MINIATURE INVERTED REPEAT TRANSPOSABLE ELEMENTS IN RICE – ORIGIN AND FUNCTION

A Dissertation

by

GUOJUN YANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2003

Major Subject: Biology

MINIATURE INVERTED REPEAT TRANSPOSABLE ELEMENTS IN RICE – ORIGIN AND FUNCTION

A Dissertation

by

GUOJUN YANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Timothy C. Hall (Chair of Committee)	Deborah A. Siegele (Member)
James W. Golden (Member)	Keerti S. Rathore (Member)
	Vincent Cassone (Head of Department)

May 2003

Major Subject: Biology

ABSTRACT

Miniature Inverted Repeat Transposable Elements in Rice – Origin and Function. (May 2003)

Guojun Yang, B.S., Sichuan Teachers' College;

M.S., Guangxi Agricultural University

Chair of Advisory Committee: Dr. Timothy C. Hall

Transposable elements (TEs) are interspersed repetitive sequences that are present in most genomes. Miniature inverted repeat transposable elements (MITEs) are the most numerous Class II elements in higher eukaryotes. Little is known about their origin, transposition and function. In this study, three novel MITE families – *Kiddo*, *MDM1* and *MDM2* –were identified in the rice genome. They bear terminal inverted repeats (TIRs) and show target site duplications (TSDs) at the insertion sites. Each family is present in hundreds of copies with lengths that range from 200 bp to 400 bp. An evolutionary relationship between *Mutator* elements and the *MDM1* and *MDM2* families was established. The absence of an observed transposition event, together with the mutated ancestral elements identified by *in silico* analysis, led to a conclusion that *Kiddo* and its autonomous elements are not presently active.

To overcome laborious and time consuming manual analysis of MITEs on a genomic scale, MAK, a computational tool kit, was developed to automatically retrieve MITE sequences, their neighboring genes and ancestral elements from genome sequences. MAK has been functionally tested and is now available to the research community.

Studies on the effect of MITE (*Kiddo* and *MDM1*) insertions into a rice ubiquitin (*rubq2*) promoter revealed a two-edged role of MITEs on gene regulation. While *Kiddo* and *MDM1* contribute ~40% to *rubq2* promoter activity, they also induce progressive silencing of this promoter. The evolutionary implications of the two-edged role of MITEs in gene regulation are discussed.

TO MY WIFE

TO MY PARENTS

TO MY SISTERS AND BROTHERS

ACKNOWLEDGMENTS

Studying at Texas A&M University has been one of the most important events in my life. I thank the Department of Biology for providing me this opportunity. During this time, tremendous support and help from many kind people made my graduate life joyful. Each of them is special, yet the high standards and qualities they share will leave a lasting mark on me.

I give my foremost thanks to Dr. Timothy C. Hall, my committee chairman and an extremely supportive mentor. When I started in his lab, Dr. Hall very generously provided me with a research assistantship despite the uncertainty (at the time) of funding for the project. Dr. Hall's distinguished guidance made me understand science from an integrated view. It is the research freedom in Dr. Hall's lab that allowed the development of the topic described in this thesis. Being unusually busy with many things, Dr. Hall is always patiently available for problem solving, research discussion, manuscript editing or for personal help. His critical thinking and serious attention to research made quality publications possible. Dr. Hall not only greatly encouraged me on scientific endeavors, he and his wife, Ms. Sandra Hall, also provided us a friendly home, where we superbly enjoyed parties, celebrations, dinners and various games.

Another most supportive professor is Dr. Deborah A. Siegele, to whom I owe multiple thanks. Her professional guidance in my course selection and degree planning made my course studies very efficient and smooth. During my studies on *E. coli* outgrowth, she taught me frontier knowledge in microbiology and cutting edge technologies, including gene

array analysis, pulse-chase labeling and 2-D gel electrophoresis which will be to my life-long benefit. Her generous financial support during my study in her lab greatly eased hardships. It would not be a complete acknowledgment to Debby without mentioning my appreciation for her unforgettable kindness to help me by serving on my new committee even after I left her lab.

Dr. James Golden has kindly served as my committee member since 1999. In addition to his careful study of my degree plan and proposals and giving thoughtful suggestions, he has been very helpful in guiding me to review the latest developments in basic aspects of biology that are easily ignored. I also thank him for his critical thinking about my study. Dr. Keerti Rathore, the outside department committee member also kindly agreed to serve as my GCR since my previous GCR, Dr. Carol Higham from the History Department, left the university. His expertise in transgenic plants and his kind permission for me to use his imaging system is a big boost in my research.

My graduate study would not have proceeded so smoothly without the tremendous help from the Graduate Advisory Committee. Dr. Mark Zoran and Dr. Duncan Mackenzie, the present and former graduate advisors, greatly helped in my graduate study planning and progression. Their consultation, advice and strong support unfailingly helped me to quickly overcome difficulties or to fulfill a goal. I also thank Dr. Kay Goldman, the academic advisor, for her zealous help and very detailed advice.

I would also like to thank Dr. Michael Manson, Dr. Deborah Siegele, Dr. James Golden, Dr. Deborah Bell-Pedersen, Dr. Rodolfo Aramayo, Dr. Jim Hu, Dr. Thomas McKnight, Dr. Allan Pepper, Dr. Ellen Collison and Dr. Jin Xiong for their wonderful

classes. Particularly, it was Dr. Jin Xiong's bioinformatics class that led me to the world of Perl and Bioperl. Teaching is an important aspect of my study and I thank Dr. Nina Caris and Ms. Tonna Harris-Haller for their guidance in teaching.

My research involved the active participation of highly skilled experts in the lab. Mr. Yiming Jiang contributed greatly to tissue culture and transgenic rice studies and I am indebted to him for instruction in techniques for rice transformation. Dr. Yeon-Hee Lee also contributed a lot to the studies on transgenic rice and downstream analysis. Ms. Addie Embry, an undergraduate student in the Genetics Program, has been a great help with her molecular biology techniques in many experiments. Mr. Larry Harris-Haller greatly contributed to oligo synthesis and DNA sequencing. Ms. Allison Myrick helped more than just in getting information, documents, materials and managing events.

Scientists in the lab including Dr. Jinjiang Dong and Dr. Magda Cervera were very helpful in getting me started on my research projects. Dr. Mahesh Chandrasekharan was particularly helpful in discussing a variety of topics and providing me some of the starting experimental materials and other help, such as providing the electronic template for this thesis. My fellow graduate students and scientists shared their knowledge and friendship. They are Dr. Guofu Li, Dr. Iyer Lakshimi, Mr. Tao Wang, Dr. Sophie Fernandez, Mr. Bin Zhou, Ms. Xiangyu Shi, Mr. Xin Zhou, Ms. Prapapan "Lee" Teerawanichpan, Dr. Rakesh Pancholy, Mr. Danny Ng and Dr. Raul Carranco.

I thank Charlie Harris at IDMB and David Reed at the Biology Department of Texas A&M University, Hao Yu and Hemanth Sundaram at the supercomputing facility at Texas A&M University for exceptional help with computing techniques. I also thank Jason Stajich

and the bioperl community for advice on bioperl modules and Dr. Tom Bureau for help in obtaining the *Tc8* sequences.

Many other people have been very helpful in improving my research and study. They include Dr. Hongbin Zhang, Dr. Jeffrey Chen, Dr. Siva Kumpatla and Dr. Vincent Cassone, the present department head.

My study was kindly supported by a graduate teaching assistantship from the Biology Department (1998, 1999, 2000), a graduate research assistantship in Dr. Deborah Siegele's lab (1998, 1999) and a graduate research assistantship in Dr. Timothy Hall's lab (2000-2003). This work is supported in part by NSF grant MCB-0110477 to Dr. Timothy Hall.

