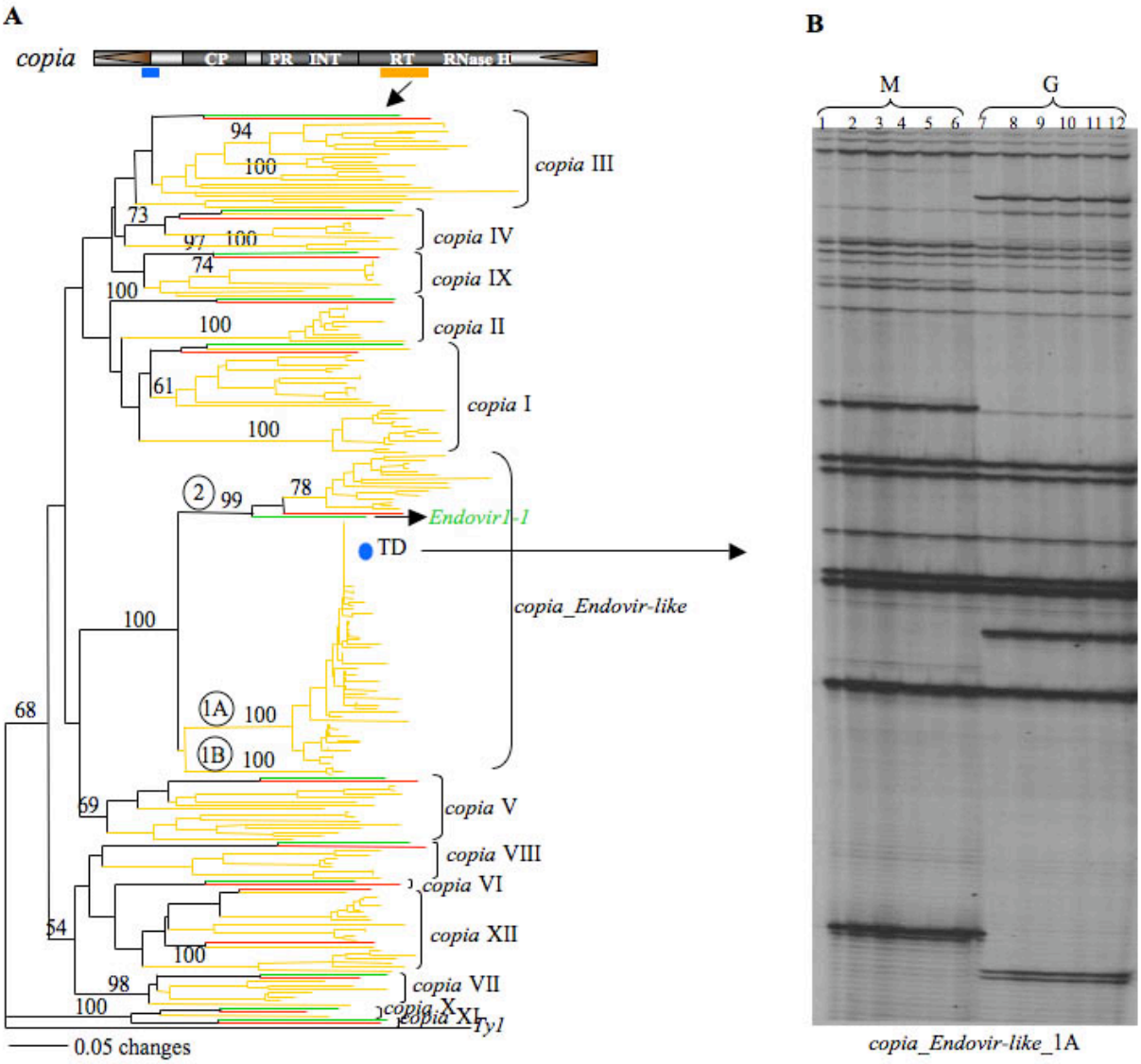

A
copia
CP
PR
INT
RT
RNase H
94
100
copia III
73
97
100
copia IV
74
100
copia IX
100
copia II
61
100
copia I
2
99
78
Endovir1-1
TD
copia_Endovir-like
100
68
1A
100
1B
100
69
copia V
copia VIII
copia VI
54
copia XII
100
98
copia VII
100
copia X
copia XI
Ty1
0.05 changes
B
M
G
1 2 3 4 5 6 7 8 9 10 11 12
copia_Endovir-like_1A

Figure 2.2. 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 lineages (*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 in Figure 2.1B.

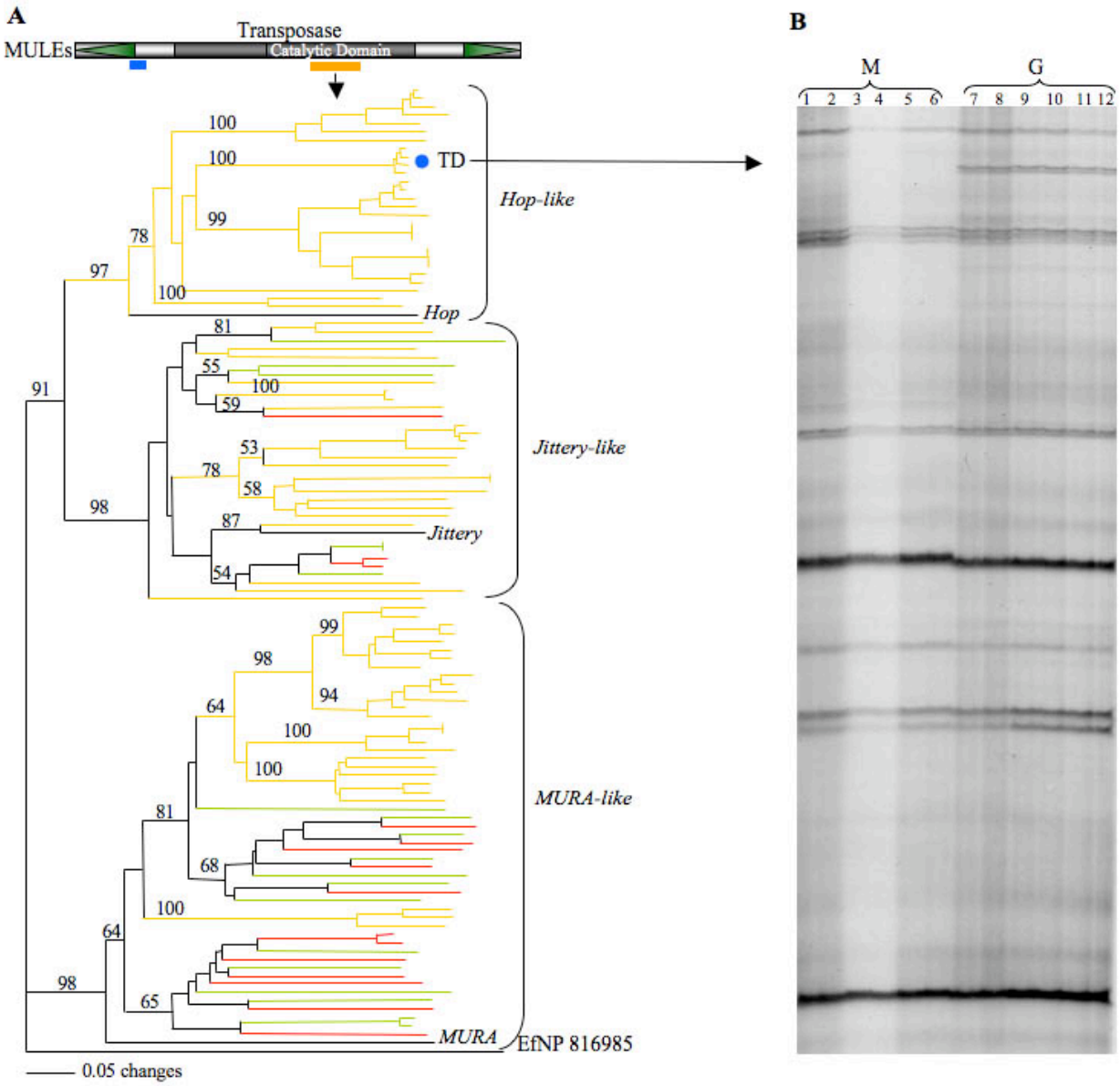

A
Transposase
MULEs
Catalytic Domain
TD
Hop-like
Hop
Jittery-like
Jittery
MURA-like
MURA
EfNP 816985
0.05 changes
B
M
G
1 2 3 4 5 6
7 8 9 10 11 12

Figure 2.3. 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 2.1B for details and DNA analyzed.

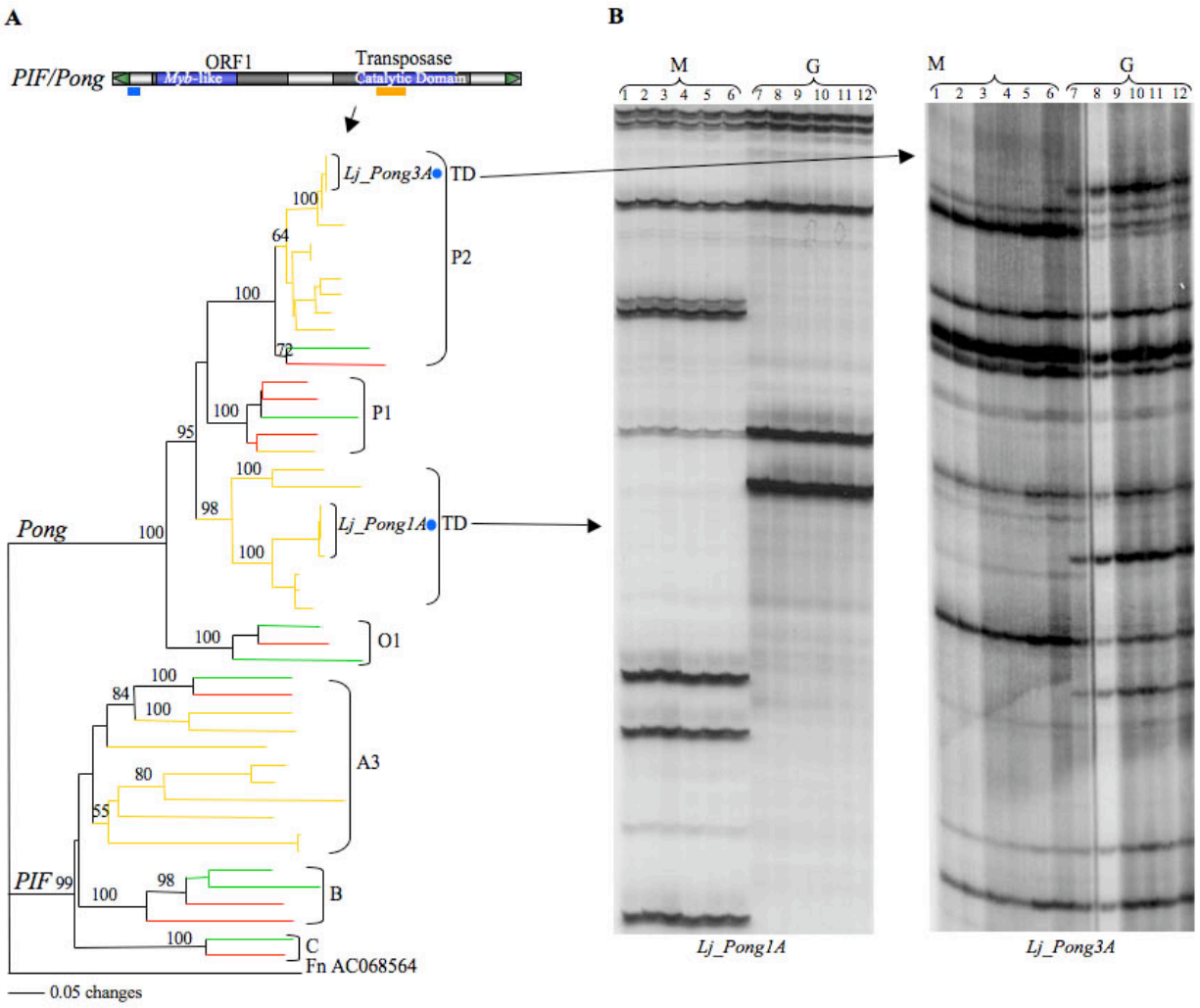

A
B
PIF/Pong
ORF1
Transposase
Myb-like
Catalytic Domain
Lj_Pong3A
TD
P2
P1
Lj_Pong1A
TD
Pong
O1
A3
PIF
B
C
Fn AC068564
0.05 changes
M
G
1 2 3 4 5 6 7 8 9 10 11 12
Lj_Pong1A
Lj_Pong3A

Figure 2.4. Phylogeny and transposon display analysis of CACTA-like elements. A) The phylogenetic tree was generated as described in earlier figures. Lineages identified in previous studies are indicated beside the brackets. The solid blue circle indicates the elements used to generate primers for transposon display analysis. **B)** Transposon display analysis of a CACTA-like lineage. See Figure 2.1B for details and DNA analyzed.

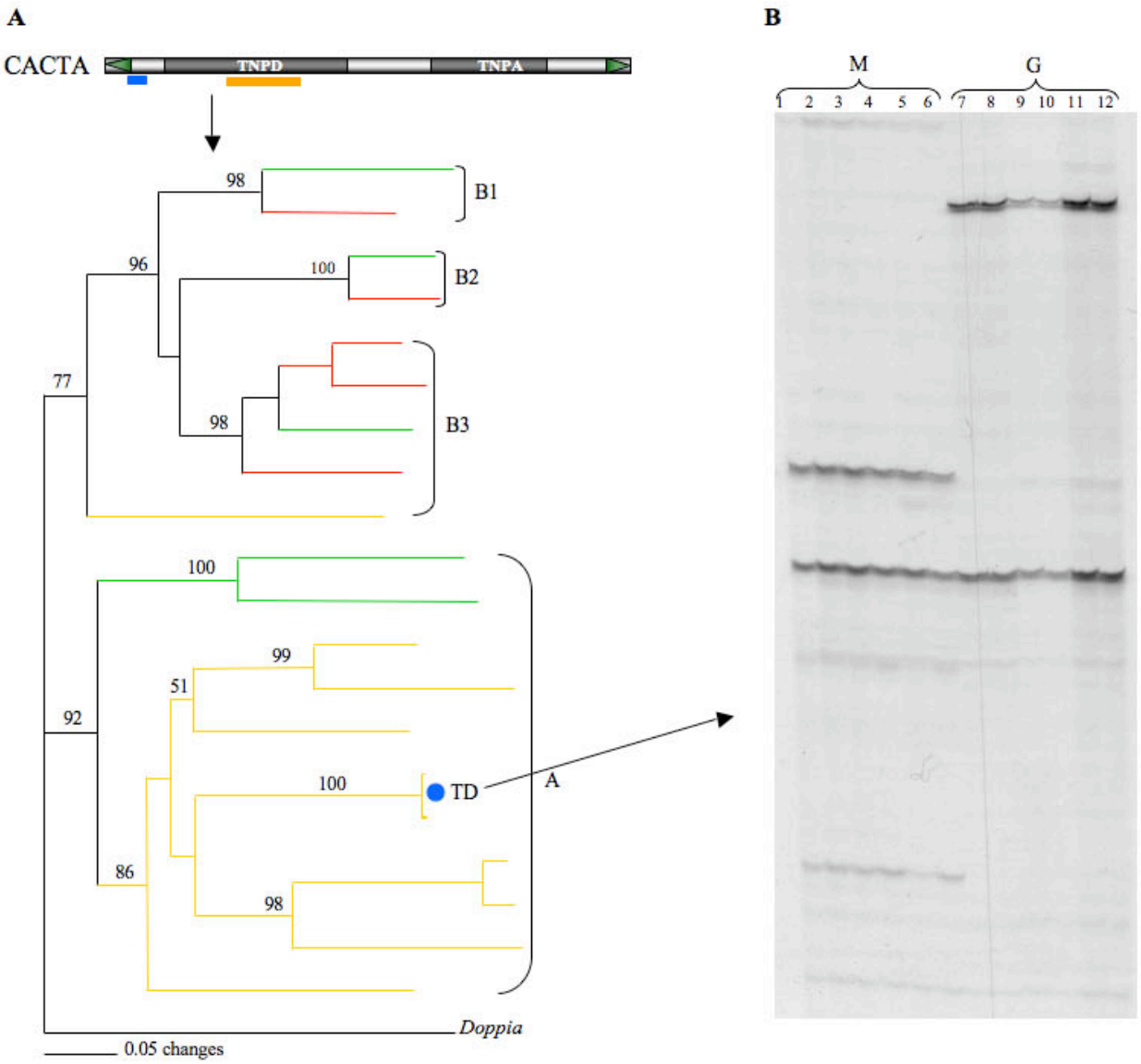

A
CACTA
TNPD
TNPA
98
B1
96
100
B2
77
98
B3
100
99
51
92
100
TD
A
86
98
Doppia
0.05 changes
B
M
G
1 2 3 4 5 6 7 8 9 10 11 12

Figure 2.5. Phylogeny and transposon display analysis of *gypsy*-like elements. A) The phylogenetic tree was generated as described in previous figures. Lineages identified in previous studies and *L. japonicus* specific lineages (*Lj_gypsy_1, Lj_gypsy_2, and Lj_gypsy_3)* are indicated next to the brackets. The solid blue circle indicates the elements used to generate primers for transposon display analysis. **B)** Transposon display analysis of a *copia*-like lineage. See Figure 2.1B for details and DNA analyzed.

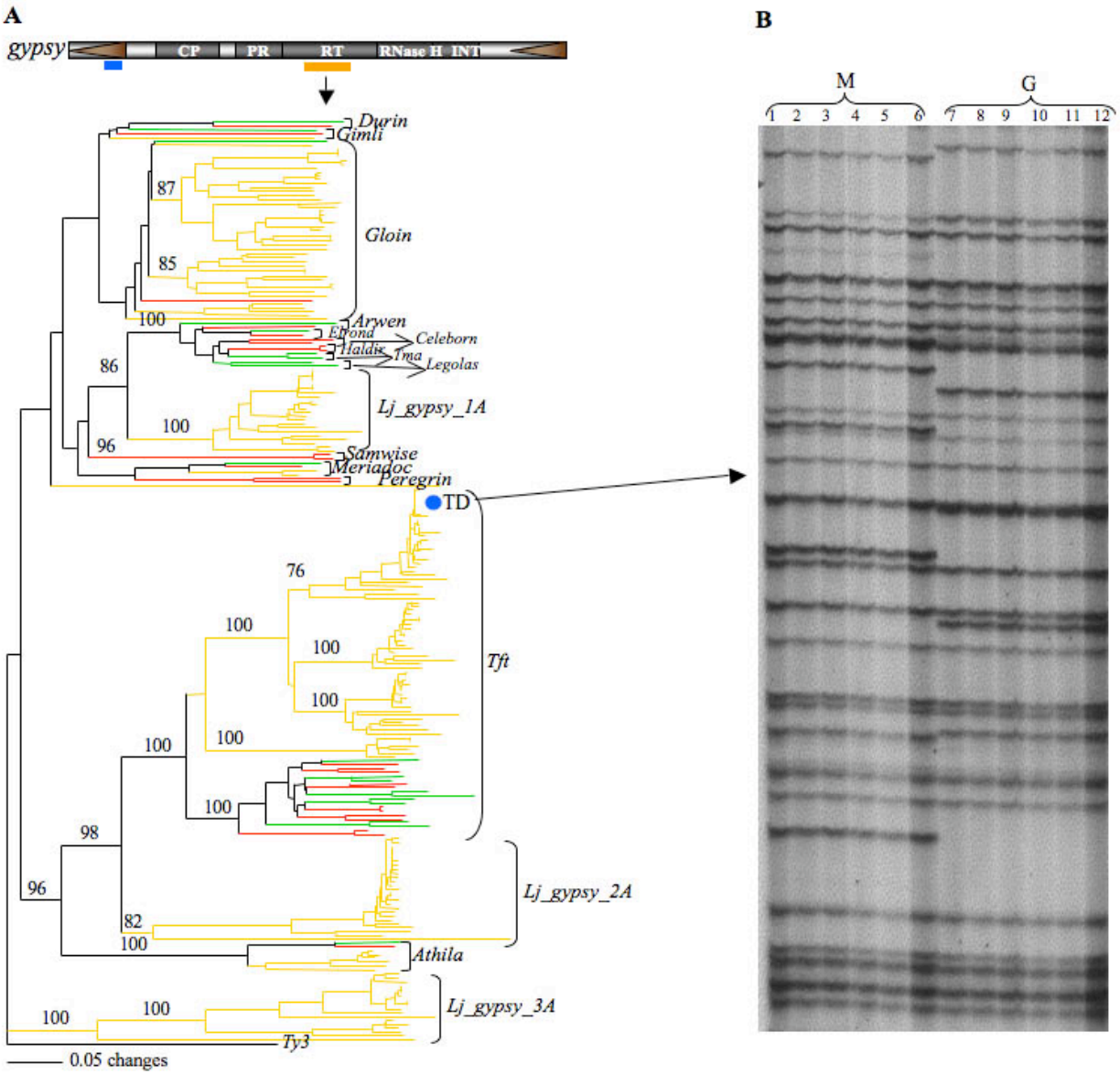

A
gypsy
CP
PR
RT
RNase H
INT
Durin
Gimli
Gloin
Arwen
Elrond
Celeborn
Haldir
Tma
Legolas
Lj_gypsy_1A
Samwise
Meriadoc
Peregrin
TD
Tft
Lj_gypsy_2A
Athila
Lj_gypsy_3A
Ty3
87
85
100
86
100
96
76
100
100
100
100
100
98
96
82
100
100
100
0.05 changes
B
M
G
1 2 3 4 5 6
7 8 9 10 11 12

Figure 2.6. Phylogeny of *hAT* -like elements. The phylogenetic tree was generated as described in previous figures. Lineages identified in previous studies (*Tag1*, *Tag2* and *Tip100*) are indicated beside the brackets.

