PLANT *MUTATOR*-LIKE TRANSPOSABLE ELEMENTS (MULES): THEIR EVOLUTIONARY DYNAMICS, INTERACTION WITH GENES, AND RECAPITULATION OF TRANSPOSITION ACTIVITY IN YEAST

By

Dongyan Zhao

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

PLANT *MUTATOR*-LIKE TRANSPOSABLE ELEMENTS (MULES): THEIR EVOLUTIONARY DYNAMICS, INTERACTION WITH GENES, AND RECAPITULATION OF TRANSPOSITION ACTIVITY IN YEAST

By

Dongyan Zhao

Transposable elements (TEs) are genomic sequences that can move from one position to another within a genome, where the movement is catalyzed by transposases. *Mutator*-like transposable elements (MULEs) belong to a highly mutagenic and widespread TE superfamily. This dissertation focuses on studying the evolution and biology of MULEs in plants, specifically in three projects described below.

While the maize genome (2,500 Mb) is six times larger than the rice genome (380 Mb), it contains fewer MULEs than the latter (12,900 vs. 32,000). The differential amplification of MULEs in the two genomes prompted us to investigate the status of MULEs containing transposase (coding-MULEs). The results indicate that they harbor similar amount of candidate coding-MULEs; however, the majority of candidate coding-MULEs in maize are defective. This is partly due to the higher amount of LTR retrotransposons that disrupt the coding region of MULEs in maize. Additionally, the candidate coding-MULEs seem to be subjected to higher indel rate in maize than that in rice, which accelerate the deterioration of maize elements. Collectively, nested insertions and accumulation of indels may explain the low abundance of MULEs in maize than that in rice.

To date, MULEs have been only shown to be active in their native hosts. In this study, transposition of a rice MULE was recapitulated in yeast, representing the first report of MULE transposition in a heterologous species. The wild-type transposase induced low transposition frequency; however, it could be improved by a variety of transposase modifications. Deletion of the N-terminal 129 amino acids led to enhanced activity as well as altered cellular localization of the transposase. Mutational analysis revealed a critical region

of the transposase, where changes of the amino acid compositions resulted in either enhanced or repressed activity. Additionally, fusion of a peptide to the N-terminal deleted transposase also enhanced transposition frequency, which is the first report of transposition activity enhancement by protein fusion. Taken together, the establishment of the MULE transposition system in yeast laid the foundation for further studying MULE biology.

MULEs have the propensity to insert into genic regions, which influence the expression of adjacent genes. To determine whether MULEs played any role in the Illinois Long-Term Selection Experiment (ILTSE) maize strains, MULE insertions that are co-segregating with either high or low protein maize strains were studied. Consistent with previous studies, most insertions (~79%) were located in low-copy regions. Interestingly, compared with MULEs in the B73 maize genome, co-segregating insertions are over-represented in exons and equally represented in the 5′ and 3′ regions of genes, which is in contrast to the 5′ insertion preference of MULEs reported in previous studies. Expression analysis revealed that out of 55 genes with adjacent co-segregating insertions, over 1/4 exhibited altered expression levels and may be associated with the selected trait. Further studies are needed to test whether there are causal relationships between these co-segregating MULE insertions and kernel protein content.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Ning Jiang, who gave me the chance to work in her lab, allowed and guided me to conduct different kinds of research, and helped me with my writing. I have grown immensely over the past few years under her guidance. Thanks to our frequent discussions, which spurred novel ideas for my projects, allowed me to see the beauty of science and know how to be a good researcher. I also thank members of my guidance committee, Dr. Cornellius Barry, Dr. Rebecca Grumet, Dr. Shin-Han Shiu, and Dr. Dechun Wang, who provided constructive and insightful ideas for my research.

I thank Dr. Steve Moose and Dr. Han Zhao (University of Illinois at Urbana-Champaign) for kindly providing materials of the Illinois Long-Term Selection Experiment maize populations; Dr. Sue Wessler and Dr. Nathan Hancock (University of Georgia) for providing me the vectors and protocols for the yeast project. Special thanks go to Dr. Robin Buell for her effort in organizing the Applied Bioinformatics Workshop, where I learned skills in analyzing next-generation sequencing data. I also thank Robin for letting me be her teaching assistant of the Plant Genomics course, which is a valuable experience. I thank Dr. Melinda Frame for helping me with the confocal microscopy experiment. Thanks also go to all the collaborators and friends for helping me with my research work, Dr. Guo-qing Song, Dr. Zhulong Chan, Dr. Pingfang Yang, Dr. John Hamilton, Dr. Kevin Childs, Dr. Yuehua Cui, Dr. Jiming Jiang (University of Wisconsin-Madison), Dr. Dave Douches, Dr. Miguel Flores, Dr. Melissa Lehti-Shiu, and Dr. Gaurav Moghe.

Thanks go to the past and current members of the Jiang lab, Dongmei Yin, Ann Armenia Ferguson, Veronica Vallejo, Stefan Cerbin, Dongying Gao, Jason Miller, Shujun, Ou, and

Scott Funkhouser. I especially thank Dongmei, Ann, and Veronica for helping me adapt to the new environment when I first joined the lab. I also thank Jason for providing me protocols for the yeast project.

The joint-lab meeting of Department of Horticulture provided me chance to present my work, which allowed me to improve my communication and presentation skills. I thank all the labs participating this meeting and all the people for their comments and suggestions. I also thank Bill Chase and Gary Winchell of the Horticulture Teaching and Research Centre for helping me with my corn field work every summer.

Finally, I would like to thank my family for their support and understanding of me being an “eternal” student (25 years) and being far away from them to pursue my study. Special thanks to my husband, who is always there listening to my complaints (mostly through video chat) and providing me moral support. Meeting and knowing him is a turning point of my life and I appreciate him being part of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction and Literature Review

1.1 Transposable Elements and Their Classification

Transposable elements (TEs), also called transposons, are segments of DNA that can move from one position to another within a genome and the process is referred to as transposition. According to a variety of criteria, TEs are divided into different groups (Wicker et al., 2007; Kapitonov and Jurka, 2008; Abrusan et al., 2009). Based on the mode of transposition, they are categorized into two major classes. Class I TEs transpose via an RNA intermediate using a "copy-and-paste" mechanism and class II TEs transpose via a DNA intermediate through a "cut-and-paste" mechanism. Both TE classes consist of autonomous and non-autonomous elements, where the former encode protein(s) responsible for catalyzing transposition of themselves and their corresponding non-autonomous elements. In most cases, non-autonomous TEs are usually deletion derivatives of the cognate autonomous elements. Common to most TEs is that they duplicate a short stretch of nucleotides (2-11 bp) of the genomic loci where they integrate into, which is called target site duplication (TSD).

1.1.1 Class I RNA TEs

Class I TEs, also called RNA TEs or retrotransposons, can be further divided into two groups: long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons, whereby the latter consists of long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (Figure 1b, c). As indicated by the name, LTR retrotransposons are characterized by the LTR sequences at their 5′ and 3′ ends, which flank sequences encoding both structural and enzymatic proteins for their transposition. Typically, autonomous LTR retrotransposons contain two genes, *i.e.*, *gag* and *pol*, where the former encodes proteins that form the virus-like particles in which the mRNAs of these elements are reverse-transcribed into cDNA followed by integration into new genomic loci. This process is catalyzed by proteins encoded by the *pol* gene, including a protease, RNase H, reverse transcriptase (RT),

and integrase (IN) (Havecker et al., 2004). Based on the organization of the proteins encoded by the *pol* gene, LTR retrotransposons are further divided into two major subgroups: the *Ty1/Copia*-like and *Ty3/Gypsy*-like elements.

Non-LTR retrotransposons harbor no LTR at their ends and are characterized with a poly adenine tract at their 3′ termini. Autonomous (functional) non-LTR retrotransposons (LINEs) usually contain two open reading frames (ORFs), encoding proteins necessary for the transposition (a nucleic acid binding protein (*gag*), an endonuclease and a reverse transcriptase). SINEs are shorter in size than LINEs. Sequences of the 3′ region of SINEs show similarity to that of LINEs while their 5′ regions are more similar to tRNA genes or 7SL RNA genes (Schmidt, 1999). They do not encode reverse transcription machinery and rely on those from related autonomous elements, *e.g.*, LINEs.

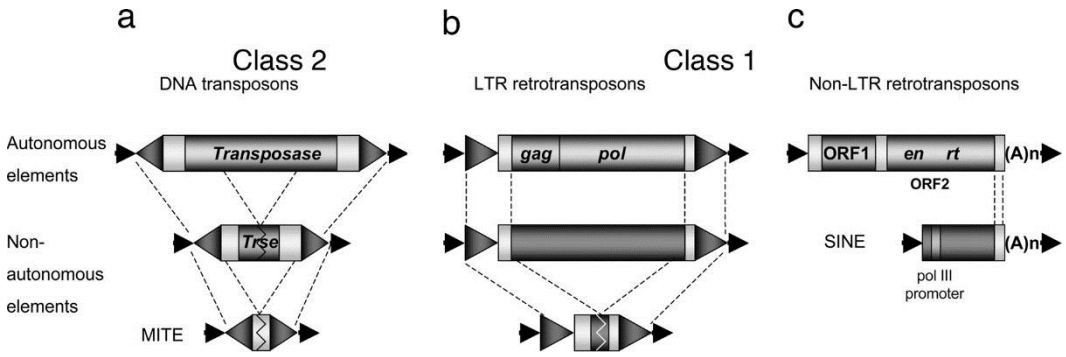

Figure 1.1 Schematic representation of TE classes (from Feschotte et al., 2002)

1.1.2 Class II DNA TEs

The majority of class II TEs (DNA TEs) are flanked by terminal inverted repeats (TIRs) (Figure. 1a). The length of TIRs varies among different DNA TE families and usually ranges from several (*e.g., hAT, Tc1/mariner* elements) to hundreds of nucleotides (such as *Mutator* elements (see below)). TIRs contain binding sites for the transposases, proteins encoded by autonomous DNA TEs and are responsible for both excision and reintegration of DNA TEs

(Wicker et al., 2007). Meanwhile, TIRs of non-autonomous elements are also recognized by transposase proteins encoded by the autonomous elements sharing similar TIR sequences. Another feature of TIRs is that some of them harbor *cis*-acting elements, such as promoters, which guide the expression of genes (including transposase) within the elements and sometimes genes in close proximity of the elements if outward promoters are within the TIRs (Barkan and Martienssen, 1991). Transposition of DNA TEs is usually non-duplicative and the increase in their copy number usually occurs through repair of the original TE locus using either sister chromatid (chromosome) or homologous sequences within the genome.

Besides DNA TEs with TIRs, there occurs another family, *Helitrons*, which contain no TIRs and transpose through a rolling-circle mechanism. They are characterized by the presence of a small hairpin structure at their 3′ end and the conserved "TC" and "CTRR" nucleotides at their 5′ and 3′ termini, respectively (Kapitonov and Jurka, 2001). Computational analysis based on these structural features revealed that *Helitrons* are ubiquitous and are present in fungi, plants and animals (Poulter et al., 2003; Hood, 2005; Yang and Bennetzen, 2009). Autonomous *Helitron* elements encode *Rep*/*helicase*-like and/or *replication protein A*-like proteins, which are involved in the transposition process. Unlike other TEs, they do not generate TSDs.

1.2 Abundance of TEs and Their Impact on Genome Size Variation

1.2.1 Abundance of TEs

Since the discovery of the first TEs (*Activator/Dissociation*) by Barbara McClintock in 1950s (Mcclintock, 1950), TEs are found to be abundant in a wide range of organisms. TEs constitute large fractions of most eukaryotic genomes sequenced to date. For example, ~85% of maize (*Zea mays*), ~61% of sorghum (*Sorghum bicolor*), and ~59% of soybean (*Glycine*

max) (Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010) are occupied by TEs. Even in the very compact *Arabidopsis thaliana* genome, TEs still account for more than 10% of the genome (Kaul et al., 2000). A myriad of studies suggest that amplification and loss of certain TE families could be the most important source for genome size and structure variations within and between different plant species (Ammiraju et al., 2007; Huang et al., 2008; Tenaillon et al., 2011).

1.2.2 Impact of TEs on Genome Size Variation

Comparative studies on genome sequences revealed that LTR-retrotransposons were the largest contributor to genome expansion in plants (Brunner et al., 2005; Piegu et al., 2006). After diverging from a common ancestor about 15-20 million years ago (mya), the maize genome size has increased at least 3 times compared with its close relative, sorghum (SanMiguel et al., 1996). The dramatic difference is attributed to the presence of retrotransposons in the intergenic regions in maize (Baucom et al., 2009). Through comparative analysis of allelic chromosomal regions between the two maize inbred lines Mo17 and B73, over 50% sequences could not align with one another, which indicate polymorphisms between the sequences (Brunner et al., 2005). A detailed study on these sequences suggested that the differences were mainly due to insertions of LTR-retrotransposons, which are mostly full-length elements and carry target site duplication (Brunner et al., 2005). Tenaillon et al. (2011) conducted a shot-gun sequencing of maize (*Zea mays* ssp. *mays*. B73) and *Zea luxurians*. Their analysis revealed that 50% of the genome size difference between B73 and *Z. luxurians* can be attributed to the composition of TEs (both class I and class II TEs). Besides maize, investigation of an LTR retrotransposon family (*RWG*: *RIRE2*, *Wallabi*, *Gran3*) in different *Oryza* species, including all the wild species and the two domesticated species (*O. sativa* and *O. glaberrima*) indicated that the RWG family

contributes to the genomic variations across these *Oryza* lineages and accounts for almost 25% of the genome enlargement of *O. granulata* compared to that of *O. sativa* (Ammiraju et al., 2007). Another study on transposon insertion polymorphisms (TIPs) between two rice subspecies, *indica* and *japonica*, discovered that more than half of the large insertions and deletions in the rice genome are caused by TIPs (Huang et al., 2008). A total of 2041 TIPs were found between 93-11 (an *indica* cultivar) and Nipponbare (a *japonica* cultivar), which account for 14% of the genomic sequence variations between them.

Like maize, *Gossypium* has also undergone a three-fold increase in genome size. Hawkins and colleagues inspected individual sequence types of transposable elements using whole genome shotgun sequencing data from three diploid members of *Gossypium* and *Gossypioides kirkii*, a phylogenetic out-group, and uncovered that the difference in genome size is largely due to proliferation of one particular group of *gypsy*-like retrotransposon, *Gorge3* (Hawkins et al., 2006).

The fact that TEs contribute a tremendous amount of genomic variation was explained partly by the burst amplification of certain TE families during a very short period (Ammiraju et al., 2007; Tenaillon et al., 2011). In addition, due to their repetitiveness, they provide raw materials for homologous and ectopic recombination between chromosomal regions with similar TE sequences, which leads to the reduction of genome size and the increase of genomic variation (Huang et al., 2008; Wang et al., 2008). Ma and colleagues conducted a comprehensive structural analysis of a representative sample of LTR retrotransposons in rice cultivar *Nipponbare*. In agreement with previous studies, unequal intra-strand homologous recombination (UHR) and illegitimate recombination (IR) are the critical mechanisms that led to the purging of these LTR-retrotransposons (Ma et al., 2004). One surprising result they obtained is that more than 190 Mb of LTR-retrotransposon sequences have been removed

from the relatively compact rice genome. In looking for counterbalancing mechanisms to genomic obesity, it was shown that IR is a major driving force for reducing the size of *Arabidopsis* genome (Devos et al., 2002). Collectively, TEs play a double-edged role with regard to genome size. One the one hand, activation and transposition of TEs can greatly increase the genome size. On the other hand, their repetitiveness provides templates for homologous recombination that leads to the reduction of genome size. As a result, genome contraction may occur when transposition activity is relatively low.

1.3 MULEs and Pack-MULEs

In 1978, Donald Robertson reported the finding of a *Mutator* (*Mu*) system in maize y9 stock, which exhibited ~30-fold increase in the mutation rate compared with two other standard maize inbred lines (M14/W22) (Robertson, 1978). One of his hypotheses is that this *Mu* may be a controlling element system, such as *Ac/Ds*. Indeed, subsequent studies proved this *Mu* belonged to a highly mutagenic TE family, which is comprised of an autonomous (*MuDR*) and seven fully sequenced non-autonomous elements (*Mu1, Mu2/Mu1.7, Mu3, Mu4, Mu5, Mu7, Mu8*) (Bennetzen et al., 1984; Freeling, 1984; Chen et al., 1987; Talbert et al., 1989; Fleenor et al., 1990; Chomet et al., 1991; Hershberger et al., 1991; Qin et al., 1991). In addition, at least four other non-autonomous elements were discovered and partially sequenced. These are *Mu10*, *Mu11*, *Mu12*, and *Mu13* (Dietrich et al., 2002; Tan et al., 2011). All these *Mu* elements share high similarity of TIR sequences (~220 bp), but contain heterologous internal sequences.

1.3.1 Widespread Distribution of MULEs

Since the discovery of the *Mu* element in maize, studies of the family have been expanded to a wide variety of organisms, including other plant species, bacteria, fungi,

animals, and viruses, where these elements were generally referred to as *Mutator*-like elements (MULEs) (Eisen et al., 1994; Yu et al., 2000; Chalvet et al., 2003; Rossi et al., 2004; Neuveglise et al., 2005; Marquez and Pritham, 2010). Using Northern blot, PCR and database search, it was found the MURA-like sequences (MURA: the transposase encoded by *MuDR*) were present in majority of the grass genomes tested, such as barley, sorghum, rice, *setaria*, and wheat (Lisch et al., 2001). In *Arabidopsis thaliana*, Yu et al. (2001) found MULEs with similar structural characteristics to *Mu* elements in maize (MURA-like sequence, long TIRs and ~9 bp TSD) and MULEs that lack the canonical long TIRs, which were named as non-TIR MULEs. Later on, non-TIR-MULEs have been discovered in other organisms, including yeast (*Yarrowia lipolytica*), *Lotus japonicus*, maize (*Zea mays*) and rice (*Oryza sativa*) (Neuveglise et al., 2005; Holligan et al., 2006; Wang and Dooner, 2006; Ferguson et al., 2013). Besides plants, MULEs, designated *Phantom*, have also been discovered in a wide range of animal genomes, providing evidence for the ubiquitous nature of the MULE family (Marquez and Pritham, 2010).

1.3.2 Insertion Preferences of MULEs

Different TEs possess different preference for their insertion sites. Some TEs more frequently integrate into chromosomally linked positions of the original site of the elements. This phenomenon is called “local hopping”. For example, in *Arabidopsis* plants transformed with *Ac/Ds* elements, 68% out of 111 transposed *Ds* elements were in tight linkage with the *Ds* donor site (Bancroft and Dean, 1993), which is similar to the rate observed in maize (59%, Dooner and Belachew, 1989). This phenomenon also applies to some TEs in animals and the *P* element in *Drosophila* is a good example. It was shown that all new integration sites recovered were within 128 kb from the starting site of the *P* elements (Tower and Kurapati, 1994). Unlike the TEs mentioned above, MULEs transpose into both genetically linked and

unlinked sites within the genome (Lisch et al., 1995), which makes them good candidates for whole genome tagging/mutagenesis research (Raizada et al., 2001; McCarty et al., 2013).

When dissecting the insertion sites of MULEs more closely, it is apparent that they exhibit preferences for genic and/or low copy regions in both naturally occurring and artificially induced MULE transpositions (Cresse et al., 1995; Das and Martienssen, 1995; Hanley et al., 2000; Raizada et al., 2001; Dietrich et al., 2002; Liu et al., 2009). In one study on *Mu* transposons, ~965,000 *Mu* flanking sequences (MFSs) were obtained from crosses between *Mu* active maize lines and various inbreds and hybrids (Liu et al., 2009). They found that 365,600 MFSs were uniquely mapped to the B73 maize genome (RefGen_v1). Further analysis of histone modifications in *Mu* insertion sites indicated that *Mu* prefers for open chromatin regions. Within genes, there appear strong preferences for MULE insertions in the 5′ regions. In analyzing *Mu*-induced *glossy8* mutant alleles, over 80% out of 75 confirmed *Mu* insertions were located in the 5′ untranslated region (Dietrich et al., 2002). A similar study revealed the preferential integration of *Mu* elements in the 5′ proximal regions of the *Mes1* locus in maize (Haun et al., 2009). Using high-throughput sequencing technology, Liu et al. (2009) showed that 31,838 uniquely mapped *Mu* insertions were located within or in the vicinity (500 bp flanking sequence of genes) of 32,477 annotated genes (B73 RefGen_v1), where a strong preference was observed for insertions into sequences immediately upstream of the transcription start site.

1.3.3 MULEs as Tagging Agents

Due to the high mutagenicity and non-specific integration sites (see above), MULEs have been used as tagging tools for gene discovery and whole genome mutagenesis in maize. A collection of gene knockout maize mutants generated by *RescueMu*, a modified *Mu1* element which contained part of a *pBluescript* plasmid in its internal region, represented an early

attempt on exploiting the transposition frequency of MULEs. The resulting mutants facilitate the understanding of MULE transposition behavior and serve as materials for study of gene function (Raizada et al., 2001; Fernandes et al., 2004). Another large mutant collection, which is still ongoing, is the development of the *UniformMu* transposon insertion lines (Settles et al., 2007; McCarty et al., 2013). The screening of mutant lines is based on the excision of *Mu1* transposon from a marker gene (*Bz1*) induced by active *MuDR* elements. As of March 2011, there were a total of 12,127 high-confidence genes containing *Mu* insertions and 5,952 lines are freely available for researchers (maizegdb.org/documentation/uniformmu/index.php)

1.3.4 Autonomous MULEs

As mentioned earlier, transposition of non-autonomous MULEs was observed in 1978; however, the regulatory element that catalyzed the transpositions was not discovered until more than 10 years later. In 1991, several groups independently reported the finding of the autonomous *Mutator* element (*MuR1* from Chomet et al., 1991; *Mu9* from Hershberger et al., 1991; *MuA2* from Qin et al., 1991), which was renamed *MuDR*, in honor of Donald Robertson who discovered the first *Mutator* element. *MuDR* encodes two genes, *mudrA* and *mudrB*, which are transcribed in a convergent orientation directed by the *cis*-sequences in the TIRs (Lisch, 2002). The protein product of *mudrA*, MURA exhibited homology to some transposases of insertion sequences in bacteria, thus is considered the transposase for *MuDR* (Eisen et al., 1994). MURB, product of *mudrB*, is suggested to be involved in element reinsertion but its exact function still remains elusive (Lisch et al., 1999; Raizada and Walbot, 2000; Woodhouse et al., 2006). Homologous sequences of MURA have been found in a wide range of organisms while the existence of MURB seems restricted to the *Zea* genus (Yu et al., 2000; Lisch et al., 2001; Rossi et al., 2004).

Besides *MuDR*, several active autonomous MULEs have been discovered, including two more elements in maize (*Jittery* and *TED*), *Hop* and *Mutyl* in fungi, *AtMu1* in *Arabidopsis*, and *Os3378* in rice (Singer et al., 2001; Chalvet et al., 2003; Xu et al., 2004; Neuveglise et al., 2005; Gao, 2012; Li et al., 2013). Except *Mutyl*, all other elements contain one open reading frame (ORF), which was mostly efficient in catalyzing transposition. *Mutyl* contains two ORFs, one is the transposase (1178 residues) while no homologous sequence was found for the second ORF (459 residues) (Neuveglise et al., 2005). These autonomous elements exhibited different characteristics in their transposition, suggesting the heterogeneity of MULEs within and among species. For example, transposition of *Mu* transposons induced by *MuDR* usually occurred in terminally differentiated somatic tissues. Furthermore, *Mu* transpositions in germ lines rarely accompanied with recovery of the donor site (Lisch et al., 1995; Raizada and Walbot, 2000). Similarly, germinal insertions occurred for *AtMu1* without losing the original copy in *Arabidopsis ddm1* (*Decrease in DNA Methylation*) mutant plants (Singer et al., 2001). In contrast to the generation of new copies, the *Jittery* element excised but no integration was observed. This unusual behavior led the authors to hypothesize that either the integration required a B function (like MURB in *MuDR*), which was missing for *Jittery*, or the imperfect 5′ TIR sequence prevented the reintegration (Xu et al., 2004). More recently, another autonomous MULE was discovered in maize, named *TED* (for Transposon Ellen Dempsey), which retains low copy number and seems to be present only in grasses (Li et al., 2013).

1.3.5 Pack-MULEs

Within the MULE family, a specific group, called Pack-*Mutator*-like transposable elements (Pack-MULEs), is distinguished from other MULE elements by having gene or gene fragments between the TIRs instead of the transposase (Jiang et al., 2004). In rice, there

are about 3000 copies of Pack-MULEs, which acquired sequences from over 1,500 non-TE genes (parental genes). Detailed exploration of available expression datasets revealed that ~22% have expression evidence and 28 have direct evidence of translation (Hanada et al., 2009). In addition, many Pack-MULEs were under purifying selection, implying functional roles of these elements (Hanada et al., 2009; Jiang et al., 2011). Subsequent studies revealed a unique impact of Pack-MULEs in affecting gene structure and genome composition in rice and maize (Jiang et al., 2011). Specifically, Pack-MULEs preferentially acquired "GC"-rich sequences and targeted their integration into 5′ regions of genes, therefore, resulted in formation of genes with negative "GC" gradient (genes with decreasing "GC" content towards their transcription termination sites). More recently, Ferguson et al. (2013) demonstrated that besides the preference for high "GC" sequences, Pack-MULEs exhibited biased acquisitions of genes with broad expression (genes expressed in more tissues and/or under more conditions). Collectively, these data indicated that the impact of Pack-MULEs on gene and genome evolution may be more significant and diversified than what is currently known.

1.4 Structure of Transposases

1.4.1 The DDE Catalytic Domain

The transposition of class II TEs depends on the transposase protein encoded by autonomous elements, which is also responsible for the movement of the corresponding non-autonomous elements. A recent analysis suggested that all eukaryotic class II transposases contain similar catalytic domain structure, which is, in most cases, at the carboxyl-terminal region (Yuan and Wessler, 2011). That is the conserved acidic amino acid triad consisting of two aspartic acids and a glutamic acid or a third aspartic acid ("DDE/D") in the catalytic core, which features an RNase H-like fold. The three amino acids are brought in close proximity to

each other by the RNase H-like fold (Nesmelova and Hackett, 2010). Besides the amino acid triad, other "signature strings" were found in the catalytic core among different class II TE superfamilies, which led the authors propose new classification of these TE superfamilies (Yuan and Wessler, 2011). A study of several transposases with available crystal structures (*e.g.*, transposases of Tn*5*, *Hermes*, bacteria phage *Mu*, and *Mos1*) revealed that the distance between the first two residues ("D") are more conserved than that between the second "D" and last "D/E" (Nesmelova and Hackett, 2010) and there appear intervening sequences in the latter with various lengths. For example, a stretch of ~90 amino acids was inserted in the second "D" and last "D/E" in Tn*5* transposase and for that of *Hermes* about 300 amino acids were in that region. Despite the variability of the length of the RNase H-like fold, the transposases form similar three dimensional structures during the catalytic process. The three amino acids are critical for the activity, which was evidenced by complete loss of transposition if they were mutated (Zhou et al., 2004; Lazarow et al., 2012). Without exception, the MULE transposases also possess the "DDE" catalytic domain in the carboxyl termini. For MURA, the transposase of *MuDR*, they are at amino acid positions 331 (the first "D"), 393 (the second "D"), and 489 (the "E"). Between the first two "D"s, there also occurs a conversed "DE" at positions 350 and 366, respectively, which can be found in many MURA-like proteins (Hua-Van and Capy, 2008).

1.4.2 The DNA-Binding Domain

Besides the catalytic domain, a competent transposase also requires a DNA-binding domain, which recognizes and binds to the TIRs of the transposon. The DNA-binding domain usually features single or multiple helix-turn-helix motif(s) (HTH), which are often located near the N-termini of the transposases (Izsvak et al., 2002; Nagy et al., 2004; Hennig and Ziebuhr, 2010). Similar to the conserved catalytic domain, most class II transposases seem to

contain HTH motif(s) in the DNA-binding domain. However, the conservation is more at the structural level of the proteins and very little conservation at the amino acid level. One of the most active *Tc1/mariner* transposons, *Sleeping Beauty*, contains two consecutive HTH motifs at its N-terminus, which were referred to as PAI and RED, respectively. Binding assays showed that the PAI domain (the first N-terminal HTH) is the major structure that contacts the TIR sequences of the element (Izsvak et al., 2002). Likewise, one bacterial insertion sequence, *IS30* also contains two HTH motifs (referred to as HTH1 and H-HTH2) at the N-terminus of its transposase (Nagy et al., 2004). Nevertheless, site-directed mutagenesis in these two motifs revealed that the second N-terminal HTH (H-HTH2) played the major role of binding to the terminal sequence of the element. The bacterial phage *Mu* transposase (MuA) contains three HTH motifs at its N-terminus, where the first motif was suggested to recognize the *Mu* enhancer and facilitate the assembly of the transposase and transposon complex while the other two HTH motifs were involved in its DNA-binding activity (Watson and Chaconas, 1996; Schumacher et al., 1997). In addition to these transposases containing modular structures of the DNA binding domain, some transposases only contain one HTH motif, such as that of Tn*5* and *Hermes* (Nesmelova and Hackett, 2010). Hence, the organization of the DNA-binding structures varies among transposases from different TE superfamilies. Using gel-shifting assay, Benito and Walbot (1997) demonstrated that MURA (*MuDR* transposase) was a DNA-binding protein, which recognized a 32-bp sequence in the TIRs of the *Mu* transposons. The transposase of the most closely related bacterial insertion sequence to the *Mu* transposons, *IS256*, was also proved to be a DNA-binding protein (Hennig and Ziebuhr, 2010). Secondary structure analysis identified two HTH structures in its N-terminal region, where the second HTH was responsible for the binding of the transposase to the transposon ends.

Common to most class II TEs, transposition starts with binding of the transposase to the terminal ends of the transposon followed by cleavage of the transposon and finishes by either integration into new genomic loci or no integration (Munoz-Lopez and Garcia-Perez, 2010). Hence, to determine whether a transposon has the potency of transposition activity, the presence of both a catalytic domain and a DNA-binding domain is prerequisite.

1.5 Regulatory Roles of TEs

TEs play important roles in both plants and animals through regulating expression of adjacent genes, which can be evidenced by large scale expression studies and small scale investigation of individual genes with TE insertions (Feschotte, 2008; Naito et al., 2009; Pereira et al., 2009). Using high-throughput sequencing and comparative microarray analysis, Naito et al. (2009) discovered 1,664 *mPing* (a DNA transposon) insertion sites in a rice variety. In this study, expression of the majority of the 710 neighboring genes were either up-regulated or had no change, suggesting a modest or positive impact of *mPing* insertion on gene transcription. In a broader evolutionary time scale, the effect of TE insertions on gene expression was studied in mouse and rat (Pereira et al., 2009). It was estimated that ~20 % of the expression profile divergence between mouse and rat may be due to TE insertions, reinforcing the important role of TE insertions in regulating gene expression. Molecular analysis of genes with TE insertions within or in their vicinity revealed the diverse effects of the insertions on those genes.

1.5.1 TE Insertions Affect Gene Expression

McGinnis et al. (1983) studied a developmentally regulated gene, *Sgs-4,* in Drosophila, and found that insertion of a DNA TE (named *hobo*) just upstream of the TATA box reduced *Sgs-4* expression 50- to 100-fold compared to the wild-type allele. Moreover, two additional

transcript splicing forms were produced that were initiated within the *hobo* element, suggesting that the element provided transcription initiation signals for the gene (McGinnis et al., 1983) (also see Section 1.5.2). In maize, a *Mu1* insertion in the first intron of *Adh1-S* (*alcohol dehydrogenase-1*) resulted in suppressed expression of alcohol dehydrogenase (40% of the wild type level) and reduced enzyme activity (Bennetzen et al., 1984). The fact that TE insertions cause null/reduced function of targeted genes have prompted generation of large collections of TE-tagging mutants in both plants and animals (Keng et al., 2005; Kolkman et al., 2005; Settles et al., 2007; Izsvak et al., 2010). These collections provided resources for discovery of gene function. In addition, there are cases that TE insertions may enhance gene expression. For example, investigation of the rice genome sequence revealed two copies of ribosomal protein L6 gene (*OsRpl6-1* and *OsRpl6-2*), where *OsRpl6-1* showed higher expression than that of *OsRpl6-2*. Sequence search established that *OsRpl6-1* contained sequence homologous to a *PIF/Harbinger* TE in its 5′ untranslated region, which may render the higher expression of this copy (Kubo et al., 2008).

1.5.2 TEs Serve as *cis*-elements of Protein-Coding Genes

TEs can be exapted by the host genome as *cis*-regulatory elements for protein-coding genes. A myriad of cases have been reported of TEs serving as promoters, enhancers, and providing alternative transcription initiation and polyadenylation sites. A comprehensive study of gene promoter regions (2 kb sequences upstream of the transcription initiation sites) in human demonstrated that ~83% of 20,193 promoter regions harbor TE insertions. Further analysis revealed that ~80% of the gene promoter regions contained TEs with regulatory elements, suggesting that these genes may be subject to the regulation of these nested TEs (Thornburg et al., 2006). In maize, there occurred many incidents that *Mu* transposons served as *cis*-acting elements. One excellent example is the *Mu* insertion alleles of *rf2a*, which

encodes a mitochondrial aldehyde dehydrogenase and loss of function of the gene resulted in male sterility in maize (Cui et al., 1996). Three *rf2a* alleles with *Mu* transposon insertions in maize were studied, where two contained *Mu* transposon in the 5′ untranslated region (UTR) and one in the 3′ UTR. In the absence of active *MuDR*, functional transcripts were produced because *Mu* elements provided alternative initiation sites in the first two alleles and alternative polyadenylation site in the latter case (Cui et al., 2003). However, when active *MuDR* was introduced into these maize lines through crossing with *Mu*-active lines, various frequencies of male-sterile progenies appeared from these crosses, suggesting the loss of function of the *rf2a* gene. The emergence of mutant phenotype caused by *Mu* insertion in the presence of active *MuDR* and loss of mutant phenotype in the absence of active *MuDR* has been observed in other *Mu*-insertion mutants; the phenomenon was called *Mu* suppression (Martienssen et al., 1989; Martienssen et al., 1990; Barkan and Martienssen, 1991; Lowe et al., 1992; Girard and Freeling, 2000).

1.5.3 TEs Serves as Splicing Signals

TE insertions can also lead to alternative RNA splicing, resulting in production of different forms of transcripts. Chen et al. (2012) found that a transposon insertion in the first intron of the *CAD2* gene (*Cinnamyl alcohol dehydrogenase*) resulted in a truncated protein, which was over 300 amino acids shorter than the wild type product. The cause is that a splicing site within the inserted transposon was used, which resulted in the incorporation of part of the transposon in the transcripts and a stop codon within the incorporated sequence led to the early maturation of the gene (Chen et al., 2012). Another example involves a CACTA element (5.7 kb) which captured sequences from five host genes (Zabala and Vodkin, 2007). Insertion of this element in the second intron of flavanone 3-hydroxylase gene resulted in generation of 10 distinct forms of transcripts, including the wild-type transcript. This is

because the inserted CACTA element provided a number of splicing sites for the gene. More recently, a comprehensive study revealed that MITEs contribute to 43 splicing sites in 40 protein coding genes in rice (Oki et al., 2008).

1.5.4 Novel Genes Derived from TEs

The importance of TEs in gene and genome evolution is evidenced by the abundance of protein coding genes with the entire or partial sequences from TEs. Bioinformatics studies discovered domesticated transposases in both plants and animals. In *Arabidopsis*, *FAR1* and *FHY3* are involved in the phytochrome A signaling pathway. Sequence analysis revealed that they were related to the MURA transposase of the *MuDR* element and their regulation of target genes resembled the binding activity of the transposase to TIR sequence of the element (Hudson et al., 2003). Recently, evidence was found that the core sequence (~600 amino acids out of the total ~1040 residues) of the *RAG1* gene, which functions in the V(D)J recombination in jawed vertebrates, has emerged from a transposase protein of the *Transib* superfamily (Kapitonov and Jurka, 2005).

Besides examples of single genes domesticated from TEs, large scale studies demonstrated the extent that TEs can be exapted and became the integral parts of many protein-coding genes. The exonization of TEs in human genome is one of the best examples. The RNA TE, *Alu* elements (SINE) account for ~10% of the human genome and have been identified in numerous mature mRNAs. The new exons (or portion of the exons) generated by these elements are often alternatively spliced and contribute to ~5% of the human transcriptome (Sorek et al., 2002). In plants, large amount of TE sequences were also found to be components of many genes. In *Arabidopsis*, 7.8% of the expressed genes harboring TE sequences and 1.2% have protein evidence of the exonization of TEs (Lockton and Gaut, 2009). One study in rice found that ~10% of the genes examined (n = 20,507) contained TEs

in their exonic regions (Sakai et al., 2007). A study specifically on MITEs revealed that they are components of more than 300 protein-coding genes in rice (Oki et al., 2008). Eighteen protein coding genes exapted MITE sequences as transcription initiation sites and 203 as poly(A) sites.

1.5.5 TEs Induce Epigenetic Changes

Plant genomes have evolved ways to keep transposons inactive, one of which is epigenetic silencing. This repression of transposon can be exerted at two levels; transcriptional silencing through DNA methylation and post-transcriptional degradation of transposon mRNA or inhibition of translation. TE insertions often accompany methylation of DNA and/or modifications of histones with repressive marks. TE methylation is negatively correlated with gene expression and methylated TEs close to genes are selected against, thus resulting in a gradual loss (Hollister and Gaut, 2009).

In studying flower pigmentation associated with TE insertions in morning glory, Iida et al. (2004) found that a MULE insertion (*ItMULE1*) in the promoter region of the *DFR-B* gene (encoding dihydroflavonol 4-reductase in anthocyanin biosynthesis) resulted in flower variegation. Interestingly, this variegation only occurred when the DNA methylation spread from *ItMULE1* to the adjacent sequences in the promoter (Iida et al., 2004). Although rare, TE insertions would lead to enhanced gene expression. The tissue-specific pigmentation of maize is mediated by the *pericarp color1* gene (*p1*), where, in some maize lines, it occurred as tandem repeated copies. A *Mu* insertion in the 5′ untranslated region of one copy resulted in ectopic expression of one adjacent uninterrupted copy (Robbins et al., 2008). It was suggested that the hypo-methylation of an enhancer upstream of the *Mu* insertion was responsible for the enhanced expression of the uninterrupted *p1* copy, which may lie close to the enhancer.

1.6 TEs in Population, Adaptation and Artificial Selection of Plants

1.6.1 TE Activation by Stress

Activation of TEs is often mediated by stress (Wessler, 1996). In other words, TEs often change from a silent state to active state under stress conditions. A comparative study showed that in a marine diatom, expression of LTR retrotransposons were highly induced by some stress conditions, including nitrate starvation and exposure to the toxic aldehydes (metabolites produced by diatom). This is also supported by the hypomethylation of the LTR retrotransposons, suggestive of active chromatin state (Maumus et al., 2009). Furthermore, differential abundance of LTR retrotransposons was proposed to contribute to the intraspecific genetic diversity among pennate diatom. Another example is the preferential insertions of some TEs into promoters of RNA Pol II-transcribed genes which are stress-inducible. Guo and Levin (2010) identified a total of 73,125 independent integration events of *Tf1* LTR retrotransposon in *S. pombe*, 76% of which were distributed in intergenic sequences that contained 31% of the promoters. Gene ontology analysis of genes downstream of 219 independent insertions revealed high percentage of the genes were stress responsive (Guo and Levin, 2010). More recently, a genome-wide study of transcriptome of the rice anther under cold condition (12 ℃) revealed that expression of 9.6% of repeat sequences (*e.g.*, MITEs) increased at least 2-fold compared to rice grown under normal temperature (Ishiguro et al., 2014). Collectively, activation of TEs under stress conditions may create new mutations, some of which are likely beneficial for the host. Indeed, it has been suggested that activation of TEs provides more regulatory sequences for the genome to work on (Jordan et al., 2003; Chenais et al., 2012).

