SEQUENCE ACQUISITION SPECIFICITY AND EVOLUTION OF TERMINAL SEQUENCES IN PLANT *MUTATOR*-LIKE ELEMENTS AND THE REPETITIVE SEQUENCE LANDSCAPE OF SACRED LOTUS (*Nelumbo nucifera*)

By

Ann Roselle Armenia Ferguson

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Breeding, Genetics and Biotechnology – Horticulture – Doctor of Philosophy

2014

ABSTRACT

SEQUENCE ACQUISITION SPECIFICITY AND EVOLUTION OF TERMINAL SEQUENCES IN PLANT *MUTATOR*-LIKE ELEMENTS AND THE REPETITIVE SEQUENCE LANDSCAPE OF SACRED LOTUS (*Nelumbo nucifera*)

By

Ann Roselle Armenia Ferguson

Transposable elements (TE) are an integral part of eukaryotic genomes; therefore, their identification, characterization and analysis remain critical in genetic and evolutionary studies. The *Mutator* superfamily (MULEs) are DNA TEs characterized by long terminal inverted repeats (TIR) and includes a special group of elements, called Pack-MULEs, that carry genes/gene fragments. This study aims to understand the role of terminal sequences in transposition of MULEs and the factors involved in sequence acquisition by Pack-MULEs. In addition, the repetitive content and diversity of an ancient eudicot genome was characterized to better understand the evolutionary role of MULEs as well as other TEs in angiosperms.

Analyses of MULEs in rice show that these elements also capture GC-rich intergenic sequences but at a much lower frequency than genes. The TIR-MULE type is predominantly involved in sequence acquisition and these elements are associated with long GC-rich TIRs which may be important in acquisition. Genes with known functions and genes with orthologs are overrepresented among parental genes of Pack-MULEs in rice, maize and Arabidopsis suggesting that Pack-MULE preferentially duplicate bona fide genes. Pack-MULEs selectively acquire and retain parental sequences through a combined effect of GC content and breadth of expression, with GC content playing a stronger role. Analysis of MULEs in maize, tomato, rice and Arabidopsis detected the formation of atypical MULEs and Pack-MULEs with multiple TIRs mostly located in tandem. The copy number of these atypical MULEs suggests their

significant mobility while their tandem TIR sequences indicate sequence conservation. The successful amplification of the Pack-MULE, PM-ZIBP, demonstrates that MULEs with tandem TIRs are functional in transposition and duplication of gene sequences.

Characterization of the repetitive sequence of sacred lotus (*Nelumbo nucifera*) shows that 50% of the genome is composed of recognizable transposable elements (TE). TE content and diversity show a comparable *Copia* and *Gypsy* LTR content, which is atypical among plants. Non-canonical LTR types comprise 15.6% of the total LTR content suggesting the need to consider other end types in annotation of LTR elements. Sacred lotus also contains the highest coverage and copy number of *hAT* elements among all sequenced genomes to date. The 1447 Pack-MULEs in the genome provide the first evidence for the GC acquisition preference by Pack-MULEs outside the grasses.

ACKNOWLEDGEMENTS

I extend special appreciation to my adviser, Dr. Ning Jiang, for her patience, mentoring, and for giving me an opportunity to learn the skills that I do now and to appreciate and be constantly in awe of those sequences in the genome many others consider as "junk". I would like to thank all my committee members, Drs. James Hancock, Douglas Schemske and Shin-han Shiu for their support and guidance.

Big thanks go to the past and current Jiang lab members (Veronica, Dongyan, Guozhu, Dongying, Dongmei and Stefan) for all the help in the lab and many wonderful discussions. Thanks to all the students, post-docs and faculty of "joint lab" and "rice group" for the insightful comments and discussions throughout the years, especially Robin Buell and Kevin Childs. I also thank the National Science Foundation and the PBGB program who have supported my research in the Jiang Lab.

I would like to show gratitude to my wonderful family, from thousands of miles away, for the support and prayers throughout this long journey. I love you all!

Thank you to my husband, Brett, for all the love, support and encouragement. I could not have done this all without you and Elliot!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

Review of Literature

Transposable Elements

Transposable elements (TEs) are genetic sequences that are capable of moving from one genomic location into another and in the process can increase their copy numbers. These genetic elements were first discovered by Barbara McClintock over 50 years ago which she referred to as "controlling elements" responsible for the altered pigmentation in mutant maize kernels (McClintock, 1951). The continued sequencing of many eukaryotic genomes has demonstrated the extensive contribution of TEs in the genomic composition of diverse plant and animal species (Adams, 2000; Lander et al., 2001; Sequencing Project, 2005; Schnable et al., 2009; Howe et al., 2013; Middleton et al., 2013).

According to the intermediate form of transposition used by the specific element, TEs are generally classified into two major groups: Class I and Class II (Figure 1.1; Feschotte et al., 2002; Wicker et al., 2007; Kapitonov and Jurka, 2008). In addition, the coding capacity of elements for proteins involved or comprising the transpositional machinery allows for further classification of elements into autonomous elements, which code for these proteins, or non-autonomous elements, which rely on the autonomous elements for movement within the genome (Figure 1.1). Furthermore, TEs may be classified into superfamilies and families, primarily based on specific sequence and/or structural features as well as replication strategy shared by the elements within the superfamily (Wicker et al., 2007). These features, among others, include the target site duplication (TSD) that is created at the flanking regions following transposition into a new genomic location and terminal repeats that specifically vary in size among different families.

Class I TEs. Elements belonging to Class I are also referred to as RNA elements or retrotransposons, and transpose via an RNA intermediate using a copy-and-paste mechanism.

The elements are transcribed forming an RNA intermediate which is then reverse-transcribed by a reverse transcriptase encoded by autonomous members. The DNA copy is then integrated into a new genomic location (Flavell et al., 1994; Wicker et al., 2007). Retrotransposons account for the majority of repetitive sequences in many plant genomes and have been implicated for genome size expansion (Bennetzen and Kellogg, 1997; Kalendar et al., 2000; Bennetzen, 2005; Schulman and Kalendar, 2005).

The presence or absence of long terminal repeats (LTRs) among Class I elements groups them into LTR retrotransposons and non-LTR retrotransposons (Figure 1.1; Feschotte et al., 2002; Kapitonov and Jurka, 2008). LTR retrotransposons are the major components of the TE fraction in plants while non-LTR retrotransposons are more predominant in animal genomes. Moreover, LTR retrotransposon insertions target specific sites in the genome (Bushman, 2003; Ammiraju et al., 2007; Linheiro and Bergman, 2012; Tsukahara et al., 2012). For instance, different LTR elements in yeast (*Saccharomyces cerevisiae*) exhibit variable targeting. *Ty3* elements target locations close to the RNA PolIII transcription initiation sites (Chalker and Sandmeyer, 1992), while *Ty5* preferentially integrates into heterochromatin of telomeres and silent mating locus HMR (Zhu, 2003).

Class II TEs. Class II elements, or DNA elements, transpose via a cut-and-paste mechanism. The element itself excises from its original genomic location and moves into a new target location. DNA elements may increase their copy numbers by transposing during DNA replication into a locus prior to the formation of the replication fork and by exploiting the gap repair process which creates a new copy and restores the original copy through repair involving the copy in the sister chromatid (Nassif et al., 1994). An exception to the typical features found

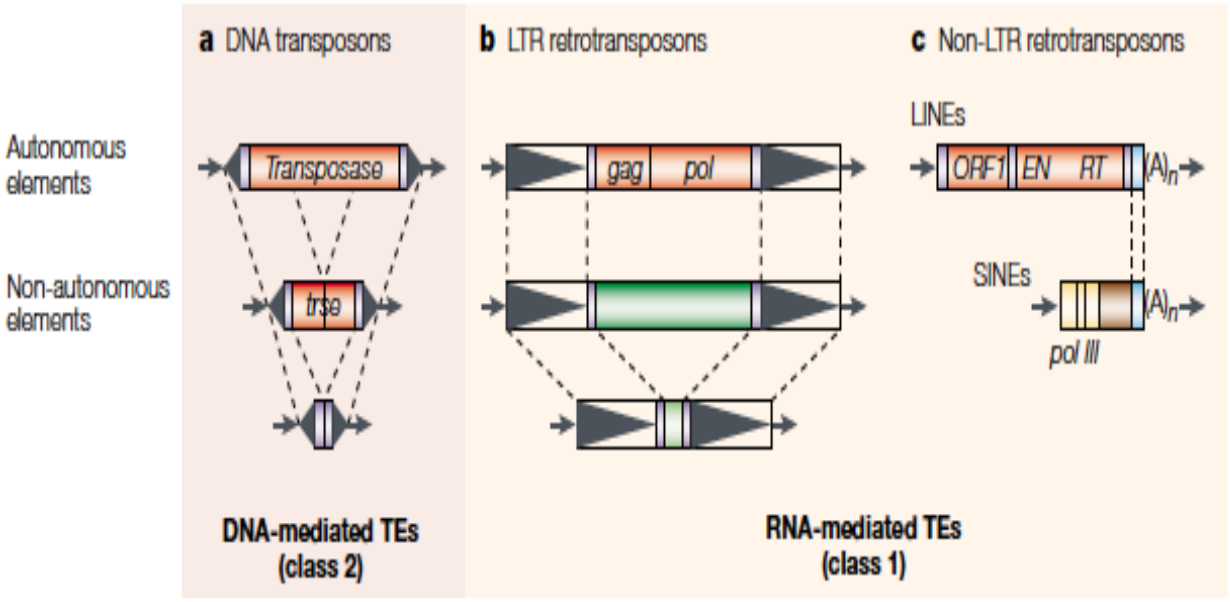


Figure 1.1 Classes of transposable elements (Feschotte et al., 2002).

in Class II elements are *Helitrons*, which is hypothesized to transpose via a rolling-circle mechanism (Kapitonov and Jurka, 2007).

An integral part of many DNA element families are the end sequences which are called Terminal Inverted Repeats (TIR) due to their reverse complementary nature. Some DNA TE superfamilies are characterized by short TIRs (<50bp) such as *hATs* and CACTA (Wicker et al., 2007). Meanwhile other superfamilies including *Mutators*, *Tc1/Mariner*, *Polintons* and *Merlin* are typically associated with long TIRs or include members with long TIRs (Li and Shaw, 1993; Lisch, 2002; Feschotte, 2004; Kapitonov, 2006; Ferguson and Jiang, 2012). This variation in TIR length partially allows for the classification of different superfamilies (Wicker et al., 2007; Kapitonov and Jurka, 2008).

The TIR plays an important role during transposition. DNA binding domains in the transposase proteins bind the TIR sequence of related elements, and in certain cases sub-terminal sequence, and this is the first step in transposition (Ichikawa et al., 1987; Becker, 1997; Benito and Walbot, 1997; Zhou et al., 2004; Loot et al., 2006). In *Osmar5*, an active and autonomous *Tc1/Mariner* element, the N-terminal binding domain of the encoded transposase that contains the helix-turn-helix (HTH) motifs which specifically binds to two sequence motifs in the TIR (Feschotte, 2005). In the bacterial *Tn3* element, the binding of transposase to the 38-bp motif sequence in the TIR facilitates the nicking at the end of *Tn3* by DNase I and initializes the transposition process (Ichikawa et al., 1987). Binding of transposase to the TIR and to the target DNA mediates the synapsis of the transposon ends and the target DNA, allowing the insertion of the element into the target sequence (Craig, 2002). In the *Osmar5* scenario, mutations in the conserved motifs in the TIR can prevent transposition (Yang et al., 2006).

Genome size variation. The "C-value" paradox refers to the apparent lack of correlation between an organism's genome size and its biological complexity (Thomas, 1971). A huge variation in genome sizes exists among organisms with the "large" genomes of this spectrum dominated by transposable elements. Due to their capacity to multiply within a host genome and their prevalence among plant and animal genomes, TEs contribute significantly to increases in genome size (Bennetzen and Kellogg, 1997; Ammiraju et al., 2007; Bennetzen, 2007; Zuccolo et al., 2007) and this may explain the differences in genome sizes without an apparent functional correlation. Recent bursts of amplification of TEs was shown to account for a two-fold increase in genome size of wild rice, *Oryza australiensis,* in comparison to the asian domesticated rice *Oryza sativa* (Piegu et al., 2006). In addition, retrotransposon families, primarily *Gorge3*, have been implicated in genome size increases in two cotton (*Gossypium*) lineages (Hawkins et al., 2006). These findings suggest that TE can play a major influence in the genome size expansion of plants.

TEs are ubiquitous among eukaryotic genomes and the Class I/retrotransposon content maintains to be the largest component of repeat content in many plants. Despite this, dramatic differences exist in the content of different TE types between organisms. For instance, some animal and insect genomes contain a higher proportion of non-LTR retrotransposons compared to LTR retrotransposons (Lander et al., 2001; Chinwalla et al., 2002; Nene et al., 2007). Meanwhile, in plants, the LTR retrotransposons coverage typically dominates the TE landscape (Rice Sequencing Project, 2005; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010). Also, the ratio of RNA to DNA elements can vary. On one extreme, the RNA TE coverage in the rice genome is 1.5-fold more abundant than the DNA elements and DNA TEs are more abundant in terms of copy number (Rice Sequencing Project, 2005) while in the opposite

extreme is the papaya genome where 99% of annotated TEs are RNA elements (Ming et al., 2008), suggesting that although plants overall share similarities in TE component, the content and ratios of specific TE families varies between species. The precise epigenetic regulating mechanisms that resulted in this variable success and lack thereof among different TE superfamilies remains to be elucidated.

Regulation of transposable elements. While TE proliferation can contribute to genome expansion, this uncontrolled activity can, in most cases, produce havoc to a genome when TE movement interrupts crucial gene function or regulation (Callinan and Batzer, 2006). Important epigenetic defense mechanisms regulate and silence TE activity, and thus TEs are largely inactive in all plant genomes (Lisch, 2009) and less than 0.05% of TEs in the human genome is able to transpose (Mills et al., 2007).

TE activity is regulated through a combination of post-transcriptional and transcriptional silencing mechanisms (reviewed in Feschotte et al., 2002; Slotkin and Martienssen, 2007). Together, the two silencing strategies can result in a repressed TE activity through DNA methylation, chromatin modifications, small RNA and reduced expression. The RNA interference (RNAi) pathway is critical in the regulation of TE. It can be involved in the generation of small RNAs that directly target the degradation of TE RNAs or through RNA-directed DNA methylation (RdDM), both of which repress expression and, therefore, activity. DNA methylation is accomplished by DNA methyltransferases such as METHYLTRANSFERASE1 (MET1), CHROMOMETHYLASE3 (CMT3) and DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) (Lindroth, 2001; Cao and Jacobsen, 2002; Kinoshita, 2004; Law and Jacobsen, 2010). In Arabidopsis, it was shown that *DDM1*, a *Snf2* nucleosome remodeling factor is important in TE regulation (Hirochika et al., 2000; Singer,

2001). In fact, *ddm1* mutants show transcriptional activation of TEs (Lippman et al., 2004) and increased transposition (Tsukahara et al., 2009). TE regulations through DNA methylation and chromatin changes via *DDM1* may involve the RNA-directed DNA methylation (RdDM), a Pol IV- and PolV-dependent process whereby small RNAs target homologous genomic DNA for cytosine methylation (Lippman et al., 2004; Law and Jacobsen, 2010). Recent study indicates that the synergistic TE silencing activity of both *DDM1* and RdDM regulate nearly all TE in Arabidopsis (Zemach et al., 2013). This control occurs when the chromatin remodeling nature of *DDM1* allows DNA methyltransferase access to otherwise inaccessible heterochromatin while RdDM silences TEs that are typically found in euchromatin. This study is a key step in understanding the variable silencing strategies that might explain the success of certain TE types or superfamilies in some organisms over others.

Domestication of transposable elements. Despite the fact that majority of TE activity is deleterious and effectively regulated by various silencing strategies, some TEs have been implicated in adaptive evolution. TE domestication refers to the process whereby TE sequences evolve as new genes or sequences with beneficial roles in the host genome (Volff, 2006). Studies have shown that transposable elements have contributed not only to the creation of new genes but also to the evolution of regulatory networks (Jordan et al., 2003; Muotri et al., 2007). Studies show that several ancient TEs have been used as noncoding functional elements in vertebrates (Bejerano et al., 2006; Kamal, 2006; Kapusta et al., 2013). Sequencing of the short-tailed opossum (*Monodelphis domestica*) genome has revealed that 20% of TE-derived non-coding elements were conserved in the eutherian lineage suggesting that these elements might be functional (Mikkelsen et al., 2007). Fifty-five human micro RNA (miRNA) genes are apparently

derived from TEs (some L2 and MIR TE family related) and can potentially regulate human genes (Piriyapongsa et al., 2006).

The catalytic core of the RAG1 and RAG2 proteins involved in the V(D)J recombination process is derived from a *Transib* transposase (DNA element) indicating that a very critical process in the jawed vertebrate immune system originally evolved from transposons (Kapitonov and Jurka, 2005). In *Drosophila*, the telomeres are not maintained by telomerase as in the case of most other organisms. In fact, *Drosophila* lacks telomerase and the conventional telomeric repeats; instead their telomeres are maintained by the *Het-A* and *TART* non-LTR retrotransposons that specifically transpososе to the ends of the chromosome (Pardue et al., 2005). In plants, one example of TE domestication is DAYSLEEPER, a transposase similar to the *hAT* superfamily of transposases, which is essential for normal plant growth and development in *Arabidopsis thaliana* (Bundock and Hooykaas, 2005). Furthermore, a *Mariner* type transposase fused to a sequence encoding a SET domain (Metnase or SETMAR) found in primates is implicated in the increased resistance to ionizing radiation and non-homologous end-joining repair of double stranded breaks in DNA (Cordaux, 2006; Shaheen et al., 2010).

Mutator **Superfamily**

The *Mutator* superfamily is one of the most active DNA elements in plants. This superfamily was initially discovered by Don Robertson in 1978 in a maize stock that generated an unusually high frequency of genetically unstable recessive mutations (Robertson, 1978). The first non-autonomous *Mutator* element characterized was *Mu1*, a 1.4 kb sequence inserted in the *Adh1* gene that resulted in an unstable mutant allele (Strommer et al., 1982; Bennetzen et al., 1984). *Mutator* maize stocks may contain between 10-50 copies of this element, whereas most other maize strains are devoid of this element (Alleman, M. and Freeling, M., 1986).

Furthermore, this copy number is maintained with a remarkably high transposition rate of 10-15 transpositions per gamete per generation. This extremely high transposition rate gives the *Mutator* superfamily an unparalleled activity among DNA transposons in plants. Since the original discovery of *Mu1*, other non-autonomous elements were later characterized from the maize genome, and subsequently in other organisms where they are referred to as *Mutator*-like elements (MULEs) (Yu et al., 2000; Chalvet, 2003; Jiang et al., 2004; Holligan et al., 2006; Marquez and Pritham, 2010; Ming et al., 2013).

The contribution of MULEs to genome composition varies among plants. The model plant with a relatively small genome size, *Arabidopsis thaliana*, contains ~1500 MULEs. However, MULE copy number is reported to be over 28,000 and 30,000 in tomato and rice, respectively (Ferguson and Jiang, 2012). In addition, maize with a genome about 6 times the rice genome contains only ~13,000 MULEs (Schnable et al., 2009). This variation in success of MULE amplification among different plants implies differential silencing between different TE types and competition from the activity of other TEs in the genome.

DNA elements belonging to the *Mutator* superfamily are generally characterized and differentiated from other classes of DNA transposons by distinct structural features such as TIR and target site duplication (TSD) (Kapitonov and Jurka, 2008). *Mutator* elements in maize share a ~220bp TIRs found on both ends of the element but the internal sequence between the TIRs varies among sub-families (Chomet et al., 1991). In addition, during transposition, these elements form a 8-11 bp (mostly 9 bp) TSD at the new location, a feature that is generally regarded as a hallmark of transposition activity.

The long TIRs of MULEs may range from 100 to 500 bp and appear to be critical for element transposition and expression. TIRs of actively transposing elements in maize contain a

~32bp conserved binding motif for the *MuDR* transposase protein A (MURA) (Benito and Walbot, 1997). In addition, two genes contained within the autonomous *MuDR* element, including the transposase MURA, are transcribed convergently from promoters located within the TIR (Hershberger et al., 1995). Furthermore, the promoters in the TIRs are also responsible for the expression of the internal regions of some coding non-autonomous elements (Jiang et al., 2004) suggesting the importance of the TIR sequence for transposition and retention of MULEs in the genome.

Transposition. Due to the prevalence of *Mutators* and MULEs in plants, understanding how these elements amplify in the host genome is essential. To this end, studies within this superfamily are limited to the active autonomous member (*MuDR*) in maize. The MURA protein binds to a conserved ~32 bp motif in both methylated and unmethylated TIRs of known active *Mutator* elements (Benito and Walbot, 1997). It is presumed that the MURA binds the TIRs, consequently bringing them together and catalyzing a double stranded break (Lisch and Jiang, 2009). In maize the integration process appears to require the MURB (*MuDR* protein B) protein, encoded by *mudrB* (Lisch et al., 1999); however, the precise molecular role of *mudrB* in transposition remains to be elucidated. Furthermore, the exact molecular processes involved in the excision and reintegration of MULEs also awaits further analyses.

Target selection. The first few detected *Mutator* elements already allowed identification of target preference. These elements were cloned due to its insertion nearby and disruption of three genes, *Adh1*, *A1* and *Bz2* (Bennetzen et al., 1984; O'Reilly et al., 1985; McLaughlin and Walbot, 1987) suggesting a preference for insertion near genic sequences. In a genome-wide mutagenesis study with *RescueMu,* a *Mu1* element containing a *pBluescript* plasmid, a strong bias against insertion in repetitive DNA was found (Fernandes et al., 2004). Only ~6% of

flanking sequences were in repeats in a recent study of over 40,000 nonredundant *Mu* insertions using the 454 technology (Liu et al., 2009). These elements also show insertion hotspots for the 5' end of genes (Dietrich et al., 2002) and particularly for the sequences directly surrounding the transcription start site and 5' of the translational start site (Liu et al., 2009). Because insertion in or near genes can typically cause gene disruption, yet certain insertions are retained in the genome. This genic insertion preference of MULEs may suggest that certain insertions can lead to mutations with adaptive traits.

Analysis of the nucleotide composition at the insertion sites has previously suggested a preference for a weak consensus insertion site (Dietrich et al., 2002). Studies with *RescueMu* insertions indicate a bias for high GC content in the TSD and a flanking dyad-symmetrical consensus: CCT-(TSD)-AGG (Fernandes et al., 2004). The weak consensus and the variation in insertion sequences have been proposed to suggest that this preference is for specific structural features rather than sequence (Lisch and Jiang, 2009). It appears that chromatin structure may play a role, and secondary structures such as transitions between high and low GC content, DNA bendability, B-DNA twist, α-philicity, protein-induced deformability (Dietrich et al., 2002) and recombination and epigenetic modifications (Liu et al., 2009) are likely important.

Epigenetic regulation. Similar to other TEs in the genome, *Mutator* activity is regulated by silencing mechanisms. The typical occurrence when *Mutator* elements become inactive in maize is methylation (Chomet et al., 1991). Although methylation is a typical hallmark of epigenetic silencing, methylation in *Mu1* and *MuDR* derivatives do not seem to result in transcriptional silencing (Barkan and Martienssen, 1991; Lisch et al., 1999). In *Arabidopsis*, mutation of the chromatin remodeling factor *DDM1 (Decrease in DNA methylation)* results in

progressive loss of heterochromatin DNA methylation, and transcriptional and transpositional activation of different MULE families (Singer, 2001).

In maize, an element that reliably silences and leads to methylation of *MuDR* is *Mu killer* (*Muk*) (Slotkin et al., 2003). Later, *Muk* was described to be a derivative of the *MuDR* element that contained a deletion and an inverted duplication of the internal *MuDR* sequence (Slotkin et al., 2005). *Muk* transcription produces a hairpin transcript that is processed into small RNAs. The initiation of silencing by *Muk* appears to be distinct from the maintenance of silencing among *Mu* elements in maize, due to its different small interfering RNA (siRNA) expression pattern and its lack of dependence on the RNA-dependent RNA polymerase (*MOP1*) (Lisch and Jiang, 2009).

Pack-MULEs

Within the *Mutator* superfamily are a special group of non-autonomous elements, called Pack-MULEs that carry genes and gene fragments (Jiang et al., 2004). In fact, the first discovered *Mutator* element (*Mu1*) is a Pack-MULE which was later found to carry a sequence conserved in all maize lines as well as closely related species (Talbert and Chandler, 1989). This sequence is expressed and referred to as *Mu*-related sequence A (MRS-A); however, its function in maize is still unknown. To date, Pack-MULEs have been characterized in both monocots and dicots, including rice, maize, *Lotus japonicus*, *Arabidopsis*, sacred lotus, and tomato (Yu et al., 2000; Jiang et al., 2004; Juretic et al., 2005; Holligan et al., 2006; Schnable et al., 2009; Ferguson and Jiang, 2012; Ming et al., 2013) suggesting their prevalence among plants and its ancient formation. Consistent with *Mutator* and MULEs, Pack-MULEs also show insertion preference for regions flanking the 5' end of genes (Jiang et al., 2011).

Mechanism of acquisition. To date, the mechanism of formation and acquisition of the genes and gene fragments by Pack-MULEs remains to be elucidated. Three models have been proposed. The first is a process similar to the acquisition of resistance genes by *IS* elements in bacteria (Talbert and Chandler, 1989). In this model, a mobile solo-TIR is suggested to encompass the *Mu*-related Sequence A (MRS-A) to form the *Mu1* element. However, to date, a mobile solo-TIR has not been reported. Secondly, Bennetzen and Springer (1994) suggested a model using an ectopic gene conversion across a nicked-cruciform structure. Based on this model, a stem-loop structure is formed with the TIR serving as the stem and the internal region as the loop. A nick within the loop results from an endonucleolytic cleavage and the subsequent repair of the gap may proceed using a genomic sequence template with short homology to the gap ends. In this process a genomic sequence not previously associated with the elements becomes incorporated in the regions internal to the TIR. Finally, the third model for acquisition proposes an aberrant gap repair process that uses ectopic sequences as template. In this model, an excision event is necessary and the acquisition of sequences occurs upon the repair of the gap at the donor site (Yamashita et al., 1999). To date, none of the proposed mechanisms of acquisition have any empirical support including the amplification in the genome by solo TIRs. The long inverted nature of the TIR seems to support the likelihood of formation of a cruciform structure. However, it is unknown whether MULEs, in fact, form these structures *in vivo* or *in vitro*. The third model requires transposition of an element and can therefore be tested using the presence or absence of an active autonomous element or a transposase protein.

Despite the lack of information to imply an acquisition mechanism, a recent study in rice showed that Pack-MULEs do not acquire fragments at random. Instead these elements preferentially acquire GC-rich sequences (Jiang et al., 2011). Plant genes can be defined by GC

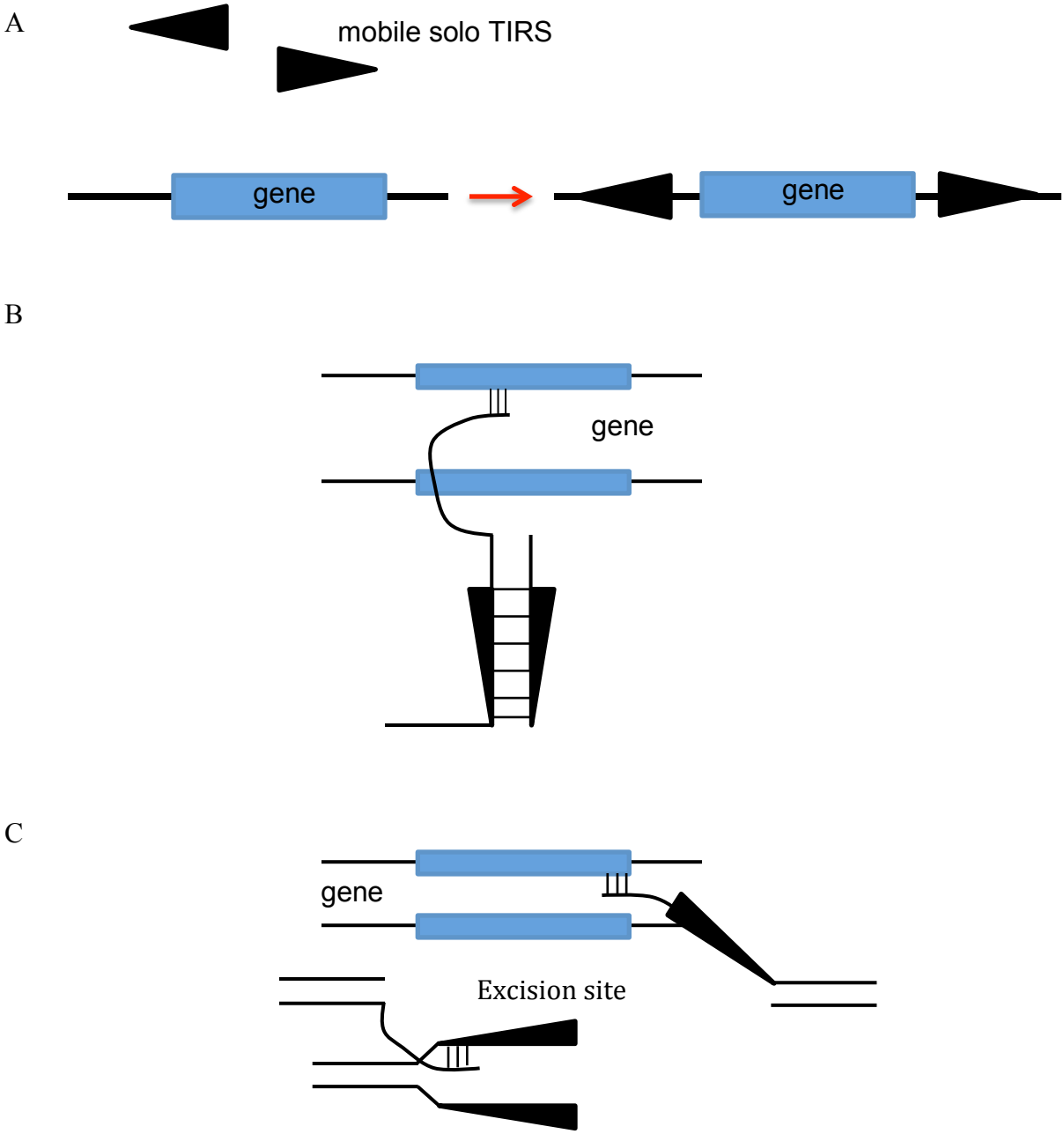


Figure 1.2 Mechanisms of sequence acquisition. (A) Capture by mobile solo-TIRs model; (B) Ectopic gene conversion using cruciform structure; (C) Aberrant gap-repair model

gradients which refers to the variation in GC content along the direction of transcription. Despite the prevalence of genes with negative GC gradients, those where the 5' half is more GC-rich than the 3' half of the gene body, in grasses, it appears that this acqusition preference is influenced by the GC gradient and not the position within the gene itself. Sequence capture via GC preference may occur via the aberrant gap-repair or ectopic gene conversion. Further genetic, biochemical and bioinformatics analyses will be necessary to begin to unravel the mystery behind the acquisition process by Pack-MULEs.

Functional capacity. Although Pack-MULEs have been discovered in many plant genomes, limited information in rice exists on the exact role and contribution of Pack-MULE encoded sequences in gene function and regulation. The first report of Pack-MULEs in rice show that they can carry gene fragments from multiple loci, which form new open reading frames (Jiang et al., 2004). In addition to coding sequences that are entirely derived from the Pack-MULE internal region, Pack-MULEs provide part of the ORF and/or UTR that fuses with downstream sequences/genes to form chimeric transcripts (Jiang et al., 2011). Both of these forms have evidence of expression based on full-length cDNA transcripts or protein expression data. Comprehensive analyses on rice Pack-MULEs showed that 22% of rice Pack-MULEs are transcribed with 28 elements having evidence of translation (Hanada et al., 2009). Interestingly, these chimeric Pack-MULEs appear to be expressed more frequently than those that acquired only a single gene.

Pack-MULEs can be expressed in either orientation and with a small subset having bidirectional transcription, that is, both sense and antisense transcripts are formed (Hanada et al., 2009). The formation of antisense transcripts suggests a role for Pack-MULE transcript in regulating the expression of parental genes, which refer to the sequences where the gene or gene

fragments are derived. Moreover, a possible feedback regulation if transcripts are expressed in aberrant quantities has been suggested (Lisch, 2005). In fact, more than half of rice Pack-MULEs are generating small RNAs (Hanada et al., 2009). In addition, parental genes that have shared small RNAs with Pack-MULEs show lower expression compared to genes without shared small RNAs.

Although a definite coding capacity of a Pack-MULE encoded gene remains to be demonstrated though biochemical and genetic analyses, a computational study has looked into the selection pressure on Pack-MULE encoded sequences. If Pack-MULE captured fragments exhibit sequence conservation, this may indicate that the sequences are conserved are likely functional. Ka/Ks ratios of Pack-MULE sequences in rice show that a considerable number of Pack-MULEs are under selective constraint and subjected to purifying selection. The selective constraints are particularly stronger for Pack-MULE transcripts in the sense orientation. These data suggest its potential in contributing new open reading frames with coding capacity (Hanada et al., 2009).

Sacred Lotus

To date, MULEs and Pack-MULEs are characterized in multiple monocots and dicots species (Yu et al., 2000; Jiang et al., 2004; Juretic et al., 2005; Holligan et al., 2006; Schnable et al., 2009; Ferguson and Jiang, 2012; Ming et al., 2013). Nevertheless, the acquisition preference for GC-rich sequences of Pack-MULEs is only evident in grasses, which are monocot plants. It raised the question whether such preference has been only evolved in monocots or it is of ancient origin. To this end, examination of the genomes of basal dicots would provide novel insight on this issue, and scared lotus (*Nelumbo nucifera*) is one of the basal dicot plants. Sacred lotus is a perennial aquatic plant that belongs to the Nelumbonaceae family and is found throughout Asia

and northern Australia (Han et al., 2007; Pan et al., 2009). It provides economic value as an ornamental and food crop in Asia and is also used for medicinal purposes (Shen-Miller, 2002; Guo, 2008). The sacred lotus has a century-old cultivation history and remarkable longevity whereby seeds are viable for up to 1300 years and the rhizomes remain healthy for more than 50 years.

The genome of sacred lotus was recently sequenced using Illumina and 454 technologies (Ming et al., 2013). This provides an excellent biological resource particularly for evolutionary analysis of transposable element biology in eudicots since lotus phylogenetically lies outside the core eudicots making it currently the most basal lineage of angiosperm sequenced. The final genome assembly (804Mb) is 86.5% of the estimated 929 Mb lotus genome (Diao et al., 2006).

Outline for this dissertation

In this dissertation, the nature of sequence acquisition by MULEs and Pack-MULEs was explored to give insights into the mechanism of capture of parental fragments and the repetitive composition of the sacred lotus was annotated. Elucidating the mechanism of sequence acquisition will be a fundamental step in understanding Pack-MULE biology and evolution and can be used to exploit the mechanism process as a tool in generating novel coding sequences. The annotation of the sacred lotus repeats is essential in dissecting its TE content and diversity. This information may be used to understand TE evolution in a basal eudicot and provide a platform for comparative TE analysis, especially acquisition mechanism of Pack-MULEs, between dicots and monocots. First, the role of TIR and sub-TIR sequence and structure on the acquisition process and the acquired sequence context was evaluated in the rice genome. Results show that these two key features of the MULEs may be involved in the acquisition process. Also, data suggest that gene GC content and ubiquity of expression play a role in the acquisition

frequency of a parental gene; moreover, very high GC content influences the retention of these fragments in the rice genome. Second, an atypical class of MULEs that possess duplicated TIRs on one or both ends of the elements was characterized. These elements were found to be present in higher copy numbers than their single or three-TIR counterparts indicating the importance of the tandem TIR structure in amplification efficiency of specific families of MULEs in plant genomes. Third, the repetitive component of the sacred lotus (*Nelumbo nucifera*) genome was analyzed. Results indicate features that sets it distinctly from other plants in terms of its TE content and diversity. Because of its location in plant evolutionary history by being the most basal angiosperm genome sequenced to date makes *N. nucifera* a model in TE analysis and evolution.