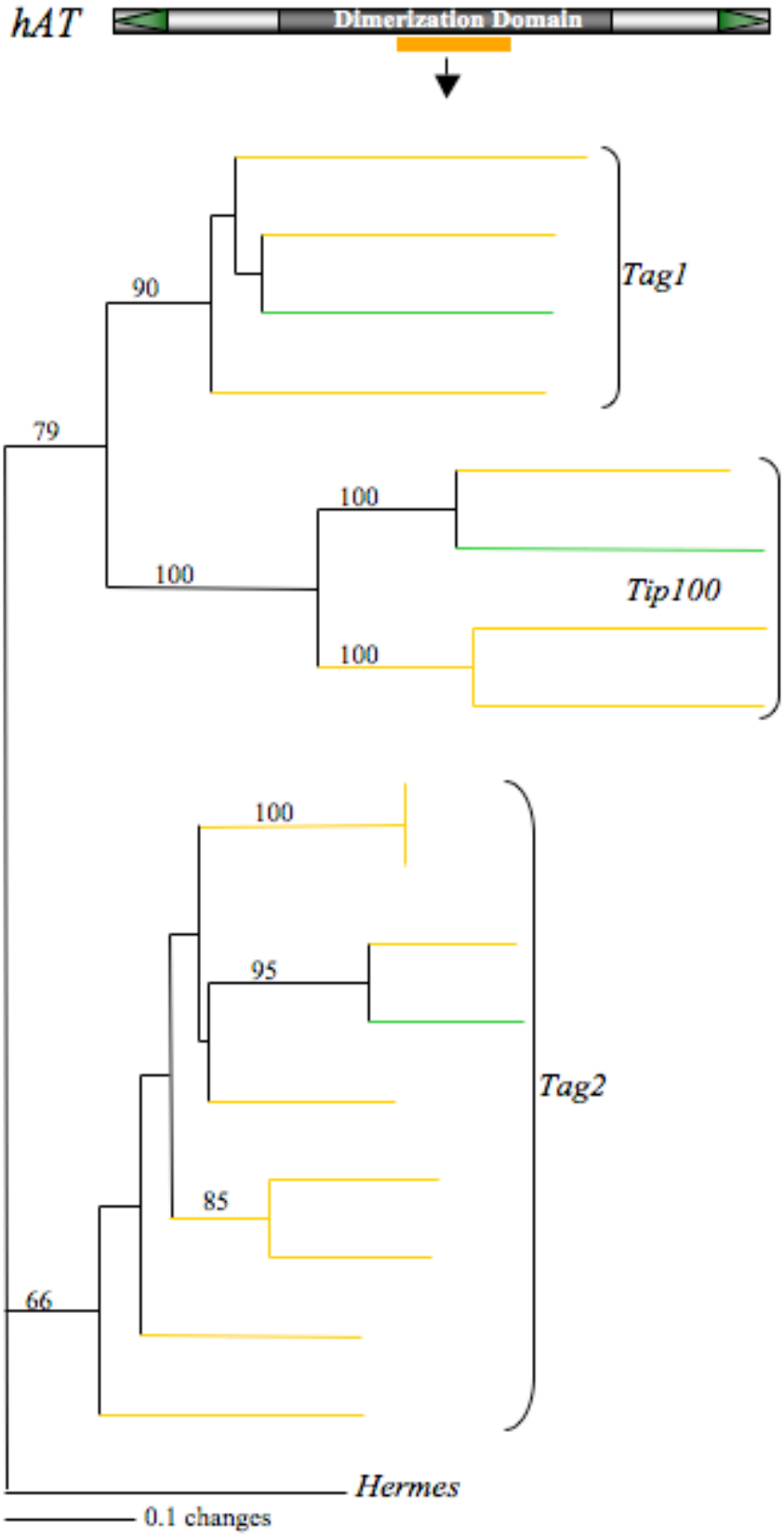
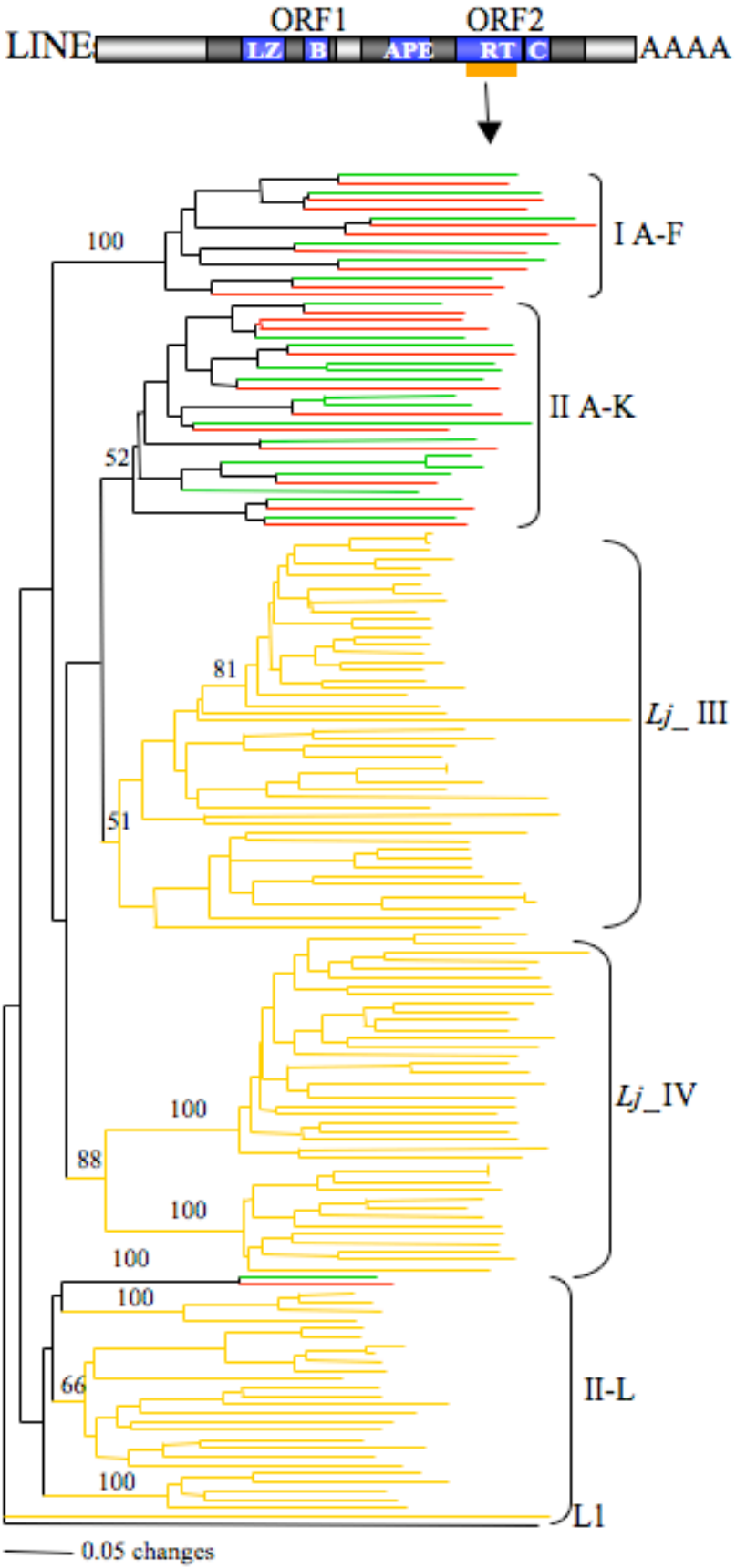

hAT
Dimerization Domain
Tag1
90
79
100
100
Tip100
100
100
95
Tag2
85
66
Hermes
0.1 changes

Figure 2.7. Phylogeny of LINEs. The phylogenetic tree was generated as described in previous figures. Lineages identified in previous studies and *L. japonicus*-specific lineages (*Lj_III and Lj_LIV*) are indicated beside the brackets.


LINE
ORF1
ORF2
LZ
B
APE
RT
C
AAAA
I A-F
II A-K
Lj_ III
Lj_IV
II-L
L1
100
52
81
51
100
88
100
100
100
66
100
0.05 changes

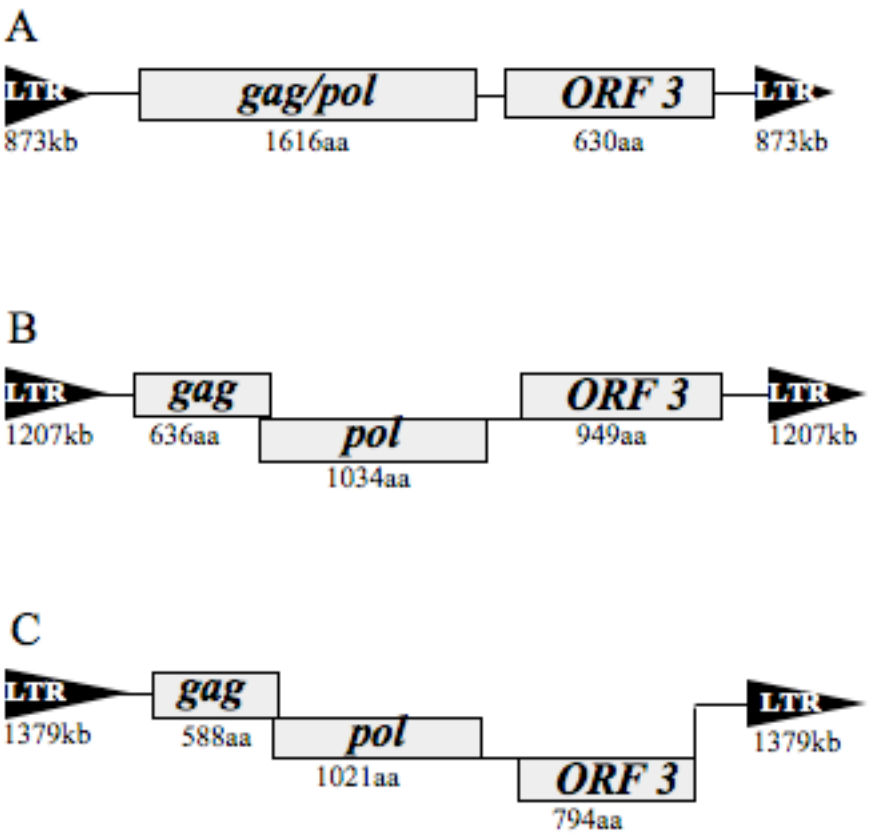
In contrast, several *gypsy*-like and LINE lineages in *L. japonicus* are not in the Brassicaceae. Of 23 Brassicaceae *gypsy*-like lineages, five [*Gimli, Gloin, Tft, Meriadoc* and *Athila,* Figure 2.5 (A)] were in the *L. japonicus* database, and the three remaining *L. japonicus gypsy*-like lineages were not in the Brassicaceae [*Lj_gypsy_1A, Lj_gypsy_2A and Lj_gypsy_3A*, Figure 2.5 (A)]. Similarly, 11 of the 12 Brassicaceae LINE lineages (including the entire clade I) were not in this subset of *L. japonicus* sequence; the majority of the *L. japonicus* LINEs (87 of 114 copies) grouped into 3 lineages (*Lj*_LIII, *Lj*_IV and II-L, Figure 2.7) 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 2.1(A)], and a group of MULEs that are related to the fungal *Hop* element (Figures 2.2, 2.9). 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 2.1). The most numerous lineage, *copia_Endovir*-like, includes ~40% of all *copia*-like elements and can be further divided into three sublineages [Figure 2.1(A)]. 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 2.8. Note that ORF3 from each sublineage varies in length (~630 aa – 950 aa) and that there is greater than 75% amino acid sequence identity within each sublineage. However, inter-sublineage sequence similarity is less than 20%. Our survey identified 82 elements, 64 from sublineage 1A (40 full-length with LTR and TSD defined), 3 from sublineage 1B (all full-length), and 15 from sublineage 2 (nine 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, two are identical, four are 99% identical (~200 mismatches) and four are ~97 - 98% identical (< 500 mismatches). For sublineage 2, the lineage most closely related to *Endovir1-1*, six of the nine 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

Figure 2.8. Structural organization of a representative *copia_Endovir*-like element from the 3 sublineages. (A) *Lj_copia_Endovir*-like_2, **(B)** *Lj_copia_Endovir*-like_1A, and **(C)** *Lj_copia_Endovir*-like_1B. For each sublineage grey boxes represent the different ORFs and black arrowheads represent the LTRs. The length of the average ORF (in aa) and LTR (in bp) is shown. For (B) and (C), the *gag* and *pol* ORFs overlap and a frameshift is presumed to occur to generate both proteins.


A
LTR
gag/pol
ORF 3
LTR
873kb
1616aa
630aa
873kb
B
LTR
gag
ORF 3
LTR
1207kb
636aa
pol
949aa
1207kb
1034aa
C
LTR
gag
LTR
1379kb
588aa
pol
1379kb
1021aa
ORF 3
794aa

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; HAVECKER *et al.* 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 Material 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.

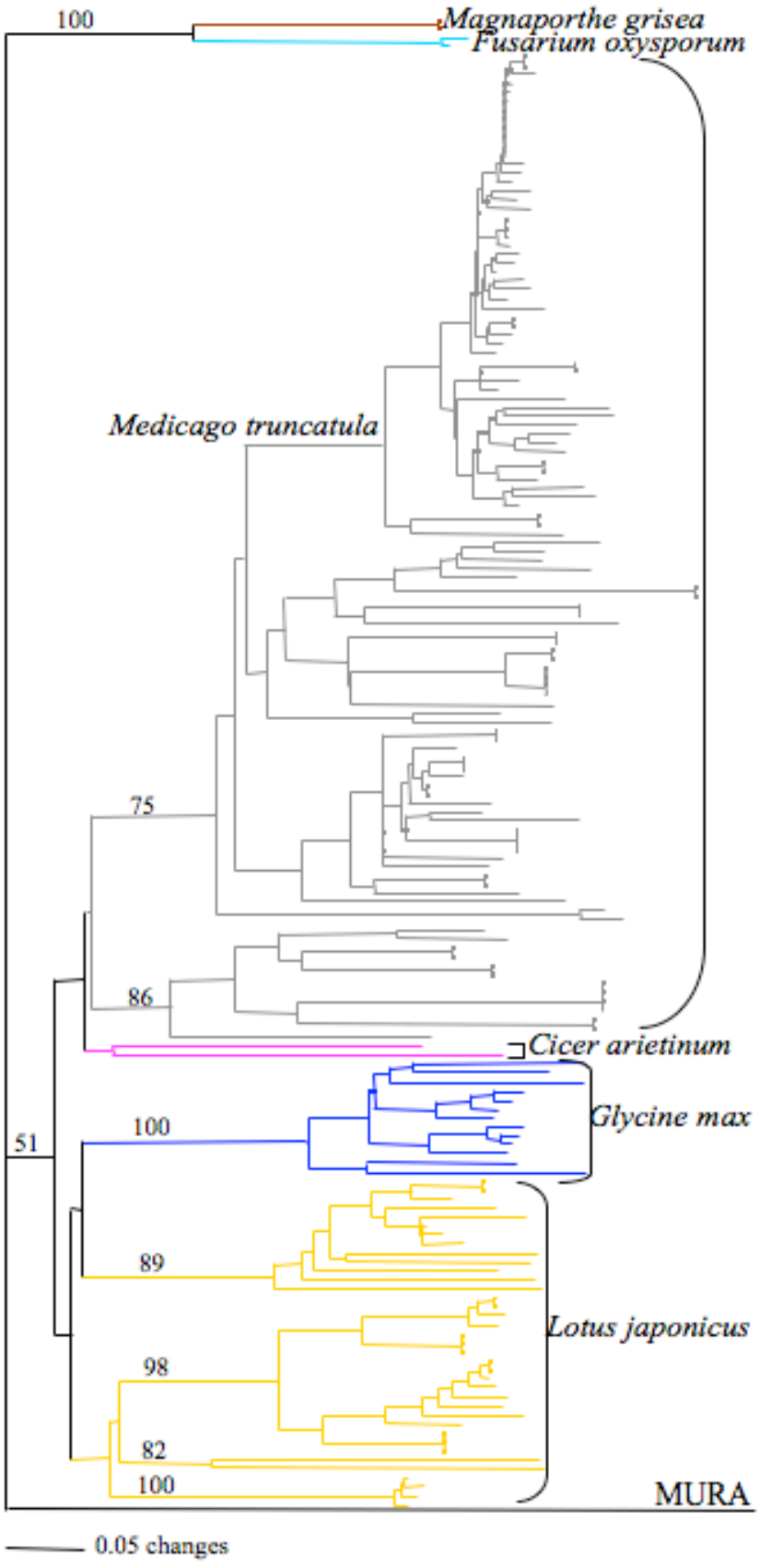
A legume-specific lineage of MULEs: MULEs represent a diverse family of TEs that are widespread in plants and are in some fungal species (YU *et al.* 2000; CHALVET 2003). While previous studies identified two major groups of plant MULEs (*MuDR*-like and *Jittery*-like) (LISCH 2002; XU *et al.* 2004) we were surprised to find a third lineage in *L. japonicus* [Figure 2.2 (A)]. 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 a MULE element from *Magnaporthe grisea*, than to any other plant element. Elements from this lineage (named herein *Hop*-like) are 3 kb to 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 80-100%. Furthermore, at least seven elements share >90% sequence homology over a 2 – 4 kb region. Overall 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 (non redundant), 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 2.9). Surprisingly, not a single *Hop*-like element was detected from any non-legume plant species, including the completely sequenced genomes of *A. thaliana, O. sativa* and *P. trichocarpa.* Furthermore, a phylogenetic tree generated from the legume *Hop*-like MULEs suggested that elements from each legume formed species-specific sublineages (Figure 2.9).

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 over 3000 were identified (JIANG *et al.* 2004a). The identification of only a few Pack-MULEs in the genome of *A. thaliana* (5 in 17 Mb of genomic

Figure 2.9. 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.


100
Magnaporthe grisea
Fusarium oxysporum
Medicago truncatula
75
86
Cicer arietinum
Glycine max
51
100
89
Lotus japonicus
98
82
100
MURA
0.05 changes

sequence, or 36 genome-wide) (YU *et al.* 2000(HOEN *et al.* 2006) suggested that gene capture by MULEs might only occur frequently 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 was 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 an 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 2.1, end of chapter). 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 2.2), indicating that some of the captured genes fragments are transcribed.

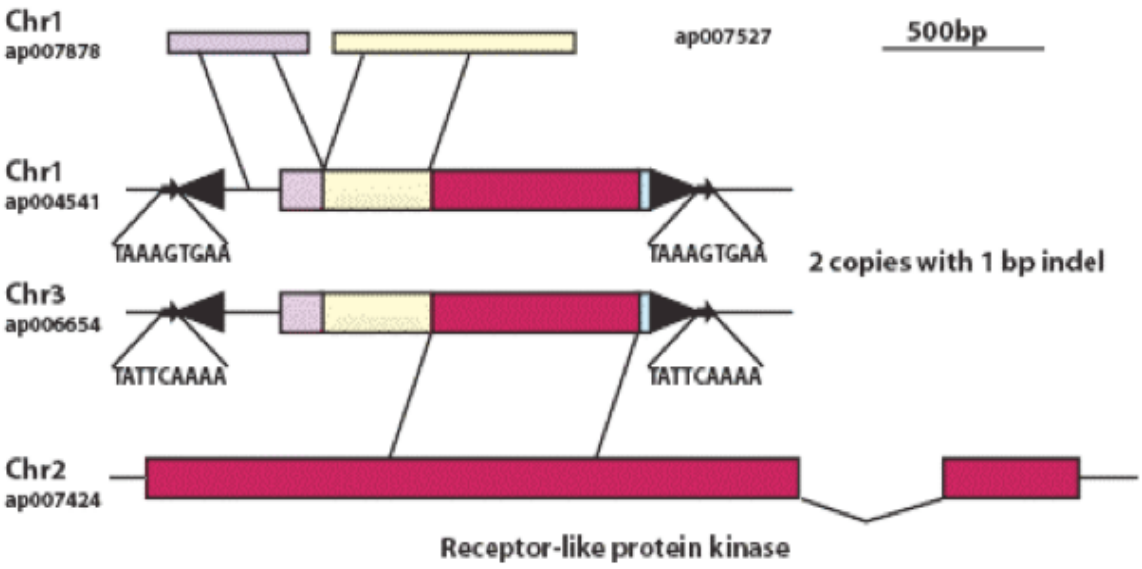
In a prior study, availability of the entire rice genomic sequence facilitated the identification of most of the rice genes whose sequences were captured by Pack-MULEs (JIANG *et al.* 2004a). To investigate the origin of the sequences captured by *L. japonicus* Pack-MULEs, the internal regions of the 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 homolog (BLASTN E < 10^{-10}) that was not flanked by MULE TIRs. Of the 71 Pack-MULEs with identified genomic homologues, 15 (9% of total Pack-MULEs) contain sequences from two or more loci (Figure 2.10, Supplemental Table 2.3). As shown in Figure 2.10, 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 2.10).

Recently amplified elements and TE insertion polymorphism: With the exception of LINEs, all other major TE types have lineages with highly similar members [Figures 2.1 (A), 2.2 (A), 2.3 (A), 2.4 (A), 2.5 (A), and 2.6]. 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 2.3 (A), *Lj_Pong3A* and *Lj_Pong1A*]. In addition, ~ 58% (48/82) of the *copia_Endovir*-like elements have a LTR identity of ~ 98% with only 1-3 mismatches over the entire length of the ~ 0.8 - 1.2 kb LTR, and ~ 20 share 99% identity overall. If these elements have amplified recently, their insertion sites may be polymorphic in different *L. japonicus* ecotypes. Polymorphic sites would not only provide evidence for recent element activity, but they 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 Material and Methods). To this end, PCR primers

were designed from the 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 2.1 (A), 2.2 (A), 2.3 (A), 2.4 (A), and 2.5 (A)], and 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 2.1 (B), 2.2 (B), 2.3 (B), 2.4 (B), and 2.5 (B). 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 for anonymous DNA markers 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 of the monomorphic bands tested represented insertions in both ecotypes.

Figure 2.10. 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 numbers AP007878 and AP007527) is not known and they are depicted as narrow boxes.


Chr1
ap007878
ap007527
500bp
Chr1
ap004541
TAAAGTGAA
TAAAGTGAA
2 copies with 1 bp indel
Chr3
ap006654
TATTCAAAA
TATTCAAAA
Chr2
ap007424
Receptor-like protein kinase

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 about 7% of the genome, we devised computational strategies to identify and characterize coding, non-coding 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 genome wide experimental analysis.

The TE landscape: As mentioned in the Materials and Methods section, the available *L. japonicus* sequence analyzed in this study was from the gene-rich regions of the genome. However, despite this limited dataset, 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 class1 elements of *L. japonicus* were more abundant than noncoding elements (~500 coding vs ~150 noncoding) and all of the noncoding elements

represented 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 2. 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 class 1 vs 8 Mb class 2). This reflects the fact that most class 1 elements are on average larger than class 2 elements.

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 appears to be predominantly near genes (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 2.1 (A)]. 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 are 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 their 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 less than 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 2.1). 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 2,300 Pack-MULEs in the genome (160 x 472 Mb/32.4 Mb).

Of the 160 characterized Pack-MULEs, 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, 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*-like 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 2003) than to *MUDR* or *Jittery*-like MULEs [Figure 2.2 (A)]. 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 non-legume species (Figure 2.9). 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 fungi such as *Fusarium oxysporum* are pathogens of legumes (ALTIER and GROTH 2005), lend 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.

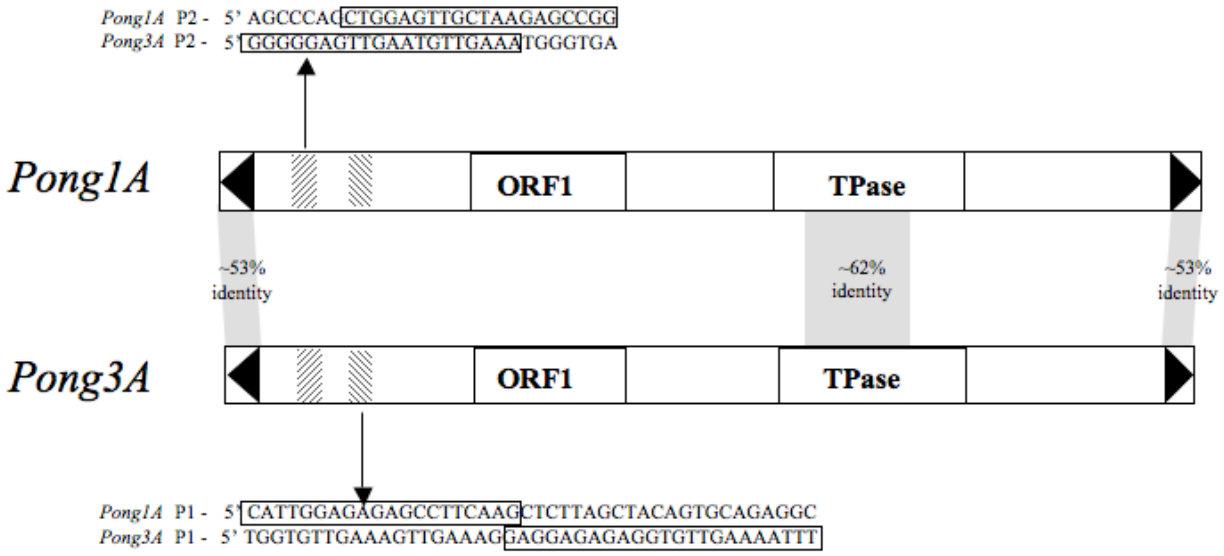
A dearth of MITEs: Less than 400 MITEs belonging to the two MITE superfamilies *Tourist* and *Stowaway* were identified in the available *L. japonicus* sequence (~200 *Tourist*-like, ~150 *Stowaway*-like; Table 2.1). This value is significantly lower, even when extrapolated to the whole genome, than the ~ 90,000 MITEs reported in *O. sativa* (TURCOTTE *et al.* 2001; JURETIC *et al.* 2004; PROJECT 2005). In order 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 2.3, 2.11). 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 based on 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. 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.

Acknowledgements

We thank Sachiko Isobe from the National Agricultural Research Center for 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 a postdoctoral research fellowship from the National Science Foundation 0107590 to E.J.P.

Figure 2.11. 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*. Black 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.


Pong1A P2 - 5' AGCCCAC CTGGAGTTGCTAAGAGCCGG
Pong3A P2 - 5' GGGGGAGTTGAATGTTGAAA TGGGTGA
Pong1A
ORF1
TPase
~53% identity
~62% identity
~53% identity
Pong3A
ORF1
TPase
Pong1A P1 - 5' CATTGGAGAGAGCCTTCAAG CTCTTAGCTACAGTGCAGAGGC
Pong3A P1 - 5' TGGTGTTGAAAGTTGAAAG GAGGAGAGAGGTGTTGAAAATTT

References

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* Newsletter **35:** 59-74.