TABLE OF CONTENTS

CHAPTER Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

EUKARYOTIC GENOMES ARE RICH IN REPETITIVE ELEMENTS

Genome size varies among different organisms. However, according to the degree and pattern of redundancy, the contents of every genome can be categorized into repetitive sequences or non-repetitive sequences. Non-repetitive sequences are present in single copies. A single copy sequence may contain repeat elements, depending on the level of analysis. These non-repetitive sequences are largely responsible for genic regions, of which only a very small proportion (e.g. 3% of the human genome) code for proteins. The remainder is intronic or spacer DNA. Repetitive sequences have more than one copy in a genome and they are present in different forms.

Classification of repetitive elements

Repetitive elements are best categorized in the human genome. Classification of repetitive elements in other organisms frequently refers to the same three categories: low repetitive sequences, moderately repetitive sequences and highly repetitive sequences. Low repetitive sequences have less than 10 copies in a genome and are sometimes merged into the non-repetitive sequences. They are often the result of gene duplication processes (1,2). Non-repetitive and low repetitive sequences make up about 60% of eukaryotic genomes (3-6). However, a very small proportion (~5%) of these sequences encode proteins.

This dissertation follows the format of *Nucleic Acids Research.*

Moderately repetitive sequences have copy numbers between 10 and 10^5 per genome; they are found throughout the euchromatic regions. These elements can be classified further into redundant genes, microsatellites, minisatellites and dispersed-repetitive DNA sequences. Microsatellites are also known as short tandem repeats (STRs). They are short runs (<100 bp) of tandemly repeated DNA with repeat units of less than 6 bp. They can be further broken down to monomeric, dimeric, trimeric, tetrameric, pentameric and hexameric repeats according to the size of their repeat units. Dinucleotide repeats of CA/GT and CT/GA are the most abundant dinucleotide repeats in the human genome. They are distributed almost ubiquitously throughout eukaryotic genomes (6). Microsatellites are highly polymorphic especially in dinucleotide repeats and thus are of high value for evolutionary information. The locations of microsatellites frequently include introns, untranslated regions (UTRs) and coding regions. Minisatellites are also known to be made up of the telomeric minisatellite group and a variable number of tandem repeats (VNTRs) (7,8). VNTRs contain repeats of 9 - 24 bp units often found clustered near telomeres and have total lengths between 0.1 and 30 kb. They are highly polymorphic but share a core sequence of GGGCAGGANG (7). Telomeric DNA sequences are tandem oligonucleotide repeats up to 10-15 kb in length depending on the specific organism, thus they also belong to minisatellite sequences. Dispersed-repetitive sequences form another important group of moderately repetitive sequences. They comprise mainly mobile genetic elements or their relics, which are described later in this chapter in more detail.

Highly repetitive sequences often refer to satellite DNA that contributes about 10% to 15% to most eukaryotic genomes (9-11). These sequences are present at more than 10^5

copies per genome and have variable length units (5 to several hundred bp) in long tracts from 100 kb up to 100 Mb. Most of these elements are located in heterochromatin regions adjacent to the centromeres or telomeres. Alpha-satellite DNA in human cells typical consists of highly repetitive sequences (6).

Repetitive sequences play major roles in cell

Although a large number of repetitive elements have been identified, the function(s) of these elements in life has been a debate. Until recently, many repetitive were thought to be junk or selfish sequences that do not provide any meaningful function in the cell. In fact, many essential functions or structures in a cell are provided by repetitive elements.

1. Redundant genes provide basic structural and functional proteins

Essential proteins are sometimes encoded by redundant genes and one strategy for obtaining large quantities of a given protein appears to be through the presence of multiple gene copies. Redundant genes are often constitutive genes required for essential functions for life forms. In eukaryotic genomes, genes coding for histones, actins, tublins, ribosomal RNAs and ribosomal proteins belong to moderately repetitive sequences.

Histones are basic proteins that constitute about half of the nuclear proteins. There are six principal types: H1, H2A, H2B, H3, H4 and H5. All of them take part in the formation of chromatin. Histone H2A, H2B, H3 and H4 form the octameric core of a nucleosome unit. Histone H1 and H5 are believed to act as linker histones. Further coiling of chromatin results in the chromosome. Thus, histones are structural components of chromosomes. Studies in recent years yielded a multitude of evidence that histones play a major role in gene regulation. Histone H1 is thought to be a part of a general repressor

mechanism, which ensures potent repression of gene expression in large chromatin fragments. In addition, H1 may be involved in controlling the transcription of individual genes by repression or stimulation. The N-terminal domain of histones H3 and H4 has been implicated in various nuclear functions, including gene silencing and activation and replication-linked chromatin assembly. Various modifications on histone amino-termi result in synergistic or antagonistic interaction affinities for chromatin associated proteins, which in turn determine the transition between the transcriptionally active or transcriptionally silent status of a chromatin locus. This observation led to a "histone code" theory that conceives modification on histone tails to be an approach to extend genetic code information (12-14).

Actin is an abundant protein and is the major component of the microfilament network of the cytoskeleton. In the past few years, it was clearly linked to nuclear processes including chromatin remodeling, transcription, and RNA splicing. Tubulins perform various essential functions. They are required for spindle formation during mitosis and meiosis, axonal transport, organelle positioning, and cilia and flagella formation.

A ribosome is built of a large and a small ribosomal subunit. The core of each ribosomal subunit is a polymer ribosomal RNA (rRNA) folded into a compact conformation. Ribosomal proteins are assembled on the RNA core. Ribosomes perform translation, a key gene expression step. They read genetic messages in mRNA and synthesize polypeptides. In addition to peptide synthesis, ribosomes are responsible for directing polypeptides to their corresponding cellular locations (15-18).

Proteins such as immunoglobulin and ubiquitin are also present as multiple copy genes; they are involved in essential functions such as defense and global regulation.

2. Satellite DNA contributes to chromosomal organization and cell division

All eukaryotic chromosomes contain centromeres and telomeres. Centromeres are key structures in chromosome organization and are attachment sites for kinetochores, which are responsible for the segregation of chromosomes during cell division. Although eukaryotic centromere DNA shares little sequence similarity between species (19), all centromeric DNA binds a histone H3-like protein (named CENP-A in humans) to form the basic building blocks of the centromeric chromatin (20-25). Lately, a subset of about 100 kb from 1-4 Mb α-satellite DNA in humans has been shown to be the major component of CENP-A-bound DNA ((26).

3. Telomeric DNA sequences protect ends of chromosomes from progressive shortening

Telomeric DNA consists of an array of DNA repeats of oligomers belonging to the moderately repetitive sequences at the very end of the chromosome and a complex array of repeats belonging to highly repetitive sequences at the subterminal regions. Telomeric DNA sequences are bound by proteins to form large complexes. These structures are thought to form nuclear domains that are important for transcriptional regulation, sister chromatid pairing during mitosis, and homologous meiotic synapsis. More importantly, extreme end repetitive sequences of telomeres are directly involved in a mechanism to prevent progressive shortening of chromosome ends, a replication paradox raised by the classical replication theory.

4. The role of dispersed-repetitive sequences

In the early years following the discovery of transposable elements (TEs), these repetitive sequences were considered 'junk' or 'selfish' DNA sequences (27). Because of their mobility

and abundance, TEs are potentially detrimental to the integrity of a genome. In contrast, further, more recent, studies have demonstrated that TEs can confer important effects on gene regulation (see Page 18). However, the significance of the effect of TEs on gene regulation in the genome is still in debate and thus not widely accepted (28-33).

Accumulation of repetitive elements during evolution

Repetitive DNA sequences exist in all cellular life forms, including eubacteria, archaebacteria, fungi, plants and animals. However, the types and abundance of repetitive sequences differ dramatically between species or even between organisms within a species. During the course of evolution, new elements are continuously formed and existing elements gradually die out. Since the mobilization of TEs is in some degree by chance, common ancestors can give rise to different fates in their progenies. The differential accumulation of types and the number of copies of a specific type of TE in individual organisms reflect the life history of the TE.