1.6.2 TEs in Plant Adaptation

The rich repertoire of TEs in various genomes combined with their activity upon stress conditions is considered to be the largest contributors to genome diversity, which enables environmental plasticity of the host organisms (Lisch, 2009). A study conducted in the "Evolution Canyon" in Israel revealed significant positive correlations between the number of *BARE*-1 retrotransposon in wild barley and the dryness of the microsites where the samples were collected (Kalendar et al., 2000). They suggested that maintenance of high copy number of complete *BARE*-1 in drier environments was a result of response to the microclimatic changes, which contributed to the increased genome sizes of wild barleys grown there. During the "Green revolution" in the 1940s to late 1960s, one of the major research achievements was the adaptation of maize to diverse geographical regions with differing day lengths and light intensity. Manipulation of flowering time facilitated the adaptation of maize and a recent study identified a locus that seemed to control the vegetative to reproductive transition of maize with microcolineality analyses revealing that a MITE insertion is likely the cause of flowering time changes in the maize variety, Gaspe Flint, one of the earliest flowering varieties (Salvi et al., 2007).

1.6.3 TEs in Plant Domestication

TEs have also been suggested to be involved in the domestication of several plants, such as maize (Bhattacharyya et al., 1990; Studer et al., 2011). The cultivated maize is proposed to be domesticated from the wild progenitor, teosinte, about 10,000 years ago (Matsuoka et al., 2002). One of the largest phenotypic changes associated with the domestication process, from multi-branched architecture to single-stalked, was suggested to result from the ectopic expression of teosinte branched1 (*tb1*) (Doebley, 1992). Recent work showed that a TE (*Hopscotch*) insertion in the *tb1* regulatory region results in decreased expression of this gene

and increased apical dominance, leading to the repressed branching in maize (Studer et al., 2011).

1.7 Outline for This Dissertation

My dissertation focused on the studies of *Mutator*-like transposable elements (MULEs) in plants, specifically in maize and rice. As mentioned before, the amount of MULEs in maize and rice is not proportional to the size of their genomes (maize with a genome size of 2,500 Mb and 12,900 MULEs; rice with a genome size of 400 Mb and 30,000 MULEs). I hypothesize this is attributed to different number and coding capacity of coding-MULEs, which have the potential to generate transposase required for transposition. Indeed, detailed analysis suggested that nested TE insertions and accumulation of insertion and deletions in coding-MULEs are negatively correlated with the abundance of MULEs in maize and rice. Subsequently, transposition of a rice MULE (Os3378-Z) was recapitulated in yeast. Os3378-Z transposase is able to catalyze excision and reinsertion, albeit, with low efficiency. Manipulation of the transposase sequence, including N-terminal deletion, substitution of some amino acids, and fusion of the enhanced yellow fluorescent protein, resulted in increased transposition frequency. The enhanced activity is neither correlated with the levels of transposase protein nor its cellular localization. Finally, the role of MULEs in the Illinois Long-Term Selection Experiment maize population (ILTSE) was investigated. The ILTSE has been selecting maize lines with extreme kernel protein and oil content and the Illinois protein maize strains were used in the study. Specifically, MULE insertions co-segregating with kernel protein content were analyzed and compared with those in the B73 maize genome. This revealed that co-segregating insertions were equally abundant in the 5′ and 3′ regions of genes in contrast to the pronounced preference for the 5′ regions observed for MULEs in the B73 genome. Additionally, expression levels of genes with adjacent co-segregating MULE

insertions were determined in immature kernels and young leaves. Several genes exhibited differential expression in immature kernels but not in young leave, suggesting their potential involvement in artificial selection.

REFERENCES

REFERENCES

Abrusan G, Grundmann N, DeMester L, Makalowski W (2009) TEclass-a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics **25:** 1329-1330

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*. Plant Journal **52:** 342-351

Bancroft I, Dean C (1993) Transposition pattern of the maize element *Ds* in *Arabidopsis thaliana*. Genetics **134:** 1221-1229

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*. Proc. Natl. Acad. Sci. USA **88:** 3502-3506

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genetics **5:** e1000732 doi:10.1371/journal.pgen.1000732

Bennetzen JL, Swanson J, Taylor WC, Freeling M (1984) DNA insertion in the 1st intron of maize *Adh1* affects message levels: Cloning of progenitor and mutant *Adh1* alleles. Proc. Natl. Acad. Sci. USA **81:** 4125-4128

Bhattacharyya MK, Smith AM, Ellis THN, Hedley C, Martin C (1990) The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell **60:** 115-122

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell **17:** 343-360

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*. Molecular Biology and Evolution **20:** 1362-1375

Chen CH, Oishi KK, Kloeckenergruissem B, Freeling M (1987) Organ-specific expression of maize *Adh1* is altered after a *Mu* transposon insertion. Genetics **116:** 469-477

Chen W, VanOpdorp N, Fitzl D, Tewari J, Friedemann P, Greene T, Thompson S, Kumpatla S, Zheng PZ (2012) Transposon insertion in a cinnamyl alcohol dehydrogenase gene is responsible for a *brown midrib1* mutation in maize. Plant Molecular Biology **80:** 289-297

Chenais B, Caruso A, Hiard S, Casse N (2012) The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. Gene **509:** 7-15

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

Cresse AD, Hulbert SH, Brown WE, Lucas JR, Bennetzen JL (1995) *Mu1*-related transposable elements of maize preferentially insert into low copy number DNA. Genetics **140:** 315-324

Cui XQ, Hsia AP, Liu F, Ashlock DA, Wise RP, Schnable PS (2003) Alternative transcription initiation sites and polyadenylation sites are recruited during *Mu* suppression at the *rf2a* locus of maize. Genetics **163:** 685-698

Cui XQ, Wise RP, Schnable PS (1996) The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize. Science **272:** 1334-1336

Das L, Martienssen R (1995) Site-selected transposon mutagenesis at the *hcf106* locus in maize. Plant Cell **7:** 287-294

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Research **12:** 1075-1079

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

Doebley J (1992) Mapping the genes that made maize. Trends in Genetics **8:** 302-307

Dooner HK, Belachew A (1989) Transposition pattern of the maize element *Ac* from the *bz-m2*(*Ac*) allele. Genetics **122:** 447-457

Eisen JA, Benito MI, Walbot V (1994) Sequence similarity of putative transposases links the maize *Mutator* autonomous element and a group of bacterial insertion sequences. Nucleic Acids Research **22:** 2634-2636

Ferguson A, Zhao D, Jiang N (2013) Selective acquisition and retention of genomic sequences by Pack-MULEs based on GC content and breadth of expression. Plant Physiology **163:** 1419-1432

Fernandes J, Dong QF, Schneider B, Morrow DJ, Nan GL, Brendel V, Walbot V (2004) Genome-wide mutagenesis of *Zea mays* L. using *RescueMu* transposons. Genome Biology **5:** R82 doi:10.1186/gb-2004-5-10-r82

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics **9:** 397-405

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: Where genetics meets genomics. Nature Reviews Genetics **3:** 329-341

Fleenor D, Spell M, Robertson D, Wessler S (1990) Nucleotide sequence of the maize *Mutator* element, *Mu8*. Nucleic Acids Research **18:** 6725-6725

Freeling M (1984) Plant transposable elements and insertion sequences. Annual Review of Plant Physiology and Plant Molecular Biology **35:** 277-298

Gao DY (2012) Identification of an active *Mutator*-like element (MULE) in rice (*Oryza sativa*). Molecular Genetics and Genomics **287:** 261-271

Girard L, Freeling M (2000) *Mutator*-suppressible alleles of *rough sheath1* and *liguleless3* in maize reveal multiple mechanisms for suppression. Genetics **154:** 437-446

Guo YB, Levin HL (2010) High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in *Schizosaccharomyces pombe*. Genome Research **20:** 239-248

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

Hanley S, Edwards D, Stevenson D, Haines S, Hegarty M, Schuch W, Edwards KJ (2000) Identification of transposon-tagged genes by the random sequencing of *Mutator*-tagged DNA fragments from *Zea mays*. Plant Journal **23:** 557-566

Haun WJ, Danilevskaya ON, Meeley RB, Springer NM (2009) Disruption of imprinting by *Mutator* transposon insertions in the 5' proximal regions of the *Zea mays Mez1* locus. Genetics **181:** 1229-1237

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biology **5:** 225 doi:10.1186/gb-2004-5-6-225

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

Hennig S, Ziebuhr W (2010) Characterization of the transposase encoded by IS*256*, the prototype of a major family of bacterial insertion sequence elements. Journal of Bacteriology **192:** 4153-4163

Hershberger RJ, Warren CA, Walbot V (1991) *Mutator* activity in maize correlates with the presence and expression of the *Mu* transposable element *Mu9*. Proc. Natl. Acad. Sci. USA **88:** 10198-10202

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

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Research **19:** 1419-1428

Hood ME (2005) Repetitive DNA in the automictic fungus *Microbotryum violaceum*. Genetica **124:** 1-10

Hua-Van A, Capy P (2008) Analysis of the DDE motif in the *Mutator* superfamily. Journal of Molecular Evolution **67:** 670-681

Huang XH, Lu GJ, Zhao Q, Liu XH, Han B (2008) Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. Plant Physiology **148:** 25-40

Hudson ME, Lisch DR, Quail PH (2003) The *FHY3* and *FAR1* genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant Journal **34:** 453-471

Iida S, Morita Y, Choi JD, Park KI, Hoshino A (2004) Genetics and epigenetics in flower pigmentation associated with transposable elements in morning glories. Advances in Biophysics **38:** 141-159

Ishiguro S, Ogasawara K, Fujino K, Sato Y, Kishima Y (2014) Low temperature-responsive changes in the anther transcriptome's repeat sequencesare indicative of stress sensitivity and pollen sterility in rice strains. Plant Physiology **164:** 671-682

Izsvak Z, Frohlich J, Grabundzija I, Shirley JR, Powell HM, Chapman KM, Ivics Z, Hamra FK (2010) Generating knockout rats by transposon mutagenesis in spermatogonial stem cells. Nature Methods **7:** 443-445

Izsvak Z, Khare D, Behlke J, Heinemann U, Plasterk RH, Ivics Z (2002) Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in *Sleeping Beauty* transposition. Journal of Biological Chemistry **277:** 34581-34588

Jiang N, Bao ZR, Zhang XY, 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

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

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. USA **97:** 6603-6607

Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. USA **98:** 8714-8719

Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. PLoS Biology **3:** e181 doi:10.1371/journal.pbio.0030181

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

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY, et al. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815

Keng VW, Yae K, Hayakawa T, Mizuno S, Uno Y, Yusa K, Kokubu C, Kinoshita T, Akagi K, Jenkins NA, et al. (2005) Region-specific saturation germline mutagenesis in mice using the *Sleeping Beauty* transposon system. Nature Methods **2:** 763-769

Kolkman JM, Conrad LJ, Farmer PR, Hardeman K, Ahern KR, Lewis PE, Sawers RJH, Lebejko S, Chomet P, Brutnell TP (2005) Distribution of *Activator* (*Ac*) throughout the maize genome for use in regional mutagenesis. Genetics **169:** 981-995

Kubo N, Fujimoto M, Arimura S, Hirai M, Tsutsumi N (2008) Transfer of rice mitochondrial ribosomal protein L6 gene to the nucleus: Acquisition of the 5'-untranslated region via a transposable element. BMC Evolutionary Biology **8:** 314 doi:10.1186/1471-2148-8-314

Lazarow K, Du ML, Weimer R, Kunze R (2012) A hyperactive transposase of the maize transposable element *Activator* (*Ac*). Genetics **191:** 747-756

Li YB, Harris L, Dooner HK (2013) *TED*, an autonomous and rare maize transposon of the *Mutator* superfamily with a high gametophytic excision frequency. Plant Cell **25:** 3251-3265

Lisch D (2002) *Mutator* transposons. Trends in Plant Science **7:** 498-504

Lisch D (2009) Epigenetic regulation of transposable elements in plants. Annual Review of Plant Biology **60:** 43-66

Lisch D, Chomet P, Freeling M (1995) Genetic characterization of the *Mutator* system in maize: Behavior and regulation of *Mu* transposons in a minimal line. Genetics **139:** 1777-1796

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

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

Liu SZ, Yeh CT, Ji TM, Ying K, Wu HY, 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 doi:10.1371/journal.pgen.1000733

Lockton S, Gaut BS (2009) The contribution of transposable elements to expressed coding sequence in *Arabidopsis thaliana*. Journal of Molecular Evolution **68:** 80-89

Lowe B, Mathern J, Hake S (1992) Active *Mutator* elements suppress the knotted phenotype and increase recombination at the *Kn1-O* tandem duplication. Genetics **132:** 813-822

Ma JX, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Research **14:** 860-869

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

Martienssen R, Barkan A, Taylor WC, Freeling M (1990) Somatically heritable switches in the DNA modification of *Mu* transposable elements monitored with a suppressible mutant in maize. Genes & Development **4:** 331-343

Martienssen RA, Barkan A, Freeling M, Taylor WC (1989) Molecular cloning of a maize gene involved in photosynthetic membrane organization that is regulated by Robertson's *Mutator*. EMBO Journal **8:** 1633-1639

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. USA **99:** 6080-6084

Maumus F, Allen AE, Mhiri C, Hu HH, Jabbari K, Vardi A, Grandbastien MA, Bowler C (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics **10:** 624 doi:10.1186/1471-2164-10-624

McCarty DR, Suzuki M, Hunter C, Collins J, Avigne WT, Koch KE (2013) Genetic and molecular analyses of *UniformMu* transposon insertion lines. Plant Transposable Elements: Methods and Protocols **1057:** 157-166

Mcclintock B (1950) The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA **36:** 344-355

McGinnis W, Shermoen AW, Beckendorf SK (1983) A transposable element inserted just 5' to a Drosophila glue protein gene alters gene expression and chromatin structure. Cell **34:** 75-84

Munoz-Lopez M, Garcia-Perez JL (2010) DNA transposons: Nature and applications in genomics. Current Genomics **11:** 115-128

Nagy Z, Szabo M, Chandler M, Olasz F (2004) Analysis of the N-terminal DNA binding domain of the IS*30* transposase. Molecular Microbiology **54:** 478-488

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature **461:** 1130-1134

Nesmelova IV, Hackett PB (2010) DDE transposases: Structural similarity and diversity. Advanced Drug Delivery Reviews **62:** 1187-1195

Neuveglise C, Chalvet F, Wincker P, Gaillardin C, Casaregola S (2005) *Mutator*-like element in the yeast *Yarrowia lipolytica* displays multiple alternative splicings. Eukaryotic Cell **4:** 615-624

Oki N, Yano K, Okumoto Y, Tsukiyama T, Teraishi M, Tanisaka T (2008) A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. *japonica*. Genes & Genetic Systems **83:** 321-329

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

Pereira V, Enard D, Eyre-Walker A (2009) The effect of transposable element insertions on gene expression evolution in rodents. PLoS ONE **4:** e4321 doi:10.1371/journal.pone.0004321

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

Poulter RTM, Goodwin TJD, Butler MI (2003) Vertebrate helentrons and other novel *Helitrons*. Gene **313:** 201-212

Qin M, Robertson DS, Ellingboe AH (1991) Cloning of the *Mutator* transposable element *MuA2*, a putative regulator of somatic mutability of the *a1-Mum2* allele in maize. Genetics **129:** 845-854

Raizada MN, Nan GL, Walbot V (2001) Somatic and germinal mobility of the *RescueMu* transposon in transgenic maize. Plant Cell **13:** 1587-1608

Raizada MN, Walbot V (2000) The late developmental pattern of *Mu* transposon excision is conferred by a cauliflower mosaic virus 35S-driven MURA cDNA in transgenic maize. Plant Cell **12:** 5-21

Robbins ML, Sekhon RS, Meeley R, Chopra S (2008) A *Mutator* transposon insertion is associated with ectopic expression of a tandemly repeated multicopy *Myb* gene *pericarp color1* of maize. Genetics **178:** 1859-1874

Robertson DS (1978) Characterization of a mutator system in maize. Mutation Research **51:** 21-28

Rossi M, Araujo PG, de Jesus EM, Varani AM, Van Sluys MA (2004) Comparative analysis of *Mutator*-like transposases in sugarcane. Molecular Genetics and Genomics **272:** 194-203

Sakai H, Tanaka T, Itoh T (2007) Birth and death of genes promoted by transposable elements in *Oryza sativa*. Gene **392:** 59-63

Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E, et al. (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus m maize. Proc. Natl. Acad. Sci. USA **104:** 11376-11381

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, MelakeBerhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-768

Schmidt T (1999) LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Molecular Biology **40:** 903-910

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

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science **326:** 1112-1115

Schumacher S, Clubb RT, Cai ML, Mizuuchi K, Clore GM, Gronenborn AM (1997) Solution structure of the *Mu* end DNA-binding Iβ subdomain of phage *Mu* transposase: Modular DNA recognition by two tethered domains. EMBO Journal **16:** 7532-7541

Settles AM, Holding DR, Tan BC, Latshaw SP, Liu J, Suzuki M, Li L, O'Brien BA, Fajardo DS, Wroclawska E, et al. (2007) Sequence-indexed mutations in maize using the *UniformMu* transposon-tagging population. BMC Genomics **8:** 116 doi:10.1186/1471-2164-8-116

Singer T, Yordan C, Martienssen RA (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

Sorek R, Ast G, Graur D (2002) *Alu*-containing exons are alternatively spliced. Genome Research **12:** 1060-1067

Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nature Genetics **43:** 1160-1163

Talbert LE, Patterson GI, Chandler VL (1989) *Mu* transposable elements are structurally diverse and distributed throughout the genus *Zea*. Journal of Molecular Evolution **29:** 28-39

Tan BC, Chen ZL, Shen Y, Zhang YF, Lai JS, Sun SSM (2011) Identification of an active new *Mutator* transposable element in maize. G3: Genes Genomes Genetics **1:** 293-302

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011) Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome Biology and Evolution **3:** 219-229

Thornburg BG, Gotea V, Makalowski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene **365:** 104-110

Tower J, Kurapati R (1994) Preferential transposition of a Drosophila *P* element to the corresponding region of the homologous chromosome. Molecular & General Genetics **244:** 484-490

Wang H, Xu Z, Yu HJ (2008) LTR retrotransposons reveal recent extensive inter-subspecies nonreciprocal recombination in Asian cultivated rice. BMC Genomics **9:** 205 doi:10.1186/1471-2164-9-565

Wang QH, 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

Watson MA, Chaconas G (1996) Three-site synapsis during *Mu* DNA transposition: A critical intermediate preceding engagement of the active site. Cell **85:** 435-445

Wessler SR (1996) Plant retrotransposons: Turned on by stress. Current Biology **6:** 959-961

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

Woodhouse MR, Freeling M, Lisch D (2006) The *mop1* (*mediator of paramutation1*) mutant progressively reactivates one of the two genes encoded by the *MuDR* transposon in maize. Genetics **172:** 579-592

Xu ZN, Yan XH, Maurais S, Fu HH, O'Brien DG, Mottinger J, Dooner HK (2004) *Jittery*, a *Mutator* distant relative with a paradoxical mobile behavior: Excision without reinsertion. Plant Cell **16:** 1105-1114

Yang LX, Bennetzen JL (2009) Structure-based discovery and description of plant and animal *Helitrons*. Proc. Natl. Acad. Sci. USA **106:** 12832-12837

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

Yuan YW, Wessler SR (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc. Natl. Acad. Sci. USA **108:** 7884-7889

Zabala G, Vodkin L (2007) Novel exon combinations generated by alternative splicing of gene fragments mobilized by a CACTA transposon in *Glycine max*. BMC Plant Biology **7:** 38 doi:10.1186/1471-2229-7-38

Zhou LQ, Mitra R, Atkinson PW, Hickman AB, Dyda F, Craig NL (2004) Transposition of *hAT* elements links transposable elements and V(D)J recombination. Nature **432:** 995-1001

CHAPTER 2

Nested Insertions and Accumulation of Indels are Negatively Correlated with Abundance of *Mutator*-Like Transposable Elements in Maize and Rice

(Zhao D, Jiang N (2014) Nested insertions and accumulation of indels are negatively correlated with the abundance of *Mutator*-like transposable elements in maize and rice. PLoS ONE 9(1): e87069. doi:10.1371/journal.pone.0087069)

2.1 Abstract

Mutator-like transposable elements (MULEs) are widespread in plants and were first discovered in maize where there are a total of 12,900 MULEs. In comparison, rice, with a much smaller genome, harbors over 30,000 MULEs. Since maize and rice are close relatives, the differential amplification of MULEs raised an inquiry into the underlying mechanism. A hypothesis was tested that the differential abundance may be partly attributed to the differential copy number of autonomous MULEs with the potential to generate the transposase that is required for transposition. To this end, the two genomes were analyzed, which resulted in detection of 530 and 476 MULEs containing transposase sequences (candidate coding-MULEs) in maize and rice, respectively. Over 1/3 of the candidate coding-MULEs harbor nested insertions and the ratios are similar in the two genomes. Among the maize elements with nested insertions, 24% have insertions in coding regions and over half of them harbor two or more insertions. In contrast, only 12% of the rice elements harbor insertions in coding regions and 19% have multiple insertions, suggesting that nested insertions in maize are more disruptive. This is because most nested insertions in maize are from LTR retrotransposons, which are large in size and are prevalent in the maize genome. The results suggest that the amplification of retrotransposons may limit the amplification of DNA transposons but not vice versa. In addition, more indels are detected among maize elements than rice elements whereas defects caused by point mutations are comparable between the two species. Taken together, more disruptive nested insertions combined with higher frequency of indels resulted in few (6%) coding-MULEs that may encode functional transposases in maize. In contrast, 35% of the coding-MULEs in rice retain putative intact transposase. This is in addition to the higher expression frequency of rice coding-MULEs, which may explain the higher occurrence of MULEs in rice than that in maize.

2.2 Introduction

Transposable elements (TEs) are genomic sequences that are capable of moving from one position to another. Based on the transposition intermediate, TEs can be divided into two classes. Class I, also called RNA or retrotransposons, transpose via an RNA intermediate. Class II, DNA transposons, transpose via a DNA intermediate. TEs can also be grouped into autonomous or non-autonomous elements, where the former encode proteins (transposase for class II elements) responsible for the transposition of themselves as well as their corresponding non-autonomous counterparts.

TEs constitute large fractions of most plant genomes sequenced to date, *i.e.*, ~85% of maize (Schnable et al., 2009), ~62% of soybean (Schmutz et al., 2010) and sorghum (Paterson et al., 2009), ~63% of tomato (Sato et al., 2012), ~43% of papaya (Ming et al., 2008), and ~35% of rice (Matsumoto et al., 2005) genomes. Comparative analysis reveals that the abundance of different classes of TEs (*e.g.*, RNA and DNA TEs) varies dramatically in different plants. In rice, the genomic coverage of RNA TEs (20%) is 1.5-fold of that of DNA TEs (13%) whereas in maize the difference is 8-fold between RNA (76%) and DNA (9%) TEs (Matsumoto et al., 2005; Schnable et al., 2009). This difference is the greatest in the papaya genome, which contains ~43% RNA TEs and very few DNA TEs (0.2%). In general, the amount of retrotransposons is correlated with plant genome size whereas such correlation is not found for DNA elements (Slotkin et al., 2012).

Mutator-like transposable elements (MULEs), first discovered in maize (Robertson, 1978; Bennetzen et al., 1984), are widespread in plants (Yu et al., 2000; Rossi et al., 2004), fungi (Chalvet et al., 2003; Neuveglise et al., 2005), and animals (Marquez and Pritham, 2010). MULEs are one of the most complex TE families, with dramatic variation in structure, sequence, size, and abundance within a genome and among different plant genomes (Yu et al.,

2000; Schnable et al., 2009; Ferguson and Jiang, 2012). Based on the similarity of their terminal inverted repeats (TIRs), MULEs can be grouped into TIR-MULEs or non-TIR-MULEs, where TIR-MULEs are characterized by long TIRs (100-500 bp) with high sequence similarity and non-TIR-MULEs have relatively short TIRs with low sequence similarity between their terminal sequences (Yu et al., 2000). Within the TIR-MULE group, a special subgroup was discovered in several plants, which contains tandem TIRs (two consecutive TIRs) flanking the internal sequence (Ferguson and Jiang, 2012). In addition, some MULEs harbor gene and/or gene fragments, which are referred to as Pack-MULEs (Jiang et al., 2004). Their high transposition frequency combined with the capability to duplicate gene fragments suggest that MULEs play important roles in genome evolution (Lisch, 2002, 2005). Individual genomes can contain all forms of MULEs, including TIR- and tandem TIR-MULEs, non-TIR-MULEs, and Pack-MULEs, albeit with differential abundance (Yu et al., 2000; Ferguson and Jiang, 2012). Rice by far contains the largest amount of known MULEs (n=32,000, 5.5% of the genome) and Pack-MULEs (n=2,924) (Ferguson and Jiang, 2012; Ferguson et al., 2013). As a close relative, maize contains 12,900 MULEs (1% of the genome) and 276 Pack-MULEs, which are relatively less abundant given that its genome size is over 5 times larger than rice (2,066 Mb vs. 370 Mb) (Matsumoto et al., 2005; Schnable et al., 2009). The papaya genome represents an extreme case, which is almost void of DNA TEs including MULEs. Despite the prevalence and importance of MULEs, the mechanism underlying the differential amplification of these elements or DNA elements in general remains largely unknown.

The first autonomous *Mutator* element discovered is *MuDR* in maize. It encodes two proteins, MURA, the major protein responsible for its transposition, and MURB, a helper protein found only in the *Zea* genus and for which the function is still unclear (Lisch et al.,

1999; Lisch, 2002; Kim and Walbot, 2003). MULE transposases belong to the DDE transposase family, which commonly consists of a helix-turn-helix (HTH) DNA-binding domain at the amino terminus and a DDE catalytic domain at the carboxyl terminus (Hua-Van and Capy, 2008; Yuan and Wessler, 2011). MURA-like transposase sequences have been discovered in many other organisms, including plants, fungi, and animals (Lisch et al., 2001; Singer et al., 2001; Xu et al., 2004). In addition to the differences in TIR length, MURA-like transposase sequences are also divergent as demonstrated by distinct subfamilies (Lisch et al., 2001; Rossi et al., 2004). Furthermore, some MULE transposase sequences, such as *FAR1* and *MUSTANG*, have been domesticated as cellular genes and are no longer associated with any mobility (Hudson et al., 2003; Cowan et al., 2005).

The abundance of TEs is a result of the interplay between the amplification through transposition, duplication, horizontal transfer and loss via excision, sequence erosion, deletion *etc*. (Pritham, 2009). Since autonomous elements are responsible for the transposition of both themselves and their corresponding non-autonomous counterparts, their abundance and activity influence the speed of amplification, and therefore contribute to the abundance of TEs in a genome. Maize and rice belong to the same family (Poaceae) with a common ancestor occurring 50-70 million years ago (Wolfe et al., 1989). As the most important crops worldwide, the genome of both maize and rice were sequenced through a hierarchical method using bacterial artificial chromosome clones (BACs) accompanied by high-density genetic maps (Matsumoto et al., 2005; Schnable et al., 2009). The availability of high quality genomic sequence and well-annotated TEs allow a comparative study of TEs in these two organisms. In this study, candidate coding-MULEs (MULEs containing transposase sequences) were detected and analyzed in the maize and rice genomes. In addition, possible factors involved in the loss of coding capacity of those elements were dissected, which

facilitates the understanding about the underlying mechanisms for differential abundance of MULEs in the two genomes.

2.3 Results

2.3.1 Candidate MULEs Containing Transposase Sequences (Coding-MULEs) in Maize and Rice

In this study, TIR-MULEs (see Section 2.2) containing MURA-related transposase sequences (candidate coding-MULEs) were detected and analyzed in the maize and rice genomes. Non-TIR MULEs (see Section 2.2) were not considered because of the difficulty in defining the boundary of their termini and target site duplication (TSD; generated upon integration of a TE into a genomic locus) with high confidence. To retrieve candidate coding-MULEs, a collection of MURA-related transposase sequences from several organisms, especially plants (Table 2.5), was used to search against the maize and rice genomic sequences (NCBI TBLASTN, $e<10^{-5}$). The following criteria (see Section 2.5 for more details) were employed for defining a candidate coding-MULE. First, the element should contain a pair of TIRs and the distance between the two TIRs should be 2 to 30 kb (including nested TE insertions inside the element). Second, the sequence located within the TIRs should have homology (NCBI TBLASTN, $e<10^{-5}$) to known MULE transposase. Third, a TSD should be immediately flanking the TIRs of a single element. With these criteria, a total of 530 and 476 candidate coding-MULEs were detected in maize and rice, respectively. Due to the presence of nested TE insertions in maize and rice (SanMiguel et al., 1996; Jiang and Wessler, 2001; Kronmiller and Wise, 2008, 2009), the candidate coding-MULEs were first masked using TE libraries and nested insertions were identified based on the output of RepeatMasker. After removing nested TE insertions, most candidate coding-MULEs (>95%) are between 2 to 10 kb in size in both maize and rice. However, the number of elements

within different size ranges exhibits different patterns in the two species. In maize, a considerable portion of candidate coding-MULEs ranges from 2 to 3.5 kb (~20%) while only 9% of the rice elements fall into this category (χ^2=25.7360, p<0.0001; Figure 2.1). The minimum size of known autonomous MULEs in plants is around 3.5 kb (Table 2.5), suggesting that elements smaller than 3.5 kb are unlikely to encode an intact transposase (transposase containing both DNA-binding and catalytic domains and no other coding defects). Most (71%) candidate coding-MULEs in rice are between 3.5 to 8 kb while only 58% of the elements in maize are within this range. As a result, candidate coding-MULEs with relatively small (2-3.5 kb) and large size (>8 kb) are more prevalent in maize (42%) than that in rice (29%) (χ^2=20.3473, p<0.0001).

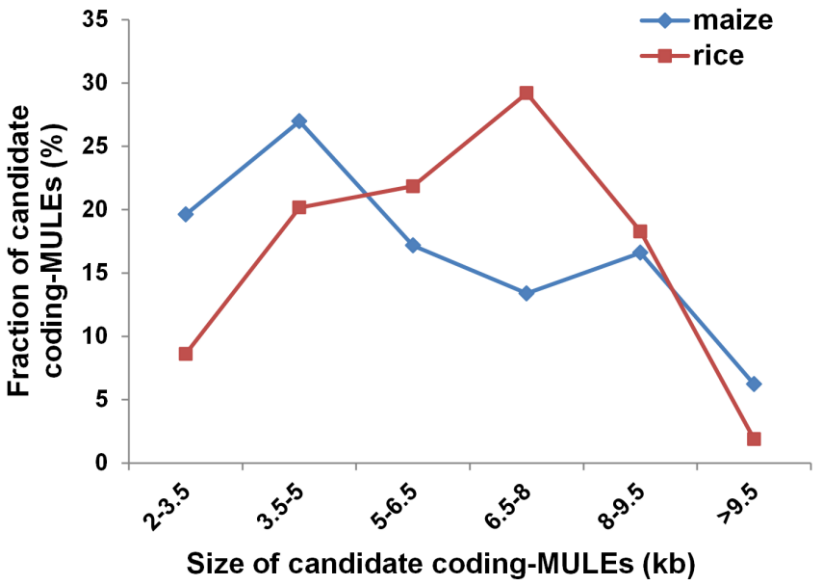

Figure 2.1 Fraction of candidate coding-MULEs within different size ranges after removing nested TE insertions.

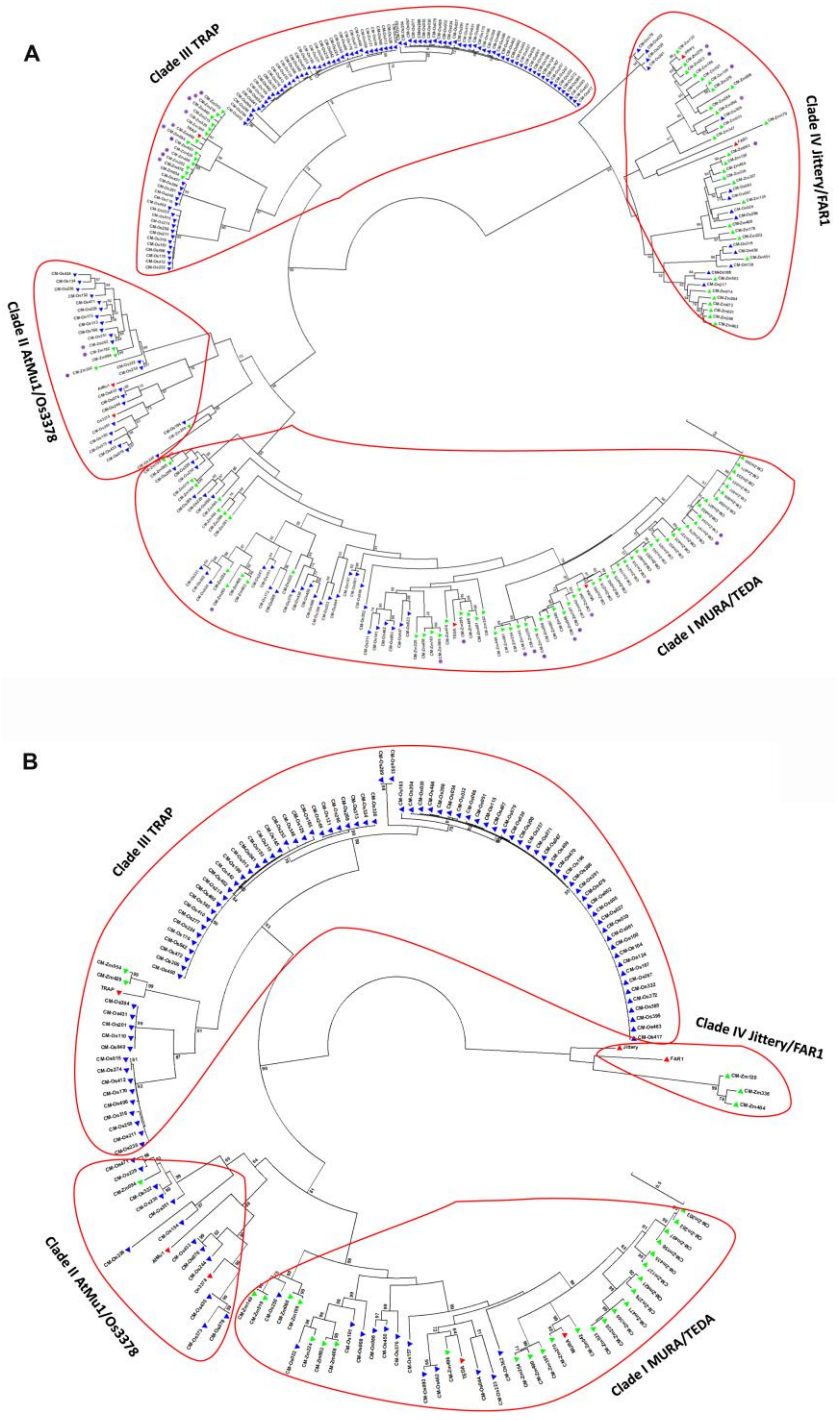

Figure 2.2 Phylogenetic analyses of the candidate coding-MULEs with a DDE domain.

(A) All representative candidate coding-MULEs containing a DDE domain.
(B) Candidate coding-MULEs with putative intact transposases.
Candidate coding-MULEs in maize are denoted by green triangles; those in rice are denoted by blue triangles; and known MULE transposases are denoted by red triangles, maize coding-MULEs with nested LTR insertions are denoted with purple round dots (A).

To determine the phylogenetic relationship of these candidates, protein sequences containing the catalytic (DDE) domain within the elements were used to construct a phylogenetic tree. The phylogenetic analysis revealed that all elements belong to four distinct phylogenetic clades (Figure 2.2A). Clade I is represented by MURA, the first known MULE transposase (Robertson, 1978); and TEDA, a recently discovered MULE in maize (Li et al., 2013). Clade II is represented by *AtMu1*, an active element in *Arabidopsis* (Singer et al., 2001) and *Os3378*, an active MULE element in rice (Gao, 2012). The representative element for clade III is *TRAP*, a MULE in maize (Comelli et al., 1999). The remainder of elements comprises clade IV, which is represented by *Jittery* and *FAR1*. *Jittery* is another active MULE from maize (Xu et al., 2004) and *FAR1* is a gene domesticated from a MULE transposase in *Arabidopsis* (Hudson et al., 2003). As shown in Figure 2.2A, clade I and IV contain more maize elements (green triangles) while rice elements (blue triangles) are more expanded in clade II and III.

2.3.2 Nested Insertions within the Candidate Coding-MULEs

As mentioned above, nested insertions of some TE families are very common in maize and rice (SanMiguel et al., 1996; Jiang and Wessler, 2001; Kronmiller and Wise, 2008, 2009). If a TE inserts into a coding-MULE, it may interrupt the transposase open reading frame (ORF) and abolish its function. To test whether this is the case, all TEs located within the candidate coding-MULEs were examined. The numbers of candidate coding-MULEs containing nested insertions are largely comparable between maize (n=191, 36%) and rice (n=195, 41%) (χ^2=2.5761, p=0.1085). However, among the 195 rice elements with nested insertions, 107 are highly similar to each other, where 40 contain an intact *Os0548* (a *Tourist*-like miniature inverted repeat transposable element (MITE) of 274 bp in length) and 67 contain a partial *Os0548* at the same site. As a result, the 107 elements are likely copies

derived from a single insertion event where 40 copies are amplifications of the element with the intact *Os0548* and meanwhile one copy may have experienced partial deletion of *Os0548* followed by proliferation to 67 copies. Apparently, the insertion of this element did not abolish the capability for further transposition. In contrast, the 191 elements in maize all harbor independent insertions. If the 107 elements containing *Os0548* were excluded from rice, the fraction of candidate coding-MULEs with nested insertions is much lower in rice ([195-107] / [476-107] * 100 ≈ 24%) than that in maize (36%) (χ^2=15.1021, p<0.0001). Moreover, the number and pattern of inserted TEs in individual candidate coding-MULE is different between maize and rice. Among the elements with nested insertions, most (~81%) contain only a single TE insertion in rice whereas that in maize is only less than 47% (χ^2=49.6349, p<0.0001; Figure 2.3A). The fact that candidate coding-MULEs in maize contain more nested TEs than that in rice is also obvious when comparing the average number of nested insertions per coding-MULE with nested TEs (2 nested insertions in maize vs. 1 in rice).