REFERENCES

REFERENCES

Adams MD (2000) The Genome Sequence of Drosophila melanogaster. Science 287: 2185–2195

Alleman, M., Freeling, M. (1986) The Mu transposable elements of maize: evidence for transposition and copy number regulation during development. Genetics 112: 107–119

Ammiraju JSS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, Walling JG, Ma J, Talag J, Brar DS, et al (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. The Plant Journal 52: 342–351

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al (2010) The genome of Theobroma cacao. Nature Genetics 43: 101–108

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, et al (2011) The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants. Science 332: 960–963

Barkan A, Martienssen RA (1991) Inactivation of maize transposon Mu suppresses a mutant phenotype by activating an outward-reading promoter near the end of Mu1. Proceedings of the National Academy of Sciences 88: 3502–3506

Bartolomé C, Bello X, Maside X (2009) Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. Genome Biology 10: R22

Becker H-A (1997) Maize Activator transposase has a bipartite DNA binding domain that recognizes subterminal sequences and the terminal inverted repeats. Molecular and General Genetics MGG 254: 219–230

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441: 87–90

Benito MI, Walbot V (1997) Characterization of the maize Mutator transposable element MURA transposase as a DNA-binding protein. Mol Cell Biol 17: 5165–75

Bennetzen J (2007) Patterns in grass genome evolution. Current Opinion in Plant Biology 10: 176–181

Bennetzen JL (2005) Mechanisms of Recent Genome Size Variation in Flowering Plants. Annals of Botany 95: 127–132

Bennetzen JL, Kellogg EA (1997) Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9: 1509–1514

Bennetzen JL, Springer PS (1994) The generation of Mutator transposable element subfamilies in maize. Theoretical and Applied Genetics. doi: 10.1007/BF00222890

Bennetzen JL, Swanson J, Taylor WC, Freeling M (1984) DNA insertion in the first intron of maize Adh1 affects message levels: cloning of progenitor and mutant Adh1 alleles. Proceedings of the National Academy of Sciences 81: 4125–4128

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491: 705–710

Bundock P, Hooykaas P (2005) An Arabidopsis hAT-like transposase is essential for plant development. Nature 436: 282–284

Bushman FD (2003) Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell 115: 135–138

Callinan PA, Batzer MA (2006) Retrotransposable Elements and Human Disease. *In* J-N Volff, ed, Genome Dynamics. KARGER, Basel, pp 104–115

Cao X, Jacobsen SE (2002) Role of the Arabidopsis DRM Methyltransferases in De Novo DNA Methylation and Gene Silencing. Current Biology 12: 1138–1144

Chalker DL, Sandmeyer SB (1992) Ty3 integrates within the region of RNA polymerase III transcription initiation. Genes & Development 6: 117–128

Chalvet F (2003) Hop, an Active Mutator-like Element in the Genome of the Fungus Fusarium oxysporum. Molecular Biology and Evolution 20: 1362–1375

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, et al (2010) Draft genome sequence of the oilseed species Ricinus communis. Nature Biotechnology 28: 951–956

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562

Chomet P, Lisch D, Hardeman KJ, Chandler VL, Freeling M (1991) Identification of a regulatory transposon that controls the Mutator transposable element system in maize. Genetics 129: 261–270

Cordaux R (2006) From the Cover: Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proceedings of the National Academy of Sciences 103: 8101–8106

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488: 213–217

Dassanayake M, Oh D-H, Haas JS, Hernandez A, Hong H, Ali S, Yun D-J, Bressan RA, Zhu J-K, Bohnert HJ, et al (2011) The genome of the extremophile crucifer Thellungiella parvula. Nature Genetics 43: 913–918

Diao Y, Chen L, Yang G, Zhou M, Song Y, Hu Z, Liu JY (2006) Nuclear DNA C-values in 12 species in Nymphales. Caryologia 59: 25–30

Dietrich CR, Cui F, Packila ML, Li J, Ashlock DA, Nikolau BJ, Schnable PS (2002) Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics 160: 697–716

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, et al (2011) De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nature Biotechnology 29: 521–527

Drinnan AN, Crane PR, Hoot SB (1994) Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). *In* PK Endress, EM Friis, eds, Early Evolution of Flowers. Springer Vienna, Vienna, pp 93–122

Du J, Grant D, Tian Z, Nelson RT, Zhu L, Shoemaker RC, Ma J (2010) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11: 113

Duke JA, Duke (2002) Handbook of medicinal herbs. CRC Press, Boca Raton, FL

Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21: i152–i158

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9: 18

Ferguson AA, Jiang N (2012) Mutator-Like Elements with Multiple Long Terminal Inverted Repeats in Plants. Comparative and Functional Genomics 2012: 1–14

Fernandes J, Dong Q, Schneider B, Morrow DJ, Nan G-L, Brendel V, Walbot V (2004) Genome-wide mutagenesis of Zea mays L. using RescueMu transposons. Genome Biol 5: R82

Feschotte C (2004) Merlin, a New Superfamily of DNA Transposons Identified in Diverse Animal Genomes and Related to Bacterial IS1016 Insertion Sequences. Molecular Biology and Evolution 21: 1769–1780

Feschotte C (2005) DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. Nucleic Acids Research 33: 2153–2165

Feschotte C, Jiang N, Wessler SR (2002) PLANT TRANSPOSABLE ELEMENTS: WHERE GENETICS MEETS GENOMICS. Nature Reviews Genetics 3: 329–341

Flavell AJ, Pearce SR, Kumar A (1994) Plant transposable elements and the genome. Current Opinion in Genetics & Development 4: 838–844

Guo HB (2008) Cultivation of lotus (Nelumbo nucifera Gaertn. ssp. nucifera) and its utilization in China. Genetic Resources and Crop Evolution 56: 323–330

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Research 38: e199–e199

Han Y-C, Teng C-Z, Zhong S, Zhou M-Q, Hu Z-L, Song Y-C (2007) Genetic variation and clonal diversity in populations of Nelumbo nucifera (Nelumbonaceae) in central China detected by ISSR markers. Aquatic Botany 86: 69–75

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu S-H, Jiang N (2009) The Functional Role of Pack-MULEs in Rice Inferred from Purifying Selection and Expression Profile. THE PLANT CELL ONLINE 21: 25–38

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Research 16: 1252–1261

Hehl R, Nacken WKF, Krause A, Saedler H, Sommer H (1991) Structural analysis of Tam3, a transposable element from Antirrhinum majus, reveals homologies to the Ac element from maize. Plant Molecular Biology 16: 369–371

Hershberger RJ, Benito M-I, Hardeman KJ, Warren C, Chandler VL, Walbot V (1995) Characterization of the major transcripts encoded by the regulatory MuDR transposable element of maize. Genetics 140: 1087–1098

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

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proceedings of the National Academy of Sciences 93: 7783–7788

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR (2006) The Transposable Element Landscape of the Model Legume Lotus japonicus. Genetics 174: 2215–2228

Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS (2011) Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proceedings of the National Academy of Sciences 108: 2322–2327

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496: 498–503

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al (2009) The genome of the cucumber, Cucumis sativus L. Nature Genetics 41: 1275–1281

Ichikawa H, Ikeda K, Wishart WL, Ohtsubo E (1987) Specific binding of transposase to terminal inverted repeats of transposable element Tn3. Proceedings of the National Academy of Sciences 84: 8220–8224

Jarvis CE, Linnean Society of London, Natural History Museum (London, England) (2007) Order out of chaos: Linnaean plant names and their types. Linnean Society of London in association with the Natural History Museum, London, London

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573

Jiang N, Ferguson AA, Slotkin RK, Lisch D (2011) Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proceedings of the National Academy of Sciences 108: 1537–1542

Jiang N, Panaud O (2013) Transposable Element Dynamics in Rice and Its Wild Relatives. *In* Q Zhang, RA Wing, eds, Genetics and Genomics of Rice. Springer New York, New York, NY, pp 55–69

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics 19: 68–72

Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res 15: 1292–1297

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH (2000) From the Cover: Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences 97: 6603–6607

Kamal M (2006) A large family of ancient repeat elements in the human genome is under strong selection. Proceedings of the National Academy of Sciences 103: 2740–2745

Kapitonov VV (2006) Self-synthesizing DNA transposons in eukaryotes. Proceedings of the National Academy of Sciences 103: 4540–4545

Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3: e181

Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Reviews Genetics 9: 411–412

Kapitonov VV, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends in Genetics 23: 521–529

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genetics 9: e1003470

Kawakami T, Strakosh SC, Zhen Y, Ungerer MC (2010) Different scales of Ty1/copia-like retrotransposon proliferation in the genomes of three diploid hybrid sunflower species. Heredity (Edinb) 104: 341–350

Kempken F, Windhofer F (2001) The hAT family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma 110: 1–9

Kinoshita T (2004) One-Way Control of FWA Imprinting in Arabidopsis Endosperm by DNA Methylation. Science 303: 521–523

Knip M, de Pater S, Hooykaas PJ (2012) The SLEEPER genes: a transposase-derived angiosperm-specific gene family. BMC Plant Biology 12: 192

Kumar A, Bennetzen JL (1999) Plant Retrotransposons. Annual Review of Genetics 33: 479–532

Kunze R, Weil CF (2002) The hAT and CACTA superfamilies of plant transposons. Mobile DNA II. pp 565–610

Kuwahara A, Kato A, Komeda Y (2000) Isolation and characterization of copia-type retrotransposons in Arabidopsis thaliana. Gene 244: 127–136

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921

Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics 11: 204–220

Lazarow K, Du M-L, Weimer R, Kunze R (2012) A Hyperactive Transposase of the Maize Transposable Element Activator (Ac). Genetics 191: 747–756

Li W, Shaw JE (1993) A variant Tc4 transposable element in the nematode *C.elegans* could encode a novel protein. Nucleic Acids Research 21: 59–67

Lindroth AM (2001) Requirement of CHROMOMETHYLASE3 for Maintenance of CpXpG Methylation. Science 292: 2077–2080

Linheiro RS, Bergman CM (2012) Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster. PLoS ONE 7: e30008

Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, Richard McCombie W, Lavine K, Mittal V, May B, Kasschau KD, et al (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430: 471–476

Lisch D (2002) Mutator transposons. Trends Plant Sci 7: 498–504

Lisch D (2009) Epigenetic Regulation of Transposable Elements in Plants. Annual Review of Plant Biology 60: 43–66

Lisch D (2005) Pack-MULEs: theft on a massive scale. BioEssays 27: 353–355

Lisch D, Girard L, Donlin M, Freeling M (1999) Functional analysis of deletion derivatives of the maize transposon MuDR delineates roles for MURA and MURB proteins. Genetics 151: 331–341

Lisch D, Jiang N (2009) Mutator and MULE transposons. *In* JL Bennetzen, S Hake, eds, Handbook of Maize. Springer New York, New York, NY, pp 277–306

Liu S, Yeh C-T, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS (2009) Mu Transposon Insertion Sites and Meiotic Recombination Events Co-Localize with Epigenetic Marks for Open Chromatin across the Maize Genome. PLoS Genetics 5: e1000733

Loot C, Santiago N, Sanz A, Casacuberta JM (2006) The proteins encoded by the pogo-like Lemi1 element bind the TIRs and subterminal repeated motifs of the Arabidopsis Emigrant MITE: consequences for the transposition mechanism of MITEs. Nucleic Acids Research 34: 5238–5246

Marquez CP, Pritham EJ (2010) Phantom, a New Subclass of Mutator DNA Transposons Found in Insect Viruses and Widely Distributed in Animals. Genetics 185: 1507–1517

Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, Matsumoto T, Sato K, Schulman A, Muehlbauer GJ, et al (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature. doi: 10.1038/nature11543

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362–367

McClintock B (1951) CHROMOSOME ORGANIZATION AND GENIC EXPRESSION. Cold Spring Harbor Symposia on Quantitative Biology 16: 13–47

McLaughlin M, Walbot V (1987) Cloning of a mutable bz2 allele of maize by transposon tagging and differential hybridization. Genetics 117: 771–776

Middleton CP, Stein N, Keller B, Kilian B, Wicker T (2013) Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity. The Plant Journal 73: 347–356

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, et al (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167–177

Mills RE, Bennett EA, Iskow RC, Devine SE (2007) Which transposable elements are active in the human genome? Trends in Genetics 23: 183–191

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, et al (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452: 991–996

Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li L-T, Zhang Q, Kim M-J, Schatz MC, Campbell M, et al (2013) Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). Genome Biology 14: R41

Morgenstern B (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15: 211–218

Muotri AR, Marchetto MCN, Coufal NG, Gage FH (2007) The necessary junk: new functions for transposable elements. Human Molecular Genetics 16: R159–R167

Nassif N, Penney J, Pal S, Engels WR, Gloor GB (1994) Efficient copying of nonhomologous sequences from ectopicsites via P-element-induced gap repair. Mol Cell Biol 14: 1613–1625

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu Z, Loftus B, Xi Z, Megy K, Grabherr M, et al (2007) Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. Science 316: 1718–1723

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497: 579–584

O'Reilly C, Shepherd NS, Pereira A, Schwarz-Sommer Z, Bertram I, Robertson DS, Peterson PA, Saedler H (1985) Molecular cloning of the a1 locus of Zea mays using the transposable elements En and Mu1. EMBO J 4: 877–882

Oliver KR, McComb JA, Greene WK (2013) Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity. Genome Biology and Evolution 5: 1886–1901

Pan L, Xia Q, Quan Z, Liu H, Ke W, Ding Y (2009) Development of Novel EST-SSRs from Sacred Lotus (Nelumbo nucifera Gaertn) and Their Utilization for the Genetic Diversity Analysis of N. nucifera. Journal of Heredity 101: 71–82

Pardue M-L, Rashkova S, Casacuberta E, DeBaryshe PG, George JA, Traverse KL (2005) Two retrotransposons maintain telomeres in Drosophila. Chromosome Research 13: 443–453

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien M-A (2010) Impact of transposable elements on the organization and function of allopolyploid genomes: Research review. New Phytologist 186: 37–45

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457: 551–556

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, et al (2006) Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Research 16: 1262–1269

Piriyapongsa J, Marino-Ramirez L, Jordan IK (2006) Origin and Evolution of Human microRNAs From Transposable Elements. Genetics 176: 1323–1337

Robertson DS (1978) Characterization of a mutator system in maize. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 51: 21–28

Robertson HM (2002) Evolution of DNA transposons. Mobile DNA II. American Society for Microbiology Press, Washington, D.C, pp 1093–1110

Sanmiguel P (1998) Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons. Annals of Botany 82: 37–44

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, et al (2010) Sequence Analysis of the Genome of an Oil-Bearing Tree, Jatropha curcas L. DNA Research 18: 65–76

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in Ecology & Evolution 25: 537–546

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326: 1112–1115

Schulman AH, Kalendar R (2005) A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. Cytogenetic and Genome Research 110: 598–605

Sequencing Project IRG (2005) The map-based sequence of the rice genome. Nature 436: 793–800

Shaheen M, Williamson E, Nickoloff J, Lee S-H, Hromas R (2010) Metnase/SETMAR: a domesticated primate transposase that enhances DNA repair, replication, and decatenation. Genetica 138: 559–566

Shen-Miller J (2002) Sacred lotus, the long-living fruits of China Antique. Seed Science Research 12: 131–143

Shen-Miller J, Aung LH, Turek J, Schopf JW, Tholandi M, Yang M, Czaja A (2013) Centuries-Old Viable Fruit of Sacred Lotus Nelumbo nucifera Gaertn var. China Antique. Tropical Plant Biology 6: 53–68

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, et al (2010) The genome of woodland strawberry (Fragaria vesca). Nature Genetics 43: 109–116

Singer T (2001) Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes & Development 15: 591–602

Singh R, Ming R, Yu Q (2013) Nucleotide Composition of the Nelumbo nucifera Genome. Tropical Plant Biology 6: 85–97

Slotkin RK, Freeling M, Lisch D (2003) Mu killer causes the heritable inactivation of the Mutator family of transposable elements in Zea mays. Genetics 165: 781–797

Slotkin RK, Freeling M, Lisch D (2005) Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. Nature Genetics 37: 641–644

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8: 272–285

Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Research 37: 7002–7013

Strommer JN, Hake S, Bennetzen J, Taylor WC, Freeling M (1982) Regulatory mutants of the maize Adh1 gene caused by DNA insertions. Nature 300: 542–544

Talbert LE, Chandler VL (1989) Characterization of a highly conserved sequence related to mutator transposable elements in maize. Mol Biol Evol 5: 519–529

Temin HM (1981) Structure, variation and synthesis of retrovirus long terminal repeat. Cell 27: 1–3

Tenaillon MI, Hollister JD, Gaut BS (2010) A triptych of the evolution of plant transposable elements. Trends in Plant Science 15: 471–478

Thomas CA (1971) The Genetic Organization of Chromosomes. Annual Review of Genetics 5: 237–256

Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, Toyoda A, Fujiyama A, Tarutani Y, Kakutani T (2012) Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. Genes & Development 26: 705–713

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T (2009) Bursts of retrotransposition reproduced in Arabidopsis. Nature 461: 423–426

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al (2006) The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science 313: 1596–1604

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, Whaley AM, et al (2011) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nature Biotechnology 30: 83–89

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, et al (2010) The genome of the domesticated apple (Malus × domestica Borkh.). Nature Genetics 42: 833–839

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, et al (2007) A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLoS ONE 2: e1326

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, et al (2013) The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics 45: 487–494

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, et al (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463: 763–768

Volff J-N (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28: 913–922

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al (2011) The genome of the mesopolyploid crop species Brassica rapa. Nature Genetics 43: 1035–1039

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al (2007) A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973–982

Wikstrom N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proceedings of the Royal Society B: Biological Sciences 268: 2211–2220

Xu Q, Chen L-L, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao W-B, Hao B-H, Lyon MP, et al (2012) The draft genome of sweet orange (Citrus sinensis). Nature Genetics 45: 59–66

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, et al (2011) Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research 35: W265–W268

Yamashita S, Takano-Shimizu T, Kitamura K, Mikami T, Kishima Y (1999) Resistance to gap repair of the transposon Tam3 in Antirrhinum majus: a role of the end regions. Genetics 153: 1899–1908

Yang G, Weil CF, Wessler SR (2006) A rice Tc1/mariner-like element transposes in yeast. Plant Cell 18: 2469–2478

Yin H, Liu J, Xu Y, Liu X, Zhang S, Ma J, Du J (2013) TARE1, a Mutated Copia-Like LTR Retrotransposon Followed by Recent Massive Amplification in Tomato. PLoS ONE 8: e68587

Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature. doi: 10.1038/nature10625

Yu Z, Wright SI, Bureau TE (2000) Mutator-like elements in Arabidopsis thaliana. Structure, diversity and evolution. Genetics 156: 2019–2031

Zemach A, Kim MY, Hsieh P-H, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell 153: 193–205

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, et al (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nature Biotechnology 30: 549–554

Zhou L, Mitra R, Atkinson PW, Burgess Hickman A, Dyda F, Craig NL (2004) Transposition of hAT elements links transposable elements and V(D)J recombination. Nature 432: 995–1001

Zhu Y (2003) From the Cover: Controlling integration specificity of a yeast retrotransposon. Proceedings of the National Academy of Sciences 100: 5891–5895

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA (2007) Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evolutionary Biology 7: 152

CHAPTER 2:

Selective Acquisition and Retention of Genomic Sequences by Pack-*Mutator*-Like Elements Based on GC Content and Breadth of Expression

Ferguson AA, Zhao D, Jiang N (2013) Selective Acquisition and Retention of Genomic Sequences by Pack-*Mutator*-Like Elements Based on GC Content and Breadth of Expression. Plant Physiology. pp.113.223271

Abstract

The process of gene duplication followed by sequence and functional divergence is important for the generation of new genes. Pack-MULEs, nonautonomous *Mutator*-like elements (MULEs) that carry genic sequence(s), are potentially involved in generating new open reading frames and regulating parental gene expression. These elements are identified in many plant genomes and are most abundant in rice (*Oryza sativa*). Despite the abundance of Pack-MULEs, the mechanism by which parental genes are captured by Pack-MULEs remains largely unknown. In this study, we identified all MULEs in rice and examined factors likely important for sequence acquisition. Terminal inverted repeat MULEs are the predominant MULE type and account for the majority of the Pack-MULEs. In addition to genic sequences, rice MULEs capture GC-rich intergenic sequences, albeit at a much lower frequency. MULEs carrying nontransposon sequences have longer terminal inverted repeats and higher GC content in terminal and subterminal regions. An overrepresentation of genes with known functions and genes with orthologs among parental genes of Pack-MULEs is observed in rice, maize (*Zea mays*), and Arabidopsis (*Arabidopsis thaliana*), suggesting preferential acquisition for bona fide genes by these elements. Pack-MULEs selectively acquire/retain parental sequences through a combined effect of GC content and breadth of expression, with GC content playing a stronger role. Increased GC content and number of tissues with detectable expression result in higher chances of a gene being acquired by Pack-MULEs. Such selective acquisition/retention provides these elements greater chances of carrying functional sequences that may provide new genetic resources for the evolution of new genes or the modification of existing genes.

Introduction

Transposable elements (TEs) are sequences in the genome that move from one location to another and in the process multiply in copy number. According to the transposition intermediate, TEs are classified into two major classes: class I or RNA elements, which transpose via an RNA intermediate using a copy-and-paste mechanism; and class II or DNA elements, which transpose via a DNA intermediate using a cut-and-paste mechanism. Based on their coding capacity for transposition machinery, both classes of TEs can be divided into autonomous and nonautonomous elements. Autonomous elements encode the protein products (transposase or reverse transcriptase) required for their transposition, whereas nonautonomous elements do not encode relevant proteins and rely on their cognate autonomous elements for transposition. TEs constitute over 50% of many plant genomes and as much as 85% of the maize (*Zea mays*) genome (Devos et al., 2005; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010; Tomato Genome Consortium, 2012; Nystedt et al., 2013; Wu et al., 2013). In addition, computational and biological analyses of genomic information have revealed critical roles of transposons in gene expression, regulation, and genome evolution (Bennetzen and Kellogg, 1997; Lippman et al., 2004; Piegu et al., 2006; Ammiraju et al., 2007; Bennetzen, 2007; Slotkin and Martienssen, 2007; Zuccolo et al., 2007; Feschotte, 2008).

The *Mutator* superfamily is a class II/DNA TE originally discovered in maize (Robertson, 1978). Since the initial discovery of *Mu1* and *MuDR* in maize (Robertson,1978; Robertson et al., 1989), similar elements were later identified from the maize genome and subsequently in other organisms including plants, animals, and fungi, where they are referred to as *Mutator*-like elements (MULEs; Yu et al., 2000; Lisch et al., 2001; Chalvet et al., 2003; Jiang et al., 2004; Holligan et al., 2006; Marquez and Pritham, 2010). MULEs are typically

characterized by an 8- to 11-bp target site duplication (TSD) flanking the element, with 9-bp TSD as the most frequent form. In addition, the majority of these elements are known for the presence of long terminal inverted repeats (TIRs), which typically range from 100 to 500 bp, a feature that largely sets them apart from other major class II TEs such as *En/Spm*, *Helitron*, *PIF/Pong*, and *Tc1/Mariner* elements. MULEs associated with long TIRs are referred to as TIR MULEs. TIR sequences appear to be important for element transposition and expression (Benito and Walbot, 1997; Raizada et al., 2001; Jiang et al., 2004). Recently, however, non-TIR MULEs have been reported in Arabidopsis (*Arabidopsis thaliana*), *Lotus japonicus*, maize, and yeast (*Yarrowia lipolytica*; Yu et al., 2000; Neuvéglise et al., 2005; Holligan et al., 2006; Wang and Dooner, 2006). Non-TIR MULEs refer to the MULEs with exceptionally short TIRs (less than 50 bp) and low similarity between the inverted terminal sequences. The detection of non-TIR MULEs in multiple plants suggests that extended long TIRs are dispensable for the transposition of MULEs in plants. Although elements belonging to the same TIR MULE family share an overall sequence similarity in their TIRs, they vary in their internal region. The *Mu* family of maize that includes multiple elements (*Mu1–Mu13*) share a 220-bp sequence in their TIRs, but the internal region between the TIRs may contain unique and unrelated sequences (Chomet et al., 1991; Lisch, 2002; Lisch and Jiang, 2009). The *Mu4* elements, for instance, have much longer TIRs compared with other *Mu* elements (530 bp long), and the TIR sequence includes a fragment from a *BRASSINOSTEROID INSENSITIVE1* gene (Lisch, 2002). Thus, in addition to differences in the internal sequence, elements within a MULE family vary in their TIR lengths.

Pack-MULEs are nonautonomous *Mutator* and MULEs that carry genes or gene fragments. Although the abundance of Pack-MULEs was not acknowledged until the availability of the entire rice (*Oryza sativa*) genomic sequence, the first *Mutator* element discovered (*Mu1*)

was, in fact, a Pack-MULE carrying a fragment of the *MRS-A* gene (Talbert and Chandler, 1988), as were the other nonautonomous *Mutator* elements (Lisch, 2002). To date, Pack-MULEs have been characterized in both monocots and dicots, including rice, maize, *L. japonicus*, Arabidopsis, tomato (*Solanum lycopersicum*), and sacred lotus (*Nelumbo nucifera* ; Yu et al., 2000; Jiang et al., 2004; Hoen et al., 2006; Holligan et al., 2006; Schnable et al., 2009; Ferguson and Jiang, 2012; Ming et al., 2013), suggesting their prevalence among plants. The genes from which gene sequences or fragments are captured are referred to as parental genes, and the captured fragment is referred to as the acquired sequence. Previous work identified 2,853 Pack-MULEs in rice that have transduced about 1,500 parental genes (Jiang et al., 2011). Comprehensive analyses showed that over 22% of rice Pack-MULEs are transcribed, with at least 28 elements having evidence of translation (Hanada et al., 2009). These elements often carry gene fragments from multiple loci, forming new open reading frames (ORFs). In addition to the formation of independent ORFs, Pack-MULEs can serve as part of the ORF and/or untranslated region that fuses with adjacent sequences/genes to form chimeric transcripts (Jiang et al., 2011). Pack-MULE transcripts are found in either orientation with regard to the transcription of the parental gene, with a small subset having bidirectional transcription. The formation of antisense transcripts suggests a critical role for Pack-MULE-derived transcripts in regulating the expression of parental genes through the activity of small RNAs (Hanada et al., 2009). In fact, over half of Pack-MULEs in rice are directly involved in the formation of small RNAs. Parental genes that have shared small RNAs with Pack-MULEs show lower expression levels compared with genes without an association with small RNAs (Hanada et al., 2009). Thus far, rice has remained exceptional in its Pack-MULE copy number load. Another advantage of studying Pack-MULEs in rice is the unparalleled quality of its reference genome sequence,

which was accomplished using the traditional bacterial artificial chromosome-by-bacterial artificial chromosome sequencing technology (International Rice Genome Sequencing Project, 2005).

Despite progress in MULE and Pack-MULE identification in sequenced higher eukaryotes, the process by which parental genes are captured by these elements remains to be elucidated. Thus far, two probable mechanisms have been proposed. Bennetzen and Springer (1994) suggested a model (model 1) similar to an ectopic gene conversion across a nicked-cruciform structure. Here, ectopic sequences are introduced into the internal region of the element during repair of the nick within the loop. According to this model, acquisition may or may not require the presence of transposase. The second model (model 2) proposes an aberrant gap-repair process that uses ectopic sequences as template during the repair of the empty site. In this model, an excision event is necessary, and the acquisition of new sequences occurs upon the repair of the gap at the donor site (Yamashita et al., 1999). As a result, the acquisition requires the presence of transposase but is not associated with transposition of the element. Both models predict the involvement of short stretches of homology between the broken ends and the new genomic sequence not previously associated with the element, which ultimately becomes incorporated in the internal region. Although neither of the two models has any empirical support at this time, computational analysis of Pack-MULEs in rice, maize, and Arabidopsis has shed some light on the acquisition process. A phenomenon that likely extends to all grass genomes, where significant GC islands and gradients persist, is the preferential acquisition of GC-rich sequences by Pack-MULEs (Jiang et al., 2011).

In this study, a comprehensive analysis of all MULEs in the rice genome, including Pack-MULEs, was performed to further understand how Pack-MULEs select and acquire parental

gene sequences. The results from this study indicate that element TIR and sub-TIR properties differ between Pack-MULEs and non-Pack-MULEs and may be involved in target selection and acquisition. Analysis of the parental genes of Pack-MULEs in rice, maize, and Arabidopsis supports the role of GC content and ubiquity in the expression of the parental genes in sequence acquisition, which explains the significant preference of MULEs to duplicate genic sequences.

Methods

Identification of MULEs and Pack-MULEs

The sequences for rice (*Oryza sativa* subsp. *japonica* 'Nipponbare') pseudomolecules and gene annotation information were downloaded from the Rice Genome Annotation Project at Michigan State University (http://rice.plantbiology.msu.edu/; release 7.0). The rice TIR library was built using repeats generated by RECON (Bao and Eddy, 2002). Prior to the identification of MULEs and Pack-MULEs, MULE TIRs were classified as TIR MULEs or non-TIR MULEs. MULE families whose TIRs are at least 50 bp in length with at least 60% similarity are considered as TIR MULEs. However, MULE families less than 150 bp in size (small elements) where TIR length is at least 40 bp and their terminal sequence is related to a TIR MULE were also classified as TIR MULEs, because the short TIR is due to deletion and not to phylogeny origin. All other families are considered non-TIR MULEs. The procedure for the annotation and identification of Pack-MULEs was similar to that described previously with some modifications (Hanada et al., 2009). Briefly, genomic sequence was masked with MULE TIR library and all possible TIR pairs within 20 kb of each other were examined. The annotation of other MULEs was similar to that of Pack-MULEs, except that there is no requirement for the internal region of MULEs to match proteins. Auto-MULEs are MULEs with matches to previously known MULE transposases. Elements with flanking 9- to 11-bp TSD with no more than two mismatches (or 1-

bp mismatch plus 1-bp insertion/deletion) were accepted for further classification and analysis. For elements with 8-bp TSD, only 1-bp mismatch or 1-bp insertion/deletion was accepted. The presence of TSD for non-Pack-MULEs was detected by custom perl scripts, and a maximum 10-bp swing from the putative element ends was allowed. For all elements with parental copies, TSD was verified by manual examination of elements and flanking sequences.

The identification of the parental origin of the sequences captured by MULEs and Pack-MULEs was conducted as described previously (Jiang et al., 2011). For an individual Pack-MULE, the sequence with the highest similarity score (BLASTN, $E = 1e^{-10}$) that was not associated with a MULE TIR was considered as the parental copy of the internal sequence in a Pack-MULE. Elements without matching any proteins that did not contain a recognizable parental genomic sequence were classified as MULE-other or non-Pack-MULEs. Elements with recognizable nongenic parental sequences were classified as MULE-intergenic. Elements with hits only to hypothetical proteins and without parental sequences were classified as MULE-HypProt.

TIR and Sub-TIR Analyses

Since the majority of Pack-MULEs belong to TIR-MULEs, TIR and sub-TIR analyses were performed only on the TIR-MULEs. To identify the TIR length of each individual element, the terminal 800-bp sequence (or half of the element if the element is shorter than 1,600 bp) of each element was aligned using DIALIGN2 (Morgenstern, 2004). A custom perl script was used to determine the length of the TIR on each side, whereby considerable sequence alignment falls off. The sub-TIR was defined as the 50-bp sequence immediately following the TIR, as determined previously. The GC content of each individual TIR and sub-TIR sequence was calculated using a custom perl script. Calculations of sub-TIR free energy were performed using

UNAFold (Markham and Zuker, 2008), available at http://mfold.rna.albany.edu. The statistical difference between each group was examined using the R package (http://www. r-project.org). A Bonferroni correction was applied to account for multiple comparisons.

Analysis of GC Content

To calculate the GC content of MULEs and Pack-MULEs, nested TE insertions were first curated and removed from the element sequence. Determination of the GC content of parental genes was conducted after masking with the rice repeat library that excluded Pack-MULEs. To calculate the GC gradient along MULE sequences, the TIR sequences (on both ends of the elements) and the internal region (the sequences between the TIRs) were divided into two and 10 equal-sized bins, respectively. A custom perl script was used to determine the GC content of each bin. Comparisons of GC content between groups were performed using the R package (http://www.r-project.org).

Gene Functional and Expression Analyses

The biological process GOSlim assignments and RNA-Seq expression data for rice genes were downloaded from the Rice Genome Annotation Project at Michigan State University (http://rice.plantbiology.msu.edu/). GOSlim categories were calculated such that a total count of 1 was generated from each gene; that is, genes with multiple GOSlim assignments were given an equal proportion totaling to 1. To classify expressed genes, only RNA-Seq libraries with calculated FPKM values were used, to avoid misclassifying background or noise reads from expression calls made using a single read to a gene. Genes were considered expressed if the FPKM values were 1 or greater in at least one expression library.

The maize (*Zea mays*) filtered gene set sequence (release 5b) and functional annotation were downloaded from the maize sequencing project (http://www.maizesequence.org). The

Arabidopsis (*Arabidopsis thaliana*) gene sequences and functional annotation were downloaded from The Arabidopsis Information Resource 10 (http://www.arabidopsis.org). The genes were classified as "known" if a functional annotation is available; otherwise, the genes were classified as unknown. Maize RNA-Seq expression data were obtained from previously published work (Davidson et al., 2011). Evaluation of the expression of maize genes was similar to rice (FPKM ≥1 in at least one expression library) from RNA-Seq, which includes 13 different expression libraries. Since a similarly comprehensive RNA-Seq expression library is not readily available for Arabidopsis, the MPSS data set from eight different expression libraries was downloaded (http://mpss. udel.edu/at/mpss_index.php; Meyers et al., 2004a, 2004b). To determine the expression patterns of genes in Arabidopsis, eight libraries were used, and a gene was classified as expressed if the TPM values were 5 or greater in at least one expression library. Comparisons of various expression parameters among groups were performed using the R package (http://www.r-project.org).

To determine the distribution of parental genes among non-TE genes with and without orthologs, Arabidopsis and rice gene orthologous data were downloaded from the Rice Genome Annotation Project at Michigan State University (http://rice.plantbiology.msu.edu/; Lin et al., 2010; Davidson et al., 2012). For maize genes, data were downloaded from the maize sequencing project (http://www.maizesequence.org; Schnable et al., 2009).

Age of Acquisition Events

To roughly estimate the age of the genic and intergenic acquisition events, Pack-MULE and MULE-intergenic sequences was aligned to parental sequences using BLASTN (M = 5, N = 211, Q = 22, R = 11, E = $1e^{-10}$, wordmask = dust, wordmask = seg, hspsepSmax = 100, hspsepQmax = 100) to determine the boundary of the alignable region. Subsequently, each pair

of alignable sequences were aligned using MUSCLE (Edgar, 2004), and the output was further processed by custom perl scripts to calculate the number of transversion events between aligned sequences as well as the transversion rate for each sequence pair. An average transversion rate was assigned for parental genes that were acquired by multiple Pack-MULEs.

Results

Rice MULEs Preferentially Acquire Genic Sequences

To understand the mechanism of sequence acquisition by Pack-MULEs, we compared Pack-MULEs with MULEs that do not carry non-TE genomic sequences. To this end, we established a procedure to collect all MULEs in the rice genome, which resulted in a total of 13,857 MULEs with TSDs (Tables 2.1 and 2.2). MULEs were categorized into TIR MULE and non-TIR MULE according to a distinct similarity and length of TIRs (see "Materials and Methods"). Among the MULE elements with TSDs, 87% were TIR MULEs, suggesting that this MULE type is more predominant than the non-TIR MULEs.

If the internal region of a MULE has a non-TE genomic homolog, we call the genomic homolog the parental copy or parental gene (if it is from the genic region; see below). According to the internal sequence contained within the TIR, MULEs were further classified into five groups: (1) Pack-MULEs, as defined previously (Jiang et al., 2004), refers to elements containing genic sequence(s) (Supplemental Table S1); (2) MULE-intergenic refers to elements with a non-TE parental copy located in intergenic regions (Supplemental Table S2); (3) MULE-other or non-Pack-MULEs are elements whose internal sequences have no identifiable parental origin/sequence (Supplemental Table S3); (4) Auto-MULEs are elements containing sequences with homology to known *Mutator*/MULE transposases (Supplemental Table S4); and (5) MULE-HypProt are elements containing annotated hypothetical genes or with homology to

hypothetical genes yet without a recognizable parental copy (Supplemental Table S5). MULE-HypProt could represent ancient sequence acquisitions where the internal regions are too diverged or evolved to allow the identification of the parental copies. Alternatively, it is a result of misannotation from an automated gene annotation pipeline. The non-Pack-MULEs in each MULE type were subsequently categorized into two groups based on whether the TIR family is involved in sequence acquisition. PMTIR refers to TIR families that contain or include Pack-MULEs, while non-PMTIR refers to TIR families that contain exclusively non-Pack-MULEs.

Among the 13,857 MULEs identified, 2,924 (21.1%) carry gene or gene fragments, suggesting that the majority of MULEs do not acquire genes (Tables 2.1 and 2.2). A total of 251 TIR families were identified in the rice genome, which included 186 TIR MULEs and 65 non-TIR MULEs. Among these TIRs, 122 were associated with sequence acquisition (referred to as PMTIR). The copy numbers of Pack-MULEs range from one to 1,002 elements/copies per TIR family (Fig. 2.1, A and C; Supplemental Table S1). The TIR family with the most family members, Os0037, has a total of 1,151 elements, with the majority being Pack-MULEs (87%). Pack-MULEs identified are predominantly of the TIR MULE type (96.2%), suggesting that MULEs with typical long TIRs are more likely to be associated with gene sequence acquisition. This is also true if the abundance of Pack-MULEs is corrected by the total copy number: 23% of the TIR MULEs are Pack-MULEs, while only 6% of the non-TIR MULEs carry gene fragments. Nevertheless, regardless of the MULE type, the composition of Pack-MULEs and non-Pack-MULEs across different MULE TIR families that vary in total copy numbers suggests that the abundance of Pack-MULEs is not correlated to the abundance of the family in the genome (Fig. 2.1, B and D). In other words, TIR families with high copy numbers are not more likely and frequently to acquire gene fragments than families with fewer copies. Meanwhile, 129 TIR

families were devoid of Pack-MULEs (non-PMTIR), comprising a total copy number of 4,953 elements (Supplemental Fig. S1, A and B; Supplemental Table S3).

From the 2,924 Pack-MULEs, 1,557 unique parental genes were identified (Supplemental Table S6). Among the Pack-MULEs, 63 also contain intergenic sequences in addition to genic sequences. In addition, 22 MULE-intergenic elements were found (Supplemental Table S2), and all of them are associated with PMTIR. The intergenic components of the 63 Pack-MULEs and 22 MULE-intergenic elements are derived from a total of 60 intergenic parental sequences, suggesting that MULEs can acquire sequences other than genes, albeit at a much lower frequency. To test whether the dearth of intergenic sequence acquisition is a result of a lower proportion of the genome being the source of this type of parental sequences, we calculated the total genic and intergenic space of the rice genome. The intergenic space (79 Mb) is roughly 68% of the size of the genic space (116 Mb). However, there are about 26 times more genic parentals compared with intergenic parentals, and among Pack-MULEs, even more elements (45 times) have acquired only genes compared with those that acquired both genic and intergenic sequences.

The underrepresentation of intergenic sequences among acquisitions by MULEs suggests that genic sequences are preferentially acquired. Alternatively, this may indicate that, compared with the genic components in Pack-MULEs, the intergenic fragments in Pack-MULEs or MULE-intergenics have less selective advantage, so their retention time is shorter. If the latter was the case, one would expect to see more intergenic sequences among newer acquisition events. To test this, the age of acquisition events was roughly estimated based on the transversion rate (the amount of transversion that has occurred between the alignable length of the acquired sequence and the parental sequence). Sequence transversion rate was chosen, as it is

a better indicator of age compared with either sequence similarity or transition rate. This is because the transition rate is correlated with GC content in addition to age. The median transversion rate of all acquired sequences (genic and nongenic) was used as the cutoff to classify relatively old (transversion rate > 2.75%) and recent (transversion rate ≤ 2.75%) events. The results show no significant difference in the number of intergenic acquisition events between recent and old acquisitions (2.91% versus 2.94%). Thus, a potential lack of selective advantage does not explain the dramatic underrepresentation of intergenic regions inside MULEs.

Structural Differences between Pack-MULEs and Non-Pack-MULEs

Since non-TIR MULEs do not have well-defined inverted terminal regions and only account for a minor portion of the Pack-MULEs, comparisons of structural differences were limited to elements classified as TIR MULEs. A variety of differences were observed when the sequences of Pack-MULEs were compared with those of non-Pack-MULEs. Overall, Pack-MULEs have a much higher GC content compared with non-Pack-MULEs (median, 58.2% versus 36.5%; $P < 2.2 \times 10^{-16}$,Wilcoxon rank-sum test [WRS]) and are much longer (1,445 versus 441 bp; $P < 2.2 \times 10^{-16}$, WRS). To evaluate the GC gradient along the elements, the TIR regions of each element were divided into two equal-sized bins, while the internal regions were divided into 10 equal-sized bins. As shown in Figure 2.2, both TIR and internal sequences of Pack- MULEs are more GC rich than those of non-Pack-MULEs. In addition, a steeper increase in GC content (15% increase) is observed from bin 1 to bin 3 of Pack-MULEs.

Furthermore, properties among previously deemed critical regions for sequence acquisition, TIR and sub-TIR, were compared between Pack-MULEs and non-Pack-MULEs.

Table 2.1. Copy numbers and percentage of different classes of TIR MULEs in the rice genome

Element type	**TIR type**	**Internal Region**		**Copy number**[a]
		Protein match	**Parental copy**	
Pack-MULEs (PM)				2812 (23.21)
PM-genic	PM TIR	known protein	genic sequence	*2755*
PM-plusintergenic	PM TIR	known protein	genic and intergenic sequence	*57*
MULE-intergenic	PM TIR	N/A	intergenic sequence	17 (0.14)
MULE-HypProt	PM TIR/non-PM TIR	hypothetical protein	N/A	1196 (9.87)
MULE-other (non-Pack-MULEs)	PM TIR	N/A	N/A	3695 (30.50)
	nonPM TIR	N/A	N/A	3915 (32.32)
AutoMULEs	PM TIR/non-PM TIR	MULE transposase	N/A	479 (3.95)
Total				12114

[a] Numbers in parenthesis represent percent of total copy number.

Table 2.2. Copy numbers and percentage of different classes of non-TIR MULEs in the rice genome

Element type	**TIR type**	**Internal Region**		**Copy number**[a]
		Protein match	**Parental copy**	
Pack-MULEs (PM)				112 (6.42)
PM-genic	PM TIR	known protein	genic sequence	*106*
PM-plusintergenic	PM TIR	known protein	genic and intergenic sequence	*6*
MULE-intergenic	PM TIR	N/A	intergenic sequence	5 (0.29)
MULE-HypProt	PM TIR/non-PM TIR	hypothetical protein	N/A	119 (6.88)
MULE-other (non-Pack-MULEs)	PM TIR	N/A	N/A	428 (24.54)
	nonPM TIR	N/A	N/A	1038 (59.52)
AutoMULEs	PM TIR/non-PM TIR	MULE transposase	N/A	41 (2.35)
Total				1743

[a] Numbers in parenthesis represent percent of total copy number.