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, Texas.

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 U S A **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 and Evol **20:** 1362-1375.

CRAIG, N. L., R. CRAIGIE, M. GELLERT and A. M. LAMBOWITZ, 2002 *Mobile DNA II*. American Society for Microbiology Press, Washington, D.C.

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.

INITIATIVE, A. G., 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

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., 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., 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.

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, Q., and P. M. GRESSHOFF, 1997 Classical and molecular genetics of the model legume *Lotus japonicus*. Mol Plant Micro Interac **10:** 59-68.

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 and Genom **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 Sys **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 U S A. **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 U S A **95:** 6897-6902.

LISCH, D., 2002 *Mutator* transposons. Trends in Plant Science **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.

PROJECT, I. R. G. S., 2005 The map-based sequence of the rice genome. Nature **436:** 793-800.

ROST, B., R. CASADIO, P. FARISELLI and C. SANDER, 1995 Prediction of helical transmembrane segments at 95% accuracy. Protein Science **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, 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 Science **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 Genom **265:** 336-344.

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.

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.

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., N. JIANG, C. FESCHOTTE and S. R. WESSLER, 2004 Distribution and evolution of *PIF*- and *Pong*-like transposons and their relationships with *Tourist*-like MITEs. Genetics **166:** 971-986.

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 U S A **101:** 5589-5594.

Supplemental Table 2.1. The 164 Pack-MULE in 32.4 Mb of genomic *L. japonicus* sequence

Accession #	Chr	Position	Left/Right	E	Hits from NCBI nr protein database and TIGR *Lotus japonicus* EST database
AP004250	5	61720-63221	TTCTTTCAA/TTCTTTTAA	-38	>gb|AAG52007.1| putative golgi transport complex protein, *Arabidopsis thaliana*
AP004467	1	33667-35074	GTATATTTA/GTATATTTA	-16	>gb|AAG27371.1| PxORF73 peptide [Plutella xylostella granulovirus]
AP004467	1	100846-101737	AAAAGATAA/AAAAGATAA	-16	>AV414925, *Lotus corniculatus* var. *japonicus* EST
AP004468	4	10199-11190	TAGAATTTT/TAGAATTTT	-13	>BP049467, *Lotus corniculatus var. japonicus* EST
AP004470	4	17980-19741	GTTTCAATAG/GTTTCAATAG	-40	>gb|AAM13275.1| unknown protein [*Arabidopsis thaliana*]
AP004471	3	79860-81302	GAGGAAAAA/GAGGAAAAA	-18	>ref|NP_973808.1| eukaryotic translation initiation factor 3 subunit 3
AP004475	1	65678-67160	TGTTAATAA/TGTTAATAA	-15	>BP036459, *Lotus corniculatus var. japonicus* EST
AP004477	1	30895-32534	TATTTTCTACT/TATTTTCTACT	-18	>BI417542, *Lotus corniculatus var. japonicus* EST
AP004481	6	31432-32835	TTTTGTATA/TTTTGTATA	-12	>gb|AAG27371.1| PxORF73 peptide [*Plutella xylostella granulovirus*]
AP004482	6	7000-8745	AAAATTATA/AAAATTATA	-23	>emb|CAA73364.1| Pge1 protein [*Lotus corniculatus var. japonicus]*
AP004482	6	80480-81915	TTATAATATT/TTATAATATT	-11	>BF646489, *Lotus corniculatus var.*

					japonicus EST
AP004483	1	51130-52604	ATAATAAGAA/ATTATAGAA	-51	>AV777838, *Lotus corniculatus var. japonicus* EST
AP004484	1	9079-10601	TTCTTAGAA/TTCTTAGAA	-30	>TC10628 similar to UP|Q8RX87 (Q8RX87) AT5g20250/F5O24_140 , partial
AP004484	1	73409-75102	TATTTTACT/TATTTTACT	-13	>ref|XP_477694.1| unknown protein [*Oryza sativa* (*japonica* cultivar-group)]
AP004484	1	93218-94965	AAAAATCAT/AAAAATCAT	-94	>TC18024 weakly similar to PIR|F86410|F86410 protein F3M18.12 [imported]
AP004487	5	36895-38452	GTTTGAATAT/GTTTGAATAT	-16	>BI417542, *Lotus corniculatus var. japonicus* EST
AP004491	1	79241-83004	TAAGAAATA/TAAGAAATA	-22	>TC8028 similar to GB|AAA30143.1|552290|TOXNTP nucleoside triphosphate hydrolase
AP004499	1	74840-75979	TAGATTTAA/TAAATTTAA	-11	>gb|AAQ91707.1| unknown [Choristoneura fumiferana
AP004502	1	39977-41107	TAAAAGTTA/TAAAAGTTA	-11	>TC15956 weakly similar to UP|LGT_CITUN (Q9MB73) Limonoid UDP-glucosyltransferase
AP004512	4	3995-10229	AAATACAAA/AAATACAAA	-23	>ref|NP_671779.1| expressed protein [*Arabidopsis thaliana*]
AP004513	5	16486-20506	TATTGCAAT/TATTGCAAT	-19	>ref|NP_671779.1| expressed protein [*Arabidopsis thaliana*]
AP004526	6	50696-52138	AAAGATAAA/AAAGATAAA	-18	>ref|NP_973808.1| eukaryotic translation initiation factor 3 subunit 3, [*Arabidopsis thaliana*]
AP004528	6	5727-7289	TAATTTATT/TAATTTATT	-17	>gb|AAM64538.1| cinnamoyl-CoA

					reductase-like protein [*Arabidopsis thaliana*]
AP004528	6	73909-75606	TTATTTTTG/TTATTTTTG	-17	>gb|AAM63777.1| unknown [*Arabidopsis thaliana*]
AP004532	1	80449-82295	ATCTAATAA/ATCTAATAA	-56	>BP030309, *Lotus corniculatus var. japonicus* EST
AP004533	3	33672-35158	GAGACGTTG/GAGACGTTA	-25	>TC19021 weakly similar to UP|Q9ZQH0 (Q9ZQH0) At2g27490 protein, partial (36%)
AP004541	1	25986-27582	TAAAGTGAA/TAAAGTGAA	-90	>gb|AAF91324.1| receptor-like protein kinase 3 [*Glycine max*]
AP004543	1	33260-35005	AAGATTTAA/AAGATTTAA	-11	>gb|AAQ91707.1| unknown [*Choristoneura fumiferana*
AP004576	2	119139-120869	GAATACGAA/GAATATGAA	-30	>TC9803
AP004625	5	29543-31032	CTTCGTATA/CTTCGTATA	-12	>ref|ZP_00315413.1| COG0745: Response regulators consisting of a CheY-like
AP004625	5	100312-101726	CTTATTTAA/CTTATTTAA	-17	>BP067127, *Lotus corniculatus var. japonicus* EST
AP004896	5	40252-41896	ATTCTATTAC/ATTCTATTAC	-29	>TC13558 similar to GB|AAD33770.1|4928919|AF138744 zinc finger protein 2
AP004906	1	23886-32312	AACTAGACCA/AACTAGACCA	-19	>ref|NP_564042.1| expressed protein [*Arabidopsis thaliana*]
AP004907	2	626-2451	ATATTTAAA/ATATTTAAA	-16	>gb|AAC59091.1| unknown [*Orgyia pseudotsugata* multicapsid nucleopolyhedrovirus]
AP004907	2	9264-9967	TTAATTATA/TTAATTATA	-11	>gb|AAF55377.2| CG5225-PA [*Drosophila melanogaster*]
AP004912	3	24854-26496	TATATATAT/TATATATAT	-12	>emb|CAB67645.1| receptor lectin kinase-like protein [*Arabidopsis*

					thaliana]
AP004915	4	29370-30673	GATGAAATA/GATGAAATA	-16	>TC19331
AP004916	2	72168-73372	TAAAAAATA/TAAAAAATA	-27	>BP046136, *Lotus corniculatus* var. *japonicus* EST
AP004926	4	65464-66612	AAATTATAT/AAATTATAT	-18	>BP040315, *Lotus corniculatus var. japonicus* EST
AP004926	4	83684-85088	TTAAATATTT/TTAAATATTT	-19	>BP032887, *Lotus corniculatus var. japonicus* EST
AP004927	2	42805-51029	AGCATCAAC/AGCATCAAC	-25	>gb|AAM62582.1| unknown [*Arabidopsis thaliana*]
AP004932	1	27296-31256	TAAAATCAA/TAAAATCAA	-22	>gb|AAD23012.1| expressed protein [*Arabidopsis thaliana*]
AP004939	1	14876-16331	TAAACCAAAT/TAAACCAAAT	-15	>TC17947 similar to UP|Q7YEU3 (Q7YEU3) ATPase subunit 6, partial (6%)
AP004939	1	91536-92953	TACATATA/TACAATATA	-81	>AV775660, *Lotus corniculatus var. japonicus* EST
AP004946	N/A	88225-90071	GGGGGGGGG/GGGGGGGGG	-25	>CB828260
AP004948	N/A	1351-2831	TATTAATTA/TATTAATTA	-12	>AV778604, *Lotus corniculatus var. japonicus* EST
AP004949	1	53992-55714	ATATAATTA/ATATAATTA	-55	>TC18816
AP004956	1	67851-69464	ATATTGTGG/AATATTGTG	-11	>AV424081, *Lotus corniculatus var. japonicus* EST
AP004962	1	81323-82887	ATAATTTAA/ATAATTTAA	-10	>ref|NP_200709.2| protein kinase-related [*Arabidopsis thaliana*]
AP004963	3	10273-16115	TATTTTATT/TATTTTATT	-129	>TC12588
AP004963	3	29363-31061	AAGTAAAAA/AA	-19	>AV422772, *Lotus*

			GTAAAAA		*corniculatus var. japonicus* EST
AP004965	1	15880-17696	TTTAATGTT/TTTAATGTT	-38	>TC18833
AP004968	1	27259-28975	TAACTTA/TAACATA	-15	>gb|AAG52091.1| putative AP2 domain transcriptional regulator
AP004968	1	130271-131828	TTTTAATTT/TTTTAATTT	-27	>emb|CAB62280.1| hydroxyproline-rich glycoprotein DZ-HRGP [Volvox carteri f.
AP004970	5	8649-10177	TAAAAGTAA/TAAAAGTAA	-59	>BP047548, *Lotus corniculatus var. japonicus* EST
AP004970	5	87213-89571	CATTGAGG/CATTGAGG	-92	>TC17255
AP004971	5	14259-15709	TCAAATTAA/TCAAATTGA	-13	>gb|AAK77898.1| root nodule extensin [*Pisum sativum*]
AP004971	5	54074-55749	TGTATAATA/TATAAAATA	-11	>gb|AAM03451.1| putative transporter NIC1 [*Arabidopsis thaliana*]
AP004978	4	78226-79720	TTTTCTCAA/TTTTCTCAA	-10	>BP079561, *Lotus corniculatus var. japonicus* EST
AP004981	3	36953-38647	TAGTTAAAA/TAGTTAAAA	-13	>ref|XP_477694.1| unknown protein [*Oryza sativa (japonica* cultivar-group)]
AP006074	3	31740-33026	ATTAAATAA/ATTAAATAA	-26	>gb|AAO50669.1| unknown protein [*Arabidopsis thaliana*]
AP006079	6	1500-3115	ATAAATAA/ATAAATAA	-28	>gb|AAQ90287.1| beta-1,3-glucanase, acidic

					[Coffea arabica]
AP006084	3	77525-78709	GATTTAAA/GATTTAAA	-33	>TC15736 similar to UP|AAR24652 (AAR24652) At5g64220, partial (6%)
AP006089	2	58758-60347	TTTTCTATT/TTTTCTATT	-77	>AV407620, *Lotus corniculatus var. japonicus* EST
AP006094	4	83242-85104	TAATTTGAT/TAATTTGAT	-24	>gb|AAC59091.1| unknown [*Orgyia pseudotsugata* multicapsid nucleopolyhedrovirus]
AP006094	4	118745-120281	AATATTTTT/AATATTTTT	-28	>TC17489 similar to UP|Q8PX92 (Q8PX92) Chemotaxis protein CheW, partial (11%)
AP006095	4	5236-6438	TAGAATGTA/TAGAATGTA	-17	>TC17489 similar to UP|Q8PX92 (Q8PX92) Chemotaxis protein CheW, partial (11%)
AP006097	1	19249-20755	TAAAATATA/TAAAATATA	-14	>TC15653 similar to UP|Q941L5 (Q941L5) bZIP transcription factor BZI-4, partial
AP006098	4	80624-81755	TTATTTTGA/TTATTTTGA	-43	>BP063249, *Lotus corniculatus var. japonicus* EST
AP006098	4	84835-86430	TTTCAAGAA/TTTCAAGAA	-20	>AV776881, *Lotus corniculatus var. japonicus* EST
AP006098	4	92182-93701	GAAAAATAA/GAAAAATAA	-86	>BP054828, *Lotus corniculatus var. japonicus* EST
AP006101	1	102633-107237	CACTCAAAC/CACTCAAAC	-24	>ref|NP_175246.1| calcineurin-like phosphoesterase family protein [*Arabidopsis*
AP006101	1	141088-143259	AAAAACTAA/AAAAATTAA	-92	>AV779895, *Lotus corniculatus var. japonicus* EST
AP006102	4	58327-60715	CTAATAAAC/CTAATAAAC	-12	>CB828260, *Lotus corniculatus var. japonicus* EST
AP006103	1	29747-30725	TATTACTTT/TATTACTTT	-15	>AV422500, *Lotus corniculatus var.*

AP006103	1	47202-49118	AAAATTGTG/AAAATTGTG	-13	*japonicus* EST >BP053577, *Lotus corniculatus var. japonicus* EST
AP006105	3	79900-84315	TAATTTTAA/TAATTTTAA	-48	>AV777083, *Lotus corniculatus var. japonicus* EST
AP006107	1	86869-88082	TTAAATATC/TTAAATATC	-17	>AV428826, *Lotus corniculatus var. japonicus* EST
AP006107	1	98037-99556	AATATATAC/AATATATAC	-30	>TC10628 similar to UP|Q8RX87 (Q8RX87) AT5g20250/F5O24_140 , partial (35%)
AP006108	4	81664-82785	TAAAGTTTT/TAAAGTTTT	-30	>TC8028 similar to GB|AAA30143.1|552290|TOXNTP nucleoside triphosphate hydrolase (Toxoplasma gondii)
AP006111	3	18607-19834	TTGTTTAAA/TTGTTTAAA	-14	>BP063249, Lotus corniculatus var. japonicus EST
AP006113	4	17905-19516	TTTAGTCTT/TTTAGTCTT	-20	>ref|NP_200709.2| protein kinase-related [*Arabidopsis thaliana*]
AP006113	4	99996-101854	TATTGAGAA/TATTGAGAA	-78	>BP073453, *Lotus corniculatus var. japonicus* EST
AP006114	3	65189-67678	ATATTAATC/ATATTAATC	-20	>TC10922 similar to PIR|T39586|T39586 rna binding protein - fission yeast
AP006118	2	45626-46567	TTTAGAGAA/TTTAGAGAA	-27	>ref|NP_196993.2| NHL repeat-containing protein [*Arabidopsis thaliana*], also called F-box family
AP006121	1	27911-29542	TTTCCCTTT/TTTCCCTTT	-11	>AV411618, *Lotus corniculatus var. japonicus* EST
AP006122	1	32607-33980	TTGATCATA/TTGATCATA	-19	>TC18024 weakly similar to PIR|F86410|F86410 protein F3M18.12 [imported], *Arabidopsis thaliana*

AP006129	2	33723-35306	TCCAGAAAC/TCCAGAAAC	-16	>emb|CAB77840.1| putative glucan synthase component [*Arabidopsis thaliana*]
AP006131	4	85908-87316	TATATTTTC/TATATTTTC	-17	>BP039310, *Lotus corniculatus var. japonicus* EST
AP006136	2	46322-50495	TATTTTAAA/TATTTTAAA	-93	>BP033552, *Lotus corniculatus var. japonicus* EST
AP006136	2	58174-59828	AAAATAAAT/AAAATAAAT	-13	>ref|NP_194343.1| expressed protein [*Arabidopsis thaliana*]
AP006138	1	101924-103209	TAGTTTCTG/TAGTTTCTG	-10	>TC14869
AP006139	3	8424-9797	AATAATTTA/AATAATTTA	-19	>TC18024 weakly similar to PIR|F86410|F86410 protein F3M18.12 [imported], *Arabidopsis thaliana*
AP006141	4	18532-20581	TTGTTCTTT/TTGTTCTTT	-20	>TC16167 similar to UP|Q9FGH3 (Q9FGH3) Dihydroflavonol 4-reductase-like
AP006141	4	94748-96479	AAATTAATA/AAATTAATA	-17	>gb|AAN13180.1| unknown protein [*Arabidopsis thaliana*]
AP006143	2	84600-86184	TATATTAGT/TATATTAGT	-64	>TC12507
AP006144	3	70733-71534	GTTGATCAC/GTTGATCAC	-15	>TC15736 similar to UP|AAR24652 (AAR24652) At5g64220, partial (6%)
AP006354	1	11065-13446	ATTGTTGTTT/ATTGTTGTTT	-93	>TC17255
AP006355	6	53596-54745	TTTAGCTTA/TTTAGCTTA	-20	>BP040315, *Lotus corniculatus var. japonicus* EST
AP006356	6	54043-56128	TAATCTAAA/TAATCTAAA	-41	>TC15044 similar to UP|Q8VYI3 (Q8VYI3) At1g76150/T23E18_38, partial (53%)
AP006361	6	56539-59215	AAAATTAAA/AAAATTAAA	-126	>TC14358
AP006364	4	36707-38193	TTATATTAA/TTA	-11	>gb|AAM64349.1| 26S

			TATTAA		proteasome non-ATPase regulatory subunit [Arabidopsis
AP006375	4	45167-46785	TTATAATTA/TTATAATTA	-15	>AV766044, *Lotus corniculatus var. japonicus* EST
AP006380	4	50032-54754	AATATGCAA/AATATGCAA	-10	>gb|AAU95079.1| resistance-like protein HAAS-2 [Glycine max]
AP006383	6	69804-71355	AATCATTTAA/AATCATTTAA	-12	>emb|CAB77840.1| putative glucan synthase component [*Arabidopsis thaliana]*
AP006386	1	14519-16601	TATTCCGCTG/TATTCCGCTG	-76	>TC18791
AP006390	4	8197-9633	AATTTATTT/AATTTATTT	-18	>gb|AAM61436.1| unknown [Arabidopsis thaliana]
AP006393	1	16431-17915	TTAAAATTA/TTAAAATTA	-17	>BP036459, Lotus corniculatus var. japonicus EST
AP006395	3	18764-21470	TAATAAATA/CAATAAATA	-22	>TC8028 similar to GB|AAA30143.1|552290|TOXNTP nucleoside triphosphate hydrolase {Toxoplasma gondii;}
AP006396	3	18322-19108	ATATTAAAA/ATATTAAAA	-13	>TC79185 similar to GP|10178187|dbj|BAB11661. gene_id:MQN23.20~unknown protein
AP006403	1	39222-47685	TTTGTGAGC/TTTGTGAGC	-24	>gb|AAM62582.1| unknown [*Arabidopsis thaliana*]
AP006405	4	1919-3561	AAACGATTA/AAACGATTA	-29	>TC12252 weakly similar to GB|AAB66486.1|2318131|AF014824 histone deacetylase
AP006407	4	1486-3233	TTTTTTTTT/TTTTTTTTT	-16	>ref|NP_193007.1| armadillo/beta-catenin repeat family protein [Arabidopsis thaliana]
AP006408	4	12096-13435	TAAATTATAG/TAAATTATAG	-14	>AV769205, *Lotus corniculatus var. japonicus* EST

AP006412	6	14836-15727	TTCTTATTT/TTCTTATTT	-16	>AV414925, *Lotus corniculatus var. japonicus* EST
AP006417	4	46862-48660	TTAAATAAA/TTAAATAAA	-44	>gb|AAG52007.1| putative golgi transport complex protein, *Arabidopsis thaliana*
AP006417	4	97428-98939	CATACTTTTA/CATACTTTTA	-19	>gb|AAO29946.1| expressed protein [*Arabidopsis thaliana*]
AP006417	4	119244-120642	ATAATTAAA/ATAATTAAA	-46	>TC10127 weakly similar to GB|BAC16475.1|23237901|AP004671 10 kDa chaperonin, *Oryza sativa*
AP006419	4	34677-36050	TTTGATTATA/ATTGATTAT	-19	>TC18024 weakly similar to PIR|F86410|F86410 protein F3M18.12 [imported], *Arabidopsis thaliana*
AP006419	4	54310-55703	TTTTATAAA/TTTTATAAA	-12	>ref|NP_172035.1| expressed protein [*Arabidopsis thaliana*]
AP006422	2	33455-34795	TTTTTATTA/TTTTTATTA	-116	>TC11815 weakly similar to UP|Q8MLU0 (Q8MLU0) CG13503-PA, partial (3%)
AP006423	5	3515-5192	AATAATCTT/AATAATCTT	-12	>TC77696 similar to GP|2673912|gb|AAB88646.1| expressed protein {Arabidopsis
AP006425	1	4713-5987	TAAAAAAAT/TAAAAAAAT	-35	>BP040912, *Lotus corniculatus var. japonicus* EST
AP006425	1	26742-28464	ATCCTTTTT/ATCCTTTTT	-11	>ref|NP_033370.1| pleckstrin homology-like domain, family A, member 1
AP006425	1	37243-38362	TATAATAAT/TATAATAAT	-31	>TC8028 similar to GB|AAA30143.1|552290|TOXNTP nucleoside triphosphate, hydrolase {Toxoplasma gondii;}
AP006427	1	58810-64985	AGATATTTGC/A	-14	>gb|AAM62582.1|

			GATATTTGC		unknown [*Arabidopsis thaliana*]
AP006428	5	28096-33913	AAATTAGAT/AAATTAGAT	-13	>TC257202 weakly similar to UP|RN12_HUMAN (Q9NVW2) RING finger protein 12
AP006535	2	67186-68683	ATCAATTAA/ATCATTTGA	-10	>ref|NP_181587.2| expressed protein [*Arabidopsis thaliana*]
AP006632	2	21052-22599	TTTATAATA/TTTATAATA	-11	>AU089509, *Lotus corniculatus var. japonicus* EST
AP006639	3	6716-14691	AAAAGAAAA/AAAAGGAAA	-141	>gb|AAM15435.1| unknown protein [*Arabidopsis thaliana*], >dbj|BAB02793.1| helicase-like protein [*Arabidopsis thaliana*]
AP006640	3	2347-3869	TAAAAACAA/TAAAAACAA	-15	>AV766044, *Lotus corniculatus var. japonicus* EST
AP006650	6	10753-12209	TACCATTAA/TACCATTAA	-12	>TC19962
AP006654	3	57167-58651	TATTTACTA/TATTTACTA	-17	>BP036459, *Lotus corniculatus var. japonicus* EST
AP006654	3	92780-94377	TTTTGAATA/TTTTGAATA	-90	>gb|AAF91324.1| receptor-like protein kinase 3 [*Glycine max*]
AP006656	1	4285-8316	TTGAATTTA/TTGAATTTA	-13	>TC15305 similar to GB|CAB42557.1|4775270|ATH131214 SF2/ASF-like splicing, modulator Srp30 {*Arabidopsis thaliana*;}
AP006659	5	73094-75153	TGTTTTTTT/TGTTTTTTT	-67	>TC18791
AP006663	2	1185-2633	TATATTTATT/ATATTTATT	-12	>TC7919 GB|AAP47421.1|32331787|AY164847 RTNLB47w {*Lotus corniculatus* var.
AP006663	2	32507-34156	GTTTTTTTT/ATTTTTTTT	-13	>AW686510 similar to PIR|A86323|A86 protein F14D16.3 [imported] -