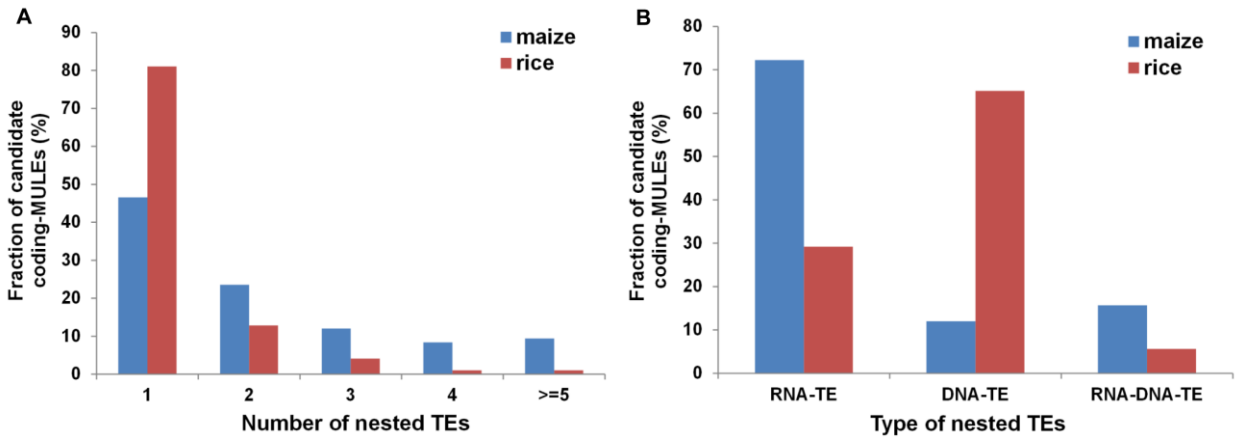

Figure 2.3 Fractions of candidate coding-MULEs containing different numbers (A) and types (B) of nested TE insertions.

Based on the class (see Section 2.2) of the inserted TEs, coding-MULEs with nested insertions were divided into three categories, *i.e.*, coding-MULEs with RNA-TE insertions (all inserted TEs were RNA TEs), coding-MULEs with DNA-TE insertions (all inserted TEs were DNA TEs), and those with RNA-DNA-TE insertions (with insertions from both RNA and DNA TEs). It turned out that over 72% of the candidate coding-MULEs in maize contained RNA-TE insertions compared with 29% of that in rice (χ^2=71.4396, p<0.0001; Figure 2.3B). In contrast, there are more candidate coding-MULEs contain DNA-TE insertions in rice (65%) than that in maize (12%) (χ^2=112.5264, p<0.0001). Furthermore, the average size of the inserted TEs in maize is about 3-fold of that in rice for DNA-TE (708 bp vs. 245 bp, p=0.0016, *t*-test) and 1.3-fold for RNA-TE (6818 bp vs. 5081 bp, p=0.0006, *t*-test) (Table 2.1). Accordingly, maize candidate coding-MULEs contain more independent and larger TE insertions, with the majority from LTR retrotransposons. Interestingly, most maize MULEs grouping with *TRAP* (clade III) (10 out of 14 vs. 26% of genome average, elements with purple dots in Figure 2.2A) are associated with nested insertions of LTR retrotransposons while they are not particularly older or distribute differently than other MULEs. This suggests there might be some structural features of these elements that attract LTR elements.

Table 2.1 Number and average length of nested TE insertions based on TE classes in the candidate coding-MULEs.

	RNA-TE*	DNA-TE*	Total
Maize	329 (6818 a†)	74 (708 a)	403
Rice	99 (5081 b)	152 (245 b)	251

* Numbers in parenthesis represent the average length (bp) of nested TEs.
† Means in parenthesis in each column followed by different letters are significantly different (p<0.005, *t*-test).

To date the insertion time of the nested TEs, all the intact LTR retrotransposons within the candidate coding-MULEs were analyzed, which are characterized with long terminal repeat (LTR) at both ends of the element. The LTRs of one element are identical when inserted into a new location and become divergent over time. As a result, the identity between the LTRs has been used to estimate the age of the insertion (SanMiguel et al., 1998). In light of this, the approximate age was calculated for 133 and 25 intact LTR elements within the candidate coding-MULEs in maize and rice, respectively (Figure 2.4). The largest fractions were estimated to be inserted within 1 million years in both maize (~65%) and rice (~69%), followed by insertions which have occurred 1 to 2 million years ago. Few LTR insertions were older than 3 million years in both genomes. Therefore, the insertion of LTR elements into the candidate coding-MULEs occurred within the same evolutionary time frame in both genomes despite the fewer insertions observed in rice.

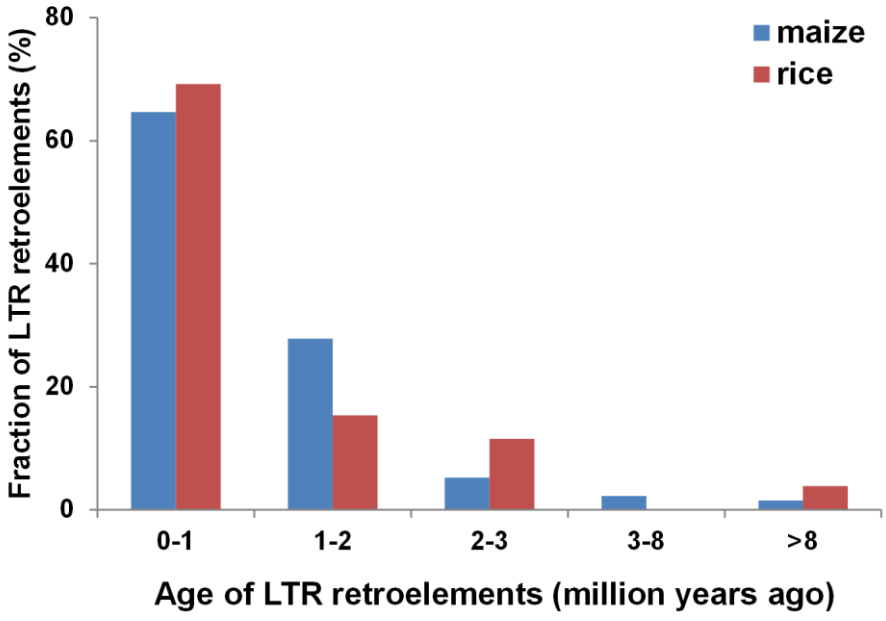

Figure 2.4 Ages of intact LTR retrotransposons in the candidate coding-MULEs in maize and rice.

As a comparison, the occurrence of coding-MULEs inserted into other TEs was also examined. It turned out that only 9% of the maize coding-MULEs and 7% of the rice elements inserted into other TEs (Table 2.2). This is in contrast to the fact that over 1/3 of the elements harbor insertions from other TEs, suggesting the coding-MULEs are more likely serving as targets for other elements rather than targeting other TEs. In addition, more maize candidate coding-MULEs (7%) inserted into RNA TEs than the rice elements (3%) (Table 2.2), which is consistent with the fact that there are many more RNA TEs in the maize genome than that in rice (Matsumoto et al., 2005; Schnable et al., 2009).

Table 2.2 Insertion preference of candidate coding-MULEs (CMs).

	Maize		Rice	
Types of TEs	RNA-TE (% of total CMs*)	DNA-TE (% of total CMs*)	RNA-TE (% of total CMs*)	DNA-TE (% of total CMs*)
Nested TE insertion events in coding-MULEs	329 (32%)	74 (10%)	99 (19%†)	46 (9%†)
Coding-MULE insertion events in other TEs	35 (7%)	9 (2%)	15 (3%)	19 (4%)

*Note that individual coding-MULE may contain different number or types of nested TEs.
† In rice, 107 candidate coding-MULEs are copies of two elements resulting from one TE insertion event (see results), which was corrected in both the number of DNA-TE insertion events and the total candidate coding-MULEs (476-107+1=370).

2.3.3 Coding Capacity of the Candidate Coding-MULEs

Previous studies indicate that class II transposases, including MULE transposases, consist of an N-terminal helix-turn-helix (HTH) DNA-binding domain and a C-terminal DDE catalytic domain (Benito and Walbot, 1997; Babu et al., 2006; Hua-Van and Capy, 2008; Yuan and Wessler, 2011). The DNA binding domain is responsible for binding to the transposon DNA especially the TIR sequences and the catalytic domain is responsible for excision and integration of the element (Hennig and Ziebuhr, 2010; Nesmelova and Hackett, 2010). To determine whether an individual candidate coding-MULE contains both domains,

the coding regions of these elements were extracted from the relevant gene annotation or annotated manually. Subsequently, the presence of a DNA-binding domain was determined using the Jpred3 program and that of the catalytic domain by examination of a multiple sequence alignment containing the MURA protein as the reference (see Section 2.5). A transposase containing both DNA-binding and catalytic domains and without other defects (*e.g.,* frameshift, premature stop codon) in its ORF is considered a potentially intact transposase.

To obtain a better view, the candidate coding-MULEs were divided into six groups based on their coding capacity. Group 1 includes coding-MULEs with putative intact transposase, *i.e.*, at least one HTH domain (some elements have more than one HTH domain, *e.g., Mos1* (Nesmelova and Hackett, 2010)) close to the amino terminus and a DDE catalytic domain close to the carboxyl terminus of the transposase protein. Group 2 is comprised of coding-MULEs with mutation of either the translation start codon (ATG) or key amino acids (D, D, E) in the catalytic domain. Group 3 and 4 include coding-MULE transposases containing frameshift and premature stop codon, respectively. Those containing both frameshift and premature stop codon were assigned to Group 5. Group 6 consists of coding-MULEs with various forms of deletions in their transposases, including deletions of the DNA-binding domain and/or the catalytic domain or other regions. Since many candidate coding-MULEs contain several types of coding defects, one element may be assigned to more than one group (Table 2.3A). Meanwhile, a more specific classification where each element is only assigned to one sub-type is available in Table 2.3B.

Table 2.3A Candidate coding-MULEs (CMs) with or without distinct defects in their coding regions (redundant grouping).

Group	ORF status	Maize			Rice		
		With TE (% of total CMs with nested TEs)	Without TE (% of total CMs without nested TEs)	Total	With TE (% of total CMs with nested TEs)	Without TE (% of total CMs without nested TEs)	Total
Group 1	ORF with putative intact transposase	2 (1.05%)	29 (8.55%)	31 (5.85%)	70 (35.90%)	98 (34.88%)	168 (35.29%)
Group 2	ORF with DDE or start codon mutation	19 (9.95%)	20 (5.90%)	39 (7.36%)	13 (6.67%)	14 (4.98%)	27 (5.67%)
Group 3	ORF with frameshift	36 (18.85%)	37 (10.91%)	73 (13.77%)	34 (17.44%)	45 (16.01%)	79 (16.60%)
Group 4	ORF with premature stop codon	82 (42.93%)	91 (26.84%)	173 (32.64%)	64 (32.82%)	95 (33.81%)	159 (33.40%)
Group 5	ORF with both frameshift and premature stop codon	19 (9.95%)	12 (3.54%)	31 (5.85%)	18 (9.23%)	33 (11.74%)	51 (10.71%)
Group 6	ORF with deletions	187 (97.91%)	288 (84.96%)	475 (89.62%)	66 (32.85%)	129 (45.91%)	195 (40.97%)

Table 2.3B Candidate coding-MULEs (CMs) with or without distinct defects in their coding regions (non-redundant grouping).

Group	ORF status		Maize			Rice		
			With TE (% of total CMs with nested TEs)	Without TE (% of total CMs without nested TEs)	Total	With TE (% of total CMs with nested TEs)	Without TE (% of total CMs without nested TEs)	Total
Group 1	ORF with putative intact transposase		2 (1.05%)	29 (8.55%)	31 (5.85%)	70 (35.90%)	98 (34.88%)	168 (35.29%)
Group 2	ORF with DDE or start codon mutation		1 (0.52%)	4 (1.18%)	5 (0.94%)	3 (1.54%)	7 (2.49%)	10 (2.10%)
Group 3	ORF with frameshift		1 (0.52%)	9 (2.65%)	10 (1.89%)	10 (5.13%)	4 (1.42%)	14 (2.94%)
Group 4	ORF with premature stop codon		0 (0%)	8 (2.36%)	8 (1.51%)	37 (18.97%)	28 (9.96%)	65 (13.66%)
Group 5	ORF with both frameshift and premature stop codon		0 (0%)	1 (0.29%)	1 (0.19%)	9 (4.62%)	15 (5.34%)	24 (5.04%)
Group 6	ORF with deletions	deletion only	81 (42.41%)	185 (54.57%)	266 (50.19%)	37 (18.97%)	64 (22.78%)	101 (21.22%)
		with DDE or start codon mutation only	8 (4.19%)	5 (1.47%)	13 (2.45%)	5 (2.56%)	5 (1.78%)	10 (2.10%)
		with frameshift	16 (8.38%)	16 (4.72%)	32 (6.04%)	6 (3.08%)	8 (2.85%)	14 (2.94%)
		with premature stop codon	63 (32.98%)	71 (20.94%)	134 (25.28%)	9 (4.62%)	34 (12.10%)	43 (9.03%)
		with both frameshift and premature stop codon	19 (9.95%)	11 (3.24%)	30 (5.66%)	9 (4.62%)	18 (6.41%)	27 (5.67%)
Total			191	339	530	195	281	476

Overall, rice contains many more candidate coding-MULEs with putative intact transposase than maize (35.29% vs. 5.85%) (χ^2=137.0185, p<0.0001; Table 2.3A), and deletion is likely the most important factor contributing to the difference. Most of the elements in maize (>89%) are associated with various deletions within the transposase compared to only 41% of the elements with deletions in rice. The presence of premature stop codon is the second most frequent defect, but the fractions are similar between maize (32.64%) and rice (33.40%) (χ^2=0.0658, p=0.7975; Table 2.3A). For elements with other defects (mutation of DDE motif or start codon, frameshift, and presence of both frameshift and premature stop codon), no dramatic difference was observed between the two species (Table 2.3A). In addition, many more maize elements (~40%) contain more than one type of defect than that in rice (< 25%) (χ^2=25.1084, p<0.0001).

When classifying these elements based on whether they contain nested TE insertions or not, it is clear that in maize there are more candidate coding-MULEs containing putative intact transposase if there are no nested TEs (Figure 2.5). For example, only 2 out of 191 (1.05%) elements with nested insertions seem to have putative intact transposase, which is in contrast to 29 out of 339 (8.55%) elements without nested insertions in maize (χ^2=12.5035, p=0.0004; Table 2.3A). In contrast, ratios of candidate coding-MULEs harboring putative intact transposases for those with (35.90%) and without (34.88%) nested TE insertions (χ^2=0.0526, p=0.8185; Table 2.3A) are comparable in rice. Nevertheless, if the high copy elements (those containing intact and partial *Os0548*) were excluded, only 12.5% candidate coding-MULEs with nested insertions harbor putative intact transposase in rice, suggesting the destructive effect of nested TE insertions in both species, or elements with nested insertions are older than these without nested insertions so more mutations have accumulated. A close examination indicated that more nested TEs (n=46) in maize directly disrupted

transposases than that in rice (n=23) (χ^2=23.0278, p<0.0001), and most are LTR retrotransposons. Again, this indicates that nested TEs are more deleterious to coding-MULEs in maize than that in rice.

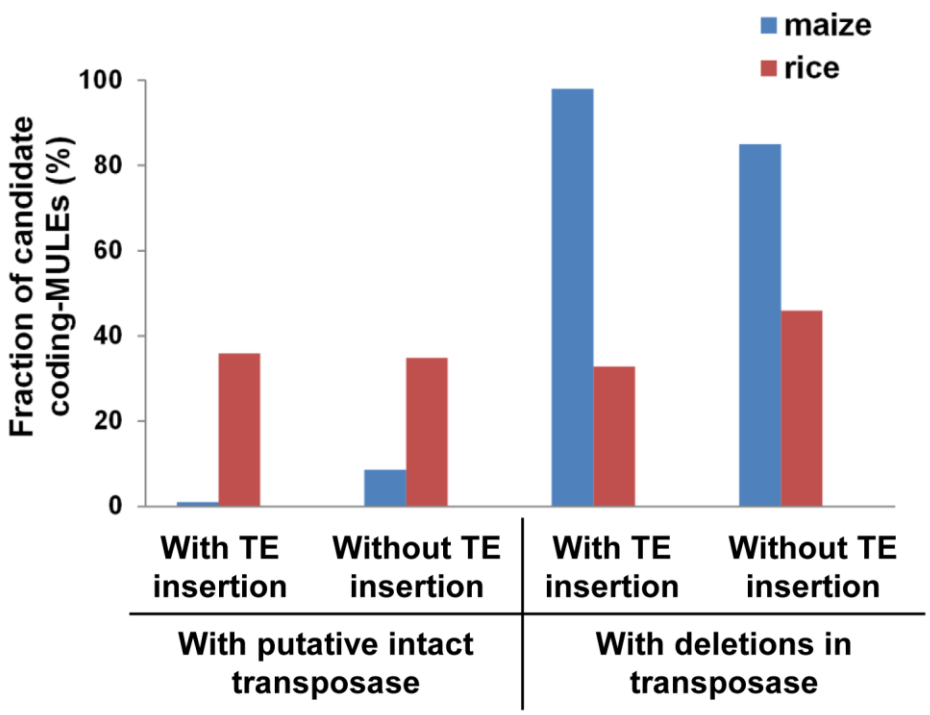

Figure 2.5 Fractions of candidate coding-MULEs containing putative intact transposases and transposases with deletions.

The fractions were calculated within each category, *e.g.*, fraction of candidate coding-MULEs with TE insertions containing putative intact transposases is based on the total number of elements with TE insertions.

A second phylogenetic tree was constructed based on the catalytic domain of the elements with putative intact transposase (Figure 2.2B). Compared with Figure 2.2A, the number of elements in all the four clades declines, which is not surprising. This decrease is more obvious in maize, as revealed by the existence of only one clade (clade I) containing more than 10 elements and the other clades having four or less elements left. In contrast, rice (Nipponbare) still has more than 10 coding-MULEs in three clades (clade I, II, and III) and

clade III seems to be active or recently active as evidenced by the short branch lengths (Figure 2.2B). In contrast, transposition activity in clades II to IV might be limited currently or in the future in maize (B73). This is consistent with the fact that there are either few maize elements (clades II and III) or many maize elements bear relatively long branches (clade IV) in these clades (Figure 2.2A), which are the signatures for loss of transposition activity.

2.3.4 Other Insertions and Deletions within the Candidate Coding-MULEs

The analysis above indicated that deletions in coding regions are prevalent in maize coding-MULEs, and this is possibly the most dominant factor for loss of coding capacity of these elements. The abundance of elements with deletions in maize could be attributed to a high deletion frequency in maize. Alternatively, the element with deletions may have an advantage in transposition (for instance, because of small size), so they have achieved relatively high copy numbers. To determine whether the frequency of insertion/deletion within the candidate coding-MULEs is similar between maize and rice, pairwise comparison was conducted using individual coding-MULEs and their most similar homologs. For accuracy, only pairs bearing higher than 95% identity in both maize and rice were considered. This portion of coding-MULE pairs were analyzed and the number and length of indels in those pairs were calculated and normalized to the average number of indels per kb (NIK) and the average length of indels per kb (LIK) (see Section 2.5). The results show that the number of indels is significantly higher ($p<0.05$, *t*-test) in maize than that in rice in all the identity ranges except 98-99% (Figure 2.6A). The number of indels is increasing when the coding-MULE pair is less similar, with a ~4.5 fold increase when the similarity is 95-96% compared to that of over 99% in both maize and rice. Meanwhile, the most significant difference between maize and rice is observed with the group containing element pairs with 95-96% identity. For this group, the NIK in maize is ~2.48 and that in rice is ~1.40, suggesting that

there is at least one more indel per kb per coding-MULE pair generated in maize than those in rice. When comparing the LIK of coding-MULE pairs between maize and rice, no significant difference was observed and the average LIK values are 87.36 bp ± 16.06 for maize and 89.24 bp ± 11.21 for rice (p=0.9235, *t*-test). To ensure the nucleotide identity reflected the divergence of homologous coding-MULE pairs, indel rates were calculated based on synonymous substitution rates (Ks) between these pairs of elements in coding regions, and a similar trend was observed (Figure 2.6B). Taken together, it seems that coding-MULEs in maize experienced more indels than that in rice.

To determine whether the higher frequency of indels in maize is specific to coding-MULEs or a generic feature for the entire genome, indel frequency in LTR sequences of individual *Copia*-like LTR elements was analyzed in the two genomes. As was observed for coding-MULEs, maize LTR sequences seem to have experienced more frequent indels than that in rice. However, the difference is only evident when the nucleotide identity is within 96-97% (p=0.0227, *t*-test). It should be noted that indel frequencies between LTRs and coding-MULEs in rice are not significantly different within the same identity range. In contrast, the average indel numbers in the maize coding-MULEs are about two-fold of those in LTRs when the sequence identity ranges from 95% to 98% (Figure 2.6A). Collectively, these results demonstrate that maize may be more prone to incidence of indels than rice at the whole genome level. More importantly, within the maize genome, coding-MULEs seem to have experienced more indels than LTRs if we assume point mutation rate is comparable across different families of TEs.

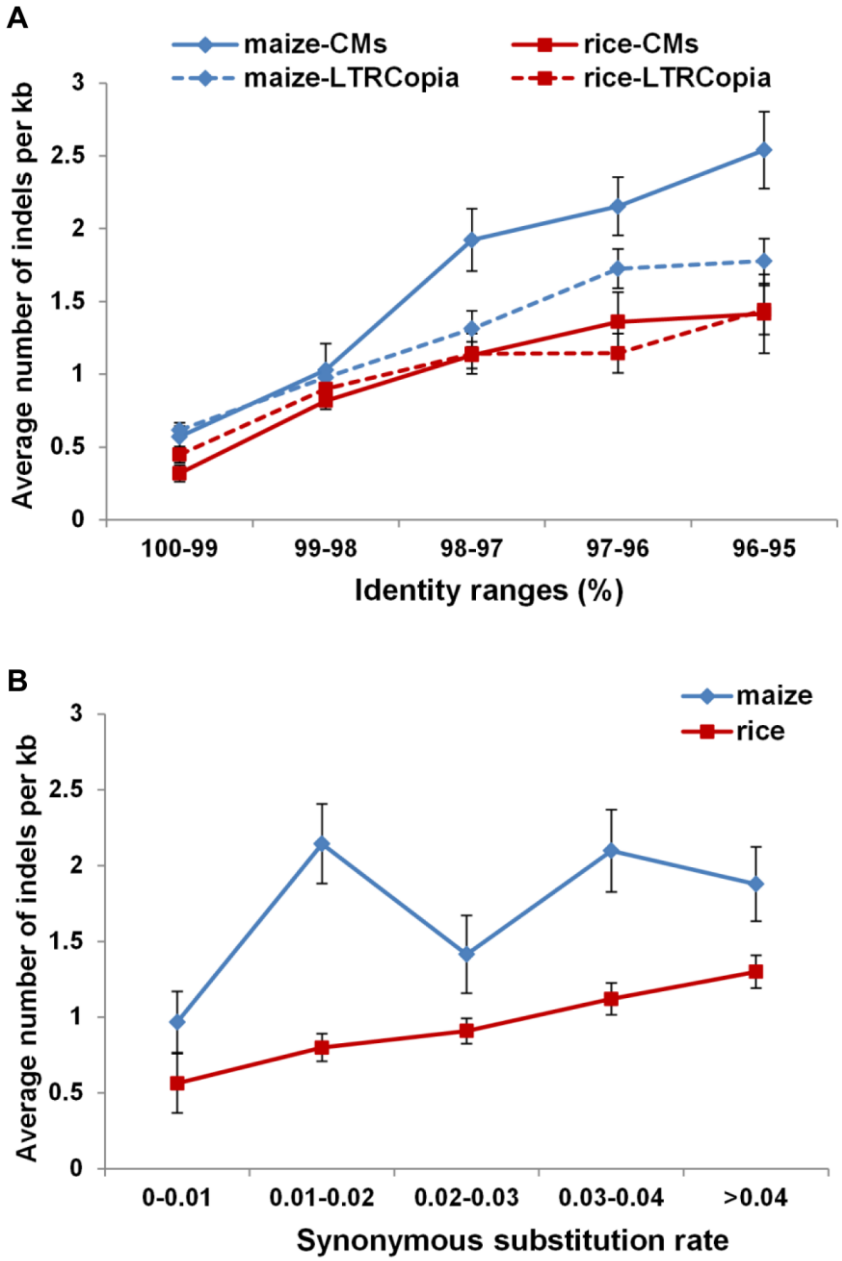

Figure 2.6 Average number of indels in maize and rice TEs.

(A) Average number of indels in candidate coding-MULEs (maize-CMs and rice-CMs) and *Copia*-like LTR elements using pairwise nucleotide identity as the grouping criterion.
(B) Average number of indels in candidate coding-MULEs using synonymous substitution rate as the grouping criterion.

2.3.5 Expression Evidence of the Candidate Coding-MULEs in Maize and Rice

A functional transposase is only available for transposition if the element is expressed. To determine how many candidate coding-MULEs are expressed, the publically available

EST and full-length cDNA (fl-cDNA) databases were searched for evidence of expression. The completeness of these datasets was roughly assessed by determining the proportion of non-TE genes that have at least one match in the database with 99.5% or higher identity over the entire matched sequence. The results revealed that 36% of non-TE genes in maize and 45% of that in rice have expression evidence, suggesting that the dataset for rice is 1.2-fold more comprehensive than that of maize. Data from next generation sequencing, such as RNA-seq, were not examined because most candidate coding-MULEs have highly similar copies in the genome and the short reads of RNA-seq experiments do not allow us to determine which specific element is expressed.

Table 2.4A Candidate coding-MULEs with expression evidence in maize and rice.

Species	Coding-MULE	With nested TE	EST/fl-cDNA*	ORF status
Maize	CM-Zm373	Yes	BT069140	deletion only
	CM-Zm391	Yes	gi|211046304|gb|FK977962.1|FK977962	deletion only
	CM-Zm138	Yes	gi|211501343|gb|FL022227.1|FL022227	deletion only
	CM-Zm216	Yes	gi|211515319|gb|FL470580.1|FL470580	deletion only
	CM-Zm352	Yes	gi|211161841|gb|FK974923.1|FK974923	deletion & premature stop codon & frameshift
	CM-Zm418	No	BT041283	deletion only
	CM-Zm091	No	BT070085	deletion only
	CM-Zm369	No	gi|211378400|gb|FL229430.1|FL229430	deletion only
	CM-Zm137	No	gi|211249173|gb|FL479070.1|FL479070	putative intact
Rice	CM-Os180	Yes	AK066496	deletion only
	CM-Os087	Yes	gi|87004716|gb|CI306800.1|CI306800	deletion only
	CM-Os392	Yes	gi|88668382|gb|CI721944.1|CI721944	deletion only
	CM-Os222	Yes	gi|88692462|gb|CI736037.1|CI736037	deletion only
	CM-Os464	Yes	gi|88846512|gb|CI373280.1|CI373280	deletion only
	CM-Os133	Yes	AK072808	deletion & premature stop codon & frameshift
	CM-Os028	Yes	AK288956	deletion & premature stop codon & frameshift
	CM-Os046	Yes	AK120202	frameshift & premature stop codon
	CM-Os181	Yes	AK067920	premature stop codon
	CM-Os252	Yes	gi|88718219|gb|CI741729.1|CI741729	putative intact
	CM-Os388	Yes	gi|88695069|gb|CI739107.1|CI739107	putative intact
	CM-Os364	Yes	gi|88695796|gb|CI737373.1|CI737373	putative intact
	CM-Os446	No	AK073736	deletion only
	CM-Os006	No	gi|58687855|gb|CK076542.1|CK076542	deletion only
	CM-Os083	No	gi|86437298|gb|CI119020.1|CI119020	deletion only
	CM-Os089	No	gi|86827995|gb|CI274907.1|CI274907	deletion only
	CM-Os422	No	gi|87030711|gb|CI361925.1|CI361925	deletion only
	CM-Os209	No	gi|88276743|gb|CI397545.1|CI397545	deletion only
	CM-Os317	No	gi|88727053|gb|CI750708.1|CI750708	deletion only
	CM-Os217	No	AK066465	deletion & frameshift
	CM-Os370	No	gi|88846132|gb|CI372900.1|CI372900	deletion & frameshift
	CM-Os350	No	gi|88297974|gb|CI555989.1|CI555989	frameshift
	CM-Os302	No	gi|29629482|gb|CB634491.1|CB634491	frameshift & premature stop codon
	CM-Os018	No	gi|88730292|gb|CI753906.1|CI753906	premature stop codon
	CM-Os395	No	gi|88279043|gb|CI400043.1|CI400043	putative intact
	CM-Os200	No	gi|5701669|gb|C28952.2|C28952	putative intact
	CM-Os044	No	AK100632	putative intact
	CM-Os352	No	gi|29632126|gb|CB637135.1|CB637135	putative intact

*EST/fl-cDNA: expressed sequence tag/full length cDNA

Table 2.4B Summary of candidate coding-MULEs with expression evidence in maize and rice.

	Coding-MULEs with putative intact transposase			Coding-MULEs with defects in transposase		
	With expression evidence	Without expression evidence	Total	With expression evidence	Without expression evidence	Total
Maize	1 (3.13%)	31 (96.88%)	32	8 (1.61%)	490 (98.39%)	498
Rice	7 (3.95%)	170 (96.05%)	177	21 (7.02%)	278 (92.98%)	299

An element was considered to be expressed if it matches an fl-cDNA or EST sequence with at least 99.5% identity over the entire matched sequence (at least 300 bp). The results revealed that there were more coding-MULEs with either EST or fl-cDNA evidence in rice (n=28) than that in maize (n=9) (χ^2=12.3933, p=0.0004; Table 2.4A). Even if the 1.2-fold enrichment of expression evidence in rice than maize were taken into account, rice still contains more elements (n=23) (χ^2=7.9969, p=0.0047) with expression evidence. In both genomes, elements with expression evidence include coding-MULEs with and without nested insertions. In rice, expression evidence was detected for 4% (n=7) of elements with putative intact transposase and 7% (n=21) of elements with defects in transposases whereas those in maize are only 3% (n=1) and 1.6% (n=8), respectively (Table 2.4B). In either case, it does not appear that elements with putative intact transposases are more frequently transcribed. Interestingly, among the 7 elements containing putative intact transposases that are expressed in rice, three have a MITE insertion (among the 107 elements containing the *Os0548*) in the C-terminal region of the relevant coding-MULEs (~2.5 kb downstream of the stop codon of the transposase ORF), which does not seem to affect the expression of the elements, at least at the transcription level (Table 2.4A). Overall, only a small subset of candidate coding-MULEs are expressed, yet rice has many more expressed elements than maize does, which may explain why the rice genome harbor many more MULEs. However, we must be cautious about interpreting this difference because it is known that expression of TEs is often induced

under stress and the higher number of coding-MULEs expressed in rice may be due to the overrepresentation of stress conditions in the rice dataset.

2.4 Discussion

2.4.1 Differential Amplification of MULEs Including Coding-MULEs in Maize and Rice

The first active MULE element, *Mutator*, was found in maize due to the high mutation frequency it caused (Robertson, 1978). With the availability of a myriad of genome sequences, analysis of TEs has been carried out at the whole genome level, resulting in the discovery of the prevalence of MULEs in most living organisms (Yu et al., 2000; Lisch et al., 2001; Chalvet et al., 2003; Rossi et al., 2004; Neuveglise et al., 2005; Marquez and Pritham, 2010). Being members of the grass family, both maize and rice contain thousands of MULEs (Schnable et al., 2009; Ferguson and Jiang, 2012). However, compared with rice (~370 Mb available sequence), maize (~2.1 Gb available sequence) contains less MULEs per unit genomic sequence (~84 per Mb in rice vs. ~8 per Mb in maize) as well as per unit coding sequence of non-TE genes (~770/Mb in rice vs. ~318/Mb in maize, based on maize genome annotation v2 and MSU rice annotation version 7). The persistence of TEs is an interaction between amplification through transposition, duplication of genomic sequences, and loss of TEs by excision and sequence erosion. Transposase, the protein encoded by autonomous elements, catalyzes transpositions, and is therefore responsible for increasing the copy number of the elements. In this study, comparable numbers (530 vs. 476) of candidate coding-MULEs were detected in maize and rice. Since all the candidate elements contain partial or complete transposases, they should have been derived from putative autonomous elements at a certain evolutionary stage. This implies that the resource for generating transposases was comparable between maize and rice in the traceable past, and the loss of

such resource has been accelerated in maize in the recent past, which led to the reduced abundance of MULEs.

Maize and rice share an ancestor about 50-70 million years ago (Wolfe et al., 1989) and should have inherited largely the same set of transposable elements including MULEs. From this point of view, it is surprising that the phylogenetic composition of coding-MULEs is different between maize and rice, in addition to the difference in the total copy numbers. The maize elements are abundant in clade I and IV, while the rice elements are amplified in clade II and III (Figure 2.2A). This suggests that either the amplification of TEs is a rather fortuitous process or different genome environments may favor the survival of distinct elements.

2.4.2 Factors Involved in Degeneration of Coding-MULEs

In this study, various factors that may lead to the loss of coding capacity of MULE transposases were dissected. Apparently, for both maize and rice, deletion is the most devastating factor (Table 2.3A), which causes loss of the coding sequence or frameshift. Deletion of sequence is much more prevalent in maize than that in rice (90% vs. 40% of the coding-MULEs in maize and rice, respectively). Point mutation is the second most important factor, which resulted in the mutation of the DDE motif, the loss of the start codon, and the formation of premature stop codon. Nevertheless, the frequency of this type of mutation (~40%; Table 2.3A; coding-MULEs containing ORFs with DDE or start codon mutations, and those with premature stop codons) is comparable between maize and rice and therefore does not explain the differential amplification of MULEs between the two species. In addition, it is likely that point mutations within regions other than the DDE motif and the start codon may also lead to dysfunctional transposase. As a result, the number of functional transposases might be overestimated in this study. The third important factor is the insertion

by other TEs, which may interrupt the ORF of the transposases or interrupt the *cis*-elements that are required for further transposition. Again, the number of elements that harbor nested insertions is comparable between maize and rice. However, the nested insertions in maize are more harmful than that in rice for a variety of reasons (see below). Thus, among the three major factors that demolish the coding capacity of MULEs, two are significantly enforced in maize.

Based on pairwise comparison of homologous elements, indels occur more frequently in maize than in rice. Certainly, this is based on the assumption that the point mutation rate is comparable in the two species. Given the fact that the number of elements with defects caused by point mutation is similar in maize and rice (see above), this is likely the case. In addition, our analysis using synonymous substitution rate led to similar results, further suggesting that functional constraint on nucleotide substitution is similar in the two species. Our analyses confirmed previous studies of the low stability of the maize genome (Vitte and Bennetzen, 2006). By comparing an orthologous region of maize, sorghum, and rice, Ilic and colleagues found that the maize genome experienced more sequence deletions than rice, leading to the conclusion that the maize genome is less stable compared with the rice genome (Ilic et al., 2003). The instability was partly attributed to the polyploidization event of maize. Several studies suggest that genome rearrangement was more frequent in species that experienced a polyploid event (reviewed in Wendel, 2000). Unequal homologous recombination and illegitimate recombination are the two most proposed mechanisms responsible for DNA removal (Bennetzen et al., 2005). It was also suggested that genes were preferentially deleted from one of the two maize homeologs possibly through a similar illegitimate recombination, which is also the primary source for TE removal in the maize genome (Woodhouse et al., 2010).

The comparison of indel frequency between *Copia*-like LTR elements and coding-MULEs shed new light on this issue. If we assume the frequency of point mutation is comparable across the two genomes, it is obvious that both types of elements demonstrate elevated indel frequency in maize, suggesting that in general indels occur more often in maize than that in rice. However, what is somehow unexpected is that coding-MULEs in maize seem to be subject to higher indel frequency than that of LTR elements and therefore contributes to their degeneration. Such difference could be attributed to the fact that they are located in different regions of the genome and the indel rate is influenced by the recombination rate of the regions because it is known that TEs in different chromosomal domains evolve differently (Tian et al., 2009). Alternatively, indels occur during transposition, so it could be due to the different transposition mechanism of different families of TEs as well. No matter what the underlying mechanism is, the difference in indel frequency explains, at least to some degree, why LTR elements are more successful than DNA TEs in maize.

In addition to the lower abundance of elements with putative intact transposase, the number of expressed candidate coding-MULEs in maize is only 1/3 of that in rice. The fewer expressed elements in maize is consistent with the low activity of MULEs in the genome, which is in contrast with the recent burst of amplification of LTR retrotransposons. It is well established that enormous variation of TE activities and compositions exist among different organisms (reviewed by Feschotte and Pritham, 2007; Huang et al., 2012). Some plant species, such as moss (*Physcomitrella patens*) and maize (*Zea mays*), exhibited relatively high LTR retrotransposon activity and low DNA TE activity while some animal species, *e.g.*, nematode (*Caenorhabditis elegans*) and brown bat (*Myotis lucifugus*), are more active in DNA TEs and less in LTR elements (Huang et al., 2012). This suggests that genomes may

have distinct mechanisms for silencing different types of elements that led to differential amplification of distinct TE families (Feschotte and Pritham, 2007). In fact, LTR elements, especially *gypsy*-like LTRs, express more frequently than DNA TEs in maize as revealed by more expressed sequence tags (ESTs) mapped to LTR retrotransposons than that to DNA TEs (Vicient, 2010). So far, it is unclear whether the lack of expression may further accelerate the degeneration of coding-MULEs in maize. For normal genes, those that are not expressed evolve more rapidly than genes that are highly expressed (Hastings, 1996; Duret and Mouchiroud, 1999, 2000; Davidson et al., 2012), and there is a correlation between expression level, Ka/Ks, and pseudogenization (Frith et al., 2006; Zou et al., 2009). As a result, it would not be surprising if the low expression frequency of the maize elements contributes to the loss of coding capacity of MULEs.

2.4.3 Interaction between Different TEs

Plant genomes harbor many distinct superfamilies of TEs and it is not known whether the amplification of some TEs impacts the amplification/survival of other TEs. Based on the genome-wide analyses of TEs, the rice genome harbors a total of 166,700 TEs (Jiang and Panaud, 2013) while that for maize is 1,283,000 (Schnable et al., 2009). If this is translated to TE insertions per unit sequence, the density of TEs in maize is 1.4-fold of that in rice (613 TEs/Mb sequence in maize vs. 439 TEs/Mb sequence in rice). From this point of view, the total number of nested TE insertions found in the candidate coding-MULEs (403 in maize vs. 251 in rice) is largely comparable to the genomic average and to each other in rice and maize. Nevertheless, nested TE insertions are more detrimental to the candidate coding-MULEs in maize than that in rice due to the following facts. First, the majority (~88%) of the candidate coding-MULEs with TE insertions in maize contains RNA or retrotransposons (RNA-TE and DNA-RNA-TE), which are usually larger than DNA-TEs and are more disruptive (Figure

2.3B). Second, more candidate coding-MULEs in maize harbor two or more TE insertions (53%) compared with the majority of elements in rice that contain only one TE insertion (81%) (Figure 2.3A). This further increases the chance to abolish the functionality of the coding-MULEs, which is consistent with the fact that in maize, the proportion of elements with their transposases interrupted by nested TEs is twice of that in rice. Collectively, these nested TE insertions impaired a large fraction of coding-MULEs in maize, which is reflected by the presence of only two elements (~1%) with potentially intact transposases among all the coding-MULEs containing nested insertions (Table 2.3A). On the other hand, few candidate coding-MULEs have inserted into other TEs including retrotransposons. This is likely because MULEs preferentially insert into low copy or genic sequences (Liu et al., 2009). Such target specificity confers certain evolutionary advantages. For instance, the elements in genic regions are more likely to be expressed. Nevertheless, it may also bring about vulnerability to these elements when other elements such as retrotransposons are actively transposing. This is because retrotransposons tend to insert into other repetitive sequences so their activity is deleterious to other elements while the amplification of MULEs or other DNA transposons rarely interrupt retrotransposons. This is consistent with the results from a previous study where it was shown that the incidents of insertions of other TEs (including LTR elements) into miniature inverted repeat transposable elements (MITEs) were 65 times more often than that of MITEs into LTR and other DNA elements (Jiang and Wessler, 2001). This indicates that MITEs are similar to coding-MULEs in terms of serving as targets for other elements rather than targeting other TEs.