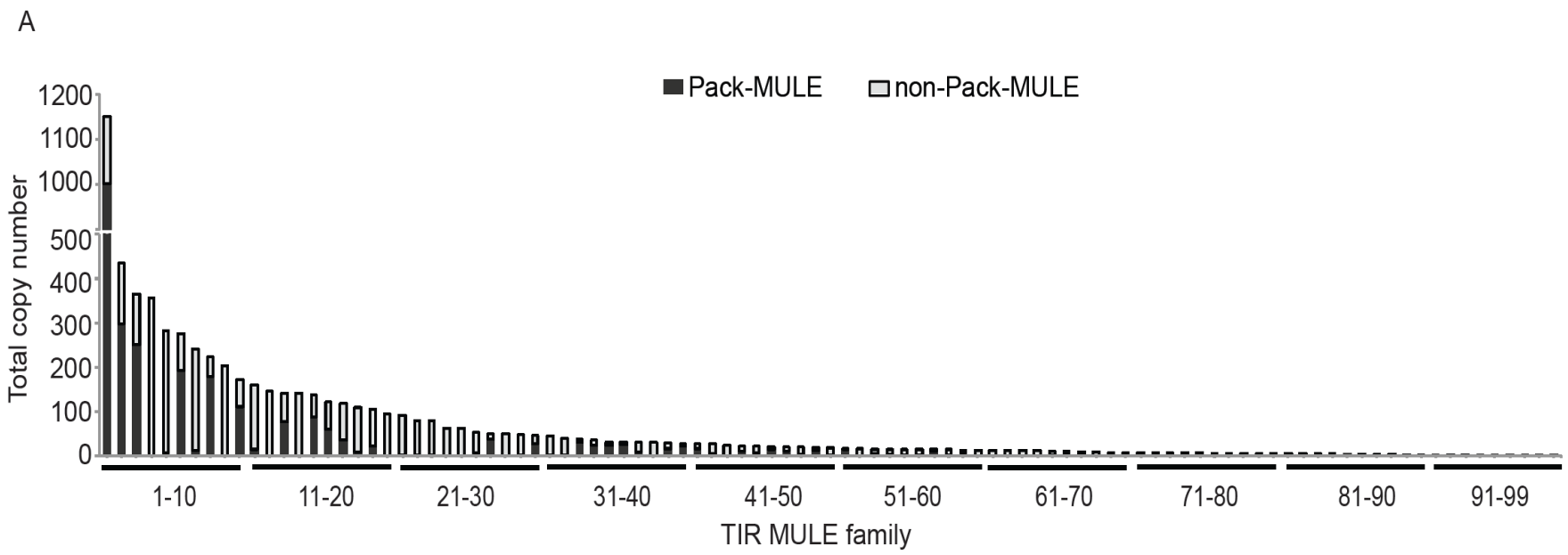


Figure 2.1. Partition of Pack-MULEs and nonPack-MULEs among TIR MULE and non-TIR MULE families in the rice genome. (A) Copy Number and Pack-MULE distribution in TIR-MULE families associated with gene acquisition. (B) Percent Pack-MULEs and non-Pack-MULEs of total copy number for TIR-MULE families associated with gene acquisition. (C) Copy Number and Pack-MULE distribution in non-TIR MULE families associated with gene acquisition. (D) Percent Pack-MULEs and non-Pack-MULEs of total copy number for non-TIR MULE families associated with gene acquisition. The corresponding names of MULE families for A and B are found on Appendix Table 1.

Figure 2.1 (cont'd)

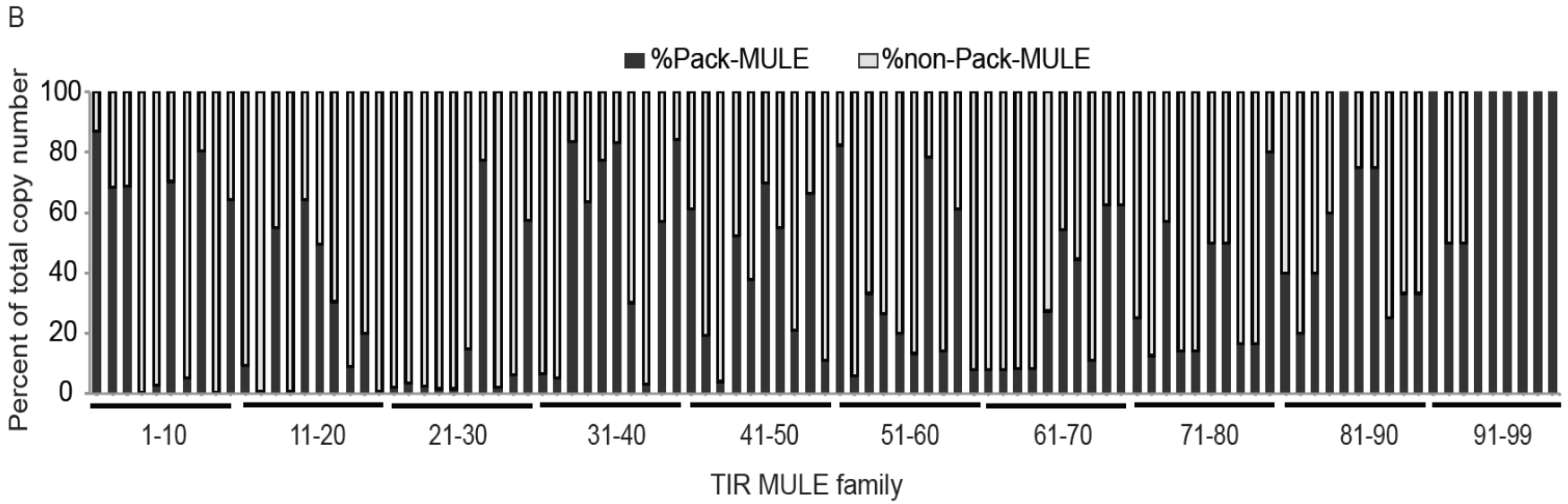
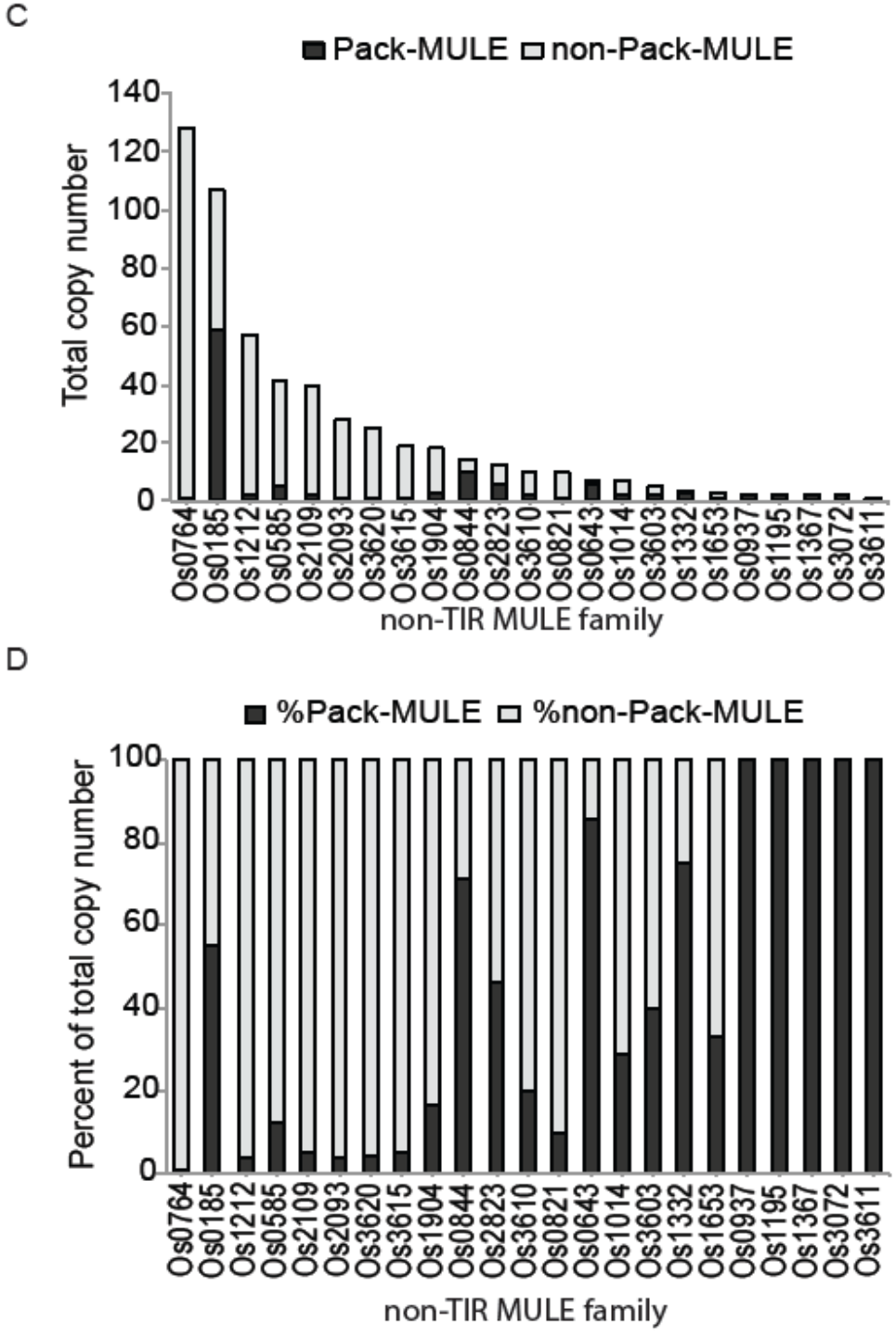


Figure 2.1 (cont'd)



In our analysis, the sub-TIR was defined as the 50-bp sequence adjacent to the TIR. As shown in Figure 2.3, Pack-MULEs have significantly longer TIRs, higher TIR and sub-TIR GC content, and stronger sub-TIR free energy than non-Pack-MULEs ($P < 2.2 \times 10^{-16}$, WRS). If only elements within PMTIR families are considered, there are more non-Pack-MULEs than Pack-MULEs (Table 2.1), yet the TIRs of Pack-MULEs are still longer, with higher GC content in TIR and sub-TIR regions (Fig. 2.3; $P < 2.2 \times 10^{-16}$, WRS). This suggests that longer TIR and higher GC content are not required or favorable for transposition but may be important in sequence acquisition. Alternatively, these differences may also be a product of a positive feedback mechanism through the acquisition of GC-rich sequences that in some cases can be converted as part of the TIR (as in the case for the *Mu4* element mentioned in the Introduction), therefore resulting in longer and more GC-rich TIRs and GC-rich sub-TIRs with stronger free energies. However, analysis of GC content using only the first 100-bp sequence of the Pack-MULE TIRs, a size more similar to the average TIR length of non-Pack-MULEs, shows that even the most terminal end of Pack-MULEs is more GC rich than TIRs of non-Pack-MULEs ($P < 2.2 \times 10^{-16}$, WRS; Fig. 2.3B). Since this region is distal to the internal region, it is unlikely that the higher GC content in this region in Pack-MULEs is a direct or an immediate consequence of acquisition. However, it could be an indirect consequence or result of selection if the higher GC content of TIRs promotes acquisition.

Since previous work has reported the importance of GC content in the acquisition of genic sequences by Pack-MULEs (Jiang et al., 2011), we tested the role of GC content in the acquisition of intergenic sequences by MULEs. Intergenic parental sequences of MULEs are significantly more GC rich than the overall TE and intergenic sequence of the genome ($P = 1.687 \times 10^{-15}$ and $P = 2.59 \times 10^{-12}$, respectively, WRS; Fig. 2.4). Similarly, Pack-MULE parental

genes are significantly more GC rich than the overall genic sequence of the genome ($P < 2.2 \times 10^{-16}$, WRS; Fig. 2.4), suggesting that the preference for GC-rich sequences applies to both genic and nongenic regions.

Underrepresentation of Genes with Unknown Function among Parental Genes

Although Pack-MULEs preferentially acquire GC-rich genes, it is not known whether they also prefer certain classes of genes or if acquisition based on gene function is random. If acquisition is random, we would expect no differences in the ratio of non-TE genes and Pack-MULE parental genes for each functional category. To test this hypothesis, the ratio of non-TE genes and rice parental genes among different GOSlim assignments of biological processes was evaluated using functional assignments and annotations made by the Rice Genome Annotation Group at Michigan State University (Kawahara et al., 2013). A total of 32 biological process categories, which includes " unknown" for genes without an assignment, were compared between Pack-MULE parental genes and non-TE genes. As shown in Figure 2.5, a slight overrepresentation of genes involved in biosynthetic and metabolic processes (X^2 test, $P < 2.2 \times 10^{-16}$) and a strong bias against genes with unknown classification among parental genes (X^2 test, $P < 2.2 \times 10^{-16}$) were observed. This slight preference for a few categories dissipates, however, when the unknown category is excluded from the analysis (Supplemental Fig. S2). A comparison was also conducted in maize and Arabidopsis to determine whether such a bias against genes with unknown function exists in other plant species where Pack-MULEs have been characterized. In both species, a significant underrepresentation of genes with unknown function among Pack-MULE parental genes was also found (X^2 test, maize, $P < 2.2 \times 10^{-16}$, Arabidopsis, $P = 0.03$; Tables 2.3 and 2.4).

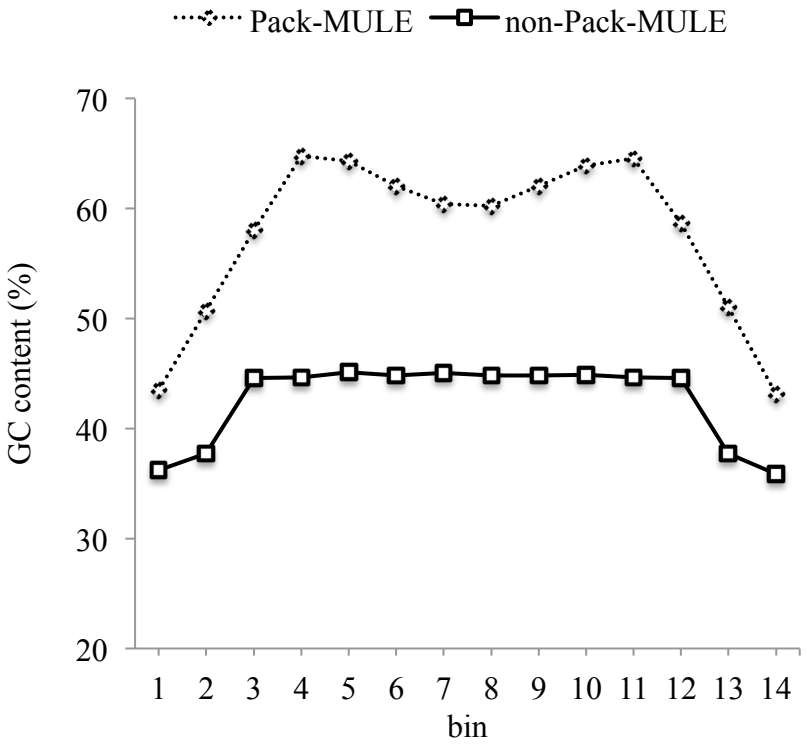


Figure 2.2. GC content along Pack-MULEs and non-Pack-MULEs. The first 2 and last 2 bins represent TIR regions and the internal sequence was divided into 10 equal-sized bins prior to determination of GC content per bin.

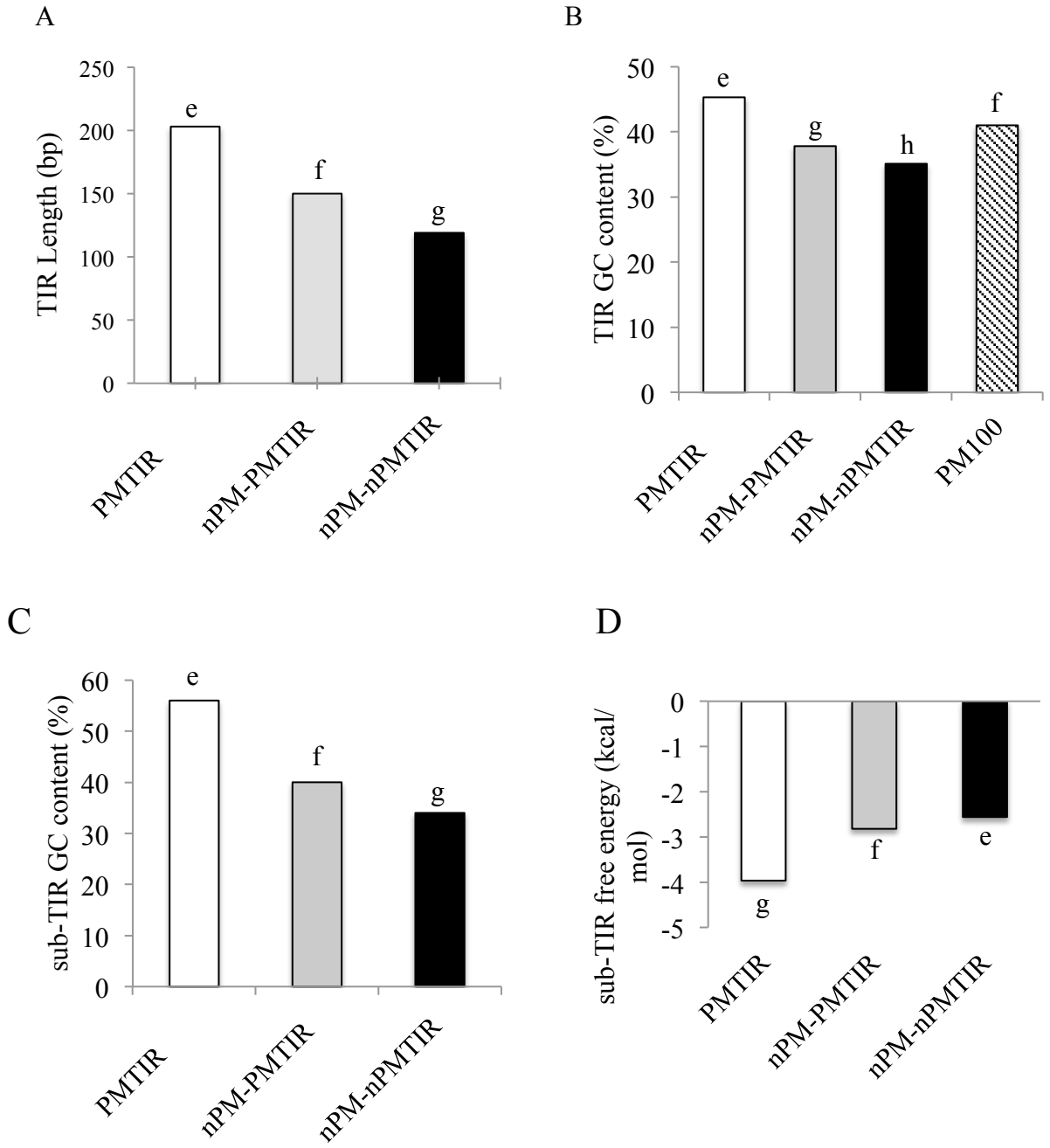


Figure 2.3. Structural difference between Pack-MULEs and non-Pack-MULEs based on TIR and sub-terminal (sub-TIR) sequences. (A) Median TIR length. (B) Median TIR GC content. (C) Median sub-TIR GC content. (D) Median sub-TIR free energy. PM: Pack-MULE; nPM-PMTIR: non-Pack-MULEs with PM associated TIRs; nPM-nPMTIRs: non-Pack-MULEs with non-Pack-MULE exclusive TIRs; PM100: using only first 100 bp sequence of Pack-MULE TIRs. Bars designated with different letters indicate their values are significantly different (α=0.008 for B and α=0.02 for A, C and D) by Wilcoxon Rank Sum Test (WRS) with Bonferroni correction.

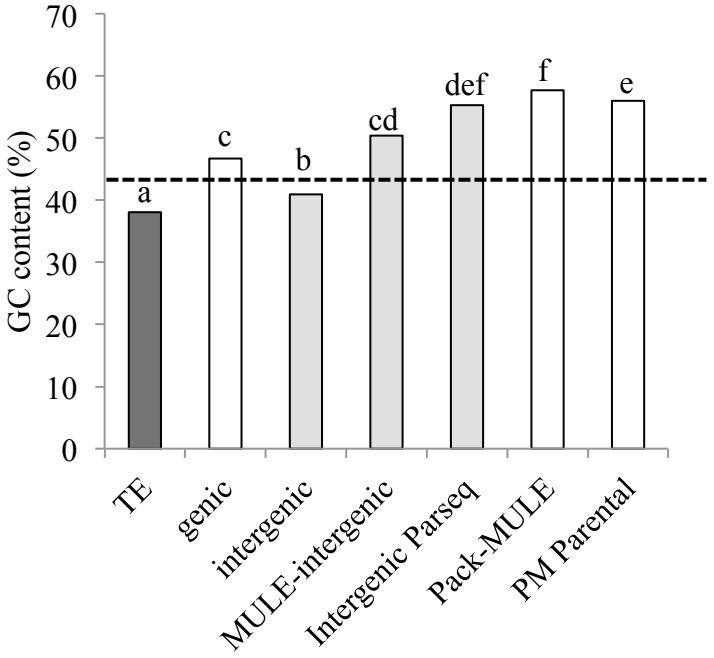


Figure 2.4. GC content of different genomic sequences in the rice genome. Genome average GC content is indicated by a dashed line. Bars designated with different letters indicate their values are significantly different (α= 0.002) by Wilcoxon Rank Sum Test (WRS) with Bonferroni correction.

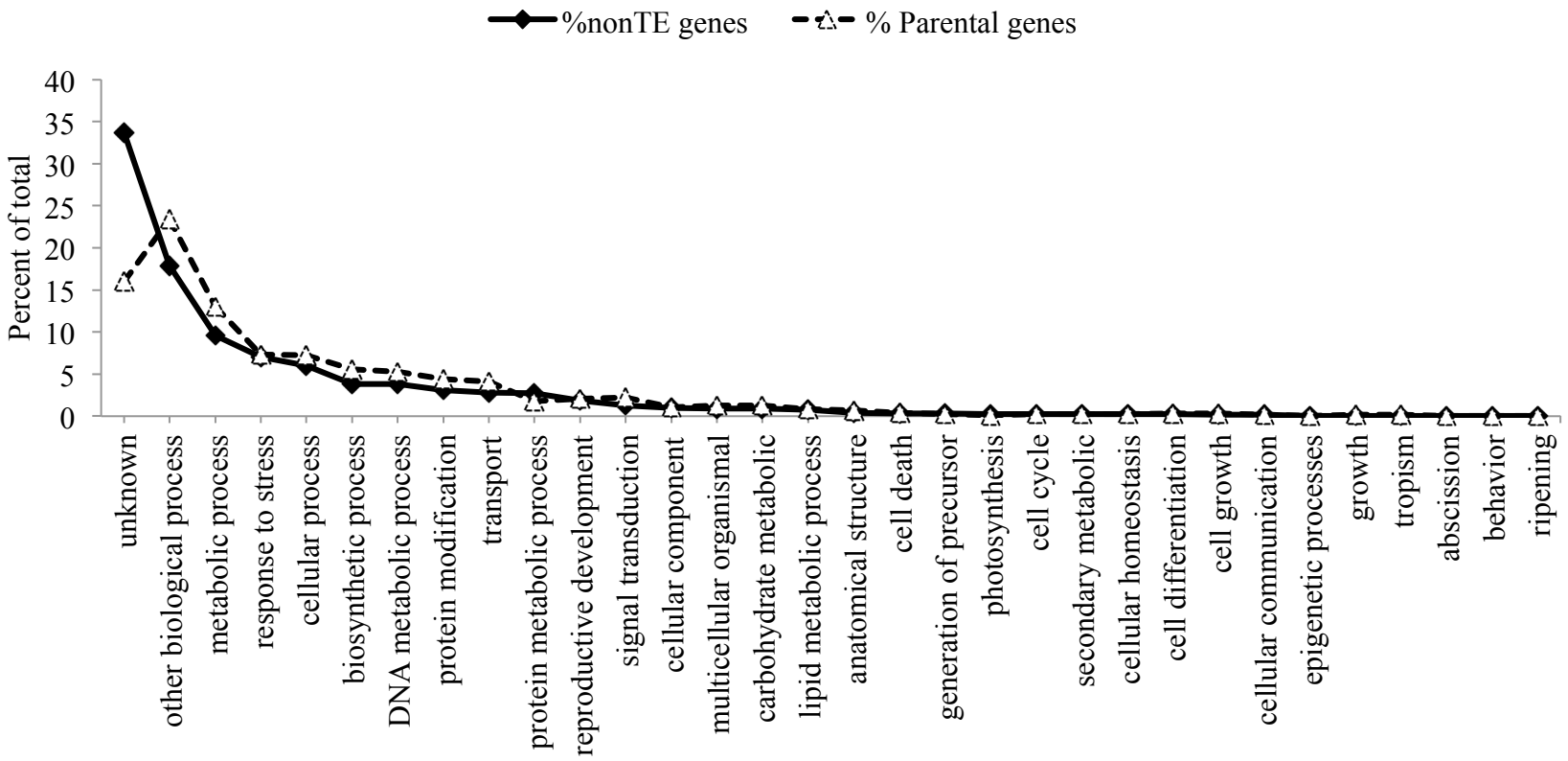


Figure 2.5. Percent of GOSlim categories of Pack-MULE parental genes and all non-TE genes of rice.

To understand the mechanism underlying the apparent bias against genes without a known function, we compared the GC content among genes with and without a GOSlim assignment, since it is known that GC richness is favored in sequence acquisition/retention. Among non-TE genes, those without a GOSlim assignment have significantly higher GC content than counterparts with a GOSlim assignment both at the genomic and coding sequence (CDS) levels (genomic, $P < 2.2 \times 10^{-16}$; CDS, $P < 2.2 \times 10^{-16}$, WRS; Table 2.5). Among Pack-MULE parental genes, the GC content difference between GOSlim genes and unknown genes was detectable at the genomic sequence level ($P = 0.001$, WRS; Table 2.5) but not significant at the CDS level. In all four comparisons, genes with unknown category have higher or comparable GC content than those with assigned function(s). These results suggest that the gene GC content does not explain the underrepresentation of genes with unknown biological function among Pack-MULE parental genes in rice. Similarly, maize non-TE genes with unknown function have significantly higher GC content than those with known function (genomic, $P < 2.2 \times 10^{-16}$; CDS, $P = 5.589 \times 10^{-16}$, WRS; Table 2.3). In Arabidopsis, genes with unknown function had significantly higher GC content than genes with known function only at the genomic level ($P = 0.01$; Table 2.4). These data show that acquisition bias against genes with unknown function is not species specific and supports the notion that GC content does not explain this finding.

Some of the genes with unknown function might be the result of misannotation. Thus, it is feasible that sequences misannotated as genes are overrepresented within the unknown group and that the apparent bias against them indicates a preference for bona fide genes. To test this, we surveyed the distribution of parental genes among non-TE genes with and without an ortholog (Schnable et al., 2009; Lin et al., 2010; Davidson et al., 2012), since genes with orthologs are more likely bona fide genes. For both rice and maize, genes with orthologs are

Table 2.3. GC content and expression information among Pack-MULE parental genes and non-TE genes in maize according to functional assignment.

Gene	Total	%GC-genomic	%GC-CDS	FPKM*	# of library*	no expression	% no expression
nonTEgene-unknown	13057	50.10b	56.50b	39.60a	11.00a	4196	32.1d
nonTEgene-known	26599	47.50a	55.90a	70.14b	12.00b	2873	10.8b
PMPar- unknown	47	57.20c	63.60ac	45.43b	12.00b	2	4.3a
PMPar- known	188	55.45c	63.50c	62.71b	12.00b	7	3.7a

*Median determined only from genes that are expressed.

PMPar – Pack-MULE parental gene; Numbers in each column followed by different letters are significantly different (α=0.008 with Bonferroni adjustment)

Table 2.4. GC content and expression information among Pack-MULE parental genes and non-TE genes in *Arabidopsis* according to functional assignment.

Gene	Total	%GC-genomic	%GC-CDS	TPKM*	# of library*	no expression	% no expression
nonTEgene-unknown	10239	39.50b	43.50a	33.83a	5.00a	4584	44.8d
nonTEgene-known	17147	39.30a	44.40b	54.33b	6.00b	3088	18.0c
PMPar- unknown	7	41.80ab	43.20ab	46.92ab	2.50ab	1	14.3b
PMPar- known	28	39.80ab	44.65ab	41.80ab	6.00ab	2	7.1a

*Median determined only from genes that are expressed.
PMPar – Pack-MULE parental gene
Numbers in each column followed by different letters are significantly different (α=0.008 with Bonferroni adjustment)

Table 2.5. GC content and expression information among Pack-MULE parental genes and non-TE genes in rice according to GOSlim assignment

Gene	Total	%GC-genomic	%GC-CDS	FPKM*	# of library*	no expression	% no expression
nonTEgene-NoSLIM	12010	51.80b	59.50b	5.81a	6.0a	6541	54.46d
nonTEgene-WithSLIM	23609	45.20a	55.00a	9.98b	8.0c	3671	15.55b
PMPar-NoSLIM	250	59.00d	68.30c	6.79a	7.0ab	60	24.29c
PMPar-WithSLIM	1334	56.40c	67.30c	9.41b	7.0b	145	11.10a

*Median determined only from genes that are expressed.
(PMPar – Pack-MULE parental gene; NoSlim – genes without a GOSlim assignment; WithSlim – genes with a GOSlim assignment
Numbers in each column followed by different letters are significantly different (α=0.008 with Bonferroni adjustment)

significantly enriched among Pack-MULE parental genes (X^2 test, rice, $P < 2.2 \times 10^{-16}$; maize, $P = 4.152 \times 10^{-14}$; Supplemental Table S7). For Arabidopsis, an enrichment is observed among parental genes (91% have orthologs; Supplemental Table S7), yet this overrepresentation is not statistically significant, most likely due to the low number of parental genes.

The Effect of Gene Expression on Sequence Acquisition and Its Interaction with GC Content

Since GC content does not explain the discrepancy in gene acquisition preference mentioned above, other factors that may influence sequence acquisition were explored. The role of gene expression in rice was tested using RNA-Seq data (Michigan State University Rice Genome Annotation Group) from 10 different rice developmental stages encompassing diverse vegetative and reproductive tissues (Davidson et al., 2012). A gene was considered expressed if the fragments per kilobase of exon per one million fragments mapped (FPKM) value was 1.0 or greater in at least one expression library; otherwise, the gene was categorized as not expressed. Over one-half (54%) of non-TE genes without a GOSlim assignment were not expressed, while only 15% of non-TE genes with known functions were not expressed (Table 2.5). Meanwhile only 24% of Pack-MULE parental genes without a GOSlim assignment and even fewer, 11%, of those with GOSlim assignments were not expressed, suggesting that gene expression may play a role in the preference for acquisition. The role of gene expression was also evaluated in maize and Arabidopsis. Maize RNA-Seq expression data were obtained from a previous study (Davidson et al., 2011), and expression was determined using parameters similar to rice. Since similarly comprehensive RNA-Seq expression data generated from a single experiment were not readily available in Arabidopsis, we utilized the massively parallel signature sequencing data set with expression levels of genes from multiple tissues (Meyers et al., 2004a, 2004b). Using only

uniquely mapping signatures, a gene was considered expressed if the transcripts per one million (TPM) value was 5.0 or greater in at least one expression library. In both species, significantly more genes with unknown function were not expressed compared with genes with known function (X^2 test, rice, $P < 2.2 \times 10^{-16}$; Tables 2.3 and 2.4). However, the number of nonexpressed genes from either category was much lower among parental genes than the genomic average, suggesting that the underrepresentation of genes with unknown function among parental genes is connected to the lack of expression of unknown genes in all three species. To further assess the role of gene expression, the level of expression, determined by the FPKM/TPM value, and the number of tissues with expression were compared among different groups of genes. In all three species, expressed genes with unknown function have significantly lower expression levels ($P < 2.2 \times 10^{-16}$, WRS) and fewer tissues with detectable expression ($P < 2.2 \times 10^{-16}$, WRS) than those with a GOSlim assignment (Tables 2.3-2.5). The expression levels of parental genes of Pack-MULEs do not significantly differ from the genomic average, with the exception of the maize parental genes with unknown function, which showed a significantly higher expression level than the genomic average ($P = 5.923 \times 10^{-6}$, WRS). Interestingly, parental genes with or without a known function were expressed in similar numbers of tissues in all three species (Tables 2.3-2.5), suggesting that the breadth of expression is critical to sequence acquisition. Thus, the high percentage of genes with no expression and genes with less ubiquitous expression explains why genes without a GOSlim assignment are underrepresented in the parental genes of Pack-MULEs.

To determine whether the roles of GC content and the breadth of gene expression on sequence acquisition are independent, we categorized rice genes into different GC content groups (low, moderate, and high) as well as different expression categories (no/low, moderate, and high) based on the number of tissues with detectable expression (FPKM $\geq$ 1.0) and

determined the proportion of Pack-MULE parental genes within each group. Although the results above suggest that expression plays a similar role in acquisition preference in maize and Arabidopsis, the analysis in this section was limited to the rice data, due to the much lower number of parental genes in the other two species. Our results in rice indicate that both GC content and the number of tissues with expression evidence play a role in sequence acquisition preference by Pack-MULEs. The ratio of parental genes among non-TE genes was used to reflect the acquisition frequency (how frequently a certain group of genes was acquired). As shown in Figure 2.6A, for low GC genes and moderate GC genes, a very minimal and modest increase in the proportion of parental genes, respectively, may be observed, with increase in the number of tissues with expression. In comparison, among high GC genes, a stronger increase in the ratio of parental genes occurs with more expression libraries. Meanwhile, when genes are categorized according to the number of tissues with expression, a more substantial increase in the ratio of parental genes is observed in all three expression groups as GC levels increase, and the increase is much greater among genes expressed in eight to 10 tissues (Fig. 2.6B). It is clear, however, that GC content plays a more dominant role than gene expression for sequence acquisition/retention. This is because variation of GC content may lead to as much as an 11-fold change in the percentage of parental genes, while that for gene expression is only 2- to 5-fold. In addition, the increase in GC content is accompanied by a boost in the percentage of parental genes, despite their expression patterns. In contrast, the effect of gene expression on the percentage of parental genes is only substantial when the genes have moderate or high GC content (Fig. 2.6A). It is also interesting that the effect of gene expression plateaued with expression in seven or more tissues (Fig. 2.6A). That explains why the median value of the number of tissues (seven tissues) with expression for parental genes with known function is

slightly lower than the genomic average (eight tissues; Table 2.5) in rice, because more ubiquitous expression (in more than seven tissues) does not confer additional advantage for acquisition.

The Enrichment of GC-Rich Sequences inside Pack-MULEs Is Due to Selective Acquisition and Preferential Retention

The apparent preference for higher GC content and relatively ubiquitous expression in sequence acquisition in rice, however, can be an artifact of selection, since sequences with higher GC content and more ubiquitous expression are more likely derived from coding regions and, thus, are more likely to be functional. If that is the case, one would expect the preference to be more dramatic among old than among recent acquisition events. Again, the age of acquisition events was roughly estimated through the transversion rate between the acquired sequence and the parental gene. Parental genes were separated into two groups: recent acquisitions, those with transversion rate of 2.75% or less; and old acquisitions, those with transversion rate of 2.75% or greater. As shown in Figure 2.7A, the two groups of parental genes show an overall similar percentage with increasing number of tissue expression, suggesting that the number of expressed tissues does not have a significant influence on the retention of their gene fragments. In contrast, there are significantly more parental genes in old acquisition events (transversion rate > 2.75%) compared with recent events among genes with a GC content of 69% to 82% (Fig. 2.7B), suggesting that selection may play a role in the apparent enrichment of parental genes with extremely high GC content. To further characterize the impact of gene GC content on the retention of the relevant gene fragments, we tested the relationship of GC content and transversion rate of all rice parental genes and found a low, albeit statistically significant, correlation (0.09; $P = 0.0003$, Spearman; Fig. 2.7C); that is, the GC content of parental genes

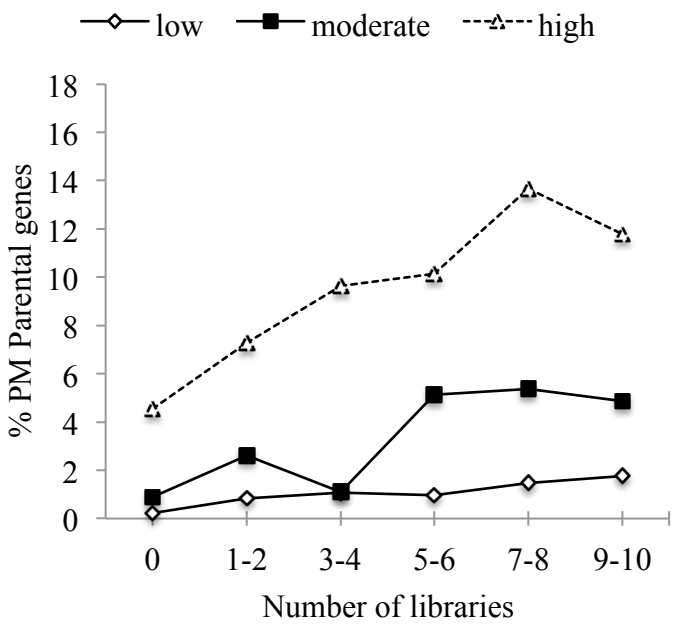
progressively increases with transversion rate between Pack-MULEs and the parental genes. Again, this indicates that selection plays a role in the retention of GC-rich genes.

To obtain the best possible assessment of the GC content of parental genes upon acquisition, we calculated the GC content of the 14 parental genes with a 0% transversion rate. Theoretically, these sequences represent the most recent acquisition events and have been subjected to little selection. The GC content of all of them is higher than 50%, and the average value is 66.1%, which is dramatically higher than the genome average GC content (45.6%) of non-TE genes. This fact, together with the minor increment of GC content of parental genes over evolutionary time (Fig. 2.7C), suggests a strong preference for GC-rich genes upon acquisition. Taken together, our results suggest that selection may play a role in the retention of fragments from different parental genes, although it is insufficient to fully explain the enrichment of GC-rich genes among parental genes of Pack-MULEs.

Discussion

The process of gene duplication followed by sequence and functional divergence (neofunctionalization) is one of the most important means for the generation of new genes (Flagel and Wendel, 2009). Studies have shown that all major families of TEs are involved in gene duplication in plants (Jiang et al., 2004; Kawasaki and Nitasaka, 2004; Morgante et al., 2005; Zabala and Vodkin, 2005; Wang et al., 2006; Schnable et al., 2009). In the rice genome, over 1,500 parental genes have been transduced by Pack-MULEs, which can generate independent or chimeric transcripts when fused with nearby sequences (Jiang et al., 2011). In addition, these transcripts may regulate parental gene expression, suggesting a very important role of Pack-MULEs in novel gene formation and evolution. It was shown previously that Pack-

A



B

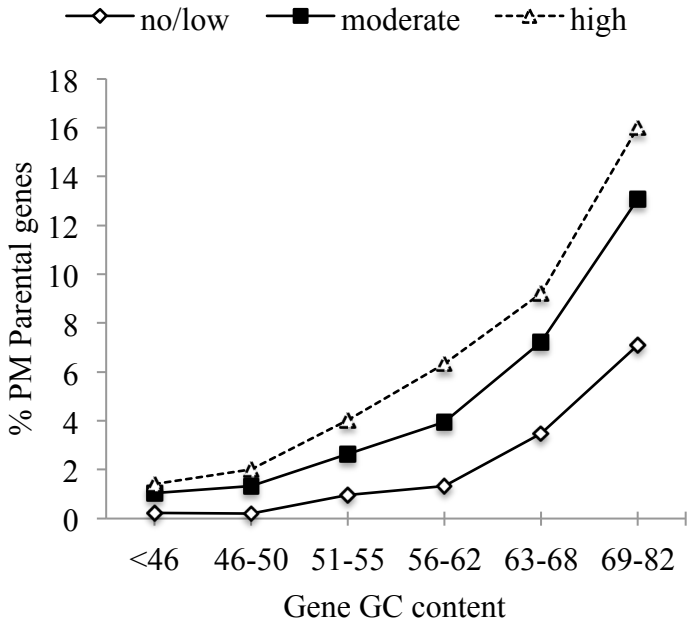


Figure 2.6. The effect of GC content and expression in gene acquisition frequency. (A) The relationship between expression breadth and ratio of parental genes among genes grouped on GC content range (low: 30-50% GC; moderate: 51-62% GC; high: 63-81% GC); (B) The relationship between gene GC content and ratio of parental genes among genes grouped on number of tissue expression range (no/low: 0 to 1 libraries; moderate: 2 to 7 libraries; high: 8 to 10 libraries).