					Arabidopsis
AP006666	3	19914-21429	TTATTTTTA/TTATTTTTA	-27	>TC15513 homologue to UP|O61085 (O61085) Coronin binding protein, partial (5%)
AP006666	3	54086-56147	GAAATAAAT/GAAATAAAT	-88	>BP030309, *Lotus corniculatus var. japonicus* EST
AP006667	5	19134-23700	TAAATAATT/TAAATAATA	-57	>AV779405, *Lotus corniculatus var. japonicus* EST
AP006667	5	67401-68880	TAAATTAAA/TAAATTAAA	-16	>ref|NP_031376.2| pleckstrin homology-like domain, family A, member 1 [*Homo sapiens*]
AP006669	2	14249-15456	TAATTTAAA/TAATTTAAA	-37	>TC11832
AP006673	3	22145-23444	GACAAAAAG/GACAAAAAG	-13	>TC18018
AP006675	2	5917-6719	TGTTCGGTT/TGTTCAGTT	-14	>TC15736 similar to UP|AAR24652 (AAR24652) At5g64220, partial (6%)
AP006675	2	10358-17469	ATAGCAGAG/ATAGCAGAG	-14	>gb|AAM62582.1| unknown [*Arabidopsis thaliana*]
AP006676	1	21576-23370	CAATTTTA/CAATTTTG	-29	>CB828260, *Lotus corniculatus var. japonicus* EST
AP006678	1	6472-8034	AAAAAAAAA/AAGAAAAAA	-71	>AV779189, *Lotus corniculatus var. japonicus* EST
AP006681	1	59953-61313	TGAAATTAT/TGAAATTAT	-12	>gb|AAS13373.1| wee1 [*Glycine max*]
AP006685	1	76101-77517	TAATATAAT/TAATATAAT	-18	>ref|NP_973808.1| eukaryotic translation initiation factor 3 subunit 3 [*Arabidopsis thaliana*]
AP006686	2	27555-28636	TATTATGAA/TATTATGAA	-32	>TC15711 weakly similar to UP|MID2_YEAST (P36027) Mating process protein MID2
AP006688	2	28274-29346	TAAATATAA/TA	-36	>TC17489 similar to

			AATATAA		UP|Q8PX92 (Q8PX92) Chemotaxis protein CheW, partial (11%)
AP006691	3	30938-38121	GAGTAAAAT/GAGTAAAAT	-69	>TC16706
AP006691	3	46605-48021	TTAATCATA/TTAATCATA	-15	>gb|AAP04031.1| putative aldolase [*Arabidopsis thaliana*]
AP006691	3	67778-69504	TAGAATTTA/TAGAATTTA	-10	>gb|AAM44925.1| putative protein kinase [*Arabidopsis thaliana*]
AP006692	2	76808-78392	TCCAGAAAC/TCCAGAAAC	-16	>emb|CAB77840.1| putative glucan synthase component [*Arabidopsis thaliana*]
AP006696	3	4828-6174	TAAAAATA/CAAAAATA	-61	>BE122502, *Lotus corniculatus var. japonicus* EST
AP006696	3	10970-12581	TATATATAA/TATATATAA	-106	>BE044806, *Lotus corniculatus var. japonicus* EST
AP006698	1	98186-99754	TAAAAATAA/TAAAAATAA	-83	>TC12317
AP006700	2	63111-64684	TTTTTTTT/TTTTTTTT	-33	>TC17489 similar to UP|Q8PX92 (Q8PX92) Chemotaxis protein CheW, partial (11%)
AP006703	2	7585-8859	ATATATATA/ATATATATA	-13	>BP049467, *Lotus corniculatus var. japonicus* EST
AP006707	3	11148-13628	TTAATATAA/TTAATATAA	-56	>AV779895, *Lotus corniculatus var. japonicus* EST
AP006712	2	42330-43473	TATTTCTTT/TATTTCTTT	-19	>BP040315, *Lotus corniculatus var. japonicus* EST
AP006712	2	87720-89413	TTTTTTTTG/TTTTTTTTT	-11	>gb|AAM91821.1| choriogenin Hminor [*Oryzias latipes*]

Supplemental Table 2.2. Pack-MULEs with perfect EST matches

MULE Accession #.	Chromosome (Chr)	Element Position	Matching EST	Sequence Identity (%)
AP004484	1	93218-94965	AV423041	100
AP004512	4	3995-10229	AV769506	99.7
AP004970	5	87213-89571	BP071088	99.8
AP004978	4	78226-79720	BP079561	99.7
AP006089	2	58758-60347	AV407620	99.5
AP006354	1	11065-13446	BP069447	99.5
AP006386	1	14519-16601	BP074812	100
AP006666	3	54086-56147	BP030309	99.8
AP006696	3	10970-12581	BE044806	100

Supplemental Table 2.3. Pack-MULEs with multiple genomic copies

MULE Accession #	Chr	Element Position	Acquired Region	Genomic copy Accession #	Chr	Matched Region
AP004250	5	61720-63221	62526-62928	AP007403	1	102686-102297
AP004467	1	33667-35074	34519-34749	AP007896	2	4153-4377
AP004468*	4	10199-11190	10429-10604	AP006382	4	95987-95819
			10790-10883	AP006382	4	94326-94229
			10665-10770	AP004625	5	125022-125127
AP004470	4	17980-19741	18390-18583	AP006699	2	95582-95357
			19086-19315	AP006699	2	93981-93753
AP004471	3	79860-81302	80726-81001	AP007357	N/A	14373-14104
AP004477	1	30895-32534	31266-32774	AP007927	2	75145-75363
AP004481	6	31432-	31758-	AP007896	2	4378-4133?

		32835	32003			
AP004482	6	7000-8745	7826-8586	AP004974	3	14362-15146
AP004482	6	80480-81915	80997-81358	AP006648	1	7515-7882
AP004484	1	73409-75102	74179-74519	AP007633	2	17434-17094
AP004484	1	93218-94965	93604-93854	AP007360	N/A	3668-3419
AP004487	1	36895-38452	37205-37414	AP007927	2	75145-75362
AP004491	1	79241-81915	80997-81358	AP006648	1	7515-7882
AP004499	1	74840-75979	75451-75740	AP007896	2	4174-4462
AP004502	1	39977-41107	40256-40350	AP007553	3	11600-11512
			40687-40956	AP007553	3	11028-10768
AP004526	6	50696-52138	51578-51770	AP007357	N/A	14388-14167
AP004541*	1	25986-27582	26226-26500	AP007878	1	49851-50128
			26489-	AP007527	N/A	31818-31517

			26791			
			26790-27439	AP007424	2	91010-90360
AP004625	5	100312-101726	100772-101042	AP008012	5	76381-76116
AP004912	3	24854-26496	25235-25948	AP006362	1	941-1651
AP004915	4	29370-30673	30070-30450	AP007647	2	67861-67482
AP004939	1	14876-16331	15359-15726	AP007535	5	145165-144786
AP004939	1	91536-92953	92317-92425	AP007734	1	50632-50525
AP004946	N/A	43306-44917	43978-44576	AP008079	1	52324-52873
AP004946	N/A	88225-90071	88874-89639	AP007656	N/A	49871-50636
AP004948	N/A	1351-2831	1995-2153	AP006352	2	25150-25005
AP004956*	1	67851-69464	68125-68538	AP008020	5	9996-9581
			68830-69149	AP008020	5	9935-10248
			68471-	AP007702	N/A	63152-63296

			68615			
AP004962	1	81323-82887	81998-82725	AP008079	1	52325-53025
AP004963	3	29363-31061	29736-29939	AP007418	N/A	59048-59252
			30023-30330	AP007418	N/A	59269-59571
AP004965*	1	155880-17696	16153-16499	AP008029	6	12602-12950
			16483-16577	AP007922	6	5885-5979
AP004968	1	130271-131828	130297-130513	AP006403	1	54781-55013
			131049-131197	AP006403	1	56088-56231
AP004970*	5	87213-89571	88017-88996	AP007702	N/A	82420-81441
			89052-89248	AP008020	5	10034-10226
AP004971*	5	54074-55749	54876-55214	AP008079	1	39903-39571
			55473-55599	AP007693	6	28172-28049

AP004981	3	36953-38647	37562-37778	AP007633	2	17019-17337
AP006074	3	31740-33026	32155-32501	AP007591	1	11584-11241
AP006079	6	1500-3115	2237-2496	AP008005	1	111823-112089
			2547-2656	AP008005	1	112083-112193
AP006084	3	77525-78709	77778-78257	AP008078	1	32200-32702
AP006094	4	83242-85104	83992-84610	AP008181	N/A	24733-25351
AP006094	4	118745-120281	118904-119120	AP007996	1	27877-28090
AP006095	4	5236-6438	5463-6145	AP008173	1	24386-25068
AP006097	1	19249-20755	19814-20240	AP006137	5	49797-50188
AP006098	4	80624-81755	81024-81358	AP008014	3	6881-6552
AP006102*	4	58327-60715	58702-59532	AP008124	N/A	31004-30164
			59502-60303	AP007804	1	20414-19618

AP006103	1	47202-49118	48254-48548	AP004929	3	4624-4918
AP006113	4	17905-19516	18068-18845	AP008079	1	52324-53025
AP006121	1	27911-29542	28930-29193	AP008136	1	7847-8103
AP006122*	1	32607-33980	32885-33057	AP007528	3	8254-8084
			33038-33634	AP007360	N/A	3427-4025
AP006144	3	70733-71534	70892-71215	AP008096	5	19306-19623
AP006354*	1	11065-13446	11389-11611	AP008020	5	10232-10013
			11642-12621	AP007702	N/A	81441-82420
AP006355	6	53596-54745	54282-54391	AP004513	5	36268-36376
AP006380	4	50032-54754	51304-51932	AP004485	2	16825-16196
AP006383	6	69804-71355	70500-71001	AP007832	3	44585-45084
AP006390*	4	8197-9633	8928-9074	AP007865	2	100925-

						100778
			9063-9329	AP007375	3	39696-39961
AP006396	3	18322-19108	18390-18646	AP008036	4	4003-3746
AP006417*	4	46862-48660	47161-47550	AP007403	1	70245-70738
			47543-48066	AP008180	N/A	1031-498
AP006417	4	97428-98939	97982-98615	AP007882	4	43531-44165
AP006419*	4	34677-36050	34991-35619	AP007460	1	4058-3418
			35601-35777	AP007528	3	8084-8254
AP006535	2	67186-68683	67787-68552	AP007802	4	14677-15439
AP006650	6	10753-12209	11211-11911	AP007493	4	11427-10755
AP006654	3	92780-94377	92943-93550	AP004898	3	23128-22520
AP006663	2	32507-34156	32705-32814	AP004496	1	34093-34009
			33827-	AP004496	1	33307-33122

			34022			
AP006675	2	5917-6719	6210-6447	AP006352	2	89571-89831
AP006676	1	21576-23370	21820-22817	AP007688	2	60249-59246
AP006685	1	76101-77517	76940-77215	AP007357	N/A	14373-14104
AP006686*	2	27555-28636	27893-28028	AP007412	N/A	27274-27383
			28322-28477	AP007996	1	27935-28090
AP006688*	2	28274-29346	28399-28593	AP007996	1	28126-27934
			28886-28996	AP007412	N/A	27383-27273
AP006691	3	46605-48021	47250-47766	AP007763	2	89909-90425
AP006692	2	76808-78392	77610-77784	AP007832	3	44707-44876
AP006696	3	79856-80642	80318-80574	AP008036	4	3743-4003
AP006700	2	63111-64684	64308-64524	AP007996	1	27876-28088
AP006712	2	42330-	42683-	AP004521	3	9214-9324

		43473	42794			
AP006712*	2	87720-89413	88590-88803	AP007633	2	17120-17337
			89053-89166	AP007900	3	108168-108290

*Pack-MULEs containing sequence from two or more predicted genes

CHAPTER 3

ANALYSIS OF RECENTLY AMPLIFIED *LOTUS PONG* ELEMENTS

Abstract

Transposable elements (TEs) are the single largest output of genome sequencing projects, accounting for 15 - 20% of the *Arabidopsis thaliana (Arabidopsis)* genome and 30% of the *Oryza sativa* (rice) genome. Although TEs make up a significant fraction of most plant genomes, only a few of these elements have been shown to be active under normal conditions. Results from a computer-assisted analysis in 32.4 Mb of *L. japonicus* (*Lotus*) sequences revealed that *Lotus* genome all the major TE types found in previously characterized plant genomes, accounting for ~30% of the available sequence. In *Lotus*, 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) suggesting that they have recently amplified and may still be actively transposing in the genome. The first and only active *Pong*-like element reported to date is the rice *Pong* element. Therefore, to examine whether the *Lotus Pong*-like elements are indeed active in *Lotus* or a heterologous plant host such as *Arabidopsis*, several experimental analyses were performed. Results from these analyses revealed that although the *Lotus Pong-like elements* appear to have recently amplified and are transcriptionally active; they are currently not active at an easily detected level in *Lotus*

Introduction

Pong-like elements are Class 2 DNA elements that belong to the *PIF/Harbinger* TE super family (JURKA and KAPITONOV 2001; LE *et al.* 2001; ZHANG *et al.* 2001; JIANG *et al.* 2003). Structurally, *Pong*-like elements contain ~ 30bp terminal inverted repeats (TIRs) and a 3bp TAA or TTA target site duplication (TSD). Unlike most DNA TEs, *Pong*-like elements contain two open reading frames (ORFs), ORF1 and ORF2 (ZHANG *et al.* 2004). ORF1 contains a domain that shares weak similarity to the DNA binding region of myb transcription factors, and is therefore predicted to be involved in DNA binding of the TE terminal sequences (JIANG *et al.* 2003; ZHANG *et al.* 2004). ORF2 encodes the transposase and contains the signature DDE motif usually found in the catalytic domain of eukaryotic transposases (ZHANG *et al.* 2004). Recent molecular and biochemical studies revealed that both ORFs are required for transposition of the rice-*Pong* (YANG *et al.* 2007).

In some plant genomes such as *A. thaliana (Arabidopsis)*, *O. sativa* (rice) and *M. truncatula* (medicago), most *Pong*-like elements are either truncated or mutated and thus inactive. The first and only active *Pong*-like element reported to date is the rice *Pong* element (JIANG *et al.* 2003). The lack of active *Pong*-like elements in other organisms could be due to the lack of full-length elements with intact coding sequences. For example, in *Arabidopsis*, of the ~30 copies, only

one full-length copy containing multiple introns in the coding sequence was identified.

However, in the *Lotus japonicus* (*Lotus*) genome, a computational analysis on 8% of the genome sequence identified 20 copies of *Pong*-like elements. Of these, 15 are full-length, and 6 of these 15 have intact coding sequences (HOLLIGAN *et al.* 2006). In addition, 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). This suggests that members within these subfamilies have recently amplified in the *Lotus* genome and are likely to still be active. Because the activity of more than one DNA TE subfamily (belonging to the same TE superfamily) is a rare occurrence, the two *Lotus Pong* elements provided the raw material to investigate whether these two subfamilies are active in the *Lotus* genome.

In this study, several experimental assays were performed to examine whether these two families of *Lotus Pong* elements are indeed active. Results from these analyses revealed that although the *Lotus Pongs* appear to have recently amplified and are transcriptionally active; they are currently not active at an easily detected level in *Lotus.*

Materials and Methods

Plant material and DNA extraction. Miyakojima (MG-20, sequenced genome) 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 from MG-20 and B-129, and purified using the DNAeasy plant mini kit (QIAGEN, Chatsworth, CA).

Plasmid construction. Full-length *Lotus Pong*1A (AP004506) and *Pong*3A (AP006430) were PCR amplified from *Lotus* genomic DNA (MG-20 ecotype) using Invitrogen Platinum Taq polymerase and cloned into the binary vector pBin-mgfp5-er (YANG *et al.* 2007). A 930bp deleted version of *Pong* 1A (600bp from the 5' terminal region and 330bp of the 3' terminal region), and a 960bp deleted version of *Pong* 3A (610bp from the 5' terminal region and 350bp of the 3' terminal region) were cloned between the *Bam*H1 and *Xba*1 site of the GFP gene located in pBin-mgfp5-er, to obtain pBin-*miniPong*1A and pbin-*miniPong*3A. Full-length copies of *Pong*1A and 3A were then cloned between the *Sbf*1 and *Hin*dIII site of pBin-*miniPong*1A and pbin-*miniPong*3A to obtain pBin-*Pong*1A-*miniPong*1A and pbin-*Pong*3A-*miniPong*3A, respectively. A 430bp deleted version of the rice *Pong* element that comprised 220bp from the 5' terminal region and 210bp of the 3' terminal region was cloned between the *Xba*1 site in the GFP gene of pBin-mgfp5-er which already contained a copy of the rice full-length *Pong* element. All primer sequences are available upon request.

***Arabidopsis* transformation and selection**. The *Arabidopsis* ecotype Columbia was transformed with *Agrobacterium tumefacience* strain GV3103

containing the constructed binary vectors according to Bechtold *et al* (BECHTOLD *et al.* 1993). The seeds of transformed plants were collected and germinated on Murashige and Skoog solid medium (0.2% phytagel) containing 150mg/liter Timentin and 50mg/liter kanamycin. After 7 – 10 days of incubation at 26 degrees celsius and 16:8 day/night cycle, transformants were observed. Transgenic plants were then imaged using a fluorescence stereoscope Lecia MZ10 F (Leica, Wetzlar, Germany).

Analysis of excision events. Evidence of excision was first screened on the basis of GFP expression in the cotyledons and then further confirmed by PCR analysis. Genomic DNA was collected from the leaves and PCR was performed with the following primers 5 -agacgttccaaccacgtcttcaaagcaag-3' (35S forward) and 5' -cctctccactgacagaaaa-tttgtgccca-3' (GFP reverse). All bands were cloned into a TOPO vector (Invitrogen, Carlsbad, CA) and sequenced.

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 subterminal sequences for each *Pong*. Final annealing temperature for selective amplification was 56 ^{0}C with the ^{33}P-labeled primers. Primer sequences were the following: BfaI1+0, 59 GACGATGAGTCCTGAGTAG-39; Bfa1+T, 59-GACGATGAGTCCTGAGTAGT-39; *Pong* P1, 59-CTTKAAGGCTCTCTCCAATG-39; *Pong* P2, 59 GGTCTTAGCAACTCCAG-39;

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.

Results and Discussion

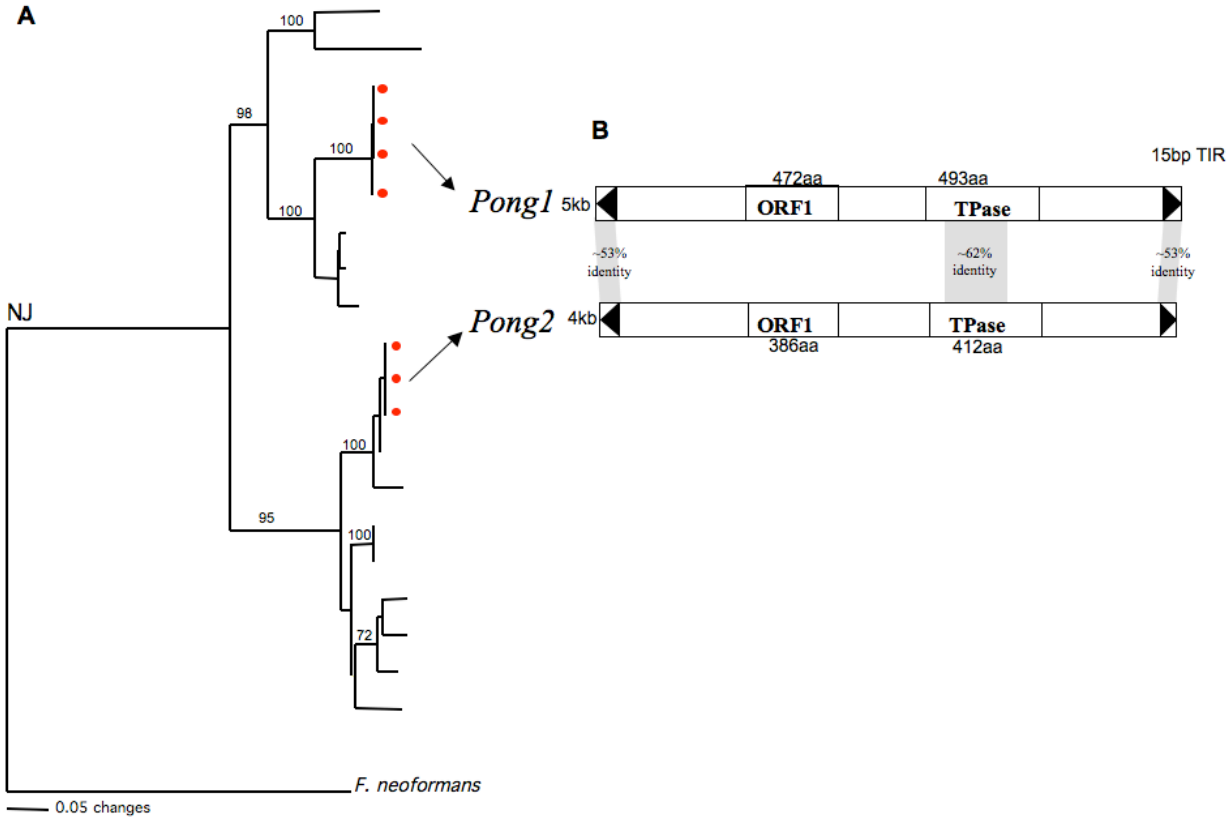
***Pong*-like elements in *Lotus*.** Two families of *Lotus Pong*-like elements have members that share ~98% nucleotide sequence similarity across the entire length of the element, and both ORF1 and ORF2 are intact (not interrupted by stop codons) (Figure 3.1). To test whether these elements are active in *Lotus*, transposon display analysis was first performed on sibling plants from the two *Lotus* ecotypes (Miyakajima and Gifu). Because the transposon display technique is able to visualize multiple bands simultaneously, new insertions should be easily detected (see Material and Methods for details). Based on this analysis, no new insertions were detected in any of the sibling plants for either subfamily, indicating that these elements might not be active under normal conditions [Figure 3.2 (A) and (B)]. Because the rice *Pong* element was shown

to be active in a rice tissue culture strain, and transposons are known to be reactivated by stress, *Lotus* callus was induced for 6 months and screened by transposon display to detect whether these *Lotus Pong* elements can transpose. However, results from this analysis did not detect any new insertions, similar to what was observed in the previous transposon display analysis using normal genomic DNA [Figure 3.3 (A) and 3.3 (B)]. It is possible that the time spent in tissue culture for the *Lotus* calli was insufficient. The tissue culture strain that was used in rice *Pong* was maintained for more than ten years. Together these results suggest that although the two *Lotus Pongs* have recently amplified, they are not presently transposing in the DNA samples tested for the *Lotus* genome.

Lotus Pong* in *Arabidopsis. Several TEs have been shown to be active when introduced into heterologous plant hosts. For example, *Tto1* from tobacco is active after transformation into both *Arabidopsis* and rice (HIROCHIKA *et al.* 2000). More recently, a study of rice *Pong* elements showed that, in *Arabidopsis*, the rice *Pong* is also active (YANG *et al.* 2007). Therefore to determine whether the *Lotus Pongs* are active in another plant system, *Lotus Pong* 1A and 3A were transformed into *Arabidopsis* using a previously described transposition assay (YANG *et al.* 2007). Briefly, a T-DNA containing the full-length *Pong* element (transposase source) and deleted version (pBin-*Pong*-*mpong*-gfp5-er, 900bp) were transformed into *Arabidopsis* [Figure 3.4 and Figure 3.5 (A)].

Figure 3.1: Phylogeny and structure of *Pong*-like elements in *Lotus*.