To date (Feb 17, 2003), sequences for 96 prokaryotic organisms have been sequenced and published (http://www.tigr.org), of which 80 are bacterial and 16 are archaean. The sizes of these genomes differ by more than 10 fold (34). In general, compared to eukaryotes, they are fairly small with the largest genome being 9.2 Mb for *Mycoplasma genitalium*. Although prokaryotic genomes are compact, they contain a significant proportion of repetitive sequences. These sequences fall into duplicated genes (e.g. tRNA, rRNA and sRNA), non-coding repeated sequences (e.g. short repeated sequences (SRRs) (35), REP (36,37) and ERIC (38)), TEs and insertion sequences (IS). In the 96 published genomes, 330 IS, 711 non-coding repeat elements, 221 sRNA, 956 rRNA, and 5211 tRNA elements were

identified. The fraction of repeated regions >200 bp in the genomes ranges from 0.23% (*R. prowazekii*) to 6.2% (*Neisseria meningitidis*), with an average of 1.7%. In addition, hundreds of non-coding repeat sequences with sizes smaller than 200 bp were identified in bacterial genomes (36-42). Recently, comparative genomic studies on prokaryotic genomes has revealed a correlation between repeat content and lifestyle (43). While free-living bacteria have large genomes with a high content of repeated sequences and self-propagating DNA, obligate intracellular bacteria have small genomes with a very low content of repeated sequences. Genome sequence analysis of a bacterium that tolerates extreme radiation treatment (*Deinococcus radiodurans*) revealed that, compared to other bacteria, the *D. radiodurans* genome is enriched in repetitive sequences including ISs and small intergenic repeats (44). Thus, the accumulation or retention of repetitive sequences seems to be needed for adaption to changing or challenging environments. The presence of microsatellites in prokaryotic coding sequences of contingency genes has been shown to provide advantages under changing environmental conditions. Phase variation in *Neisseria gonorrhoeae* is dependent upon the antigenic profile variation caused by changes in the length of pentamer microsatellites contained in the outer membrane proteins.

Although VNTR analogs are present in prokaryotic genomes (35), prokaryotes are not generally considered to contain minisatellites or satellites. Retrotransposable elements are another group of repetitive sequences, but they have not been seen in prokaryotes.

In contrast to the small genomes for prokaryotes, eukaryotic genomes have much larger genome sizes. The large variation in size of eukaryotic genomes appears to have little relation to differences in organismal complexity or numbers of genes. This phenomenon is

called C-value paradox. Much of this variation is caused by the different amount of repetitive sequences in different genomes. All the three major categories of repetitive elements are found in almost every eukaryotic genome. Satellite DNA constitutes 10-15% of eukaryotic genomes and moderately repetitive sequences constitute more than 30%. The proliferation of repetitive elements in eukaryotes relies on three types of events: increase in the types of repetitive elements at the category, subcategory, superfamily, subfamily and family level; increase in the number of elements in each type of repetitive element; increase in the size of elements.

TRANSPOSABLE ELEMENTS IN EUKARYOTIC GENOMES

The first transposable element (TE) was reported by Babara McClintock in the 1940's (45). As a significant proportion of moderately repetitive sequences, transposable elements are interspersed in a genome thus are also named interspersed repetitive elements. TEs are the largest component of most eukaryotic genomes. TEs comprise more than 35% of the human genome, 14% of the *Arabidopsis* genome and 50-80% of grass genomes (46-49).

Classification of transposable elements

According to their mechanism of transposition, TEs can be grouped into two classes: class I being the retrotransposable elements and class II being the DNA transposable elements. Each class contains various superfamilies, which in turn contain a variety of families or subfamilies.

1. Retrotransposable elements

Retrotransposable elements are related to retroviruses, the major difference being the absence of functional envelope genes (*env*) in retrotransposable elements. This class of TE requires the RNA intermediate to be transcribed from the element for transposition and amplification. They are usually present in very large copy numbers in eukaryotic genomes. Depending on the presence of long terminal repeats (LTRs) flanking the coding sequences, they can be divided into LTR retrotransposons and non-LTR retrotransposons (Fig. 1.1).

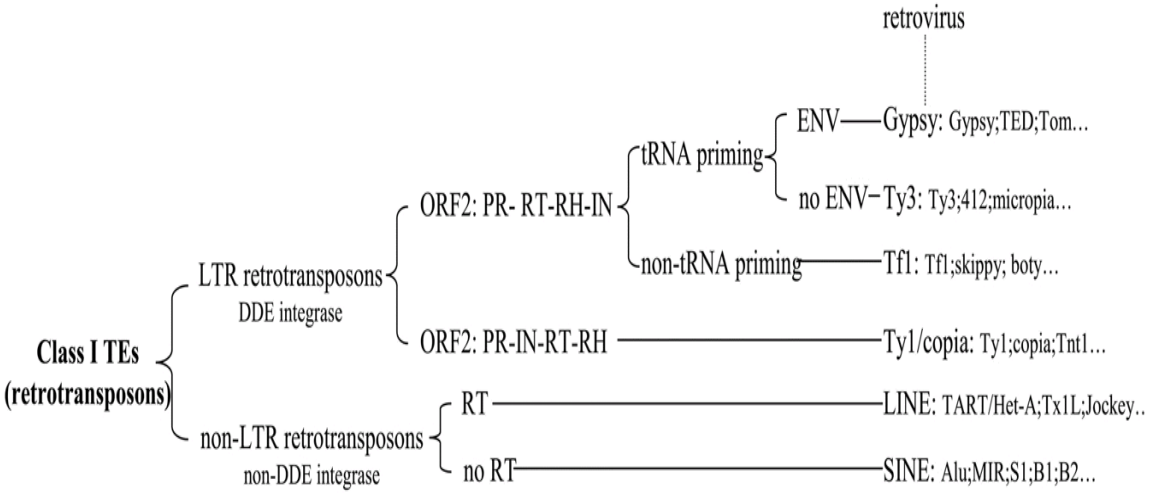

Figure 1.1. Classification of class I transposable elements (retrotransposons).

LTR retrotransposons contain LTRs similar to those of the retroviruses. However, non-LTR elements do not contain LTR sequences.

LTR retrotransposons are classified into three superfamilies: the Ty1-*copia* superfamily, the Tf1 superfamily, and the Ty3-*gypsy* superfamily. As shown in Fig. 1.2, while Ty1-*copia* elements have an integrase gene (*int*) between the protease gene (*pr*) and

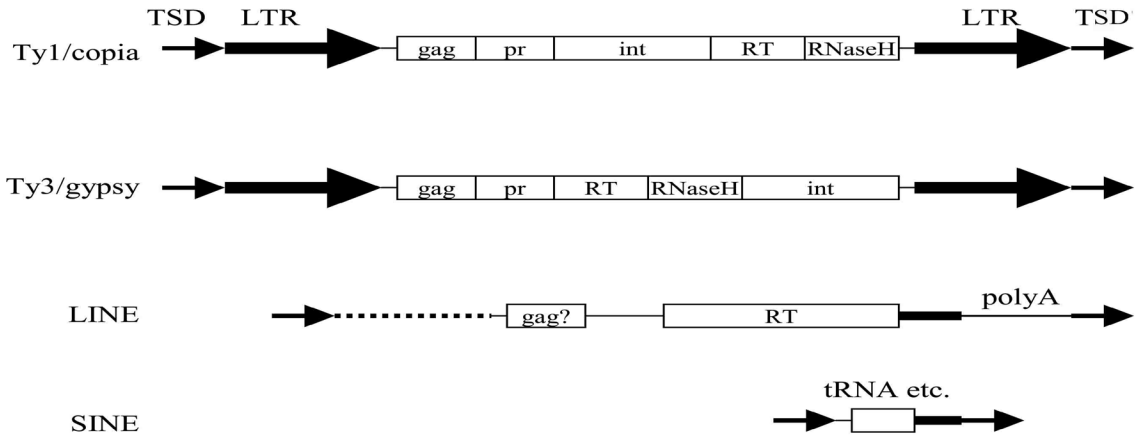

Figure 1.2. Diagrams for typical retrotransposons (see text).

the RT gene, the *int* gene is at the 3' end of the polyprotein cassette in Ty3-*gypsy* elements.