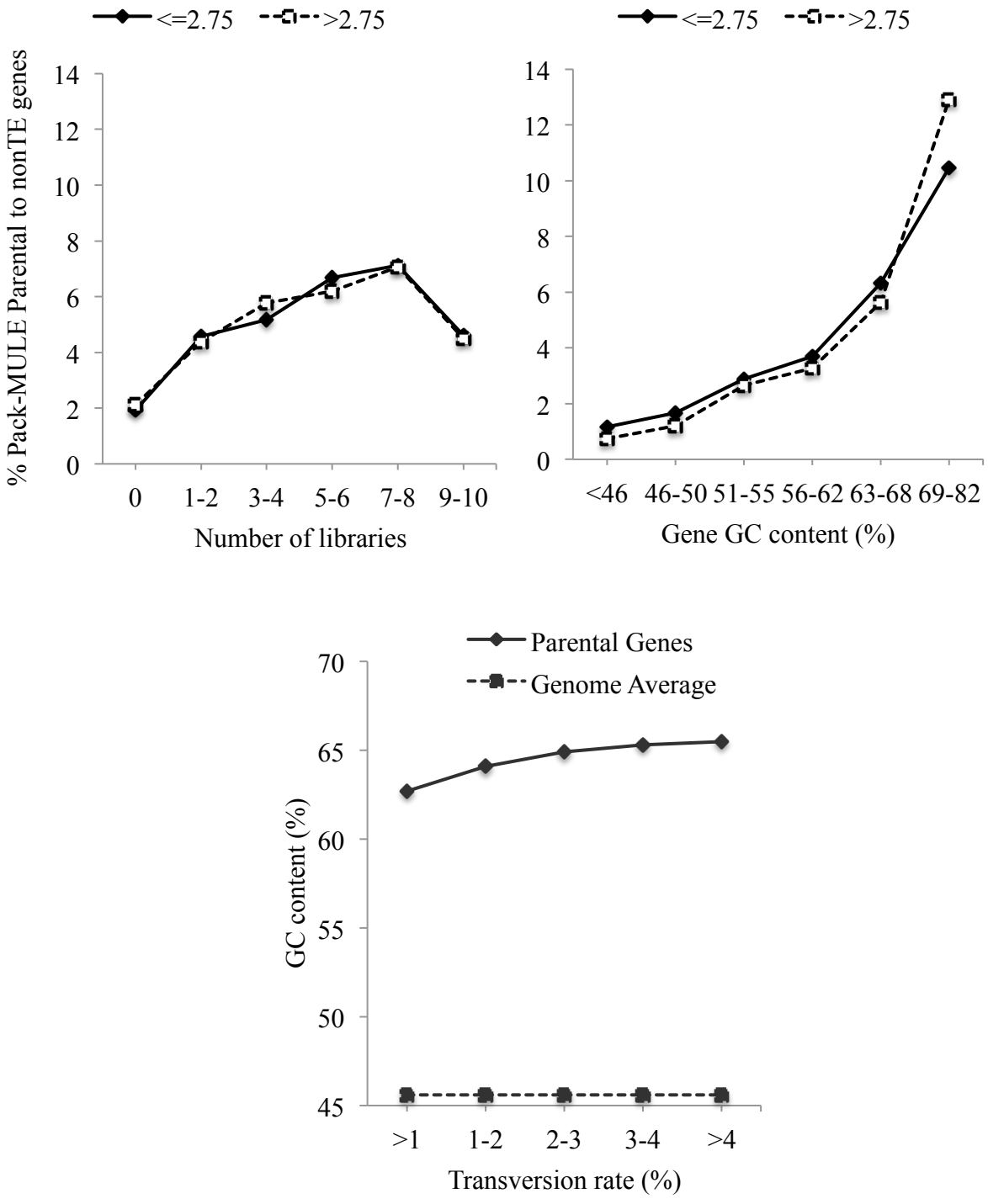


Figure 2.7. Comparison of GC content and breadth of expression of parental genes between recent and old acquisition events estimated by transversion rate. (A) The effect of acquisition age on breadth of gene expression; (B) The effect of acquisition age on the GC content; (C) The relationship between GC content of parental genes and transversion rate.

MULEs preferentially acquire GC-rich sequences, a phenomenon only seen in grasses. Aside from that, the process by which these sequences are selected and captured by Pack-MULEs remains largely an enigma.

Our findings in this study show that rice TIR MULEs have a higher propensity to acquire genomic sequences compared with non-TIR MULEs, and this bias may be related to differences in structural properties such as TIR length and TIR GC content. It remains unclear at this stage whether the capacity of sequence acquisition among TIR MULEs contributes to the overall success of TIR MULEs versus non-TIR MULEs, since 87% of the rice MULEs belong to TIR MULEs. MULEs and Pack-MULEs in rice are capable of acquiring both genic and intergenic sequences, although the acquisition preference for genic sequences is much more pronounced compared with intergenic sequences (Table 2.1). This may suggest that Pack-MULEs are either more competent to acquire genes or that genes are more readily acquired by Pack-MULEs over other sequences in the genome. Consistent with previous work, GC content was a factor in the preferential acquisition or retention of intergenic fragments, because the GC content of the intergenic sequences inside MULEs or Pack-MULEs is much higher than the genomic, TE, and intergenic GC contents. Interestingly, the GC-rich internal sequences of Pack-MULEs are accompanied by higher GC content of TIRs and sub-TIRs in Pack-MULEs compared with that of non-Pack-MULEs. One of the models for sequence acquisition suggests the formation of a cruciform structure during the process, with the TIRs forming the stem of the hairpin (Bennetzen and Springer, 1994). In this model, an endonucleolytic attack occurs in the single-stranded loop, aided by sequences containing homology to a parental sequence, and initiates repair through illegitimate recombination. Our results are consistent with this model in the following respects. On the one hand, the GC-rich internal regions and GC-rich sub-TIRs seem to imply that

sequence homology between the element and the acquisition target likely plays a role in acquisition. If this is the case, one would expect AT-rich sequences to be acquired as well if the sub-TIR sequence is also AT-rich. This was not observed, since non-Pack-MULEs have relatively AT-rich sub-TIRs (Fig. 2.2) but they do not carry any recognizable genomic sequences. This is possibly because the pairing of AT-rich sequences is not as stable as that of GC-rich sequences to initialize the repair process. On the other hand, a long GC-rich TIR would lead to a more stable cruciform that may facilitate the acquisition process. This hypothesis may also explain why MULEs are more frequently associated with sequence acquisitions than other "cut-and-paste" DNA transposons, in that most MULEs have extended long TIRs.

Although a particular functional category of genes does not seem to be more preferentially acquired by Pack-MULEs, a bias against genes with unknown function is obvious, and this was not species specific. Hypothetical genes and genes with unknown function can often result from misannotation. These genes, in most cases, are generated by gene prediction programs and, therefore, may lack supporting expression data. As a result, it is conceivable that there are more false positive annotations within this group compared with genes with known function. Such bias may reflect the preference of Pack-MULEs for bona fide genes. In other words, Pack-MULEs might be better than gene annotation programs in distinguishing genuine genes from other sequences. The underrepresentation of genes without a known biological function prompted the analysis of expression among annotated genes, which showed that a relatively ubiquitous expression throughout development may play a role in sequence acquisition: Pack-MULEs preferentially acquire genes that are expressed in multiple tissues/developmental stages (Fig. 2.6A). Our analysis also shows an enrichment of genes with orthologs among parental genes (Supplemental Table S7). More importantly, our data explain the

strong propensity of Pack-MULEs to transduce genes over other nongenic sequences in the genome.

Gene GC content and the ubiquity of expression show an additive effect on preferential selection and acquisition by Pack-MULEs. Interestingly, it appears that selection acts differentially on GC content and the ubiquity of expression. The preference for expression ubiquity seems to be largely at the acquisition level, since it is evenly distributed among old and recent acquisition events (Fig. 2.7A). In contrast, there is a detectable level of selection that favors the retention of fragments from highly GC-rich genes (Fig. 2.7, B and C). Such differentiation is understandable from a mechanistic point of view: once a gene fragment is acquired by a Pack-MULE, the characteristic of expression is no longer associated with the fragment. Since the acquired fragments may not be expressed in the same pattern as their parental copies, there is no basis for selection for or against the expression pattern of the parental genes. This is consistent with the fact that Pack-MULEs are often associated with different tissue specificity from their parental genes (Hanada et al., 2009). On the other hand, GC content is always associated with the fragment. High GC content could induce a series of genetic and epigenetic changes in the genome. Genetically, it may modify the 5' end of the adjacent genes and intensify the negative GC gradient (Jiang et al., 2011). Epigenetically, GC-rich sequences offer more methylation targets that could influence the chromatin structure and expression of the nearby genes (Kalisz and Purugganan, 2004; Tatarinova et al., 2010; see below). All these features may form the basis for selection. Apparently, the high GC content is favored here, which may imply that it has provided certain benefits for the organism. Despite the possible selection for high GC content over evolutionary time, the degree of selection seems too moderate to explain the dramatic difference in the GC content between parental genes and all non-TE genes

(Fig. 2.7C). Accordingly, it is likely that the preferential acquisition by Pack-MULEs for GC-rich genes is also responsible, or more important, for the enrichment of GC-rich genes among parental genes. In addition, we cannot rule out the possibility that variation in GC content over evolutionary time is due to a change in acquisition preference. This could occur, for example, when different MULE families have slightly different acquisition preferences and their amplification rate has not been constant in each time range. Future computational and biochemical analyses are required to test whether acquisition preference varies among different MULE families.

To our knowledge, our study is the first to elucidate the direct involvement of gene expression in sequence duplication by DNA transposons. Furthermore, our data suggest that GC richness offers a more dominant effect in this process. This is because, as discussed above, high GC content is very likely favored by both acquisition and selection. Studies to determine the relationship between GC content and expression level and the breadth of expression are conflicting, with studies reporting a strong positive correlation (Lercher et al., 2003; Kudla et al., 2006) and those reporting weak or unclear correlation (Gilbert et al., 2004; Sémon et al., 2005). Nevertheless, these and other studies established the association of GC-rich sequences with open chromatin (Vinogradov, 2003). The open chromatin provided by GC-rich sequences potentially allows these sequences to be more accessible by host enzymes during interrupted gap repair (model 2; see the introduction) or during internal strand repair of cruciform structures (model 1; see the Introduction). Expression level, as measured by FPKM/TPM values, does not appear critical to the likelihood of a gene being transduced by a Pack-MULE (Tables 2.3-2.5). Nevertheless, we cannot rule out the possibility that parental genes were expressed at a level higher than average prior to the formation of relevant Pack-MULEs. This is because Pack-

MULEs have a negative regulatory effect on the expression level of parental genes, which may render the difference no longer detectable after the acquisition. In contrast, preferential acquisition is positively correlated with the breadth of expression when the genes are expressed in seven or fewer tissues (Fig. 2.6). The open chromatin configuration during active transcription may allow access to a sequence for duplication. Consequently, the greater number of tissues with detectable expression allows the gene greater chances of being transduplicated by Pack-MULEs.

Conclusion

The unprecedented copy number of Pack-MULEs, the massive duplication of thousands of genes in the rice genome, combined with their biased acquisition for GC-rich genes and insertion in 5' regions of genes suggest an evolutionary importance of these elements in gene evolution and regulation. Our findings in rice show that sequence acquisition by Pack-MULEs relies on structural/sequential properties of the elements and the acquisition targets. TIR MULEs are the predominant MULE type in the rice genome and account for the majority of the Pack-MULEs. Although Pack-MULEs can duplicate both genic and intergenic sequences, a much stronger preference for genic sequences exists. Pack-MULEs exhibit a non-species-specific bias against genes with unknown function and enrichment of parental genes with orthologs, suggesting its preferential acquisition for bona fide genes. Structural properties of elements, GC content, and the breadth of expression of parental genes influence the selection and acquisition of sequences. Increased GC content and number of tissues with detectable expression results in a higher likelihood of a gene being acquired by a Pack-MULE. Moreover, GC-rich sequences acquired by Pack-MULEs are preferentially retained compared with sequences that are not so GC rich. Although the molecular mechanism for how Pack-MULEs locate and duplicate intergenic and genic sequences remains to be empirically evaluated, our study demonstrates that

the activity of Pack-MULEs leads to the selective duplication/retention of CDSs, because CDSs are more GC rich and have a wider breadth of tissue expression. Such selection enables them to carry the most likely functional sequences instead of "junk" and so provide new resources for the evolution of new genes or the modification of existing gene.

APPENDIX

Table A1. Order of TIR MULE families found in Figure 2.1 A and B.

Bar number	TIR name
1	Os0037
2	Os2580
3	Os0166
4	Os0086
5	Os0243
6	Os0205
7	Os0229
8	Os0949
9	Os0105
10	Os0284
11	Os0335
12	Os0301
13	Os2537
14	Os0312
15	Os0053
16	Os2088
17	Os1129
18	Os0297
19	Os0222
20	Os0372
21	Os0378
22	Os0138
23	Os0182
24	Os0230
25	Os0319
26	Os0115
27	Os2471
28	Os0385
29	Os0219
30	Os0202
31	Os3333
32	Os3372
33	Os1693
34	Os1617
35	Os0709
36	Os0513
37	Os2788
38	Os0547
39	Os2159

Table A1 (cont'd)

40	Os0208
41	Os1121
42	Os2701
43	Os3609
44	Os2269
45	Os0116
46	Os3617
47	Os0610
48	Os0853
49	Os1224
50	Os3284
51	Os1272
52	Os2766
53	Os3366
54	Os3625
55	Os3605
56	Os3621
57	Os0874
58	Os1886
59	Os1104
60	Os1308
61	Os0584
62	Os3608
63	Os0454
64	Os2835
65	Os0570
66	Os0892
67	Os3614
68	Os1015
69	Os1810
70	Os3624
71	Os3375
72	Os1057
73	Os3369
74	Os2033
75	Os3604
76	Os3602
77	Os3618
78	Os1455
79	Os3601

Table A1 (cont'd)

80	Os3613
81	Os3371
82	Os3622
83	Os1404
84	Os2398
85	Os3616
86	Os3337
87	Os3612
88	Os2623
89	Os1838
90	Os3355
91	Os3383
92	Os1228
93	Os3619
94	Os3238
95	Os3600
96	Os3606
97	Os3623
98	Os3626
99	Os3627

All supplemental tables and figures are available at:

http://www.plantphysiol.org/content/early/2013/09/12/pp.113.223271.short

Figure S1. Copy number of MULE TIR families not associated with gene acquisition. (A) TIR-MULE families, (B) non-TIR MULE families

Figure S2. Percent of GOSlim categories of Pack-MULE parental genes and non-TE genes of rice excluding genes without a functional assignment.

Table S1. List of rice Pack-MULEs.

Table S2. List of rice MULEs with intergenic sequence acquisition (MULE-intergenic).

Table S3. List of rice non-Pack-MULEs.

Table S4. List of rice autonomous MULEs.

Table S5. List of rice MULE-HypProt.

Table S6. List of parental genes of Pack-MULEs.

Table S7. Distribution of non-TE and parental genes among genes with and without orthologs

Sequence-File S1. Terminal sequences of TIR-MULE families.

Sequence-File. S2. Terminal sequences of non-TIR-MULE families.

Sequence-File S3. Element sequences of small MULEs (<150 bp).

Sequence-File S4. Sequences of intergenic parental copies.

REFERENCES

REFERENCES

Ammiraju JS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, Walling JG, Ma J, Talag J, Brar DS, et al (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. Plant J 52: 342–351

Bao, Z and Eddy, SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 12:1269-1276

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

Bennetzen JL (2007) Patterns in grass genome evolution. Curr Opin Plant Biol 10: 176–181

Bennetzen JL, Kellogg EA (1997) Do plants have a one-way ticket to genomic obesity? Plant Cell 9: 1509–1514

Bennetzen JL, Springer PS (1994) The generation of *Mutator* transposable element subfamilies in maize. Theor Appl Genet 87: 657–667

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

Chomet P, Lisch D, Hardeman KJ, Chandler VL, Freeling M (1991) Identification of a regulatory transposon that controls the *Mutator* transposable element system in maize. Genetics 129: 261–270

Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu SH, Jiang N, Buell CR (2012) Comparative transcriptomics of three *Poaceae* species reveals patterns of gene expression evolution. Plant J 71: 492–502

Davidson RM, Hansey CN, Gowda M, Childs K, Lin H, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Jiang N, et al (2011) Utility of RNA-seq for analysis of maize reproductive transcriptomes. Plant Genome 4: 191–203

Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL (2005) Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. Proc Natl Acad Sci USA 102: 19243–19248

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797

Ferguson AA, Jiang N (2012) *Mutator*-like elements with multiple long terminal inverted repeats in plants. Comp Funct Genomics 2012: 695827

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9: 397–405

Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants. New Phytol 183: 557–564

Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell 118: 555–566

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu SH, Jiang N (2009) The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile. Plant Cell 21: 25–38

Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, Cowan RK, Bureau TE (2006) Transposon-mediated expansion and diversification of a family of ULP-like genes. Mol Biol Evol 23: 1254–1268

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR (2006) The transposable element landscape of the model legume *Lotus japonicus*. Genetics 174: 2215–2228

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

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573

Jiang N, Ferguson AA, Slotkin RK, Lisch D (2011) Pack-*Mutator*-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proc Natl Acad Sci USA 108: 1537–1542

Kalisz S, Purugganan MD (2004) Epialleles via DNA methylation: consequences for plant evolution. Trends Ecol Evol 19: 309–314

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, et al (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice 6: 4

Kawasaki S, Nitasaka E (2004) Characterization of *Tpn1* family in the Japanese morning glory: *En/Spm*-related transposable elements capturing host genes. Plant Cell Physiol 45: 933–944

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol 4: e180

Lercher MJ, Urrutia AO, Pavlícek A, Hurst LD (2003) A unification of mosaic structures in the human genome. Hum Mol Genet 12: 2411–2415

Lin H, Moghe G, Ouyang S, Iezzoni A, Shiu SH, Gu X, Buell CR (2010) Comparative analyses reveal distinct sets of lineage-specific genes within *Arabidopsis thaliana*. BMC Evol Biol 10: 41

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, et al (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430: 471–476

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

Lisch D, Jiang N (2009) *Mutator* and *MULE* transposons. In JL Bennetzen, S Hake, eds, Handbook of Maize: Genetics and Genomics. Springer, New York, pp 277–306

Lisch DR, Freeling M, Langham RJ, Choy MY (2001) *Mutator* transposase is widespread in the grasses. Plant Physiol 125: 1293–1303

Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol 453: 3–31

Marquez CP, Pritham EJ (2010) *Phantom*, a new subclass of *Mutator* DNA transposons found in insect viruses and widely distributed in animals. Genetics 185: 1507–1517

Meyers BC, Lee DK, Vu TH, Tej SS, Edberg SB, Matvienko M, Tindell LD (2004a) Arabidopsis MPSS: an online resource for quantitative expression analysis. Plant Physiol 135: 801–813

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S (2004b) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome Res 14: 1641–1653

Ming R, Vanburen R, Liu Y, Yang M, Han Y, Li LT, Zhang Q, Kim MJ, Schatz MC, Campbell M, et al (2013) Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biol 14: R41

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by *Helitron*-like transposons generate intraspecies diversity in maize. Nat Genet 37: 997–1002

Morgenstern B (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res 32: W33–W36

Neuvéglise C, Chalvet F, Wincker P, Gaillardin C, Casaregola S (2005) *Mutator*-like element in the yeast *Yarrowia lipolytica* displays multiple alternative splicings. Eukaryot Cell 4:

615–624

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497: 579–584

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551–556

Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, 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

Raizada MN, Benito MI, Walbot V (2001) The *MuDR* transposon terminal inverted repeat contains a complex plant promoter directing distinct somatic and germinal programs. Plant J 25: 79–91

Robertson D, Woessner JP, Gillham NW, Boynton JE (1989) Molecular characterization of two point mutants in the chloroplast *atpB* gene of the green alga *Chlamydomonas reinhardtii* defective in assembly of the ATP synthase complex. J Biol Chem 264: 2331–2337

Robertson DS (1978) Characterization of a *Mutator* system in maize. Mutat Res 51: 21–28

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115

Sémon M, Mouchiroud D, Duret L (2005) Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. Hum Mol Genet 14: 421–427

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8: 272–285

Talbert LE, Chandler VL (1988) Characterization of a highly conserved sequence related to *Mutator* transposable elements in maize. Mol Biol Evol 5: 519–529

Tatarinova TV, Alexandrov NN, Bouck JB, Feldmann KA (2010) GC3 biology in corn, rice, sorghum and other grasses. BMC Genomics 11: 308

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641

Vinogradov AE (2003) DNA helix: the importance of being GC-rich. Nucleic Acids Res 31: 1838–1844

Wang Q, Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. Proc Natl Acad Sci USA 103: 17644–17649

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18: 1791–1802

Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, Khan MA, Tao S, Korban SS, Wang H, et al (2013) The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res 23: 396–408

Yamashita S, Takano-Shimizu T, Kitamura K, Mikami T, Kishima Y (1999) Resistance to gap repair of the transposon *Tam3* in *Antirrhinum majus*: a role of the end regions. Genetics 153: 1899–1908

Yu Z, Wright SI, Bureau TE (2000) *Mutator*-like elements in *Arabidopsis thaliana*: structure, diversity and evolution. Genetics 156: 2019–2031

Zabala G, Vodkin LO (2005) The *wp* mutation of Glycine max carries a gene fragment-rich transposon of the CACTA superfamily. Plant Cell 17:2619–2632

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA (2007) Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. BMC Evol Biol 7: 152

CHAPTER 3:

Mutator-like Elements with Multiple Long Terminal Inverted Repeats in Plants

Ferguson AA, Jiang N (2012) *Mutator*-like elements with multiple long terminal inverted repeats in plants. Comp Funct Genomics 2012: 695827

Abstract

Mutator-like transposable elements (MULEs) are widespread in plants and the majority have long terminal inverted repeats (TIRs), which distinguish them from other DNA transposons. It is known that the long TIRs of *Mutator* elements harbor transposase binding sites and promoters for transcription, indicating that the TIR sequence is critical for transposition and for expression of sequences between the TIRs. Here, we report the presence of MULEs with multiple TIRs mostly located in tandem. These elements are detected in the genomes of maize, tomato, rice, and *Arabidopsis*. Some of these elements are present in multiple copies, suggesting their mobility. For those elements that have amplified, sequence conservation was observed for both of the tandem TIRs. For one MULE family carrying a gene fragment, the elements with tandem TIRs are more prevalent than their counterparts with a single TIR. The successful amplification of this particular MULE demonstrates that MULEs with tandem TIRs are functional in both transposition and duplication of gene sequences.

Introduction

Transposable elements (TEs) are DNA fragments that are capable of moving from one genomic location to another and increasing their copy numbers. Based on their transposition mechanisms, TEs fall into two classes: (1) Class I elements, or retrotransposons, that use the element-encoded mRNA as the transposition intermediate and (2) Class II elements, or DNA transposons, that transpose through a DNA intermediate. Autonomous transposons encode transposases that are responsible for the transposition of themselves and their cognate nonautonomous elements that do not encode transposases.

A common feature for DNA transposons, with a few exceptions, is the presence of a terminal inverted repeat (TIR) at each terminus of the element. As an essential structural

component of the element, TIR plays important roles in transposition. For example, the transposase encoded by the bacterial *Tn3* element specifically binds to its TIR (38 bp in length), and this then facilitates the nicking at the end of *Tn3* by DNase I and initializes the transposition process [1]. In eukaryotes, it was shown that the transposase of the *Hermes* element binds to its imperfect TIRs and excises the element. This process is accompanied by the formation of a hairpin structure in the flanking donor sequence, resembling the V(D)J recombination process [2]. Binding of transposase to the TIR and to the target DNA mediates the synapsis of the transposon ends and the target DNA, allowing the insertion of the element into the target sequence [3]. The presence of the TIR sequence also influences the target specificity of the element. For instance, the deletion of a 4 bp sequence within the binding region of the TIR in *Tn3* abolishes its transposition immunity, that is, the phenomenon whereby *Tn3* avoids insertion into another *Tn3* element [4].

The TIRs of DNA transposons are usually less than 50 bp in length and can be as short as 8 bp [5]. Nevertheless, there are a few transposon families with exceptionally long TIRs and among these is the *Mutator* superfamily of transposable elements. First discovered in maize in 1978 [6], *Mutator* and *Mutator*-like elements (MULEs) appear to be prevalent among eukaryotes. Subsequent to the initial discovery in maize, MULEs have been found in other plant genomes such as *Arabidopsis*, rice, and *Lotus japonicus* [7–9], as well as in fungal and animal genomes [10, 11]. In addition to their unparalleled activity in maize [12], MULEs include a special subgroup of elements referred to as Pack-MULEs which are nonautonomous MULEs carrying genes or gene fragments. The sequence acquisition by Pack-MULEs may result in the formation of new open reading frames and the potential to regulate the expression of the parental genes from which the fragments are derived [8, 13, 14].

Mutator and MULEs are distinguished from other DNA transposable elements by having a 9–11 bp target site duplication (TSD) which flanks the element and are formed during transposition into a new genomic location. The TIRs of MULEs, typically ranging from 100 to 500 bp, appear to be critical for element transposition and expression. *Mutator* TIRs contain binding sites for the transposase where MURA protein was shown to bind a conserved ~32 bp sequence motif in active *Mutator* elements in maize [15]. In addition, two convergent genes contained within the maize autonomous *MuDR* element, including the transposase MURA, are transcribed from promoters located within the TIRs [16]. Furthermore, the *MuDR* TIR contains plant cell cycle enhancer motifs which program a 20-fold upregulated expression in reproductive organs as compared with leaves [16]. The promoters in the TIRs are also responsible for the expression of the internal regions of Pack-MULEs [8], suggesting the importance of the TIR sequence for transposition and retention of MULEs in the genome. In this study, we report on the identification and characterization of MULEs and Pack-MULEs with multiple TIRs in plants and the possible role of tandem TIRs in element amplification.

Methods

Plant Genomic Sequences and Construction of the Tomato MULE TIR Library

The sequence for the tomato (*Solanum lycopersicum*) genome was downloaded from the International Tomato Genome Sequencing Consortium Library. The sequence (http://www.solgenomics.net/organism/Solanumlycopersicum/genome/; release 2.40). The sequence for rice (*Oryza sativa* ssp. *japonica* cv. Nipponbare) pseudomolecules was downloaded from the rice annotation group at Michigan State University (http://rice.plantbiology.msu.edu/, release 6.0). Maize (*Zea mays* cv. B73) chromosome sequences (4a.53) were downloaded from the maize sequencing project (http://www.maizesequence.org/, B73 RefGen v1 [17]). The

sequence for potato (*Solanum tuberosum* cv. DM) was downloaded from the Potato Genome Sequencing Consortium (http://potatogenomics.plantbiology.msu.edu, release 3.0 [18]). The *Arabidopsis* genome sequence (TAIR10) was downloaded from the Arabidopsis Information Resource (http://www.arabidopsis.org/). For identifying elements with additional TIRs, MULE TIR libraries that were generated previously were used for rice, maize, and *Arabidopsis* genomes [19].

The tomato MULE TIR library was built with an iterative process that uses Pairwise Alignment of Long Sequences (PALS 1.0) [20] to identify long inverted repeats (minimum length = 100 bp; minimum similarity = 80%). A custom python script was used to identify pairs of inverted repeats from the output of PALS, extract flanking sequences, and identify a 9–11 bp TSD. Manual curation was done to verify terminal inverted pairs with overall high sequence similarity in at least 100 bp sequence and having intact TIR ends and presence of a 9–11 bp TSD immediately flanking the ends of the TIR. Pairs that passed the above criteria were added to a TIR sequence library of tomato which was later filtered for redundancy. If two TIR sequences share 80% or higher similarity in at least 80% of their length, the two sequences are considered redundant, and one of the sequences is excluded. Among a redundant group of TIR sequences, the TIR sequence from the element with the highest TIR identity is retained in the nonredundant library.

Estimation of MULE Copy Number and Identification of Elements with Multiple TIRs from Plants

To determine the abundance of MULEs related to different TIR families (Table 3.1), copy number was estimated by considering one pair of TIRs as one element. To estimate how many elements are associated with a TSD, the presence of a TSD was verified using a pipeline

consisting of perl scripts that search for 9-10 bp direct repeat with no more than 2 mismatches flanking the ends of the TIR sequences. The copy number of autonomous MULE elements was estimated from all elements retrieved from the previous step having significant match to known MULE transposases ($E=10^{-5}$, BLASTX) after filtering for low complexity.

To search for MULEs with multiple TIRs, elements from each genome sequence were identified using RepeatMasker (using default parameters; http://repeatmasker.genome.washington.edu/) with the rice, maize, tomato, and *Arabidopsis* MULE TIR libraries. A custom python script was used to identify elements flanked by 2 similar TIRs on one or both sides of the element. The following criteria were used to filter the results: (1) distance between the external TIRs is not larger than 20 kb and there is no sequencing gap between the TIRs, (2) TIRs must be at least 50 bp long, (3) truncations at the external ends of TIRs must be no more than 15 bp, (4) the two TIRs on one or both ends are less than 600 bp apart, and (5) presence of a 9–11 bp TSD with no more than 2 mismatches. Custom perl and python scripts were used in combination to extract the sequences of putative multiple TIR elements and their flanking sequences, and all elements were manually verified for the presence of a TSD. To define whether an element has multiple copies in a genome, the following criteria were applied: for individual elements, if the TIRs of two elements (with different TSDs) can be aligned (BLASTN, $E=10^{-10}$), and if > 70% of the sequence between the TIRs can be aligned (BLASTN, $E=10^{-10}$), then the two elements are defined as copies.

To obtain an approximate location of the elements in the tomato chromosomes, their coordinates in the pseudomolecules (release 2.4) were used to find nearby flanking SGN-markers as indicated in the Tomato Genome Browser (http://www.solgenomics.net/). The tomato FISH map was used as the basis of chromosome structure indicating centromere, euchromatin, and

heterochromatin regions of each chromosome [21, 22]. The relative position of the flanking SGN-markers were identified in EXPEN-2000 physical map [23, 24] which have been linked to the FISH map by sequenced BACs. These maps are available through the Sol Genomics Network.

Phylogenetic Analysis of TIRs and Internal Sequences

To generate multiple sequence alignments, 220 bp of the external and internal TIRs at both ends (Figure 3.2) of the PM-ZIBP elements (the element containing a gene fragment from a gene encoding a zinc-ion binding protein, see Results) were used and resolved into lineages by generating phylogenetic trees. Multiple sequence alignment was performed by CLUSTALW (http://www.ebi.ac.uk/clustalw/) with default parameters. Phylogenetic trees were generated on the basis of the maximum likelihood method [25] with Kimura-2 parameter distances [26] using the MEGA 4 program (http://www.megasoftware.net). Support for the internal branches of the phylogeny was assessed using 1000 bootstrap replicates. To compare the TIR of type 2-46 elements, 130 bp of the external and internal TIRs of type 2-46 elements were used to generate sequence alignment and a phylogenetic tree employing the same parameters and methods as the TIR comparisons for PM-ZIBP. Similar methods and parameters were used to generate sequence alignment and a phylogenetic tree for the acquired fragments in PM-ZIBP, the parental gene (SGN-U574419) in tomato, and gene sequences from potato, tobacco, and pepper.

Annotation of Pack-MULEs and Frequency of Element Sizes in Tomato

The procedure for the annotation of Pack-MULEs in the tomato genome was similar to that described previously [13]. Candidate Pack-MULEs and MULEs with multiple TIRs identified in this process were masked with all available tomato repeat sequences using

RepeatMasker (http://repeatmasker.genome.washington.edu/ version open-3.0) with default parameters. The repeat library was built by combining repeats collected previously [27] with sequences matching transposase sequences in the RepeatMasker package (version open-3.0). The masked outputs were queried against the Solanaceae Unigene database (http://solgenomics.net/, version 5, BLASTN E= 10^{-10}) and the nonredundant protein database in NCBI (http://www.ncbi.nlm.nih.gov/Blast.cgi E=10^{-05}) to identify gene fragments inside elements. To estimate element size, the elements were masked with all available tomato repeat sequences excluding MULE-related sequences using RepeatMasker to identify nested insertion of other transposons inside Pack-MULEs. The element size was then calculated using a custom perl script which excluded the masked sequence inside Pack-MULEs. Elements larger than 2 kb were not plotted due to their low abundance.

TIR Sequence Analysis and Conservation Test

The external and internal TIR sequences were compared using the "gap" program available from the GCG package (version 11.0, Accelrys Inc., San Diego, CA) to identify repeats found in the sequences. To determine the conservation of the TIR at nucleotide level, the first 700 bp sequence of each element was extracted and compared using DIALIGN2-2 [28] with default parameters. The normalized local sequence similarity scores as determined from DIALIGN2-2 were then used to determine an average similarity score for every 5 nucleotides and then plotted.

Results

Types of MULEs with Multiple TIRs

A typical MULE contains one pair of TIRs, which refers to similar or identical terminal inverted sequences found on the opposite ends of the element (Figure 3.1, see above). In this

study, we detected some atypical MULEs with two pairs of TIRs (type 1 and type 2, Figure 3.1), which will be referred to as external TIR and internal TIR, respectively. The external and internal TIRs have extended sequence similarity in at least 100 bp of the TIR region. In some cases, there is only one additional terminal sequence instead of one pair of terminal sequences (type 3, 4, 5, and 6, Figure 3.1), and this additional terminal sequence will be called a solo TIR to distinguish it from paired TIRs. Analysis of the tomato genome sequence revealed the presence of 61 MULEs with at least one additional TIR. All these elements are associated with a distinguishable TSD, a hallmark of transposition, suggesting that these elements are derived from transposition and not recombination. Furthermore, there is no recognizable TSD flanking the internal TIR, indicating that these elements are not formed through nested insertion of the same type of elements.

As shown in Figure 3.1, the majority of these elements (48 out of 61 elements, 79%) are type 1 and type 2 which contain external and internal TIRs located in tandem on both ends of the element (Figure 3.1). Some elements carry recognizable gene fragments (type 1 and type 5) and are therefore classified as Pack-MULEs. Thirteen elements (out of 61, 21%) are associated with one additional solo TIR only on one end of the element (type 3) or in the internal region (type 4, 5, and 6). Elements with tandem TIRs are more abundant than elements containing a solo TIR (48 versus 13). Three elements (out of 6) with tandem TIRs have multiple copies (elements are considered to be copies if they have similar TIR and similar internal region). The element with

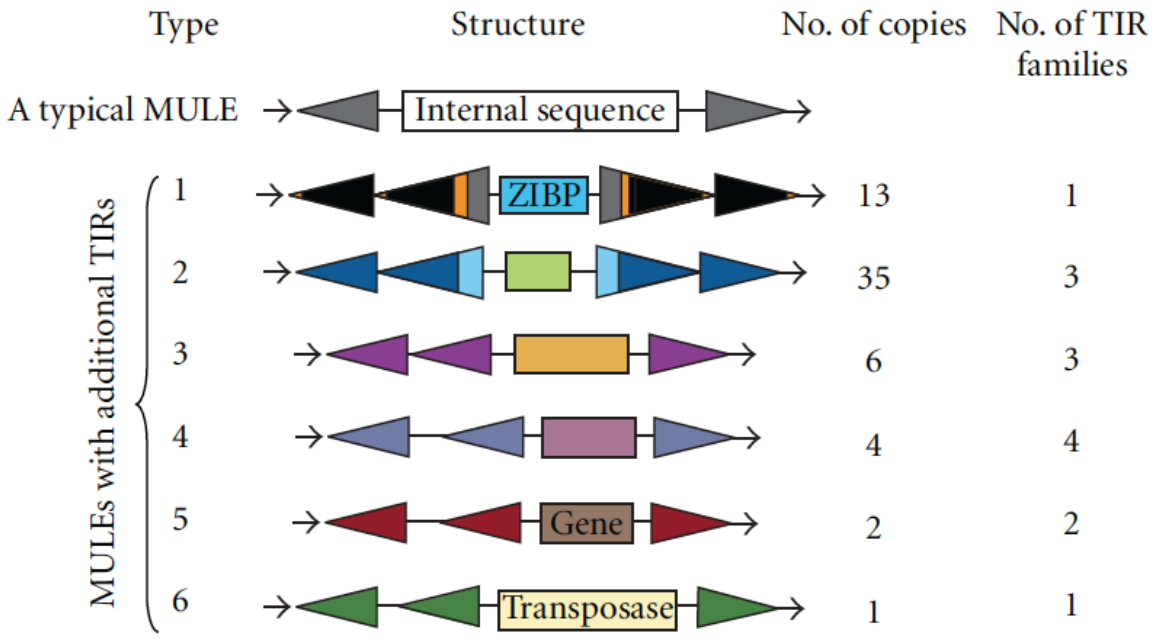


Figure 3.1. The structure of distinct types of MULEs with additional TIRs in tomato including their copy numbers and the number of TIR families involved in formation. Black horizontal arrows indicate target site duplication (TSD); solid colored triangles indicate Terminal Inverted Repeat (TIR); colored boxes indicate internal sequence and are labeled accordingly if sequences are annotated as genes or have similarity to MULE transposases. Objects with different colors indicate unrelated sequences. ZIBP – zinc ion binding protein

the highest copy number is flanked by two SLMULE46 TIRs (29 copies, type 2, and will be referred to as type 2-46 element), followed by a second group, an element with 13 copies that makes up the type 1 class. In contrast, among the 12 distinct elements containing one additional solo TIR, only one has another copy in the genome, and this element belongs to type 3 with SLMULE46 TIR. In fact, this element is the 3-TIR version of type 2-46 because it has a similar internal region (see below). This suggests that elements with solo TIRs may be dramatically less competent in transposition than elements with tandem TIRs on both ends. In tomato, a total of 59 MULE TIR families have been identified with approximately 28,000 total copies of MULEs (see Table 3.1 for distribution of copies in each family). Among them, 10 TIR families (17%) are involved in the formation of elements with additional TIR sequences (Figure 3.1, Table 3.2). These 10 families have moderate copy numbers and a fraction of putative autonomous elements comparable to many other TIR families that are not associated with the formation of elements with multiple TIRs (Table 3.1). Eight families are represented in the formation of elements with solo TIRs, yet only three TIR families are involved in the formation of elements with tandem TIRs (type 1 and type 2, Table 3.2). In addition, most of the elements with this atypical TIR feature are non-autonomous elements in that they do not encode the proteins essential for transposition. The only exception is an element with an additional solo TIR, whose internal region was found to have sequence similarity to MULE transposases (type 6 in Figure 3.1). However, a close examination indicates the presence of numerous premature stop codons and frameshifts in the coding region, suggesting the element is unlikely a currently functional autonomous element.

Table 3.1. Distribution of elements among tomato MULE TIR families.

TIR Family	Copy number	No. with TSD	% with TSD	Autonomous Elements	% of Autonomous
MULETIR056	8	1	12.5	0	0.0
MULETIR022	11	3	27.3	0	0.0
MULETIR028	13	4	30.8	0	0.0
MULETIR007	25	5	20.0	3	12.0
MULETIR048	26	7	26.9	0	0.0
MULETIR050	34	2	5.9	0	0.0
MULETIR049	40	1	2.5	2	5.0
MULETIR034	42	1	2.4	0	0.0
MULETIR025	49	13	26.5	9	18.4
MULETIR052	56	11	19.6	0	0.0
MULETIR044	62	2	3.2	1	1.6
MULETIR038	68	22	32.4	10	14.7
MULETIR031	71	5	7.0	3	4.2
MULETIR039	88	26	29.5	3	3.4
MULETIR054	100	28	28.0	12	12.0
MULETIR043	106	63	59.4	35	33.0
MULETIR047	109	25	22.9	0	0.0
MULETIR017	112	4	3.6	5	4.5
MULETIR045	112	12	10.7	2	1.8
MULETIR020	114	36	31.6	0	0.0
MULETIR053	116	8	6.9	3	2.6
MULETIR010	118	3	2.5	35	29.7
MULETIR011	130	43	33.1	7	5.4
MULETIR059	144	49	34.0	2	1.4
MULETIR042	156	6	3.8	1	0.6
MULETIR027	164	72	43.9	19	11.6
MULETIR029	164	64	39.0	1	0.6
MULETIR016	167	40	24.0	4	2.4
MULETIR030	176	31	17.6	1	0.6
MULETIR051	210	64	30.5	1	0.5
MULETIR023	233	84	36.1	1	0.4
MULETIR004	250	4	1.6	2	0.8
MULETIR058	262	66	25.2	2	0.8
MULETIR026*	266	56	21.1	0	0.0
MULETIR013	273	73	26.7	94	34.4
MULETIR055	275	87	31.6	2	0.7
MULETIR057*	337	178	52.8	10	3.0
MULETIR040	361	173	47.9	20	5.5
MULETIR036	373	100	26.8	17	4.6
MULETIR024*	381	96	25.2	4	1.0
MULETIR041	413	126	30.5	3	0.7

Table 3.1 (cont'd)

MULETIR046*	418	62	14.8	2	0.5
MULETIR032	444	199	44.8	62	14.0
MULETIR015	455	130	28.6	9	2.0
MULETIR018*	498	198	39.8	11	2.2
MULETIR003	546	193	35.3	10	1.8
MULETIR037*	635	361	56.9	40	6.3
MULETIR008*	644	133	20.7	4	0.6
MULETIR009*	760	425	55.9	26	3.4
MULETIR035	823	407	49.5	30	3.6
MULETIR014	1144	236	20.6	1	0.1
MULETIR021	1301	73	5.6	7	0.5
MULETIR033*	1310	548	41.8	16	1.2
MULETIR005*	1341	110	8.2	7	0.5
MULETIR006	1621	816	50.3	18	1.1
MULETIR002	1685	748	44.4	1	0.1
MULETIR012	1798	748	41.6	5	0.3
MULETIR019	2386	914	38.3	3	0.1
MULETIR001	4017	2609	64.9	3	0.1
total	28041	10604	38%	569	2%

* MULE TIR families involved in TIR duplication.