A) The phylogenetic tree was generated from the catalytic domain of the transposase using the neighbor-joining method, and bootstrap values were calculated from 250 replicates. Red circles are used to indicate individual *Pong*-like elements not apparent from the short branch-lengths. B) Structure of *Pong1A* and *Pong 3A*. The ORF1 and the transposase (TPase) regions are indicated for each *Pong*. Black arrows represent terminal inverted repeats and shaded areas share sequence homology between the two *Pongs*

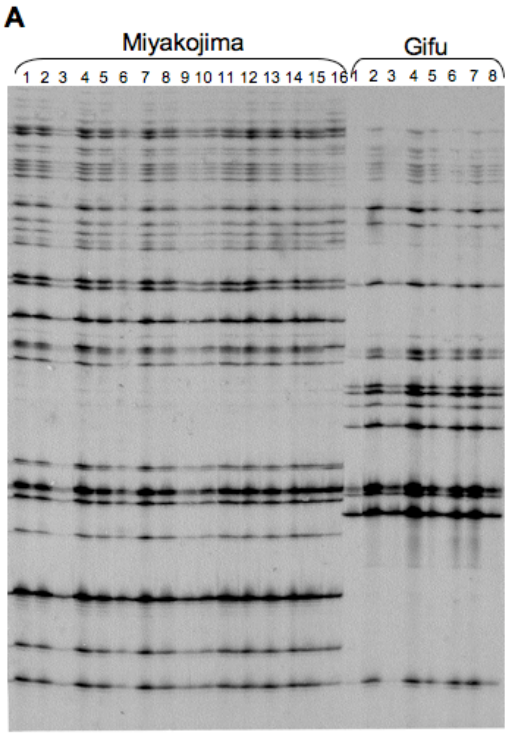

A
100
98
100
100
NJ
100
95
100
72
F. neoformans
0.05 changes
Pong1
Pong2
B
15bp TIR
472aa
493aa
5kb
ORF1
TPase
~53%
identity
~62%
identity
~53%
identity
4kb
ORF1
386aa
TPase
412aa

The deleted *Pong* is inserted into the 5'UTR of the GFP gene, thus if excision occurred, seedlings will exhibit GFP sectors. However, after screening of numerous transformants, no GFP spots were observed [Figure 3.5 (B)]. Therefore to further confirm that no excision event occurred, a PCR analysis using primers flanking the deleted version of the *Pong* elements (inserted in the GFP gene), and also flanking the full-length *Pongs* was performed. No excision products were detected for any of the 5 transformants analyzed for both the deletion derivative and full-length *Pongs* (Figure 6 and data not shown). Because it is possible that excision did not occur in the primary transformants (T1 generation) due to low transposase protein level, seedlings from the next generation (T2) were screened for excision. However no evidence of excision was observed for either *Lotus Pong* family (Figure 3.7 and data not shown). Together these results suggest, as was observed in *Lotus,* that these *Pong* elements are not active in *Arabidopsis*.

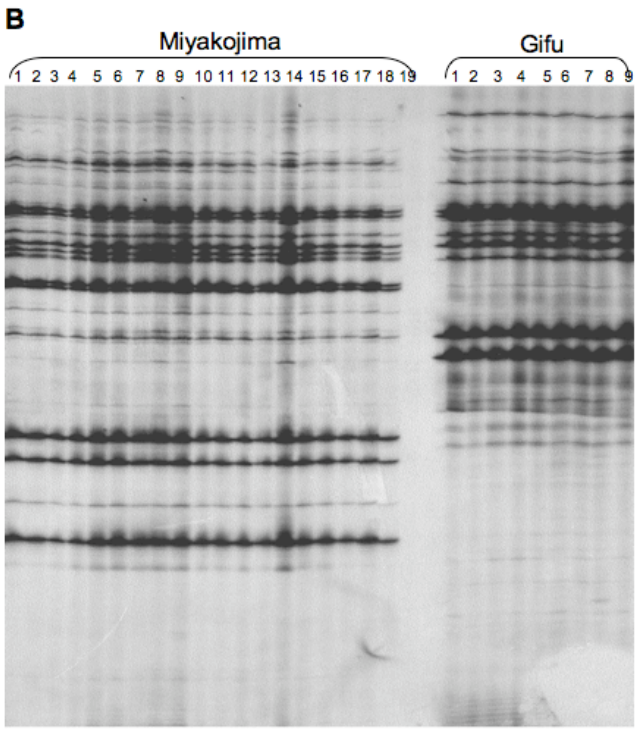
To further investigate and explain the lack of transposition for these *Lotus Pong* elements, several analyses were performed. It is known that for most coding TEs, transcription is a prerequisite step for transposition (HIROCHIKA *et al.* 2000). Therefore in order to determine whether the lack of *Lotus Pong* transposition in *Arabidopsis* could be due to lack of expression of the coding sequences, an RT-PCR analysis was performed. Based on this analysis both ORFs from the two *Lotus Pongs* are expressed in *Arabidopsis* seedlings (Figure 3.8).

Figure 3.2: Transposon display of *Lotus Pong* 1A and 3A.

Sublineage specific primers were designed based on Figure 1 and transposon display analysis was performed with these primers along with a *Bfa*1+T primer, and resolved on a 6% polyacrylamide gel. A) *Lotus Pong* 1A. Lanes 1 - 16: genomic DNAs from individual (siblings) plants from Miyakojima (M), Lanes 1 - 8: individual plants from Gifu (G). B) *Lotus Pong* 3A. Lanes 1 - 19: genomic DNAs from individual (siblings) plants from Miyakojima (M), Lanes 1 - 9: individual plants from Gifu (G).



Pong 1A



Pong 3A

Figure 3.3: Transposon display analysis *Pong*1A and 3A in *Lotus* tissue culture.

Transposon display was performed as described in Figure 3.2. A) *Lotus Pong* 1A. B) *Lotus* 3A. Lane 1, genomic DNA from Miyakojima; Lane 2 - 4, genomic DNA from *Lotus* callus.

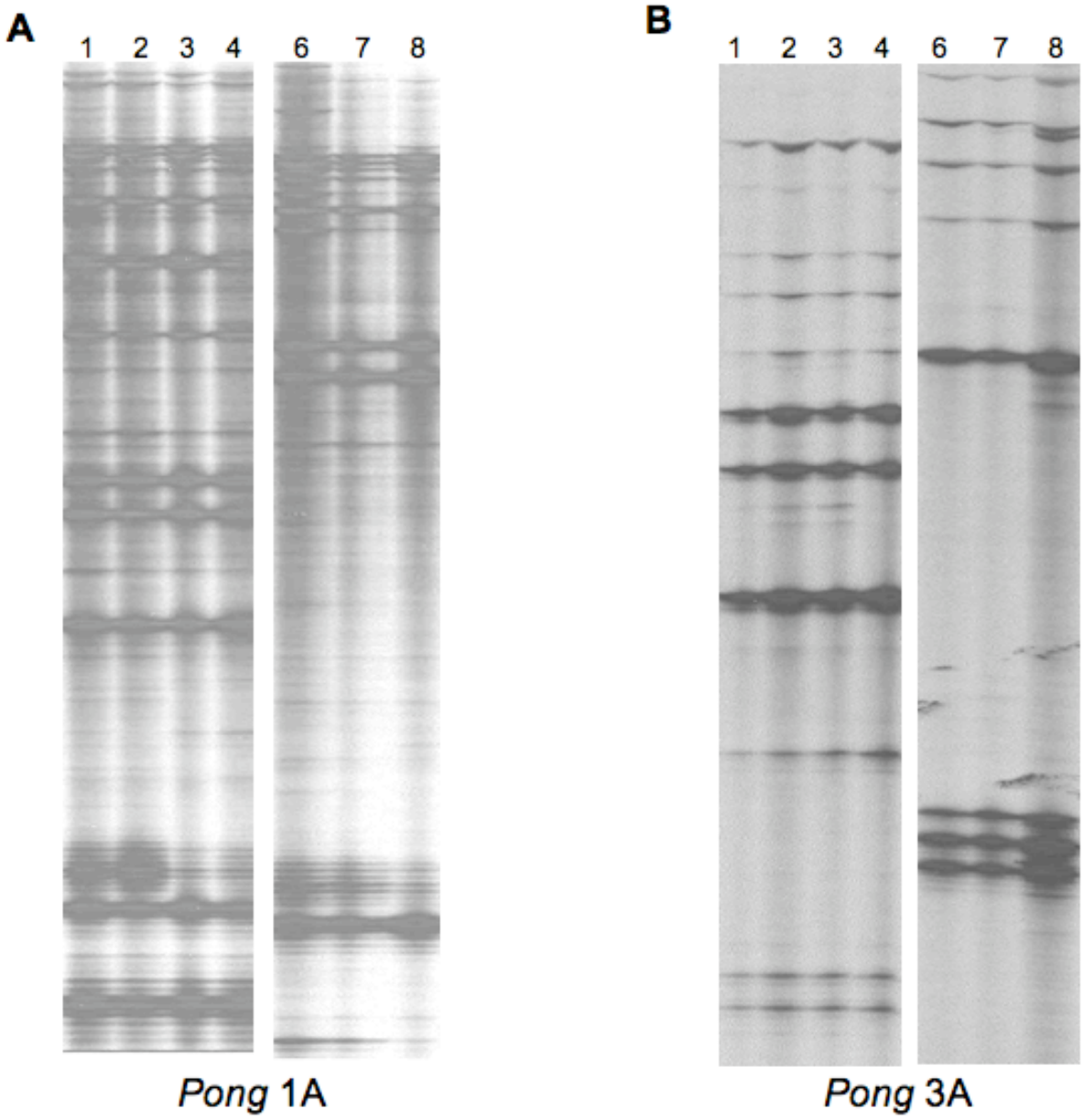
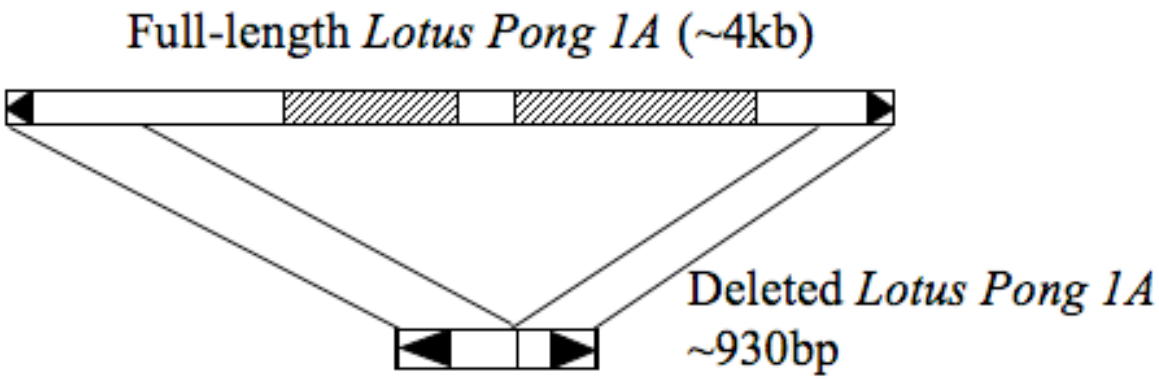

A
1 2 3 4 6 7 8
Pong 1A
B
1 2 3 4 6 7 8
Pong 3A

Figure 3.4: Structures of *Lotus Pongs* and *miniPongs*.

A and B represent the two *Lotus Pongs*, *Pong* 1A and *Pong* 3A. Predicted ORFs (ORF1 and ORF2) are shown as hatched boxes. The black triangles represent TIRs. Numbers in brackets represent the length of the full length sequence for each *Pong*. Deleted *Lotus Pong* 1A and 3A are referred to as *miniPong* throughout the text and construction details are found in the Materials and Methods section.

A


Full-length *Lotus Pong 1A* (~4kb)
Deleted *Lotus Pong 1A*
~930bp

B

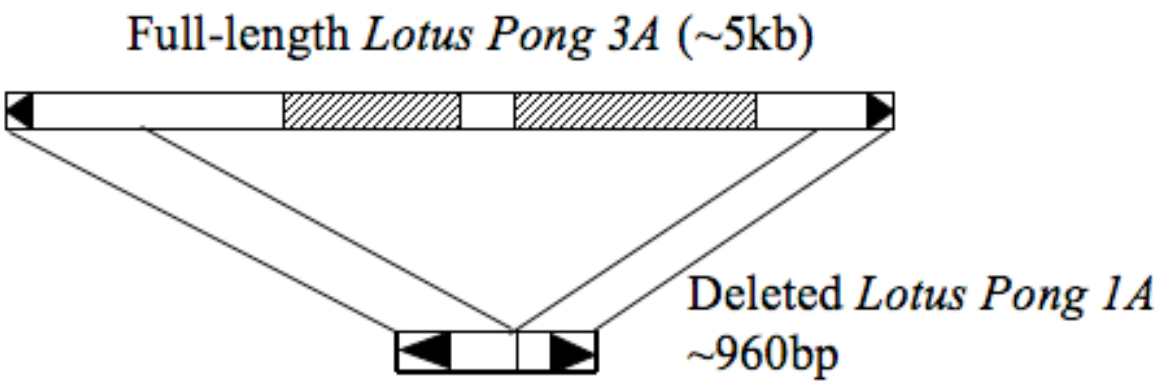

Full-length *Lotus Pong 3A* (~5kb)
Deleted *Lotus Pong 1A*
~960bp

Figure 3.5: T-DNA constructs transformed into Arabidopsis.

A) Constructs depicting both *Lotus Pong1A* and *3A* for simplicity. See Figure 3.4 for construct details. B) To the right of each construct are representative images of transformants with red fluorescence from chlorophyll. npt II, neomycin phosphate transferase; RB and LB, right and left borders of T-DNA; Pnos, promoter of nopaline synthase gene; Tnos, terminator of nopaline synthase gene; P35S, promoter of CaMV 35S gene.

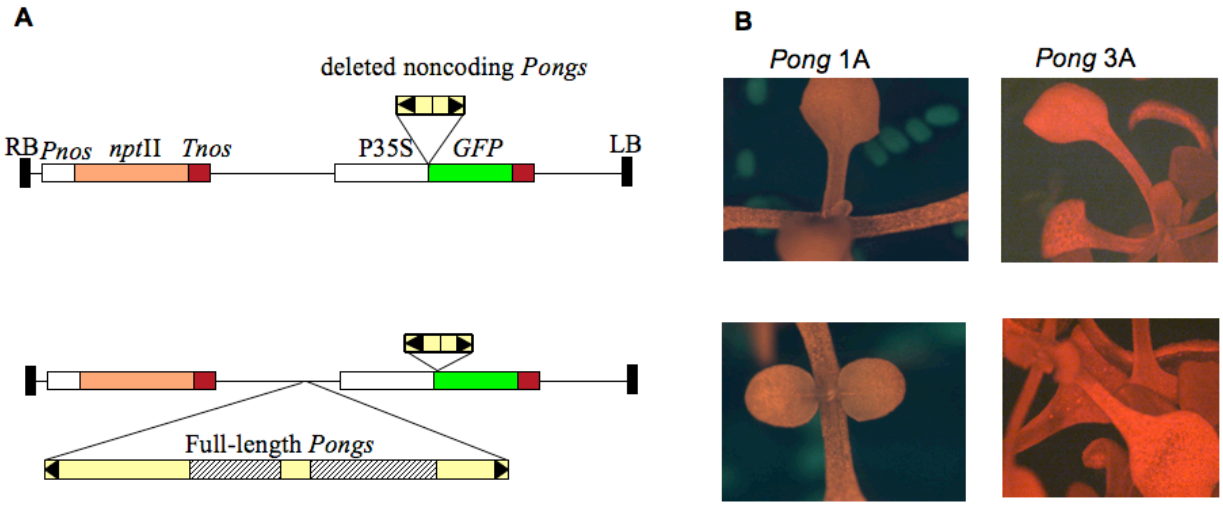
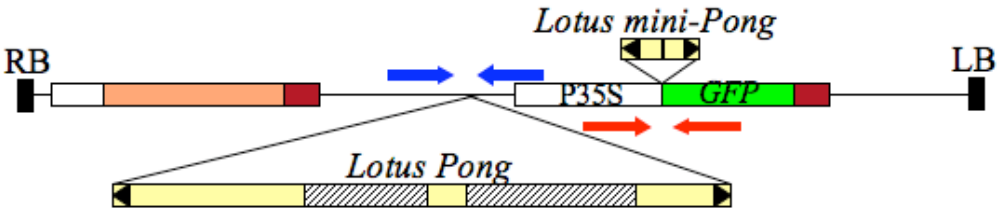
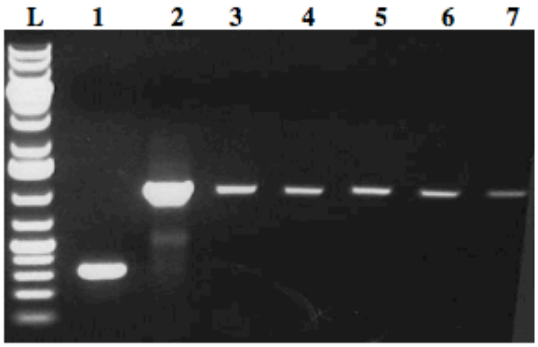

A
deleted noncoding Pongs
RB
Pnos
nptII
Tnos
P35S
GFP
LB
Full-length Pongs
B
Pong 1A
Pong 3A

Figure 3.6: PCR analysis of *Lotus mini-Pongs* excision in Arabidopsis.

A) Constructs used as described in Figure 3.5. The red arrowheads indicate PCR primer locations for *mini-Pongs* excision analysis. (B) Agarose gel of PCR products in transgenic seedlings. Five independent transformants are shown for each construct. L represents the size marker, a 1-kb ladder (New England Biolabs). Lane 1: empty plasmid/vector, Lane 2: plasmid control with *mini-Pongs*, Lane 3 - 4: DNA from transformants (T1) with *Lotus mini-Pong* but lacking *Lotus Pong*, Lane 5 - 7: DNA from transformants with both *Lotus mini-Pong* and *Lotus Pong* inserts.


Lotus mini-Pong
RB
LB
P35S
GFP
Lotus Pong


L 1 2 3 4 5 6 7

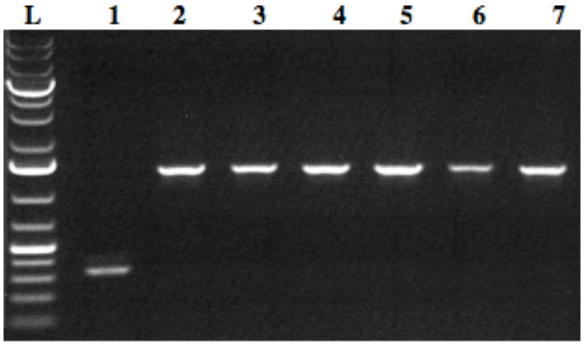
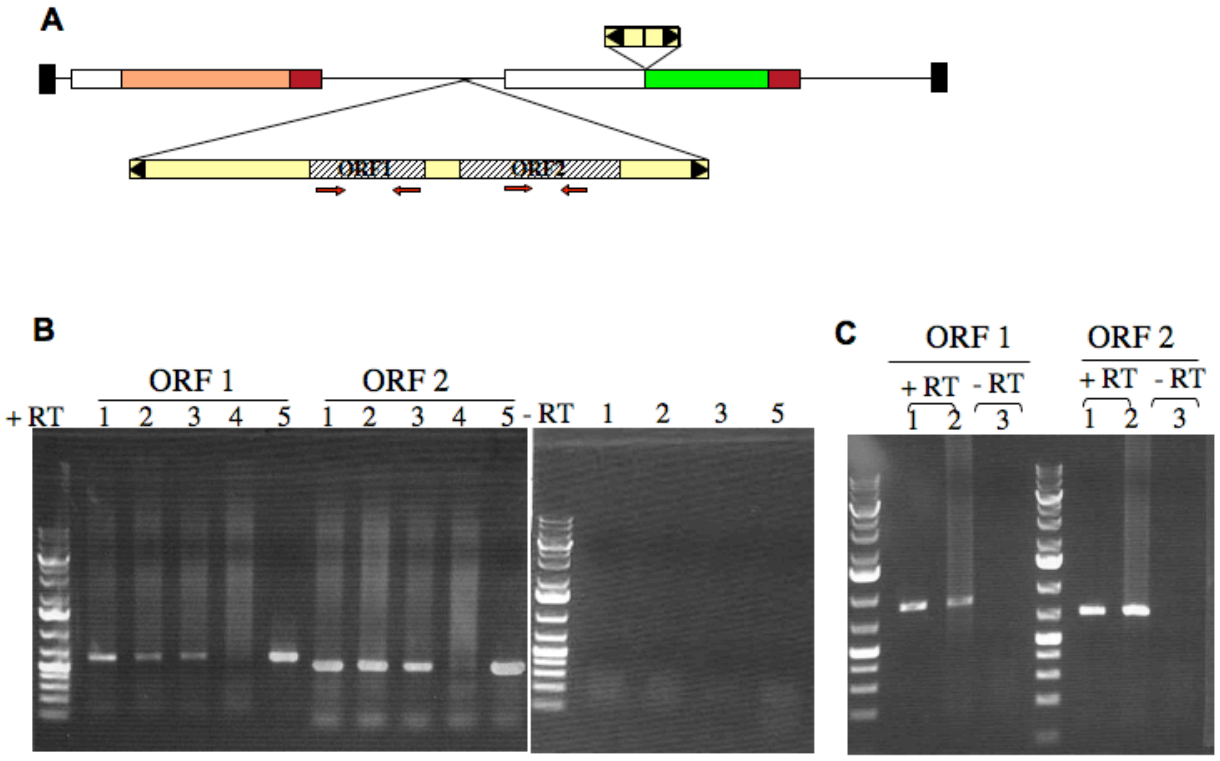

L 1 2 3 4 5 6 7

Figure 3.7: RT-PCR analysis of ORF1 and ORF2 from *Lotus Pongs.*

A) Constructs used as described in Figure 3.5. The red arrowheads indicate PCR primer locations for the *ORF* expression analysis. (B) Agarose gel of RT-PCR of *Lotus Pong 1A* ORFs in transgenic seedlings. +RT, represents with reverse transcriptase and –RT without reverse transcriptase. L represents the size marker, a 1-kb ladder (New England Biolabs). Lane 1 - 4: cDNA from transformants, Lane 5: genomic DNA from transformants. C) Agarose gel of RT-PCR of *Lotus Pong 3A* ORFs in transgenic seedlings. L represents the size marker, a 1-kb ladder. Lane 1: cDNA from transformants, Lane 2: genomic DNA from transformants. Lane 3: RNA (-RT).


A
ORF1
ORF2
B
ORF 1
ORF 2
+ RT 1 2 3 4 5 1 2 3 4 5 - RT 1 2 3 5
C
ORF 1
ORF 2
+ RT - RT
+ RT - RT
1 2 3
1 2 3

These results rule out the possibility that the lack of transposition is due to no transcription. Although expression was detected, it is possible that the proteins are not being translated or are translated but not functional due to missense mutations.

However, another explanation is that the lack of transposition in *Arabidopsis* might be the due to the fact that an artificial deletion derivative of *Lotus Pongs* is being used as a reporter. No active direct deletion derivatives of *Pong*-like elements have ever been reported. Therefore, to test whether, a coding full-length *Pong* can excise its own deletion derivative; a deletion derivate of the active rice *Pong* element (*mpong*, 430bp) was generated and transformed into *Arabidopsis*. Primary transformants were screened and several showed GFP sectors. PCR analysis and sequencing of the products confirmed excision of the deleted *Pong* (Figure 3.9). This provided evidence that a *Pong*-like element can mobilize its own deletion derivative. However, the size of the deletion derivative from the rice *Pong* is almost half the length of the deleted *Lotus Pong* elements (430bp vs 930bp and 960bp). Therefore, it is possible that a smaller version of the *Lotus Pong* element might be more likely to transpose, based on the fact that the size of the element in a previous study in yeast appears to be crucial for excision to occur (G. Yang *et al*, manuscript in preparation). However, due to the sequence quality (high GC rich content) for both *Lotus Pong* elements, a smaller version (<500bp) of these elements was not feasible. In addition, a PCR assay to detect excision of the full-length *Lotus Pong* copy, did not detect excision of these elements.

In summary, several assays were performed to examine the activity of two recently amplified *Lotus Pong* elements in *Lotus* and *Arabidopsis*. However, our analysis did not detect the activity of these elements, suggesting that although they have recently amplified, they are not presently active. Alternatively, these elements might be active in the *Lotus* genome at undetectable (low) frequency. As more genomic sequences become available and possibly more active TEs identified, the transposition mechanism for more TE families will be better understood.

Acknowledgements

I would like to thank Guojun Yang and Feng Zhang for their assistance in plasmid construction and technical support.

Figure 3.8: Analysis of *Lotus mini-Pongs* excision in the T2 generation. A) Constructs are as described in Figures 3.4 and 3.5. The red arrowheads indicate PCR primer locations for *mini-Pong 1A* excision analysis. B) The dark red seedlings with true leaves are transgenic with red fluorescence from chlorophyll. C) Agarose gel of PCR products in transgenic seedlings. Four independent transformants are shown for each construct. L represents the size marker, a 1-kb ladder. P: empty plasmid/vector, Lane 1: plasmid control with *mini-Pong 1A*. Lane 2 - 4: DNA from T2 generation (T2) with both *Lotus mini-Pong 1A* and *Lotus Pong 1A* inserts.