Both theoretical modeling and empirical results demonstrated that mating systems play potential roles in shaping the abundance and diversity of transposable element within a genome (Wright and Schoen, 1999; Wright and Finnegan, 2001; Lockton and Gaut, 2010;

Boutin et al., 2012). A recent study showed that mode of reproduction contributed to the different transposon profiles in self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata* (Lockton and Gaut, 2010). This and other studies led to the conclusion that a reduced efficacy of natural selection against TE insertions in selfing populations (Wright et al., 2001; Tam et al., 2007; Dolgin et al., 2008). Such reduced pressure against TE insertions provides more advantage for DNA TEs than RNA TEs because the insertion of DNA TEs are more likely to be in genic regions which are subject to more intensive selection. Our comparison between rice (a selfing plant) and maize (an outcrossing plant) provides additional understanding about the possible influence of mating system on the dynamics of TEs. This is because outcrossing offers continuous opportunity for stochastic introduction of novel autonomous elements that may initialize new transposition activity. If a new retrotransposon is introduced, it is conceivable that the increased amplification of retrotransposons gradually abolishes the activity of MULEs or other DNA TEs through insertion into them. In contrast, the introduction of DNA TEs is not as harmful for retrotransposons. Taken together, the genomes of outcrossing plants are less favorable to DNA TEs due to the elevated efficacy of natural selection as well as the increased chances being attacked by retrotransposons. From this point of view, plants experiencing significant outcrossing are less likely to contain abundant DNA TEs and papaya is an excellent example.

2.5 Materials and Methods

2.5.1 Genomic Sequences and TE Libraries

The B73 maize genomic sequence RefGen v2 was downloaded from the MaizesSquence.org (http://www.maizesequence.org/) and the Nipponbare rice genomic sequence Release 7 was downloaded from the Rice Genome Annotation Project at Michigan State University (http://rice.plantbiology.msu.edu/index.shtml). TE library for rice and

MULE TIR libraries (MULE TIR) for both maize and rice were constructed and curated by the Jiang Lab. The TE library for maize was downloaded from the Maize Transposable Element Database (http://maizetedb.org/~maize/) in August 2011.

2.5.2 Identification of Candidate Coding-MULEs

To maximize the possibility of detecting coding-MULEs, all the transpositionally active MURA-related transposases (*e.g.* MURA and those of *Hop*, *Jittery*, *Os3378*, *AtMu1*) (Robertson, 1978; Bennetzen et al., 1984; Singer et al., 2001; Chalvet et al., 2003; Xu et al., 2004; Gao, 2012) and some MURA-related transposases with conceptual translations were collected and used in the NCBI TBLASTN search ($e<10^{-5}$, http//www.ncbi.nlm.nih.vov/blast/) against the maize and rice genomes, respectively. Detailed information about the MULEs used as queries is provided in Table 2.5. Sequences producing significant alignments ($e<10^{-5}$) were retained and their 50 kb flanking sequences were retrieved. The resulting sequences were masked by the MULE TIR libraries of maize and rice, respectively, using the RepeatMasker program (www.repeatmasker.org). A candidate coding-MULE must satisfy the following criteria. First, the pairwise identity between its two TIRs should be higher than 75% and in an inverted orientation with the TIR-ends facing outwards. Second, sequence homologous to transposases (NCBI BLASTX, $e<10^{-5}$) must be located within the two TIRs. Lastly, a TSD (8-11 bp) immediately flanking the TIRs must be present. For TSDs of 8 bp, one mismatch or indel (one nucleotide) is allowed and for those equal or larger than 9 bp, a maximum of 2 mismatches (or one mismatch plus one indel) is allowed.

Table 2.5 MURA and MURA-related transposases used in the study.

Element	GenBank GI /Accession	Species	Size of putative transposase (amino acids)	Element size (bp)	Reference
Jittery	7673677/AAF66982	Zea mays	709	3916	Xu et al., 2004
MuDR	540581/AAA21566	Zea mays	823	4942	Hershberger et al., 1991
TRAP	5690095/CAB51950	Zea mays	863	6393	Comelli et al., 1999
AtMu1	2565011/AAB81881	Arabidopsis thaliana	761	3645	Singer et al., 2001
Os3378	52353379/AAU43947	Oryza sativa	866	4394/4395	Gao, 2012
FAR1	240255849/ NP 567455	Arabidopsis thaliana	827*	4079*	Hudson et al., 2003
Hop	30421204/AAP31248	Fusarium oxysporum	836	3299	Chalvet et al., 2003
Mutyl	50553866/XP 504344	Yarrowia lipolytica	1178	7413	Neuveglise et al., 2005
RMUA	156723167/BAF79582	*Oryza sativa*	707	4374	N/A†
MURA-like	194689672/ACF78920	Zea mays	601	N/A	N/A
MURA-like	223950329/ACN29248	Zea mays	751	N/A	N/A
MURA-like	12322384/AAG51216	Arabidopsis thaliana	826	N/A	N/A
MURA-like	5734742/AAD50007	Arabidopsis thaliana	622	N/A	N/A
MURA-like	7523705/AAF63144	Arabidopsis thaliana	726	N/A	N/A
MURA-like	17380908/AAL36266	Arabidopsis thaliana	749	N/A	N/A
MURA-like	41469647/AAS07370	Oryza sativa	747	N/A	N/A
MURA-like	22094356/AAM91883	Oryza sativa	896	N/A	N/A
MURA-like	15209152/AAK91885	Oryza sativa	959	N/A	N/A
MURA-like	29788811/AAP03357	Oryza sativa	907	N/A	N/A
MURA-like	51477400/AAU04773	Cucumis melo	807	N/A	N/A
Hypothetical protein	242096428/XP 002438704	Sorghum bicolor	720	N/A	N/A
Predicted protein	224122824/XP 002318925	Populus trichocarpa	580	N/A	N/A
Hypothetical protein	225432189/XP 002268620	Vitis vinifera	746	N/A	N/A

* *FAR1* has no identifiable terminal inverted repeats (TIRs), the length of putative transposase refers to the protein sequence of the gene and element size is the gene length.
† MURA-related transpsases with only NCBI depository and no detailed study are denoted as N/A for the element size and reference.

2.5.3 Detection of Nested TE Insertions in the Candidate Coding-MULEs and Estimation of Ages of Nested LTR Retroelements

The candidate coding-MULEs were masked using the maize and rice TE libraries, respectively, using the RepeatMasker program (www.repeatmasker.org) and the resulting output file was used to determine the number and type of nested TEs in the candidate coding-MULEs. The coordinates of the coding regions (as defined in the following section) and the inserted TEs were compared in order to determine whether the inserted TEs interrupt the open reading frame of the candidate coding-MULEs. If the coordinates of the inserted TE are within the coordinates of the coding region, the TE is considered to be inserted in the coding region of the candidate coding-MULE.

Estimation of the insertion time of intact long terminal repeat (LTR) retrotransposons was based on the divergence of the two LTRs. Sequences of the two LTRs of one element were aligned using the MUSCLE program (Edgar, 2004) and the number of substitutions per site between the two LTRs was obtained using MEGA 5.2.2. Nucleotide substitution rate 1.3 x 10^{-8} per site per year was used to calculate the approximate age of the LTR elements (Ma et al., 2004).

2.5.4 Determination of the Coding Capacity of the Candidate Coding-MULEs

To determine the coding region of the candidate coding-MULEs, nested insertions were first masked using the maize and rice TE libraries (without MULE sequences), respectively, using the RepeatMasker program (www.repeatmasker.org). The masked sequences were used to search (NCBI BLASTX with e<10^{-5}) against the total protein datasets of maize and rice, which contain annotated TE proteins and were downloaded from the MaizeSequence.org (http://www.maizesequence.org/) and the Rice Genome Annotation Project

(http://rice.plantbiology.msu.edu/index.shtml), respectively. To ensure that the annotated proteins producing significant alignments (smallest e-value) with the candidate coding-MULEs were indeed MULE transposase, these protein sequences were used to search against the above mentioned MURA and MURA-related transposase sequences using the NCBI BLASTP ($e<10^{-5}$) program. If the annotated proteins had significant alignments ($e<10^{-5}$) with MURA or MURA-related sequences, the coding region was defined according to the annotated protein. For the annotated proteins without any significant alignment with known transposases, it suggested that the transposase sequence was not annotated in an ORF because of some defects (*e.g.*, frameshift, premature stop codon, or large deletions). In this case, the nucleotide sequences of the elements were used to search (NCBI BLASTX with $e<10^{-5}$) against the MURA and MURA-related transposase sequences directly and the regions with significant alignments with those transposase sequences were defined as the coding region. For each candidate coding-MULE, the alignment profiles were manually examined and custom python scripts were used to retrieve the sequences encoding transposases of the candidate coding-MULEs. After obtaining the transposase coding sequences of the candidate coding-MULEs, their coding capacity was evaluated for the presence of DNA binding and catalytic domains. The HTH DNA binding domain was determined using the Jpred3 program (http://www.compbio.dundee.ac.uk/www-jpred/), which was capable of detecting all the experimentally defined HTH domains of some well-annotated transposases (*e.g.*, those of *Hermes*, *Tc3*, *Mos1*, Phage *Mu*) (Nesmelova and Hackett, 2010). To locate the DDE domain, a multiple sequence alignment was conducted using the transposase coding sequences of the candidate coding-MULEs in maize and rice separately, with the MURA protein sequence as a reference. The conserved catalytic domain was established by comparing candidate coding-MULEs with that of MURA and positions of the DDE amino acids in the transposases were recorded. Specifically, if the three amino acids (DDE) are all present, and the length from the

first “D” to the “E” is longer than 100 amino acids, without premature stop codon(s) and frameshift(s), the element was considered to contain a catalytic domain.

2.5.5 Determination of Indels in the Candidate Coding-MULEs

The number and length of indels between two homologous coding-MULEs were determined using the following procedure. First, sequences of the candidate coding-MULEs (after removing nested TEs) were used to conduct an all vs. all search using the NCBI BLASTN program ($e<10^{-10}$). Second, after the self-match was excluded, the element that produced the most significant alignment (the two sequences with the longest alignment) was retained to form a pair with the query coding-MULE, for which pairwise alignment was conducted using the “gap” program available from the GCG package (version 11.0, Accelrys Inc., San Diego, CA). Lastly, the number and length of indels were normalized based on the total length of the two elements to make them comparable between maize and rice. That is, the number of indels was divided by the total length of the two elements, giving a value of the average number of indels per kb (NIK). Similarly, the normalized length of indels is the average length of indels per kb (LIK). Grouping of different coding-MULE pairs was based on two parameters, *i.e.*, pairwise nucleotide identities and synonymous substitution rates (Ks), respectively. To calculate the Ks values, coding sequences (as defined in the previous section) of these element pairs were aligned based on amino acid codons, premature stop codons and frameshifts were removed to achieve correct coding frame. The resulting aligned sequences in fasta format were changed to .axt format using a custom Python script and Ks values were determined using the KaKs Calculator program (Zhang et al., 2006). Comparison of NIK and LIK of coding-MULE pairs with similar identity or Ks between maize and rice was conducted using the SAS9.3 program at the High Performance Computing Centre of

Michigan State University. Nested insertions and one copy of the TSD were excluded upon alignment.

In addition, the indel numbers in the LTR sequences of the *Copia*-like LTR retrotransposon in the two genomes were calculated. The sequences of LTRs in rice and maize were from the rice and maize TE libraries mentioned above. The LTR sequences were used to search against the maize and rice genomes using the RepeatMasker program (www.repeatmasker.org). Custom Python scripts were used to extract the coordinates of intact LTRs and retrieve their sequences. Pairwise sequence identity and determination of average indel numbers were similar to that conducted for coding-MULEs.

2.5.6 Phylogenetic Analysis

Amino acid sequences of the catalytic region (corresponding to 331-489 amino acids of MURA) of the candidate coding-MULEs were used to conduct the phylogenetic analysis in MEGA 5.2.2 (Tamura et al., 2011). Frameshifts were corrected and premature stop codons were excluded to ensure appropriate alignment. Figure 2.2A contains known MULEs and representative elements with catalytic domains in the two genomes. The representatives were chosen as following: if two elements shared 80% identity in 80% of the element region, only one element was retained. In this way, 211 rice MULEs and 195 maize MULEs with catalytic regions were excluded. Figure 2.2B contains known MULEs in Figure 2.2A and all MULEs with putative intact transposases, with the exception of one element family in rice. This element family contains 108 members and 59 of them contain putative intact transposases, and only four representative elements were chosen to be included in Figure 2.2B (clade III) in order to attain a manageable size. The Maximum Likelihood method was used to infer the evolutionary history and the tree with the highest log likelihood value is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5

categories (+G, parameter = 4.0540)). All positions with higher than 95% site coverage were used in the analysis and only bootstrap (500 replicates) values equal or higher than 50% were shown.

2.5.7 Determination of the Expression Status of the Candidate Coding-MULEs

Expression Sequence Tags (ESTs) for maize (516,425 ESTs) and rice (1,253,557 ESTs) were downloaded from the NCBI dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/index.html) on January 4th, 2013. The full-length cDNA (flcDNA) library for maize (27,455) was downloaded from the Maize Full Length cDNA Project (http://www.maizecdna.org/) on February 2nd, 2013 and that for rice (37,139) was from the Knowledge-based Oryza Molecular Biological Encyclopedia (http://cdna01.dna.affrc.go.jp/cDNA/) on Oct 1st, 2008. The EST and flcDNA libraries were concatenated into one file (EST-flcDNA) for both maize and rice, respectively. The candidate coding-MULE sequences were searched against the EST-flcDNA library using the NCBI BLASTN program ($e<10^{-5}$). A coding-MULE is considered to have expression evidence if the identity of matched sequence between the coding-MULE and EST/flcDNA is higher than 99.5% over the entire length of the EST/flcDNA.

2.6 Acknowledgments

We thank Ann Ferguson and Stefan Cerbin (Michigan State University) for critical reading of the manuscript.

REFERENCES

REFERENCES

Babu MM, Iyer LM, Balaji S, Aravind L (2006) The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. Nucleic Acids Research **34:** 6505-6520

Benito MI, Walbot V (1997) Characterization of the maize *Mutator* transposable element *MURA* transposase as a DNA-binding protein. Molecular and Cellular Biology **17:** 5165-5175

Bennetzen JL, Ma JX, Devos K (2005) Mechanisms of recent genome size variation in flowering plants. Annals of Botany **95:** 127-132

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. Proc Natl Acad Sci U S A **81:** 4125-4128

Boutin TS, Le Rouzic A, Capy P (2012) How does selfing affect the dynamics of selfish transposable elements? Mobile DNA **3**

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*. Molecular Biology and Evolution **20:** 1362-1375

Comelli P, Konig J, Werr W (1999) Alternative splicing of two leading exons partitions promoter activity between the coding regions of the maize homeobox gene *Zmhox1a* and *Trap* (transposon-associated protein). Plant Molecular Biology **41:** 615-625

Cowan RK, Hoen DR, Schoen DJ, Bureau TE (2005) *MUSTANG* is a novel family of domesticated transposase genes found in diverse angiosperms. Molecular Biology and Evolution **22:** 2084-2089

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

Dolgin ES, Charlesworth B, Cutter AD (2008) Population frequencies of transposable elements in selfing and outcrossing *Caenorhabditis nematodes*. Genet Res **90:** 317-329

Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, *Arabidopsis*. Proc. Natl. Acad. Sci. USA **96:** 4482-4487

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Molecular Biology and Evolution **17:** 68-74

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

Ferguson A, Zhao D, Jiang N (2013) Selective acquisition and retention of genomic sequences by Pack-MULEs based on GC content and breadth of expression. Plant Physiology **163:** 1419-1432

Ferguson AA, Jiang N (2012) *Mutator*-like elements with multiple long terminal inverted repeats in plants. Comparative and Functional Genomics **2012:** 695827 doi:10.1155/2012/695827

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annual Review of Genetics **41:** 331-368

Frith MC, Wilming LG, Forrest A, Kawaji H, Tan SL, Wahlestedt C, Bajic VB, Kai C, Kawai J, Carninci P, et al. (2006) Pseudo-messenger RNA: Phantoms of the transcriptome. PLoS Genetics **2:** 504-514

Gao DY (2012) Identification of an active *Mutator*-like element (MULE) in rice (*Oryza sativa*). Molecular Genetics and Genomics **287:** 261-271

Hastings KEM (1996) Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. Journal of Molecular Evolution **42:** 631-640

Hennig S, Ziebuhr W (2010) Characterization of the transposase encoded by IS*256*, the prototype of a major family of bacterial insertion sequence elements. Journal of Bacteriology **192:** 4153-4163

Hershberger RJ, Warren CA, Walbot V (1991) *Mutator* activity in maize correlates with the presence and expression of the *Mu* transposable element *Mu9*. Proc. Natl. Acad. Sci. USA **88:** 10198-10202

Hua-Van A, Capy P (2008) Analysis of the DDE motif in the *Mutator* superfamily. Journal of Molecular Evolution **67:** 670-681

Huang CRL, Burns KH, Boeke JD (2012) Active transposition in genomes. Annual Review of Genetics **46:** 651-675

Hudson ME, Lisch DR, Quail PH (2003) The *FHY3* and *FAR1* genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant Journal **34:** 453-471

Ilic K, SanMiguel PJ, Bennetzen JL (2003) A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. Proc. Natl. Acad. Sci. USA **100:** 12265-12270

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

Jiang N, Panaud O (2013) Transposable element dynamics in rice and its wild relatives. *In* Zhang Q,Wing RA, eds, Genetics and Genomics of Rice, Plant Genetics and Genomics: Crops and Models. Springer Science+Business Media New York

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

Kim SH, Walbot V (2003) Deletion derivatives of the *MuDR* regulatory transposon of maize encode antisense transcripts but are not dominant-negative regulators of *Mutator* activities. Plant Cell **15:** 2430-2447

Kronmiller BA, Wise RP (2008) TEnest: automated chronological annotation and visualization of nested plant transposable elements. Plant Physiology **146:** 45-59

Kronmiller BA, Wise RP (2009) Computational finishing of large sequence contigs reveals interspersed nested repeats and gene islands in the *rf1*-associated region of maize. Plant Physiology **151:** 483-495

Li YB, Harris L, Dooner HK (2013) *TED*, an autonomous and rare maize transposon of the *Mutator* superfamily with a high gametophytic excision frequency. Plant Cell **25:** 3251-3265

Lisch D (2002) *Mutator* transposons. Trends in Plant Science **7:** 498-504

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 the *MURA* and *MURB* proteins. Genetics **151:** 331-341

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

Liu SZ, Yeh CT, Ji TM, Ying K, Wu HY, 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 doi:10.1371/journal.pgen.1000733

Lockton S, Gaut BS (2010) The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. BMC Evolutionary Biology **10**

Ma JX, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Research **14:** 860-869

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

Matsumoto T, Wu JZ, Kanamori H, Katayose Y, Fujisawa M, Namiki N, Mizuno H, Yamamoto K, Antonio BA, Baba T, et al. (2005) The map-based sequence of the rice genome. Nature **436:** 793-800

Ming R, Hou SB, Feng Y, Yu QY, 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

Nesmelova IV, Hackett PB (2010) DDE transposases: Structural similarity and diversity. Advanced Drug Delivery Reviews **62:** 1187-1195

Neuveglise C, Chalvet F, Wincker P, Gaillardin C, Casaregola S (2005) *Mutator*-like element in the yeast *Yarrowia lipolytica* displays multiple alternative splicings. Eukaryotic Cell **4:** 615-624

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

Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. Journal of Heredity **100:** 648-655

Robertson DS (1978) Characterization of a mutator system in maize. Mutation Research **51:** 21-28

Rossi M, Araujo PG, de Jesus EM, Varani AM, Van Sluys MA (2004) Comparative analysis of *Mutator*-like transposases in sugarcane. Molecular Genetics and Genomics **272:** 194-203

SanMiguel P, Gaut B, Tikhonov A, Nakajima Y, Bennetzen J (1998) The paleontology of intergene retrotransposons of maize. Nature **20:** 43-45

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, MelakeBerhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-768

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

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

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science **326:** 1112-1115

Singer T, Yordan C, Martienssen RA (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

Slotkin RK, Nuthikattu S, Jiang N (2012) The evolutionary impact of transposable elements on gene and genome regulation. *In* Greilhuber J,Wendel J, eds, Molecular Biology and Evolution of the Plant Genome. Springer Press, pp 35-58

Tam SM, Causse M, Garchery C, Burck H, Mhiri C, Grandbastien MA (2007) The distribution of *Copia*-type retrotransposons and the evolutionary history of tomato and related wild species. Journal of Evolutionary Biology **20:** 1056-1072

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony methods. Molecular Biology and Evolution **28:** 2731-2739

Tian ZX, Rizzon C, Du JC, Zhu LC, Bennetzen JL, Jackson SA, Gaut BS, Ma JX (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Research **19:** 2221-2230

Vicient CM (2010) Transcriptional activity of transposable elements in maize. BMC Genomics **11:** 601 doi:10.1186/1471-2164-11-601

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc. Natl. Acad. Sci. USA **103:** 17638-17643

Wendel JF (2000) Genome evolution in polyploids. Plant Molecular Biology **42:** 225-249

Wolfe KH, Gouy ML, Yang YW, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. USA **86:** 6201-6205

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biology **8:** e1000409 doi:10.1371/journal.pbio.1000409

Wright S, Finnegan D (2001) Genome evolution: Sex and the transposable element. Current Biology **11:** R296-R299

Wright SI, Quang HL, Schoen DJ, Bureau TE (2001) Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. Genetics **158:** 1279-1288

Wright SI, Schoen DJ (1999) Transposon dynamics and the breeding system. Genetica **107:** 139-148

Xu ZN, Yan XH, Maurais S, Fu HH, O'Brien DG, Mottinger J, Dooner HK (2004) *Jittery*, a *Mutator* distant relative with a paradoxical mobile behavior: excision without reinsertion. Plant Cell **16:** 1105-1114

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

Yuan YW, Wessler SR (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc. Natl. Acad. Sci. USA **108:** 7884-7889

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics **4:** 259-263

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH (2009) Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. Plant Physiology **151:** 3-15

CHAPTER 3

Transposition of a Rice *Mutator*-Like Element in the Yeast *Saccharomyces cerevisiae*

3.1 Abstract

Mutator-like transposable elements (MULEs) are wide-spread in plants and are present in animals and fungi. MULEs are well known for their high transposition activity as well as their capability to duplicate and amplify genomic sequences including gene fragments; thus they play important roles in genome evolution. Despite their abundance and importance, few MULEs have been shown to be currently active. Moreover, there is no report about MULE activity in non-native hosts and it has been suggested that host factors are involved in the transposition of *Mutator* elements in maize. In this study, the transposition activity of a rice MULE, *Os3378*, was tested in yeast. The fact that *Os3378* is capable of excising and reinserting into the yeast genome suggests that yeast harbors all the host factors for transposition of MULEs. The transposition activity induced by the wild-type transposase is low; nevertheless, it can be altered through modification of the transposase sequences including deletion, fusion and substitution. N-terminal deletion altered excision frequency, expression level, and cellular localization of the transposase. The finding that fusion of a fluorescent protein to a transposase significantly enhanced the transposition activity represents the first report of enhancement of transposition activity through protein fusion. Furthermore, the establishment of a MULE transposition system in yeast provides the foundation for further studying the transposition mechanism of MULEs as well as how they duplicate and acquire genomic sequences.

3.2 Introduction

Transposable elements (TEs) are genomic sequences that can move from one position to another within a genome. With their repetitive nature, TEs usually constitute large fractions of plant genomes. Based on the intermediate of transposition, TEs are divided into two major classes. Class I TEs or retrotransposons transpose via an RNA intermediate, where the

genomic copy is transcribed into mRNA followed by reverse transcription of mRNA into cDNA and integration into a new genomic location. Class II or DNA TEs transpose via a DNA intermediate, where the TE excises directly from its original position and integrates in another position. Both TE classes consist of autonomous and non-autonomous elements, where the former encode proteins (transposases) that are responsible for the transposition of themselves and their corresponding non-autonomous elements.

Mutator (*Mu*)-like transposable elements (MULEs) are class II TEs, for which the initial discovery in maize was attributed to their high mutagenicity (Robertson, 1978). The founder elements include the autonomous element, *MuDR* and the non-autonomous elements (*Mu1-Mu8*, *Mu10-Mu13*), where the former generates the transposition machinery for movement of all these elements (Bennetzen et al., 1984; Freeling, 1984; Chen et al., 1987; Talbert et al., 1989; Fleenor et al., 1990; Chomet et al., 1991; Hershberger et al., 1991; Qin et al., 1991; Lisch et al., 1995; Dietrich et al., 2002; Tan et al., 2011). All *Mu* elements share high sequence similarity in their terminal inverted repeats (TIRs, ~220 bp) but containing distinct internal sequences. Upon integration into a new position, they usually generate a 9-bp duplication of the target site sequence, which is called target site duplication (TSD).

The founder elements of MULEs, *MuDR*/*Mu*, possess several characteristics that make them excellent candidates for genome-wide mutagenesis studies in maize. First, their insertions occur in both genetically linked and unlinked sites (Lisch et al., 1995), which entails the possibility of genome-wide mutagenesis. Second, they preferentially integrate into low-copy genomic regions, especially 5′ regions of genes (Cresse et al., 1995; Liu et al., 2009). Third, insertions occur in late germinal cells, which can be transmitted to the progeny. Taking advantage of *Mu*-active maize lines, large collections of transposon-insertion lines have been developed by several research groups, providing materials for reverse genetics

studies (Raizada et al., 2001; Fernandes et al., 2004; Settles et al., 2007; McCarty et al., 2013). In addition to *Mu* elements in maize, MULEs have been found in a wide range of organisms, including other plants, fungi, and animals (Yu et al., 2000; Lisch et al., 2001; Singer et al., 2001; Chalvet et al., 2003; Rossi et al., 2004; Neuveglise et al., 2005; Marquez and Pritham, 2010). In organisms where TEs have been examined at the whole genome level, MULEs can reach thousands of copies. The two important food crops, rice and maize contain 30,475 and 12,900 MULEs, respectively (Schnable et al., 2009; Ferguson and Jiang, 2012). The fleshy fruit crop, tomato, contains 28,041 MULEs (Ferguson and Jiang, 2012). Among the MULEs, there occur a special group of elements that capture gene and/or gene fragments, which are known as Pack-MULEs (Jiang et al., 2004). They are best studied in rice, where 2,924 Pack-MULEs were discovered (Ferguson et al., 2013). They are considered to be one of the sources for genome evolution in terms of formation of new genes and regulation of related sequences through the siRNA silencing pathway (Hanada et al., 2009). Although important, elucidation of the mechanism for their formation is hindered partly because of the lack of a MULE transposition system (or a species) that can be readily manipulated.

Despite their abundance in plants, only a few active MULE elements have been discovered and little is known about the transposition mechanism. *MuDR*, which was the first active autonomous MULE element discovered, encodes two proteins, MURA and MURB. Transcription of MURA and MURB is initiated within their adjacent TIRs (Lisch, 2002). MURA is the transposase protein that is responsible for catalyzing excision of *Mu* elements from the original locations (donor sites) accompanied with or without reintegration into new genomic loci (reinsertion sites). In contrast, the function of MURB remains unclear (Lisch, 2002). Subsequent to the discovery of *MuDR*, a number of active MULEs were reported in several species, such as *Jittery* in maize (Xu et al., 2004), *Os3378* in rice (Gao, 2012), and

AtMu1 in *Arabidopsis* (Singer et al., 2001), which are all distantly related to *MuDR*. More recently, another active MULE (*TED*) was discovered in maize, which is relatively close to *MuDR* based on their transposase sequences (Li et al., 2013). In all cases, transposition of MULEs has only been observed in their native hosts. As such, mechanistic studies are hindered and it is unclear whether any host factors are required for their transposition. Establishment of a transposition system in a maneuverable organism would facilitate studies of MULEs, especially the acquisition mechanism of Pack-MULEs. Yeast is an ideal species for this purpose because of its small size, short generation cycles, and successful recapitulation of transpositions of other TEs (Yang et al., 2006; Hancock et al., 2010).

In this study, a rice MULE, *Os3378*, successfully transposed in yeast. The transposition activity induced by the wild type transposase is rather low, but can be improved through a variety of approaches, including deletion of an N-terminal region and fusion with an enhanced yellow fluorescent protein. This study provides new insights about the function of transposase and regulation of transposition activity as well as the mechanism underlying formation of Pack-MULEs.

3.3 Results

3.3.1 The *Os3378* Transposase

The *Os3378* element has 4 copies with over 98% similarity at the nucleotide level in Nipponbare, the sequenced *japonica* rice cultivar. All of the four copies are annotated as transposon proteins (LOC_Os04g28290, LOC_Os04g28350, LOC_Os05g31510, and LOC_Os11g44180) in MSU version 7 rice pseudomolecules. Three of the *Os3378* copies are associated with a terminal inverted repeat (TIR) of 196-bp in length. One of them (LOC_Os11g44180) is only associated with one terminal sequence, yet the coding region of

this element seems to be intact (Gao, 2012). To verify the gene structure, total RNA was extracted from a *japonica* somaclonal line called Z418, where the expression of *Os3378* was previously detected (Gao, 2012). The amplified cDNA sequence of *Os3378* from Z418 (called Os3378-Z) differs from that of LOC_Os5g31510 (computer annotation) at the splicing sites of intron 2 and 3 (Figure 3.1A and Figure 3.8A). Exon 2 of Os3378-Z is 288 nucleotides (nt) longer than that of LOC_Os5g31510, while exon 4 is 6-nt shorter than that of LOC_Os5g31510. The total length of Os3378-Z coding region is 886 amino acids (aa), which is 90 aa longer than that of LOC_Os05g31510. Except this difference, the rest of the sequences are the same between Os3378-Z and LOC_Os5g31510. The Os3378-Z transposase protein is also longer than that of three known active MULE elements in maize, MURA of *MuDR* (823 aa), JITA of *Jittery* (709 aa), and TEDA of *TED* (785 aa) (Chomet et al., 1991; Feldmar and Kunze, 1991; Hershberger et al., 1991; Xu et al., 2004; Li et al., 2013).

A functional transposase protein usually consists of a DNA binding domain (for binding transposon DNA) and a catalytic domain (for element excision and reinsertion). Through comparison to known DNA transposases (encoded by *IS256* and *Hermes*) (Hennig and Ziebuhr, 2010; Nesmelova and Hackett, 2010) and locating helix-turn-helix motifs, a putative DNA binding domain was identified between 300 and 400 aa of Os3378-Z, and a catalytic domain was located between 440 to 610 aa (Figure 3.1B; 445D-510D-610E). In addition, one nuclear export signal (NES) was predicted to be within the catalytic domain. Three nuclear localization signals (NLS) were predicted, with one located upstream of DNA binding domain and two located at the C-terminus of the protein. Interestingly, one of the NLSs at the C-terminal region is encoded by the additional sequence of exon 2 in Os3378-Z (compared to LOC_Os5g31510).

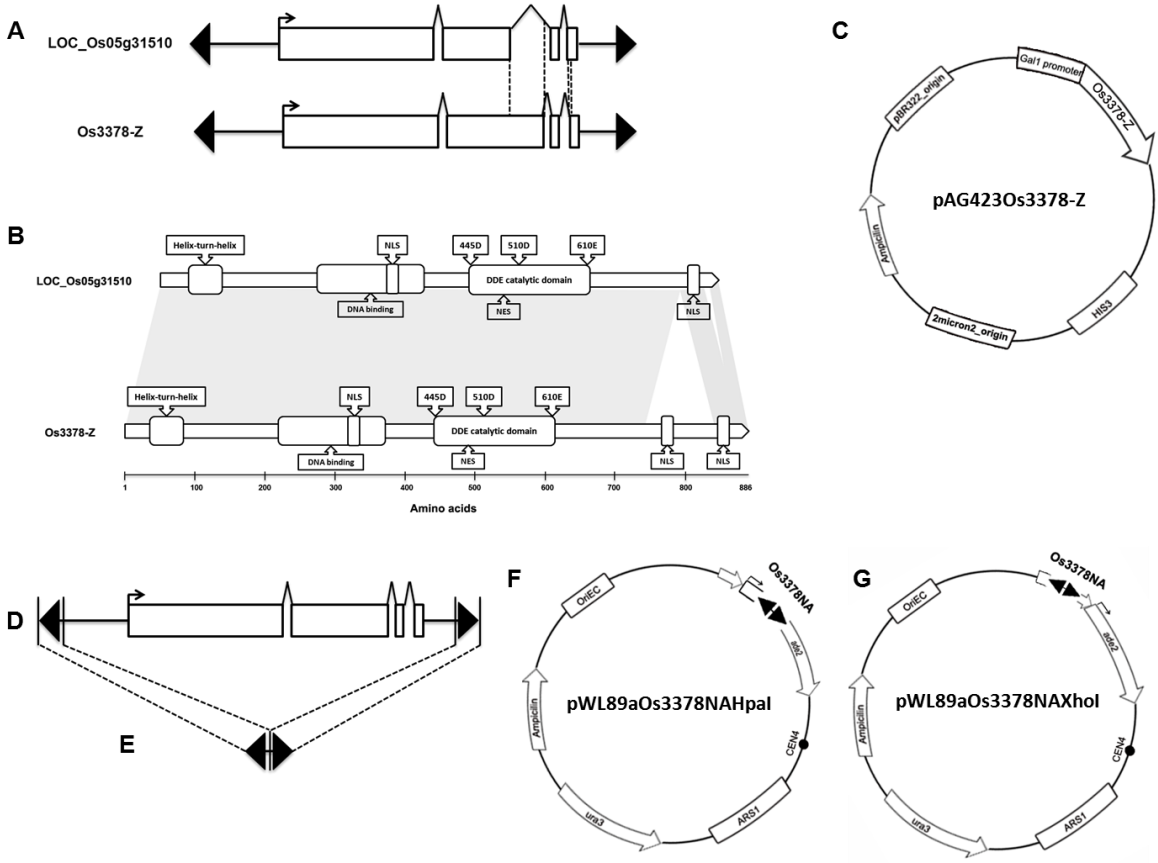

Figure 3.1 Schematic structures of Os3378 and constructs used in this study.

(A) Structural comparison between the annotated transposase (LOC_Os05g31510) and Os3378-Z at the nucleotide level. Black triangles: terminal inverted repeat (TIR); empty boxes: exons; lines linking empty boxes: introns; arrows: transcription start sites and orientation; dashed lines indicate splicing site variations between LOC_Os05g31510 and Os3378-Z.
(B) Structural comparison between LOC_Os5g31510 and Os3378-Z at the amino acid level. NLS: nuclear localization signal; NES: nuclear export signal. Shaded regions are in common between these two proteins.
(C) The expression vector with Os3378-Z transposase fused downstream of the GAL1 promoter.
(D and E) Os3378-Z and the artificial non-autonomous *Os3378* (*Os3378NA*) which includes the 5′ and 3′ TIRs and partial sub-terminal sequence. Symbols are the same as in (A).
(F and G) The reporter vector with Os3378NA inserted in the coding sequence and the promoter region of the ADE2 gene.

3.3.2 Os3378-Z is Capable of Induction of Excision and Reinsertion with Low Frequency in Yeast

The transposition assay consisted of two vectors, *i.e.*, an expression vector containing the Os3378-Z transposase coding sequence downstream of the GAL1 promoter, and a reporter vector with an artificial non-autonomous *Os3378* element (*Os3378NA*) inserted into either the coding region (pWL89aOs3378NAHpaI) or the promoter region (pWL89aOs3378NAXhoI) of the ADE2 gene (Figure 3.1C-G). *Os3378NA* is comprised of the 5′ and 3′ TIRs of *Os3378* with a short linker sequence (~30 bp) between them. The *Os3378NA* in the coding region of ADE2 is flanked by a TSD (TTAATTTAA), while the one in the promoter region is not.

Excision of *Os3378NA* and recovery of a functional ADE2 gene allows growth of yeast on culture medium lacking adenine. In general, the excision frequency was rather low (~0.055 event/10^{-6} cells on 2% galactose). The excision events were further characterized by amplifying the donor site following excisions using primers in the flanking region of *Os3378NA* insertion sites and sequencing the amplicons. For the reporter vector with *Os3378NA* inserted in the coding region of ADE2 (pWL89aOs3378NAHpaI), five colonies were tested and all of them exhibited recovery of the original ADE2 sequence, suggesting the selective pressure for ADE2 function. When the reporter vector pWL89aOs3378NAXhoI (with *Os3378NA* in the promoter region) was used, variation of sequences at the donor site following excisions was observed, which were classified into three types (Figure 3.2). Type I includes those that restored the original ADE2 promoter sequence. Type II sequences consist of deletions of one or both flanking regions of the donor site, while type III sequences retained one terminal sequence on one side accompanied with or without small deletions on the other side of the donor site (Figure 3.2). Out of 14 cases studied, type III is the most

frequent, accounting for 7 cases (50%). The length of the retained terminal sequence is between 122-188 bp, which contained more than 62% sequence of an entire *Os3378* terminus (196 bp). Surprisingly, 6 out of the total 7 cases showed a preference for retention of the terminal sequence at the 3′ end, suggesting such orientation may favor the expression of the ADE2 gene. In addition, five type III events contained exactly the same sequences at the donor site, which is featured by retention of 122 bp of the 3′ terminal sequence and 34 bp deletion of the upstream flanking sequence of the *Os3378NA*. This is unlikely contamination among different yeast ADE2 revertant colonies because they were either from different sectors (one plate was divided into several sectors) or plates. Furthermore, one of the five events was detected in one experiment while the other four were from another experiment. Nevertheless, for the four events from the same experiment, the possibility of an early excision event cannot be excluded, which may have occurred before cells were plated on the medium lacking adenine. As a result, the 7 cases were at least derived from 3 independent excision events, suggesting that retention of a single terminal sequence after excision was not a rare event. However, search of the Nipponbare genomic sequence did not recover any single terminus of *Os3378* and no additional copy was detected in Z418 as well (Gao, 2012).

Figure 3.2 Sequences of the donor site following excision of *Os3378NA*.

The original donor site sequence (promoter region of the ADE2 gene) prior to the incorporation of *Os3378NA* is shown on the top followed by the donor site sequence with *Os3378NA*. Nucleotides underlined are the restriction site (*Xho*I) used to insert *Os3378NA*. Black triangles denote the TIR of *Os3378*. Nucleotides flanking *Os3378NA* in the rectangle are sequences from the cloning vector between the *Xho*I and *Eco*RV restriction sites (see Section 3.5.3). Numbers in parenthesis are yeast revertant colonies containing that type of sequences at the donor site following excisions.

Using the transposon display technique, reinsertions of *Os3378NA* after its excision from the coding region of the ADE2 gene were identified. Cloning of two reinsertion sites indicated that both inserted into 35S rRNA genes, but within different locations. One element inserted into the internal transcribed spacer region (ITS1-1) and the other inserted into the 18S rRNA sequence (RDN18-1). Both insertions were associated with a 9-bp TSD (AAAATTTAA; TTGAAAAAA) immediately flanking the element ends. Thus, Os3378-Z is capable of inducing both excision and reinsertion of *Os3378NA* in yeast, albeit the frequency is low.