Table 3.2. MULE TIR families involved in TIR duplication in tomato.

Type	MULE TIR	Number of copies
1	SLMULE18	13
2	SLMULE18	1
	SLMULE33	1
	SLMULE46	33
3	SLMULE26	1
	SLMULE46	4
	SLMULE57	1
4	SLMULE05	1
	SLMULE08	1
	SLMULE26	1
	SLMULE46	1
5	SLMULE09	1
	SLMULE24	1
6	SLMULE37	1

A similar analysis of rice and maize genomes to detect elements with multiple TIRs identified fewer elements compared to the tomato genome (Tables 3.3 and 3.4). Two elements belonging to the type 3 and four elements (type 3 and 4) were uncovered in rice and maize genomes, respectively. In these two genomes, the presence of tandem TIR on both ends (type 1 or type 2) was not detected. In contrast, in *Arabidopsis*, a species that has previously been reported as having few MULEs and Pack-MULEs compared to rice and maize [17, 19, 29], eleven elements with tandem TIRs on both ends were found, including one Pack-MULE. This suggests that the formation of tandem TIR elements is not related to the abundance of MULEs and Pack-MULEs in the genome, and dicot genomes harbor more tandem TIR elements than genomes of monocots.

A Pack-MULE Family with Tandem TIRs

The elements comprising type 1 elements in tomato are copies of a Pack-MULE that harbors a fragment from a gene encoding a zinc-ion binding protein (Figure 3.2). This Pack-MULE family has 13 copies with tandem SLMULE18 TIRs located on both ends of the element and a single copy with one pair of TIRs (Figure 3.2), resembling that of a typical Pack-MULE. This family will be referred to as PM-ZIBP hereafter (Table 3.5). To dissect the relationship between the single TIR element and elements with tandem TIRs, a phylogenic tree was built using the acquired region, the parental gene in tomato, and the corresponding regions from other related plants including potato, pepper, and tobacco. The PM-ZIBP elements are grouped with the putative parental gene from tomato. Moreover, related elements are not present in the genome of potato suggesting that the acquisition of the gene fragment and the formation of the Pack-MULE may have occurred after the divergence of tomato and potato. Among the PM-ZIBP elements,

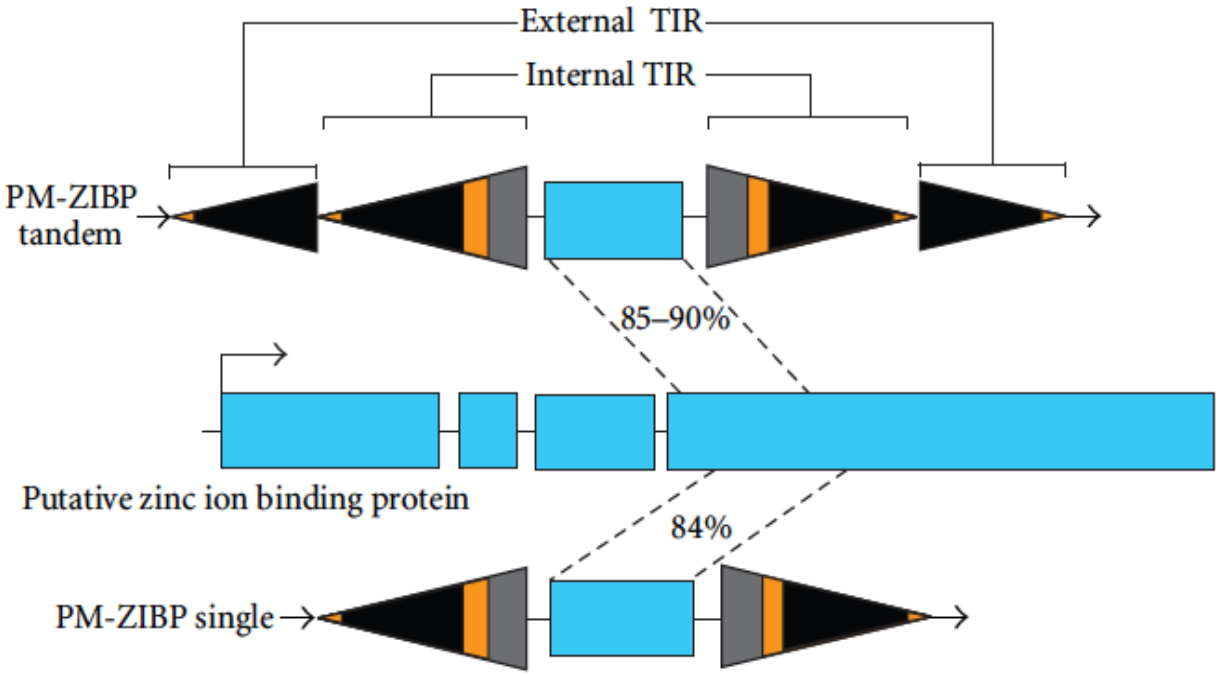


Figure 3.2. PM-ZIBP elements with single and tandem TIRs that contain a gene fragment. Solid triangles indicate TIRs and blue boxes indicate exons of gene SGN-U574419 and fragment acquired by the Pack-MULE. Introns are depicted as lines connecting exons.

Table 3.3. List of the multiple TIR elements in Arabidopsis, maize and rice.

Element ID	Genome	Chromo-some	Position Start	Position End	TIR family	Type*
AtMULEdt-1	Arabidopsis	1	15918026	15919043	At000317	2
AtMULEdt-2	Arabidopsis	1	16260326	16261097	At000824	2
AtMULEdt-3	Arabidopsis	2	5670143	5671281	At000317	2
AtMULEdt-4	Arabidopsis	2	6446333	6451005	At000317	1
AtMULEdt-5	Arabidopsis	3	16141980	16143074	At000317	2
AtMULEdt-6	Arabidopsis	4	1230491	1231738	At000800	2
AtMULEdt-7	Arabidopsis	4	6700242	6701357	At000800	2
AtMULEdt-8	Arabidopsis	4	2341448	2342412	At000317	2
AtMULEdt-9	Arabidopsis	5	9709514	9710845	At000800	2
AtMULEdt-10	Arabidopsis	5	17479616	17480806	At000317	2
AtMULEdt-11	Arabidopsis	5	18980159	18981488	At000800	2
OsMULEdt-1	Rice	1	20126005	20126814	Os0182	3
OsMULEdt-2	Rice	9	8210481	8216333	Os1455	3
ZmMULEdt-1	Maize	1	239115417	239115828	Zm15155	3
ZmMULEdt-2	Maize	5	17121166	17121599	Zm00411	4
ZmMULEdt-3	Maize	8	94462286	94472999	Zm28610	4
ZmMULEdt-4	Maize	8	121663900	121667746	Zm00411	4

*Based on the same classification as Figure 3.1.

Table 3.4. Frequency of MULEs, Pack-MULEs and MULEs with multiple TIRs in four different plant genomes.

Element type	Arabidopsis	Tomato	Rice	Maize
MULEs	1576	28041	30475	12900 [a]
Pack-MULEs	46[b] (2.92)	220 (1.6)	2853 [b] (9.4)	276 [b] (2.1)
MULEs with multiple TIRS	11 (0.70)	61 (0.22)	4 (0.01)	2 (0.02)

Number in parenthesis indicate percentage from total number of MULEs.
[a] Schnable et al., 2009 [16]
[b] Jiang et al., 2011 [18]

the element with a single TIR (PMZIBP-1) forms a branch with the longest length (Figure 3.3). If the mutation rate is comparable among this group of elements, this implies that PM-ZIBP-1 is the most ancient element that acquired this gene fragment (or the ancestor of this element acquired the gene fragment), and elements with tandem TIRs are putative derivatives of PM-ZIBP-1. This is consistent with the fact that PM-ZIBP-1 is associated with the lowest TIR identity (Table 3.5), since young elements often have identical or highly similar TIRs. Our results show no evidence that elements with tandem TIRs are capable of acquiring gene sequence.

The 14 copies of PM-ZIBP were mapped onto the tomato chromosomes (Table 3.5). All the elements with tandem TIRs are mapped to distinct chromosomal loci with most of them in euchromatic regions. However, the single-TIR copy mapped to Chromosome 0 which consists of sequenced fragments that cannot be physically mapped to any of the chromosomes. Sequence analysis of the contig containing this element showed that it was highly repetitive suggesting its location in heterochromatic regions of the genome. This raised the question as to whether the long branch length associated with PM-ZIBP-1 element is an artifact of accelerated mutation in heterochromatic regions. To test this notion, the five elements with tandem TIRs that are located in heterochromatic regions (Figure 3.4) were examined, and these elements were found to be associated with both long and short branches (Figure 3.3). Moreover, none of them has a branch that is longer than that of PM-ZIBP-1. As a result, the location of PM-ZIBP-1 does not fully explain its branch length, and it is likely the oldest PM-ZIBP element. However, we cannot rule out the possibility that there are other unknown factors responsible for the unusually high mutation rate (in both TIR and internal regions) in PM-ZIBP-1 that are not correlated with its

Table 3.5. List of the Pack-MULEs that captured a fragment of a putative zinc-ion binding protein.

Element ID	Chromosome	Position Start	Position End	Size	TSD sequence	% outer TIR identity	% inner TIR identity
PM-ZIBP-1	0	12038207	12039150	944	TTTTAAATT	87	N/A
PM-ZIBP -2	2	29724087	29725458	1372	TAAATTATA	94	92
PM-ZIBP -3	2	33667391	33668737	1347	TACATTTTAA	92	90
PM-ZIBP -4	3	50199015	50200383	1369	TTAAAATTA	91	93
PM-ZIBP -5	4	4315606	4316970	1365	TATTATAAA	95	90
PM-ZIBP -6	4	58126752	58128120	1369	GTCAGGTTAA	93	91
PM-ZIBP -7	5	10380074	10381444	1371	ATAAAAGAT	93	92
PM-ZIBP -8	6	29844899	29846272	1374	CTTCGAGAC	91	92
PM-ZIBP -9	6	41871484	41872853	1370	TTTATTTAC	90	89
PM-ZIBP -10	6	42030690	42032062	1373	TTAAAAAAA	92	92
PM-ZIBP -11	6	7121877	7123250	1374	TTAAAAGAA	90	90
PM-ZIBP -12	8	14577081	14578450	1370	GAATAATAA	93	91
PM-ZIBP -13	8	4530978	4532348	1371	TTTTGGGAA	93	89
PM-ZIBP -14	12	9912832	9914208	1377	TATTTTTAT	92	90

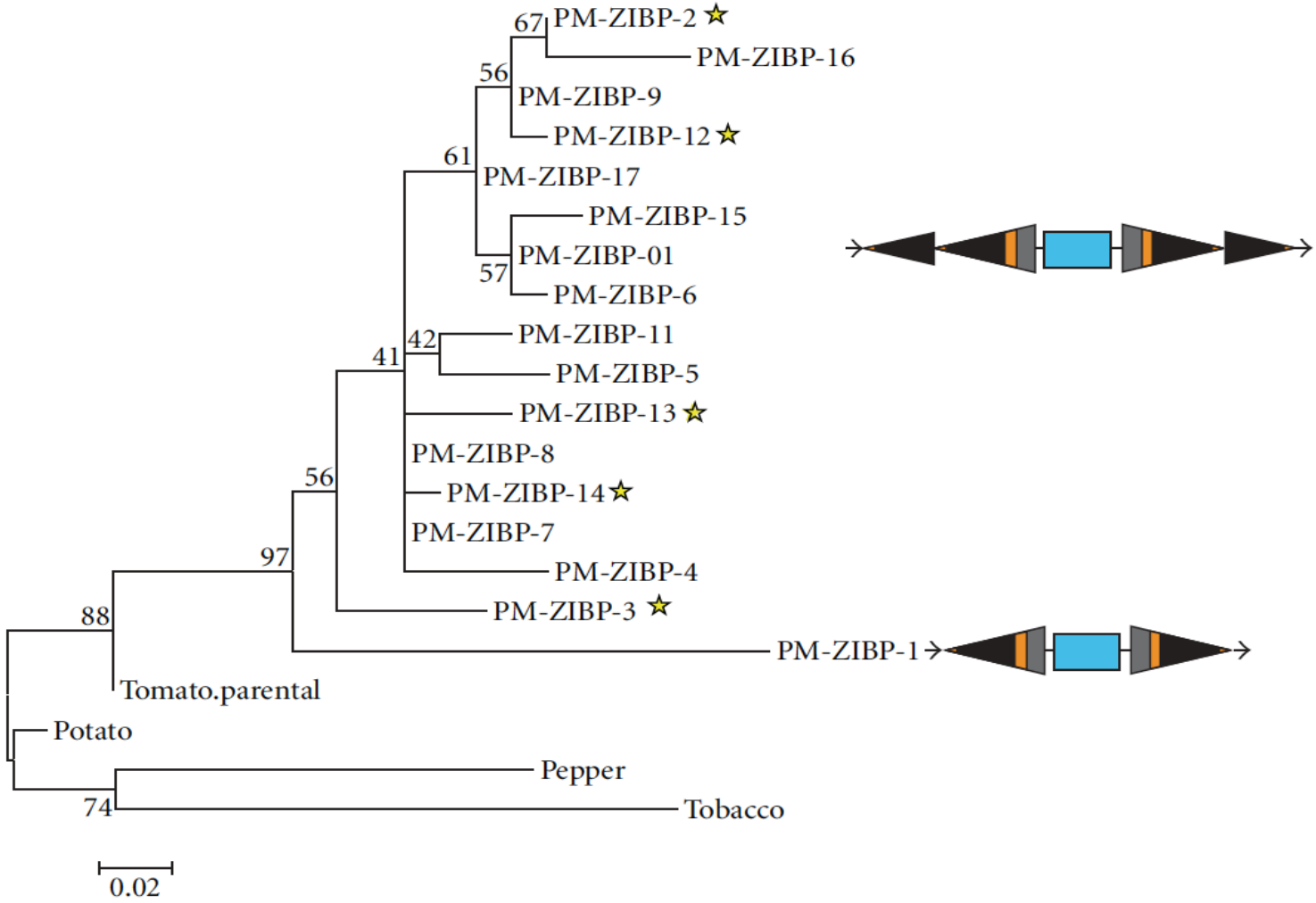


Figure 3.3. Phylogenetic analysis based on the acquired fragment in PM-ZIBP from SGN-U574419 and related sequences (SGN-U273862, SGN-U20267, and SGN-U506815). Sequences were aligned using ClustalW and phylogenetic reconstruction used the maximum likelihood method with Kimura-2 parameter distances implemented in the MEGA program. Bootstrap values are indicated as a percentage of 1000 replicates (40% majority rule consensus). Elements mapping to heterochromatic regions are indicated by a star symbol.

age. Despite this limitation, it is obvious that PM-ZIBP-1 has been present in the genome for a substantial amount of time, without being amplified.

Sequence Features with Elements Carrying Multiple TIRs

To understand the mechanism involved in the formation of PM-ZIBP elements in tomato, careful analysis, and comparison between the external and internal TIR sequences were performed for the single and tandem TIR copies. Three motifs with high sequence similarity were found in PMZIBP with tandem TIRs and two of these, motif-I and motif-II, were also found in the single TIR of PM-ZIBP-1 (Figure 3.5). The presence of the repetitive motifs in the TIR suggests that the additional TIR could be formed via a DNA replication slippage process involving a single TIR element. According to this model, when the replication proceeds to motif-II, the DNA polymerase slips from the DNA template and subsequently reattaches at motif-I so that the sequence between motif-I and motif-II is duplicated. If this is the case, the external and internal TIR should originate from the same template. To test this notion, phylogenetic analysis of the internal and external TIRs from the tandem TIR PM-ZIBP and the TIR of the single TIR element from both 5' end and 3' end was performed using ClustalW and MEGA (Figure 3.6). This analysis demonstrates a separate grouping of the internal, external and single TIRs, which seems to contradict with the slippage hypothesis. However, this is not definitive evidence against the slippage hypothesis since the bootstrap values are relatively low and separation of the external and internal TIR could be due to their distinct role in transposition (also see below).

The examination of another element with tandem TIRs (type 2-46) failed to identify the presence of similar motifs, suggesting that the presence of recognizable repetitive motif is not essential for the formation of tandem TIRs. Phylogenetic analysis between the external and internal TIRs of this family showed four fundamental groups (5' internal, 5' external, 3' internal,

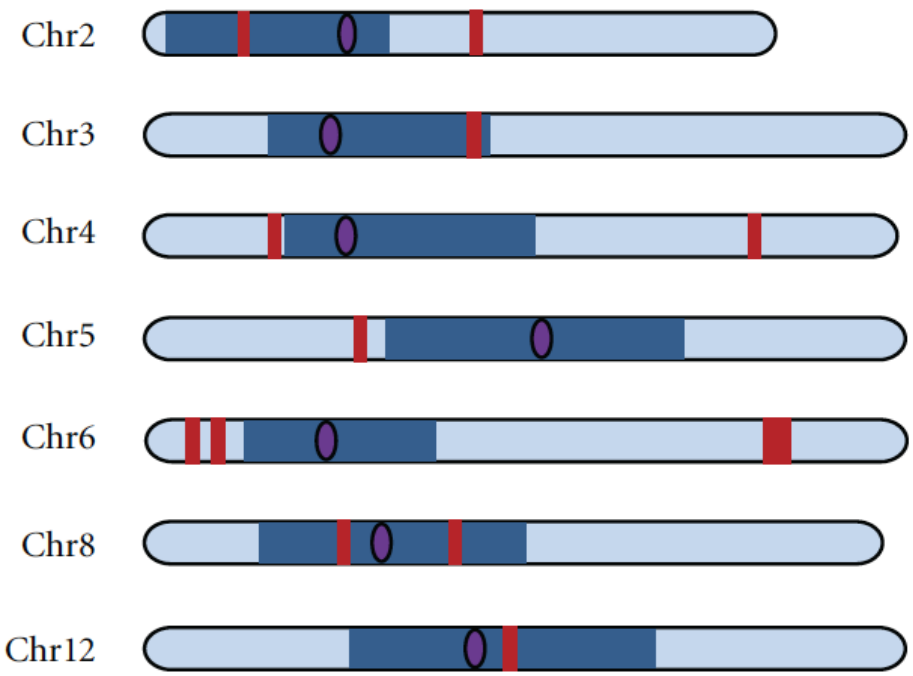


Figure 3.4. Chromosomal distribution of tandem TIR PM-ZIBP in tomato. Dark blue blocks represent heterochromatin and light blue regions represent euchromatin. Individual elements are represented by dark red vertical bars, and the purple ovals indicate the location of the centromere.

Figure 3.5. Alignment of TIR sequences from a PM-ZIBP with tandem TIRs and that from PM-ZIBP-1 illustrating the location of the 3 repetitive motifs found in the TIRs.

3' external) (Figure 3.7) with distinct branch lengths, suggesting they have been amplifying for an extended period. However, there are several TIRs intermingled with other groups and the bootstrap values for the major branches are rather low. As a result, it is difficult to make a clear-cut interpretation about the origin of the TIR duplication in this family. Interestingly, the two 3-TIR elements do not form an independent group. Instead, their TIRs intermingled with different elements with tandem TIRs. The branch length of the 3-TIR elements is comparable to other type 2-46 elements so they might have formed within a similar time frame.

An alternative mechanism for the formation of elements with tandem TIRs is through recombination between elements with related TIRs. If this is the case, the initial parental elements are not expected to harbor a TSD. To test this hypothesis, we screened all 3-TIR elements and tandem TIR elements that are not associated with a TSD. A closer examination of these candidates indicates all of them have a certain level of truncation at the very termini of the TIR. As a result, it is not clear whether the lack of TSD is due to recombination or due to truncation. Thus, it remains an open question whether recombination played a role in the formation of these elements.

The Putative Role of the Tandem TIRs in Amplification of the Elements

As mentioned above, the Pack-MULE element with single and tandem TIR PM-ZIBP copies share terminal and internal sequences, yet the elements with tandem TIRs have many more copies than the single TIR PM-ZIBP-1 (13 versus 1). For the type 2-46 element with SLMULE46 TIRs, we failed to identify a corresponding element with single TIR and exactly the same internal sequence. However, other non-autonomous MULEs with single SLMULE46 TIR and associated with a TSD were identified and the copy number of none of them is as high as that of type 2-46 (29 copies). The copy numbers of these elements range from 1 to 14, with an

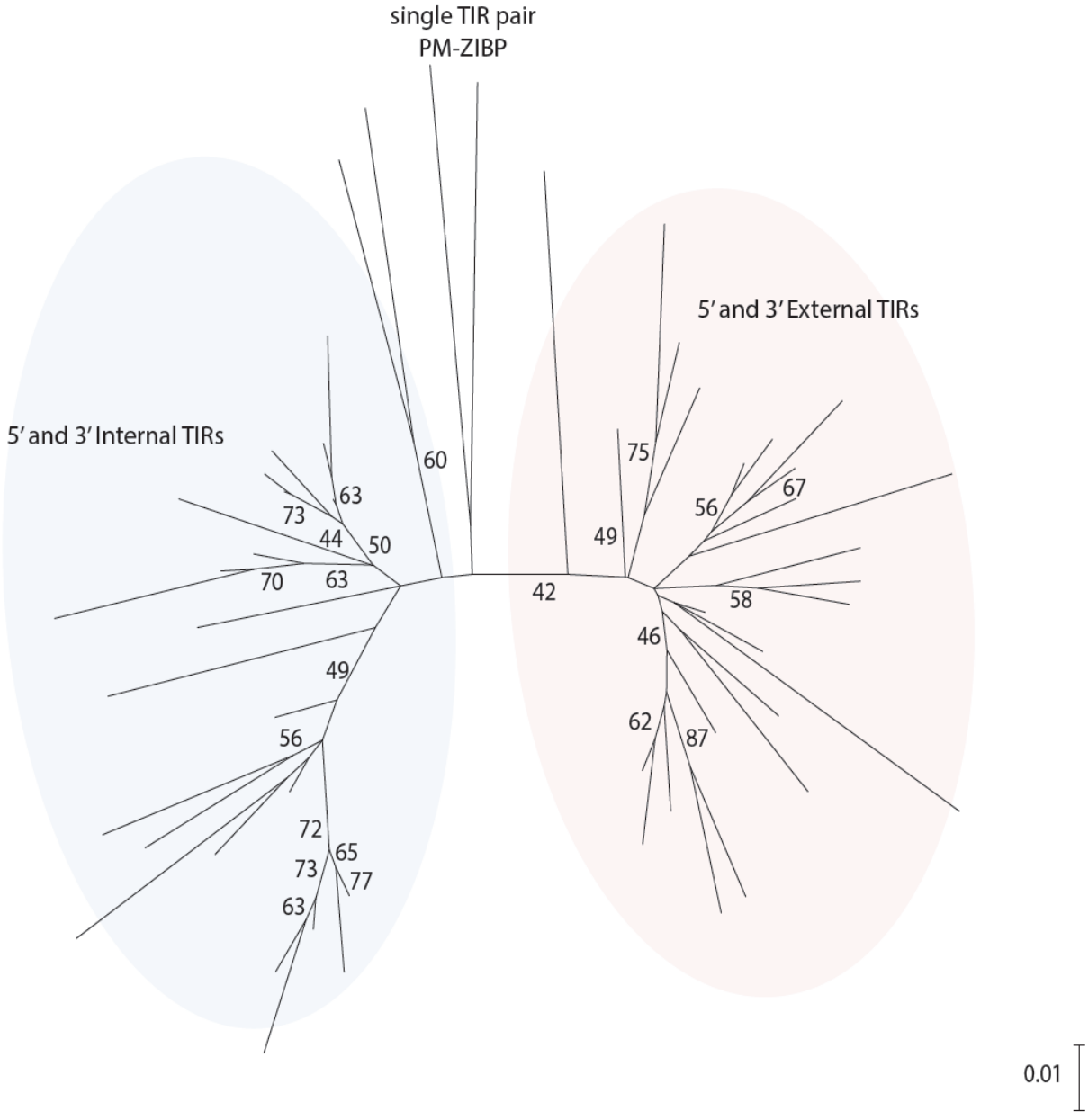


Figure 3.6. Phylogenetic analysis of sequences of the external, and internal TIRs of PM-ZIBP elements. Sequences were aligned and phylogeny was reconstructed as described for Figure 3.3

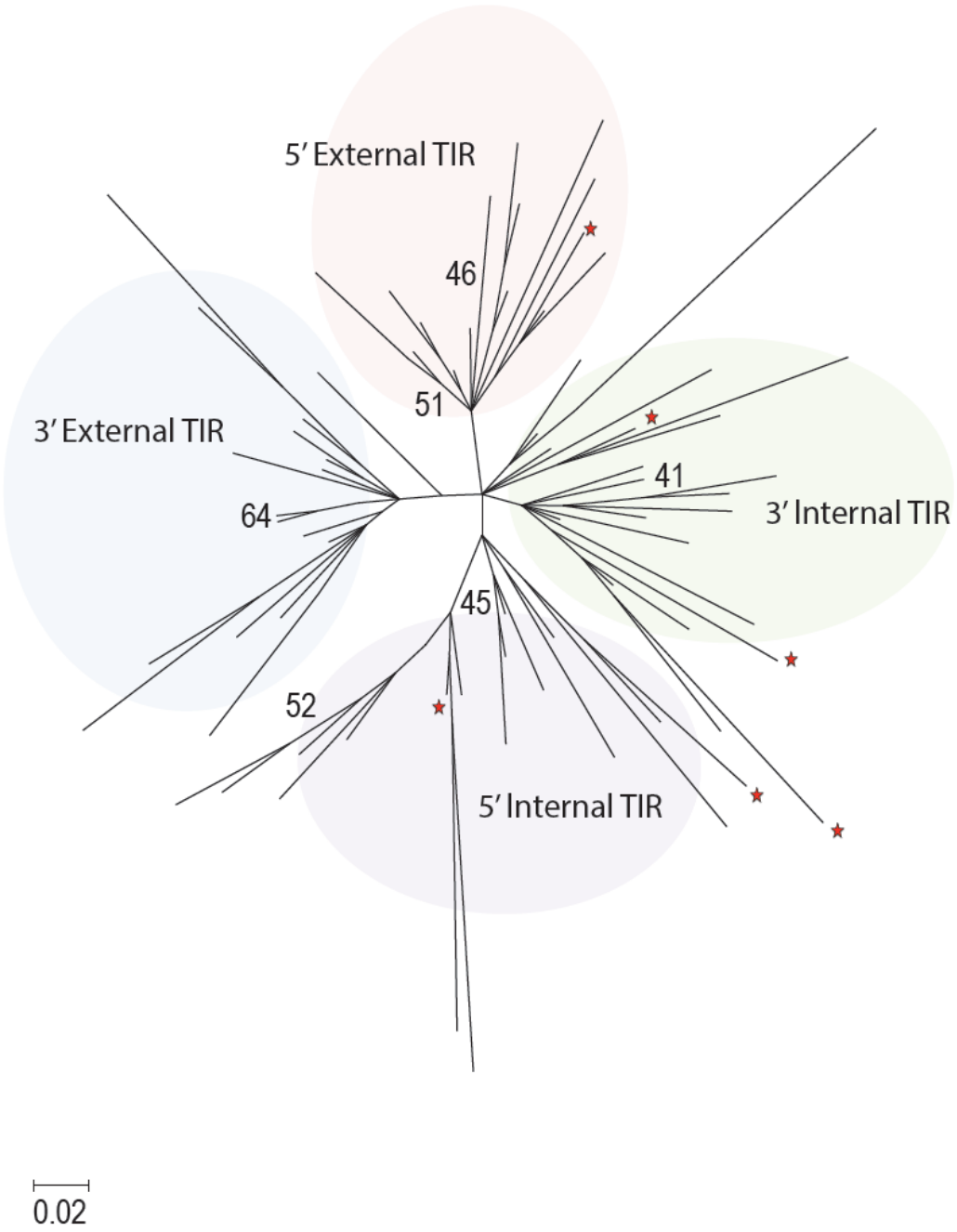


Figure 3.7. Phylogenetic analysis of external and internal TIRs from type 2-46. TIRs from the 3-TIR elements are indicated by a star symbol. Sequences were aligned and phylogeny was reconstructed as described for Figure 3.3

average of 3 copies. A parsimonious explanation for such phenomenon is that the presence of the second TIR confers some advantage for transposition. The internal TIR may simply function as a filler DNA that allows the element to achieve an optimum size. To evaluate the role of size in transposition efficiency of PM-ZIBP, the distribution of Pack-MULE sizes in tomato was examined to identify a range that would correspond to optimal sizes for successful movement and amplification in the genome. The Pack-MULEs in tomato were grouped according to size at 100 bp increments. As shown in Figure 3.8, Pack-MULEs are most abundant with size ranging from 1000–1200 bp which is very close to that of the single TIR PM-ZIBP-1 (944 bp). There are elements from 9 TIRs with 24 different types of internal regions so the presence of this maximum is not due to the amplification of one or two element families. Meanwhile, a minor maximum was observed at 1300–1400 bp (composed of seven TIR families with 10 different internal sequences), which coincides with the size of PM-ZIBP associated with tandem TIRs. Interestingly, the sizes of the elements with tandem SLMULE46 TIRs that predominantly compose the type 2 elements (type 2-46) fall into the same peak as the PM-ZIBP with tandem TIRs (Figure 3.8). The presence of tandem TIRs (over 1 kb in total) and internal sequence makes it highly unlikely for a single element to be within 1000–1200 bp in size. As a result, 1300–1400 bp could be the optimal size for elements with tandem TIRs, regardless of the presence or absence of gene fragments in the internal region.

An additional explanation for the abundance of tandem TIR elements over their single TIR counterpart is an advantage conferred by the tandem TIR resulting in increased frequency of recognition by the transposase or enhanced interaction between the element and the transposition machinery. If this is the case, one would expect significant sequence conservation in both TIRs. Comparison of sequence identity between the external TIRs and the internal TIRs of PM-ZIBP

shows that the initial part of the internal TIR is slightly less conserved compared to that of the external TIR. However, the majority of the TIR sequence has a similar level of conservation, indicating that both TIRs may play functional roles (Figure 3.9A). The most conserved region is motif II and its adjacent region (orange and grey region, Figure 3.9A), which is not present in the external TIR, suggesting the importance of this region. This is in contrast to the low conservation level in regions between the TIR and the acquired gene fragment. Interestingly, the conservation level of the acquired gene fragment is comparable or slightly higher than that of TIRs, suggesting that the gene fragments might be functional. The divergence of the internal TIR around motif-M may be a result of selection to ensure that precise cleavage occurs in the external TIR instead of the internal TIR upon excision of the element. This is in concordance with the fact that no element with single TIR appears to be derived from elements containing tandem TIRs among PM-ZIBP elements. Compared to the internal region, the TIRs of the type 2-46 elements have considerable level of conservation, which is similar to that of PM-ZIBP (Figure 3.9B). However, unlike the PM-ZIBP elements, the most internal region of the internal TIR does not demonstrate an elevated level of conservation, suggesting the variation in location of important *cis* elements among different families of TIR sequence. Furthermore, a low level of conservation was observed in the internal region of this group of elements, which is consistent with the lack of gene fragments in its internal region.

Discussion

The Formation and Amplification of MULEs with Additional TIR Sequence

The TIR sequences of DNA elements contain *cis* elements that are responsible for interaction with and recognition by the relevant transposases. It also contributes to the selection

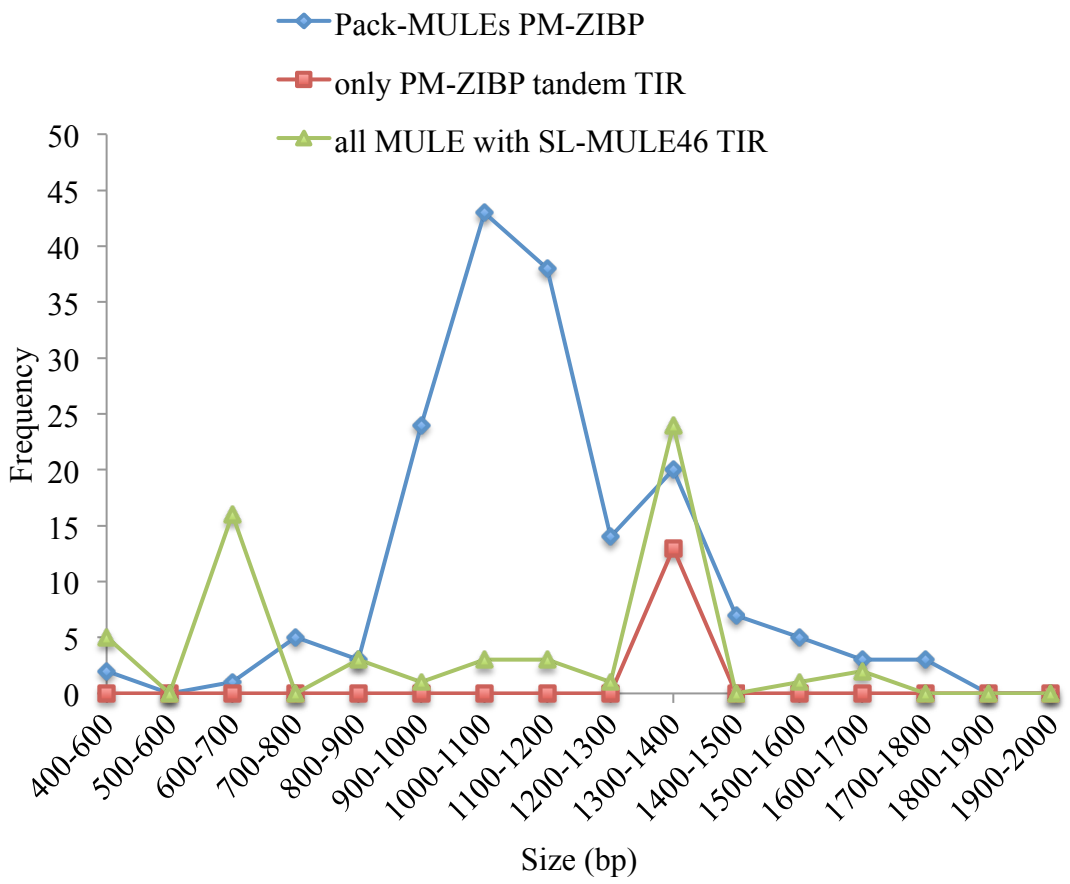
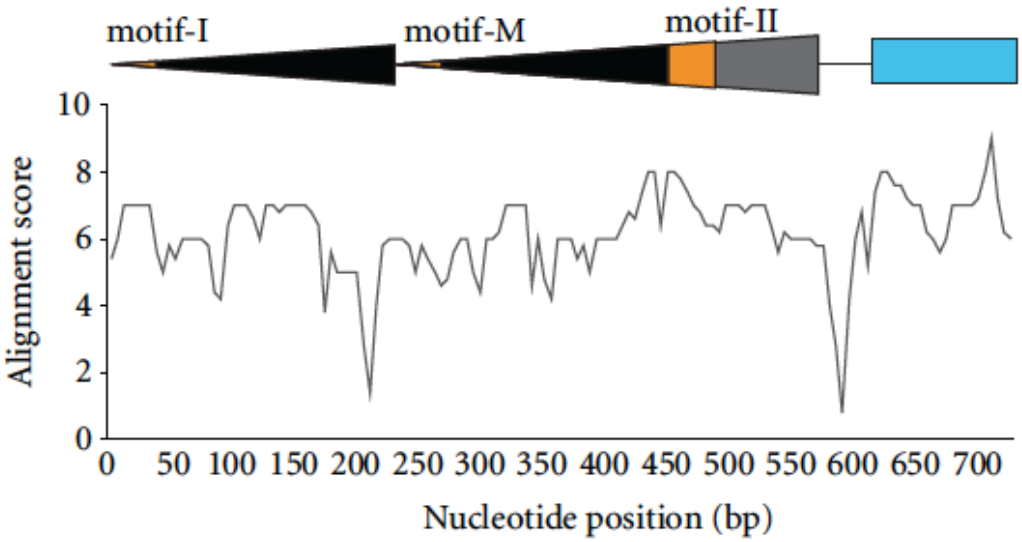


Figure 3.8. The frequency of different element sizes. Elements that are less than 2 kb are plotted.

A



B

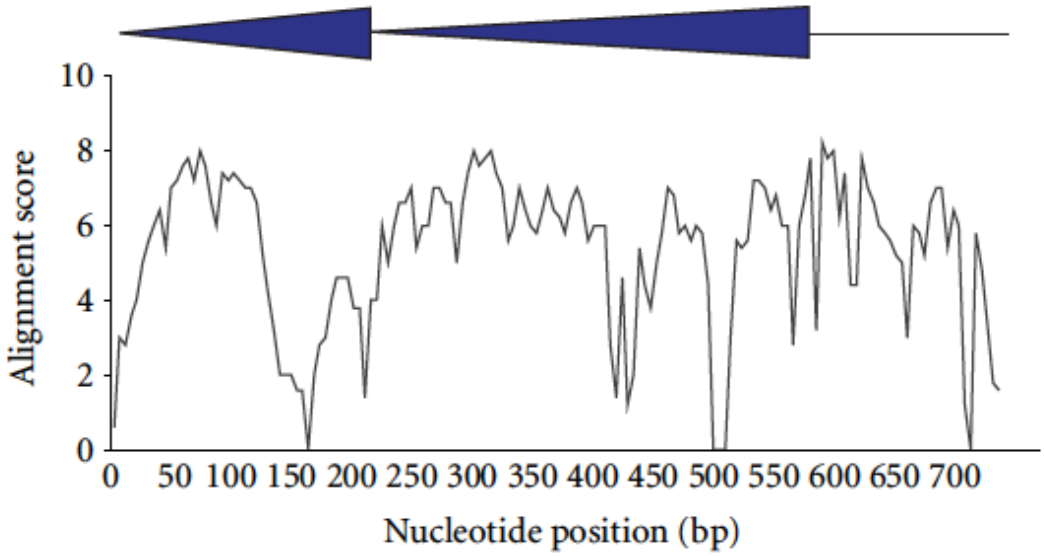


Figure 3.9. Nucleotide conservation across the two tandem TIRs. (A) Tandem TIRs from PM-ZIBP. (B) Tandem TIRs from type 2-46. The nucleotide conservation scores are calculated as an average of 5 nucleotide position scores from the copies of the element. Colored or black triangles represent the TIR. In Figure 3.9A, the orange regions indicate the 3 repetitive motifs (see text). Colored box indicates part of the acquired gene fragment.

of insertion site as well as serving as the target for epigenetic regulation [4, 30]. As a result, the TIR sequences play a critical role in the successful amplification of relevant TEs. For many DNA transposons, short repetitive motifs are present in the TIR or subterminal regions, either in direct or inverted orientations. Nevertheless, the duplication of an entire TIR (or almost an entire TIR) is unusual and has not been studied previously. In this study, 59 MULE TIR families in tomato were examined and 10 of them are associated with TIR duplication. This indicates that certain MULE TIR families have a propensity to form duplicate TIRs over others, and the frequency is not correlated with the total copy number of the particular TIR family. Among the elements with multiple TIRs, some are only associated with one solo TIR (type 3, 4, 5, and 6, with 8 TIR families) and others are associated with duplicated TIRs on both ends (type 1 and 2, with 3 TIR families). Obviously, few TIR families are associated with the formation of duplicated TIRs on both ends, suggesting this is a less frequent event. However, only one element containing a solo TIR has an additional copy, and it is uncertain whether the two copies are derived from each other. This suggests the destiny of "death on birth" for elements with a solo TIR. It is possible that the presence of one additional TIR resulted in a lack of structural symmetry which interferes with transposition. In other words, the presence of one additional TIR sequence could have negative impact on transposition competency. In contrast, the elements with tandem TIRs on both ends are more successfully amplified, despite their low frequency of initial formation.