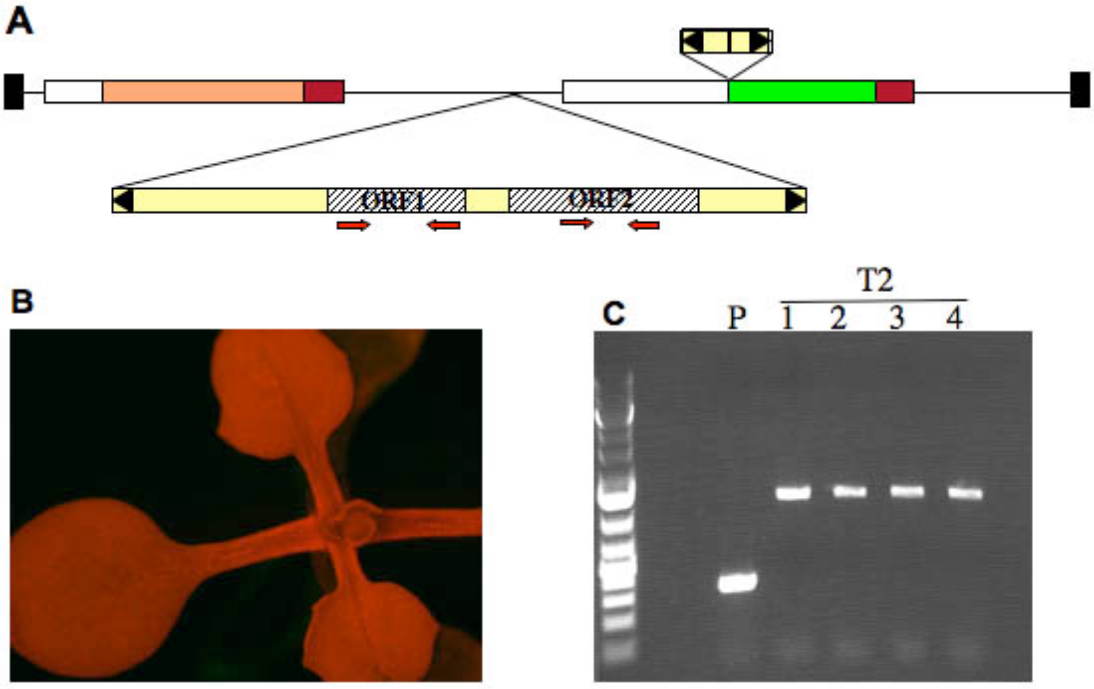
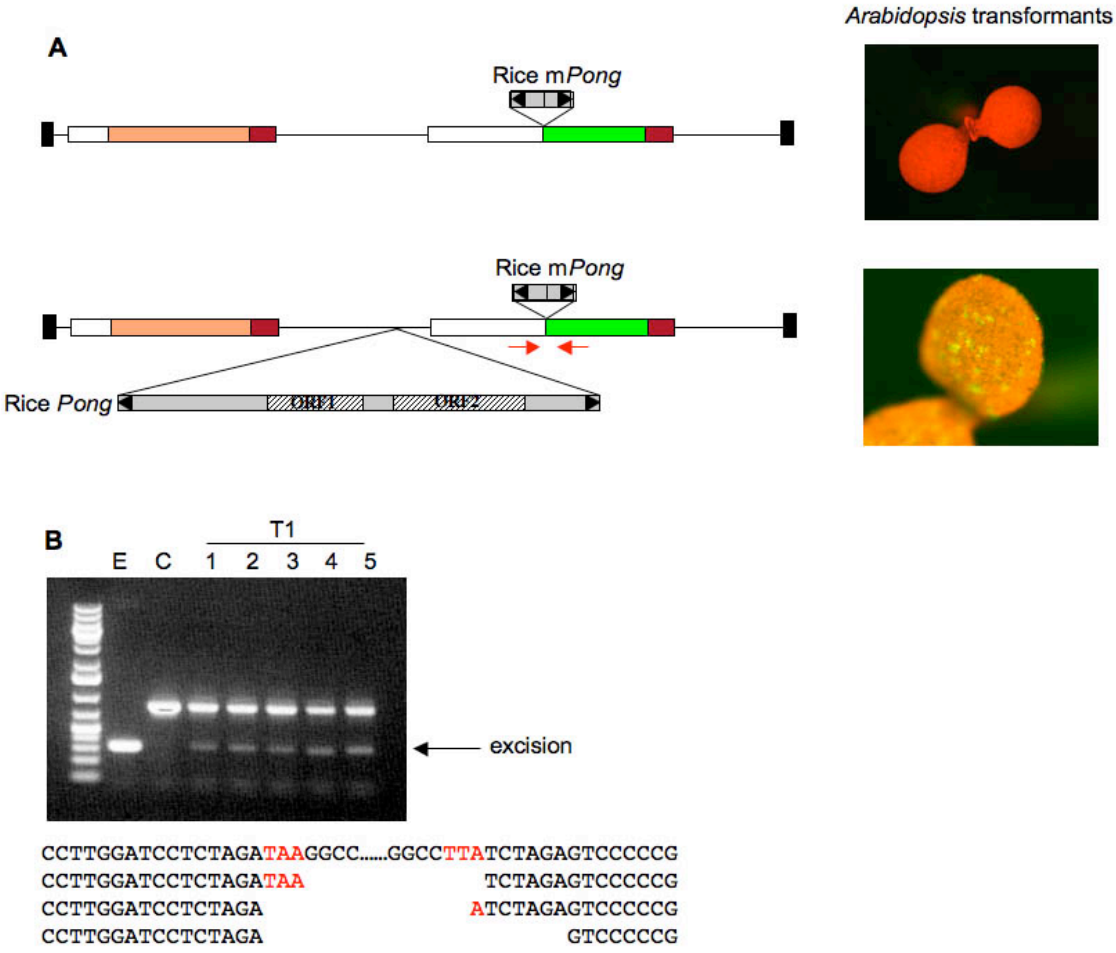

A
ORF1
ORF2
B
C
T2
P
1
2
3
4

Figure 3.9: Analysis of rice *Pong* and *mPong* in *Arabidopsis*.

A) Constructs are as described in Figures 3.4 and 3.5. See Material and Methods for rice *mPong* construction details. The red arrowheads indicate PCR primer locations for *mPong* excision analysis. The dark red seedlings with true leaves are transgenic with red fluorescence from chlorophyll and green fluorescence from GFP. (B) Agarose gel of PCR products in transgenic seedlings. Five independent transformants are shown. L represents the size marker, 1-kb ladder. E: empty plasmid/vector, C: plasmid control with *mPong*. Lane 1 - 2: DNA from T1 plants with *mPong* only. Lane 3 – 5: DNA containing both rice *Pong* and *mPong* inserts. Black arrow indicates excision products. C) (B) Sequences of *mPong* donor sites after excision events. The sequence before excision of *mPong* is shown at the top along with the the 3-bp TSD sequence (in red) (Lane 1) and the three remaining sequences show the excision sequences recovered for Lane 3 - 5.


A
Rice mPong
Rice mPong
Rice Pong
ORF1
ORF2
Arabidopsis transformants
B
T1
E C 1 2 3 4 5
excision
CCTTGGATCCTCTAGATAAGGCC.....GGCCTTATCTAGAGTCCCCCG
CCTTGGATCCTCTAGATAA TCTAGAGTCCCCCG
CCTTGGATCCTCTAGA ATCTAGAGTCCCCCG
CCTTGGATCCTCTAGA GTCCCCCG

References

BECHTOLD, N., J. ELLIS and G. PELLETIER, 1993 In planta Agrobacterium-mediated gene transfer by infiltration of adult Arabidopsis thaliana plants. . C.R. Acad. Sci. **316:** 1194-1199.

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 U S A **97:** 10083-10089.

HIROCHIKA, H., H. OKAMOTO and T. KAKUTANI, 2000 Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. Plant Cell **12:** 357-369.

HOLLIGAN, D., X. ZHANG, N. JIANG, E. J. PRITHAM and S. R. WESSLER, 2006 The transposable element landscape of the model legume Lotus japonicus. . Genetics **174:** 2215-2228.

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.

JURKA, J., and V. V. KAPITONOV, 2001 PIFs meet Tourists and Harbingers: A superfamily reunion. Proc Natl Acad Sci U S A **98:** 12315-12316.

LE, Q. H., K. TURCOTTE and T. BUREAU, 2001 *Tc8*, a *Tourist*-like Transposon in *Caenorhabditis elegans*. Genetics **158:** 1081-1088.

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

SWOFFORD, D. L., 1999 *PAUP*: phylogenetic analysis using parsimony and other methods.* Sinauer, Sunderland, MA.

YANG, G., F. ZHANG, C. N. HANCOCK and S. R. WESSLER, 2007 Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana. Proc Natl Acad Sci U S A. **104:** 10962-10967.

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.

ZHANG, X., N. JIANG, C. FESCHOTTE and S. R. WESSLER, 2004 Distribution and evolution of *PIF*- and *Pong*-like transposons and their relationships with *Tourist*-like MITEs. Genetics **166:** 971-986.

CHAPTER 4

DECIPHERING THE TRANSPOSITION AND AMPLIFICATION MECHANISM OF RICE MITES USING A YEAST ASSAY SYSTEM

ABSTRACT

Miniature inverted repeat transposable elements (MITEs) are considered unique among TEs because of their small size and high copy number. Because MITEs do not encode the proteins necessary for transposition and because some MITEs are not deletion derivates of coding elements in the genome, several studies have focused on understanding the origin and amplification mechanism of MITEs. Using the results from a previous study on rice *Stowaway*-like MITEs and their putative transposase source (*Osmar* elements, members of the *Tc1/Mariner* superfamily), an analysis was performed to understand why *Osmar14* transposase is able to mobilize a *Stowaway*-like MITE at higher frequencies than a deleted version of *Osmar14*. Results from this analysis provide insight into the best substrates for excision by the *Tc1/Mariner*-like transposase and the cis- requirements for *Stowaway*-like MITE transposition in rice.

Introduction

MITEs are a unique group of noncoding DNA transposons that are characterized by their high copy number and small size (FESCHOTTE *et al.* 2002b). The basic architecture consists of terminal inverted repeats that vary in length depending on the family, and all MITEs generate target site duplications upon insertion (CRAIG *et al.* 2002). Most MITEs in plant genomes are either *Tourist*-like or *Stowaway*-like, a distinction that is based on TIR and target site duplication sequences (BUREAU and WESSLER 1992; BUREAU and WESSLER 1994). While some MITEs do not share any significant sequence similarity with coding elements, others such as the active rice MITE *mPing* are direct deletion derivatives of coding elements (CRAIG *et al.* 2002; FESCHOTTE *et al.* 2002a). Furthermore, unlike other small noncoding DNA elements (deletion derivatives) that are present in low copy numbers, MITEs are highly abundant in eukaryotes (FESCHOTTE *et al.* 2002b; JIANG *et al.* 2004; NENE *et al.* 2007). For example, the rice genome contains ~60,000 *Tourist*-like MITEs and ~30,000 *Stowaway*-like MITEs (JIANG *et al.* 2004). Furthermore, the recently published mosquito genome (*A. aegypti*) revealed that MITEs account for ~16% of the genome (NENE *et al.* 2007). This is interesting since MITEs have been shown to preferentially insert into or near genes and can amplify to high copy numbers in a very short period of time (FESCHOTTE *et al.* 2002b; OKI *et al.* 2008).

Ever since the discovery of MITEs, one intriguing question concerns their abundance, specifically, how are they are able to amplify to high copy numbers given the known (conservative) mechanisms of transposition for DNA transposons. It has been proposed that the ability of MITEs to amplify to high copy numbers is due to at least two reasons. First, MITEs might be able to use the transposase source necessary for transposition from several somewhat related coding elements (FESCHOTTE *et al.* 2002b; JIANG *et al.* 2004). Second, MITEs might possess special features that are not present in typical deletion derivates.

The transposase source responsible for the amplification of *Tourist*-like MITEs was initially proposed to belong to the *PIF/Harbinger* superfamily based on sequence and structure similarity (JURKA and KAPITONOV 2001; ZHANG *et al.* 2004). However, this relationship has since been experimentally established by an active family of *Tourist*-like MITEs in rice (JIANG *et al.* 2003). The transposase source for *Stowaway*-like MITEs was deduced based on sequence and structure similarity and some indirect experimental results to the *Tc1/Mariner*-like superfamily (FESCHOTTE *et al.* 2005). However, this experimental evidence only demonstrated binding of the *Tc1/Mariner*-like transposase to the *Stowaway*-like MITE TIRs in vitro, not actual transposition of the MITE (FESCHOTTE *et al.* 2005). Therefore, further experimental analyses on this group of MITEs is essential for understanding the amplification and transposition mechanisms of MITEs in general.

The rice genome contains tens of thousands of *Stowaway*-like MITEs, making it a valuable resource for MITE analyses (FESCHOTTE *et al.* 2003). In rice, ~ 50 *Tc1/Mariner*-like elements (*Osmars)* have been identified and classified into 25 families, containing ~ 30,000 *Stowaway*-like MITEs (FESCHOTTE *et al.* 2003). *Osmar* and *Stowaway families* share ~10 bp terminal sequences and a target site duplication of the dinucleotide TA. However, none of the ~30,000 rice *Stowaway*-like MITEs are obvious deletion derivatives of the ~ 40 rice *Osmar* elements. Furthermore, there are no reports of a naturally active *Tc1/Mariner*-like element or *Stowaway*-like MITE in plants (FESCHOTTE *et al.* 2003). In order to understand the amplification mechanism of *Stowaway*-like MITEs and establish a functional relationship between *Tc1/Mariner*-like elements and *Stowaway*-like MITEs in rice, an active element must be identified along with an assay system for transposition.

In a prior study using a yeast excision assay system, the rice *Osmar5* transposase was shown to excise a deletion derivative (*Osmar5NA*) of the full-length *Osmar5* element but not any of the *Stowaway*-like MITEs tested. However, in a more recent study performed by a former postdoctoral fellow, Guojun Yang, a genome wide screen in yeast for functional interactions between rice *Stowaway*-like MITEs and *Osmars* demonstrated the excision of a *Stowaway*-like MITE (*stowt*35) by the rice *Tc1/Mariner*-like superfamily (*Osmar*14). The results from this analysis confirmed the long held hypothesis that *Tc1/Mariner*-like elements are indeed the transposase source for *Stowaway*-like MITEs in rice. However, from these studies, the most dramatic interaction

observed was the high excision frequency (~10,200 ± 2700) observed between *Osmar*14 transposase and *stow35* MITE. In contrast, Osmar14NAS excised at a much lower frequency (2.30 ± 1.33) (Table 4.1; with permission of G. Yang, manuscript in preparation). The fact that *Osmar*14 transposase can excise a *Stowaway*-like MITE better than its own coding element suggested that *stow35* is a much better substrate for excision by *Osmar*14 transposase than *Osm14NAS*.

In an effort to determine what features of *stow35* made it a better substrate for excision by Osmar14 transpose, several regions (TIR, subTIR and internal) were swapped between *Osm14NAS* and *stow35*, and a similar experimental assay was performed (Supplemental Figure 4.1). The results from this analysis showed that the most efficient substrate for excision by *Osmar*14 transposase contained the TIRs of *Osm14* and the internal region of *stow35* (Supplemental Figure 4.1).

I joined this project after the previous analysis to further investigate whether the internal region of *stow35* contains special cis-sequences for excision by *Osmar*14 transposase. Therefore, for the study in the rest of this Chapter, experiments were designed together with G. Yang and performed by me using the previous yeast assay system. The results from this analysis determined the best substrates for excision by *Osmar*14 transposase, along with the cis-requirements for *stow35* MITE transposition in yeast. Together these findings have provided insight into the transposition and amplification mechanisms of MITEs.

Results

In a previous study, it was proposed that the rice *Stowaway*-like MITE *stow35* might contain special internal sequences that allow it to be excised at higher frequency by *Osmar*14 transposase than the deletion derivative (*Osm14NAS*) of *Osmar*14. In addition, because none of the rice *Stowaway*-like MITEs are deletion derivates of rice *Osmars*, and the fact that *Osm14NAS* and *stow35* share sequence similarity in only the most terminal regions of their TIRs further suggested that the high excision frequency observed for *stow35* is probably due to its internal region. However, the highest excision frequency in the previous study was observed when the TIRs of *Osm14NAS* were combined with *stow35* internal region. Taken together, these results suggested that *Osm14NAS* TIRs contributed to the high excision frequency. Therefore, to determine whether the internal region of *stow35* plays a major role in its high excision frequency, and to further investigate the difference in excision frequency observed between *stow35* and Os*m14NAS,* several structural regions were swapped and tested for excison in the yeast assay. Briefly, regions (TIR, subTIR and internal) were swapped between *stow35* and *Osm14NAS*, along with *stow16*, another *Stowaway*-like MITE that is also mobilized at a higher frequency by *Osmar*14 transposase compared to *Osm14NAS* (16.5 $\pm$ 16.5 vs 2.30 $\pm$ 1.33; Table 4.1), and three random sequences (see Materials and Methods for details).

Chimeric constructs were then co-transformed with *Osmar*14 transposase and without (control) in a yeast excision assay (Figure 4.1, 4.2) (Yang *et al*. 2006).

The results from this analysis revealed several interesting findings. First, as was observed by G.Yang in a previous assay, *Osm14NAS* TIRs with *stow35* middle show the highest excision frequency (Figure 4.3 (C); G. Yang, manuscript in preparation). Second, constructs containing *Osm14NAS* TIRs with the internal regions of *stow35*, and a random sequence (GFP1), showed a higher excision frequency than constructs containing both *Osm14NAS* TIRs and subTIR regions with the similar internal regions [Figure 4.3 (C) and (F) vs (D) and (G)]. Furthermore, *Osm14NAS* TIRs showed a higher excision frequency when combined with the internal regions of *stow35* and a random sequence (GFP1), compared to *stow35* TIRs fused to the internal regions of *Osm14NAS* and the same random sequence (GFP1) [Figure 4.3 (C) and (F) vs (H) and (J)]. In addition, constructs bearing *stow16* internal region fused to *OsmNAS14* TIRs showed a much higher excision frequency when compared to the original *stow16* element [105.1 $\pm$ 6.7 vs 16.5 $\pm$ 16.5; Table 4.1, Figure 4.3 (E)].

Finally, unlike the higher excision frequencies observed for *Osm14NAS* TIRs versus the subTIRs; constructs containing only the TIRs of *stow35* showed a much lower excision frequency when compared to those with both *stow35* TIR and subTIR regions [Figure 4.3 (H), (J), (L) and (N) vs (I), (K), (M) and (O)]. Taken together, the results presented here showed that the 32bp TIRs of *Osm14NAS* and the internal region of *stow35* extending into the subTIRs provide the best substrate for excision by *Osmar*14 transposase.

Discussion

The ability of MITEs to amplify to much higher copy numbers than other noncoding elements in the genome suggested that there could be something special about MITE structure. In this study, the availability of an active *Stowaway*-like MITE, a functional transposase and an established yeast assay system, provided a unique opportunity to investigate whether MITEs are indeed special, and what these special features could be.

In all chimeric constructs examined that included the internal region of *stow35* MITE, a higher excision frequency was observed when compared to the original constructs. Although the internal region of *stow35* appears to enhance the excision of *Osm14NAS* TIRs, this enhancing effect appears to be concentrated in the subTIR portion of the internal region. This is demonstrated with constructs containing the internal region extending into the subTIRs or lacking the subTIR regions [Figure 4.3 (C) and (D)]. Furthermore, this is also illustrated by comparing several constructs containing only the *stow35* TIRs or together with the subTIRs [Figure 4.3 (H), (J), (L) and (N) vs (I), (K), (M) and (O)]. However, the functional significance of this region in the transposition of *stow35* MITE is not known. For other noncoding DNA transposons, such as the maize *Ds* element, sites in the subterminal region are required for excision of the element during transposition (CHATTERJEE and STARLINGER 1995).

As shown in the Figure 4.3, the highest excision frequency was obtained when *Osm14NAS* TIRs were fused to the *stow35* internal region [Figure 4.3 (C)].

This suggests that either the TIRs of *Osm14NAS* or the internal region of *stow35* contributes to high excision. The results based on several comparisons between different constructs strongly suggest that the TIRs of *Osm14NAS* are also ideal substrate for high transposition by *Osmar*14 transpose. For example, when the *Osm14NAS* TIRs were combined with the internal regions of *stow16*, it enhanced the excision frequency compared to the original element [Figure 4.3 (E); Table 4.1). In addition, even when *Osm14NAS* TIRs are combined with random sequences, the excision frequency is significantly higher than that of *Osm14NAS* and similar to that of *stow35* [Figure 4.3 (B), (E), (F)]. This suggests that very little sequence is needed to generate an artificial MITE with an excision frequency that is as good as the endogenous rice MITE used in this assay.

Finally, based on previous reports, the TIRs and subTIR regions of *Tc1/Mariner*-like elements are critical for transposition because they are bound by the transposase during transposition (ZHANG *et al.* 2001). In this study, in all cases when the *Osm14NAS* TIRs and subTIRs were included, the excision frequency was much lower than when only the TIRs were present [Figure 4.3 (D) and (G) vs (C) and (F)]. This suggests that the *Osm14NAS* subTIR sequences might be repressing excision of the element by *Osmar14* transposase, and this hypothesis was experimentally confirmed by G.Yang based in a mutational analysis of this region. However, the actual mechanism of inhibition remains unknown.

In conclusion, the results presented here provide experimental evidence to support that a rice MITE contain special internal regions that are ideal substrates

for transposition by a particular transposase. Whether the internal region of MITEs contributes to the high amplification of MITEs observed in eukaryotes, will require further experimental analyses to first understand the functional interaction between MITEs and their transposase source.

Materials and Methods

Plasmid construction. *Osmar*14 transposase sequence was amplified from rice genomic DNA before being pieced together to obtain a full length coding sequence. *Osm14NAS* (240 bp) was PCR from an ~1 kb version (*Osm14NA*) used in a previous analysis (G.Yang, manuscript in preparation). *stow35* and *stow16* were amplified from previously cloned plasmids using element-specific primers. *Osmar*14 transposase was cloned into a pRS413 based vector as previously described (YANG *et al.* 2006).

Chimeric constructs. To obtain constructs with swapped TIRs, subTIRs and internal regions, PCR primers were designed to include the terminal 32 bp of one element (e.g. *stow35*) as 5' overhangs attached to the oligos priming to the subterminal region of the other element (e.g. *Osm14NAS*). Similarly, to obtain constructs with swapped terminal regions, 65bp terminal sequences (include TIR and subTIR) were attached as 5' overhangs to the oligos priming to the internal sequences of the other element. Random internal sequences (GFP1, GFP2, and

GFP3) were PCR amplified from a pBin-mgfp5-er plasmid with sequences corresponding to three different regions of the GFP sequence. Chimeric PCR products (240 bp) were then cloned into the *Hpa*I site in the *ade2* reporter gene on pWL89A and used for the yeast excision assay (Figure 3.2; (YANG *et al.* 2006). All oligo sequences are available upon request.

Yeast excision assay. Plasmids containing the chimeric constructs and *Osmar*14 transposase described above were transformed into yeast and analyzed as follows. Transformation reactions [50 μL of competent cells, 5.8 μL of 5 mg/mL denatured salmon sperm DNA, 1 μL (~200 ng) each of plasmids and 400 μL of 50% PEG-3500 buffer (Gietz and Woods, 2002)] were incubated at 42^0 C for 45 min. Cells were then plated on plates containing complete supplement mixture (CSM) (Q-BIOgene), but lacking histidine and uracil, and incubated at 30^0 C. Colonies appeared after 3 to 4 days of incubation and were grown to saturation at room temperature (~10 days). To select ADE2 revertants from the double transformants, and calculate the excision frequency, three colonies from plates lacking histidine and uracil were picked and placed into 50 μL of water, of which 49 μL was plated onto CSM plates containing 2% galactose but lacking adenine, and 1 μL was used for 10^6 dilutions. Of the diluted yeast cell suspension, 49 μL was plated on YPD (yeast extract/peptone/dextrose) to calculate the total number of live yeast cells in the cell suspension. The revertant frequency (excision frequency) was calculated as the number of ADE2 revertants per cell [(ADE/YPD, Figure 3.2);(YANG *et al.* 2006)].

Acknowledgements

I would like to thank Guojun Yang for permission to use some of his previous data and constructs for this project. This work was supported by a grant from the National Science Foundation to S.R.W.

Table 4.1. ADE2 reversion frequency of *Osm14NAS*, *stow35 and stow16.*

Donor	Transposase source – *Osmar*14
Osm14NAS	2.30 ± 1.33
*stow*35	10200±2700
*stow*16	16.5±16.5[†]

Frequency: 10^{-9}; Error, standard error of the mean for three independent events.

[†] *stow*16 excision frequency was calculated based on one revertant colony obtained from three independent events. .