3.3.3 Deletion of the N-Terminal Sequence of Os3378-Z Transposase Enhanced Excision Frequency Which is Responsive to Galactose Concentrations

Previous studies showed that deletion of the N-terminal region of *Ac* (Kunze et al., 1995) and phage *Mu* (MuA) transpoases (Kim and Morrison, 2009) resulted in increased

transposition activity. Comparison of their amino acid compositions revealed no common sequence feature. However, when comparing their secondary structures, a helix-turn-helix domain was detected in the deleted regions of both transposases (*Ac* transposase: 10-50 aa; MuA: 1-77 aa). Analysis of Os3378-Z transposase sequence suggested that it also contains a helix-turn-helix domain at the N-terminus (40-100 aa). The structural similarity of the N-terminal regions of all three transposases prompted us to test whether deletion of the N-terminal region would enhance the excision frequency of Os3378-Z. To this end, five truncated transposases were constructed: Os3378-Z-80, Os3378-Z-105, Os3378-Z-130, Os3378-Z-161, and Os3378-Z-168, which contain deletions of the N-terminal 79, 104, 129, 160, and 167 aa, respectively (Figure 3.3A). In the following studies, reporter vector with *Os3378NA* integrated in the coding region of ADE2 gene was used to test the transposition activity of different forms of Os3378-Z transposase.

As mentioned above, the Os3378-Z transposase is under the control of the GAL1 promoter, which is induced by galactose. Therefore, culture media with various galactose concentrations (0.05%, 0.1%, 0.5%, 1%, 2%) were used to modulate the expression levels of the transposase and Western blot analysis using the His antibody was conducted to determine the protein levels. Not surprisingly, increasing galactose concentrations led to increased protein levels as evidenced by both Os3378-Z and Os3378-Z-130 (Figure 3.3C, D), although it seems that the wild-type Os3378-Z is more responsive to high galactose concentrations.

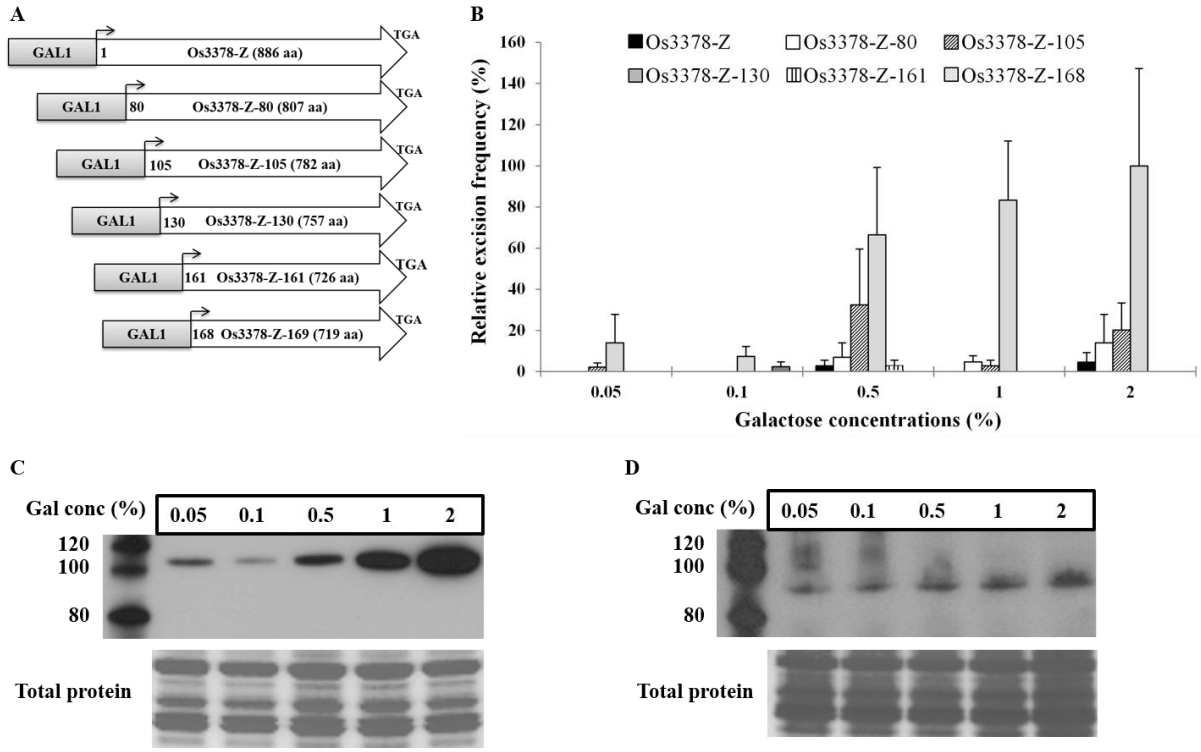

Figure 3.3 N-terminal deleted Os3378-Z transposases.

(A) Schematic representation of full-length and N-terminal deleted Os3378-Z transposases
(B) Relative excision frequency of full-length and N-terminal deleted Os3378-Z transposases in response to various galactose concentrations. The highest excision frequency (*i.e.*, Os3378-Z-130 at 2% galactose concentration) was set as 100%.
(C) Protein levels of Os3378-ZCFH under various galactose concentrations. CFH: C-terminal FLAG-His_6 tag. A partial image of the total protein is shown at the bottom (see Section 3.5.3).
(D) Protein levels of Os3378-Z-130CFH under various galactose concentrations. A partial image of the total protein is shown at the bottom.

The excision frequency increased with elevated galactose concentrations for both wild type and truncated transposases, where the trend for Os3378-Z-130 is the most prominent (Figure 3.3B). For Os3378-Z, Os3378-Z-80, Os3378-Z-105, Os3378-Z-161, and Os3378-Z-168, few excision events were observed at low galactose concentrations (0.05% and 0.1%). Even at the highest galactose level (2%), the excision frequency is still low. In contrast, the excision frequency induced by Os3378-Z-130 is significantly enhanced ($p<0.0001$, *t*-test) at higher galactose concentrations (≥0.5%) compared with lower concentrations (0.05% and 0.1%). Moreover, Os3378-Z-130 resulted in higher excision frequency than the other five

transposases, especially at 1% and 2% galactose levels ($p<0.0001$, *t*-test). With 2% galactose, the excision frequency with Os3378-Z-130 is about 20-fold higher than that with the wild-type transposase (~1.2 vs. 0.055 events per 10^6 cells). As such, deletion of the N-terminal 129 aa of the Os3378-Z transposase significantly increased the excision frequency if galactose concentration is 0.5% or higher.

3.3.4 Excision Frequency Altered by Substitutions of Amino Acids within 105 to 130 aa of Os3378-Z Transposase

The differential excision frequency caused by Os3378-Z-130 and Os3378-Z-105 prompted us to study the amino acid composition between 105-130 aa region, which is the only sequence/structural difference between the two truncated transposases. Multiple sequence alignment revealed that this region was not conserved among different known MULE transposases, including MURA, TEDA, and JITA. Analysis of the amino acid composition indicated that it contains 35% (9 out of 26 aa) acidic residues and only two basic residues, and the entire region is very hydrophilic based on the hydropathy analysis (Kyte and Doolittle, 1982). To determine whether the amino acid composition of 105-129 aa affected the excision activity, some amino acids in this region were mutated and four mutant transposases were constructed (Figure 3.4A). In Os3378-Z-105Ala, most of the acidic residues were substituted by alanine; in Os3378-Z-105Neutral, all the acidic residues were substituted by their corresponding neutral amino acids; in Os3378-Z-105Basic, all the acidic residues were substituted by basic amino acids with similar molecular weight, and in Os3378-Z-105Hydrophobic, 9 hydrophilic residues were replaced by hydrophobic ones (see Section 3.5.3). Excision assay showed that both Os3378-Z-105Ala and Os3378-Z-105Neutral resulted in much higher activity than the original Os3378-Z-105 ($p<0.01$, *t*-test) and the excision frequency was comparable to that by Os3378-Z-130 ($p>0.81$, t-test) (Figure 3.4B).

However, Os3378-Z-105Basic exhibited no obvious difference from the original Os3378-Z-105 (p=0.3204, *t*-test) and mutation of hydrophilic amino acids to hydrophobic ones completely abolished the activity at all galactose concentrations (Figure 3.4B), suggesting chemical and physical properties of amino acids in this region (105-130 aa) are critical to the activity of the Os3378-Z transposase.

A

Amino acid substitutions	Transposase name	Sequence
Change to small amino acids	Os3378-Z-105Original:	I**L**KHNE**DD**NV**V**LSNGVSD**D**SDV**ED**D
	Os3378-Z-105Ala:	I**A**KHNE**AA**NV**A**LSNGVSD**A**SDV**AA**D
Change to neutral amino acids	Os3378-Z-105Original:	ILKHN**EDD**NVVLSNGVS**DD**S**D**V**EDD**
	Os3378-Z-105Neutral:	ILKHN**QNN**NVVLSNGVS**NN**S**N**V**QNN**
Change to basic amino acids	Os3378-Z-105Original:	ILKHN**EDD**NVVLSNGVS**DD**S**D**V**EDD**
	Os3378-Z-105Basic:	ILKHN**RKK**NVVLSNGVS**KK**S**K**V**RKK**
Change to hydrophobic amino acids	Os3378-Z-105Original:	IL**KHN**EDD**N**VVLS**N**GV**S**D**DS**DVE**D**D
	Os3378-Z-105Hydrophobic:	IL**IFL**EDD**L**VVLS**I**GV**V**D**LV**DVE**I**D

B

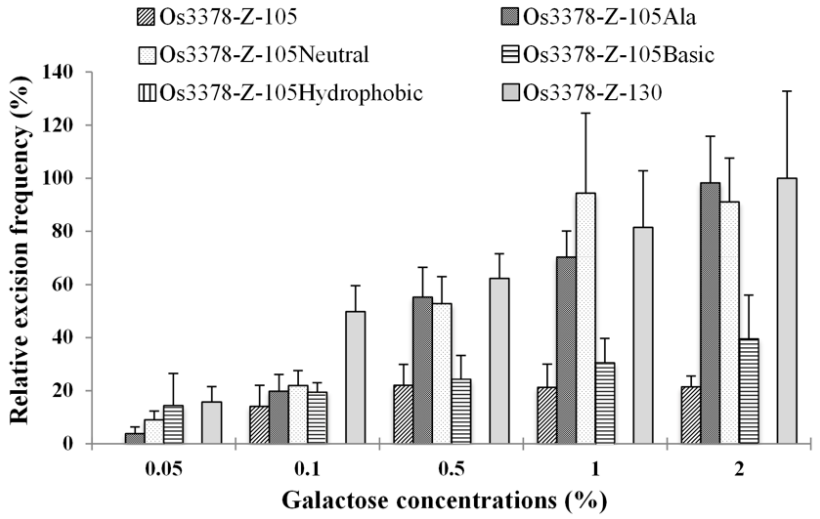

Figure 3.4 Os3378-Z-105 transposases with amino acid substitutions and their excision frequency.

(A) Amino acid substitutions between 105-130 aa of Os3378-Z-105
(B) Relative excision frequency of Os3378-Z-105 and its mutant forms under various galactose concentrations. The highest excision frequency (*i.e.*, Os3378-Z-130 at 2% galactose concentration) was set as 100%.

3.3.5 Cellular Localization of Os3378-Z Transposases with an N- or C-Terminal EYFP Fusion as well as its Effect on Excision Frequency

To determine whether the increased excision rate of Os3378-Z-130 was due to its cellular localization, an EYFP-tag (enhanced yellow fluorescent protein) was fused to the N- or C-terminus of these transposases (Figure 3.10B). Since the size of the EYFP is relatively large (~260 aa and ~29 kD), whether the fusion would affect the excision frequency was assessed at 2% galactose concentration. For the wild-type Os3378-Z transposase, the EYFP fusion at the N-terminus exhibited no obvious effect while no excision event was observed if the fusion is at the C-terminus (Figure 3.5B). However, this should be interpreted cautiously because the basal excision frequency is low for Os3378-Z. Interestingly, addition of the EYFP exerted different effects on the N-terminal truncated transposases with regard to the fusion positions. For Os3378-Z-105, fusion of EYFP at its N-terminus enhanced its activity whereas C-terminal fusion of EYFP suppressed its activity (Figure 3.5B). For Os3378-Z-130, both N- and C-terminal fusions led to reduced activity. Since the N-terminal fusion of EYFP on Os3378-Z-105 demonstrated positive effect on transposition, the amino acid sequence that was introduced through fusion, which is directly attached to the transposase, was examined. The 25 amino acid peptide upstream of the transposase (C-terminal sequence of the EYFP) is hydrophobic, containing 4 basic amino acids with no acidic amino acids. In contrast, the sequence upstream of 105 aa (80-104 aa) in Os3378-Z contain 6 acidic and only 1 basic amino acids.

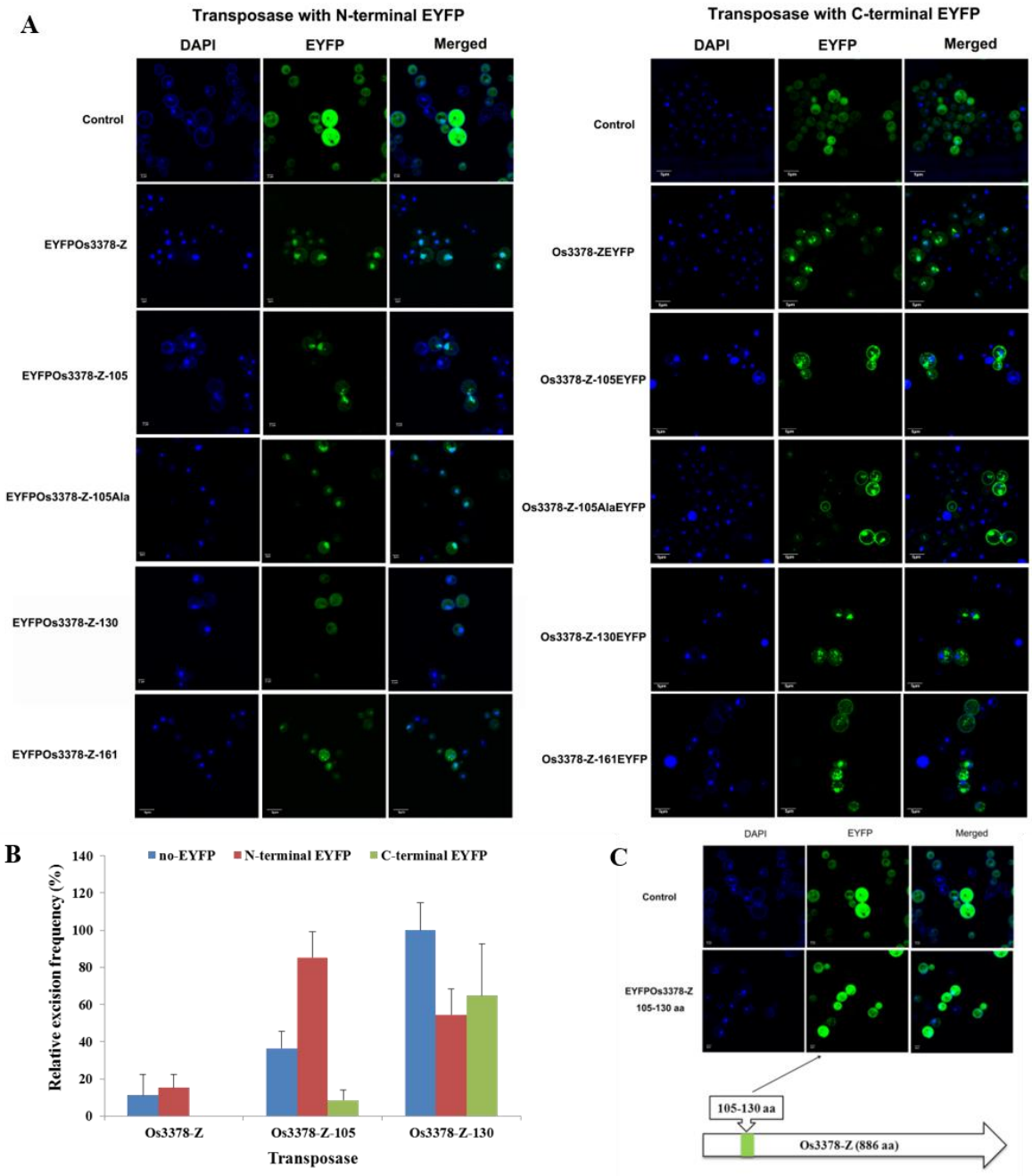

Figure 3.5 Cellular localization and excision frequency of N- or C-terminal EYFP-tagged transposases.

(A) Cellular localization of full-length and N-terminal deleted transposases with EYFP tag at the N-termini (left panel) and C-termini (right panel). DAPI (4', 6-diamidino-2-phenylindole) is a fluorescent stain that binds to DNA.
(B) Relative excision frequency of different transposases at 2% galactose concentration.
(C) Cellular localization of EYFP with and without the Os3378-Z 105-130 protein sequence.

Confocal microscopy examination indicated that transposases with N- and C-terminal EYFP fusions had different cellular localizations. When the fusion was at the N-terminus, both EYFPOs3378-Z and EYFPOs3378-Z-105 were concentrated in the nucleus as revealed by the co-localization of the DAPI-stained nuclear DNA and the EYFP-tagged transposases (Figure 3.5A). In contrast, EYFPOs3378-Z-130 protein was almost evenly distributed in the nucleus and cytoplasm (Figure 3.5A). When the fusion was at the C-terminus, all fusion proteins formed aggregates in the cytoplasm (Figure 3.5A). This may explain the reduced activity of both Os3378-Z-130 and Os3378-Z-105 upon fusion with EYFP at their C-termini.

The distinct cellular localization between EYFPOs3378-Z-105 and EYFPOs3378-Z-130 (with EYFP at the N-termini) raised the question whether the 105-130 aa region contains a cryptic nuclear localization signal. To this end, this region was fused downstream of an EYFP tag and its cellular localization was observed using confocal microscopy. As shown in Figure 3.5C, the resulting EYFP fusion protein was present over the entire cells, suggesting that this region alone does not act as a nuclear localization signal. This finding was further confirmed by the nuclear localization of EYFPOs3378-Z-105Ala, which contains 7 amino acid substitutions between 105-130 aa (Figure 3.5A) with a similar cellular localization to EYFPOs3378-Z-105. Similar to other transposases (Os3378-Z, Os3378-Z-105, and Os3378-Z-130), C-terminal fusion of EYFP resulted in aggregation of Os3378-Z-105Ala (Figure 3.5A).

3.3.6 Protein Levels of Different Forms of Os3378-Z Transposase are not Correlated with Excision Frequency

To determine whether excision frequency of full-length and N-terminal deleted transposases is associated with protein levels in yeast cells, Western blot analysis was conducted. For such analysis, a FLAG-His_6 dual-tag (Zanetti et al., 2005) was fused to the C-

termini of the transposases (CFH), which were referred to as Os3378-ZCFH, Os3378-Z-105CFH, Os3378-Z-105AlaCFH, and Os3378-Z-130CFH. Unlike EYFP, CFH-tag is short in length (25 aa including spacer residues) and has small molecular weight (~2 kD), which theoretically would not affect the activity of the transposases as much as EYFP. Indeed, analysis of the excision frequency revealed no significant difference between the CFH-tagged and control transposases (*e.g.*, Os3378-Z-130 vs. Os3378-Z-130CFH; p=0.9535, *t*-test) (Figure 3.6A). Western blot analysis indicated that protein levels of Os3378-ZCFH and Os3378-Z-105AlaCFH were higher than that of Os3378-Z-105CFH and Os3378-Z-130CFH under the same galactose concentration (2%) (Figure 3.6B). In general, no correlation was observed between protein levels and excision frequency among different forms of Os3378-Z transposase (p=0.7183; Figure 3.6C).

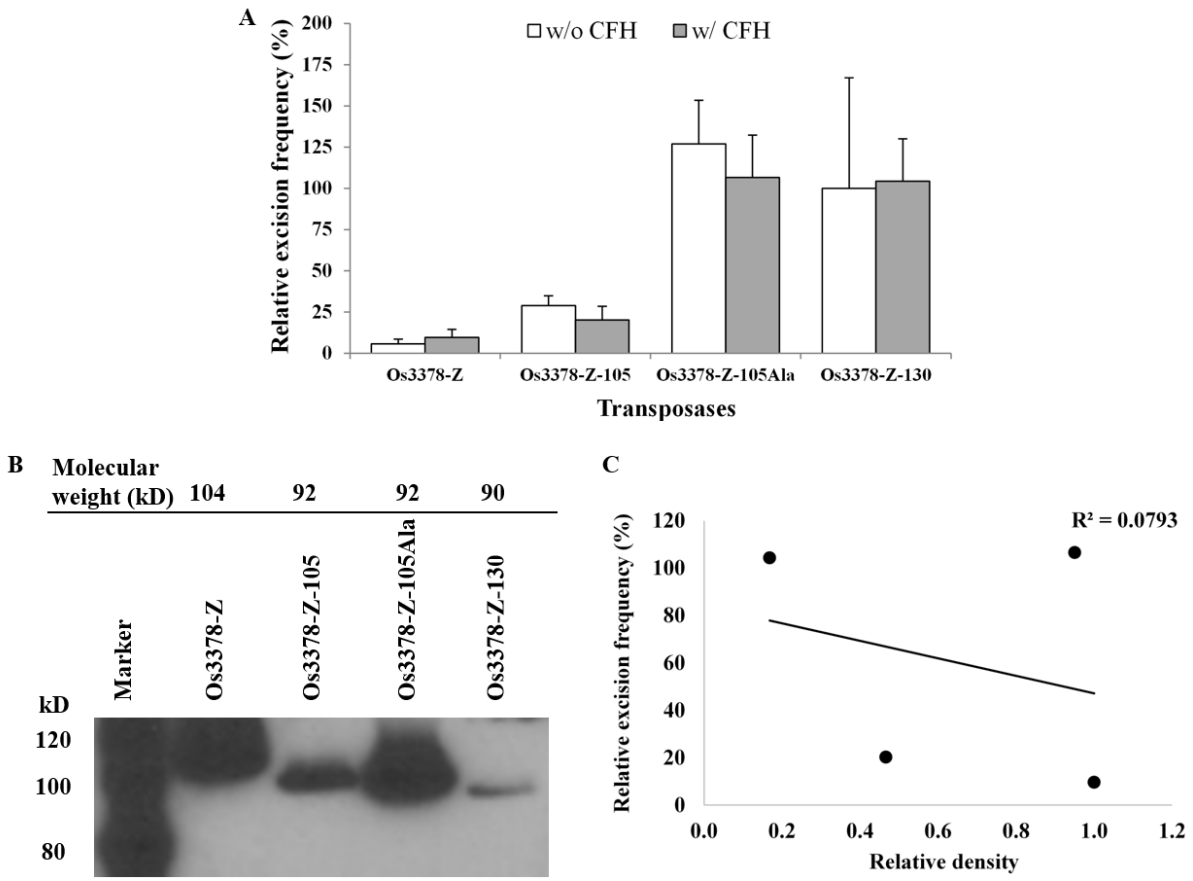

Figure 3.6 Western blot analysis and excision frequency of transposases with and without C-terminal FLAG-His$_6$ tag (CFH).

(A) Relative excision frequency of transposases with and without CFH tag at 2% galactose concentration.
(B) Protein levels of transposases at 2% galactose concentrations. Numbers on the top are molecular weight of full-length and N-terminal deleted Os3378-Z transposases with CFH tag.
(C) Relative excision frequency (B) as a function of protein levels (C) (relative density) of various forms of Os3378-Z transposase with CFH tag.

3.3.7 Single *Os3378* Terminus is Immobile

In 1988, Talbert and Chandler proposed that the presence of genes inside *Mu* elements could be caused by the mobility of *Mu* termini (Talbert and Chandler, 1988). If two independent *Mu* terminus from the same family encompass a gene, a Pack-MULE is formed. To determine whether *Os3378* terminus was capable of excision, a reporter construct containing only the 5′ terminal sequence of *Os3378* in the coding region of the ADE2 gene

was transformed into yeast along with the expression construct containing Os3378-Z-130. Out of 30 replicates, no ADE2 revertants were observed on culture medium without adenine, suggesting no excisions were induced by the transposase. This indicates that single *Os3378* terminus is unlikely to be mobile.

3.3.8 Reinsertion of *Os3378NA*

To determine whether the factors influencing excision frequency also affect the reinsertion frequency, Southern blot experiment was conducted using total DNA extracted from the ADE2 revertant yeast colonies. In Figure 3.7A, the hybridization signals shared by all samples represent the *Os3378NA* that is in the reporter vector and other signals correspond to reinserted *Os3378NA* elements. Similar number of colonies with reinsertions were observed between EYFPOs3378-Z-105 and EYFPOs3378-Z-130 at 1% (7 vs. 8, χ^2=0.1101, p=0.74) and 2% (8 vs. 10, χ^2=0.4222, p=0.5158) galactose. Different numbers of reinsertions were observed at 0.1% (9 vs. 5, χ^2=1.8095, p=0.1786) and 0.5% (6 vs.11, χ^2=2.6611, p=0.1028) galactose concentrations. However, the differences were not statistically significant (Figure 7A). Given the fact that the total number of colonies with reinsertions is comparable between these two forms of transposase (30 vs. 34), it is likely the reinsertion frequency remains unchanged. Overall, 38-49% of the excision events are accompanied with reinsertion events. Since EYFPOs3378-Z transposase did not induce many excision events, the ADE2 revertant colonies from all the galactose concentrations were tested in a single blot, and the result revealed that 6 (32%) out of 19 yeast colonies were associated with reinsertions (Data not shown), which is largely comparable to that from other transposases (EYFPOs3378-Z-105 and EYFPOs3378-Z-130).

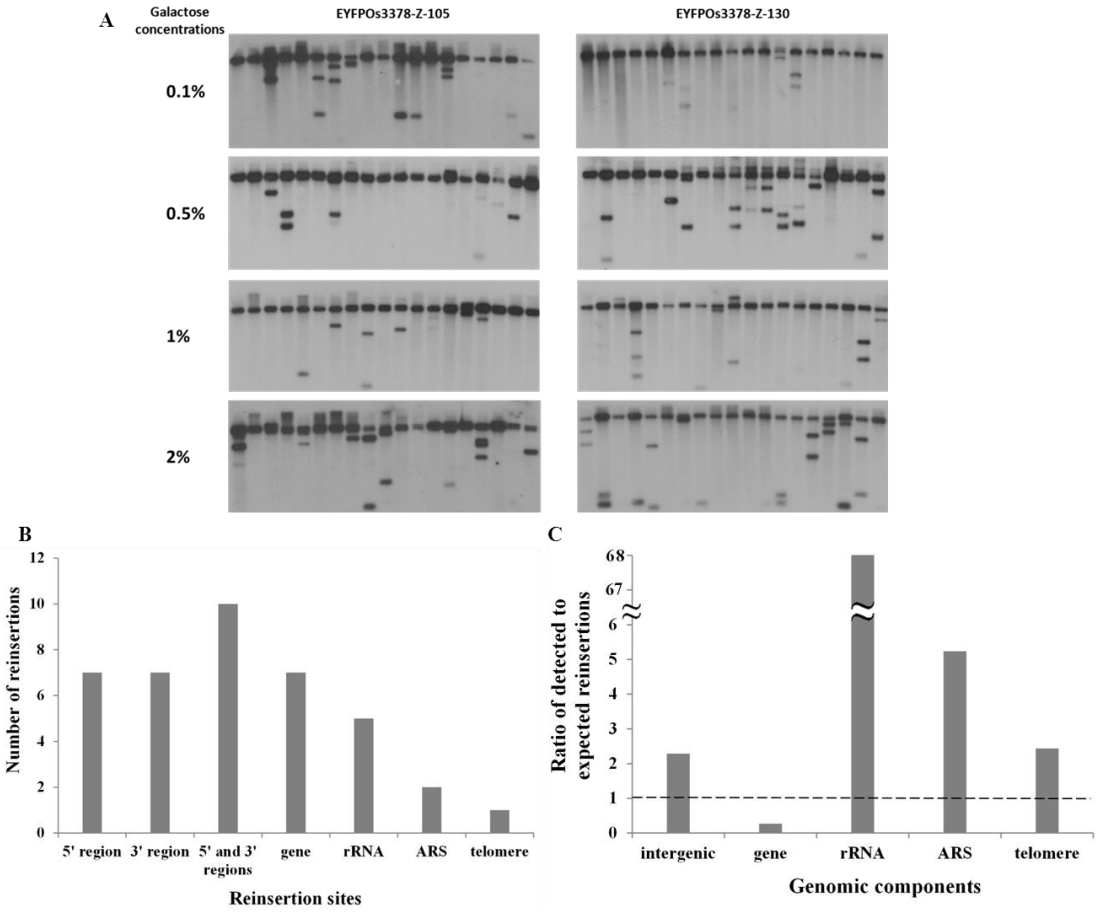

Figure 3.7 Reinsertions of *Os3378NA*.

(A) Southern blot analysis of reinsertions of *Os3378NA.*
(B) Distribution of reinsertion sites of *Os3378NA* with regard to genomic sequence features.
(C) Ratios of observed distribution of reinsertion sites to that of fraction of each genomic feature in the yeast genome (S288C).
5′ region: reinsertions located at the upstream region of transcription start site (TSS) of protein coding genes on both sides of reinsertion site; 3′ region: *Os3378NA* insertions downstream of the transcription termination site (TTS) of protein coding genes on both sides of reinsertion site; 5′ and 3′ regions: reinsertions located at the upstream region of the TSS of one gene and the downstream region of the TTS of another; ARS: autonomously replicating sequence; intergenic: including insertions in both 5′ and 3′ regions of genes.

To determine the location of new insertions, forty-four reinsertion sites were cloned and sequenced, where five (11%) of them had no match to either the yeast genome (S288C) or vector sequences. This is likely due to the sequence polymorphisms between the two yeast strains (S288C vs. DG2523). Thirty-nine sequences have significant alignments with the yeast genomic sequence (Appendix Table 3.1). The majority of insertions (62%) are located in gene-rich regions where equal numbers (n=7) of insertions are located in the 5′ and 3′ regions of protein-coding genes (Figure 3.7B, see Section 3.5.8). Additionally, ten insertions are located in the 5′ region of one gene and 3′ region of another (see Section 3.5.8). Compared with the genomic fraction of intergenic sequence (~24%), reinsertions in these regions are ~2-fold higher than expected if the insertion is random in the genome (Figures 3.7C and 3.8). On the contrary, insertions within gene bodies are only one fourth of that expected based on the fraction of genes in the genome. This is likely because of the more deleterious effect of insertions inside genes. Hence, reinsertions in genes are likely higher than what was observed. In addition, five insertions are located in rRNAs genes, which is 68-fold of the genomic fraction of rRNA genes. The overrepresentation of insertions in rRNA genes is likely due to the functional redundancy but not the "TA" content because the average "TA" content of rRNA genes is similar to that of the entire genome (61.56% vs. 61.85%). Reinsertions were also detected in autonomously replicating sequences (ARS; n=2) and the telomere (n=1). However, no insertion was found in LTR retrotransposons although they occupy twice more genomic space than both ARS and the telomeres of yeast (Figures 3.7C and 3.8). For those mapped reinsertion sites, the 9-bp sequences immediately flanking *Os3378NA* are highly "TA"-rich, with an average "TA" content of 87%, which is much higher than the genomic average (62%) of yeast. This target site preference is similar to what was observed for *Os3378* in rice (~86%) (Gao, 2012).

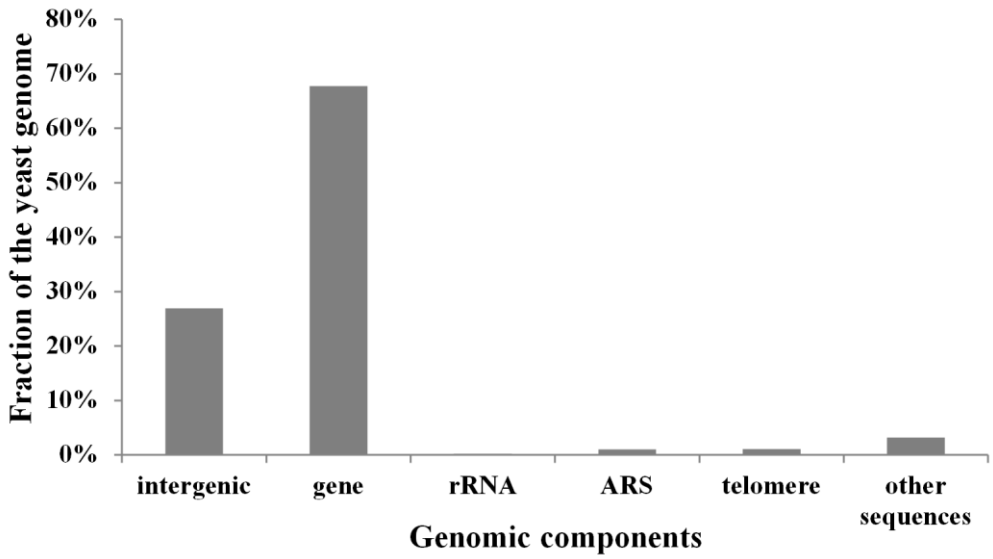

Figure 3.8 Fractions of genomic sequence features of the yeast (S288C).

3.4 Discussion

In this study, transposition of a rice MULE, *Os3378*, was recapitulated in yeast. To our knowledge, this is the first report of MULE transposition in a heterologous species. It should be noted that the actual excision frequency should be higher than that was shown in the results because the data only represent yeast colonies that experienced excision and recovery of a functional ADE2 gene through the repair system of yeast. Those with excisions but without correct repair of the ADE2 gene would not survive on the medium lacking adenine.

3.4.1 The Low Copy Number of *Os3378* in Rice and Its Relatives is Likely Due to Its Low Transposition Activity of the Relevant Transposase

In a previous study, it was shown that *Os3378* is present in most rice cultivars tested as well as wild relatives with AA genomes (Gao, 2012). However, *Os3378* remains low copy number (less than 10) in all the species tested. The transcripts of *Os3378* were readily detectable in the reproductive tissues of Z418 (Gao, 2012). Nevertheless, additional new

insertions of *Os3378* were never detected in Z418 populations (Gao, 2012; Ferguson and Jiang, unpublished). The reasonable expression level and lack of amplification or new transposition seems to be puzzling. This study shows that the wild-type Os3378-Z transposase is not highly competent in inducing transposition events in a heterologous system, and such low competency cannot be significantly altered through its expression level. This may explain why the copy number of *Os3378* is uniformly low among a wide range of species and cultivars surveyed (Gao, 2012).

Prior to this study, it was already demonstrated that several wild-type transposases were not the forms optimal for transposition activity (Kunze et al., 1995; Zayed et al., 2004; Pledger and Coates, 2005; Lazarow et al., 2012). For example, mutation of three different amino acids in the transposase of *Himar1* (a *Mariner*-like transposon) from horn fly led to a 4 to 50-fold increase in transposition activity (Lampe et al., 1999). Given those instances, it is conceivable that high transposition activity might have been selected against because transposon insertions are more often linked to deleterious effects than favorable effects. As a result, low transposition activity and low copy numbers, as were demonstrated by *Os3378*, might be one of the strategies for long-term success. On the other hand, the transposition activity could be significantly enhanced by a variety of mutations, such as point mutations or truncations. Nevertheless, the fact that *Os3378* is not amplified in any cultivar or any species tested suggested either such mutations are rare or are selected against.

3.4.2 Transposition of *Os3378* in a Heterologous Organism

To date, transposition in non-host organisms other than their natural hosts has been proved for a number of TEs (Kunze et al., 1995; Plasterk et al., 1999; Weil and Kunze, 2000; Yang et al., 2006; Yang et al., 2007). In animals, the *Tc1*/*mariner* family is widespread and many elements have been shown to be able to transpose in other organisms (Plasterk et al.,

1999). One of the most promising TEs for gene therapy is the *Sleeping Beauty* transposon (*SB*). It was originally from fish genomes and was capable of transposition in human cells, mouse and rat (Ivics et al., 1997; Fischer et al., 2001; Keng et al., 2005). In plants, several TEs belonging to different families (*e.g.*, *hAT*, *Tc1/mariner*, *PIF/Harbinger*) have been tested for transposition in heterologous species. The maize *Ac/Ds* elements transposed in transgenic *Arabidopsis*, rice, tobacco, potato, yeast, and zebrafish (Knapp et al., 1988; Grevelding et al., 1992; Kunze et al., 1995; Enoki et al., 1999; Weil and Kunze, 2000; Emelyanov et al., 2006). *Osmar5*, a rice *Tc1/mariner*-like element transposed in yeast (Yang et al., 2006). A *PIF/Harbinger* element (*mPing*) transposed in both *Arabidopsis* (Yang et al., 2007) and yeast (Hancock et al., 2010). The within and trans-kingdom transposition of these TEs indicated that they do not require host-specific factors or these factors are conserved across the wide range of species. Unlike the TEs above, some elements seem to require host factors for transposition or optimal activity. Integration host factor (IHF) and Dam methylase affect the transposition frequencies of IS*50* and Tn*5*, where IHF is required for high activity in Dam^- cells (Makris et al., 1990). It was also shown that the TIR sequence of a MULE, *Mu1* in maize contains binding sites for nuclear proteins, suggesting host factors may be involved in its transposition process (Zhao and Sundaresan, 1991). In this study, the recapitulation of transposition in yeast suggests that *Os3378* either does not require host specific factors or these factors are conserved between rice and yeast, despite the fact that the yeast genome only harbors retrotransposons (Kim et al., 1998).

3.4.3 A Single Transposase of *Os3378* Catalyzes Both Excision and Reinsertion Events

Among known autonomous MULE elements, *Os3378* transposase is more closely related to *AtMu1*, an element from *Arabidopsis*, than to *MuDR*, *Jittery*, and *TED* from maize (Zhao and Jiang, 2014). In maize, *MuDR* encodes two proteins, with MURA as the transposase and

being sufficient to induce excision. The role of MURB remains enigmatic and it was suggested that it may be involved in element insertions (Lisch et al., 1999; Raizada and Walbot, 2000; Woodhouse et al., 2006). Homologous sequences of MURA have been found in a wide range of species; however, no MURB sequences have been found in any species other than those in the *Zea* genus (Lisch, 2002). In contrast, *Jittery*, *AtMu1*, *TED* and *Os3378* all encode a single protein. Except *Jittery*, whose reinsertion was not observed, both *AtMu1* and *TED* can induce excision and reinsertion even though no MURB-related sequence was found (Singer et al., 2001; Li et al., 2013). Similarly, *Os3378* is also capable of both excision and reinsertion without a MURB-like protein. From a maximum parsimonious point of view, it is likely that the requirement of MURB for reinsertion was derived later for the *Mu* system in maize and the ancient MULE only encodes a single transposase.

3.4.4 Comparison of Target Specificity between Rice and Yeast

For the target site duplication, the average "TA" content for that of *Os3378* in rice is about 86%, which is similar to that of *AtMu1* in *Arabidopsis* (89%, Singer et al., 2001). This preference also retains for *Os3378* in yeast, where the average "TA" content of the 9-bp flanking sequences of reinsertion sites is 87%. Considering the genome-wide average "TA" content (rice: ~56%; yeast: 62%), this "TA" preference becomes more evident, suggesting that *Os3378* is highly specific in targeting "TA"-rich sequences.