The Mechanism Involved in the Formation of Duplicated TIRs

At present it is not clear how the TIR sequence was duplicated in these atypical MULEs. DNA replication slippage is considered a common mechanism to cause deletion or duplication of sequences when repetitive motifs are present in adjacent regions. This seems to apply to the PM-

ZIBP elements due to the presence of repetitive motifs inside the TIR sequence. Nevertheless, this hypothesis is not unambiguously supported by phylogenetic analyses of internal and external TIR sequences. Furthermore, not all elements with additional TIR have significant repetitive motifs inside the TIR. As a result, there may be other mechanisms involved in the duplication of TIRs. This may include duplication by recombination or through nested insertion followed by loss of TSD for the internal element. If recombination is the main factor that drives the formation of elements with multiple TIRs, one would expect those elements to be overrepresented among the TIR families with highest copy numbers in the genome. Nevertheless, it does not seem to be the case (Table 3.1).

Another question about the formation of elements with tandem TIR is whether the duplication of TIR at both ends is a single event or a step-wise process. Based on the fact that there are elements with 3 TIRs, it is possible that the duplication is a step-wise process. The coexistence of type 2-46 with two corresponding 3-TIR elements seems to suggest this is the case. However, the phylogenic analysis does not support the notion that the 3-TIR elements are older than all type 2-46 elements (Figure 3.7). Thus, it is unlikely that the 3-TIR elements are the direct progenitor of elements with tandem TIRs, and the true ancestor may have been lost from the genome. On the other hand, since the two 3-TIR elements are more closely related to type 2-46 elements than to each other, it seems to imply they are not derivatives of each other. In this case, an alternative scenario is that the two 3-TIR elements are derivatives of distinct type 2-46 elements through aberrant transposition. This may occur, for example, when one external TIR and one internal TIR in type 2-46 are recognized for transposition. This is consistent with the fact that none of the other 3-TIR elements has a duplicated copy, so the duplication of this particular 3-TIR element may have not arisen through the transposition of itself.

It is known that non-autonomous MULEs are capable of acquiring genomic sequences including genes. The frequency of acquisition of genes by MULEs seems to be higher than that of other DNA elements with shorter TIRs [17]. Moreover, the acquired sequences can be integrated into extended TIR sequences [13, 31]. Given this fact, it is conceivable that the additional TIR could also be introduced through acquisition. Unfortunately, the mechanism of sequence acquisition is yet to be understood.

The comparison of copy number of elements with tandem TIRs in different genomes may provide additional insights into this question. Considering the abundance of MULEs and Pack-MULEs in the genomes of maize and rice, it is striking that only a few MULEs with additional TIRs are found in these genomes. However, if we assume that tandem TIRs are formed through sequence acquisition, the phenomenon can be readily explained. The genomes of maize and rice contain substantially more GC-rich sequences (or a more significant GC gradient) than that of Arabidopsis [32, 33]. Pack-MULEs in rice and maize demonstrate a strong preference for acquiring GC-rich sequences [19, 34]. Since the GC content of Pack-MULE TIR is similar or lower than the genomic average level [19], the acquisition of additional TIR sequence would be discriminated against in the genomes of rice and maize due to their relatively low GC content compared to gene sequences. In contrast, the GC content of sequences in dicot genomes is less variable [19, 35], such that TIR sequences are more likely to be acquired than their counterparts in the genomes of monocots. This might explain why there are more elements with tandem TIRs in tomato and Arabidopsis than in maize and rice.

Possible Competency Conferred by Tandemly Duplicated TIR

Among the elements with duplicated TIRs, two tomato elements have amplified to a certain degree. The PM-ZIBP elements have 13 copies with an additional copy that is associated

with only a single TIR. The type 2-46 elements have 29 copies without a corresponding copy with a single TIR, yet this particular TIR family is associated with single TIR elements harboring distinct internal regions with a lower copy number. Due to the coincidence of elements with tandem TIRs and single TIRs, it is clear that the presence of duplicated TIRs is not required for transposition, at least for these two TIR families. This raises the question whether the additional TIR has any role in transposition or successful amplification of these elements.

There are several explanations for the overrepresentation of elements with tandemly duplicated TIRs among the PMZIBP elements. Our analysis excludes the possibility that the duplicated TIR is acting as a filler DNA to allow the element to achieve an optimal size for amplification. It is worth pointing out that the "optimum size" might be present due to reasons other than size. If that is case, it also implies that the failure of PM-ZIBP-1 to amplify is unlikely attributable to its size. An alternative possibility points to the role of the internal region of PM-ZIBP since the acquired region appears more conserved than other internal sequence. According to this model, the PM-ZIBP is amplified because of the functional role of the acquired fragment. The element with single TIR (PM-ZIBP-1) failed to amplify due to its genomic location that is likely in heterochromatic region and not accessible for the transposition machinery. Nevertheless, due to the excision activity of DNA transposons and insertion polymorphism in the population, only a small subset of the transposons formed by transposition will be retained in the genome. If this model is valid, this may imply that PM-ZIBP-1 is the sole copy with a single TIR that has ever been present in the genome and one of the elements with duplicated TIRs must have been directly derived from PM-ZIBP-1. If this was the case, this would require PM-ZIBP-1 to be accessible in a certain way, which contradicts the original assumption of this model. Alternatively, there were other copies of PM-ZIBP with single TIR in the genomic location with

more open chromatin that gave birth to the element with tandem TIRs. In either case, one or more of the elements with single TIR was in an accessible location but failed to amplify while their counterparts with tandem TIRs significantly increased their copy number. In addition, the type 2-46 element has amplified to 29 copies without an apparently functional internal region, suggesting that the presence of gene fragments is not required for the amplification of elements with duplicated TIRs. Taken together, the overrepresentation of PM-ZIBP and type 2-46 elements with tandem TIRs likely reflects an elevated competency for transposition for these two specific MULE families. This could be achieved by increased recognition of the element by the transposase and/or interaction with transposase. This is in accordance with the fact that sequence conservation was observed for both internal and external TIRs.

Conclusion

Transposable elements are the major components of plant genomes. MULEs play important roles in plant genome evolution due to their high activity and potential to acquire and amplify gene fragments. In this study, we uncovered that formation of duplicated TIRs might have contributed to the success of some specific MULE elements. The availability of genomic sequences from multiple plant genomes allows us to conduct a comprehensive analysis which led to the following conclusions: (1) the formation of elements with additional TIR is not a rare event but only elements with duplicated TIRs on both terminus have significant mobility; (2) the genome of dicots harbor more elements with duplicated TIRs than that of monocots, and such difference might be attributed to the presence of GC-rich sequences in the genomes of monocots; (3) distribution of size versus copy number of MULEs (or Pack-MULEs) is periodic, suggesting the distance between the TIRs or the relative spatial position of TIRs may have a role in transposition; (4) in the elements with tandem TIRs, both TIRs appear to be subject to certain

constraints, and the presence of duplicated TIRs may confer certain mechanistic advantages for transposition. Such features may be utilized to create elements with elevated transposition activity.

REFERENCES

REFERENCES

[1] H. Ichikawa, K. Ikeda, W. L. Wishart, and E. Ohtsubo, "Specific binding of transposase to terminal inverted repeats of transposable element *Tn3*," Proceedings of the National Academy of Sciences of the United States of America, vol. 84, no. 23, pp. 8220–8224, 1987.

[2] L. Zhou, R. Mitra, P.W. Atkinson, A. B. Hickman, F. Dyda, and N. L. Craig, "Transposition of *hAT* elements links transposable elements and V(D)J recombination," Nature, vol. 432, no. 7020, pp. 995–1001, 2004.

[3] N. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, Mobile DNA II, ASM Press, Washington, DC, USA, 2nd edition, 2002.

[4] C. J. Huang, F. Heffron, J. S. Twu, R. H. Schloemer, and C. H. Lee, "Analysis of *Tn3* sequences required for transposition and immunity," Gene, vol. 41, no. 1, pp. 23–31, 1986.

[5] D. A. O'Brochta and P. W. Atkinson, "Transposable elements and gene transformation in non-drosophilid insects," Insect Biochemistry and Molecular Biology, vol. 26, no. 8-9, pp. 739–753, 1996.

[6] D. S. Robertson, "Characterization of a *Mutator* system in maize," Mutation Research, vol. 51, no. 1, pp. 21–28, 1978.

[7] Z. Yu, S. I.Wright, and T. E. Bureau, "*Mutator*-like elements in *Arabidopsis thaliana*: structure, diversity and evolution," Genetics, vol. 156, no. 4, pp. 2019–2031, 2000.

[8] N. Jiang, Z. Bao, X. Zhang, S. R. Eddy, and S. R. Wessler, "Pack-MULE transposable elements mediate gene evolution in plants," Nature, vol. 431, no. 7008, pp. 569–573, 2004.

[9] D. Holligan, X. Zhang, N. Jiang, E. J. Pritham, and S. R. Wessler, "The transposable element landscape of the model legume *Lotus japonicus*," Genetics, vol. 174, no. 4, pp. 2215–2228, 2006.

[10] F. Chalvet, C. Grimaldi, F. Kaper, T. Langin, and M. J. Daboussi, "*Hop*, an active *Mutator*-like element in the genome of the fungus *Fusarium oxysporum*," Molecular Biology and Evolution, vol. 20, no. 8, pp. 1362–1375, 2003.

[11] C. P. Marquez and E. J. Pritham, "*Phantom*, a new subclass of Mutator DNA transposons found in insect viruses and widely distributed in animals," Genetics, vol. 185, no. 4, pp. 1507–1517, 2010.

[12] M. Alleman and M. Freeling, "The *Mu* transposable elements of maize: evidence for transposition and copy number regulation during development," Genetics, vol. 112, no. 1, pp. 107–119, 1986.

[13] K. Hanada, V. Vallejo, K. Nobuta et al., "The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile," Plant Cell, vol. 21, no. 1, pp. 25–38, 2009.

[14] N. Juretic, D. R. Hoen, M. L. Huynh, P. M. Harrison, and T. E. Bureau, "The evolutionary fate of MULE-mediated duplications of host gene fragments in rice," Genome Research, vol. 15, no. 9, pp. 1292–1297, 2005.

[15] M. I. Benito and V. Walbot, "Characterization of the maize *Mutator* transposable element MURA transposase as a DNA binding protein," Molecular and Cellular Biology, vol. 17, no. 9, pp. 5165–5175, 1997.

[16] M. N. Raizada, M. I. Benito, and V. Walbot, "The *MuDR* transposon terminal inverted repeat contains a complex plant promoter directing distinct somatic and germinal programs," Plant Journal, vol. 25, no. 1, pp. 79–91, 2001.

[17] P. S. Schnable, D. Ware, R. S. Fulton et al., "The B73 maize genome: complexity, diversity, and dynamics," Science, vol. 326, no. 5956, pp. 1112–1115, 2009.

[18] The Potato Genome Sequencing Consortium, "Genome sequence and analysis of the tuber crop potato," Nature, vol. 475, no. 7355, pp. 189–195, 2011.

[19] N. Jiang, A. A. Ferguson, R. K. Slotkin, and D. Lisch, "Pack-*Mutator*-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition," Proceedings of the National Academy of Sciences of the United States of America, vol. 108, no. 4, pp. 1537–1542, 2011.

[20] R. C. Edgar and E. W. Myers, "PILER: identification and classification of genomic repeats," Bioinformatics, vol. 21, no. 1, pp. i152–i158, 2005.

[21] S. B. Chang, L. K. Anderson, J. D. Sherman, S. M. Royer, and S. M. Stack, "Predicting and testing physical locations of genetically mapped loci on tomato pachytene chromosome 1," Genetics, vol. 176, no. 4, pp. 2131–2138, 2007.

[22] D. Szinay, S. B. Chang, L. Khrustaleva et al., "High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6," Plant Journal, vol. 56, no. 4, pp. 627–637, 2008.

[23] S. D. Tanksley, M. W. Ganal, J. P. Prince et al., "High density molecular linkage maps of the tomato and potato genomes," Genetics, vol. 132, no. 4, pp. 1141–1160, 1992.

[24] T. M. Fulton, R. Van der Hoeven, N. T. Eannetta, and S. D. Tanksley, "Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants," Plant Cell, vol. 14, no. 7, pp. 1457–1467, 2002.

[25] J. Felsenstein, “Evolutionary trees from DNA sequences: a maximum likelihood approach,” Journal of Molecular Evolution, vol. 17, no. 6, pp. 368–376, 1981.

[26] M. Kimura, “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences,” Journal of Molecular Evolution, vol. 16, no. 2, pp. 111–120, 1980.

[27] N. Jiang, D. Gao, H. Xiao, and E. Van Der Knaap, “Genome organization of the tomato *sun* locus and characterization of the unusual retrotransposon *Rider*,” Plant Journal, vol. 60, no. 1, pp. 181–193, 2009.

[28] B. Morgenstern, “DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment,” Bioinformatics, vol. 15, no. 3, pp. 211–218, 1999.

[29] T. Sasaki, “The map-based sequence of the rice genome,” Nature, vol. 436, no. 7052, pp. 793–800, 2005.

[30] D. Lisch and N. Jiang, “*Mutator* and MULE transposons,” in Handbook of Maize: Genetics and Genomics, pp. 277–306, Springer, New York, NY, USA, 2009.

[31] D. Lisch, “*Mutator* transposons,” Trends in Plant Science, vol. 7, no. 11, pp. 498–504, 2002.

[32] N. Carels and G. Bernardi, “Two classes of genes in plants,” Genetics, vol. 154, no. 4, pp. 1819–1825, 2000.

[33] G. K. S.Wong, J.Wang, L. Tao et al., “Compositional gradients in Gramineae genes,” Genome Research, vol. 12, no. 6, pp. 851–856, 2002.

[34] A. A. Ferguson and N. Jiang, “Pack-MULEs: recycling and reshaping genes through GC-biased acquisition,” Mobile Genetic Elements, vol. 1, no. 2, pp. 1–4, 2011.

[35] P. F. Cavagnaro, D. A. Senalik, L. Yang et al., “Genomewide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.),” BMC Genomics, vol. 11, no. 1, article 569, 2010.

CHAPTER 4:

Repetitive Sequence Landscape of the Basal Eudicot Genome Sacred Lotus (*Nelumbo nucifera*)

Abstract

Transposable elements (TEs) are pervasive among eukaryotes and are often the largest component in these genomes. These sequences can introduce genetic variation that possesses adaptive and evolutionary potential; and therefore, their identification remains an integral part of genome studies. The sequencing of the genome of sacred lotus, a basal eudicot, has allowed us to characterize its repetitive content. Here, we report that 59% of the genome is composed of repetitive sequences, and the majority of them (50% of the genome) are identifiable TEs. Analysis of the TE content and diversity in sacred lotus revealed unique composition of TEs compared to other plant genomes characterized so far. Among LTR elements, *Copia*-like and *Gypsy*-like elements demonstrate comparable coverage and copy number, a distribution not typical of numerous plant genomes where *Gypsy*-like elements are the dominant type of LTR elements. The relative abundance of *Copia*-like elements is in part due to the presence of families with non-canonical LTR ends (15.6% of total LTR content), many of which have not been described previously. Sacred lotus also contains the highest coverage and copy number of *hAT* elements among all genomes sequenced to date. In addition, the genome contains 1447 Pack-MULEs and provides the first evidence for the acquisition preference of GC-rich sequences by Pack-MULEs outside the grasses.

Introduction

Sacred lotus (*Nelumbo nucifera*) is a perennial aquatic plant that belongs to the Nelumbonacee family and is found throughout Asia and northern Australia. It provides economic value as an ornamental and food crop in Asia. In addition, parts of lotus such as the flowers, roots and rhizomes have been used for medicinal purposes (Duke, 2002; Shen-Miller, 2002; Shen-Miller et al., 2013). The genome of sacred lotus variety "China Antique" was recently sequenced (Ming et al., 2013) using Illumina and 454 technologies. The final genome assembly (804 Mb) is 86.5% of the estimated 929 Mb lotus genome (Diao et al., 2006). This provides an excellent resource for the evolutionary analysis of eudicots and comparative studies between dicots and monocots since sacred lotus phylogenetically lies outside the core eudicots. Phylogenetic comparisons between grape and sacred lotus suggests that it is a better model than the grape genome for inferences about the common ancestors of eudicots (Ming et al., 2013). Genomic analysis reveals that the sacred lotus genome lacks the γ triplication event seen in all core eudicots and shows a remarkably low substitution rate and a higher retention of duplicated genes compared with most other angiosperm genomes (Ming et al., 2013).

Transposable elements (TEs) are genetic sequences first discovered 50 years ago by Barbara McClintock. TEs move from one genomic location into another and in the process increase their copy numbers. According to the intermediate form of transposition used by the specific element, TEs are generally classified into two major groups: Class I or RNA elements which transpose via an RNA intermediate using a copy-and-paste mechanism, and Class II or DNA elements that transpose via a DNA intermediate using a cut-and-paste mechanism (Wicker et al., 2007; Kapitonov and Jurka, 2008). In addition, the coding capacity of elements for proteins involved or comprising the transpositional machinery allows for further classification of

elements into autonomous elements, which code for these proteins, or non-autonomous elements, which rely on the cognate autonomous elements for movement within the genome.

Due to their capacity to multiply within a host and their prevalence among plant and animal genomes, TEs have previously been implicated to contribute significantly to increases in genome size (Bennetzen and Kellogg, 1997; Ammiraju et al., 2007; Bennetzen, 2007; Zuccolo et al., 2007). In some instances, TE may constitute the largest part of the genome (Schnable et al., 2009; Brenchley et al., 2012; Mayer et al., 2012). Dramatic differences exist in the content and diversity of different TE types between organisms. While animal and insect genomes typically contain a higher proportion of non-LTR retrotransposons (Lander et al., 2001; Chinwalla et al., 2002; Nene et al., 2007), the LTR retrotransposons typically dominates the TE landscape in plants (Rice Sequencing Project, 2005; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010). Various computational and biological analyses of genomic information have demonstrated the critical roles of transposons in many aspects of genome evolution, gene expression and regulation (Jordan et al., 2003; Muotri et al., 2007). Widespread in plant genomes is a special type of DNA element called Pack-MULE which carry genes or gene fragments (Jiang et al., 2004). These elements have been implicated in the generation of new open reading frames and the regulation of the expression of their parental genes (Hanada et al., 2009; Jiang et al., 2011). Taken together, these studies indicate that TEs are prevalent in plants and are actively interacting with other components in their host genomes.

Prior to the sequencing of the sacred lotus genome, no information was available regarding any aspect of its repetitive sequence content. Due to its position in the eudicot phylogeny, sacred lotus may offer important biological contributions in terms of its TE content, structure and diversity. Here we report the results of a comprehensive computer-assisted

identification and analysis of the repetitive content in the available sequence of sacred lotus. Our results show the exceptional contribution of atypical *Copia* LTR families with non-canonical end sequences. In addition, the genome appears to have unprecedented amplification of the *hAT* superfamily and suggests that GC-preferential acquisition by Pack-MULEs occurs in eudicots.

Methods

Construction of repeat library

The current assembly of 804 Mb scaffold contains 707 Mb of contig sequence and 97 Mb of sequencing gaps. The 707 Mb contig sequence was downloaded from Genbank at the NCBI database (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AQOG01) for further analysis. Repetitive sequences were mined using a variety of approaches. To identify LTR elements, the program LTRharvest (Ellinghaus et al., 2008) was used (parameters: -minlenltr 80 -maxlenltr 6000 -mindistltr 300 -mintsd 4 -maxtsd 6 -motif tgca -similar 90) and the resulting elements were further screened using LTRdigest (Steinbiss et al., 2009) to determine the presence of a poly purine tract (PPT) or primer binding site (PBS). Only elements that contain a PPT or PBS were retained for further analysis. To determine the precise boundary of LTR elements, 100 bp of flanking sequences (5' and 3' ends) were retrieved and aligned using DIALIGN2 (Morgenstern, 1999). Elements wherein ≥ 50 bp of the flanking sequences were alignable (with similarity score ≥ 60%) were excluded. This is because for most LTR elements, only the LTRs, not the flanking sequences, should be alignable. To reduce the redundancy, examplar elements were selected using the "examplar_maker.pl" script from the MITE-Hunter package (Han and Wessler, 2010). The above procedure was initially performed to recover only elements with terminal sequences 5'-TG..CA-3'. However, after the recovery of elements with non-canonical terminal sequences during manual curation (see below), the procedure was repeated without the

definition of terminal motif to consider other motifs and additional examplars of LTR elements with non-canonical ends were manually verified. To search for LTR elements with non-canonical ends in grape (*Vitis vinifera*), its genomic sequence was downloaded from Phytozome (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Vvinifera/).

Non-autonomous DNA elements were mined using the MITE-Hunter package with parameters as recommended. Terminal inverted repeats (TIRs) of *Mutator*-like elements were identified using Pairwise Alignment of Long Sequences (PALS; Edgar and Myers, 2005) and manual curation as described previously (Ferguson and Jiang, 2012). The sequences of exemplars of LTR elements, non-autonomous DNA elements, and MULE TIRs were then used to mask the genomic sequence and the repetitive sequences in the unmasked portion of the genomic DNA were further identified in a second mining step using RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html). The output of RepeatModeler contains both known and unknown repeats. The resulting sequences were first filtered to remove putative gene families using BLASTX and sequences matching non-TE genes proteins ($E < 10^{-5}$) were removed. The remaining sequences where the copy number is > 1000 or the genome coverage is $\geq$ 0.05% were manually curated to determine their identity and the 5' and 3' boundaries. This was done in a stepwise process. First, the relevant sequences were initially used to search and retrieve at least 10 hits (BLASTN, $E < 10^{-10}$) with the corresponding 100 bp of 5' and 3' flanking sequences. Second, recovered sequences were then aligned using DIALIGN2 (Morgenstern, 1999), to determine the possible boundary between elements and their flanking sequences. In this case, a boundary was defined as the position to which sequence homology is conserved over more than half of the aligned sequences. Finally, sequences with defined boundaries were examined for the presence of target site duplication (TSD). To classify the relevant TEs, features

in the terminal ends and TSD were used. Each transposon family is associated with distinct features in their terminal sequences and TSD which can be used to identify an element (Wicker et al., 2007). To identify inverted or direct repeats in the terminal sequence, the putative terminal sequences were aligned using "gap" in the GCG package. Fragmented sequences identified by RepeatModeler were joined to derive a complete element.

Manually curated sequences were compared to the unknown repeats using RepeatMasker. Sequences matching the curated sequences were excluded if they belong to the same family. Families were defined as follows: elements that share 80% or higher similarity across at least 90% of their length. If a repetitive sequence matches the curated sequences without reaching the above criteria, this sequence was retained and was considered to belong to a new family within the same superfamily.

Estimation of copy number and genomic coverage

The entire repeat library, containing elements with curated and non-curated ends, was used to mask the genomic sequence to determine TE coverage and copy number. If an element in the genomic sequence matched a sequence in the repeat library over the entire sequence, or if the truncation was less than 20 bp on each end, this copy was considered to be intact. Otherwise it was considered as a truncated sequence or half of a copy. Copy number is reported only for families that are curated, that is, the ends are known and verified manually. The genome coverage of TEs was estimated as the total sequence masked by each superfamily with overlapping regions between different entries only calculated once.

Pack-MULEs

Pack-MULE elements (*Mutator*-like elements carrying genes) were identified as described previously (Hanada et al., 2009). To determine which Pack-MULEs are expressed, the

element sequences were compared with EST database (Ming et al., 2013). A Pack-MULE is considered expressed if it matches an EST sequence with ≥97% similarity and the Pack-MULE coordinate is the best hit for the EST in the genome. The identification of the parental origin of the sequences captured by Pack-MULEs was conducted as described previously (Jiang et al., 2011). For an individual Pack-MULE, the sequence with the highest similarity score (BLASTN, $E=1e^{-10}$), and was not associated with a MULE terminal inverted repeat (TIR), was considered as the parental copy of the internal sequence in a Pack-MULE.

To calculate the GC content of MULEs and Pack-MULEs, nested TE insertions were first curated and removed from the element sequence. Determination of GC content of parental genes was conducted after masking with the repeat library that excluded Pack-MULEs. To calculate GC gradient along MULE sequences, the TIR sequences (on both ends of the elements) and the internal region (the sequence between the TIRs) were divided into two and 12 equal-sized bins, respectively. A custom perl script was used to determine the GC content of each bin. Comparisons of GC content between groups were performed using the R package (http://www.r-project.org). Arabidopsis (*Arabidopsis thaliana*) gene sequences were downloaded from The Arabidopsis Information Resource 10 (http://www.arabidopsis.org) while the *Lotus japonicus* gene sequences were downloaded from *Lotus japonicus* Genome Sequencing Project (http://www.kazusa.or.jp/lotus/). Rice, Arabidopsis and *Lotus japonicus* genes were classified as negative, positive or moderate genes based on GC content along the direction of transcription as described previously (Jiang et al., 2011).

Phylogenetic analysis

To search for autonomous *hAT* elements, both curated and non-curated autonomous *hAT* families were used to search the genome. Copies that are truncated by no more than 15 bp on

each end of the element or are over 2.5 Kb in length were retained and compared against a database containing known *hAT* transposase to identify elements containing the conserved motif 3 (Kempken and Windhofer, 2001; Lazarow et al., 2012).

Intact LTR elements were mined using the sequences in repeat library and verified to contain the internal sequence, flanked by LTR and were associated with a 5-bp target site duplication. The conserved integrase core domain, which contains the RNase H fold catalytic motif, of representative LTR elements from each family was retrieved.

Sequences of conserved integrase core domain or motif 3 from LTR elements and *hAT* transposase, respectively, 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. Phylogenetic trees were generated using the maximum-likelihood method. Support for the internal branches of the phylogeny was assessed using 500 bootstrap replicates using MEGA (http://www.megasoftware.net).

Results

Repeat content and diversity

Identification of repeats was performed using the 707 Mb contig sequence of the sacred lotus genome that excluded the sequencing gaps contained in the 804 Mb final assembly. The contig sequences account for 76% of the estimated lotus genome (929 Mb) (Ming et al., 2013). Transposable elements (TEs) and other repetitive sequences were mined using a combination of structure-based and homology-based approaches (see methods). Approximately 59.3% of the genome is derived from repetitive sequences and 50% of it is composed of recognizable TEs (Table 4.1). The 353 Mb of repetitive genomic sequence included at least 1.26 million copies of TEs. This copy number is underestimated since families whose precise boundaries cannot be defined were not included in the copy number estimation which was done to avoid grossly overestimating the copy number from fragmented families whose ends are not known.

The majority of recognizable TE is contributed by Class I/RNA elements (65.6%), a familiar phenomenon across the plant kingdom where the amplification of LTR elements has been suggested to contribute to genome size expansion. The majority of LTR retrotransposons are classified into two major superfamilies: *Copia* and *Gyspy* depending on the arrangement of the genes in the *pol* region. In sacred lotus, the LTR retrotransposon content (26.4%) is comprised by a comparable number and coverage of *Copia* and *Gypsy* elements (77951 and 72624 copies, respectively; Table 4.1). This is not typical among plant genomes wherein the *Gypsy* content usually outnumbers the *Copia* content (Table 4.2A). Among the 29 genomes with available *Gypsy*:*Copia* ratio, *Gypsy* elements occupy twice as much of the genomic fraction as that for *Copia* elements in 17 (59%) genomes. The only other plant genomes except lotus that seem to share a *Gypsy:Copia* ratio close to 1.0 are *Brassica rapa*, peach, strawberry and sweet

orange (1.10, 1.16, 1.20 and 1.25 respectively). However, the ratio is still closest to 1.0 in lotus. Genomes with the highest *Gypsy*:*Copia* ratio are *Selaginella moellendorffii* and papaya (7.81 and 5.05, respectively). Two plant genomes, banana and date palm, display the opposite extreme observed in most plants, where the *Copia* content outnumbers the *Gypsy* content (Table 4.2A). In addition, the scared lotus genome contains a high coverage of non-LTR retrotransposons (6.4%), which are predominantly contributed by LINEs (Table 4.2A). The two other plant genomes containing such a high non-LTR retrotransposon load are that of apple (7.95%) and banana (5.41%) (Velasco et al., 2010; D'Hont et al., 2012). Taken together, these results suggest a high activity and/or retention of non-LTR and *Copia* retrotransposons in the evolution of sacred lotus genome in comparison to most other plants. This may result from a massive amplification combined with a lack or low silencing from the genome for these TE types, overall conferring non-LTRs an efficient evasion strategy for various silencing mechanisms.

Class II elements comprised about 17% of the genome. This level of DNA TE content is notable and only observed before in rice, wheat and soybean (Schmutz et al., 2010; Brenchley et al., 2012; Jiang and Panaud, 2013). The largest contributors to DNA element content are *hATs* (~7% of the genome), followed by *Helitrons* (3.8%). At least 170635 copies of the *hAT* elements were detected, making sacred lotus the most abundant in both genome *hAT* coverage and copy number among all plant genomes sequenced to date (Table 4.2B). Although most major DNA transposon families were identified, the *Tc1/Mariner* superfamily is absent. The absence of *Tc1/Mariner* elements has been reported in three other plant genomes: banana, grape, and *Selaginella* (Velasco et al., 2007; Banks et al., 2011; D'Hont et al., 2012), suggesting that the loss of this superfamily in plants is not unusual.

LTR elements with non-canonical ends

In both major classes of LTR elements (*Copia* and *Gypsy*), the canonical LTR repeats found at the ends of these elements typically start with 5'-TG and ends in CA-3' (Kumar and Bennetzen, 1999), which forms a short inverted repeat. In fact, many computer-assisted searches use this criterion in *de novo* searches for LTRs (McCarthy and McDonald, 2003; Steinbiss et al., 2009). However, in sacred lotus, 84 LTR families comprising eight different non-canonical LTR ends were found (Table 4.3). While the majority of these elements are *Copia*, two *Gypsy* and one TRIM families contained these atypical ends. Among all groups of LTR elements with non-canonical ends, four of them (TGCT, TGGA, TACA and TGTA) harbor mutations in one nucleotide, and the remainder (TGGT, TACT, TATA and TGTT) harbor mutations in two nucleotides. Obviously, the ones with a single mutation are more abundant than those with two mutations (Table 4.3). In addition, variations in constraints are observed in terms of the mutation allowed at different sites: a) no mutation was detected at the most 5' nucleotide (always "T") b) second nucleotide at 5' end is a purine (G or A), c) the second nucleotide at 3' end is the most flexible, and can be C, G, or T and d) the last nucleotide can be either "A" or "T".

The most abundant non-canonical end type is found in elements where the LTR starts with the canonical 5'-TG but ends in CT-3' (referred to as TGCT LTR), where the most terminal nucleotide is not inverted. This LTR end type includes 25 *Copia* families making up an estimated 5658 copies or 2.3% of the genome. These elements are flanked by LTR sequences with the non-canonical ends that range from 199 to 408 bp in size, a 5 bp TSD at the insertion site, a primer binding site (PBS) and a polypurine tract (PPT) (Figure 4.1). The PBS, located immediately downstream of the 5' LTR, and the PPT, located immediately upstream of the 3' LTR, are conserved sequences motifs which are important for replication and amplification of the element.

Table 4.1. Repeat content of the sacred lotus genome.

Class	Subclass	Superfamily	Copy number[a]	Copy number[b]	Genome Coverage
Class I	LTR	*Copia*	29928	77951	12.551
		Gypsy	26562	72624	13.654
		Other LTR	n/a	n/a	0.194
	non-LTR	LINE	5621	42331	6.192
		SINE	5717	8551	0.129
		Other non-LTR	n/a	n/a	0.083
Total Class I			67828	203734	32.803
Class II		CACTA	44	1038	0.315
		hAT	83043	170635	7.315
		MULE	51450	93733	2.634
		PIF/Tourist	62126	103755	3.213
		Helitron	10740	72489	3.757
		DNA/unknown	n/a	n/a	0.082
Total Class II			207403	441650	17.317
Total TE			275231	645384	50.120
unknown			165921	324267	6.517
simple sequence			118452	118452	0.886
low complexity			174823	174823	1.791
Total repeats			734427	1262926	59.314

[a] Copy number based on detected elements that contain terminal sequences. Elements with non-curated ends were not included in copy number estimation

[b] Copy number based on all detected fragments. Elements with non-curated ends were not included in copy number estimation

Table 4.2A. RNA TE information of various sequenced plant genomes.

Common name	*Scientific name*	*Copia* (%)	*Gypsy* (%)	*Gypsy*:*Copia* ratio	Non-LTR RT (%)
DICOTS					
tomato	*Lycopersicum esculentum*	6.3	19.7	3.13	0.9
Potato	*Solanum tuberosum*	3.8	15.2	4.0	1.0
soybean	*Glycine max*	12.47	29.52	2.37	0.25
Apple	*Malus domestica*	6.72	30.98	4.61	7.95
pigeon pea	*Cajanus cajan*	6.22	11.79	1.89	1.17
sacred lotus	***Nelumbo nucifera***	**12.55**	**13.65**	**1.09**	**6.4**
castor bean	*Ricinus communis*	4.77	11.45	2.4	0.14
poplar	*Populus trichocarpa*	1.79	6.96	3.89	0.54
papaya	*Carica papaya*	5.5	27.8	5.05	1.1
canola	*Brassica rapa*	4.85	5.34	1.1	3.28
Peach	*Prunus persica*	8.6	9.97	1.16	0.63
Jatropha curcas	*Jatropha curcas*	8.03	19.6	2.44	n/a
barrel medic	*Medicago trunculata*	4.1	5.7	1.39	2.3
Grape	*Vitis vinifera*	6.12	17.96	2.93	0.82
cucumber	*Cucumis sativus*	n/a	n/a	n/a	1.75
strawberry	*Fragaria vesca*	5.33	6.39	1.2	0.45
sweet orange	*Citrus sinensis*	7.84	9.77	1.25	0.4
Cacao	*Theobroma cacao*	n/a	n/a	n/a	n/a
thale cress	*Arabidopsis thaliana*	3.56	5.74	1.61	1.49
saltwater cress	*Thellungiella parvula*	1.92	2.46	1.28	1.1
MONOCOTS					
date palm[a]	*Phoenix dactylifera*	3.1	1.4	0.45	n/a
Maize	*Zea mays*	23.7	46.4	1.96	1.0

Table 4.2A (cont'd)

Barley	*Hordeum vulgare*	8.46	17.96	2.12	0.53
wheat[a]	*Triticum aestivum*	17.39	44.03	2.53	0.82
sorghum	*Sorghum bicolor*	5.18	19	3.67	0.04
Rice	*Oryza sativa*	3.6	15.5	4.31	1.9
banana	*Musa acuminata*	25.58	11.45	0.45	5.41
foxtail millet	*Setaria italica*	7.18	22.14	3.03	1.98
Brachypodium	*Brachypodium distachyon*	4.86	16.05	3.3	1.94
GYMNOSPERM					
Norway spruce[a]	*Picea abies*	16.0	35.0	2.19	1.0
LYCOPHYTE					
S. moellendorfii	*Sellaginella moellendorfii*	2.7	21.1	7.81	n/a

[a] Data is based on results from unassembled reads
n/a – data not available for specific family

Table 4.2B. Genome information and DNA TE content of various sequenced plant genomes.

Common name	*Scientific name*	Assembled size (Mb)	Total TE (%)	DNA TE (%)	*hAT* (%)	Reference
DICOTS						
tomato	*Lycopersicum esculentum*	760	63	0.9	0.1	(Sato et al., 2012)
Potato	*Solanum tuberosum*	727	62	3.9	0.2	(Xu et al., 2011)
soybean	*Glycine max*	973	59	16.5	0.04	(Schmutz et al., 2010)
Apple	*Malus domestica*	604	52	6.6	0.35	(Velasco et al., 2010)
pigeon pea	*Cajanus cajan*	606	52	4.5	n/a	(Varshney et al., 2011)
sacred lotus	***Nelumbo nucifera***	**804**	**50**	**17.3**	**7.32**	
castor bean	*Ricinus communis*	350	50	0.91	n/a	(Chan et al., 2010)
poplar	*Populus trichocarpa*	485	44	2.4	0.02	(Tuskan et al., 2006)
papaya	*Carica papaya*	370	43	0.21	n/a	(Ming et al., 2008)
canola	*Brassica rapa*	284	40	3.2	2.87	(Wang et al., 2011)
Peach	*Prunus persica*	227	37	9.1	n/a	(Verde et al., 2013)
Jatropha curcas	*Jatropha curcas*	285	37	2.0	n/a	(Sato et al., 2010)
barrel medic	*Medicago trunculata*	262	31	3.4	0.2	(Young et al., 2011)
Grape	*Vitis vinifera*	477	29	1.8	1.03	(Velasco et al., 2007)
cucumber	*Cucumis sativus*	244	24	1.2	n/a	(Huang et al., 2009)
strawberry	*Fragaria vesca*	210	23	6.4	0.64	(Shulaev et al., 2010)
sweet orange	*Citrus sinensis*	320	20	2.3	0.36	(Xu et al., 2012)
Cacao	*Theobroma cacao*	327	16	n/a	n/a	(Argout et al., 2010)
thale cress	*Arabidopsis thaliana*	119	16	5.25	0.53	(Hollister et al., 2011)
saltwater cress	*Thellungiella parvula*	137	7	1.2	n/a	(Dassanayake et al., 2011)
MONOCOTS						
date palm[a]	*Phoenix dactylifera*	380	n/a	n/a	n/a	(Al-Dous et al., 2011)

Table 4.2B (cont'd)

Maize	*Zea mays*	2048	84	8.6	1.1	(Schnable et al., 2009)
Barley	*Hordeum vulgare*	4560	84	5.0	0.2	(Mayer et al., 2012)
wheat[a]	*Triticum aestivum*	n/a	79	14.9	0.04	(Brenchley et al., 2012)
sorghum	*Sorghum bicolor*	730	62	7.5	0.02	(Paterson et al., 2009)
Rice	*Oryza sativa*	374	45	20.4	1.6	(Jiang and Panaud, 2013)
banana	*Musa acuminata*	523	44	1.24	n/a	(D'Hont et al., 2012)
foxtail millet	*Setaria italica*	423	40	9.4	0.59	(Zhang et al., 2012)
Brachypodium	*Brachypodium distachyon*	272	28	4.8	0.24	(Vogel et al., 2010)
GYMNOSPERM						
Norway spruce[a]	*Picea abies*	12000	70	1.0	n/a	(Nystedt et al., 2013)
LYCOPHYTE						
spikemoss	*Sellaginella moellendorfii*	212	37	1.8	0.02	(Banks et al., 2011)

[a] Data is based on results from unassembled reads
n/a – data not available

An example of this LTR end type is the *Copia* family NN00209 and is present in at least 424 copies (Figure 4.2 and Table 4.3). In total, the non-canonical *Copia* elements contribute to 3.74% of the genome, which account for 30% of all *Copia* elements. The two families of *Gypsy* elements with TACA terminal sequence) only contribute to 0.038% of the genome, so they are not significantly amplified.

To determine the relationship between the non-canonical elements and other elements, a phylogenetic analysis was performed using the conserved integrase catalytic domain of TGCT LTR elements as well as other LTR elements in lotus and other characterized *Copia* elements in other species. Our analysis shows that the majority of the TGCT LTR families in sacred lotus are closely related to each other and form a single clade (Figure 4.3). Although the TGCT LTR families are predominant in this clade, it also contains a few families with the canonical TGCA end, four of the five non-canonical TGGT, and one TGGA LTR families. This suggests that mutations within this clade may give rise to other non-canonical end families and a reversion to the typical TGCA LTR type. It appears that this clade is distantly related to the *Tos17* element from rice which was the first reported LTR element with a non-canonical end, TGGA (Hirochika et al., 1996). Meanwhile, four TGCT LTR families (NN00215, NN00219, NN00221 and NN00225) grouped separately in a clade that contained *AtRE1* element (TATA end) from Arabidopsis (Kuwahara et al., 2000). In addition, this clade includes other sacred lotus LTR families with five non-canonical LTR types (TACA, TACT, TATA, TGTA and TGTT; Figure 4.3). This suggests a high frequency of mutation has given rise to multiple types of non-canonical LTR families within this clade. Since grape is the closest eudicot sequenced genome to sacred lotus, *Copia* elements from grape including five TGCT families identified in this study were also analyzed. Overall, the TGCT *Copia* families in grape grouped distinctly from those in

Table 4.3. Family, copy number and genome coverage of different LTR ends.