Non-LTR retransposons are classified into two superfamilies: long interspersed elements (LINEs) and short interspersed elements (SINEs). LINEs have the typical structure of a eukaryotic gene, consisting of 5' untranslated regions (5' UTRs), coding sequences, 3' UTRs and poly A tails. They do not contain *int* genes but instead contain relics of the *env* gene. While the 3' ends of LINE families are stable, their 5' regions are usually variable.

SINEs are usually short (80 -500 bp) and they do not contain *gag-pol* protein coding regions. Instead, they carry internal sequences originating from some other genes or tRNA genes. Thus, SINEs are usually subdivided into tRNA SINEs and non-tRNA SINEs. Interestingly, a SINE family contains a 3' end region that shares high similarity (up to 90%

identity) to a certain LINE family from the same genome (50). Lately, it has been established that SINEs are dependent on their corresponding LINEs for transposition (51,52).

2. DNA transposable elements

DNA transposable elements contain terminal inverted repeats (TIRs). They transpose by a cut-and-paste approach in which the original element is excised and inserted into another locus. Depending on the motif feature in the catalytic domains of their transposases, transposons can be divided into two subclasses: the ones contain a DDE motif and the ones that do not (Fig. 1.3).

The DDE subclass includes the *Tc-mariner* superfamily and the *ISa* superfamily. The *ISa* superfamily is mainly made up of prokaryotic insertion elements, including *IS2*, *IS3*, *IS4*, *IS6* and *IS30*. The *Tc-mariner* superfamily contains TE families mainly from eukaryotic genomes. *Tc1* elements from *C. elegans* and *Mariner* elements from *Drosophila* are the founding families of this superfamily. In the past several years, a large novel superfamily

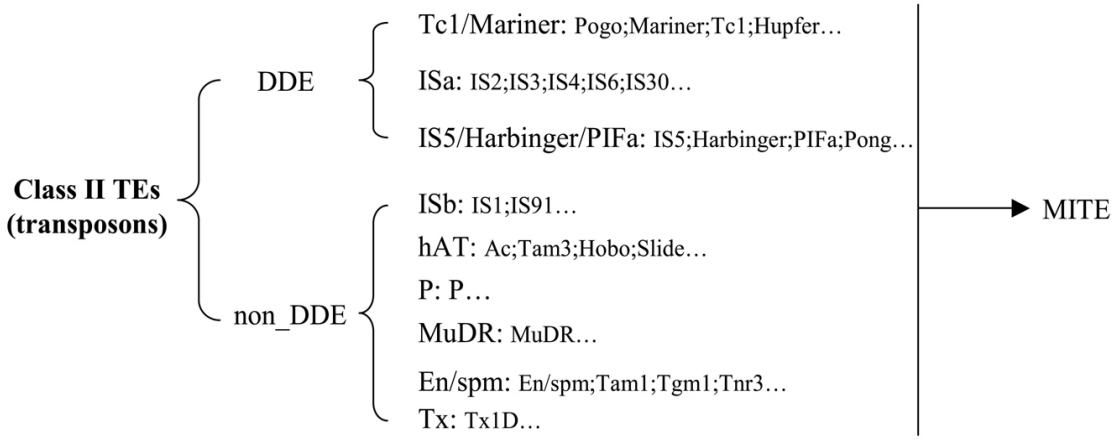

Figure 1.3. Classification of class II transposable elements (DNA transposons).

named *IS5/Harbinger/PIFa* has been discovered and is presently being actively studied (53-56).

The majority of known transposons fall into the non-DDE subclass. The hAT superfamily contains the first identified transposable element family *Ac/Ds*. The *Ac* superfamily has a target site duplication (TSD) of 8 bp and a TIR of 11 bp. The *P* element superfamily, originally identified in *Drosophila*, is the best documented family in animal systems. Evidence for horizontal transmission has been shown for *P* elements (57,58). The *Mutator* superfamily contains numerous decayed sequences and they form very long TIRs (up to 250 bp) (59,60). *Mutator*s have a TSD length of 9 bp. The *En/Spm* superfamily is also known as the CACTA superfamily because of the strong conservation of this motif in its TIR. It has a TSD length of 3 bp (61-64). A superfamily, namely *ISb*, containing elements mainly from bacteria also belongs to the DDE subclass. A new superfamily of Tx, identified from *Xenopus laevis*, promises to be an interesting example of a composite TE family. Tx elements frequently carry members of a non-LTR retrotransposon family, Tx1L. The entire composite units are mobilized through collaboration between the *pol*-like gene product and the TIRs (65-67).

Transposition mechanisms

The two classes of TEs take different approaches to transpose and they have different consequences. Being structurally similar to retroviruses, retrotransposon transposition takes a similar approach to that of retroviral integration. The master elements of a retrotransposon family are transcribed from a promoter inside the U3 regions of 5' LTRs. Reverse transcriptases synthesized from the retrotransposon mRNAs use the RNA intermediates as

templates to synthesize cDNA strands. The newly synthesized DNA strands insert into a new location. Because the original copy of the retrotransposon is not excised, the insertions of retrotransposons are permanent. Since reverse transcriptase is error-prone, the new copies of the retrotransposons frequently contain mutations that may damage their coding capacity. However, they may still be able to be copied and pasted to new locations by the enzymes produced from functional master elements.

Since LINEs have dramatic structural differences from LTR retrotransposons (51,68-70), their transposition mechanism is thought to be different from that of the typical retrovirus. The autonomous LINE elements encode a protein with both endonuclease and reverse transcriptase activity (71). This enzyme is thought to guide transposition through target primed reverse transcription (TPRT) (72). In TPRT, the transcripts of LINEs produce proteins that are required for transposition and DNA strand cleavage. One strand of the target site DNA is thought to be cleaved by the endonuclease activity, leaving a free 3'-hydroxyl group at the nick site. The 3' ends of the LINE transcripts pair with the nicked DNA and the 3' OH ends are used as primers to guide reverse transcription. Upon the completion of cDNA strand synthesis, the other DNA strand at the target site is cleaved and the RNA moiety of the hybrid molecule is removed. Subsequent synthesis of the second DNA strand of the LINE sequence and the repairing of the nick site results in variable length TSDs flanking the new LINE element.

The similarity of the 3' end regions between SINEs and corresponding LINEs is considered to be a critical feature that underlies the functional importance of the 3' end

regions. It is proposed that the transposition of SINEs is dependent upon their corresponding autonomous LINEs (50).

Unlike retrotransposons, transposons do not use RNA intermediates for transposition. In the conservative transposition approach, the DNA sequences of transposons are physically excised by transposases expressed from the corresponding autonomous elements at the border between the TEs and flanking genomic sequences. The ends of the excised elements are bound by proteins. The target site in a new location is nicked or double strand cleaved by the endonuclease activity of the transposase proteins. Insertion of the excised sequence and subsequent repair of the cleaved target site results in TSDs that flank the newly inserted element. Since the DNA element is only cut from one place and pasted to another, point mutation frequency is relatively low although truncation of the sequences is often seen. Ligation and repair of the excised donor site often results in a footprint that consists of two copies of the previous TSD sequences (73,74). Replicative transposition, in which a replication recombination event is responsible for transposition and amplification of phage *Mu*, is seen in prokaryotes but has not been reported in eukaryotic organisms. However, when a transposition event occurs at a replication fork, it may be responsible for the amplification of eukaryotic transposons in a duplicative way (75). In this event, a transposon on one of the two daughter molecules immediately behind a replication fork excises and inserts into another location. The excised locus is repaired by a homologous recombination using the other daughter molecule as the template, resulting in a net gain in the total number of transposons.