It is not surprising that the majority (~62%) of the recovered reinsertions occurred in intergenic regions since those insertions usually do not cause complete loss of function of the genes in their vicinity. About 13% of the reinsertions are in rRNA genes (genomic fraction of 0.19%), likely because these rRNAs have multiple copies in tandem so there is significant functional redundancy. In addition, less reinsertion occurred in repetitive regions, including telomere, LTR retrotransposons, which is in agreement with the target specificity of MULEs

in other studies (Raizada et al., 2001; Liu et al., 2009). Previous studies showed that MULEs prefer to insert into the 5′ region of genes with decreasing frequencies towards the 3′ region (Dietrich et al., 2002; Liu et al., 2009; Jiang et al., 2011). Among the four copies of *Os3378* in Nipponbare, only one of them is located in proximity to a gene and it is located at the 5′ region (Gao, 2012), which is insufficient to determine whether it has the same preference as other MULEs. In this study, equal number of *Os3378NA* insertions was found at the 5′ and 3′ regions of genes. As a result, it is unclear whether *Os3378* has a preference for 5′ region of genes or whether such preference is not recapitulated in yeast.

3.4.5 A Critical Region for Modification of Transposition Activity through Deletion and Substitutions

Previous studies have shown that deletion of selected regions of transposases can enhance their transposition activity. For example, deletion of the first 102 amino acids of the maize *Ac* transposase (a total of 807 aa) led to an increase of excision activity, while further deletion (up to 188 aa) aborted the transposase activity (Kunze et al., 1993). This is because *Ac* transposase has a basic DNA binding domain around 200 aa, and deletion of the first 188 aa lead to the loss of DNA binding activity (Feldmar and Kunze, 1991). In addition, a 10-fold "PQ" (Pro-Gln or Pro-Glu) repeat is present in 109-128 aa, which is essential for transposase activity (Kunze et al., 1993). In addition, deletion of C-terminal sequences (53 and 98 aa) of the *Ac* transposase abrogated excision activity, suggesting the C-terminus is essential for transposition. Beside *Ac* transposase, deletion of the first 77 aa of the MuA (transposase of phage *Mu*) also resulted enhanced transposition activity (MuA, 663 aa in length) (Kim and Morrison, 2009). However, unlike the *Ac* transposase, deletion of part of the C-terminal sequence of MuA also led to the enhancement of transposition activity, albeit to a less degree compared to that derived from N-terminal deletion. The MuA transposase with deletion of

both termini, on the other hand, is only associated with a minor improvement of transposition activity which is inferior to that from deletion of either end.

Since N-terminal deletion led to higher transposition activity for both *Ac* and phage *Mu* transposases, a similar strategy was employed to enhance the transposition activity of wild-type transposase of Os3378-Z. The consequence of deletion resembles that for *Ac* transposase, *i.e.*, the effective region for deletion is located in 105-130 aa. No significant alteration on transposition frequency was observed if the deletion is too short (79 aa or 104 aa) or too long (160 aa). As a result, it is likely that a short deletion does not significantly change the property of the transposase while a long deletion may lead to the loss of essential components for the transposition machinery.

Since the deletion of first 129 aa does not abort the transcription activity, it is intriguing what role of this portion of transposase plays. Obviously, this portion is not essential for transposition. Within this region, a 25 aa peptide (105-129 aa) seems to be critical for both transposition activity and the cellular localization of the transposase. With this peptide, the transposase is located in the nucleus (EYFPOs3378-Z, EYFPOs3378-Z-105, EYFPOs3378-Z-105Ala), whereas the transposase is located in both nucleus and cytoplasm when this peptide was removed (EYFPOs3378-Z-130). Since this peptide alone fails to direct the EYFP protein to nucleus, it is clear that this peptide does not serve as a nuclear localization signal itself (Figure 3.5C). In this case, the likelihood is that the absence of this peptide influences the conformation of the remainder of the transposase so that the nuclear export signal (NES) is exposed. Alternatively, the change in conformation buried the NLSs so they are less effective in the transport process.

In addition to the impact on cellular localization, the conformation change due to the deletion or replacement of the short peptide may lead to the structural change in the DNA binding domain or catalytic domain of transposase so that the transposase is more efficient. This is consistent with the fact that replacement of the amino acids in 105-129 aa with different physical properties significantly altered the excision activity. The original 105-129 aa peptide contains 36% acidic aa, 56% neutral aa, and 8% basic aa, which results in an overall acidic and hydrophilic status. During protein folding, hydrophobic peptides tend to form internal regions while hydrophilic region are often at the surface of the protein (Kyte and Doolittle, 1982). This implies that, at physiological conditions, this peptide is negatively charged and likely located in external regions of the protein molecule. As a result, either the negative charges carried by the peptide interfere with the function (for example, the interaction between transposase and transposon ends) or the presence of this peptide at the surface of the protein prevents the exposure of the reaction center. Substitution of some (Os3378-Z-105Ala) or all (Os3378-Z-105Neutral) of the acidic residues with neutral ones resulted in enhanced excision frequency whereas substitution of acidic residues to basic residues did not bring any benefit. This suggests the presence of any significant net charge in this region might be detrimental to transposition activity. Meanwhile, barely any excision event was observed for Os3378-Z-105Hydrophobic, which has similar amount of neutral residues in this peptide as that of Os3378-Z-105Ala, but very high level of hydrophobicity. It is possible that the introduction of high level of hydrophobicity led to gross structural alteration which abolishes the function of the transposase. Overall, neutrality and low hydropathy of this peptide seem to be critical for the transposition activity, although the underlying mechanism is not well understood (also see below).

3.4.6 Protein Levels and Transposition Activity of Various Forms of Os3378-Z Transposase

Apparently, different forms of Os3378-Z transposase are associated with different levels of proteins at 2% galactose concentration (Figure 3.6B). This suggests that the N-terminal sequence is likely to play a role in the expression or stability of the transposase, in addition to its role in transposition. Specifically, deletion of N-terminal sequence resulted in reduced level of protein, so the presence of the N-terminal sequence either promotes expression (transcription and/or translation) or enhances the stability of the transcripts or proteins. Nevertheless, it is clear that the enhancement of transposition activity through deletion and mutation is unlikely due to the change in protein level since both Os3378-Z-105CFH and Os3378-Z-130CFH are associated with low (and comparable) levels of proteins but with Os3378-Z-130CFH induced significantly more excisions than Os3378-Z-105CFH at 1% and 2% galactose concentrations (Figure 3.6). Instead, the protein property or structure may play a major role. On the other hand, the protein level of Os3378-Z-105AlaCFH (with high activity) was much higher than that of Os3378-Z-105CFH, again suggesting the amino acid composition at the N-terminus affects expression efficiency or transcript/protein stability. However, it is unclear whether the enhanced activity of Os3378-Z-105AlaCFH is due to protein level or structural/chemical properties. In other words, the high excision activity achieved by different forms of transposases in this study could be due to distinct mechanisms.

3.4.7 The Effect of EYFP Fusion on Transposition Activity as well as Cellular Localization

Fluorescent proteins, including EYFP have been widely used to characterize protein trafficking, lipid metabolism, protein-protein interaction, and so forth. Adverse effects of the fluorescent protein on target proteins were reported, such as change of cellular localization and formation of dysfunctional protein. This study demonstrated that fusion of a protein (EYFP) to the transposase may significantly improve or inhibit the transposition activity. As was observed with deletions of Os3378-Z, the effect of the fusion depends on the location of the fusion because fusion to the wild-type transposase resulted in little difference. This might be because the fusion location is too distant to the functional domains (such as DNA binding domain or catalytic domain). Again the region around 105-129 aa is very critical in terms of alteration of transposase activity through fusion of an additional protein. Fusion of EYFP to the N terminus of Os3378-Z-130 largely offset the positive effect on transposition activity through deletion. As discussed above, significant net charge around 105-129 aa seems to be detrimental to transposition activity. When EYFP is fused to Os3378-Z-105, the introduction of 4 additional basic amino acids, together with 2 basic amino acids originally present in 105-129 aa may render this region a neutral state at physiological pH which may be the cause for the enhanced transposition activity. In contrast, when EYFP is fused to Os3378-Z-130, it introduces net charge to this region through the basic amino acids, which may inhibit the transposition activity. As a result, it is possible that the fusion of EYFP accidentally changes electronic dynamics in this region thereby altered the transposition activity of the resulting fusion protein.

Unlike N-terminal EYFP-tagged transposases, fusion of EYFP to the C-termini of all forms (wild-type, N-terminal deleted, and those with amino acid substitutions) of

transposases resulted in reduced excision frequency, which is likely due to aggregation of the fusion proteins in the cytoplasm. EYFP protein alone did not form aggregates as indicated by the even distribution of EYFP in the entire yeast cells (Figure 3.5A), so it is likely the fusion at C-termini causes structural changes of the proteins which triggers aggregation. It was reported that a conserved domain at the C-terminus of the maize *Ac* transposase may be involved in dimerization of transposases during transposition (Essers et al., 2000). Meanwhile, when overexpressed, this domain was also involved in the formation of aggregates, resulting in inactive transposase. However, mutations in the most conserved amino acids did not abolish the dimerization property but those less conserved residues did. Comparison of the C-terminal sequences of Os3378-Z and *Ac* transposases did not reveal any common features. Nevertheless, it cannot be ruled out the conservation is at structural level not sequence level. Further investigation is needed in order to elucidate the mechanism underlying the aggregation of C-terminal tagged Os3378-Z transposases.

As shown in Figure 3.5A, the majority of C-terminal EYFP-tagged transposases was located in the cytoplasm. Since Os3378-Z-130 is still associated with a considerable level of excision activity when the C-terminus is tagged, it is obvious that the fusion did not fully abolish the function of NLS (Figure 3.5A, B). Given this fact, the location of the C-terminal EYFP-tagged proteins is more likely due to the aggregation, which prevents the transportation into nucleus, or the aggregation blocks the exposure of NLS or promotes the exposure of NES. Although the underlying mechanism remains unclear, this and other studies suggest there are multiple ways to induce aggregation of transposases that could readily trigger the suppression of their activity (Heinlein et al., 1994; Essers et al., 2000). This may be attributed to post-translational control of transposon activity.

3.4.8 Retention of Single TIR after Excision Suggests a Potential Mechanism for Formation of Pack-MULEs

The mechanism underlying sequence acquisition by Pack-MULEs or other DNA TEs remains an enigma. Talbert and Chandler (1988) proposed that *Mu* termini may function independently of their internal sequences. Specifically, a single *Mu* terminus may transpose alone and two separate but similar *Mu* termini may fortuitously insert into adjacent genic regions. If the two termini are close enough and in the right orientation, the entire sequence (two termini and the genes inside) could be moved together thus a Pack-MULE is formed. If this is the case, no direct repeats (TSD) would be generated in the flanking sequence of the initial Pack-MULE element (Figure 3.9). In this study, no mobility was observed for the single terminus of *Os3378*. However, analysis of the sequences at the donor site of *Os3378NA* revealed that retention of a single terminus after excisions is not rare for *Os3378* in yeast. A similar phenomenon has been observed in maize (Raizada et al., 2001). Seven out of 45 donor sites following excision of *RescueMu*, a modified *Mu1* element, contained terminal sequences with various lengths (41 to 211 bp out of 215 bp TIR), which was proposed to result from interrupted homology-dependent gap repair. Hence, retention of single terminal sequence following excisions is not restricted to yeast, and it is more likely a common phenomenon. The reason that no single terminus of *Os3378* was detected in rice is possibly due to its low activity, where barely any double-stranded breaks are available for this to happen. This prompted us to propose a "retention of single terminus mechanism" for Pack-MULE formation, which starts with the excision of a MULE element followed by formation of a single terminus at the donor site and in the meantime the same phenomenon happens to another MULE which shares similar terminal sequence. If the two elements are close to each other with single terminus in the appropriate orientation, and there is a gene between them, then a new Pack-MULE can be formed (Figure 3.9). The difference between

the “single terminus movement mechanism” and the “retention of single terminus mechanism” is that movement of single terminus is a requisite for new element formation for the former, which seems unlikely as evidenced from no excision of single terminus of *Os3378*. On the other hand, interrupted gap repair following excision of TEs seems relatively common, which renders the formation of single terminus with high possibility (Raizada et al., 2001). However, more analyses are needed to evaluate this mechanism.

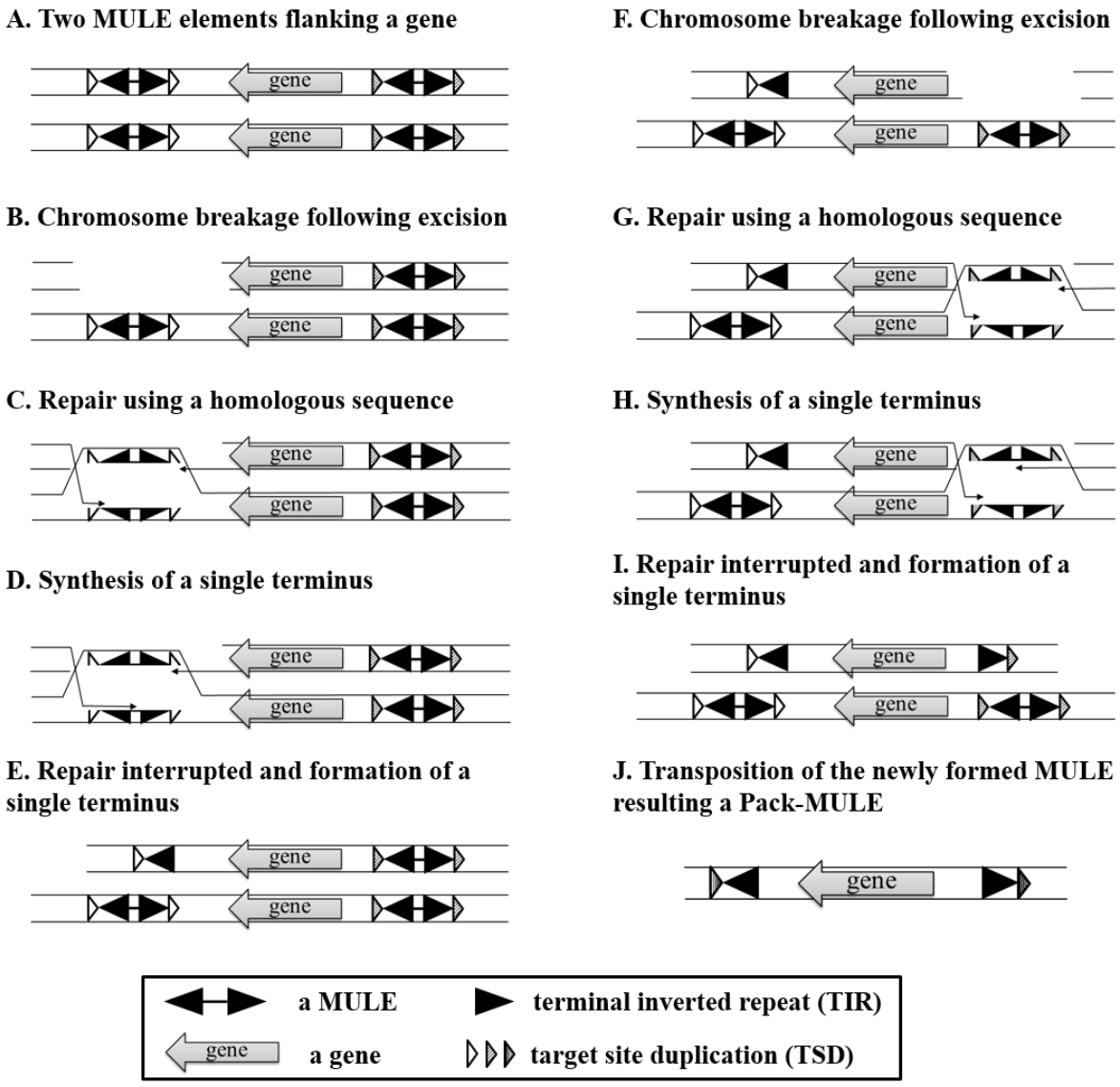

Figure 3.9 Schematic representation of a potential mechanism of Pack-MULE formation.

3.5 Materials and Methods

3.5.1 Cloning of the Coding Sequence of *Os3378* Transposase

Expression of *Os3378* was detected in a rice somaclonal mutant, Z418 (Gao, 2012). These rice plants were grown in growth chamber with temperature set at 28 ℃/24 ℃ (day/night) and a photoperiod of 14 h/10 h (light/dark). Two pairs of primers (Appendix Table 3.2) were used to amplify the coding sequence of *Os3378*. Total RNA was extracted from Z418 young panicles using TRIzol reagent (Invitrogen, Grand Island, NY, USA) followed by DNase treatment (RNase-free DNase set, Qiagen, Hilden, Germany) and purified using the RNeasy Mini Kit (Qiagen). RNA (4 μg) was reverse transcribed into cDNA using the GoScript Reverse Transcription system (Progema, Madison, WI, USA). The cDNA was used as template for amplification of the *Os3378* transcripts using the Platinum Taq DNA Polymerase High Fidelity kit (Invitrogen, Grand Island, NY, USA). PCR was performed with an initial denature step at 95 ℃ for 2 min, followed by 30 cycles of 94 ℃ 30 s, 58 ℃ 30 s, and 68 ℃ 1 min, and a final elongation step at 68 ℃ for 5 min. The resulting amplicons were sequenced at the Research Technology Support Facility (RTSF) of Michigan State University. The sequences of the two PCR products were concatenated to obtain the entire coding sequence of *Os3378*, which was referred to as Os3378-Z.

3.5.2 Computational Characterization of Functional Domains of the Os3378-Z Transposase

The coding-sequence of Os3378-Z transposase was translated into protein sequence using the ExPASy translation tool (http://web.expasy.org/translate/). Nuclear localization signal (NLS) was predicted using the PSORT II Prediction program (http://psort.hgc.jp/form2.html), which gave 4-residue and 7-residue NLS and the latter was presented in Figure 3.1B. DNA binding domain was defined through a combination of

secondary structure analysis (Jpred3, http://www.compbio.dundee.ac.uk/www-jpred/) and software prediction for helix-turn-helix DNA-binding motif (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hth.html). To locate the catalytic domain, multiple sequence alignment was performed on MURA-related transposases (*i.e.*, MURA, JITA, Os3378-Z) and the conserved triad amino acids was obtained through comparison to the defined "DDE" in the known proteins. Nuclear export signal (NES) was identified by manually checking Os3378-Z transposase for the conserved pattern of sequence (L–x(2,3)–[LIVFM]–x(2,3)–L–x–[LI]) (Bogerd et al., 1996).

A

	1st exon	1st intron	2nd exon	2nd intron	3rd exon	3rd intron	4th exon
Os3378-Z	1539	88	957	76	75	85	90
LOC_Os05g31510	1539	88	669	364	75	79	96

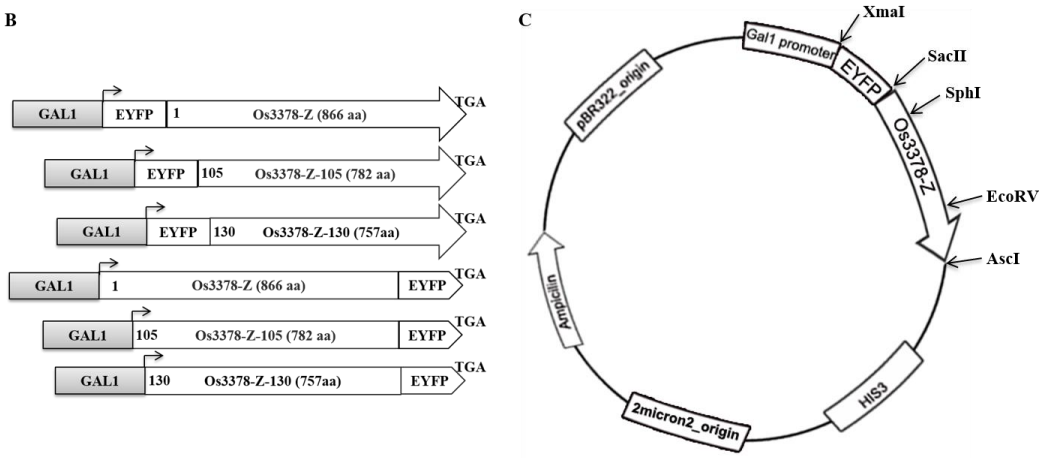

Figure 3.10 Os3378-Z transposase.

(A) The size (bps) of exons and introns of Os3378-Z and its most similar annotated locus in rice (Nipponbare).
(B) Schematic representation of full-length and N-terminal deleted Os3378-Z transposases with either N- or C-terminal EYFP tag.
(C) Schematic representation of restriction sites used in this study. *Xma*I is 5-bp upstream of the translation start site (TSS) of EYFP; *Sac*II is 14-bp upstream of the TSS of Os3378-Z transposase; *Sph*I is at 504-509 bp within Os3378-Z transposase; *Eco*RV is at 2469-2474 bp within Os3378-Z transposase; *Asc*I is 7-bp downstream of the translation termination site of Os3378-Z transposase.

3.5.3 Construction of Reporter and Expression Constructs

All the restriction sites used in this study were shown in Figure 3.10C. Sequences of all the constructs were verified by sequencing at the RTSF of Michigan State University.

3.5.3.1 Reporter Constructs

The reporter vector, pWL89a, was kindly provided by Nathan Hancock (University of Georgia). The non-autonomous *Os3378* (*Os3378NA*) was obtained by ligating two PCR fragments of the 5′ (213 bp) and 3′ (236 bp) TIR and sub-terminal sequences of the *Os3378* copy on chromosome 5 in Nipponbare (Gao, 2012) using Z418 genomic DNA as template. Two forms of reporter constructs were built, which differed in the insertion sites of *Os3378NA*. For one form, the *Os3378NA* was integrated into the promoter region of ADE2 through an *Xho*I restriction site, resulting in pWL89aOs3378NAXhoI. For the other form, the *Os3378NA* was integrated into the coding region of ADE2 through an *Hpa*I restriction site, resulting in pWL89aOs3378NAHpaI. To construct the reporter vector with a single terminus of *Os3378*, the 5′ terminal sequence was amplified with a pair of primers with *Hpa*I restriction sites in them. The resulting amplicons and pWL89a were digested with *Hpa*I and ligated together to form pWL89aOs3378singleTIRHpaI.

3.5.3.2 Expression Constructs of the Os3378-Z Transposase with N-terminal EYFP Tag

Gateway vector, pAG423GAL-EYFP-ccdB (Addgene plasmid 14341, EYFP: enhanced yellow fluorescent protein), was used to build the expression constructs. The coding sequence of the transposase obtained above was synthesized by Genscript (Piscataway, NJ, USA) and transferred to an entry vector (pENTR-D-TOPO, Invitrogen, Grand Island, NY, USA) through BR recombination reaction using the Gateway® BP Clonase™ Enzyme Mixes (Invitrogen, Grand Island, NY, USA). The final transfer of the transposase from the entry

vector to the expression vector was accomplished through LR recombination reaction using the Gateway® LR Clonase™ Enzyme Mixes (Invitrogen, Grand Island, NY, USA). The resulting construct was referred to as pAG423EYFPOs3378-Z. In this construct, the EYFP and Os3378-Z formed a fusion protein, whereby its transcription was controlled by the GAL1 promoter.

The constructs with N-terminal deleted Os3378-Z transposase were obtained by modifying pAG423EYFPOs3378-Z. The construct was digested using *Sac*II and *Sph*I, where *Sac*II recognizes a site between the EYFP and the transposase, and *Sph*I makes a single cut at 508 bp (170 aa) within the transposase (Figure 3.10C). The larger fragment consisting of the vector backbone and the partial transposase sequence (~170-886 residues) was purified. Primers containing *Sac*II and *Sph*I restriction sites and "ATG" start codon were designed to amplify 80-170 amino acids (aa) for Os3378-Z-80, 105 to 170 aa for Os3378-Z-105, 130-170 aa for Os3378-Z-130, and 161-170 aa for Os3378-Z-161 (Appendix Table 3.2). The amplicons were digested with *Sac*II and *Sph*I before ligating to the purified expression vector pre-digested by *Sac*II and *Sph*I. The resulting constructs, pAG423EYFPOs3378-Z-80, pAG423EYFPOs3378-Z-105, pAG423EYFPOs3378-Z-130, and pAG423EYFPOs3378-Z-161, lack the N-terminal 79, 104, 129, and 160 residues, respectively.

To prepare a construct containing EYFP with Os3378-Z-105-130 peptide (25 aa), pAG423EYFPOs3378-Z was digested with *Sac*II (between EYFP and Os3378-Z) and *Asc*I (7-bp downstream of the Os3378-Z stop codon) restriction enzymes to remove Os3378-Z transposase. The vector backbone was purified. Primers containing restriction sites (*Sac*II in forward primer and *Asc*I in reverse primer) were used to amplify the Os3378-Z 105-130 peptide and amplicons was digested with *Sac*II and *Asc*I followed by ligating to the purified vector fragment.

3.5.3.3 Expression Constructs of the Os3378-Z Transposase without EYFP Tag

Expression constructs without the EYFP-tag was obtained by removing the EYFP tag in the constructs prepared above. First, the constructs were digested with *Xma*I (5-bp upstream of the transcription start codon of EYFP) and *Sac*II (between EYFP and Os3378-Z) to release the EYFP tag (Figure 3.10C). Second, the sticky ends of the vectors were repaired using the Quick Blunting Kit (New England Biolabs, Ipswich, MA, USA) and a self-ligation reaction was incubated at 4 ℃ overnight. Standard bacteria (DH5α) transformation was conducted and positive clones were identified through restriction enzyme digestion. The resulting constructs were referred to as pAG423Os3378-Z, pAG423Os3378-Z-80, pAG423Os3378-Z-105, pAG423Os3378-Z-130, and pAG423Os3378-Z-161.

3.5.3.4 Expression Constructs of the Os3378-Z Transposase with C-terminal EYFP Tag

Os3378-Z transposase with C-terminal EYFP tag was obtained by transferring the EYFP sequence from the N-terminus to the C-terminus of the transposase in each construct. The construction of pAG423Os3378-ZEYFP (wild-type transposase with C-terminal EYFP) was used as an example to explain the detailed procedures. To initialize the construction, pAG423EYFPOs3378-Z was digested using *Xma*I and *Sac*II to separate the EYFP sequence and the rest of the construct. A filler DNA was used to seal the staggered ends of the construct and prevent the recovery of the *Xma*I site (Appendix Table 3.2). The vector DNA was self-ligated which resulted in pAG423Os3378NoXmaI. The original stop codon at the end of transposase was removed and new *Xma*I and *Sac*II sites were created downstream of the transposase by amplifying a segment of the transposase (2471-2658 bp) using a forward primer specific to 2471-2505 bp with an *Eco*RV site and a reverse primer complementary to the last 27 bp of the transposase (no stop codon, with an additional linker containing *Xma*I, *Sac*II, and *Asc*I sites). The resulting PCR product and pAG423Os3378NoXmaI were digested

using *Eco*RV and *Asc*I followed by ligating them together. The resulting sequence was then digested by *Xma*I and *Sac*II followed by ligating the EYFP sequence into it (pAG423Os3378-ZEYFP). Other C-terminal EYFP constructs (*e.g.*, pAG423Os3378-Z-105EYFP, pAG423Os3378-Z-130EYFP), were prepared in the same way.

3.5.3.5 Expression Constructs Containing Amino Acid Substitutions in Os3378-Z 105-129

Amino acid substitutions in the 105-129 peptide were made so that the peptide, as a whole piece, had different chemical and physiological properties (Figure 3.4A). These are Os3378-Z-105Ala with five acidic and two hydrophobic residues mutated to alanine, Os3378-Z-105Neutral with nine acidic residues changed to neutral aa, Os3378-Z-105Basic with nine acidic residues changed to basic aa, and Os3378-Z-105Hydrophobic with certain hydrophilic residues changed to hydrophobic (Figure 3.4A). Codons for these amino acids were chosen according to their corresponding usage frequency in the wild type Os3378-Z transposase. For each construct, a pair of primers was used to amplify 105-170 amino acid segment. The forward primer (~110 nt) contained a start codon ("ATG"), an alanine residue, mutated amino acids, and a *Sac*II restriction site while the reverse primer contains a *Sph*I restriction site (Appendix Table 3.2). pAG423Os3378-Z was used as the template for PCR amplification and amplicons were digested with *Sac*II and *Sph*I. Meanwhile, the expression constructs without EYFP tag were digested using the same restriction enzymes. The digested expression constructs and the amplicons were ligated, which resulted in four constructs, pAG423Os3378-Z-105Ala, pAG423Os3378-Z-105Neutral, pAG423Os3378-Z-105Basic, and pAG423Os3378-Z-105Hydrophobic.

3.5.3.6 Expression Constructs with C-terminal FLAG-His_6 Dual Tag

Expression constructs (without tag) generated above were digested using *Eco*RV and *Asc*I (Figure 3.10C), which cut off part of the C-terminal sequence of the transposases. This sequence was amplified using a pair of primers (Zanetti et al., 2005), where the forward primer contained an *Eco*RV restriction site and the reverse primer contains the FLAG-His_6 sequence and an *Asc*I site (Appendix Table 3.2). The resulting amplicons were digested with *Eco*RV and *Asc*I to generate corresponding ends, which was ligated to the pre-digested (with *Eco*RV and *Asc*I) expression constructs.

3.5.4 Transformation of Reporter and Expression Constructs into Yeast and Selection for ADE2 Revertants

The reporter construct was first transformed into the yeast haploid strain, DG2523 (*MATalpha ura3-167 trp 1-hisG leu2-hisG his3-del200 ade2-hisG*) (provided by Nathan Hancock and Sue Wessler). Yeast colonies growing on synthetic defined medium without uracil (SD/-Ura) were verified for the presence of the reporter vector by PCR. Competent cells of the confirmed yeast transformants were prepared using the Frozen-EZ Yeast Transformation II kit (Zymo Research Corporation, Irvine, CA) followed by transforming the expression constructs. As a control, the empty expression vector was also transformed into the yeast cells at the same time.

Transformed yeast cells were grown on SD plates lack uracil and histidine with 3% raffinose as the carbon source. After recovery for 6 to 7 days, colonies were plated on SD medium without uracil, histidine and adenine (SD/-Ura-His-Ade), and supplied with 3% raffinose and various amount of galactose (0.05%, 0.1%, 0.5%, 1%, 2%) to induce expression of the transposase. To measure the excision frequency, yeast colonies were suspended in 50

μL sterile dH_2O, where 49 μL was plated on SD/-Ura-His-Ade medium and 1 μL was diluted to 10^6-fold and 49 μL was plated on YPD plates to obtain the total number of viable cells in each colony. Excision of *Os3378NA* and restoration of a functional ADE2 gene allowed the growth of yeast cells on the SD/-Ura-His-Ade medium; the surviving colonies were called ADE2 revertants.

3.5.5 Determination of the Sequences of the Donor Site Following Excisions

PCR primers flanking the insertion sites of *Os3378NA* in ADE2 were used to amplify the sequences of the donor site following excisions. All PCR products were cloned into the pCR2.1-TOPO vector (Invitrogen, Grand Island, NY, USA) and sequenced.

3.5.6 Extraction of Yeast Total Protein and Determination of Protein Levels of Transposases

Yeast colonies containing both expression (those with FLAG-His_6 tag) and reporter constructs were streaked on SD/-Ura-His plates supplied with various galactose concentrations. After growing for 5 days, the cells were collected and lysed using extraction buffer (200 mM Tris pH8.0, 150 mM ammonium sulfate, 10% glyceol, 1 mM EDTA, 0.2 mM dithiothreitol, 10 mM phenylmethanesulfonyl fluoride, and 1X Yeast/Fungal Protease Arrest, BD Biosciences, San Jose, CA, USA). Glass beads were added into the mixture followed by vigorous vortexing to break the cells. Centrifugation was used precipitate cell debris and the supernatant contained the total crude proteins. For transposases whose transcription was induced by various galactose concentrations, 2.5 μg total protein extract for Os3378-Z and 5 μg for Os3378-Z-130 was used. For all forms of transposes under 2% galactose level, 2.5 ug crude protein extract were used. To ensure the usage of similar amount of total protein, these protein extracts were separated on a 4-10% SDS-PAGE gel (Invitrogen,

Grand Island, NY, USA) and developed using the Pierce Silver Stain Kit (Thermo Scientific, Waltham, MA, USA). For Western blot analysis, these protein extracts were separated on 4-10% SDS-PAGE gel followed by protein transfer to PVDF transfer membrane (Millipore, Billerica, MA, USA). The membrane was first probed using the first antibody, THE His tag antibody (Mouse) (Genscript, Piscataway, NJ, USA) and then second antibody, anti-mouse IgG, horseradish peroxidase-conjugated (Cell Signaling Technology, Danvers, MA). Signal detection was performed using the SuperSignal West Pico Chemiluminesent Substrate (Thermo Scientific, Waltham, MA, USA) and exposed to a film. Relative protein levels (relative density) of various forms of transposases were quantified using the ImageJ software (http://imagej.nih.gov/ij/).

3.5.7 Determination of the Cellular Localization of Os3378-Z Transposases

The cellular localization of the transposases was determined using confocal microscopy in the Centre for Advanced Microscopy at Michigan State University. Yeast cells containing both expression (those with EYFP tag) and reporter constructs were cultured in SD/-Ura-His medium containing 1% galactose at 30 ℃ and 250 rpm shaking overnight. The culture was diluted to ~0.4 OD600 units and incubated at 30 ℃ at 250 rpm for 3 h before collecting the cells by centrifugation (3000 rpm). Cells were washed using 1X PBS buffer (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4) and pelleted before treating cells with 1X PBS containing 1 μg/mL^{-1} DAPI (4',6-Diamidino-2-Phenylindole, Dihydrochloride), a fluorescent stain for the nucleus. Confocal images were acquired using an Olympus FluoView FV1000 Laser Scanning Confocal Microscope (Center Valley, PA) configured on an automated IX81 inverted microscope with a 100X UPLSAPO (NA 1.4) oil objective. EYFP fluorescence was excited with the 515nm Argon laser line, while emission was collected with a 535-565 nm band pass filter. Nuclei were counterstained blue with

DAPI excited with a 405nm diode laser while emission was collected with a 430-470 nm band pass filter.

3.5.8 Analysis of Reinsertions of *Os3378NA* Following Excisions

DNA blot assay was conducted to determine reinsertions of *Os3378NA* after excisions. For EYFPOs3378-Z-105 and EYFPOs3378-Z-130, nineteen yeast colonies for each galactose concentration were tested except 0.05% at which few excision events occurred. Genomic DNA of ADE2 revertant yeast colonies were extracted using Zymolase digestion of yeast cell wall, PCIAA (phenol: chloroform: isoamyl alcohol) and isopropanol precipitation method. Genomic DNA (200 ng) was digested with BamHI and XhoI, where the latter recognizes a sequence within *Os3378NA*. The digested DNA was resolved on 1% agarose gel for 3 h, followed by transfer of DNA from the gel to positively charged nylon membrane (GE Healthcare, Pittsburgh, PA, USA) using capillary flow overnight. The digoxigenin-labeled probe specific to the 5′ TIR of *Os3378* was used to bind to the fragments of *Os3378NA*, which was detected by anti-Digoxigenin-AP, Fab fragments (Roche Applied Science, Indianapolis, IN, USA).

To determine the reinsertion locations, transposon display (a modified AFLP method) was conducted. Specifically, genomic DNA of ADE2 revertant yeast colonies were digested using HinP1I restriction enzyme and adapters were ligated to these genomic fragments. Nested PCRs were conducted using one primer specific to the TIR sequence of *Os3378* and another to the adapter sequence, which resulted in amplicons consisting of part of the TIR and flanking sequences of the reinsertion sites. The resulting products were resolved on a polyacrylamide gel and polymorphic fragments were recovered by PCR and sequenced. The flanking sequences were mapped to the yeast genome (S288C, release R64-1-1, http://yeastgenome.org/) and the reinsertion locations were determined with regard to the

closest genomic features (such as genes, rRNAs based on the file containing genomic sequence features, coordinates and annotations, which was downloaded in February, 2011 from http://yeastgenome.org/). Insertions that are within 1 kb from genes were considered to be in flanking regions. If the insertion are between two genes which were in the opposite transcription orientations ("head-to-head" or "tail-to-tail"), the insertions were considered to be in either the 5′ or 3′ regions. Otherwise, they were assigned to the group of "5′ and 3′ regions" ("head-to-tail" or "tail-to-head").

APPENDIX

APPENDIX

Table 3.1 Reinsertion sites.

Coordinate of insert site	Coordinates of adjacent genomic feature (A)	Insert position (A)	Coordinates of adjacent genomic feature (B)	Insert position (B)
chrXII_NC_001144: 502262	chrXII_501260_502162	5′ region	chrXII_502421_504247	5′ region
chrXIII_NC_001145: 56408	chrXIII_55265_56269	5′ region	chrXIII_56773_57453	5′ region
chrXVI_NC_001148: 539082	chrXVI_535820_538936	5′ region	chrXVI_539385_541970	5′ region
chrVII_NC_001139: 433848	chrVII_433214_433579	5′ region	chrVII_435625_436351	5′ region
chrXIII_NC_001145: 307476	chrXIII_307489_308682	5′ region	chrXIII_303236_305593	5′ region
chrXV_NC_001147: 710335	chrXV_710446_711576	5′ region	chrXV_710201_710272	3′ region
chrI_NC_001133: 71592	chrI_71786_73288	5′ region	chrI_70257_70489	NA
chrII_NC_001134: 414051	chrII_411054_413981	3′ region	chrII_414186_415261	3′ region
chrIII_NC_001135: 130688	chrIII_130745_131542	3′ region	chrIII_128470_130281	3′ region
chrV_NC_001137: 295069	chrV_295301_295732	3′ region	chrV_293050_294813	3′ region
chrV_NC_001137: 339527	chrV_337949_339472	3′ region	chrV_339864_342167	3′ region
chrV_NC_001137: 79571	chrV_79636_79977	3′ region	chrV_78053_79456	3′ region
chrXIV_NC_001146: 352777	chrXIV_352820_355042	3′ region	chrXIV_352414_352530	3′ region
chrIII_NC_001135: 125623	chrIII_126011_126730	3′ region	chrIII_124134_124465	NA
chrIII_NC_001135: 172874	chrIII_172950_173440	3′ region	chrIII_170886_172424	5′ region
chrXII_NC_001144: 143042	chrXII_141073_142923	3′ region	chrXII_143201_146041	5′ region
chrXIII_NC_001145: 819953	chrXIII_818827_819948	3′ region	chrXIII_820256_822454	5′ region
chrXIV_NC_001146: 37497	chrXIV_34696_37422	3′ region	chrXIV_37700_38554	5′ region
chrXIV_NC_001146: 731134	chrXIV_731618_733303	3′ region	chrXIV_728426_730186	5′ region
chrVII_NC_001139: 982227	chrVII_979765_982068	5′ region	chrVII_982482_984272	3′ region
chrXI_NC_001143: 557534	chrXI_557677_558954	5′ region	chrXI_556518_557342	3′ region
chrIV_NC_001136: 1422659	chrIV_1422763_1424688	5′ region	chrIV_1421157_1422485	3′ region
chrIV_NC_001136: 431604	chrIV_431108_431517	5′ region	chrIV_432308_432631	3′ region
chrV_NC_001137: 347619	chrV_343320_347612	5′ region	chrV_347912_348400	3′ region
chrVII_NC_001139: 203990	chrVII_203914_204155	ARS	NA	NA
chrXI_NC_001143: 196257	chrXI_196042_196287	ARS	NA	NA
chrI_NC_001133: 149952	chrI_147594_151166	gene	NA	NA
chrV_NC_001137: 246198	chrV_243810_246503	gene	NA	NA
chrVI_NC_001138: 135910	chrVI_134521_137157	gene	NA	NA
chrVIII_NC_001140: 349931	chrVIII_349574_352453	gene	NA	NA
chrX_NC_001142: 566066	chrX_559416_566828	gene	NA	NA
chrXV_NC_001147: 297465	chrXV_297078_298838	gene	NA	NA
chrIV_NC_001136: 216460	chrIV_216158_216489	gene	NA	NA
chrXII_NC_001144: 456483	chrXII_455933_457732	rRNA	NA	NA
chrXII_NC_001144: 457514	chrXII_455933_457732	rRNA	NA	NA
chrXII_NC_001144: 457574	chrXII_455933_457732	rRNA	NA	NA
chrXII_NC_001144: 458146	chrXII_457733_458432	rRNA	NA	NA
chrXII_NC_001144: 458212	chrXII_457733_458432	rRNA	NA	NA
chrXII_NC_001144: 288	chrXII_76_5661	telomere	NA	NA

Table 3.2 Primers used in this study.