Figure 4.1. Structure of *Osm14NAS* and *Osmar*14 transposase.

White boxes represent *Osmar*14 coding exons, and shaded regions represent noncoding sequences. Dotted box represents an intron. The three Asp residues (D195, D318, and D358) constitute the putative DD39D motif. The 32bp TIRs are shown as black triangles. Dashed lines indicate shared regions. The length of *Osm14NA* and *Osm14NAS* is noted to the right in base pairs (bp).

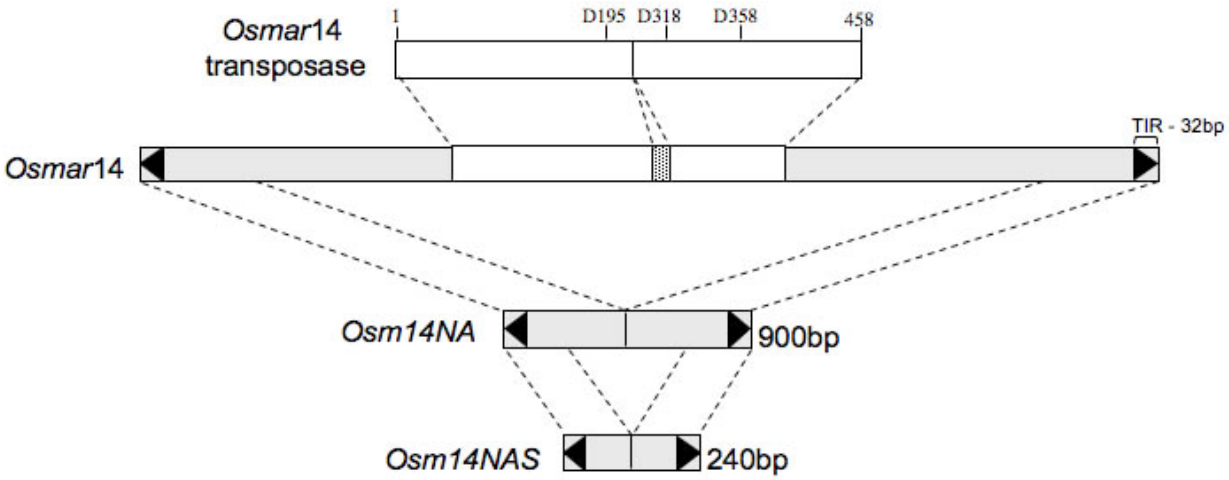

Osmar14
transposase
1
D195
D318
D358
458
TIR - 32bp
Osmar14
Osm14NA
900bp
Osm14NAS
240bp

Figure 4.2. Yeast transposition assay constructs and protocol.

pOsm14Tp, pRS413 vector with Osmar14 transposase. Empty pRS413, control vector similar to pOsm14Tp but without the transposase. pWL89A, reporter plasmid. The grey rectangle box represents the constructs shown in Figure 3.3. See Materials and Methods. amp, ampicillin resistance gene; ARS1, autonomous replication sequence1; ARS H4, autonomous replication sequence of the H4 gene; CEN6 and CEN4, centromere sequences of yeast chromosomes 6 and 4, respectively; *cyc1* ter, terminator of yeast cyclin gene *cyc1*; OriEC, E. coli replication origin; *Pgal1*, yeast *gal1* promoter.

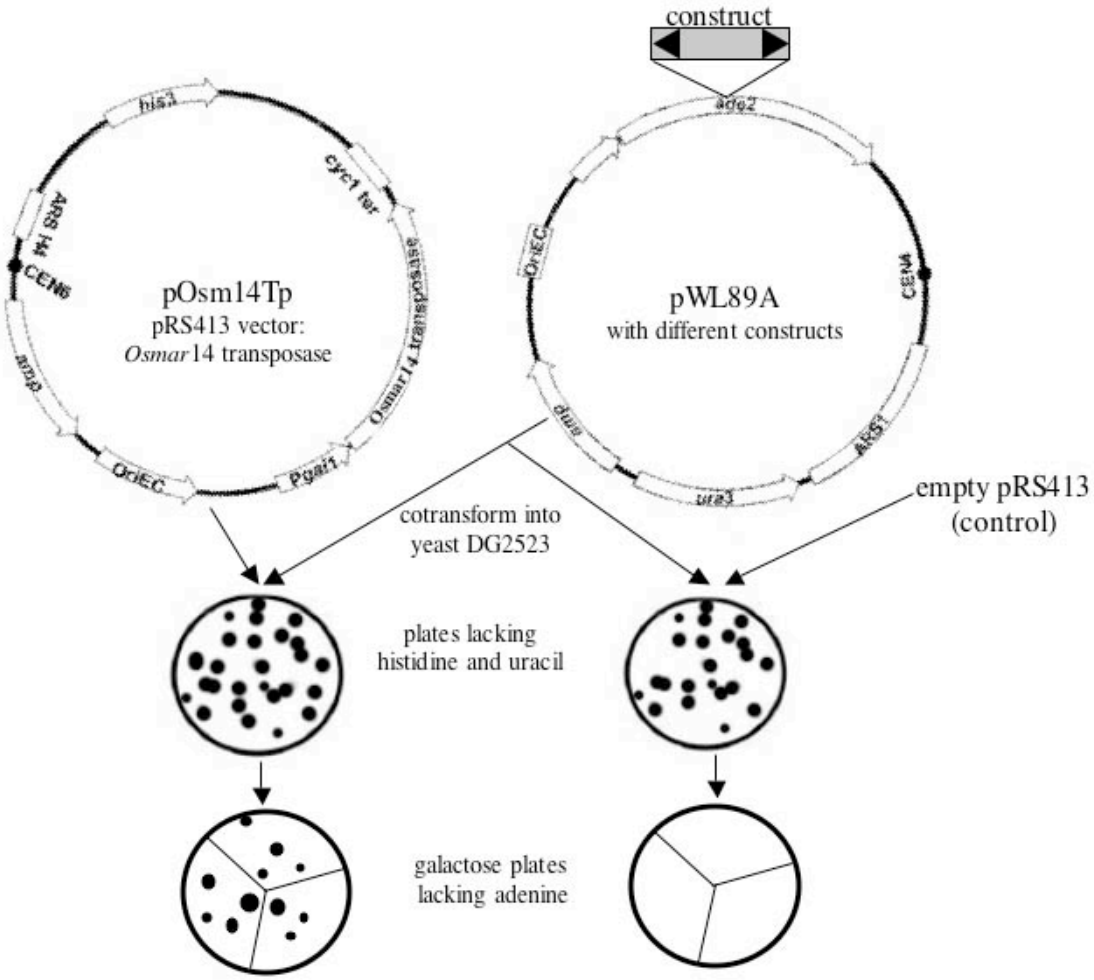
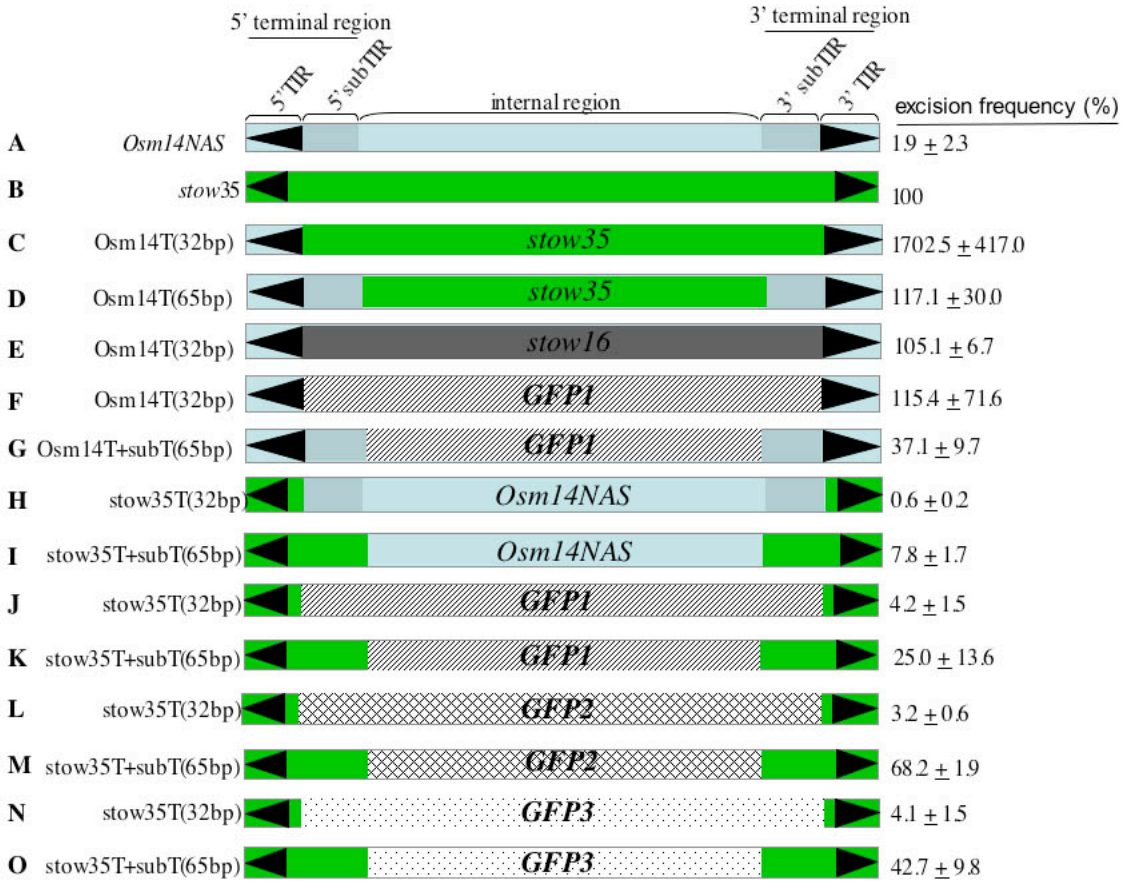

construct
his3
cyc1 ter
Osmar14 transposase
Pgal1
ori EC
amp
ori EC
CEN6
ARS H4
pOsm14Tp
pRS413 vector:
Osmar14 transposase
ade2
ori EC
CEN4
pWL89A
with different constructs
ARS1
ura3
amp
cotransform into
yeast DG2523
empty pRS413
(control)
plates lacking
histidine and uracil
galactose plates
lacking adenine

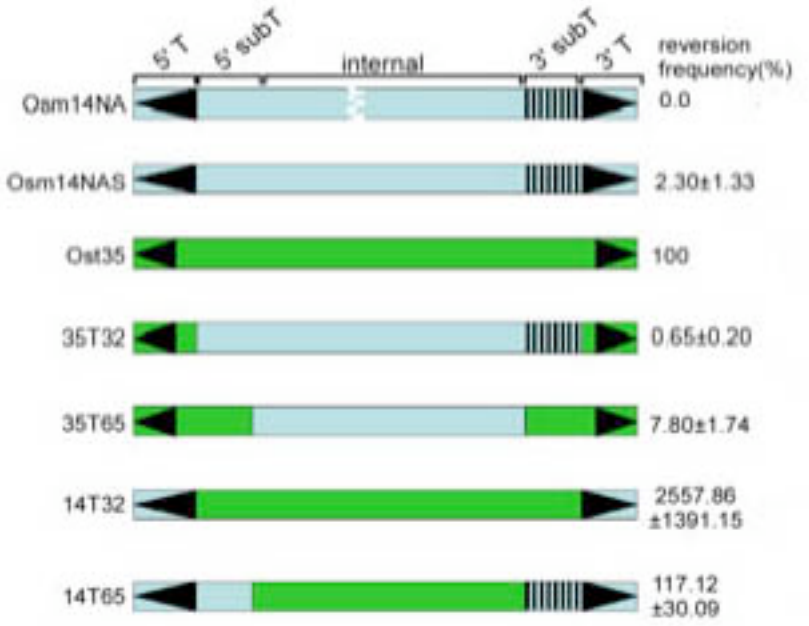
Figure 4.3: Chimeric constructs used in this study.

The 5' and 3' TIRs (T) represent the 32 bp terminal inverted repeat sequences for each construct. The 5' and 3' subTIRs (subT) contain the 33bp sequences immediately following the 32bp TIRs. The terminal region represents the combined 65bp TIRs and subTIRs sequences. The internal region represents all sequences between the 5' and 3' terminal regions. *Osm14NAS*, a 240bp deleted version of the full-length *Osmar*14 element; *stow35* and *stow16* are two rice *stowaway* elements. GFP1, GFP2 and GFP3 are three different random sequences (see Material and Methods for details). The name and a unique letter for each chimeric construct is also listed on the left with the length of sequence in base pairs (bp) from the TIRs represented in brackets. Excision frequencies are shown to the right of each construct.


5' terminal region
3' terminal region
5'TIR
5' subTIR
internal region
3' subTIR
3' TIR
excision frequency (%)
A Osm14NAS 1.9 ± 2.3
B stow35 100
C Osm14T(32bp) stow35 1702.5 ± 417.0
D Osm14T(65bp) stow35 117.1 ± 30.0
E Osm14T(32bp) stow16 105.1 ± 6.7
F Osm14T(32bp) GFP1 115.4 ± 71.6
G Osm14T+subT(65bp) GFP1 37.1 ± 9.7
H stow35T(32bp) Osm14NAS 0.6 ± 0.2
I stow35T+subT(65bp) Osm14NAS 7.8 ± 1.7
J stow35T(32bp) GFP1 4.2 ± 1.5
K stow35T+subT(65bp) GFP1 25.0 ± 13.6
L stow35T(32bp) GFP2 3.2 ± 0.6
M stow35T+subT(65bp) GFP2 68.2 ± 1.9
N stow35T(32bp) GFP3 4.1 ± 1.5
O stow35T+subT(65bp) GFP3 42.7 ± 9.8

Supplemental Figure 4.1. Region swapping assay.

ADE2 reversion frequency (excision frequency) is shown to the right of each construct. ADE2 revertant selection plates are shown for *Osm14NA*, *Osm14NAS*, *Ost*35 (*stow*35) and 14T32. 35T32, terminal 32 bp of *Osm14NAS* replaced by that of *stow*35; 35T65, terminal 65 bp of *Osm14NAS* replaced by that of *stowt*35; 14T32, terminal 32 bp of *stow*35 replaced by that of *Osm14NAS*; 14T65, terminal 32 bp of *stow*35 replaced by that of *Osm14NAS*; Tpase, *Osmar*14 transposase source provided. This analysis was performed by G.Yang and permission was granted to use in this Chapter.


5' T
5' subT
internal
3' subT
3' T
reversion frequency(%)
Osm14NA
0.0
Osm14NAS
2.30±1.33
Ost35
100
35T32
0.65±0.20
35T65
7.80±1.74
14T32
2557.86 ±1391.15
14T65
117.12 ±30.09

References

BUREAU, T. E., and S. R. WESSLER, 1992 *Tourist*: a large family of inverted-repeat element frquently associated with maize genes. Plant Cell **4:** 1283-1294.

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

CHATTERJEE, S., and P. STARLINGER, 1995 The role of subterminal sites of transposable element Ds of *Zea mays* in excision. Mol. Gen. Genet. **249:** 281-288.

CRAIG, N. L., R. CRAIGIE, M. GELLERT and A. M. LAMBOWITZ, 2002 *Mobile DNA II*. American Society for Microbiology Press, Washington, D.C.

FESCHOTTE, C., N. JIANG and S. R. WESSLER, 2002a Plant transposable elements: where genetics meets genomics. Nat Rev Genet **3:** 329-341.

FESCHOTTE, C., M. T. OSTERLUND, R. PEELER and S. R. WESSLER, 2005 DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. Nucleic Acids Res. **33:** 2153-2165.

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.

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.

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., C. FESCHOTTE, X. Y. ZHANG and S. R. WESSLER, 2004 Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Current Opinion in Plant Biology **7:** 115-119.

JURKA, J., and V. V. KAPITONOV, 2001 PIFs meet Tourists and Harbingers: A superfamily reunion. Proc Natl Acad Sci U S A **98:** 12315-12316.

NENE, V., J. R. WORTMAN, D. LAWSON, B. HAAS and E. AL., 2007 Genome sequence of Aedes aegypti, a major arbovirus vector. Science **316:** 1703-1704.

OKI, N., K. YANO, Y. OKUMOTO, T. TSUKIYAMA, M. TERAISHI *et al.*, 2008 A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica. Genes Genet Syst **83:** 321-329.

YANG, G., C. F. WEIL and S. R. WESSLER, 2006 A rice Tc1/Mariner-like element transposes in yeast. Plant Cell **18:** 2469-2478.

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., N. JIANG, C. FESCHOTTE and S. R. WESSLER, 2004 Distribution and evolution of *PIF*- and *Pong*-like transposons and their relationships with *Tourist*-like MITEs. Genetics **166:** 971-986.

CHAPTER 5

CONCLUSIONS

In conclusion, the research described in this dissertation involves a combined computational and experimental approach to study transposable elements (TEs) in plant genomes.

TEs in Lotus. As mentioned in the Introduction and Literature review section, 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 less than 10% of the *Lotus* genome allows us to approximate its TE landscape. Furthermore, the availability of complete BAC sequences covering ~25% of *M. 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 (Arabidopsis).* As an example, when compared with the TE content in *Arabidopsis*, the most abundant class 1 (*Ty1/copia*-like) and class 2 (MULEs) TE type in the *Lotus* dataset appears to be similar to what was observed for *Arabidopsis* (ZHANG and WESSLER 2004). More importantly, the high quality of *the Lotus* 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 (Figure 2.1, 2.9). Experimental strategies

based on the use of PCR primers developed from full-length *Lotus* elements allowed us to test for activity of recently amplified elements (Chapter 3). Furthermore, experimental analyses on other recently amplified elements in *Lotus* promise to enrich our understanding of TEs and how they interact with host genomes.

MITE Amplification. The research described in Chapter 4 was designed to address questions surrounding the amplification mechanism of MITEs. The data obtained from this study suggests that rice MITEs contain special internal features that might provide an advantage for their transposition compared to other noncoding DNA transposons. This was nicely demonstrated for two different families of *Stowaway*-like MITEs in this study [Figure 4.3 (C), (D), (E)]. In addition, this can be further supported by the fact that besides the most terminal 10 bp sequence that is shared between the rice *Stowaway*-like MITEs and full-length *Tc1/Mariners*, the internal sequence shares no similarity. Since this study was done in rice for particular *Stowaway*-like MITE families, it will be interesting to analyze other MITE families belonging to both groups of MITEs and their related transposase sources to determine if this observation can generalized.

Another intriguing question about MITEs beside their amplification mechanism is their origin. In this study we created several artificial *Stowaway*-like MITEs by combining random internal sequences with *Osm14NAS* and *Ost*35 terminals (Figure 4.3). This suggests that even with suboptimal ends, very little sequences are needed to generate an artificial MITE with an excision frequency

that is almost as good as the endogenous MITE in this assay. Therefore, a model is supported that some MITEs might be derived from random sequences in the genome. However, further experimental analysis with other MITE families in different organisms is required to confirm the significance and functional role of MITE internal regions, because this will add to our understanding of how MITEs originated.

In conclusion, the results described in both Chapters of this dissertation have not only added to our understanding of TE abundance, diversity, and amplification in plants, but have also generated novel and interesting findings that can be investigated in future studies.

References

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 U S A **101:** 5589-5594.

APPENDIX A

ANALYSIS OF THE FIRST ACTIVE *TC1/MARINER*-LIKE ELEMENT IN PLANTS

Introduction

The *Tc1/Mariner* superfamily consists of a diverse group of elements that have been found in most eukaryotes studied to date (FESCHOTTE and WESSLER 2002; PLASTERK and VAN LUENEN 2002). *Tc1/Mariner*-like elements contain terminal inverted repeats (TIR) flanked by a conserved target duplication of TA. The internal coding sequence contains a single ORF, which encodes the transposase protein, required for transposition (CRAIG *et al.* 2002). This transposase contain a DDE or DDD motif that is typically found in most eukaryotic transposases and integrases. The *Tc1/Mariner*-like superfamily is usually present in low copy numbers (< 100 copies/genome) and members of this superfamily have been shown to transpose in diverse organisms. To date, seven members have been shown to be active under normal conditions, *Tc1* and *Tc3* from *C. elegans*, *Himar*, *Minos*, and *Mos1*, from *Drosophila (D.mauritiana)*, and *Impala* and *Fot1* from a fungus (*F. oxysporum*) (EMMONS *et al.* 1983; BRYAN *et al.* 1987; COLLINS *et al.* 1989; DABOUSSI *et al.* 1992; FRANZ *et al.* 1994; LANGIN *et al.* 1995; ROBERTSON and LAMPE 1995). Furthermore, in vertebrates, two transposases, *Sleeping Beauty* from fish and *Frog Prince* from frog, were reconstructed from inactive copies and shown to transpose not only in vertebrates but also in primates (IVICS *et al.* 1997; MISKEY *et al.* 2003).

However, though *Tc1/Mariner*-like elements are also widespread in flowering plants, no active elements in plants have been reported to date. In rice, the *Tc1/Mariner*-like family comprise ~34 elements of which seven *Osmar1A, Osmar5A, Osmar5Bi, Osmar9A*, *Osmar15Bi, Osmar17A,* and *Osmar19*) appear to encode intact transposases (no stop codons) (FESCHOTTE *et al.* 2003). Of these elements, *Osmar5* appears to be the most recently amplified element for which three full-length copies are 99.5% identical. In a previous study, the rice *Osmar5* transposase was shown to mobilize a deletion derivative of itself in the budding yeast *Sacchromyces cerevisiae*, suggesting that the transposition of plant *Tc1/Mariner*-like elements could occur without host-specific factors (YANG *et al.* 2006).

Therefore, in this study, we utilized a previously described transposition assay in an attempt to mimic the amplification of an artificially created deleted copy derived from the rice *Tc1/Mariner*-like element, *Osmar5* in *Arabidopsis* (YANG *et al.* 2006; YANG *et al.* 2007). Results showed that when a T-DNA containing both the transposase source and *Osm5NAS*-gfp reporter were transformed into *Arabidopsis*, GFP sectors were observed, indicating transposition of the *Osm5NAS* element. A PCR analysis revealed excision products of *Osm5NAS* and subsequent sequencing showed the footprints of *Osm5NAS*. The insertion sites of *Osm5NAS* were confirmed to be exclusively at TA dinucleotides and unlinked. This is the first report of transposition activity of a *Tc1/Mariner*-like element in plants.

Results and Discussion

Transposition of *Osmar5* in *Arabidopsis*. In plants, there has been no report of an active *Tc1/Mariner*-like element to date. Although the rice *Osmar5* element has not been shown to be active in rice, it is active in yeast, suggesting that it might also be found to be active if transformed into *Arabidopsis*. Therefore, in this study, experiments were designed to test whether the rice *Osmar5* can transpose in a heterologous plant host. Constructs were designed based on a T-DNA-containing vector which includes a selectable marker and a reporter gene composed of the green fluorescent protein (gfp) coding region fused to a 35S promoter (pBin-mgfp5-er) (see Materials and Methods for details). Two constructs, one containing only *Osm5NAS* (pBin-Osm5NAS) and another construct containing *Osm5NAS* with *Osmar5* transposase, were both transformed into *Arabidopsis* (Figure 2A). Plants containing GFP sectors were observed only in the transformants with the *Osmar5* transposase, indicating that *Osm5NAS* had indeed excised in response to the transgenic transposase source (Figure 2B). To confirm that the GFP expression was due to excision of *Osm5NAS* from the reporter, a PCR analysis on genomic DNA from one control seedling (*Osm5NAS* reporter construct without transpose source) and five GFP expressing seedlings was performed with primers flanking

Figure 1: Structure of *Osmar5, Osm5NA and Osm5NAS*.

The predicted transposase ORF is shown as a dark grey box in *Osmar5*. The black triangles represent TIRs flanked by TA target site duplications. Numbers in brackets represents the length of *Osm5NA* and *Osm5NAS*. For construct details, see the Material and Methods section.