Evolution of transposable elements

Retrotransposons and transposons seem to be different; nevertheless they have some common characteristics. The most obvious common feature is their ability to translocate. In addition, both types of element need to integrate into new locations. The mechanism of transposition appears to be very similar for transposases used by transposons and for integrases used by retrotransposons (76). The relationship between these two classes of elements is supported by studies on the evolution of TEs (Fig. 1.4) (77,78).

It is thought that non-DDE elements are derived from a variety of ancestral sources. The *P* and *Tn3* elements seem to be derived from a common ancestor bearing resolvase activity. Piv and some *IS* elements may have a common ancestor containing a ancestral

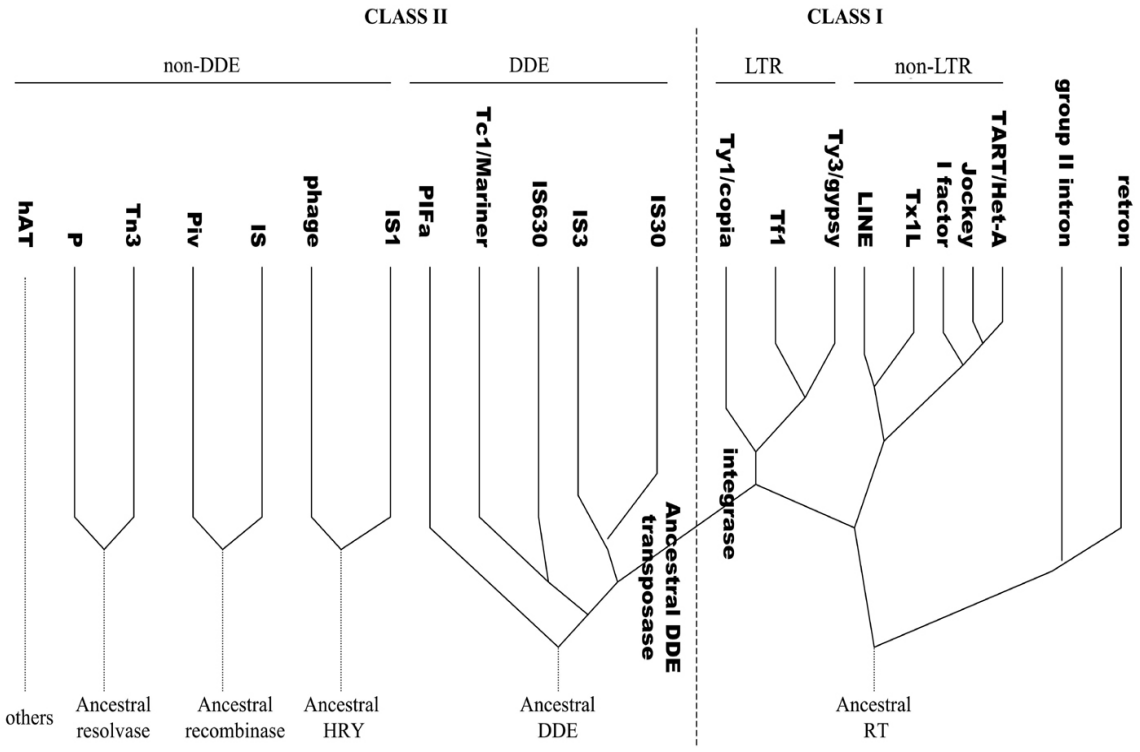

Figure 1.4. Evolution of transposable elements (see text).

recombinase gene. Some bacterial phages and *IS1* appear to be derived from ancient proteins containing *HRY* domain. Ironically, no evolutionary origin has been suggested for *Ac* superfamily, the first identified transposable element.

All the DDE elements are thought to be derived from a common ancestral element. The highly similar catalytic domains of D(50-70)D(35)E of transposases and integrases support the evolutionary relationship between transposons and retrotransposons. The common ancestor of retrotransposons is believed to be an ancestral reverse transcriptase (*RT*) protein. This ancestor may have also given rise to modern group II introns and retrons. On the way to LTR-retrotransposons, ancestral non-LTR elements branched out and differentiated into modern day *LINEs*. *SINEs* are thought to relate to *LINEs* in the 3' end regions. Acquisition of *gag* and *RNase* H genes by the ancestral *RT* protein, followed by incorporation of the *int* gene, is thought to have given rise to ancestral LTR retrotransposons. Further acquisition of the *env* gene by *Ty3-gypsy* like elements may have resulted in the ancestral retroviruses. However, another school of thought puts the ancestral retroviruses as the starting point and retrotransposons as the derivatives of those primitive retroviruses (79).

The role of transposable elements in genomes

In canonical genetic studies, the genome was pictured as an essentially stable network that had evolved to yield a viable organism. The components inside an existing genome are postulated to have been selected against the natural environment. The discovery of TEs led to the concept of repetitive elements as selfish DNA, junk DNA or parasitic DNA (27,29,31,32,80-82). With more discoveries, the selfish DNA theory is under challenge since TEs have been shown to contribute to essential aspects of life (83-87). One issue being

debated is wether the prevalence of TEs in a genome is primarily because of their replicative advantage over the host genome or because of selected retention as a result of their contributions to the genome. Nevertheless, both sides of the debate are rigorously checking interactions between TEs and the rest of the genome.

1. Challenges to genomes imposed by TEs

The overwhelming presence of TEs in a genome may impose disadvantages on the genome. The basis for potential damage to the genome is their ability to insert at new locations. As summarized by Bennetzen (88), negative outcomes resulting from transposition events include an increase in genome size, the creation of intronless pseudogenes, the movement of genes or gene segments to new sites, the creation of new chromosomal folding patterns and chromosomal rearrangement or breakage. The majority of these changes must be deleterious, as is the case for other forms of mutation. As a consequence, host extinction or reshaping of the structure of the genetic system may occur (29).

2. Contributions of TEs to genome evolution

Since the presence of TEs in genomes is a 'fact of life', all existing genomes represent the successful co-evolution of the TEs and host genome. Many essential genetic structures and functions have been shown to be the result of interaction between TEs and host genomes. Looking back into evolutionary history, TEs contributed to the formation of modern genomes. The contributions include formation of new genes, genome reorganization, formation of gene silencing mechanisms, speciation, telomere maintenance, immune system development, heterochromatin formation, manipulation of host sex and gene expression modulation (4,32,33,89-91).

3. TEs and gene regulation

Differential TE insertion can affect the expression of flanking genes in various ways, and even the same insertion can have different effects depending on the activity of the TE family. The following types of effects have been observed to date (92):

• *TE insertion alters the level of expression*

An insertion of a *Ty1* element into the promoter region of the *CYC-7* locus has been shown to result in a 20-fold increase in expression of the gene as well as of the *Ty1* element (93).

• *TE insertion regulates expression in different ploidy*

The enhancement described by Errede *et al.* (93) is specific to haploid cells. It is repressed in diploid cells by the products of the MAT locus (94). The presence of SV40 enhancer-like sequences (95) and mating-type responsive elements (96) in the LTR of the *Ty* element may be responsible for the observed phenomena.

• *TE induced tissue specific expression*

The *Drosophila* element *tom* has been shown to drive ectopic eye expression of the *Om(1D)* homeobox gene (97). Sequences inside the *tom* element are believed to be responsible.

• *Orientations of TE insertion have different effects on expression*

A *Tam3* insertion in the *pleana* intron resulted in a loss of function phenotype when the *Tam3* element inserted in the same orientation of the *pleana* gene. However, expression of the gene was seen when the *Tam3* insertion was in the opposite orientation (98).