End	# *Copia* families	# *Gypsy* families	# TRIM families	Total families	Total copy number	Total coverage
TGCA	112	111	2	225	48009	20.75
TGCT	25	0	0	25	5658	2.29
TGGA	6	0	1	7	2210	0.31
TACA	13	2	0	15	697	0.34
TGTA	3	0	0	3	566	0.34
TGGT	5	0	0	5	400	0.22
TACT	8	0	0	8	359	0.19
TATA	7	0	0	7	186	0.13
TGTT	1	0	0	1	18	0.01
total	180	113	3	296	58103	24.58


```
5'LTR     TCATGGTGTCTGTTGAAGATTAATTAGTTTTATGATGTTTTTCTGAGT
          | | || |  ||||||||||||||||||||||||||||||||||||||
3'LTR    CTTGAGGGGGAGTGTTGAAGATTAATTAGTTTTATGATGTTTTTCTGAGT

5'LTR    AGATGTTTTATGTTTGTTGTTAGATAGGAGGTAGGGACCACAGTAGTTAT
         ||||  |||||||||  ||||||||||||||||||  |||||||  | |||||||
3'LTR    AGATATTTTATGTTAGTTGTTAGATAGGAGGTGGGGACCAAAATAGTTAT

5'LTR    GTGAGTCCAAATTTTATTATGTTTTTTGAG...TTTATATGCTGTTTTAT
         ||||||||||||||||||||  ||||||||   | ||||||  ||||||||
3'LTR    GTGAGTCCAAATTTTATTATGCTTTTTGAGTGTTATATATGTTGTTTTAT

5'LTR    GTTTTGCCGCCCTATAAGGGCATAGTTGTCTTTGTACAATTCTATTATTT
         ||||||||  | ||||||||||||||||||||||||||   ||||||||||
3'LTR    GTTTTGCCCCTCTATAAGGGCATAGTTGTCTTTGTACGGCTCTATTATTT

5'LTR    TAGTATAAATAATAAGTGAAGGGGGCGGCCAGATCATTGCTCTCTTTTCT
         ||||||||||||||||||||||||||||||||||||||||||||||||||
3'LTR    TAGTATAAATAATAAGTGAAGGGGGCGGCCAGATCATTGCTCTCTTTTCT

5'LTR    TTCAACTTCTCTTCTTCTCTTCTTTTCCTTCTCTCTTCATTTCTTCTTCA
         ||||||||||||||||  |||||  ||||||||||||  ||   ||||||
3'LTR    TTCAACTTCTCTTCTTCTTTTCTTCTCCTTCTCTCTTCCTTCTCTCTTCA

5'LTR    AGAAACTACTTTCTTTCATATTTTTTAATTCAGCAGATTTCTACTTGGTA
         ||||||||||||||||||||||||||||||||||||||||||||  ||
3'LTR    AGAAACTACTTTCTTTCATATTTTTTAATTCAGCAGATTTCTACTGTGTC

5'LTR    TCAAAGCATCGGCTAGGGTTTGTTGAAGCWTTCTCATGCAGCCCTAGATC
         |||  |
3'LTR    TCAGA
```

Figure 4.1. Sequence alignment of NN00206 illustrating the LTR sequence, TSD, primer binding site (PBS) and polypurine tract (PPT). Blue text indicates 10bp initial LTR sequence at the each terminal and red indicates 5 bp TSD. The shaded text indicates the sequence that matches a Gly-tRNA.

```
NN00206-7930    CACTAGAATCTGTTGA---------TCTACTGAATCTTTCA
NN00206-9299    CAAAGAAAAATGTTGA---------TCTACTAAAAAAGTTG
NN00206-10517   GTAATCATTTTGTTGA---------TCTACTCATTTTAATA
NN00206-10869   GTTAGAAATGTGTTGA---------TCTACTAAATGCATCA
NN00206-14103   AAAGAATGTATGTTGA---------CCTACTATGTATAAAT
NN00206-14546   TCATGGTGTCTGTTGA---------TCTACTGTGTCTCAGA
NN00206-18031   AAATGGAAGTTGTTGA---------TCTACTGAAGTGGTAT
NN00206-20935   GATGGTTTAGTGTTGA---------CCTACTTTTAGAATAA
NN00206-21176   CCTAGCACTCTGTTGA---------TCTACTCACTCTATTG
NN00206-24944   GTATGCCTCTTGTTGA---------TCTACTCCTCTTGTTT
```

Figure 4.2. Sequence alignment of LTR ends and TSD of the *Copia* family NN00206. Text in blue indicates 6 bp outermost LTR sequence and red indicates 5 bp TSD.

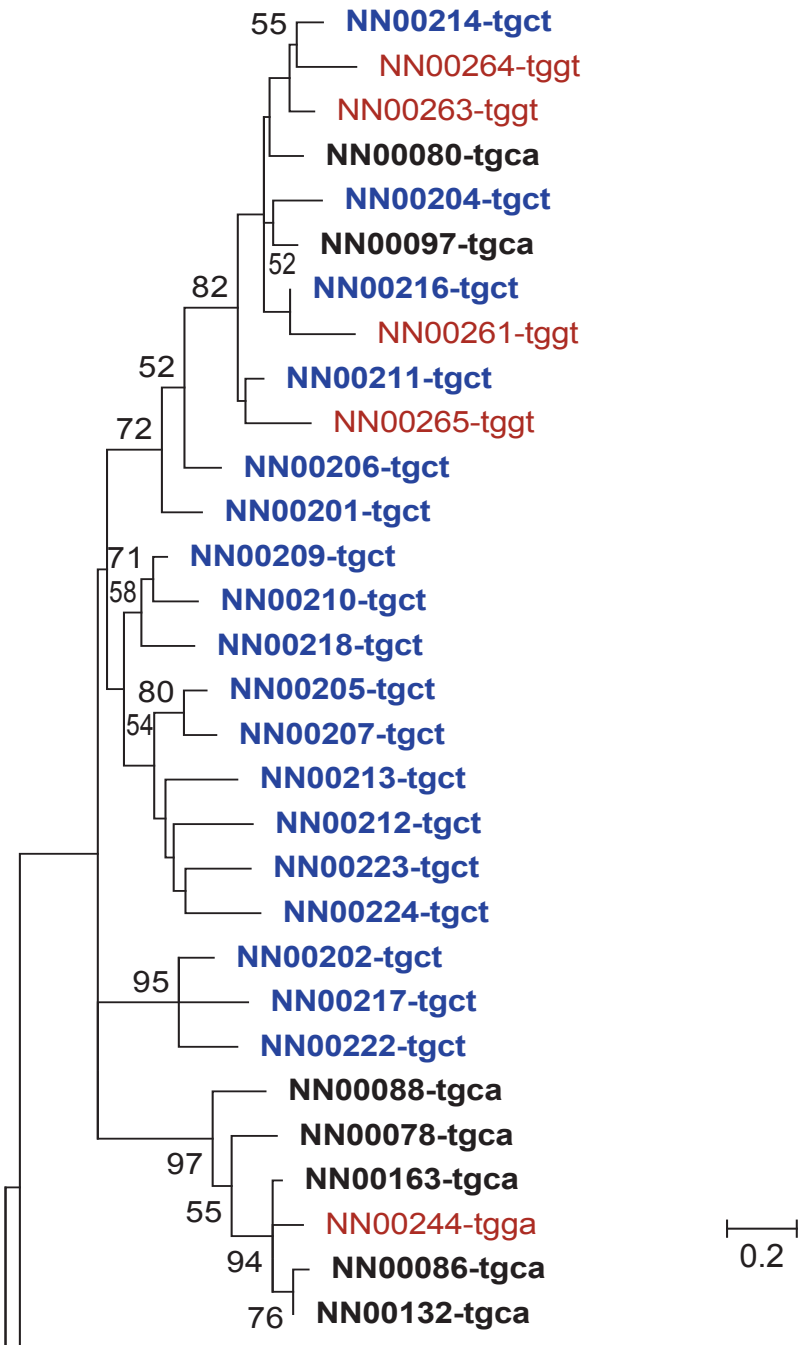


Figure 4.3. Phylogenetic analysis of TGCT LTR using the conserved integrase catalytic core domain. Bootstrap values are indicated as a percentage of 500 replicates. Shown are: sacred lotus sequences TGCT LTR (blue bolded), TGCA LTR (black bolded), other types of non-canonical LTR (red), Sireviruses (green), TGCT grape LTR (light blue), TGCA grape LTR (brown), other species LTR (black).

Figure 4.3 (cont'd)

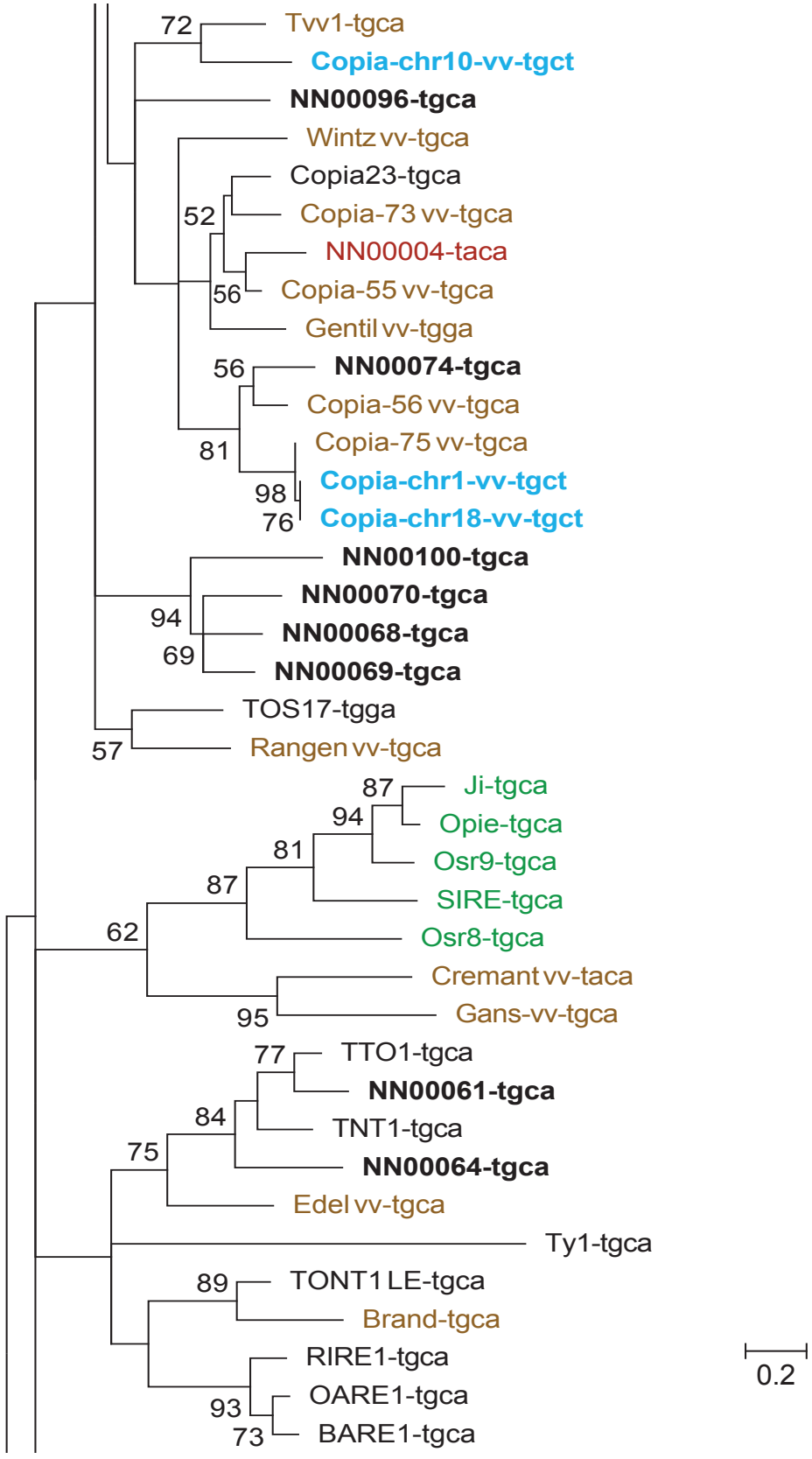
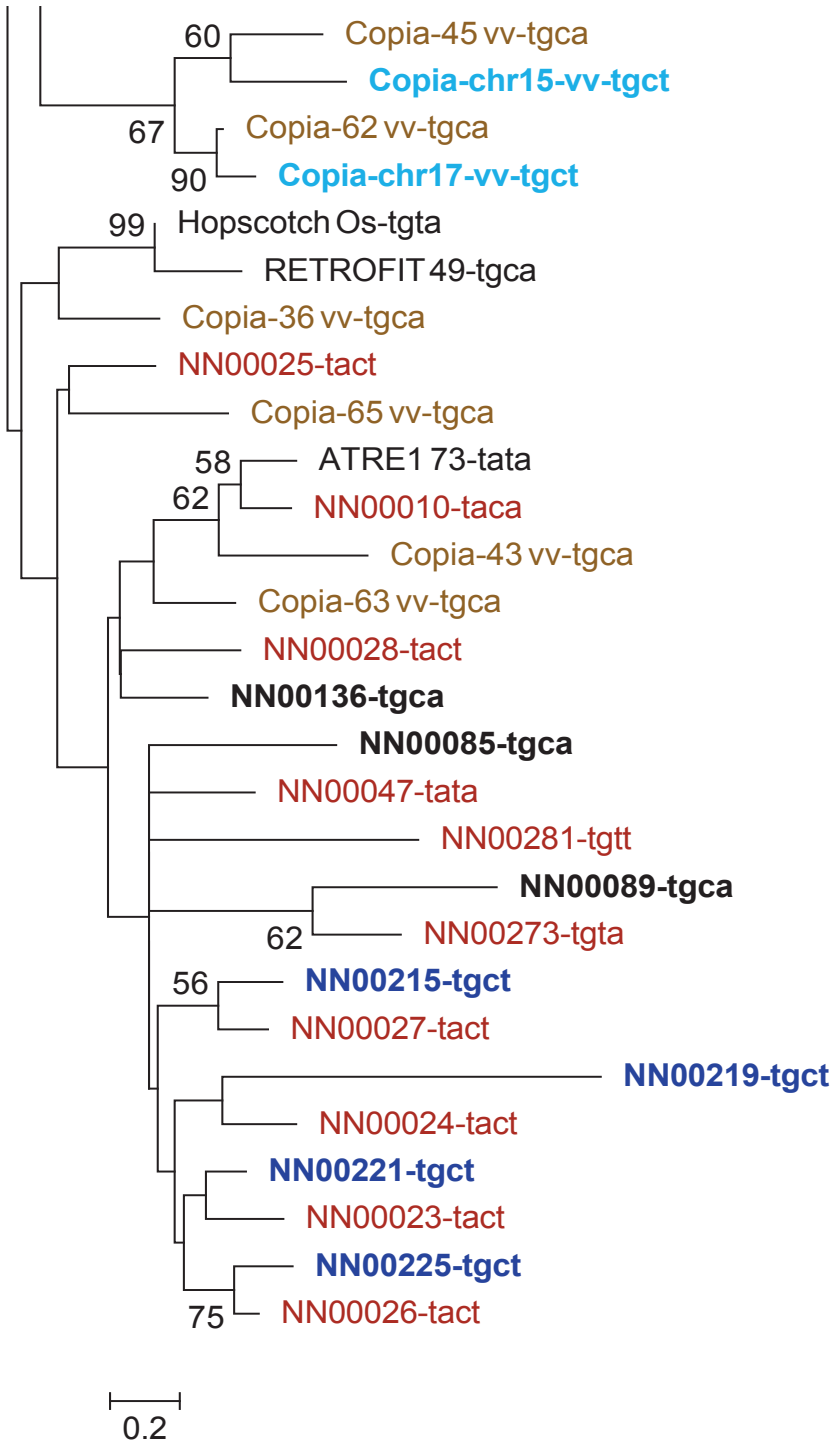


Figure 4.3 (cont'd)



sacred lotus. In contrast, some *Copia* families in grape with canonical ends do group together with their lotus counterparts (Figure 4.3), suggesting the TGCT elements in the two species may not have a common origin. To our knowledge, two copies of a curated LTR element with a similar non-canonical end (5'-TG.CT-3') is present in the soybean genome (Du et al., 2010). However, these elements have been classified as *Gypsy*. Phylogenetic analysis of the conserved reverse transcriptase domain supports that the soybean elements are *Gypsy* type (Figure A1). These results suggest that formation of non-canonical LTR ends can occur in both *Gypsy* and *Copia* LTRs but it appears it occurs more often to *Copia* elements.

Another non-canonical LTR type present in considerable copy number (2210 estimated copies, Table 4.3) are those that ends in GA-3' (referred to as TGGA LTR). Further analysis of copy numbers among the families with TGGA LTR suggests that a single TRIM family, NN00293, which has an estimated copy number of 1264 elements, contributes the majority of the copies. Terminal-repeat retrotransposons in miniature (TRIM) elements are structurally classified as LTR retrotransposons despite their lack of coding sequences. Using the LTR and conserved internal region sequence of this element to mine for intact elements in the sacred lotus repeat database, 304 full copies (see Methods) were mined. NN00293 TRIM elements in sacred lotus are characterized by 73 bp LTR and 142 bp internal sequences, representing an LTR element with the smallest LTR characterized so far.

***hAT* elements**

DNA transposable elements belonging to the *hAT* superfamily are widespread in plant and animal genomes have been widely used in gene tagging and functional genomics studies (Kunze and Weil, 2002). Although widespread in plants, its contribution to genomic repeat is typically low (≤1% of the genome, Table 4.2B). As stated above, the sacred lotus genome is

exceptional in its *hAT* content among plant genomes (>7%). The copy number and coverage of this superfamily indicates its successful amplification that can be a result of high transposition or retention rates, or both. Overall, 325 *hAT* families were identified. However, defined ends are only manually verified and defined for 75 families, therefore the estimated copy number for this superfamily is greatly underestimated.

To evaluate whether specific families have expanded in the genome, the coverage of individual families was determined. Results indicate that overall, the high abundance of *hAT* elements in sacred lotus was not due to the massive amplification of a single family. However, the five most abundant *hAT* families make up to 1.52% of the sacred lotus genome, which is an exceptional success for specific DNA element family in plants. All the five most abundant families were non-autonomous elements that range in size from 293 to 658 bp. Since these elements do not encode the transposase protein, these must have relied on a cognate autonomous partner(s) in the genome for movement.

To test whether levels of diversity in *hAT* transposase may reflect its successful amplification in the sacred lotus genome, phylogenetic analysis of the most conserved domain (motif 3 which contains the E region of the catalytic DD/E motif) of the *hAT* transposase (Kempken and Windhofer, 2001; Lazarow et al., 2012) among autonomous copies was performed. Our results show substantial diversity among the *hAT* transposases found in the genome. Among the 7 clades present in plants (*Ac/Tam3 and Tag1*), animals (*hobo, hopper, Charlie and Tip100*) and fungi (*restless*) (Kempken and Windhofer, 2001; Robertson, 2002), the sacred lotus genome contained autonomous *hATs* from the two clades in plants: *Ac/Tam3* and *Tag1* (Figure 4.4). Majority of the autonomous elements with a recognizable motif 3 that groups within the *Ac/Tam3* clade shows a wide spectrum of diversity wherein various subgroups found

are more closely related to *hAT* proteins from other plant species than *hAT* transposase within the genome. The most expanded subgroup is closely related to the *Tam3* transposase in *Antirrhinum majus* (Hehl et al., 1991). Overall, these results suggest that diversity within autonomous elements may have contributed to the success of the *hAT* superfamily in sacred lotus.

Manual examination for the three conserved motifs that contain the transposases catalytic domain shows that out of 162 putative autonomous element copies, belonging to 15 autonomous *hAT* families, only 2 of these do not contain premature stop codons and are likely to code functional transposase. This data indicates that the majority of autonomous *hATs* and their corresponding non-autonomous partners can currently be considered "dead". Alternatively, active autonomous copies that carry functional *hAT* transposase may be present in multiple and highly similar copies which may make them difficult to be assembled using the next generation sequencing technique.

Pack-MULEs

A total of 1447 Pack-MULEs were identified in *N. nucifera*. Out of 53 MULE TIR families found in the genome, 43 families were involved in gene sequence acquisition. Three of these TIRs constituted over 20% of the Pack-MULEs found. Analysis using the *N. nucifera* EST library generated by this study indicated that at least 10 Pack-MULEs are expressed (Table 4.4). To determine the source of the acquired genic sequences in Pack-MULEs, the internal sequences were used to search against the genomic sequence. This search identified 996 parental genes for the acquired fragments. Preferential acquisition of GC-rich sequences has previously been reported in the grasses, rice and maize (Jiang et al., 2011; Ferguson et al., 2013). Analysis in *Arabidopsis* was inconclusive due to the low copy number of Pack-MULEs (Jiang et al., 2011). To evaluate if the GC preference phenomenon previously observed in monocots also extends to

dicots, the GC gradient of Pack-MULEs in sacred lotus was analyzed. Our results indicate that internal regions of Pack-MULEs are overall more GC-rich than the genome average (Figure 4.5). Moreover, the acquired fragments are significantly more GC-rich than the genome average ($P < 2.2 \times 10^{-16}$; WRS). This trend is not consistent with the *Arabidopsis* Pack-MULEs wherein the internal sequences, TIRs and genomic average have a similar GC content (Jiang et al., 2011). Comparison of the GC content of Pack-MULE parental genes and all non-TE genes in sacred lotus show that Pack-MULE parental genes have significantly higher GC content than other non-TE genes (Figure 4.6). This results supports data from rice Pack-MULEs where parental genes are associated with significant GC-content (Jiang et al., 2011) suggesting that the preferential acquisition based on GC content also occurs in some eudicots.

In grasses, a higher proportion of GC-rich genes are found in comparison to Arabidopsis (Jiang et al., 2011). In addition, the Pack-MULE acquisition preference for GC-rich sequences combined with their insertion specificity at 5' end of genes is considered at least partly responsible for the presence of genes with negative GC-gradients (Jiang et al., 2011), where the 5' end of genes are more GC-rich than their 3' end. Again, this correlation is not obvious with Arabidopsis, possibly due to its low activity of Pack-MULEs. To test whether the abundance of Pack-MULEs in sacred lotus is accompanied by increased negative GC-gradient of genes, we compared the number of positive, moderate and negative genes between the Arabidopsis and sacred lotus. The proportions of positive genes, those where the 3' half is more GC-rich than the 5' half, in the two genomes are comparable (Arabidopsis: 7.3% and lotus: 5.9%). However, the proportion of negative genes in sacred lotus, those where the 5' half is more GC-rich than the 3' half, is more than double that found in Arabidopsis (25.4% vs. 11.7%; Figure 4.7). *Lotus*

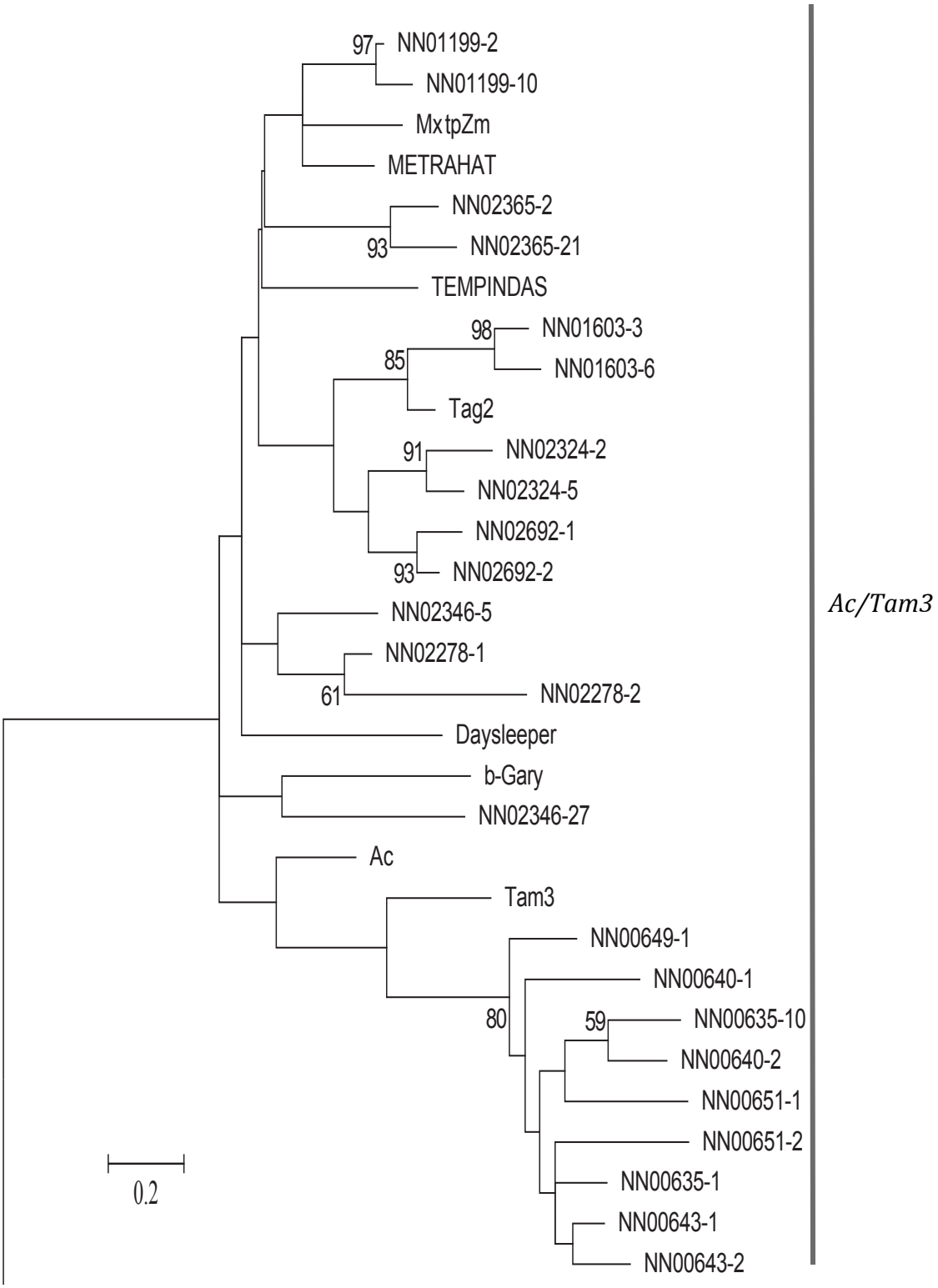


Figure 4.4 Phylogenetic analysis of the conserved domain 3 of *hAT* transposase. Bootstrap values are indicated as a percentage of 500 replicates. Names in red represent non-plant *hAT* transposase.

Figure 4.4 (cont'd)

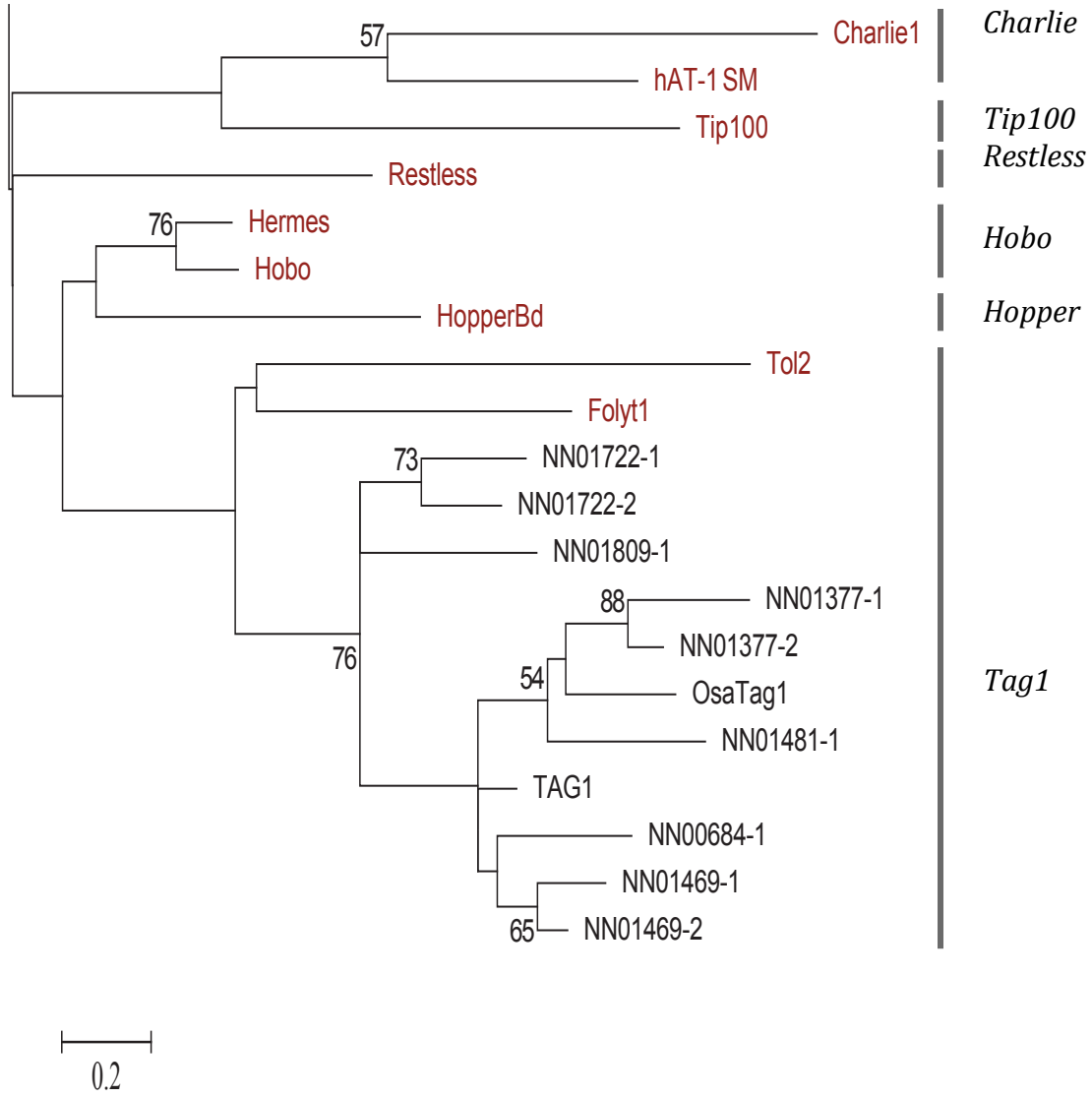


Table 4.4. Pack-MULEs with EST evidence of expression.

Megascaffold number	Coordinate		EST contig	Putative annotation
	Start	End		
1	48058897	48060093	contig04254	rhamnosyltransferase 1
2	95879186	95881535	contig08224	Kinase
3	3060656	3063250	contig00026	maturase K
3	7141300	7143148	contig04546	probable exonuclease V-like
6	30463674	30464922	contig13509	U-box domain-containing protein
7	1373862	1375844	contig11719	CCT/B-box zinc finger protein
7	1382403	1395660	contig06447	hypothetical protein
8	4585905	4587696	contig14084	hypothetical protein
10	8693634	8701482	contig10382	conserved hypothetical protein
79	8508	10373	contig07837	seed imbibition protein (*Sip1*)

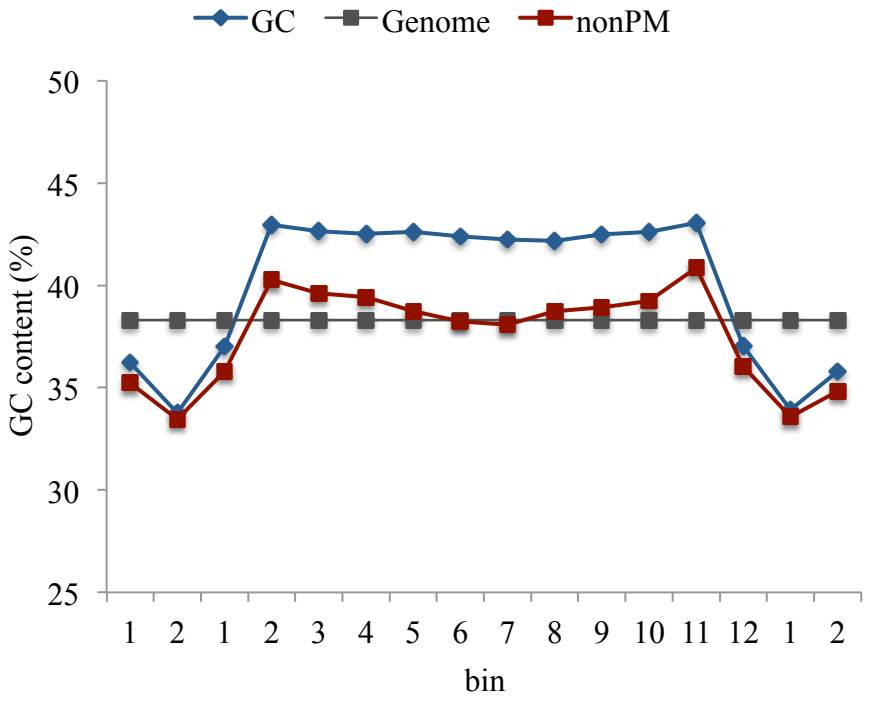


Figure 4.5. GC content along Pack-MULEs and non-Pack-MULEs. The first 2 and last 2 bins represent TIR regions and the internal sequence was divided into 12 equal-sized bins prior to determination of GC content per bin.

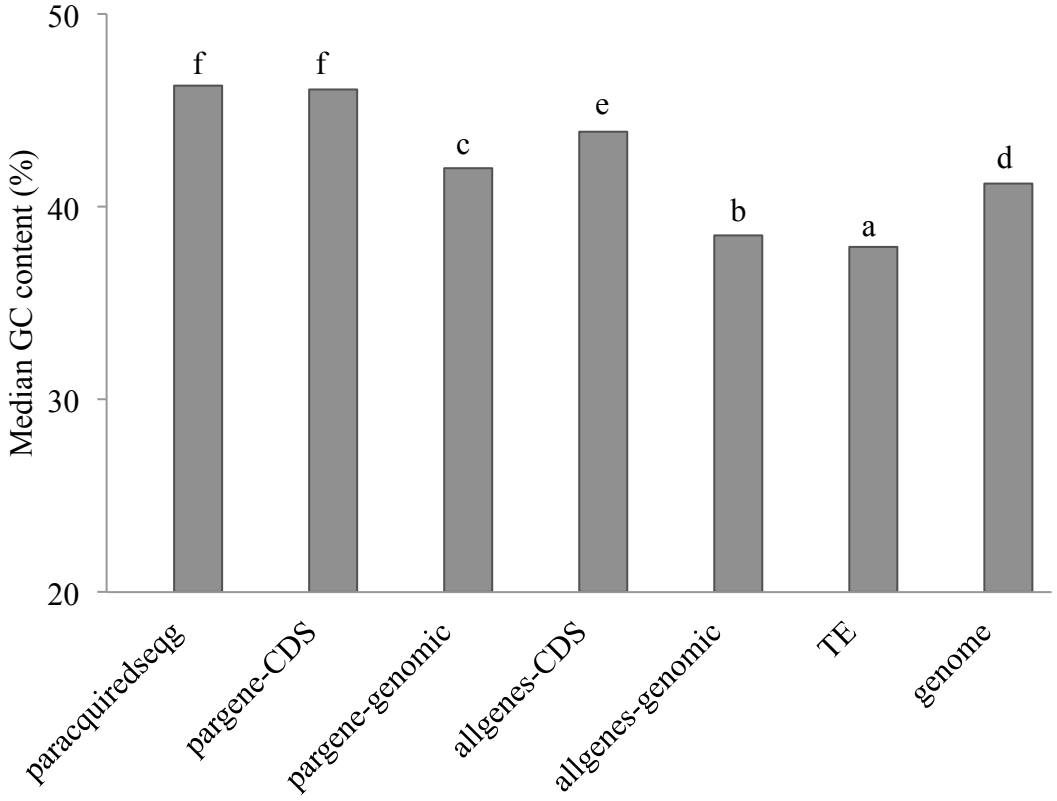


Figure 4.6. GC content of different genomic sequences in the sacred lotus genome. Bars designated with different letters indicate their values are significantly different (α= 0.0025) by Wilcoxon Rank SumTest (WRS) with Bonferroni correction.

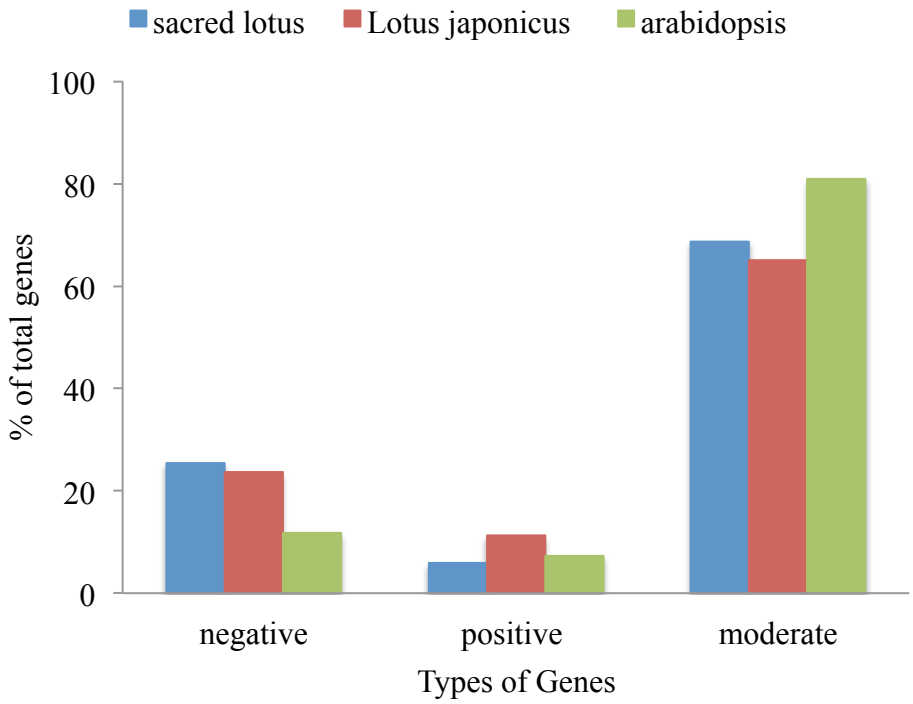


Figure 4.7. Distribution of different types of genes based on GC gradient in sacred lotus, *Lotus japonicus* and Arabidopsis.

japonicus, another dicot containing >1000 Pack-MULE (Holligan et al., 2006), has negative genes equally as frequent (23.6%) as sacred lotus (Figure 4.7). Taken together, our results suggest that Pack-MULEs may contribute to the negative GC-gradient of genes in dicots with high Pack-MULE activity.

Discussion

Angiosperm is the most dominant plant taxon containing as many as 400,000 species (Jarvis et al., 2007) and ranks second to insects in species richness. Two major groups within angiosperms are monocots and dicots whose split was dated 150-130 million years ago (MYA) (Wikstrom et al., 2001). At present, among the dicots, the eudicot clade represents ~75% of the species diversity in angiosperms (Drinnan et al., 1994) and lotus occupies a key position in studies of angiosperm evolution particularly in the monocot-dicot split. The divergence from its closest sister lineage was dated at 137-125 MYA (Wikstrom et al., 2001). Therefore, it is the most basal dicot sequence released, replacing the grape genome (Velasco et al., 2007; Ming et al., 2013), which diverged from its sister lineage about 118-108 MYA (Wikstrom et al., 2001) making it the only sequenced dicot genome that is closest to monocots. In addition, it lies outside the core eudicot clade and lacks a duplication event that is shared by all other sequenced eudicots (Ming et al., 2013).

Transposable elements are an integral part of eukaryotic genomes and studies have demonstrated its significant contribution in the evolution of angiosperms (reviewed in Oliver et al., 2013). Notably, TEs have been implicated in the genome size expansion of plants (Sanmiguel, 1998; Hawkins et al., 2006; Piegu et al., 2006) and have contributed to the domestication of new genes (Piriyapongsa et al., 2006; Mikkelsen et al., 2007) and evolution of regulatory elements in many genomes (Jordan et al., 2003; Muotri et al., 2007). The TE-thrust

hypothesis proposes that both active and inactive TE can introduce genetic changes that may provides adaptive and evolutionary potential (Oliver et al., 2013). These changes include gene modifications, altered gene expression, gene duplication and creation of novel genes (as in the case of Pack-MULEs). Thus, their identification and characterization remains crucial in genome studies.