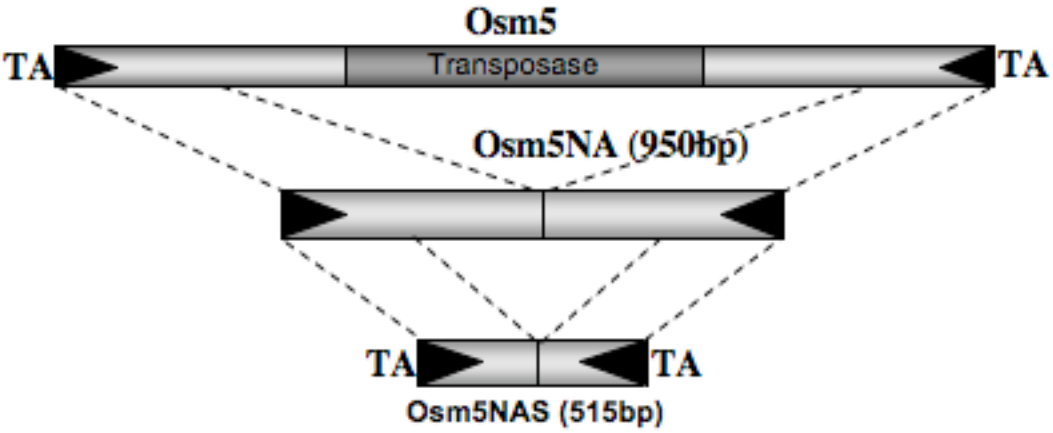
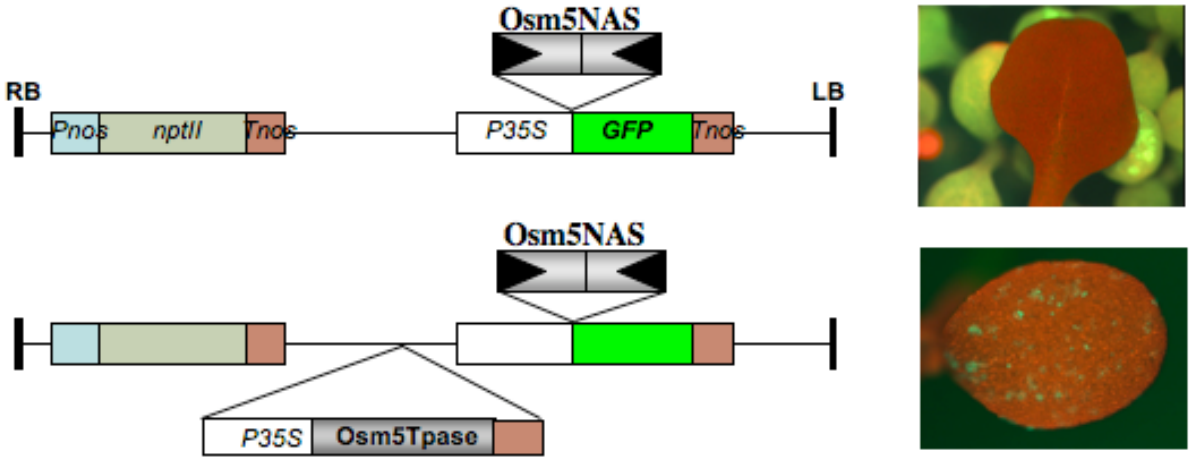

Osm5
TA
Transposase
TA
Osm5NA (950bp)
TA
TA
Osm5NAS (515bp)

Figure 2: T-DNA constructs transformed into Arabidopsis.

A) Constructs depicting *Osm5NAS* and *Osmar*5 transposase (Osm5Tpase). See Figure 1 and Material and Methods for additional construct details. B) To the right of each construct are representative images of transformants with red fluorescence from chlorophyll and green fluorescence from GFP expression. GFP, green fluorescence protein; npt II, neomycin phosphate transferase; RB and LB, right and left borders of T-DNA; Pnos, promoter of nopaline synthase gene; Tnos, terminator of nopaline synthase gene; P35S, promoter of CaMV 35S gene.


Osm5NAS
RB
LB
Pnos
nptII
Tnos
P35S
GFP
Tnos
Osm5NAS
P35S
Osm5Tpase

Osm5NAS insertion in the reporter construct (Figure 3A). PCR products that were consistent with the excision of *Osm5NAS* were obtained from the primary transformants. For the control, the PCR product showed a single band representing the presence (no excision) of *Osm5NAS* in the GFP reporter. The PCR products obtained from the GFP expressing seedlings showed two bands, an upper band representing no excision of *Osm5NAS* in some cells and a lower band indicating the somatic excision of *Osm5NAS* (Figure 3B).

Subsequent sequencing of the lower band to further confirm that excision had in fact occurred revealed the footprints of *Osm5NAS*, which varied between 4 and 12 base pairs (Figure 3C). Five of the nine sequences had footprints with the TA duplication intact and along with additional nucleotides that appeared to be derived from *Osm5NAS* ends. Furthermore, none of these excision events (footprints) showed precise excision (removal of the entire element leaving one copy of the TA target site duplication) consistent with *Osm5NA* transposition in yeast and in contrast to what has been observed for some other DNA elements (C. Nathan Hancock, manuscript in preparation).

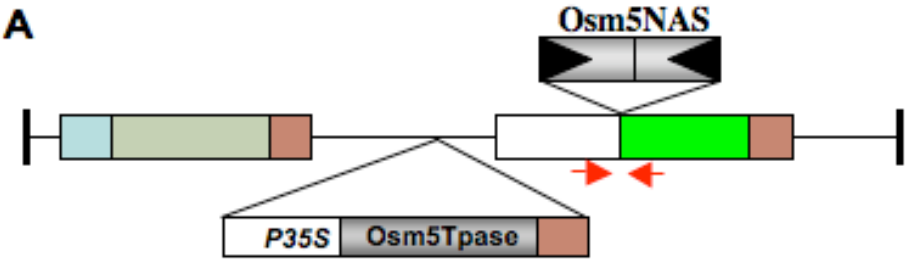
Osm5NAS insertion sites.

Because local transposition has been demonstrated for other *Tc1/Mariner*-like elements such as *Sleeping Beauty* and *Osmar5* transposition in yeast, but showed no obvious preference for local transposition, an analysis was performed to examine the insertion sites for *Osm5NAS* in the *Arabidopsis* genome (IVICS *et al.* 1997; YANG *et al.* 2006).

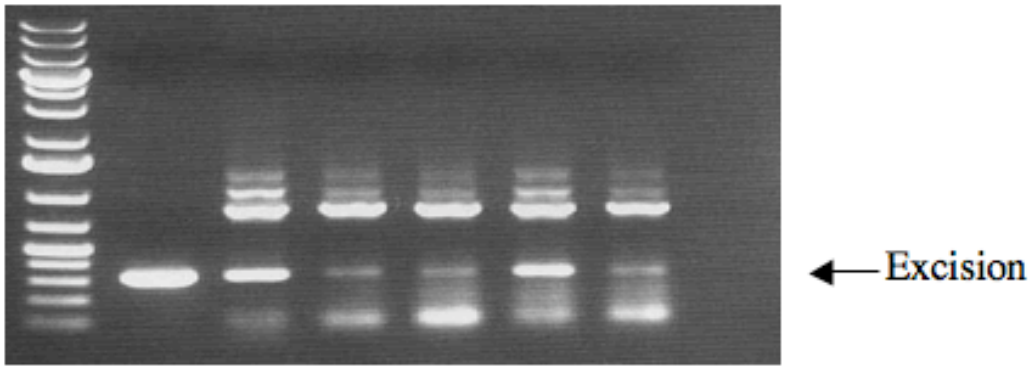
Figure 3: Analysis of *Osm5NAS* excision in *Arabidopsis*.

A) Constructs used as described in Figures 1 and 2. The red arrowheads indicate PCR primer locations for *Osm5NAS* excision analysis. (B) Agarose gel of PCR products in transgenic seedlings. Five independent transformants are shown. L represents the size marker, a 1-kb ladder. E: empty plasmid/vector; Lane 1 - 5: DNA from T0 plants containing both *Osm5NAS* and Osm5Tpase. Black arrow indicates excision products. C) Sequences of *Osm5NAS* donor sites after excision events. The sequence before excision of *Osm5NAS* is shown at the top along with the 2-bp (TA) TSD sequence (in red) and the nine remaining sequences show the excision sequences recovered for Lane 1 - 5.

A



B



C

```
CACGGGGGACTCTAGAGGATCCTACTCC...GGAGTAGGATCCAAGG

CACGGGGGACTCTAGAGGATCCTACTC.......GTAGGATCCAAGG
CACGGGGGACTCTAGAGGATCCTACT........GTAGGATCCAAGG
CACGGGGGACTCTAGAGGATCCTACT.........TAGGATCCAAGG
CACGGGGGACTCTAGAGGATCCTAC........AGTAGGATCCAAGG
CACGGGGGACTCTAGAGGATCCTAC........AGTAGGATCCAAGG
CACGGGGGACTCTAGAGGATCCTA............AGGATCCAAGG
CACGGGGGACTCTAGA..............GGAGTAGGATCCAAGG
CACGGGGGACTCTAGAG......AACACGGGGGAGTAGGATCCAAGG
CACGGGGGACTCTAGA.....................GGATCCAAGG
```

To determine the target site for new *Osm5NAS* insertions, transposon display analysis was performed and eight new insertions were detected and shown to be flanked exclusively by TA dinucleotides (Figure 4A and data not shown). In addition, 3 of 8 insertions were preferentially located in linked sites (inserted in other regions of the vector) and the remaining 5 were inserted in different locations in the *Arabidopsis* genome, similar to what was observed for *Osmar5* in yeast (Figure 4B).

Continued activity of *Osm5NAS* in *Arabidopsis*. In plants, transposition of TEs in the host genome is highly regulated by the host defense mechanism (DNA methylation) (FESCHOTTE and PRITHAM 2007; ZHANG 2008). This regulation has also been observed in plants after introduction of a TE into a heterologous host genome (e.g., *Tos17*) (HIROCHIKA *et al.* 2000). Therefore, to determine whether *Osmar5NAS* is active throughout the development of the plant and in consecutive generations, transgenic seedlings were screened at two weeks and again during floral bud development for excision events. GFP sectors were observed in both two-week-old seedlings and floral buds. Large sectors representing somatic excision events that occurred early in plant development were also observed in floral buds (Figure 5). This observation was consistent for three consecutive generations indicating that *Osm5NAS* was excising at later developmental stages in several generations (Figure 5). The activity of *Osm5NAS* in subsequence generations was screened by transposon display analysis to detect new insertions. In all three generations, new insertions of

Osm5NAS were observed, confirming that *Osm5NAS* activity (insertion and excision) is maintained for at least three generations (Figure 6).

In conclusion, as mentioned in the Introduction, *Tc1/Mariner*-like elements are active in several animal genomes and have been well studied mechanistically. In a few cases, these elements have been adopted for use as efficient tagging tools. Here, we demonstrate for the first time an active *Tc1/Mariner*-like element in any plant species. The availability of an active *Tc1/Mariner-like* element in plants provides a unique opportunity to study and compare the transposition mechanism of plant *Tc1/Mariner*-like elements with their animal relatives.

Materials and Methods

Plasmid construction. A 515bp deleted version of the rice *Osmar5* was shortened from the 950bp element on pOsm5NA (YANG *et al.* 2006).The 515bp product (*Osm5NAS*) was then cloned between the *Bam*H1 and *Xba*I site in the GFP gene of pBin-mgfp5-er to obtain pBin-*Osm5NAS*. Osm5Tpase was PCR amplified from a plasmid containing *Osmar5* transposase using *pfu* DNA polymerase (Stratagene, La Jolla, CA) and cloned into the binary vector pBin-mgfp5-er between *BamHI* and *SacI*. The fragment containing Osm5Tpase flanked by P35S and *Tnos* was then amplified using *pfu* DNA polymerase, and cloned between the *SbfI* and *Hin*dIII site of pBin-Osm5NAS to obtain pBin-Osm5Tpase-Osm5NAS. Primer sequences are available upon request.

Figure 4: Insertion sites of transposed *Osm5NAS* in the *A. thaliana* genome.

(A) Autoradiograph of a transposon display gel of DNA from primary transformants. Samples are from those shown in Figure 3, Lane 1 – 4. W: wild type *Arabidopsis* DNA; C: control DNA from transformants with Osm5NAS only; Lane 1- 4: DNA from transformants containing both *Osm5NAS* and Osm5Tpase. Red stars represent somatic insertions recovered and sequenced and the blue star represents the T-DNA copy. (B) Distribution of *Osm5NAS* insertions on *Arabidopsis* chromosomes. The black arrowheads indicate somatic *Osm5NAS* insertions obtained from the analysis in section A.

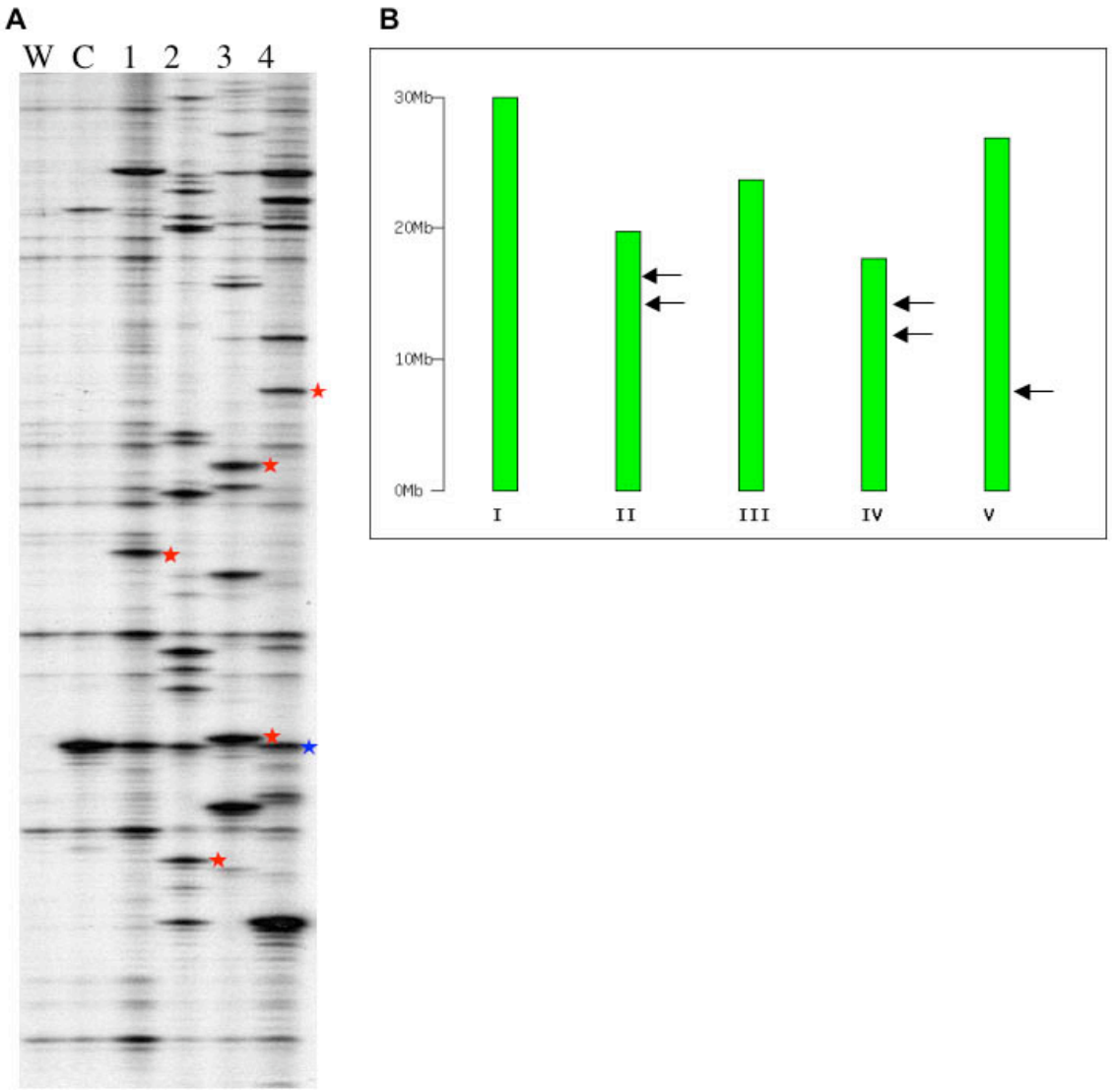
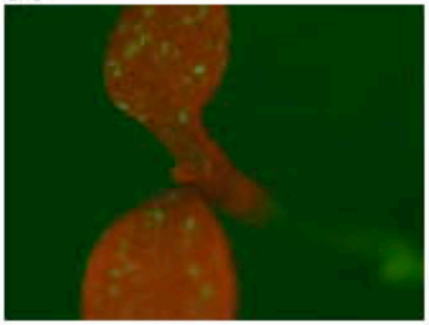

A
W C 1 2 3 4
B
30Mb
20Mb
10Mb
0Mb
I
II
III
IV
V

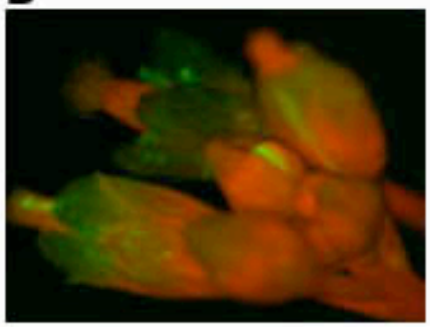
Figure 5: Analysis of *Osm5NAS* excision in three subsequent generations and late developmental stages.

The dark red seedlings with true leaves are transgenic with red fluorescence from chlorophyll and green fluorescence from GFP expression. 1) T2 generation seedlings were derived from T1 transformants (primary transformants) after two weeks (A) and after 4 weeks (B). 2) T3 generation seedlings derived from T2 transformants after two weeks (C) and after 4 weeks (D). 3) T4 generation seedlings derived from T3 transformants after three weeks (E) and after 4 weeks (F).

A



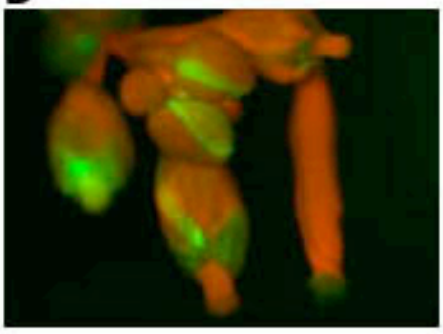
B



1. T2 generation

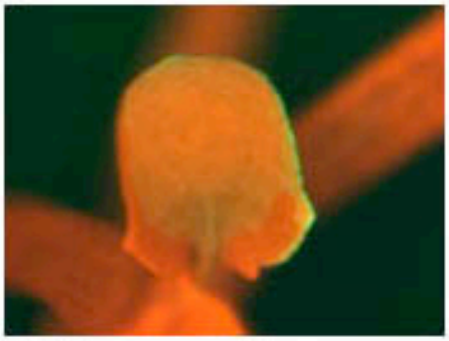
C

D

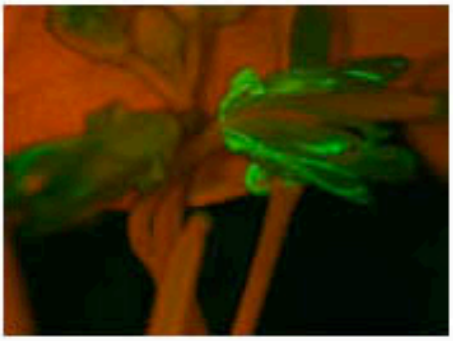


2. T3 generation

E



F



3. T4 generation

Figure 6: Insertion analysis of rice *Osm5NAS* in three generations *of* *Arabidopsis* transgenic plants.

Transposon display was performed as described in Figure 4. DNA from two-week-old seedlings was used from T2, T3 and T4 generation plants described in Figure 5. P represents plasmid control containing *Osm5NAS* insertion.


P
T2
T3
T4

***Arabidopsis* transformation and selection.** The *Arabidopsis* ecotype Columbia was transformed with *Agrobacterium tumorfacience* strain GV3103 containing the constructed binary vectors (BECHTOLD *et al.* 1993). The seeds of transformed plants were collected and germinated on Murashige and Skoog solid medium (with 0.2% Phytagel) containing 150mg/liter Timentin and 50mg/liter kanamycin. After ~7 – 10 days of incubation at 26 degrees Celsius and 16:8 day/night cycle, transformants were detectable. Transgenic plants were observed and imaged using a fluorescence stereoscope Zeiss SteREO Discovery.V12.

Analysis of excision and insertion sites. Genomic DNA was extracted from leaves expressing GFP spots and PCR was performed using the following primers: 5'-agacgttccaaccacgtcttcaaagcaag-3' and 5'-cctctccactgac-agaaaatttgtgccca-3' to confirm excision of *Osm5NAS*. The PCR products were then cloned into TOPO vector (Invitrogen, Carlsbad, CA) and sequenced to recover the excision sites.

Transposon Display. To determine the insertion sites, transposon display was performed with the following primers: *Bfa1+0*, 5'-gacgatgagtcctgagtag-3'; P1, 5'-gtacaaatgctgtaaat-gacagc-3'; and P2, 5'-ggacaatccaggggcggtg-3', using genomic DNA from 5 transgenic seedlings. Unique bands observed between seedlings on the transposon display gel were excised, reamplified, cloned and sequenced (CASA *et al.* 2000). Locations of the

new insertions were determined by mapping the flanking sequences using a BLAST search against the annotated *Arabidopsis* genome database at Gramene (www.gramene.org; TAIR, version 6). All sequencing was done by the Molecular Genetics Instrumentation Facility (University of Georgia).

Acknowledgements

The constructs and plasmids used in this analysis were generated by Guojun Yang.

References

BECHTOLD, N., J. ELLIS and G. PELLETIER, 1993 In planta Agrobacterium-mediated gene transfer by infiltration of adult Arabidopsis thaliana plants. . C.R. Acad. Sci. **316:** 1194-1199.

BRYAN, G. J., J. W. JACOBSON and D. L. HARTL, 1987 Heritable somatic excision of a *Drosophila* transposon. Science **235:** 1636-1638.

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 U S A **97:** 10083-10089.

COLLINS, J., E. FORBES and P. ANDERSON, 1989 The Tc3 family of transposable genetic elements in Caenorhabditis elegans. Genetics **121:** 47-55.

CRAIG, N. L., R. CRAIGIE, M. GELLERT and A. M. LAMBOWITZ, 2002 *Mobile DNA II*. American Society for Microbiology Press, Washington, D.C.

DABOUSSI, M.-J., T. LANGIN and Y. BRYGOO, 1992 Fot1, a new family of fungal transposable elements. Mol. Gen. Genet. **232:** 12-16.

EMMONS, S. W., L. YESNER, K. S. RUAN and D. KATZENBERG, 1983 Evidence for a transposon in Caenorhabditis elegans. Cell **32:** 55-65.

FESCHOTTE, C., and E. J. PRITHAM, 2007 DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. **41:** 331-368.

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.

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.

FRANZ, G., T. G. LOUKERIS, G. DIALEKTAKI, C. R. THOMPSON and C. SAVAKIS, 1994 Mobile *Minos* elements from *Drosophila hydei* encode a two-exon transposase with similarity to the paired DNA-binding domain. Proc. Natl. Acad. Sci. USA **91:** 4746-4750.

HIROCHIKA, H., H. OKAMOTO and T. KAKUTANI, 2000 Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. Plant Cell **12:** 357-369.

IVICS, Z., P. B. HACKETT, R. H. PLASTERK and Z. IZSVAK, 1997 Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell **91:** 501-510.

LANGIN, T., P. CAPY and M. J. DABOUSSI, 1995 The transposable element impala, a fungal member of the Tc1-mariner superfamily. Mol. Gen. Genet **246:** 19-28.

MISKEY, C., Z. IZSVAK, R. H. PLASTERK and Z. IVICS, 2003 The Frog Prince: A reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. Nucleic Acids Res. **31:** 6873-6881.

PLASTERK, R. H. A., and H. G. VAN LUENEN, 2002 The Tc1/*mariner* family of transposable elements, pp. 519-532 in *Mobile DNA II*, edited by N. L. CRAIG, R. CRAIGIE, M. GELLERT and A. M. LAMBOWITZ. American Society for Microbiology Press, Washington D.C.

ROBERTSON, H. M., and D. J. LAMPE, 1995 Distribution of transposable elements in arthropods. Annu Rev Entomol **40:** 333-357.

YANG, G., C. F. WEIL and S. R. WESSLER, 2006 A rice Tc1/Mariner-like element transposes in yeast. Plant Cell **18:** 2469-2478.

YANG, G., F. ZHANG, C. N. HANCOCK and S. R. WESSLER, 2007 Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana. Proc Natl Acad Sci U S A. **104:** 10962-10967.

ZHANG, X., 2008 The Epigenetic Landscape of Plants. Science **320:** 489-492.