• *TE insertion can result in suppressible alleles*

Gypsy insertions into the 5' region of genes act as insulators that block gene activation by distal enhancers (99-103). The phenotype caused by this insertion requires binding of the

ubiquitous nuclear protein *Su(Hw)* to the 12 octomers on the 5' UTR region of the gypsy element (100,104,105).

The effects of *Mutator* insertion into *knotted1*, *rough sheath1* and *liguleless3* intron or UTR regions are dependent on its activity (106). The presence of mutant phenotypes is directly correlated with *Mutator* activity. The mechanism of this suppressible expression remains to be determined and a mechanism related to gene silencing may be one possibility.

Unlike *Mutator* insertions, *Spm* insertions correlate with insertion orientation. When the insertion of a *Spm* element into an exon is in the opposite orientation to that of the target gene, the target gene is expressed if *Spm* is active. However, the target gene is expressed because of the aberrant splicing of the *Spm* element when *Spm* is not active (107). The 5' UTR of *Spm* is thought to be capable of serving as an insulator against position effects (108).

• *TEs can act as promoters*

TEs including *Mutator*, *Tam3*, and the retrotransposon *IAP* have been shown to drive transcription of target genes from their terminal or subterminal sequences (109-111).

MINIATURE INVERTED REPEAT TRANSPOSABLE ELEMENTS

Discovery of miniature inverted repeat transposable elements

Following the discovery of the *Ac* transposon by McClintock (45,112), detailed studies have been undertaken on the *Ac/Ds* superfamily. Among the identified elements, *Ds1* elements are particularly interesting. They are of 405 bp and have only the 5'-terminal 11 bp and 3'-terminal 26 bp in common with *Ac* (113-120). The internal regions contain transposase binding sites and do not share significant similarity to *Ac*. More interestingly, *Ds1* is not only

transposed by *Ac* but also by *Uq*, which does not mobilize other large *Ds* elements (121,122). They exist in a high copy number in the *Maize* genome.

Reports of small TEs have accumulated rapidly since 1992. They are present in almost every higher eukaryotic genome, including human, plant, insect and fish (54,123-146). Their characteristics include small size (100-500 bp), TIRs and a large copy number. They cause TSDs at their insertion locus.

MITE identification and analysis method

Early MITE families were identified following the observation that some alleles from different lines of a species contain short insertions that bear TIRs and cause TSDs (113,142,144). Subsequently, Southern blots were used to check their abundance. With advances in genome sequencing, many more MITE families were identified by database searches (123,124,139,146-150). To better understand these MITE families, database searches were carried out to retrieve MITE sequences, find neighboring genes and predict autonomous elements.

These analyses are currently undertaken manually with the help of the BLAST program (151). However, because of the huge copy number of elements, manipulating these elements on a genome scale is laborious and time-consuming. Automation of this process is very important for the analysis of MITEs on a genomic scale.

Mysteries surrounding MITEs

Although a variety of MITE families have been identified, they are largely not well understood. The following questions remain to be answered:

1. Origins of MITEs

Since MITEs do not appear to encode protein(s) for their transposition, they probably depend on *trans* factors provided by other loci. This dependence requires the recognition of the features on MITEs by the *trans* factors. Most autonomous TEs require both terminal and subterminal regions for efficient transposition. Since MITEs only show similarity to known DNA transposons at TIRs, are MITEs the abbreviated versions of other cognate long elements or they are genomic sequences armed with TIRs?

2. Amplification of MITEs

All MITEs bear TIRs and TSDs that are reminiscent of DNA transposon characteristics. However, their large copy numbers resemble those of the retrotransposable elements. Although TSD footprints have been reported, no MITE has been shown experimentally to excise and insert. Since the internal sequences of MITEs do not share significant similarity to any widely known transposons, what is responsible for their transposition? As different MITE families have different sequences both at TIRs and internal regions, are all of these MITE families dependent on a single transposition mechanism, or does each MITE family have its own mechanism?

The large copy numbers seen for MITEs are the result of amplification through transposition. How have these elements achieved such high copy numbers and which mechanism of transposition do they use?

3. Function of MITEs in a genome

Some MITE families are preferentially associated with genes (142,144), while others exhibit no preference for genic regions. In either case, many elements are found very close to genes,

either on promoters, introns or UTR regions. Some elements are even found in coding sequences (Yang and Hall, unpublished data). Because of their overwhelming presence in a genome, it is hard to imagine that they do not have any role in gene regulation or evolution. Up to date, no convincing example has been reported for the effect, even a disruptive effect, of MITEs on genes.

CHAPTER II

DISCOVERY OF MITE FAMILY *KIDDO**

INTRODUCTION

Following the discovery of transposable elements (TEs) in maize by Barbara McClintock (45), TEs have been found to be ubiquitous in biological organisms. DNA TEs are classified as transposons and elements that transpose through an RNA intermediate are classified as retrotransposons (152). A typical transposon family has two components – an autonomous part which expresses a transposase and a non-autonomous part which depends on the transposase for transposition. DNA element structures include terminal inverted repeats (TIRs) at both ends, target site duplications (TSDs) in genomic DNA that border the transposon, and internal sequences that usually share substantial similarity among members of a given transposon family. Excision of the element from genomic DNA leaves the two TSDs positioned together. A few nucleotides are sometimes removed at the insertion site by nucleases during the repair process (88). The excised transposon can then integrate into another genomic DNA location. The insertion event usually results in direct duplication of the target locus sequence.

In recent years, numerous small transposon-like elements have been found in many organisms. Collectively, they are called miniature inverted repeat transposable elements (MITEs) (123). Each MITE family has a distinctive TIR and TSD, and a similar internal

---

*Reprinted with permission from "*Kiddo*, a new transposable element family closely associated with rice genes" by G. Yang, J. Dong, M. B. Chandrasekharan, T. C. Hall, (2001),. *Mol Gen Genomics*, 266, 417-424. Copyright 2001 by the Springer-Verlag.

sequence; however, unlike *Ac/Ds* and *En/Spm*, they do not appear to contain an autonomous element. Examples from plants include: *Tourist* (144,145), *Stowaway* (146), *Alien* (153), *Bitfoot* (127), *Amy/LTP*, *p-Sine*, *Explorer*, *Gaijin*, *Castaway*, *Ditto*, *Wanderer* (123), *Emigrant* (125), *Krispie*, *Snap*, *Crackle*, *Pop*, *Snabo-1*, *Snabo-2*, *Snabo-4*, *Truncator* (154), *Hbr*, *mPIF*, *Olo* (138). It is likely that additional MITE families remain to be discovered. Here we report a new MITE family, which we have named *Kiddo*.

MITEs have the characteristics of both DNA and RNA elements (155). They contain typical DNA element structures, such as TIRs and TSDs. However, their very high copy number and lack of evidence for excision is more characteristic of retrotransposons. Classification of MITEs as DNA elements depends on the existence of DNA intermediates during transposition or evidence for their excision from genomic DNA. Unfortunately, no direct transposition evidence has been reported for MITEs and the only excision footprint observed thus far is in *Hbr* (138). Here we report evidence for the excision of a *Kiddo* element from rice genomic DNA.

Some MITE families of grasses, for example *Tourist* and *Hbr*, were originally found in maize. MITE family members in other grasses such as rice, wheat and sorghum have been found either by using the maize MITE sequences as primers to probe their genomic DNA, or by the use of maize MITE sequences as queries for database searches (123,156). Many MITE families originally found in maize also contain members from other grasses (138,146,154). However, this approach will not reveal MITE families, such as *Kiddo*, that exist only in rice and, for rice-specific MITEs, the study of individual gene polymorphisms remains an important way to discover new families (123).