The characterization of transposable elements in a basal eudicot genome such as lotus may be used as a tool in understanding the role of TE in the monocot-dicot split as well as in the diversification of eudicots. In this study, computer-assisted and manual approaches were used to mine and characterize the repeat content and diversity of sacred lotus (*Nelumbo nucifera*). Over half of the genome sequence is repetitive with the bulk of this composed by recognizable transposable elements. This is an underestimate because the unassembled region or sequencing gaps are usually more enriched in repetitive sequences. The total amount of TEs in sacred lotus is comparable to other plants with similar genome size, such as tomato, potato, sorghum, soybean, and apple (Table 4.2B). Like other plant genomes, majority of the sacred lotus genome repeat sequence is contributed by LTR retrotransposons. However, the content and diversity of some of the TE families contained in the genome show features unique to sacred lotus.

Multiple factors are involved in the success or failure of a TE family in a host genome. Novel TEs may be introduced into a host through many pathways including but not limited to hybridization and polyploidy (Kawakami et al., 2010; Parisod et al., 2010), and horizontal transfer (Bartolomé et al., 2009; Schaack et al., 2010). Unless quickly lost, novel TEs may become prolific until they are recognized by the host and subsequently silenced. TEs can be lost or become undetectable when inactive TEs accumulate mutations over time until they are no longer recognizable as such, or are eliminated from the genome through drastic processes such as

recombination and random deletions (Tenaillon et al., 2010).

The lotus TE content contains a notable gain and loss in DNA elements. The genome contains all major types of DNA TEs in plants except *Tc1/Mariner* family, a widespread DNA element family found in many plant genomes, making it the fourth reported genome devoid of these elements (D'Hont et al., 2012; Jaillion et al., 2007). The loss of this TE superfamily in sacred lotus suggests that specific silencing and/or insertion targeting mechanisms can be critical in the failure or success of TE families in a host. In this particular case, it is unknown which factor played a more important role in its demise. In contrast, the *hAT* superfamily seems to have massively proliferated, accounting for as much as 7.3% of the genome. This represents more than the total Class II/DNA element content of many dicot plant genomes with the exception of peach and soybean (Schmutz et al., 2010; Verde et al., 2013). In addition, this is a level of coverage typically seen only from retrotransposons among eukaryotes. This indicates an overall unparalleled activity of many *hAT* families in the evolution of sacred lotus, more than any other genome whose repeat content has been characterized. Since amplification and success of a specific TE family in a genome involves a struggle against many other TE types in the genome, this implies that *hAT* elements in sacred lotus may have possessed certain advantages. This can include a combination of higher transposition efficiency and more successful targeting preference that allows it to escape intense genomic regulation. The diversity in *hAT* transposase proteins may play a role in this regard. However, in this study, analysis of conserved domains from the *hAT* transposase encoded by autonomous elements suggests that many autonomous *hAT* elements found by this study are presently inactive, suggesting they are subject to significant silencing and might be proceeding to the end of their life cycle.

The *hAT* superfamily includes the *Ac*/*Ds* elements which were the very first transposons discovered (McClintock, 1951) and has since been paramount in the study of TE domestication. In a study of 65 instances of traits in angiosperms generated by TE, *hAT* elements account for 20% of these events (Oliver et al., 2013). These include the *SLEEPER* genes from a domesticated *hAT* transposase that functions as transcriptional regulators of plant development unique to angiosperms (Knip et al., 2012). Because of their prevalence in the angiosperm lineage and its role in plant development, these data suggest the potential impact of domesticated *hAT* genes in angiosperm evolution. The unparalleled activity of *hATs* in lotus compared with other organisms may prompt studies to evaluate its effect in gene regulation as well as TE domestication.

LTR retrotransposons are the most abundant TEs in plants and are considered to be responsible for the expansion of plant genomes. Most LTR elements start with 5'-TG and end in CA-3'. Such terminal sequence was believed to be important for element integration (Temin, 1981). Prior to this study, four incidents of LTR elements with non-canonical ends were reported or annotated. These include three *Copia*-like elements: the *Tos17* in rice, AtRE1 in Arabidopsis and *TARE1* in tomato with 5'-TG.GA-3', 5'-TA.TA-3' and 5'-TA.CA-3' LTR ends, respectively (Hirochika et al., 1996; Kuwahara et al., 2000; Yin et al., 2013). In addition, three copies of *Gypsy*-like elements with 5'-TG.CT-3' are annotated in soybean (Du et al., 2010). However, all these elements are low copy number elements in natural populations, despite the fact that *Tos17* can achieve high activity artificially through tissue culture (Hirochika et al., 1996). As a result, it is unclear whether they represent transient mutations or they can successfully amplify and be retained in natural populations. In sacred lotus, we found 8 different non-canonical ends that comprise at least 10000 copies and 3.8% of the genome (16% of total LTR content and 30% of

total *Copia*-like elements), 4 of these (TGTA, TACT, TGGT, TGTT) are reported for the first time. Our analysis indicates that among the two terminal nucleotides on each side of the LTR, only the first nucleotide at the 5' end is not replaceable. All other nucleotides can be substituted without complete abolishment of the transposition activity. The second nucleotide at the 3' end is the most flexible and can be C, G, or T. The differential constraint at the two ends suggests that they play distinct roles in integration.

Thus far, the high copy TGCT LTR covers over 2% of the sacred lotus genome and is the first non-canonical LTR type reported with this level of coverage and copy number in plants. This end type is found in the soybean genome but is present in only 3 copies (one of which is a solo LTR), suggesting that it may not be highly competent for transposition. Similarly, the two families of *Gypsy*-like elements with non-canonical ends detected in the genome of sacred lotus demonstrated limited amplification. Therefore, it appears that non-canonical ends are more successful among *Copia*-like elements. The origin of the TGCT elements in sacred lotus could be explained in two ways. It is possible that there is an ancient lineage of *Copia*-like elements the integrase of which has higher affinity to TGCT than TGCA end. If that is the case, one would expect the TGCT elements group with the same elements in other species but that is not observed. Alternatively, the TGCT elements are derived from relatively recent mutations in sacred lotus and somehow have achieved significant success. Our phylogenetic analysis indicates this is more likely the case.

Unlike the *hAT* elements (Figure 4.4), where most elements group with counterparts in other species, the majority of TGCT LTR families are closely related, and may represent a specific clade of LTR elements with frequent formation of non-canonical LTR ends. For the tomato *TARE1* elements, it was postulated that the change in the LTR sequence was due to a

mutation in the 3' LTR end from 'G' to 'A' prior to the transposition of the element (Yin et al., 2013). It is known that the reverse transcription reaction is error-prone so it is not surprising that mutations arise prior to transposition. The critical question is whether the mutation is competent for further transposition. If not, the mutation would rapidly disappear from the genome. Our analysis clearly demonstrated that 3 out of the 4 terminal nucleotides can be altered (a maximum of 2 can be mutated simultaneously) to retain the transposition activity. The presence of 8 different mutant ends suggests the high degree of flexibility of the integrase of *Copia*-like elements to cooperate with different element ends. On the other hand, the presence of few elements ending with TGCA end among the TGCT clade (Figure 4.3) suggests that the mutation might be reversible but the integrase of these elements might have been evolved specificity to this type of end (TGCT). If this is the case, the success of TGCT elements could be explained by the long-term co-evolution between the element and the transposition machinery in the lineage of sacred lotus. Computer-assisted techniques to mine LTR retrotransposons during *de novo* genomic searches usually involve structure-based algorithms. Two widely used programs are LTR-STRUC and LTR_FINDER which uses defining features of LTR elements (McCarthy and McDonald, 2003; Xu and Wang, 2007), one of which is the typical LTR which starts with 5'-TG and ends in CA-3'. Due to this, LTR elements with atypical ends can be inherently missed by automated searches. Our analysis provides guidance for future annotation of LTR elements in plant genomes.

Pack-MULEs are *Mutator*-like elements (MULE) that carry gene fragment(s). These elements are particularly important as they may generate new open reading frames and/or regulate the expression of the parental genes from which the captured sequences are derived (Jiang et al., 2004; Hanada et al., 2009). Recent work shows that gene GC content and

expression play a combined role in the acquisition preference of Pack-MULEs (Ferguson et al., 2013). Our annotation also allowed additional analyses in regards to sequence acquisition preference by Pack-MULEs. To date, the sacred lotus is part of a very small subset of plants (along with rice and *Lotus japonicus*) that contain thousands of Pack-MULEs (Jiang et al., 2004; Holligan et al., 2006). Analysis of GC content of Pack-MULEs and their parental genes shows that a high GC content preference by Pack-MULEs also extends to dicots, a notion that was previously inconclusive using Arabidopsis data (Jiang et al., 2011). Recent work on the nucleotide composition of sacred lotus shows that its GC content is intermediate to that of eudicots and grasses and exhibits negative GC gradient in the GC3 content (Singh et al., 2013). In combination with the presence of a higher proportion of negative genes than that in the Arabidopsis genome, it may indicate that significant activity of Pack-MULEs may cause detectable variation of GC gradient of genes. Our results suggest that sacred lotus may serve as an additional model for Pack-MULE studies. In addition, it may serve in comparative studies in the role and impact of Pack-MULE formation in gene duplication and evolution for monocots and dicots.

APPENDIX

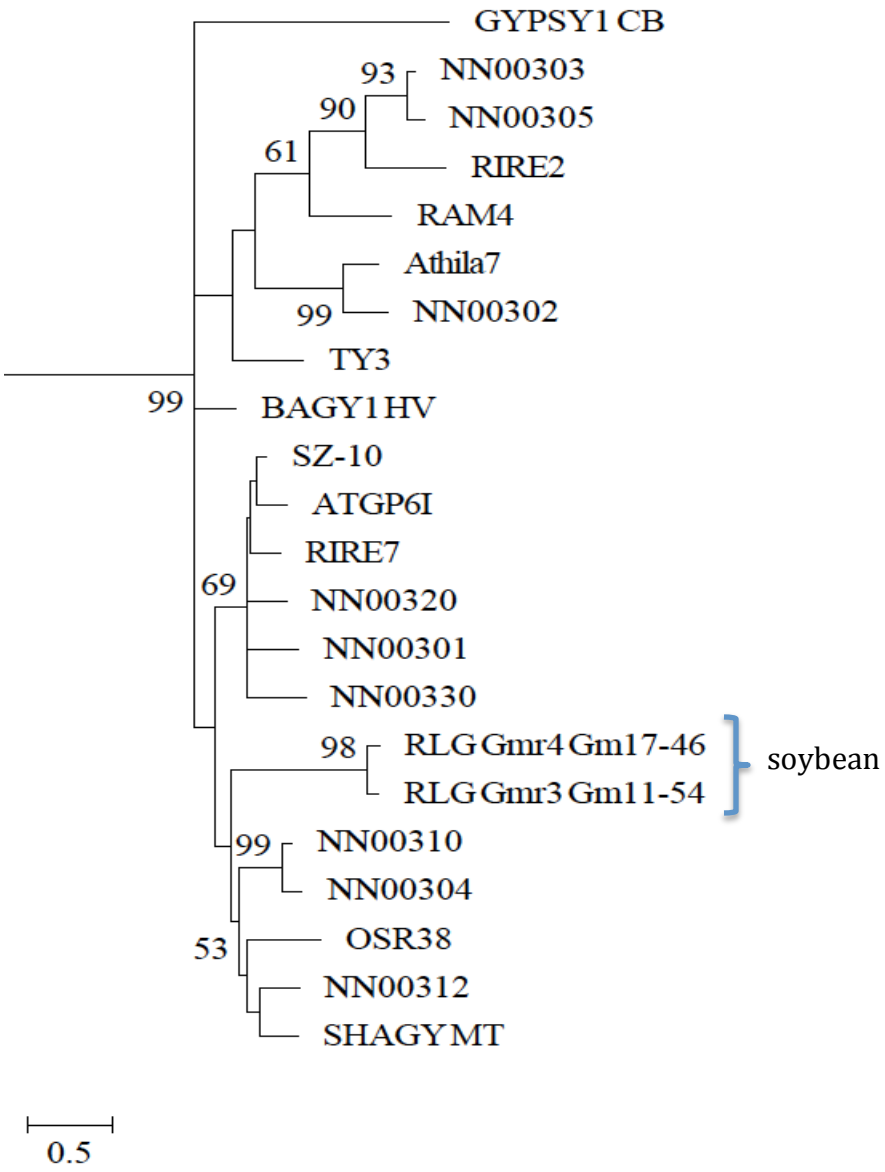


Figure A1. Phylogenetic analysis of conserved domain in the reverse transcriptase gene from Gypsy LTR. Bootstrap values are indicated as a percentage of 500 replicates.

REFERENCES

REFERENCES

Adams MD (2000) The Genome Sequence of Drosophila melanogaster. Science 287: 2185–2195

Alleman, M., Freeling, M. (1986) The Mu transposable elements of maize: evidence for transposition and copy number regulation during development. Genetics 112: 107–119

Ammiraju JSS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, Walling JG, Ma J, Talag J, Brar DS, et al (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. The Plant Journal 52: 342–351

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al (2010) The genome of Theobroma cacao. Nature Genetics 43: 101–108

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, et al (2011) The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants. Science 332: 960–963

Barkan A, Martienssen RA (1991) Inactivation of maize transposon Mu suppresses a mutant phenotype by activating an outward-reading promoter near the end of Mu1. Proceedings of the National Academy of Sciences 88: 3502–3506

Bartolomé C, Bello X, Maside X (2009) Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. Genome Biology 10: R22

Becker H-A (1997) Maize Activator transposase has a bipartite DNA binding domain that recognizes subterminal sequences and the terminal inverted repeats. Molecular and General Genetics MGG 254: 219–230

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441: 87–90

Benito MI, Walbot V (1997) Characterization of the maize Mutator transposable element MURA transposase as a DNA-binding protein. Mol Cell Biol 17: 5165–75

Bennetzen J (2007) Patterns in grass genome evolution. Current Opinion in Plant Biology 10: 176–181

Bennetzen JL (2005) Mechanisms of Recent Genome Size Variation in Flowering Plants. Annals of Botany 95: 127–132

Bennetzen JL, Kellogg EA (1997) Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9: 1509–1514

Bennetzen JL, Springer PS (1994) The generation of Mutator transposable element subfamilies in maize. Theoretical and Applied Genetics. doi: 10.1007/BF00222890

Bennetzen JL, Swanson J, Taylor WC, Freeling M (1984) DNA insertion in the first intron of maize Adh1 affects message levels: cloning of progenitor and mutant Adh1 alleles. Proceedings of the National Academy of Sciences 81: 4125–4128

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491: 705–710

Bundock P, Hooykaas P (2005) An Arabidopsis hAT-like transposase is essential for plant development. Nature 436: 282–284

Bushman FD (2003) Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell 115: 135–138

Callinan PA, Batzer MA (2006) Retrotransposable Elements and Human Disease. *In* J-N Volff, ed, Genome Dynamics. KARGER, Basel, pp 104–115

Cao X, Jacobsen SE (2002) Role of the Arabidopsis DRM Methyltransferases in De Novo DNA Methylation and Gene Silencing. Current Biology 12: 1138–1144

Chalker DL, Sandmeyer SB (1992) Ty3 integrates within the region of RNA polymerase III transcription initiation. Genes & Development 6: 117–128

Chalvet F (2003) Hop, an Active Mutator-like Element in the Genome of the Fungus Fusarium oxysporum. Molecular Biology and Evolution 20: 1362–1375

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, et al (2010) Draft genome sequence of the oilseed species Ricinus communis. Nature Biotechnology 28: 951–956

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562

Chomet P, Lisch D, Hardeman KJ, Chandler VL, Freeling M (1991) Identification of a regulatory transposon that controls the Mutator transposable element system in maize. Genetics 129: 261–270

Cordaux R (2006) From the Cover: Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proceedings of the National Academy of Sciences 103: 8101–8106

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488: 213–217

Dassanayake M, Oh D-H, Haas JS, Hernandez A, Hong H, Ali S, Yun D-J, Bressan RA, Zhu J-K, Bohnert HJ, et al (2011) The genome of the extremophile crucifer Thellungiella parvula. Nature Genetics 43: 913–918

Diao Y, Chen L, Yang G, Zhou M, Song Y, Hu Z, Liu JY (2006) Nuclear DNA C-values in 12 species in Nymphales. Caryologia 59: 25–30

Dietrich CR, Cui F, Packila ML, Li J, Ashlock DA, Nikolau BJ, Schnable PS (2002) Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics 160: 697–716

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, et al (2011) De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nature Biotechnology 29: 521–527

Drinnan AN, Crane PR, Hoot SB (1994) Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). *In* PK Endress, EM Friis, eds, Early Evolution of Flowers. Springer Vienna, Vienna, pp 93–122

Du J, Grant D, Tian Z, Nelson RT, Zhu L, Shoemaker RC, Ma J (2010) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11: 113

Duke JA, Duke (2002) Handbook of medicinal herbs. CRC Press, Boca Raton, FL

Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21: i152–i158

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9: 18

Ferguson AA, Jiang N (2012) Mutator-Like Elements with Multiple Long Terminal Inverted Repeats in Plants. Comparative and Functional Genomics 2012: 1–14

Fernandes J, Dong Q, Schneider B, Morrow DJ, Nan G-L, Brendel V, Walbot V (2004) Genome-wide mutagenesis of Zea mays L. using RescueMu transposons. Genome Biol 5: R82

Feschotte C (2004) Merlin, a New Superfamily of DNA Transposons Identified in Diverse Animal Genomes and Related to Bacterial IS1016 Insertion Sequences. Molecular

Biology and Evolution 21: 1769–1780

Feschotte C (2005) DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. Nucleic Acids Research 33: 2153–2165

Feschotte C, Jiang N, Wessler SR (2002) PLANT TRANSPOSABLE ELEMENTS: WHERE GENETICS MEETS GENOMICS. Nature Reviews Genetics 3: 329–341

Flavell AJ, Pearce SR, Kumar A (1994) Plant transposable elements and the genome. Current Opinion in Genetics & Development 4: 838–844

Guo HB (2008) Cultivation of lotus (Nelumbo nucifera Gaertn. ssp. nucifera) and its utilization in China. Genetic Resources and Crop Evolution 56: 323–330

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Research 38: e199–e199

Han Y-C, Teng C-Z, Zhong S, Zhou M-Q, Hu Z-L, Song Y-C (2007) Genetic variation and clonal diversity in populations of Nelumbo nucifera (Nelumbonaceae) in central China detected by ISSR markers. Aquatic Botany 86: 69–75

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu S-H, Jiang N (2009) The Functional Role of Pack-MULEs in Rice Inferred from Purifying Selection and Expression Profile. THE PLANT CELL ONLINE 21: 25–38

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Research 16: 1252–1261

Hehl R, Nacken WKF, Krause A, Saedler H, Sommer H (1991) Structural analysis of Tam3, a transposable element from Antirrhinum majus, reveals homologies to the Ac element from maize. Plant Molecular Biology 16: 369–371

Hershberger RJ, Benito M-I, Hardeman KJ, Warren C, Chandler VL, Walbot V (1995) Characterization of the major transcripts encoded by the regulatory MuDR transposable element of maize. Genetics 140: 1087–1098

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

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proceedings of the National Academy of Sciences 93: 7783–7788

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR (2006) The Transposable Element Landscape of the Model Legume Lotus japonicus. Genetics 174: 2215–2228

Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS (2011) Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proceedings of the National Academy of Sciences 108: 2322–2327

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496: 498–503

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al (2009) The genome of the cucumber, Cucumis sativus L. Nature Genetics 41: 1275–1281

Ichikawa H, Ikeda K, Wishart WL, Ohtsubo E (1987) Specific binding of transposase to terminal inverted repeats of transposable element Tn3. Proceedings of the National Academy of Sciences 84: 8220–8224

Jarvis CE, Linnean Society of London, Natural History Museum (London, England) (2007) Order out of chaos: Linnaean plant names and their types. Linnean Society of London in association with the Natural History Museum, London, London

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573

Jiang N, Ferguson AA, Slotkin RK, Lisch D (2011) Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proceedings of the National Academy of Sciences 108: 1537–1542

Jiang N, Panaud O (2013) Transposable Element Dynamics in Rice and Its Wild Relatives. *In* Q Zhang, RA Wing, eds, Genetics and Genomics of Rice. Springer New York, New York, NY, pp 55–69

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics 19: 68–72

Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res 15: 1292–1297

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH (2000) From the Cover: Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences 97: 6603–6607

Kamal M (2006) A large family of ancient repeat elements in the human genome is under strong selection. Proceedings of the National Academy of Sciences 103: 2740–2745

Kapitonov VV (2006) Self-synthesizing DNA transposons in eukaryotes. Proceedings of the

National Academy of Sciences 103: 4540–4545

Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3: e181

Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Reviews Genetics 9: 411–412

Kapitonov VV, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends in Genetics 23: 521–529

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genetics 9: e1003470

Kawakami T, Strakosh SC, Zhen Y, Ungerer MC (2010) Different scales of Ty1/copia-like retrotransposon proliferation in the genomes of three diploid hybrid sunflower species. Heredity (Edinb) 104: 341–350

Kempken F, Windhofer F (2001) The hAT family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma 110: 1–9

Kinoshita T (2004) One-Way Control of FWA Imprinting in Arabidopsis Endosperm by DNA Methylation. Science 303: 521–523

Knip M, de Pater S, Hooykaas PJ (2012) The SLEEPER genes: a transposase-derived angiosperm-specific gene family. BMC Plant Biology 12: 192

Kumar A, Bennetzen JL (1999) Plant Retrotransposons. Annual Review of Genetics 33: 479–532

Kunze R, Weil CF (2002) The hAT and CACTA superfamilies of plant transposons. Mobile DNA II. pp 565–610

Kuwahara A, Kato A, Komeda Y (2000) Isolation and characterization of copia-type retrotransposons in Arabidopsis thaliana. Gene 244: 127–136

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921

Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics 11: 204–220

Lazarow K, Du M-L, Weimer R, Kunze R (2012) A Hyperactive Transposase of the Maize Transposable Element Activator (Ac). Genetics 191: 747–756

Li W, Shaw JE (1993) A variant Tc4 transposable element in the nematode *C.elegans* could encode a novel protein. Nucleic Acids Research 21: 59–67

Lindroth AM (2001) Requirement of CHROMOMETHYLASE3 for Maintenance of CpXpG Methylation. Science 292: 2077–2080

Linheiro RS, Bergman CM (2012) Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster. PLoS ONE 7: e30008

Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, Richard McCombie W, Lavine K, Mittal V, May B, Kasschau KD, et al (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430: 471–476

Lisch D (2002) Mutator transposons. Trends Plant Sci 7: 498–504

Lisch D (2009) Epigenetic Regulation of Transposable Elements in Plants. Annual Review of Plant Biology 60: 43–66

Lisch D (2005) Pack-MULEs: theft on a massive scale. BioEssays 27: 353–355

Lisch D, Girard L, Donlin M, Freeling M (1999) Functional analysis of deletion derivatives of the maize transposon MuDR delineates roles for MURA and MURB proteins. Genetics 151: 331–341

Lisch D, Jiang N (2009) Mutator and MULE transposons. *In* JL Bennetzen, S Hake, eds, Handbook of Maize. Springer New York, New York, NY, pp 277–306

Liu S, Yeh C-T, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS (2009) Mu Transposon Insertion Sites and Meiotic Recombination Events Co-Localize with Epigenetic Marks for Open Chromatin across the Maize Genome. PLoS Genetics 5: e1000733

Loot C, Santiago N, Sanz A, Casacuberta JM (2006) The proteins encoded by the pogo-like Lemi1 element bind the TIRs and subterminal repeated motifs of the Arabidopsis Emigrant MITE: consequences for the transposition mechanism of MITEs. Nucleic Acids Research 34: 5238–5246

Marquez CP, Pritham EJ (2010) Phantom, a New Subclass of Mutator DNA Transposons Found in Insect Viruses and Widely Distributed in Animals. Genetics 185: 1507–1517

Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, Matsumoto T, Sato K, Schulman A, Muehlbauer GJ, et al (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature. doi: 10.1038/nature11543

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362–367

McClintock B (1951) CHROMOSOME ORGANIZATION AND GENIC EXPRESSION. Cold Spring Harbor Symposia on Quantitative Biology 16: 13–47

McLaughlin M, Walbot V (1987) Cloning of a mutable bz2 allele of maize by transposon tagging and differential hybridization. Genetics 117: 771–776

Middleton CP, Stein N, Keller B, Kilian B, Wicker T (2013) Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity. The Plant Journal 73: 347–356

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, et al (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167–177

Mills RE, Bennett EA, Iskow RC, Devine SE (2007) Which transposable elements are active in the human genome? Trends in Genetics 23: 183–191

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, et al (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452: 991–996

Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li L-T, Zhang Q, Kim M-J, Schatz MC, Campbell M, et al (2013) Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). Genome Biology 14: R41

Morgenstern B (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15: 211–218

Muotri AR, Marchetto MCN, Coufal NG, Gage FH (2007) The necessary junk: new functions for transposable elements. Human Molecular Genetics 16: R159–R167

Nassif N, Penney J, Pal S, Engels WR, Gloor GB (1994) Efficient copying of nonhomologous sequences from ectopicsites via P-element-induced gap repair. Mol Cell Biol 14: 1613–1625

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu Z, Loftus B, Xi Z, Megy K, Grabherr M, et al (2007) Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. Science 316: 1718–1723

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497: 579–584

O'Reilly C, Shepherd NS, Pereira A, Schwarz-Sommer Z, Bertram I, Robertson DS, Peterson PA, Saedler H (1985) Molecular cloning of the a1 locus of Zea mays using the

transposable elements En and Mu1. EMBO J 4: 877–882

Oliver KR, McComb JA, Greene WK (2013) Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity. Genome Biology and Evolution 5: 1886–1901

Pan L, Xia Q, Quan Z, Liu H, Ke W, Ding Y (2009) Development of Novel EST-SSRs from Sacred Lotus (Nelumbo nucifera Gaertn) and Their Utilization for the Genetic Diversity Analysis of N. nucifera. Journal of Heredity 101: 71–82

Pardue M-L, Rashkova S, Casacuberta E, DeBaryshe PG, George JA, Traverse KL (2005) Two retrotransposons maintain telomeres in Drosophila. Chromosome Research 13: 443–453

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien M-A (2010) Impact of transposable elements on the organization and function of allopolyploid genomes: Research review. New Phytologist 186: 37–45

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457: 551–556

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, et al (2006) Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Research 16: 1262–1269

Piriyapongsa J, Marino-Ramirez L, Jordan IK (2006) Origin and Evolution of Human microRNAs From Transposable Elements. Genetics 176: 1323–1337

Robertson DS (1978) Characterization of a mutator system in maize. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 51: 21–28

Robertson HM (2002) Evolution of DNA transposons. Mobile DNA II. American Society for Microbiology Press, Washington, D.C, pp 1093–1110

Sanmiguel P (1998) Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons. Annals of Botany 82: 37–44

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, et al (2010) Sequence Analysis of the Genome of an Oil-Bearing Tree, Jatropha curcas L. DNA Research 18: 65–76

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable

elements and why it matters for eukaryotic evolution. Trends in Ecology & Evolution 25: 537–546

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326: 1112–1115

Schulman AH, Kalendar R (2005) A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. Cytogenetic and Genome Research 110: 598–605

Sequencing Project IRG (2005) The map-based sequence of the rice genome. Nature 436: 793–800

Shaheen M, Williamson E, Nickoloff J, Lee S-H, Hromas R (2010) Metnase/SETMAR: a domesticated primate transposase that enhances DNA repair, replication, and decatenation. Genetica 138: 559–566

Shen-Miller J (2002) Sacred lotus, the long-living fruits of China Antique. Seed Science Research 12: 131–143

Shen-Miller J, Aung LH, Turek J, Schopf JW, Tholandi M, Yang M, Czaja A (2013) Centuries-Old Viable Fruit of Sacred Lotus Nelumbo nucifera Gaertn var. China Antique. Tropical Plant Biology 6: 53–68

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, et al (2010) The genome of woodland strawberry (Fragaria vesca). Nature Genetics 43: 109–116

Singer T (2001) Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes & Development 15: 591–602

Singh R, Ming R, Yu Q (2013) Nucleotide Composition of the Nelumbo nucifera Genome. Tropical Plant Biology 6: 85–97

Slotkin RK, Freeling M, Lisch D (2003) Mu killer causes the heritable inactivation of the Mutator family of transposable elements in Zea mays. Genetics 165: 781–797

Slotkin RK, Freeling M, Lisch D (2005) Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. Nature Genetics 37: 641–644

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8: 272–285

Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Research 37: 7002–7013

Strommer JN, Hake S, Bennetzen J, Taylor WC, Freeling M (1982) Regulatory mutants of the maize Adh1 gene caused by DNA insertions. Nature 300: 542–544

Talbert LE, Chandler VL (1989) Characterization of a highly conserved sequence related to mutator transposable elements in maize. Mol Biol Evol 5: 519–529

Temin HM (1981) Structure, variation and synthesis of retrovirus long terminal repeat. Cell 27: 1–3

Tenaillon MI, Hollister JD, Gaut BS (2010) A triptych of the evolution of plant transposable elements. Trends in Plant Science 15: 471–478

Thomas CA (1971) The Genetic Organization of Chromosomes. Annual Review of Genetics 5: 237–256

Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, Toyoda A, Fujiyama A, Tarutani Y, Kakutani T (2012) Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. Genes & Development 26: 705–713

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T (2009) Bursts of retrotransposition reproduced in Arabidopsis. Nature 461: 423–426

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al (2006) The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science 313: 1596–1604

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, Whaley AM, et al (2011) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nature Biotechnology 30: 83–89

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, et al (2010) The genome of the domesticated apple (Malus × domestica Borkh.). Nature Genetics 42: 833–839

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, et al (2007) A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLoS ONE 2: e1326

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, et al (2013) The high-quality draft genome of peach (Prunus

persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics 45: 487–494

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, et al (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463: 763–768

Volff J-N (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28: 913–922

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al (2011) The genome of the mesopolyploid crop species Brassica rapa. Nature Genetics 43: 1035–1039

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al (2007) A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973–982

Wikstrom N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proceedings of the Royal Society B: Biological Sciences 268: 2211–2220

Xu Q, Chen L-L, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao W-B, Hao B-H, Lyon MP, et al (2012) The draft genome of sweet orange (Citrus sinensis). Nature Genetics 45: 59–66

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, et al (2011) Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research 35: W265–W268

Yamashita S, Takano-Shimizu T, Kitamura K, Mikami T, Kishima Y (1999) Resistance to gap repair of the transposon Tam3 in Antirrhinum majus: a role of the end regions. Genetics 153: 1899–1908

Yang G, Weil CF, Wessler SR (2006) A rice Tc1/mariner-like element transposes in yeast. Plant Cell 18: 2469–2478

Yin H, Liu J, Xu Y, Liu X, Zhang S, Ma J, Du J (2013) TARE1, a Mutated Copia-Like LTR Retrotransposon Followed by Recent Massive Amplification in Tomato. PLoS ONE 8: e68587

Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature. doi: 10.1038/nature10625

Yu Z, Wright SI, Bureau TE (2000) Mutator-like elements in Arabidopsis thaliana. Structure, diversity and evolution. Genetics 156: 2019–2031

Zemach A, Kim MY, Hsieh P-H, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell 153: 193–205

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, et al (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nature Biotechnology 30: 549–554

Zhou L, Mitra R, Atkinson PW, Burgess Hickman A, Dyda F, Craig NL (2004) Transposition of hAT elements links transposable elements and V(D)J recombination. Nature 432: 995–1001

Zhu Y (2003) From the Cover: Controlling integration specificity of a yeast retrotransposon. Proceedings of the National Academy of Sciences 100: 5891–5895

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA (2007) Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evolutionary Biology 7: 152

CONCLUDING REMARKS

Transposable elements (TE) are an integral part of eukaryotic genomes and in some organisms they occupy a substantial portion. TE classification can be complex as multiple levels of classes and subclasses exist. Based on the transposition intermediates used, TEs are categorized into two major groups (RNA and DNA TEs). In addition, they belong to different superfamilies distinguished by structural features, protein conservation and replication strategies, intrinsic to the group (Wicker et al., 2007). These features include the terminal sequences (TIR or LTR), target site duplication (TSD) and proteins required for mobilization. Further, within each superfamily are families that consist of phylogenetically related copies sharing high sequence similarity (Wicker et al., 2007).

The activity of TE in a genome can be deleterious, neutral or beneficial. Thus, different silencing strategies exist to regulate its activity and many TEs in the host genome are currently inactive (Mills et al., 2007; Lisch, 2009). Regardless, numerous studies illustrate how TEs can introduce various forms of genetic variations with adaptive and evolutionary potentials (reviewed in Oliver and Greene, 2009; Oliver and Greene, 2009; Kejnovsky et al., 2012). The myriad of roles, ranging from structural to regulatory, played by transposable elements in a host makes their annotation an important component to every genome sequencing project. As more genomes with annotated repeat content are available, various types of comparative analysis may be performed to address important biological and evolutionary questions such as, time and mechanism of genome expansions and development of gene regulatory networks both of which can be influenced by TE activity.

In this research, the repetitive content of a basal eudicot genome (sacred lotus) was annotated and characterized. Our results show that its genome shows typical similarities to most

previously sequenced angiosperms including the high TE contribution to the genome size. This makes the sacred lotus TE information useful not only in downstream gene annotation work but also in further understanding the role of TEs in gene and genome evolution. Being a perennial adaptive to aquatic environments, combined with the many beneficial adaptive roles that TEs play, it will not be surprising to find TEs involved in genome expansion, gene regulation and sequence exaptation that may have been critical to its growth habit. Moreover, results from sacred lotus may be used in comparative studies with other eudicots and monocots to identify the role TEs may play in the diversification within this lineage. Oliver et al. (2013) proposes that TE activity in concert with other factors including hybridization, polyploidization, horizontal transfer and stress may have promoted angiosperm diversity and dominance.

Certain unique features of sacred lotus TE also may make it valuable in TE biology studies. For instance, sacred lotus possesses an unprecedented content of the *hAT* superfamily. Its *hAT* coverage is not only exceptional among DNA TE superfamilies but its success outperforms activity by all DNA TEs in many angiosperms sequenced to date. Since this study has shown that majority of the autonomous members identified in this superfamily are likely inactive and non-functional, a *hAT* transposase can be molecularly reconstituted similar to work with *Sleeping Beauty* (Ivics et al., 1997). This functional autonomous copy or transposase may then be used in transposition and silencing studies to determine how these aspects may have played a role in the success of this superfamily. This superfamily includes the *Ac/Ds* elements which were the first TEs discovered (McClintock, 1951) as well as the domesticated SLEEPER genes which are unique and conserved in angiosperms (Knip et al., 2012). Therefore, analysis of remnant *hAT* related sequences in the genome may be done to determine its role in gene domestication and gene regulation which might explain its retention despite experiencing silencing and inactivation.

Another successful TE group in sacred lotus are the non-canonical *Copia*-type LTR elements that have amplified despite the lack of the conserved "TG-CA" end sequence previously suggested as important for integration (Temin, 1981). These TEs are abundant in sacred lotus suggesting either a unique capacity for transposition and retention of these elements in sacred lotus or that most previous TE annotations may have missed these types of sequences due to their non-canonical end sequences which are not typically used in automated annotation programs. In either case, it will be necessary to elucidate the integration process for this specific group of *Copia* elements as well as the evolution of the catalytic domain in the integrase protein that accommodates for mutations in the end regions.

Meanwhile, the *Tc1/Mariner* superfamily suffered the opposite outcome where it has been eliminated from the genome. The contrasting fates of these two superfamilies in scared lotus may reflect the effect of differential amplification, silencing and retention between different TE types. Since sacred lotus is the fourth genome reported to be devoid of these elements (Velasco et al., 2007; Banks et al., 2011; D'Hont et al., 2012) , it can be used in comparative studies to understand the apparent lack of success of *Tc1/Mariner* elements in these genomes which may be related to intense genome silencing and removal.

A beneficial genetic variation that TEs provide to its host is the exaptation of coding sequences that can generate new adaptive functions. A few TE superfamilies reported to duplicate host genes are *LINEs*, *Helitrons*, *CACTA* and Pack-MULEs (Moran, 1999; Jiang et al., 2004; Morgante et al., 2005; Zabala, 2005). In rice, close to 3000 Pack-MULEs are found that transduplicated about 1500 parental genes. These elements show sequence conservation and regulation of expression of their parental genes (Hanada et al., 2009). This research shows that terminal and subterminal sequences may play a role in the sequence acquisition process. Also,

GC content and breadth of expression plays an additive role in the preferential acquisition by Pack-MULEs, with the GC effect being stronger. These results indicate that Pack-MULEs are "clever elements" in the genome that appear to distinguish and preferentially capture genes rather than other genomic sequences. This unique capability allows these elements increased chances of carrying functional sequences that may provide new genetic resources for the evolution of new genes or the modification of existing genes. In addition, this capability may have contributed to its survival and retention in the genome. However, these new data is just the beginning for understanding how Pack-MULEs form and the precise molecular mechanism of sequence capture remains to be elucidated. Nevertheless, the understanding of preferential acquisition is vital in manipulating future experiments to catch a Pack-MULE capture "in action". The understanding of how the gene fragments are captured is also important in understanding its biology and will be useful in potential usage of the capture process as a useful mutational tool to generate novel coding sequences with new adaptive potentials. Furthermore, this study illustrates that certain MULE families are capable of duplicating Terminal Inverted Repeat (TIR) sequences to generate elements with tandem TIRs on both ends, although the mechanism for this duplication is still unknown. This atypical structure seems to play a role in enhanced transposition efficiency for these specific MULE families suggesting that TIR duplication can provide a benefit in terms of a TE's successful colonization in a host. *In vivo* and *in vitro* studies to test transposition efficiency conferred by duplicated TIRs will be tested in the future.

Taken together, this research provides novel and supporting information of how TEs play a role relevant to gene and genome evolution. The unique TE components in sacred lotus illustrate the need to investigate the reasons behind the differential success and failure of various

TE types in a host genome. While results from MULEs and Pack-MULEs analyses indicate the apparent extraordinary nature of TEs to duplicate host sequences and generate functional sequences.

REFERENCES

REFERENCES

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, et al (2011) The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants. Science 332: 960–963

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488: 213–217

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu S-H, Jiang N (2009) The Functional Role of Pack-MULEs in Rice Inferred from Purifying Selection and Expression Profile. THE PLANT CELL ONLINE 21: 25–38

Ivics Z, Hackett PB, Plasterk RH, Izsvák Z (1997) Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. Cell 91: 501–510

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573

Kejnovsky E, Hawkins JS, Feschotte C (2012) Plant Transposable Elements: Biology and Evolution. *In* JF Wendel, J Greilhuber, J Dolezel, IJ Leitch, eds, Plant Genome Diversity Volume 1. Springer Vienna, Vienna, pp 17–34

Knip M, de Pater S, Hooykaas PJ (2012) The SLEEPER genes: a transposase-derived angiosperm-specific gene family. BMC Plant Biology 12: 192

Lisch D (2009) Epigenetic Regulation of Transposable Elements in Plants. Annual Review of Plant Biology 60: 43–66

McClintock B (1951) CHROMOSOME ORGANIZATION AND GENIC EXPRESSION. Cold Spring Harbor Symposia on Quantitative Biology 16: 13–47

Mills RE, Bennett EA, Iskow RC, Devine SE (2007) Which transposable elements are active in the human genome? Trends in Genetics 23: 183–191

Moran JV (1999) Exon Shuffling by L1 Retrotransposition. Science 283: 1530–1534

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nature Genetics 37: 997–1002

Oliver KR, Greene WK (2009) Transposable elements: powerful facilitators of evolution. BioEssays 31: 703–714

Oliver KR, McComb JA, Greene WK (2013) Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity. Genome Biology and Evolution 5: 1886–1901

Temin HM (1981) Structure, variation and synthesis of retrovirus long terminal repeat. Cell 27: 1–3

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, et al (2007) A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLoS ONE 2: e1326

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al (2007) A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973–982

Zabala G (2005) The wp Mutation of Glycine max Carries a Gene-Fragment-Rich Transposon of the CACTA Superfamily. THE PLANT CELL ONLINE 17: 2619–2632