Primers for amplifying Os3378-Z coding sequence (5’-3’)	
Os3378ORF-F	CACCATGGATAATTTGGATGTACTTTG
Os3378Exon1R	GAGGATCAATATTCAGATCTTTTTGAAG
Os3378Exon2-4F	TGAATATTGATCCTCATGGAGCTG
Os3378ORF-R	TCACTTCTTTGTTGTGTTTGTTCTTG
Primers for cloning the artificial non-autonomous Os3378 (Os3378NA) (5’-3’)	
Os3378AHpaIF	CAGTTAACTTAATTTAAGGAAAAAGTCAGTTTTACTCC
Os3378AR	AATCTCCCAGGTTCTCTTTCTCCACG
Os3378BF	CCTCATCACCTGGCAAGTAAAACCTG
Os3378BHpaIR	GTGTTAACTTAAATTAAGGAAAAAGTCCGTTTTACTCC
Os3378AXhoIF	GTTAACGGAAAAAGTCAGTTTTACTCCCCTCAAGTATG
Os3378BXhoIR	GTTAACGGAAAAAGTCAGTTTTACTCCCCTCAAGTATG
Primers for amplifying the donor site following excision of Os3378NA (sequences were obtained from Yang et al., 2006) (5’-3’)	
ADE2XhoI-F	CTGACAAATGACTCTTGTTGCAGGGCTACGAAC
ADE2XhoI-R	TGGAAAAGGAGCCATTAACGTGGTCATTGGAG
ADE2HpaI-F	CTTTGTACGCCGAAAAATGG
ADE2HpaI-R	CGCATAACATAAGTCACAAAT
Primers for cloning partial N-terminal sequence of the Os3378-Z transposase (5’-3’)	
Os3378A80F	ATAATACCCGGGATGCTATCAGATGATAAGGAATGC
Os3378A105F	TCGTCACCGCGGCAATGGCTATACTCAAGCACAACGAAG
Os3378A130F	TCGTCACCGCGGCAATGGCTATTGAGGAAGAATATGAC
Os3378AANterR	CTTGTTGGCATGCTCTCTATTTTCTTGTGCTTCCTC
Os3378A168linkerF	GGCCGCCCCCTTCACCATGGAGCATG
Os3378A168linkerR	CTCCATGGTGAAGGGGGCGGCCGC
Primers for mutating some amino acids between Os3378-Z 105-130 region(5’-3’)	
A105-130AlaF	ATAATACCCGGGATGGCTATAGCCAAGCACAACGAAGCCGCCAATGTTGCACTTAGCA ATGGAGTCAGTGATGCAAGTGATGTTGCCGCCGACATTGAGGAAGAATATGAC
A105-130NeutralF	ATAATACCCGGGATGATACTCAAGCACAACCAAAATAACAATGTTGTGCTTAGCAATG GAGTCAGTAATAACAGTAATGTTCAAAACAATATTGAGGAAGAATATGACAGTC
A105-130BasicF	ATAATACCCGGGATGATACTCAAGCACAACCGCAAGAAAAATGTTGTGCTTAGCAATG GAGTCAGTAAGAAAAGTAAGGTTCGCAAGAAAATTGAGGAAGAATATGACAGTC
A105-130HydrophobicF	ATAATACCCGGGATGATACTCATCTTCCTCGAAGATGATCTGGTTGTGCTTAGCATCGG AGTCGTGGATCTCGTAGATGTTGAAATCGACATTGAGGAAGAATATGACAGTC
Primers used to clone Os3378-Z transposase with C-terminal EYFP tag (5’-3’)	
XmaISacIIfillF	CCGGACAAAGGTGAAGC
XmaISacIIfillR	TTCACCTTTGT
2471EcoRV-F	GACCTAGATATCAGTACAAGAGAGAGCAAGAGAGGTTCAGG
3378C-XSacIIR	TGGGTCGGCGCGCCTCATGCCGCGGATCCTCATTTTCTCACCGCCCTACCATTGCTCTT TGTCCCGGGCCTTCTTTGTTGTGTTTGTTCTTGCAAC
Primers used to clone Os3378-Z transpoase with C-terminal FLAG-His_6 tag (5’-3’)	
2471EcoRV-F	GACCTAGATATCAGTACAAGAGAGAGCAAGAGAGGTTCAGG
HF-CtermR	TGGGTCGGCGCGCCTCAATGGTGATGATGGTGATGACCTCCACCCTTATCATCATCAT CCTTATAATCTCCACCTCCACCTCCACCTCCCATCTTCTTTGTTGTGTTTGTTCTTGCAAC

REFERENCES

REFERENCES

Bennetzen JL, Swanson J, Taylor WC, Freeling M (1984) DNA insertion in the 1st intron of maize *Adh1* affects message levels: Cloning of progenitor and mutant *Adh1* alleles. Proc. Natl. Acad. Sci. USA **81:** 4125-4128

Bogerd HP, Fridell RA, Benson RE, Hua J, Cullen BR (1996) Protein sequence requirements for function of the human T-cell leukemia virus type 1 Rex nuclear export signal delineated by a novel in vivo randomization-selection assay. Molecular and Cellular Biology **16:** 4207-4214

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*. Molecular Biology and Evolution **20:** 1362-1375

Chen CH, Oishi KK, Kloeckenergruissem B, Freeling M (1987) Organ-specific expression of maize *Adh1* is altered after a *Mu* transposon insertion. Genetics **116:** 469-477

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

Cresse AD, Hulbert SH, Brown WE, Lucas JR, Bennetzen JL (1995) *Mu1*-related transposable elements of maize preferentially insert into low copy number DNA. Genetics **140:** 315-324

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

Emelyanov A, Gao Y, Naqvi NI, Parinov S (2006) Trans-kingdom transposition of the maize *Dissociation* element. Genetics **174:** 1095-1104

Enoki H, Izawa T, Kawahara M, Komatsu M, Koh S, Kyozuka J, Shimamoto K (1999) *Ac* as a tool for the functional genomics of rice. Plant Journal **19:** 605-613

Essers L, Adolphs RH, Kunze R (2000) A highly conserved domain of the maize *Activator* transposase is involved in dimerization. Plant Cell **12:** 211-223

Feldmar S, Kunze R (1991) The ORFa protein, the putative transposase of maize transposable element *Ac*, has a basic DNA-binding domain. EMBO Journal **10:** 4003-4010

Ferguson A, Zhao D, Jiang N (2013) Selective acquisition and retention of genomic sequences by Pack-MULEs based on GC content and breadth of expression. Plant Physiology **163:** 1419-1432

Ferguson AA, Jiang N (2012) *Mutator*-like elements with multiple long terminal inverted repeats in plants. Comparative and Functional Genomics **2012:** 695827 doi:10.1155/|2012/695827

Fernandes J, Dong QF, Schneider B, Morrow DJ, Nan GL, Brendel V, Walbot V (2004) Genome-wide mutagenesis of *Zea mays* L. using *RescueMu* transposons. Genome Biology **5**

Fischer SEJ, Wienholds E, Plasterk RHA (2001) Regulated transposition of a fish transposon in the mouse germ line. Proc. Natl. Acad. Sci. USA **98:** 6759-6764

Fleenor D, Spell M, Robertson D, Wessler S (1990) Nucleotide sequence of the maize *Mutator* element, *Mu8*. Nucleic Acids Research **18:** 6725-6725

Freeling M (1984) Plant Transposable Elements and Insertion Sequences. Annual Review of Plant Physiology and Plant Molecular Biology **35:** 277-298

Gao DY (2012) Identification of an active *Mutator*-like element (MULE) in rice (*Oryza sativa*). Molecular Genetics and Genomics **287:** 261-271

Grevelding C, Becker D, Kunze R, Vonmenges A, Fantes V, Schell J, Masterson R (1992) High-rates of *Ac/Ds* germinal transposition in *Arabidopsis* suitable for gene isolation by insertional mutagenesis. Proc. Natl. Acad. Sci. USA **89:** 6085-6089

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

Hancock CN, Zhang F, Wessler SR (2010) Transposition of the *Tourist*-MITE *mPing* in yeast: An assay that retains key features of catalysis by the class 2 *PIF/Harbinger* superfamily. Mobile DNA **1:** 5 doi:10.1186/1759-8753-1-5

Heinlein M, Brattig T, Kunze R (1994) *In vivo* aggregation of maize *Activator* (*Ac*) transposase in nuclei of maize endosperm and *Petunia* protoplasts. Plant Journal **5:** 705-714

Hennig S, Ziebuhr W (2010) Characterization of the transposase encoded by IS*256*, the prototype of a major family of bacterial insertion sequence elements. Journal of Bacteriology **192:** 4153-4163

Hershberger RJ, Warren CA, Walbot V (1991) *Mutator* activity in maize correlates with the presence and expression of the *Mu* transposable element *Mu9*. Proc. Natl. Acad. Sci. USA **88:** 10198-10202

Ivics Z, Hackett PB, Plasterk RH, Izsvak 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 ZR, Zhang XY, 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

Keng VW, Yae K, Hayakawa T, Mizuno S, Uno Y, Yusa K, Kokubu C, Kinoshita T, Akagi K, Jenkins NA, et al. (2005) Region-specific saturation germline mutagenesis in mice using the *Sleeping Beauty* transposon system. Nature Methods **2:** 763-769

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998) Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Research **8:** 464-478

Kim YC, Morrison SL (2009) N-terminal domain-deleted *Mu* transposase exhibits increased transposition activity with low target site preference in modified buffers. Journal of Molecular Microbiology and Biotechnology **17:** 30-40

Knapp S, Coupland G, Uhrig H, Starlinger P, Salamini F (1988) Transposition of the maize transposable element *Ac* in *Solanum tuberosum*. Molecular & General Genetics **213:** 285-290

Kunze R, Behrens U, Couragefranzkowiak U, Feldmar S, Kuhn S, Lutticke R (1993) Dominant transposition-deficient mutants of maize *Activator* (*Ac*) transposase. Proc. Natl. Acad. Sci. USA **90:** 7094-7098

Kunze R, Kuhn S, Jones JDG, Scofield SR (1995) Somatic and germinal activities of maize *Activator* (*Ac*) transposase mutants in transgenic tobacco. Plant Journal **8:** 45-54

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology **157:** 105-132

Lampe DJ, Akerley BJ, Rubin EJ, Mekalanos JJ, Robertson HM (1999) Hyperactive transposase mutants of the *Himar1* mariner transposon. Proc. Natl. Acad. Sci. USA **96:** 11428-11433

Lazarow K, Du ML, Weimer R, Kunze R (2012) A hyperactive transposase of the maize transposable element *Activator* (*Ac*). Genetics **191:** 747-756

Li YB, Harris L, Dooner HK (2013) *TED*, an autonomous and rare maize transposon of the *Mutator* superfamily with a high gametophytic excision frequency. Plant Cell **25:** 3251-3265

Lisch D (2002) *Mutator* transposons. Trends in Plant Science **7:** 498-504

Lisch D, Chomet P, Freeling M (1995) Genetic characterization of the *Mutator* system in maize: Behavior and regulation of *Mu* transposons in a minimal line. Genetics **139:** 1777-1796

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

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

Liu SZ, Yeh CT, Ji TM, Ying K, Wu HY, 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 doi:10.1371/journal.pgen.1000733

Makris JC, Nordmann PL, Reznikoff WS (1990) Integration host factor plays a role in IS*50* and Tn*5* transposition. Journal of Bacteriology **172:** 1368-1373

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

McCarty DR, Suzuki M, Hunter C, Collins J, Avigne WT, Koch KE (2013) Genetic and molecular analyses of *UniformMu* transposon insertion lines. Plant Transposable Elements: Methods and Protocols **1057:** 157-166

Nesmelova IV, Hackett PB (2010) DDE transposases: Structural similarity and diversity. Advanced Drug Delivery Reviews **62:** 1187-1195

Neuveglise C, Chalvet F, Wincker P, Gaillardin C, Casaregola S (2005) *Mutator*-like element in the yeast *Yarrowia lipolytica* displays multiple alternative splicings. Eukaryotic Cell **4:** 615-624

Plasterk RHA, Izsvak Z, Ivics Z (1999) Resident aliens: The *Tc1/mariner* superfamily of transposable elements. Trends in Genetics **15:** 326-332

Pledger DW, Coates CJ (2005) Mutant *Mos1 mariner* transposons are hyperactive in *Aedes aegypti*. Insect Biochemistry and Molecular Biology **35:** 1199-1207

Qin M, Robertson DS, Ellingboe AH (1991) Cloning of the *Mutator* transposable element *MuA2*, a putative regulator of somatic mutability of the *a1-Mum2* allele in maize. Genetics **129:** 845-854

Raizada MN, Nan GL, Walbot V (2001) Somatic and germinal mobility of the *RescueMu* transposon in transgenic maize. Plant Cell **13:** 1587-1608

Raizada MN, Walbot V (2000) The late developmental pattern of *Mu* transposon excision is conferred by a cauliflower mosaic virus 35S-driven MURA cDNA in transgenic maize. Plant Cell **12:** 5-21

Robertson DS (1978) Characterization of a mutator system in maize. Mutation Research **51:** 21-28

Rossi M, Araujo PG, de Jesus EM, Varani AM, Van Sluys MA (2004) Comparative analysis of *Mutator*-like transposases in sugarcane. Molecular Genetics and Genomics **272:** 194-203

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science **326:** 1112-1115

Settles AM, Holding DR, Tan BC, Latshaw SP, Liu J, Suzuki M, Li L, O'Brien BA, Fajardo DS, Wroclawska E, et al. (2007) Sequence-indexed mutations in maize

using the *UniformMu* transposon-tagging population. BMC Genomics **8:** 116 doi:10.1186/1471-2164-8-116

Singer T, Yordan C, Martienssen RA (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

Talbert LE, Chandler VL (1988) Characterization of a highly conserved sequence related to *Mutator* transposable elements in maize. Molecular Biology and Evolution **5:** 519-529

Talbert LE, Patterson GI, Chandler VL (1989) *Mu* transposable elements are structurally diverse and distributed throughout the genus *Zea*. Journal of Molecular Evolution **29:** 28-39

Tan BC, Chen ZL, Shen Y, Zhang YF, Lai JS, Sun SSM (2011) Identification of an active new *Mutator* transposable element in maize. G3: Genes Genomes Genetics **1:** 293-302

Weil CF, Kunze R (2000) Transposition of maize *Ac/Ds* transposable elements in the yeast *Saccharomyces cerevisiae*. Nature Genetics **26:** 187-190

Woodhouse MR, Freeling M, Lisch D (2006) The *mop1* (*mediator of paramutation1*) mutant progressively reactivates one of the two genes encoded by the *MuDR* transposon in maize. Genetics **172:** 579-592

Xu ZN, Yan XH, Maurais S, Fu HH, O'Brien DG, Mottinger J, Dooner HK (2004) *Jittery*, a *Mutator* distant relative with a paradoxical mobile behavior: Excision without reinsertion. Plant Cell **16:** 1105-1114

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

Yang GJ, Zhang F, Hancock CN, Wessler SR (2007) Transposition of the rice miniature inverted repeat transposable element *mPing* in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **104:** 10962-10967

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

Zanetti ME, Chang IF, Gong FC, Galbraith DW, Bailey-Serres J (2005) Immunopurification of polyribosomal complexes of *Arabidopsis* for global analysis of gene expression. Plant Physiology **138:** 624-635

Zayed H, Izsvak Z, Walisko O, Ivics Z (2004) Development of hyperactive *Sleeping Beauty* transposon vectors by mutational analysis. Molecular Therapy **9:** 292-304

Zhao DY, Jiang N (2014) Nested insertions and accumulation of indels are negatively correlated with abundance of *Mutator*-like transposable elements in maize and rice. PLoS ONE **9:** e87069 doi:10.1371/journal.pone.0087069

Zhao ZY, Sundaresan V (1991) Binding-sites for maize nuclear proteins in the terminal inverted repeats of the *Mu1* transposable element. Molecular & General Genetics **229:** 17-26

CHAPTER 4

Insertions of *Mutator*-Like Transposable Elements and their Impact on Gene Expression in the Illinois Long-Term Selection Experiment Maize Strains

4.1 Abstract

Mutator-like elements (MULEs) belong to a highly mutagenic and active DNA transposon family. MULEs have the propensity to insert into low copy and genic regions, which may influence the expression of adjacent genes as well as the relevant developmental processes. The Illinois Long-Term Selection Experiment (ILTSE) maize strains are elite experimental materials for a variety of studies, especially studies on the effects of selection on the kernel chemical composition (oil or protein). However, little is known about the molecular mechanism underlying the selection. Particularly, it is not clear whether transposons have played any role in the artificial selection. To this end, MULE insertions that are co-segregating with either high or low protein maize strains (IHP and ILP) were studied. Out of 200 co-segregating insertions detected, 191 produced significant alignments with the B73 maize genomic sequence. Consistent with previous findings, ~79% of the 191 insertions were located in low-copy genomic regions. Interestingly, similar numbers of co-segregating insertions were located at the 5′ and 3′ regions of protein-coding genes, which is in contrast to the 5′ region preference of MULEs reported in previous studies. Compared with that in the B73 genome, the fraction of co-segregating MULE insertions in exons is over one fold higher, suggesting exonic insertions may be involved in the selection. Additionally, expression levels of genes with adjacent co-segregating MULE insertions in immature kernels and leaves were compared between the IHP and ILP maize strains. In immature kernels, ten out of 55 genes (~18%) associated with MULE insertions showed reduced expression levels and five (~9%) showed enhanced expression levels compared to the alleles without MULE insertions. Eight of the differentially expressed genes were further tested in young leave, where half exhibited similar expression pattern as that in immature kernels but the alteration associated with MULE insertions was not observed with the other half of the genes, suggesting they might be involved in the selected trait.

4.2 Introduction

Transposable elements (TEs), also called transposons, are genomic sequences that can move from one place to another within a genome. Based on the transposition intermediate, TEs can be divided into two classes; class I (RNA elements) or retrotransposons which transpose via RNA intermediates and class II or DNA transposons which transpose via DNA intermediates. The availability of sequenced genomes and bioinformatics analyses reinforce the concept that TEs not only contribute to genome size variations, but also have functions in many biological processes in both plants and animals (Feschotte and Pritham, 2007; Feschotte, 2008; Martinez and Slotkin, 2012).

Mutator elements, which belong to the DNA transposons, represent a highly mutagenic transposon system. They are characterized by the long terminal inverted repeats (TIRs) and 9-11 bp duplication of the target site sequence (TSD) (Lisch, 2002). They were first discovered in a maize stock which exhibited higher mutation frequency than spontaneous mutation (Robertson, 1978). Subsequently, similar elements were found to be widespread in other plant species, including both monocots and dicots, where they were referred to as *Mutator*-like transposable elements (MULEs) (Lisch and Jiang, 2009). Lisch and colleagues demonstrated that the autonomous MULEs were widespread in grasses by using DNA-blot hybridization and PCR amplification and these elements are under purifying selection which suggests potential activity (Lisch et al., 2001). Pack-MULEs are MULEs carrying gene(s) or gene fragments. In rice, about 3,000 Pack-MULEs were discovered (Jiang et al., 2004) and some were suggested to have functions in regulation of gene expression (Hanada et al., 2009). In maize, there are a total of 262 Pack-MULEs among 12,900 MULEs in the B73 genome (Schnable et al., 2009).

TEs can regulate gene expression in many ways, at both the transcriptional and post-transcriptional levels (Feschotte, 2008). At the transcriptional level, TEs can introduce new *cis*-regulatory elements (*e.g.*, promoters, alternative transcription start sites or polyadenylation sites), disrupt existing *cis*-regulatory elements, interfere with gene expression by antisense transcription of the inserted TEs and induce heterochromatin formation. At the post-transcriptional level, TEs can serve as binding sites for miRNA or RNA binding proteins, and so forth. Increasing evidence indicates that a large number of class II TEs, including MULEs, are transcribed and their position and transcription can also influence the expression of other genes (Diao and Lisch, 2006; Jiao and Deng, 2007; Naito et al., 2009). The massive amplification of *mPing*, the first active miniature inverted transposable element (MITE) to be discovered, led to unexpected consequences on gene expression in the rice strain EG4 (Naito et al., 2009). The vast majority of *mPing* insertions either led to up-regulation or no detectable changes on the expression of near-by genes. In contrast, a study of 4606 TEs (including both RNA and DNA TEs) in *Arabidopsis* demonstrated that genes close to TE insertions were less expressed compared with the genome-wide pattern of gene expression (Hollister and Gaut, 2009), suggesting the effect of TEs on gene expression could be either positive or negative. Like other TEs, MULEs can regulate transcription of adjacent genes through the mechanisms mentioned above. For example, Bennetzen and colleagues detected reduced expression level of the mutant allele of *alcohol dehydrogenase-1* which had a *Mu1* insertion in its first intron compared to that of the wild type allele (Bennetzen et al., 1984). Moreover, in some maize mutagenesis projects targeting specific metabolic pathways by using *Mu* as the mutagen, it is evident that the insertion(s) of this element greatly altered, and can even diminish the expression of related genes and disrupt respective metabolic pathways (Blauth et al., 2001; Blauth et al., 2002; Cossegal et al., 2008).

The Illinois Long-Term Selection Experiment (ILTSE) was initiated in 1896, in which kernel protein and oil content were selected (Moose et al., 2004; Dudley, 2007). The selection procedure and criteria varied throughout the 100-year experiment. However, the main strategy was similar, which involved cross-pollination within populations with a similar trait (see Figure 4.1A) and selection for the target trait from within the population (*i.e.*, protein or oil). As shown in Figure 4.1A, the experiment was initialized with 163 ears (organ in which maize seeds develop) from the open-pollinated (which allows self-pollination and cross-pollination) variety, Burr's white. The ears with the highest protein content (if protein was the target trait) served as the starting material for the high protein population (IHP) and the ears with the lowest protein content served as the starting material for the low protein population (ILP). During the first 9 cycles, plants within the same population were grown together and seeds were derived from open-pollination. Subsequently, cross-pollination was enforced by de-tasseling (removing the male flower) plants in alternate rows within the same population and seeds were harvested only from the de-tasseled plants. After each harvest, ears with the highest protein content from IHP block and ears with the lowest protein content from ILP block were used for next cycle of experiment and selection. As a result, genetic variability was mainly introduced from different individual lines within the same population. Nevertheless, due to the fact that maize is an out-crossing plant, and that the populations were not grown in absolute isolation, occasional gene introgression from other populations or varieties during the selection process cannot be ruled out.

Over 100 years of selection has created four major maize populations with extreme kernel protein and oil content, *i.e.*, Illinois High Protein (IHP), Illinois Low Protein (ILP), Illinois High Oil (IHO) and Illinois Low Oil (ILO), which have been a major source for elite experimental materials in plant breeding and genetics research (Moose et al., 2004). The long

term divergent selection for kernel protein content increased the mean protein content from 10.93% to 30.92% in the IHP population while that in the ILP population decreased from 10.93% to 4.26% (Moose et al., 2004). The continual gain of selection in these maize populations suggests that a large amount of genetic variability exists, providing a unique set of germplasm to understand the underlying mechanisms for continued variability and selection potential.

Previous studies focused on identifying molecular markers and quantitative trait loci (QTLs) in these populations by using crosses of IHP × ILP or random-mated lines (Goldman et al., 1993; Dudley et al., 2004; Wassom et al., 2008). However, different results were obtained with the different methods and locations used in the research. In one study, 6 loci accounted for over 60% of the total kernel protein content variation in the Illinois protein populations (Goldman et al., 1993). In another study, ~173 loci were suggested to be responsible for the variation in protein content in IHP versus ILP at cycle 90 (Dudley and Lambert, 1992). Long-term selection would increase coupling phase linkage and lead to high local linkage disequilibrium of genomic regions that contribute to the selected trait (Moose et al., 2004). To break the linkage blocks and improve QTL studies in these maize populations, random mating between IHP and ILP was used, which yielded at least 40 QTLs with small effects for protein, starch, oil, and kernel weight separately, which is consistent with the "Infinitesimal Model", *i.e.*, a large number of QTLs contributing to the trait with each constituting a marginal effect (Dudley, 2007).

Despite many years of classic genetic studies on the ILTSE maize strains, little research on possible influence of transposable elements on these strains has been reported. To date, there is only one report studying an insertion of a DNA TE called *ILS_1* in the ILTSE maize strains (Alrefai et al., 1994). The *ILS-1* element inserted in the 13th intron of the *Shrunken2*

gene in the ILP maize. Nevertheless, there is no evidence showing whether the insertion is related to the selected trait. The abundance of TEs in the maize genome (>85%) and the regulatory and evolutionary roles that TEs play in several model organisms (reviewed by Feschotte, 2008) prompted us to study the effect of TEs, especially MULEs, on the Illinois Long-Term Selection Experiment maize strains.

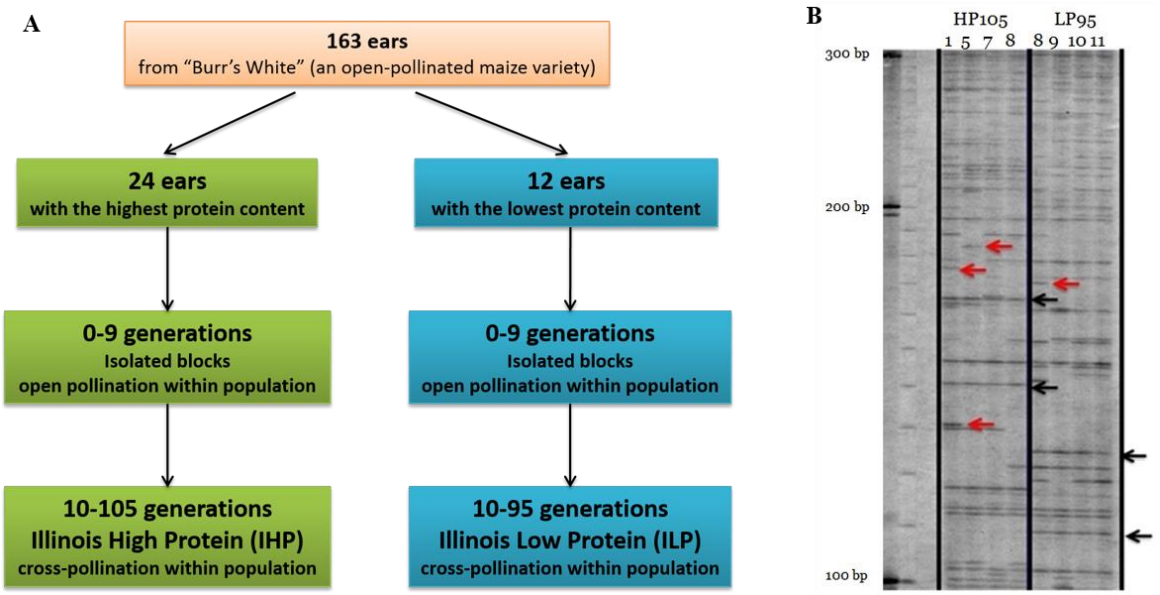

Figure 4.1 Overview of selection procedures and characterization of MULE insertions in the Illinois Long-Term Selection Experiment maize populations.

(A) Diagram of forward selection procedures of the Illinois High Protein (IHP) and Low Protein (ILP) maize populations.
(B) Image of transposon display with a maize MULE in the IHP and ILP maize populations. Four individual plants from the IHP65 and ILP65 were examined. Black arrows point to MULE insertions co-segregating with kernel protein content; red arrows point to MULE insertion that does not co-segregate with protein content.

4.3 Results

4.3.1 MULE Insertions

4.3.1.1 MULE Insertions Segregating with Kernel Protein Content

To determine whether MULE insertions are involved in the artificial selection, insertions that co-segregate with the selected trait (*i.e.*, kernel protein content) were identified. A co-segregating MULE insertion refers to a MULE insertion only present in the IHP population but not in the ILP maize strains, or vice versa (see Figure 4.1B). To this end, ten IHP maize lines (from IHP cycle 65 and 105; IHP65 and IHP105) and ten ILP maize lines (from ILP cycle 65 and 95; ILP65 and ILP95) were used to identify co-segregating MULE insertions through transposon display, a modified AFLP method (Van den Broeck et al., 1998). This method uses nested PCR to amplify MULE flanking sequences with primers specific to the MULE terminal inverted repeat (TIR) and adaptor primers followed by separation of the PCR products using polyacrylamide gel electrophoresis (see Section 4.5.2).

Over 200 co-segregating MULE insertions were detected in these maize lines and flanking sequences of 191 insertions produced significant alignments in the B73 genome. The 10 unmapped insertions may be due to sequence polymorphisms between the Illinois maize and the B73 maize or they represent unsequenced portion of the maize genome. Out of the 191 co-segregating insertions, 57% were from the IHP maize strains, which was significantly higher than that from the ILP maize strains (χ^2=7.6335, p=0.0057). Furthermore, analysis of the mapped flanking sequences indicated that ~79% are located in low-copy genomic regions (only one or two matches from B73 RefGen_v2, BLASTN, e<10^{-5}) and ~42% are within 1 kb of an annotated gene (Table 4.1). Not surprisingly, few insertions (n=12; 6.28%) were detected in exons of protein-coding genes and most of which (n=11) were in the untranslated regions (UTR). There were slightly more insertions in introns than exons of genes, however,

the difference was not statistically significant (10.47% vs. 6.28%; χ^2=2.1829, p=0.1396). Consistent with previous studies, many insertions (~26%) were within 1 kb flanking sequences of genes. However, no significant difference was observed when comparing insertions in the 5′ and 3′ regions of genes (14.14% vs. 11.52%; χ^2=0.5853, p=0.4443), which is in contrast to the previous reports that MULEs preferentially insert into 5′ regions (Liu et al., 2009; Jiang et al., 2011).

Table 4.1 Comparison of co-segregating MULE insertions in the Illinois protein maize strains and MULEs in the B73 maize genome

	Genic regions					Other regions		
	5′ region*	exon	intron	3′ region†	sub-total	intergenic region‡	repetitive region§	total
Co-segregating MULE insertions	27 (14.14%)	12 (6.28%)	20 (10.47%)	22 (11.52%)	81 (42.41%)¶	69 (36.13%)	41 (21.47%)	191
MULEs in the B73 maize genome	2605 (13.11%)	559 (2.81%)	1914 (9.64%)	1727 (8.69%)	6805 (34.26%)¶	13058 (65.74%)		19863

*5′ region: ≤1000 bp upstream of transcription start site
†3′ region: ≤1000 bp downstream of transcription termination site
‡intergenic region: MULE insertions are not within 1000 bp flanking sequences of any non-TE genes
§repetitive region: sequences producing >2 hits (BLASTN, e<10^{-5})
¶fractions in this column are significantly different (χ^2=5.5715, p=0.0183)

4.3.1.2 Comparison of Co-segregating MULE Insertions with MULEs in the B73 Maize Genome

If the co-segregating MULE insertions are associated with the selected trait, their distribution with regard to different genomic sequence features is likely different from that of other maize MULEs. To test this hypothesis, the distribution of MULEs in the B73 genome was analyzed and compared with that of the co-segregating MULEs in the Illinois protein maize strains (Table 4.1). First of all, among the co-segregating MULEs, there are more in genic regions (genes plus one kb flanking sequences) than that in B73 (42.41% vs. 34.26%, χ^2=5.5715, p=0.0183). Nevertheless, the over-representation is not even in different regions

of genes. The fraction of the co-segregating MULEs located in exons is over twice of that is observed in the B73 genome (6.28% vs. 2.81%; χ^2=8.2273, p=0.0041). In contrast, the percent of insertions in introns is comparable between the co-segregating MULEs and that of the B73 genome (10.47% vs. 9.64%; χ^2=0.1514, p=0.6972). Another difference is the ratios of MULE insertions in the 5′ and 3′ regions of genes, where no significant difference was observed for the co-segregating MULEs in the Illinois protein maize strains in contrast to the significant difference observed in the B73 genome (13.11% vs. 8.69%; χ^2=199.7311, p<0.0001).

Table 4.2A Summary of expression of genes with co-segregating MULE insertions in their vicinity.

	Decreased expression	Increased expression	No difference	Genes tested
Immature kernels	10	5	40	55
Young leaves	4	0	4	8

Table 4.2B Summary of expression of genes with co-segregating MULE insertions in their flanking regions and UTRs in immature kernels.

	Decreased expression	Increased expression	No difference	Genes tested
5′ region and 5′ UTR	6	2	14	22
3′ region and 3′ UTR	1	1	11	13

4.3.2 Expression of Genes with Co-segregating MULE Insertions

4.3.2.1 Expression in Immature Kernels

To determine whether the co-segregating MULE insertions affected gene expression, reverse-transcription polymerase chain reaction (RT-PCR) was conducted on genes adjacent to MULE insertions in immature kernels (16 days after pollination). A total of 55 genes were tested for their expression levels in the Illinois protein maize strain (Table 4.2A), where ten exhibited decreased expression levels from the alleles harboring adjacent MULE insertions

compared with the alleles without the MULE insertion (Tables 4.2, 4.3). Among these genes, five MULE insertions were specifically in the IHP maize strains and five were in the ILP maize strains. Not surprisingly, six out of these ten genes harbor MULE insertions in the upstream regions of translation start site (5′ regions and 5′ UTRs), where the promoter or other *cis*-acting elements often reside (Table 4.2B). For example, the allele of a guanine nucleotide-binding protein-like gene (GNUP-like) with a MULE insertion 43 bp upstream of the transcription start site showed no detectable expression in contrast to the allele without the insertion (Figures 4.2A, B). Similarly, for a gene encoding a sterile alpha motif (SAM) domain family protein, weak expression was detected in the ILP cycles 65 and 95, in which the allele contains a MULE insertion 116 bp upstream of its transcription start site compared with the high expression levels from the allele without MULE insertion in the IHP cycles 105 and 65 (Figures 4.3A, B). In addition, reduced expression levels were also observed for three genes containing MULE insertions in introns and one in the 3′ UTR (Table 4.3A). For instance, the allele of the chaperone protein dnaJ with a MULE insertion in its 3′ UTR exhibited reduced expression levels compared with the allele without the insertion (Figures 4.4A, B). However, the reduction seems not as dramatic as genes containing MULE insertions in the upstream regions of genes (Figures 4.2B, 4.3B).

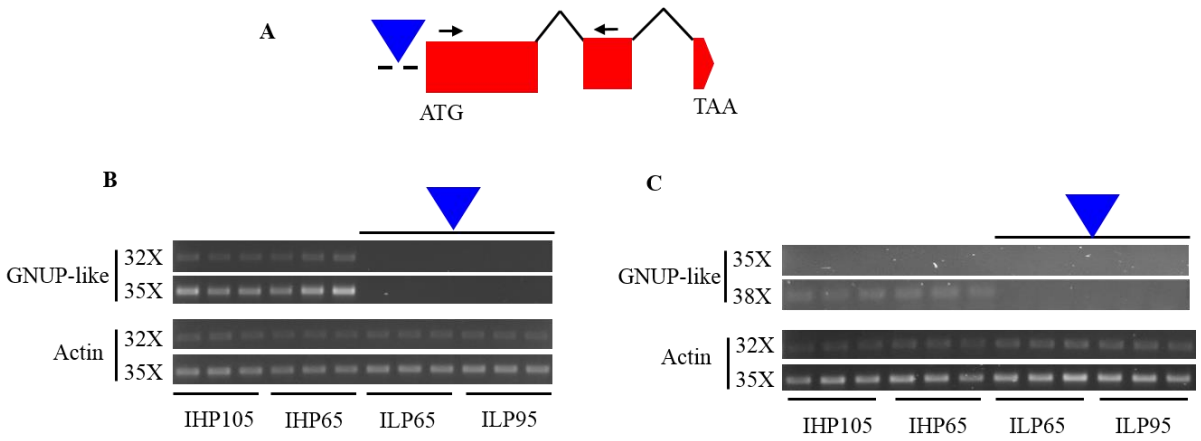

Figure 4.2 Schematic representation (A) and expression levels of a guanine nucleotide-binding protein-like gene (GNUP-like) in immature kernels (B) and young leaves of the IHP and ILP maize strains.

Blue triangles denote the MULE element; red rectangles denote exons and lines connecting them are introns; the dashed line indicates the 5′ region of the gene; black arrows indicate the positions of primers used in RT-PCR. The ILP maize strains (pointed by blue triangles) contain an insertion of Zm00266 at 43 bp upstream of the transcription start site of the gene while the IHP maize strains do not (B and C). Expression of the actin gene serves as internal control for the total input RNA. The numbers next to the gene name or actin indicate the number of cycles of amplification in RT-PCR.

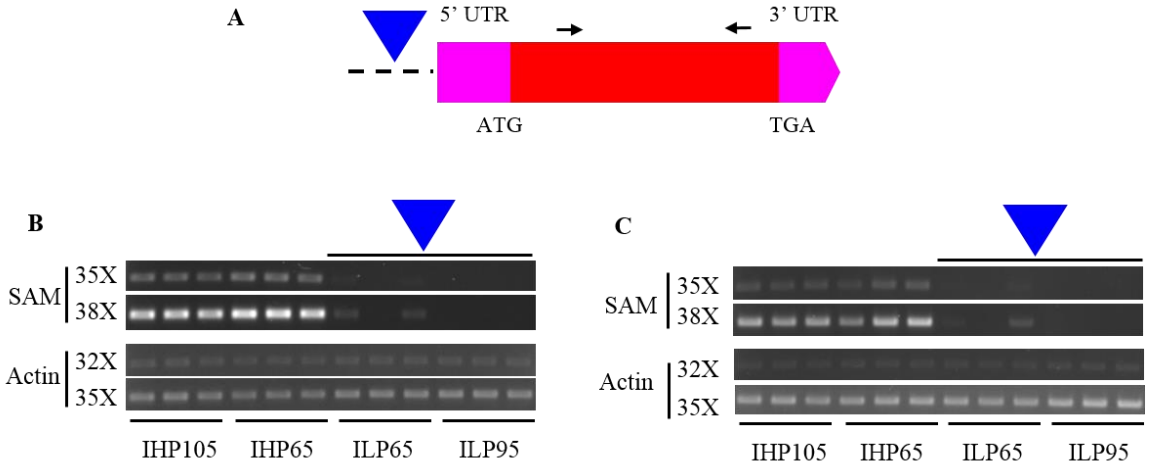

Figure 4.3 Schematic representation (A) and expression levels of an SAM domain family protein in immature kernels (B) and young leaves of the IHP and ILP maize strains. Items are depicted similarly as that in Figure 4.2.