MITEs are useful in systematics and are potentially useful as molecular markers because several are preferentially associated with genes (138,145,146). For example, *Hbr* has been successfully used as a molecular marker in maize because of its association with genic regions and the sequence homogeneity that exists within this family (130). The *Kiddo* family, which exists in a high copy number, also appears to have value as a new molecular marker since each subgroup has over 90% sequence similarity and ~80% of the 18 family members identified thus far from annotated sequences lie within 530 bp from cDNA sequences or from introns.

MATERIALS AND METHODS

PCR amplification of the *rubq2* promoter

Genomic DNA of rice lines T309 or IR24 (kindly supplied by Anna M. McClung, TAES, Beaumont, TX) was extracted using a hexadecyltrimethyl ammonium bromide method (157). DNA was digested with *Hin*dIII to facilitate PCR amplification of the desired product as there is no *Hin*dIII site in the target region. PCR reactions were conducted at various temperatures between 45° and 65°C with cloned *Pfu* (Stratagene) or *Taq* DNA polymerase (Promega, Madison, WI). The primers used were: 5'-aagcttacggaaggaaaacaaattcgg-3' and 5'-tctagatgcgaggagaggagatgag-3'.

Cloning and sequencing of the PCR product

PCR products were cloned into pPCR-Script™ Amp SK (+) and transformed into Epicurian Coli® XL10-Gold® ultracompetent cells (Stratagene). White colonies were selected on X-Gal LB plates with IPTG. Positive transformants were identified by agarose gel electrophoresis

after *Sac*II enzymatic digestion of isolated plasmids. Automated sequencing of the cloned PCR fragments was done by the Texas A&M Gene Technologies Laboratory using T3 and T7 promoter-specific primers.

DNA blot analysis

DNA (2 μg) from *Arabidopsis*, maize, tobacco, rice T309, rice IR24, wheat and *Camptotheca acuminata* was digested with *Hin*dIII overnight at 37°C. After electrophoretic separation of DNA on 1% agarose gel for 12-15 hours at constant 23 volts, the DNA was transferred to Hybond^{TM}-N+ nylon membrane (Amersham, Piscataway, NJ). Genomic DNA blot analysis was as described by Buchholz *et al.* (158). [^{32}P]dCTP-labeled probes were made using a DECAprime^{TM} II DNA labeling kit (Ambion, Austin, TX). Membranes were washed with 2×SSC [1×SSC = 0.15 M sodium chloride /0.015 M sodium citrate (pH7)]/ 0.1% SDS at 65°C for 1 hr (low stringency), or with 0.3×SSC /0.1% SDS at 65°C for 1 hr (moderate stringency).

Data mining and alignment

Initially, a BLASTN search was undertaken against GenBank, EMBL, DDBJ, PDB, EST, STS, GSS, and HTGS databases, current as of 12/23/2000 (National Center for Biotechnology Information, Bethesda, MD), using the 270 deletion sequence in the IR24 *rubq2* promoter region as a query. The retrieved sequences were then used as queries to do BLASTN and TBLASTX searches against the databases. Protein databases were used to retrieve any possible translation product and BLASTX searches were used to study the putative coding capacity of the retrieved elements. All retrieved sequences with expectation values lower than 1×10^{-3} and with a length >150bp were studied further. After elimination

of duplicate sequences, complete sequences were aligned using Vector NTI suite 6.0 AlignX (InforMax, Inc., Bethesda, MD) with a gap penalty of 5 and 15 and gap extension penalties of 1 and 6.66 for pairwise alignment and multiple alignment, respectively. Alignment was visualized using BOXSHADE (http://www.ch.embnet.org/software/BOX_form.html). A phylogenetic tree was produced using Vector NTI 6.0 AlignX. Folding of DNA fragments was carried out using M-fold (http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi).

RESULTS

Discovery of *Kiddo*

Oryza sativa (L.) cv. *Taipei* T309 genomic DNA, digested with *Hin*dIII, was used as a template for PCR amplification of the rice *ubiquitin2* (*rubq2*) promoter region. A single discrete product was obtained, but was smaller (684 bp) than expected (954 bp) from the database sequence (accession no. AF184280). The amplified fragment was cloned into PCRscript-Amp SK (+) and sequencing confirmed that, except for a 270bp deletion, the sequence was identical to the published sequence of *rubq2* (159). The cloned fragment was confirmed to be derived from the *rubq2* gene by genomic DNA blot analysis (data not shown).

Since the original *rubq2* sequence in GenBank was obtained from a BAC clone derived from the rice line IR24, we obtained that line and repeated the PCR amplification using the same primers as those used for T309. Cloning and sequencing of the amplified product confirmed it to be identical to the expected 954 bp of *rubq2*. New genomic DNA extracts were made and independent replicate experiments confirmed that PCR products of dissimilar

sizes were obtained from T309 and IR24 using *Taq* DNA polymerase (Fig. 2.1A), indicating an insertion of 270 bp in *rubq2* from IR24 (or a deletion in T309). Interestingly, *Pfu* DNA polymerase failed to amplify the *rubq2* fragment from IR24 under the conditions used (Fig. 2.1B), suggesting that the insertion might contain a secondary structure inimical to the polymerase. The location of the insertion in the *rubq2* promoter of IR24 is shown in Fig. 2.1C.

Inspection of the sequence of the 270 bp region present in *rubq2* from IR24 revealed that its ends constituted 14 bp terminal inverted repeats. To examine the possibility that the 270

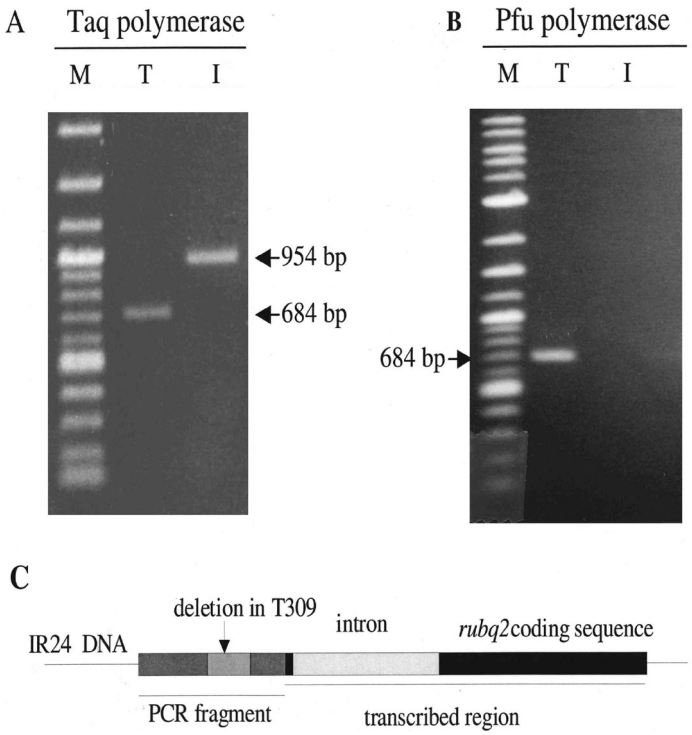

Figure 2.1. PCR amplification of the *rubq2* promoter. (**A**) Using *Taq* DNA polymerase and *Hind*III-digested genomic DNA, a 684 bp fragment was PCR-amplified from T309 (T) and a 954 bp fragment was amplified from IR24 (I). (**B**) PCR using *Pfu* DNA polymerase amplified a 684 bp fragment from T309 DNA. (**C**) Diagram of IR24 *rubq2* showing the location of the PCR-amplified region containing the *Kiddo* element.