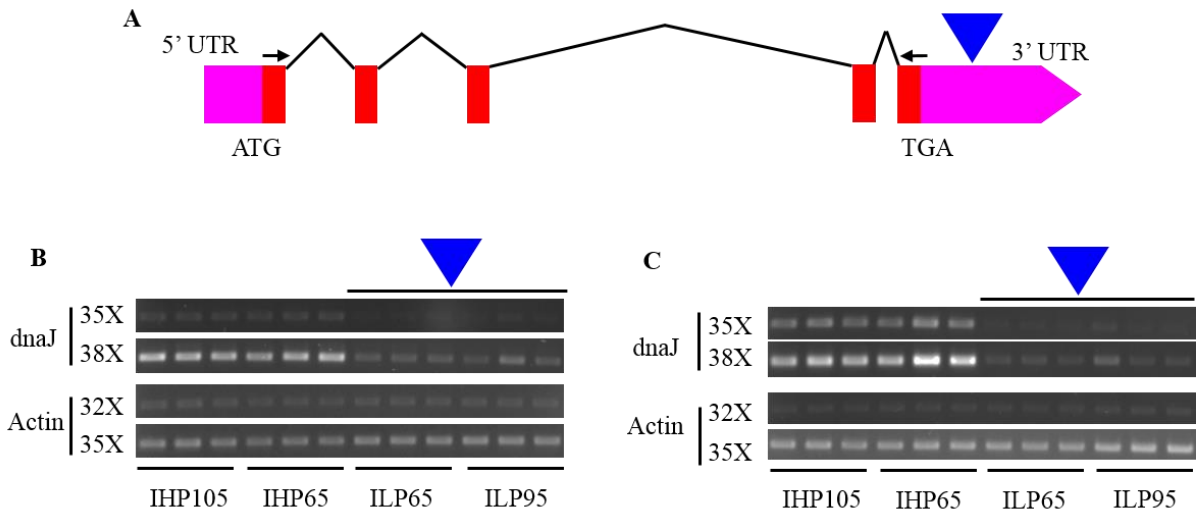

Figure 4.4 Schematic representation (A) and expression levels of the chaperone protein dnaJ in immature kernels (B) and young leaves of the IHP and ILP maize strains. Items are depicted similarly as that in Figure 4.2.

Additionally, five genes exhibited enhanced expression levels from the alleles near the MULE insertion in comparison with the alleles without the MULE insertion (Table 4.2A), and four of them are associated with high protein populations. These include two genes with MULE insertions in the 5′ upstream regions, two in introns and one in the 3′ UTR. For example, in IHP cycles 65 and 105, the allele of a gene encoding a putative penta-tricopeptide (PPR) repeat-containing protein harbors a MULE insertion in its 5′ UTR and its expression was higher compared to the allele without the insertion in the ILP cycles 65 and 95 (Figures 4.5A, B). Another example is a gene encoding AMP deaminase, where the allele in the IHP maize strains harbors a MULE insertion in the 8th intron and exhibited enhanced expression levels compared with that in the ILP maize strains which do not contain the insertion (Figures 4.6A, B). Given the fact that the *Mutator* elements have been used as powerful agents to knock-out genes, it is intriguing to observe that the incidents with enhanced expression was not rare in the Illinois protein maize strains.

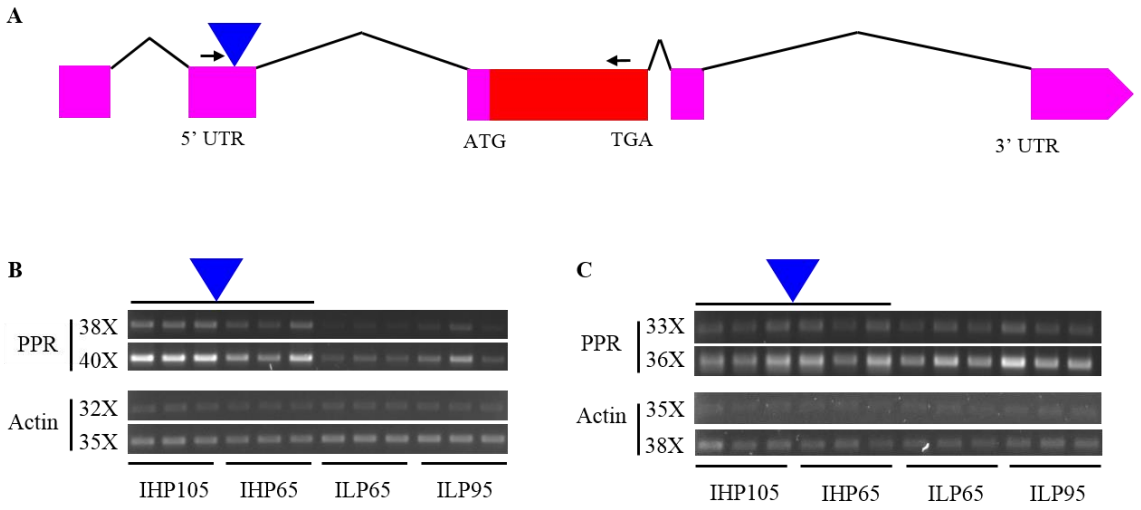

Figure 4.5 Schematic representation (A) and expression levels of a putative penta-tricopeptide repeat-containing (PPR) protein in immature kernels (B) and young leaves of the IHP and ILP maize strains. Items are depicted similarly as that in Figure 4.2.

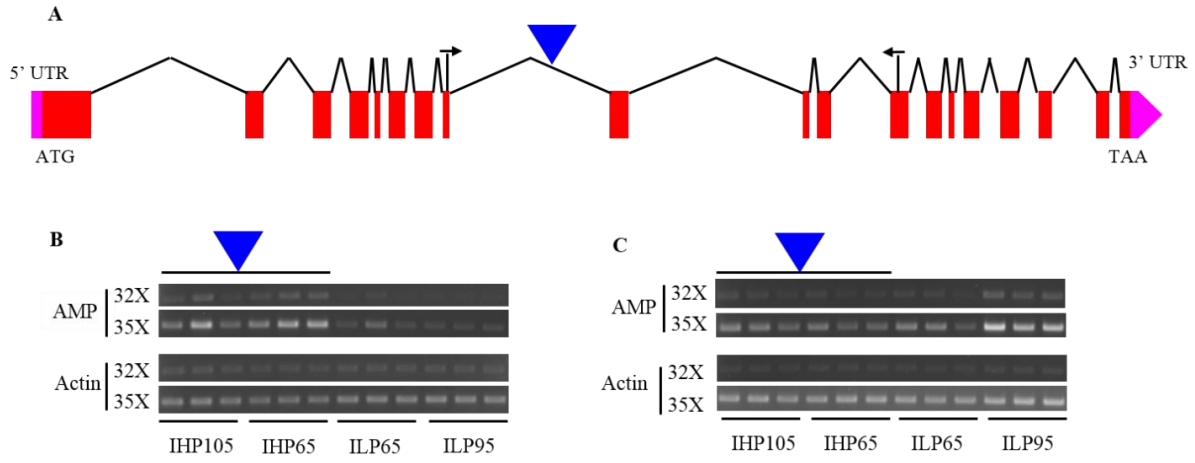

Figure 4.6 Schematic representation (A) and expression levels of an AMP deaminase in immature kernels (B) and young leaves of the IHP and ILP maize strains. Items are depicted similarly as that in Figure 4.2.

Table 4.3A Genes with altered expressions in immature kernels between maize lines with and without MULE insertions.

	MULE insertion	Insertion	Adjacent gene	Insert site	High average*	Low average†	*t*-value‡	p-value
Reduced expression	Zm00754BMspI_H1	high protein maize	phosphatase 2C	-1720 bp	295039	680710	-6.52	<.0001
	Zm00754HinP1I_L3	low protein maize	SAM-domain protein	-116 bp	222324	114627	13.72	<0.0001
	Zm00266BHinP1I_L1	low protein maize	guanine_nucleotide-binding_protein-like	-43 bp	592539	269829	8.64	<0.0001
	Zm00257HinP1ITA_H	high protein maize	10-deacetylbaccatin III 10-O-acetyltransferase	5′ UTR	107689	124347	-4.01	0.0025
	Zm00754MspI_L2	low protein maize	isochorismatase_hydrolase_family_protein	5′ UTR	4219806	3702855	3.01	0.0132
	Zm17242HinP1I_L	low protein maize	Late embryogenesis abundant protein	5′ UTR	529654	228653	3.97	0.0027
	Zm00266BHinP1I_H1	high protein maize	protein_farnesyltransferase	1st intron	162474	171962	-2.73	0.0213
	Zm00851HinP1I_H2	high protein maize	glycolipid_transporter	2nd intron	314879	406086	-4.48	0.0012
	Zm00460AMspI_H2§	high protein maize	Hypothetical protein	3nd intron	1982507	3573824	-3.72	0.004
	Zm00754HinP1I_L2	low protein maize	Chaperon protein dnaJ	3′ UTR	187933	166229	2.91	0.0156
Increased expression	Zm00851MseI_H2	high protein maize	putative_ferulate_5-hydroxylase	-297 bp	70683	45527	4.14	0.0025
	Zm754AMseI_HI§	high protein maize	pentatricopeptide repeat-containing protein	5′ UTR	258474	166992	4.8	0.0007
	Zm01654_L1	low protein maize	Unknown protein	1st intron	707398	1364408	-4.45	0.0012
	Zm00460A_H1§	high protein maize	AMP deaminase	8th intron	787802	468204	5.08	0.0005
	Zm00851HinP1I_H1§	high protein maize	Mitochondrial transcription termination factor-related	3′ UTR	6477647	4615584	2.23	0.0497

*High average: the average PCR band intensity of six high protein maize lines (see Section 4.5)
†Low average: the average PCR band intensity of six low protein maize lines (see Section 4.5)
‡t-value: the value calculated through the formula for the t-test, which is related to the size of the difference between the "high average" and "low average" for each gene
§No expression difference was detected in the young leaves (see Table 4.3B)

Table 4.3B Comparison of expression of genes in young leaves between maize lines with and without MULE insertions.

	MULE insertion	Insertion	Adjacent gene		High average	Low average	*t*-value	p-value
Reduced expression	Zm00754BMspI_H1	high protein maize	phosphatase 2C	-1720 bp	319285	369094	-6.03	0.0001
	Zm00754HinP1I_L3	low protein maize	SAM-domain protein	-116 bp	507129	382592	8.55	<0.0001
	Zm00257HinP1ITA_H	high protein maize	10-deacetylbaccatin III 10-O-acetyltransferase	5′ UTR	560774	661795	-2.92	0.0153
	Zm00754HinP1I_L2	low protein maize	Chaperon protein dnaJ	3′ UTR	530117	312552	5.38	0.0003
No difference	Zm754AMseI_HI	high protein maize	Penta-tricopeptide repeat-containing protein	5′ UTR	375394	430769	-1.32	0.2155
	Zm00460AMspI_H2	high protein maize	Hypothetical protein	3nd intron	482201	440403	1.78	0.106
	Zm00460A_H1	high protein maize	AMP deaminase	8th intron	370620	400951	-1	0.3417
	Zm00851HinP1I_H1	high protein maize	Mitochondrial transcription termination factor-related	3′ UTR	789877	1030583	-2.05	0.0674

Among the genes exhibiting differential expression levels between IHP and ILP maize strains, eight out of fifteen harbor co-segregating MULE insertions in either the 5′ regions or 5′ UTRs, and only two contain insertions in the 3′ UTRs (Tables 4.2A, B). This suggests that MULE insertions upstream of translation start codon may alter gene expression levels more often than insertions downstream of genes. To test this hypothesis, all the genes tested for expression levels (n=55) were analyzed. As shown in Table 4.2B, 8 out of 22 genes (36.36%) containing MULE insertions upstream of translation start codon exhibited altered expression levels while only 2 out of 13 (15.38%) containing insertions downstream of translation stop codon of genes showed altered expression levels (Table 4.2B). However, the difference is not significant (χ^2=1.7622, p=0.1843), which is likely due to the small sample size.

4.3.2.2 Expression in Young Leaves

Differential expression levels of genes with adjacent MULE insertions in the immature kernels implied that MULE insertions may exert effects on these genes. If these genes were indeed involved in kernel protein content and the alteration correlated with MULE insertion has been selected for the trait, we may expect that differential expression occurs more significantly in the kernels. To test this hypothesis, expression levels of eight genes showing differential expression in the immature kernels were determined in the young leaves. Indeed, the difference of expression levels between allele with or without MULE insertion was not detectable in young leaves for three genes showing enhanced expression levels and one gene showing reduced expression in immature kernels (Tables 4.3A, B). For example, in young leaves, for the gene encoding a putative penta-tricopeptide repeat-containing (PPR) protein, the IHP maize strains harboring a MULE insertion in the 5′ UTR exhibited similar expression levels to the ILP maize strains which do not have the insertion (Tables 4.3A, B; Figure 4.5C). Similarly, no enhanced expression was observed in young leaves for the allele containing a

MULE insertion in the 8th intron of the gene encoding an AMP deaminase in contrast to that in immature kernels (Tables 4.3A, B; Figure 4.6C). Besides genes showing no differential expression, four genes exhibited similar expression patterns in young leaves to that in immature kernels (Table 4.3B; Figures 4.2C, 4.3C, and 4.4C). Taken together, it seems that the enhancement associated with MULEs is kernel specific, while the suppression is more consistent between the two tissues.

4.4 Discussion and Future Work

The ultimate source for evolution is mutation. Transposons can cause a wide range of mutations, some of which possess unique features that cannot be reproduced by other mutations, such as point mutations, deletions, and segmental duplications. For example, transposon insertion may cause epigenetic modification of the chromatin they insert into, which is unlikely to be mimicked by other mutations in the genome. From this point of view, TEs are a driving force for evolution and provide raw materials for natural selection and diversification. This is consistent with the fact that TEs are widespread and abundant in plant genomes. Nevertheless, artificial selection is different from natural selection in that it does not select for fitness. Instead, it selects for a particular trait which is often quantitative. As a result, it is unclear whether TEs play any role in this process. Whereas this study did not provide an unambiguous answer for this question, it resulted in several interesting observations and provided a good foundation for future studies.

4.4.1 The Effect of Co-segregating MULE Insertions on Gene Expression

In this study, a total of 15 (out of 55 tested, 27%) co-segregating MULE insertions are correlated with significant difference of expression levels of genes in immature kernels. This ratio is very high compared with the number of differential expressed genes among different

maize varieties. In a previous study, 13,495 maize genes were surveyed and the differential expressed genes between different varieties ranged from 326 (2.4%) to 1,144 (8.5%) (Stupar et al., 2008). This suggests that co-segregating MULE insertions are frequently associated with differential gene expression. While both suppression and enhancement were observed, it is apparent that suppression was more frequent. Interestingly, most of the cases of enhancement were associated with insertions in high protein maize lines and that expression is specifically elevated in kernels. Furthermore, insertions in 5′ regions are more likely associated with differential expression than that in 3′ regions.

One of the key questions is whether the correlation between MULE insertions and the differential expression reflects a causal relationship or the MULE insertions only serve as "markers" for other underlying mutations that are responsible for differential gene expression. If such relationship is simply an association, it would be very difficult to explain the bias mentioned above. For example, if the MULE insertions did not directly alter gene expression, it is difficult to explain why the frequency of altered expression would vary with their insertion positions (5′ region or 3′ region). Given those facts, it is likely that at least some of the co-segregating MULE insertions caused the differential expression between the IHP and ILP lines.

It should be noted that the altered expression levels of genes with adjacent MULE insertions were simply associated with those maize strains and may not be the cause of the selected trait. Nevertheless, the fact that some of the differential expression was only detected in immature kernel but not leaf seem to suggest the involvement of these genes in the selected trait. To further determine whether the co-segregating MULE insertions contribute to the kernel protein content, more experiments need to be conducted. First, the ILTSE maize population consists of reverse protein maize strains, which were formed by selecting for low

protein content in IHP strains (RHP) at cycle 48 and selecting for high protein content in ILP strains (RLP) at cycle 48 (Moose et al., 2004). If the MULE insertions are involved in the protein content, we may observe some insertions co-segregating with the IHP maize strains were also present in the RLP strains, suggesting the recurrent acquisition of the relevant insertions during the selection. Indeed, such incidence was observed in the preliminary test (Data not shown). More systematic examination using more reverse maize lines is necessary in order to obtain more conclusive results. Another approach is to use *Arabidopsis* T-DNA plants which contain insertion mutations in the orthologous genes of those exhibiting differential expression levels between maize line with and without MULE insertions. The availability of *Arabidopsis* mutant libraries (www.arabidopsis.org) would make this analysis possible. Finally, maize tagging lines containing mutations in relevant genes can be used to test the role of these genes in kernel protein content.

4.4.2 The Distribution Pattern of Co-segregating MULE Insertions

The co-segregating MULE insertions may be derived from three sources: 1) existing insertion polymorphisms or heterozygosity that were present prior to the selection; 2) additional insertions from other maize varieties that were introduced to the population due to cross-pollination during the selection; 3) new transpositions occurred during the selection. Obviously, new insertions generated during selection are of very recent origin. The presence of the polymorphic insertions in 1) and 2) could be due to the insertions generated after the divergence of individuals or cultivars, or they represent insertions that don't provide great fitness benefit so they failed to spread across the population. For these reasons, the co-segregating MULE insertions should be younger than other MULEs in the genome as a group.

Among the co-segregating MULEs, more insertions were detected in IHP lines than in the ILP (57% vs. 43%) lines. At first glance this seems to be puzzling because one would

assume the reduction of protein content could be achieved through disruptive effects such as knock-out or knock-down mutations. Given the fact that MULE insertions are associated with more suppressive effects on gene expression, it would predict that the ILP lines are associated with more insertions than the IHP lines. This was not observed because disruptive effect could be saturated rather rapidly. For example, if a key enzyme in the pathway of protein synthesis or nitrogen transportation is knocked out, any additional disruptive mutation may not further reduce the protein content and would not be selected for. Moreover, the selection would be terminated when the protein content was too low to support the fitness of the seed. In contrast, there could be numerous ways to increase protein content. These include both "constructive" mutations, such as increasing protein synthesis or uptake of nitrogen to kernels, and "disruptive" mutations, such as reduction of starch synthesis so that the relative content of protein is elevated. As a result, selection for high protein content could be more progressive, and require more constructive mutations. This is in agreement with the observation that the reduction of protein content largely plateaued in ILP lines after cycle 65 while the protein content has been steadily increased in IHP lines (Dudley and Lambert, 1992; Moose et al., 2004). This is also consistent with the fact that most of the enhancement of gene expression was observed in IHP. Thus the presence of more co-segregating MULEs may actually reflect the more intensive selection in these lines and imply that they are selected for.

Compared with the distribution of MULEs in B73, the co-segregating MULE insertions seem to be overrepresented in genic region especially exons of protein-coding genes. This suggests that either the exonic insertions are related to the trait or that the co-segregating MULE insertions are relatively young and have not been subjected to sufficient natural selection for elimination. A third possibility is that the unique population structure of the

ILTSE has favored the retention of MULE insertions in exons. These hypotheses are not mutually exclusive.

In addition to the enrichment of exonic insertions among co-segregating MULEs, it is worth mentioning that the insertion in 3′ region of genes is enriched compared to that in B73 (11.52% vs. 8.69%) even though the difference is not significant. An alternative description of this phenomenon is that the well-known preference for insertion into 5′ region of genes was not observed for co-segregating MULEs. Such discrepancy cannot be explained by the younger age because the preference was observed for both new and existing MULE insertions (Liu et al., 2009; Jiang et al., 2011). As discussed above, the selection for disruptive effects should be more rapidly saturated than the constructive effects in this specific population. If that is the case, insertions in 5′ regions should be selected against among co-segregating MULEs because in general they are more likely associated with suppression of gene expression. This explains why among the co-segregating MULE insertions the preference for 5′ region is not observed and provides another piece of evidence that some of the co-segregating MULE insertions have been selected for. Further studies involving other MULE insertions in the Illinois maize strains are needed to elucidate this phenomenon.

4.5 Materials and Methods

4.5.1 Plant Materials and DNA Extraction

Seeds of the Illinois High Protein maize strains at cycles 105 and 65 (IHP105 and IHP65), the Illinois Low Protein maize strains at cycles 95 and 65 (ILP95 and ILP65) were kindly provided by Dr. Moose (University of Illinois at Urbana-Champaign). Plants were grown in the field at the Horticulture Teaching and Research Centre of Michigan State University from

early May to late October, 2009-2013. Leaf tissues were collected from two-week old seedlings. Emerging ears (before the emergence of silks) were covered with envelops and hand pollinated using pollens from the same plant. Immature kernels were collected 16 days after pollination (DAP). The Cetyltrimethyl Ammonium Bromide (CTAB) DNA extraction protocol (Clarke, 2002) was used to extract genomic DNA from young leaves followed by RNase treatment.

4.5.2 Detection of MULE Insertions

MULE insertions in maize genome were detected using transposon display, a modified AFLP method (Van den Broeck et al., 1998). Genomic DNAs from maize plants were digested using restriction enzymes (*Bfa*I, *Mse*I, *Msp*I and *Hin*P1I) separately so that different digestion patterns can be obtained to increase the chance of covering as many loci as possible in these genomes. After digestion, a double-stranded adapter was ligated to the digested genomic fragments, this serves as the annealing site for one of the PCR primers (adapter primer). By using this adapter primer and another primer specific to MULE TIRs, genomic fragments containing part of the MULE TIR and a portion of MULE flanking sequence are amplified. Nested PCRs were used to increase the specificity of amplicons, which were resolved on 6% acrylamide gel. To recover desired insertions, DNA fragments were excised from the acrylamide gels and placed in 15 µl distilled water over night at 4 ℃ and 7 µl resulting solution was used for PCR amplification with a total volume of 20 µl PCR cocktail. After purification of the PCR products using the Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, WI, USA), fragments were cloned using TA-cloning vector and sequenced at the Research and Technology Sequencing Facility (RTSF) of Michigan State University. BLAST analysis against the B73 genomic sequence revealed the locations of the

MULE insertion sites. B73 RefGen_v2 filtered gene database (www.maizegdb.org) is examined to determine whether MULE insertion sites are within or close to genes.

4.5.3 Determination of Gene Expression Levels

Plant tissues were collected in the field and immediately placed in liquid nitrogen. Thereafter the tissues were stored in -80 ℃ freezer prior to RNA extraction. Total RNA was extracted from immature maize kernels and young leaves using the TRIzol Reagent (Life Technologies, Carlsbad, CA, USA). Specifically, tissues were ground in liquid nitrogen and transferred to RNase-free tubes containing 1 mL TRIzol Reagent. After incubating the homogenized sample for 5 minutes at room temperature, 0.2 mL of chloroform was added to each sample. The mixture was centrifuged at 14,000 rpm for 15 minutes at 4 ℃ and the colorless upper aqueous phase was transferred to a new tube, to which 0.5 mL isopropanol was added to precipitate the RNA. Centrifugation was conducted to pellet the RNA followed by washing with 75% ethanol. The RNA samples were treated with RNase-free DNase (Qiagen, Hilden, Germany) and run on a 1% agarose gel to monitor the integrity. Complementary DNA (cDNA) was synthesized using the GoScript Reverse Transcription System (Promega, Madison, WI, USA). The expression of the actin gene was used to standardize the input volume of synthesized cDNA for all the samples. Products of the actin and target genes were amplified through an initial denaturing step at 94 ℃ for 2 min, followed by different cycles (based on different genes) of 94 ℃ for 30 sec, 58 ℃ for 30 sec, and 72 ℃ for 30 sec. Equal volume of PCR products were run on 1% agarose gel. To compare the expression levels, intensity values of the amplified fragments on the agarose gel were measured using the Image Lab software for the Gel Doc EZ Imager (Bio-Rad Laboratories, Hercules, CA USA) and were evaluated using the student *t*-test. If the

difference of expression level between the two alleles (with or without MULE insertion) is significant ($p<0.05$), it was considered the gene expression is enhanced or suppressed.

4.5.4 Detection of MULE Insertions in the B73 Maize Genome

The maize MULE library was constructed and maintained by the Jiang lab at Michigan State University. The left-most 300 nucleotides (nt) of each MULE element (individual length >300 bp) and the entire sequence of those elements with smaller size (≤300 nt) were used to mask the B73 maize genome (RefGen_v2, www.maizegdb.org) using the RepeatMasker program (www.repeatmasker.org). Masked regions with less than 50 nt truncations at the external ends of MULE TIRs were retained for further analysis. The coordinates of genomic features (genes, exons, introns) were based on the RefGen_v2 annotation of the filtered gene set of the B73 genome (www.maizegdb.org). For genes with multiple transcript isoforms, only the first one of each was used in the analysis. If there was overlap between the coordinates of the terminal sequences of MULE TIRs and exons, the MULE element was considered to be inserted in the exon. Two TIRs were considered as one insertion event if they satisfied the following criteria: 1) the distance between two TIRs was less than 5 kb; 2) two TIRs were in inverted orientation with external ends outwards, as that in a normal element; 3) they shared at least 75% identity over their TIR sequences. TIRs that did not meet these criteria were considered as independent insertion event for each TIR. Similar analysis was done for MULE insertions in the introns and 1 kb flanking sequences of genes.

REFERENCES

REFERENCES

Alrefai R, Orozco B, Rocheford T (1994) Detection and sequencing of the transposable element *ILS-1* in the Illinois Long-Term Selection maize strains. Plant Physiology **106:** 803-804

Bennetzen JL, Swanson J, Taylor WC, Freeling M (1984) DNA insertion in the 1st intron of maize *Adh1* affects message levels: Cloning of progenitor and mutant *Adh1* alleles. Proc. Natl. Acad. Sci. USA **81:** 4125-4128

Blauth SL, Kim KN, Klucinec J, Shannon JC, Thompson D, Guiltinan M (2002) Identification of *Mutator* insertional mutants of starch-branching enzyme 1 (*sbe1*) in *Zea mays* L. Plant Molecular Biology **48:** 287-297

Blauth SL, Yao Y, Klucinec JD, Shannon JC, Thompson DB, Guilitinan MJ (2001) Identification of *Mutator* insertional mutants of starch-branching enzyme 2a in corn. Plant Physiology **125:** 1396-1405

Cossegal M, Chambrier P, Mbelo S, Balzergue S, Martin-Magniette ML, Moing A, Deborde C, Guyon V, Perez P, Rogowsky P (2008) Transcriptional and metabolic adjustments in ADP-glucose pyrophosphorylase-deficient *bt2* maize kernels. Plant Physiology **146:** 1553-1570

Diao XM, Lisch D (2006) *Mutator* transposon in maize and MULEs in the plant genome. Acta Genetica Sinica **33:** 477-487

Dudley JW (2007) From means to QTL: The Illinois long-term selection experiment as a case study in quantitative genetics. Crop Science **47:** S20-S31

Dudley JW, Dijkhuizen A, Paul C, Coates ST, Rocheford TR (2004) Effects of random mating on marker-QTL associations in the cross of the Illinois High Protein x Illinois Low Protein maize strains. Crop Science **44:** 1419-1428

Dudley JW, Lambert RJ (1992) 90-generations of selection for oil and protein in maize. Maydica **37:** 81-87

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics **9:** 397-405

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annual Review of Genetics **41:** 331-368

Goldman IL, Rocheford TR, Dudley JW (1993) Quantitative trait loci influencing protein and starch concentration in the Illinois Long-Term Selection maize strains. Theoretical and Applied Genetics **87:** 217-224

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

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Research **19:** 1419-1428

Jiang N, Bao ZR, Zhang XY, 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

Jiao YL, Deng XW (2007) A genome-wide transcriptional activity survey of rice transposable element-related genes. Genome Biology **8:** R28 doi:10.1186/gb-2007-8-2-r28

Lisch D (2002) *Mutator* transposons. Trends in Plant Science **7:** 498-504

Lisch D, Jiang N (2009) *Mutator* and MULE transposons. *In* Bennetzen JL,Hake S, eds, Mazie Handbook, Vol II. Springer Science + Business Media LLC, pp 277-306

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

Liu SZ, Yeh CT, Ji TM, Ying K, Wu HY, 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 doi:10.1371/journal.pgen.1000733

Martinez G, Slotkin RK (2012) Developmental relaxation of transposable element silencing in plants: Functional or byproduct? Current Opinion in Plant Biology **15:** 496-502

Moose SP, Dudley JW, Rocheford TR (2004) Maize selection passes the century mark: A unique resource for 21st century genomics. Trends in Plant Science **9:** 358-364

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature **461:** 1130-1134

Robertson DS (1978) Characterization of a mutator system in maize. Mutation Research **51:** 21-28

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science **326:** 1112-1115

Stupar RM, Gardiner JM, Oldre AG, Haun WJ, Chandler VL, Springer NM (2008) Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. BMC Plant Biology **8:** 33 doi:10.1186/1471-2229-8-33

Van den Broeck D, Maes T, Sauer M, Zethof J, De Keukeleire P, D'Hauw M, Van Montagu M, Gerats T (1998) Transposon display identifies individual transposable elements in high copy number lines. Plant Journal **13:** 121-129

Wassom JJ, Wong JC, Martinez E, King JJ, DeBaene J, Hotchkiss JR, Mikkilineni V, Bohn MO, Rocheford TR (2008) QTL associated with maize kernel oil, protein, and starch concentrations; kernel mass; and grain yield in Illinois high oil x B73 backcross-derived lines. Crop Science **48:** 243-252

CHAPTER 5

Conclusions and Perspectives

For most organisms, TEs are an integral part of the genome; as a consequence, the activity of TEs not only contributes to the success of themselves, but also the evolution of the host genome. In this dissertation, I studied the amplification of MULEs and their interaction with their genomes. My study covers the intrinsic features that control the amplification of TEs – the activity of transposase (and the modification of such activity) as well as external factors that determine the retention of TEs. These factors include how the host genome limits the abundance of MULEs through deletion and mutation, how MULEs interact with other TEs, and how MULE insertions are selected by humans. The findings described in this dissertation facilitate the understanding of MULE evolution, and the transposition system established provides the foundation for further studies of MULE transposition mechanisms as well as how they acquire gene and/or gene fragments.

5.1 Differential Indel Rate between LTR Retrotransposons and Coding-MULEs in Maize

In Chapter 2, mechanisms underlying differential abundance of MULEs in maize and rice were investigated, where one mechanism is the accelerated indel rate in maize. Interestingly, MULEs seem to be subjected to more frequent indels than LTR retrotransposons (specifically *Copia*-like LTR retrotransposons) in maize. This may be due to the differences in target site preference between LTR retrotransposons and MULEs. It is well known that LTR retrotransposons are mostly located in pericentromeric heterochromatin. Although some *Copia*-like elements were found to prefer gene-rich regions, they actually inserted into the heterochromatic regions near genes in maize (Liu et al., 2007). In contrast, MULEs preferentially target low-copy euchromatic regions, especially sequences in close proximity of genes, which has been demonstrated by previous studies as well as the findings in this dissertation (Dietrich et al., 2002; Liu et al., 2009; Jiang et al., 2011). Hence, the

different properties associated with heterochromatic and euchromatic regions, such as lower recombination rate of heterochromatic regions, may contribute to the lower indel rate of LTR retrotransposons compared with that of MULEs in maize.

Another possibility is the different activities between LTR retrotransposons and MULEs in maize. Successful amplification of retrotransposons makes them the largest mass of the B73 genome (>75%) (Schnable et al., 2009). Detailed sequence analysis of LTR retrotransposons in gene-free regions revealed two peaks of amplifications, around 1.5-2 million year ago (mya) and within the last 500,000 years (Liu et al., 2007). In addition, expression analysis based on EST databases revealed higher transcriptional activity of LTR retrotransposons compared to other TE families (Vicient, 2010). Another global gene expression analysis in maize found that some retrotransposon ESTs were up-regulated more than 10-fold in the shoot apical meristem relative to that in seedlings (Ohtsu et al., 2007). Previous studies showed accelerated evolution of genes with no transcriptional activity (Hastings, 1996; Zou et al., 2009). Hence, the recent amplification and transcriptional activity of LTR retrotransposons may contribute to the low indel rate compared to that of coding-MULEs, which only have few elements with expression evidence.

The third possibility is the different transposition strategies between LTR retrotransposons and MULEs. As mentioned before, LTR retrotransposons transpose through an RNA intermediate, which is reverse transcribed into cDNA and becomes integrated at a new genomic locus. This mode of transposition will not result in double strand breaks at the locus of the original copy because the element does not excise. In contrast, DNA transposons, including MULEs, transpose through a “cut-and-paste” mechanism, which requires the excision of the elements and generates double strand breaks at the original loci. Previous studies revealed that transcription slippage during the gap repair process of the double strand

breaks contributed to the instability of repeat tracts (*e.g.*, deletions) in yeast (Jankowski et al., 2000). This phenomenon was also observed in maize by the presence of deletion derivatives of *MuDR* (Lisch et al., 1999). Hence, it is conceivable that MULEs contain more indels because of the inefficient repair of the double strand breaks created following their excision. In contrast, LTR retrotranspons do not excise, therefore, are void of defects generated during double strand gap repair.

Of course, the above hypotheses are not mutually exclusive. The genomic niche where diverse TE families co-exist is not static. Hence, to further elucidate this phenomenon, analysis of other TE families may be helpful.

5.2 Maintenance of Low Activity of Wild-Type Transposases May be a Tactic for TE Persistence

In this dissertation, the wild-type transposase of the rice MULE Os3378 only exhibited limited activity in yeast. Sequence and structural modifications of the transposase resulted in enhanced activity, suggesting the potential of the element to be highly active and maintenance of low activity of the wild type may be beneficial for its survival in the rice genome. A similar phenomenon has been seen for transposases from other TEs, where minor manipulations of the transposase sequence resulted in increased activity. For example, mutation of the serine at position 129 to alanine of the *Drosophila P*-element transposase led to two-fold elevation of activity compared with the wild-type transposase (Beall et al., 2002). In an attempt to generate hyperactive transposase for mammalian applications, Yusa et al. (2011) conducted site-directed mutagenesis of the *piggyBac* transposase and found 18 mutants exhibited enhanced activity (Yusa et al., 2011). These 18 mutants are comprised of one or two amino acid substitutions in an individual transposase protein. It is known that most TEs are under epigenetic control of the host. When copy number increases through

active transposition of a TE, it likely triggers the host defense system, which will result in tight control of the TE. Thus, maintenance of low activity may render the TE less pressure from host surveillance system, such as RNA interference.

5.3 A System for Studying Transposition Mechanism of MULEs

The successful transposition of the rice MULE (Os3378) in yeast provides the possibility of studying MULE transposition in a system that is easy to handle. As shown by previous studies, the MULEs are a very diverse family, which consists of elements with varied sequence compositions (Yu et al., 2000; Lisch and Jiang, 2009; Ferguson and Jiang, 2012). The canonical elements contain one pair of long TIRs (100-500 bp) at the ends of elements, which is one of the features (such as TSDs and transposase) that distinguish MULEs from other TE families. Subsequent studies discovered MULEs whose terminal sequences did not share high similarity, which were referred to as non-TIR MULEs (Yu et al., 2000). In addition, Ferguson and Jiang (2012) found a special group of MULEs with multiple long TIRs in several plant species. It is known that TIRs contain sequences that transposases recognize and bind to during the transposition process. Nevertheless, it is not known how MULE transposases deal with such diverse terminal structures/sequences. Hence, it will be interesting to test whether non-TIR MULEs can be mobilized and whether MULEs with multiple TIRs are more competent in transposition than MULEs with one pair of TIRs.

Additionally, the mechanisms of sequence acquisition by Pack-MULEs may be investigated using the yeast system. The finding that retention of single terminal sequence occurred after the excision of the Os3378 element already showed a potential way for the formation of Pack-MULEs (see Chapter 3). Again, it is not known whether the long TIRs or multiple TIRs play any role in the acquisition process. In rice, the average copy number of Pack-MULEs is three while the one Pack-MULE family with tandem TIRs in tomato has 13

copies, suggesting that the tandem TIRs may render the element more competent in amplification (Jiang et al., 2004; Ferguson and Jiang, 2012). While testing the formation of Pack-MULEs using MULEs with one pair of TIRs, it would also be interesting to use MULEs with multiple TIRs.

5.4 The Evolutionary Origin of Co-segregating MULE Insertions Associated with Altered Gene Expression

As determined from RT-PCR analysis, 15 out of 55 genes exhibited differential expression levels between alleles with and without adjacent co-segregating MULE insertions, implying they may be involved in the changes of kernel protein content. Since this selection occurred during the past 100 years, it is interesting to know when these MULE insertions were selected into the high or low protein maize strains. The co-segregating MULE insertions may be derived from three sources: 1) existing insertion polymorphisms or heterozygosity that were present prior to the selection; 2) additional insertions from other maize varieties that were introduced to the population during selection due to cross-pollination; 3) new transpositions that occurred during selection. If an insertion is derived from a new transposition, it should only be present in this population. If it is from other varieties, it should be present in other maize cultivars. Indeed, some but not all of the co-segregating MULE insertions are indeed observed in the B73 maize genome, suggesting they either represent ancient polymorphisms or were derived from cross-hybridization. For this study, a wide collection of maize varieties, such as the 25 inbred parents used in generating the Nested Association Mapping (NAM) population (McMullen et al., 2009), may be used to determine the relative insertion time frame based on the presence and absence information. In the case that some MULE insertions may be lost during the domestication process, it will informative to include some teosinte species/subspecies that are closely related to

domesticated maize, *i.e.*, *Z. mays* ssp. *mexicana*, *Z. mays* ssp. *huehuetenangensis*, and *Z. mays* ssp. *parviglumis.*

REFERENCES

REFERENCES

Beall EL, Mahoney MB, Rio DC (2002) Identification and analysis of a hyperactive mutant form of Drosophila *P*-element transposase. Genetics **162:** 217-227

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

Ferguson AA, Jiang N (2012) *Mutator*-like elements with multiple long terminal inverted repeats in plants. Comparative and Functional Genomics **2012:** 695827 doi:10.1155/2012/695827

Hastings KEM (1996) Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. Journal of Molecular Evolution **42:** 631-640

Jankowski C, Nasar F, Nag DK (2000) Meiotic instability of CAG repeat tracts occurs by double-strand break repair in yeast. Proc. Natl. Acad. Sci. USA **97:** 2134-2139

Jiang N, Bao ZR, Zhang XY, 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

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

Lisch D, Jiang N (2009) *Mutator* and MULE transposons. *In* Bennetzen JL,Hake S, eds, Mazie Handbook, Vol II. Springer Science + Business Media LLC, pp 277-306

Liu RY, Vitte C, Ma JX, Mahama AA, Dhliwayo T, Lee M, Bennetzen JL (2007) A GeneTrek analysis of the maize genome. Proc. Natl. Acad. Sci. USA **104:** 11844-11849

Liu SZ, Yeh CT, Ji TM, Ying K, Wu HY, 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 doi:10.1371/journal.pgen.1000733

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li HH, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, et al. (2009) Genetic properties of the maize nested association mapping population. Science **325:** 737-740

Ohtsu K, Smith MB, Emrich SJ, Borsuk LA, Zhou RL, Chen TL, Zhang XL, Timmermans MCP, Beck J, Buckner B, et al. (2007) Global gene expression analysis of the shoot apical meristem of maize (*Zea mays* L.). Plant Journal **52:** 391-404

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science **326:** 1112-1115

Vicient CM (2010) Transcriptional activity of transposable elements in maize. BMC Genomics **11:** 601 doi:10.1186/1471-2164-11-601

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

Yusa K, Zhou LQ, Li MA, Bradley A, Craig NL (2011) A hyperactive *piggyBac* transposase for mammalian applications. Proc. Natl. Acad. Sci. USA **108:** 1531-1536

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH (2009) Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. Plant Physiology **151:** 3-15