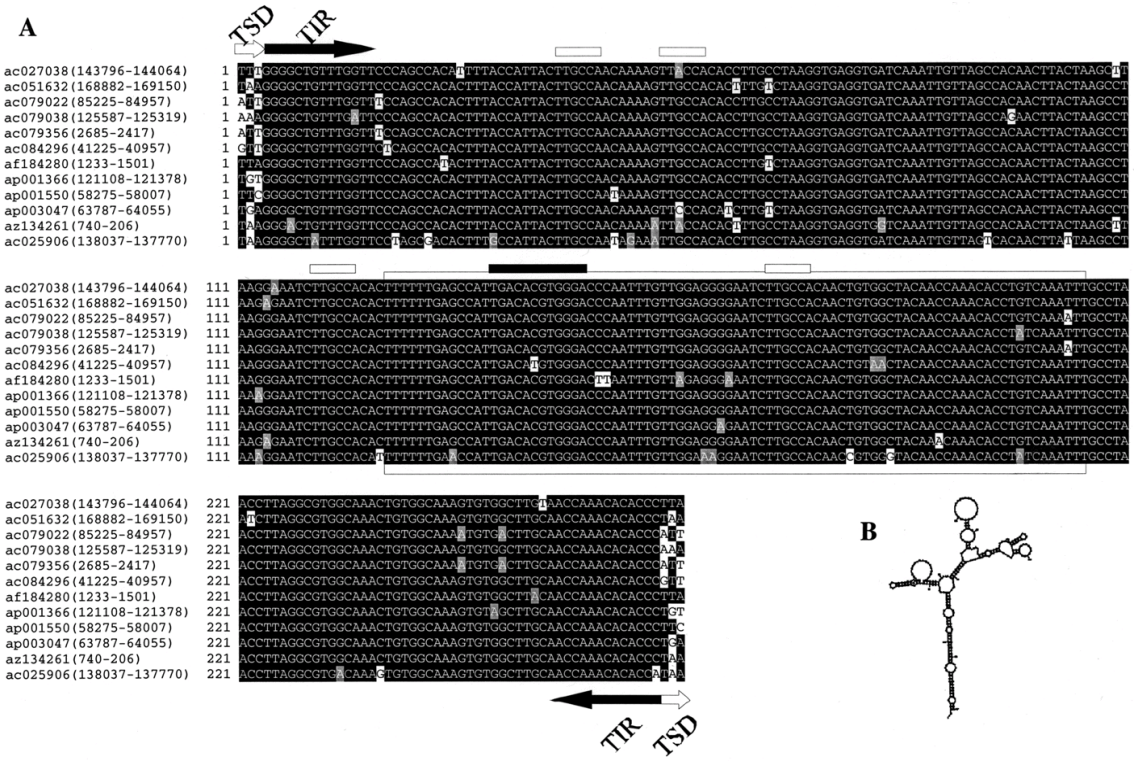

Figure 2.2. Sequence alignment of twelve genomic group A *Kiddo* members (see Appendix C). The sequences were aligned using Vector NTI 6.0 AlignX and visualized with BOXSHADE (**A**). Filled arrows refer to the TIRs and empty arrows denote TSD sequences. The black bar indicates a G box factor binding site. The empty bars denote TTGCCA repeats. The boxed region represents sequences flanked by TTTTTTGA and TCAAATTT. Folding of the 270 bp insertion fragment from IR24 *rubq2* promoter using M-fold (**B**). $\triangle G$ = -37.6 kcal/mol at 37°C.

bp insertion represented a TE, it was used as a query sequence to search the database (see *Materials and Methods*). Twelve rice sequences with high (>90%) similarity were retrieved that all had MITE characteristics of small size, TIR and TSD (Fig. 2.2A). All were 269 bp in length and, since their internal sequences had no detectable similarity to those of known TEs, we classified them as a new MITE family named *Kiddo*. This assembly of sequences (*Kiddo* group A) bear the consensus TIR, GGGGCTGTTTGGTT, with 2 mismatches at the 4th and 5th nucleotides. No polyA/polyT elements were found in the subterminal regions. The TSDs consisted of three nucleotides. The first nucleotide was typically T, the second was

usually T or A, and the third was preferentially T or A. Some 10 bp of A/T rich sequence flanked the TSDs. Four copies of the motif TTGCCA and one plant G-box factor binding site (TGACACGTGGG; Tfsitescan program expectation value 1.28e-03) were typically present within the internal sequence. An additional TE-like sequence flanked by TTTTTTGA and TCAAATTT was found nested inside the group A sequences (Fig. 2.2A). As shown in Fig. 2.2B for *Kiddo* in IR24 *rubq2*, all of the sequences in this group have the potential to form hairpin structures. This could result in the formation of cruciform structures in rice genomic DNA; since such structures are known to be capable of affecting transcriptional regulation (160), the presence of *Kiddo* MITES may affect expression of adjacent genes, such as the IR24 *rubq2* gene.

Excision of *Kiddo* from the T309 *rubq2* promoter region

T309 belongs to the *Japonica* subspecies of rice while IR24 is a member of the *Indica* subspecies (161). These subspecies are thought to have started to diverge some 2-3 million years ago (162). The question arises whether the *Kiddo* element inserted into IR24 after divergence of the subspecies or if it was present before divergence and left the locus in T309 due to transposition following divergence. In the case of a *de novo* insertion, one can postulate that the target sequence was TTA|GA prior to insertion of a TE and TTA*e*TTAGA after insertion (*e* represents the TE insert) (Fig. 2.3). In the case of a TE excision from this insertion, one would expect TTA|TTAGA after net excision or TTA|TAGA if a single nucleotide (T) was lost at the right border TSD (73). Inspection of the sequences in rice lines shows that the IR24 sequence TA*e*TTAGA is in agreement with retention of a TE insertion and that the T309 sequence TTA|TAGA is that expected after loss of the TE; i.e. the TE

(*Kiddo*) was present prior to divergence. An alternative postulation is that the target sequence was TTA|TAGA and that the T309 sequence reflects lack of any TE insertion. However, it is then difficult to see how insertion of a TE into this sequence could occur (as

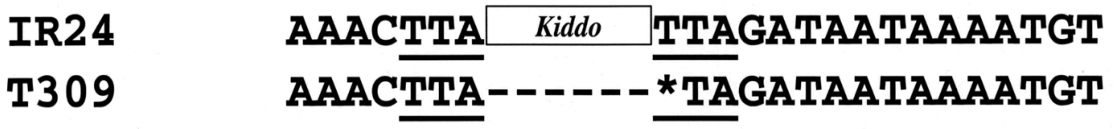

Figure 2.3. *Kiddo* is lost from the T309 *rubq2* promoter. The *Kiddo* in IR24 (rectangle) is flanked by TSDs (underlined). It is absent from this locus in T309 plants (dashed line) and a T residue is lost (*) from the right border TSD.

evidently did occur in IR24) since the expected sequence would be TTA*e*TTATAGA, which is not observed.

Existence of four groups of *Kiddo*-related elements in rice

After several iterations of database search, 36 *Kiddo*-related sequences were retrieved that included both TIRs. These were sorted (see *Materials and Methods*) into four groups (Fig. 2.4). They all had MITE characteristics, with internal sequences that are flanked by AT-rich micro-regions (data not shown). However, individual groups had slightly different consensus TIRs and TSDs. The derived phylogenetic tree (Fig. 2.4) indicates that the *Kiddo* family may have common ancestor. Presumably, during evolution, some sequence changes inactivated the ability to transpose while others had little or positive effects on transposition that permitted their propagation throughout rice genomes. Groups A, B, C and D have multiple copies of highly similar sequences, and are thus likely to be still active. Within each of the four groups, sequences are >90% identical. Similarity between groups ranges from 65 to 75%.

***Kiddo* is a rice-specific element**

In addition to the 35 complete *Kiddo* sequences, 24 incomplete sequences were retrieved with 75-93% similarity to queries (see *Materials and Methods*). That 58 out of 59 retrieved sequences came from rice suggested that the *Kiddo* MITE family was essentially restricted

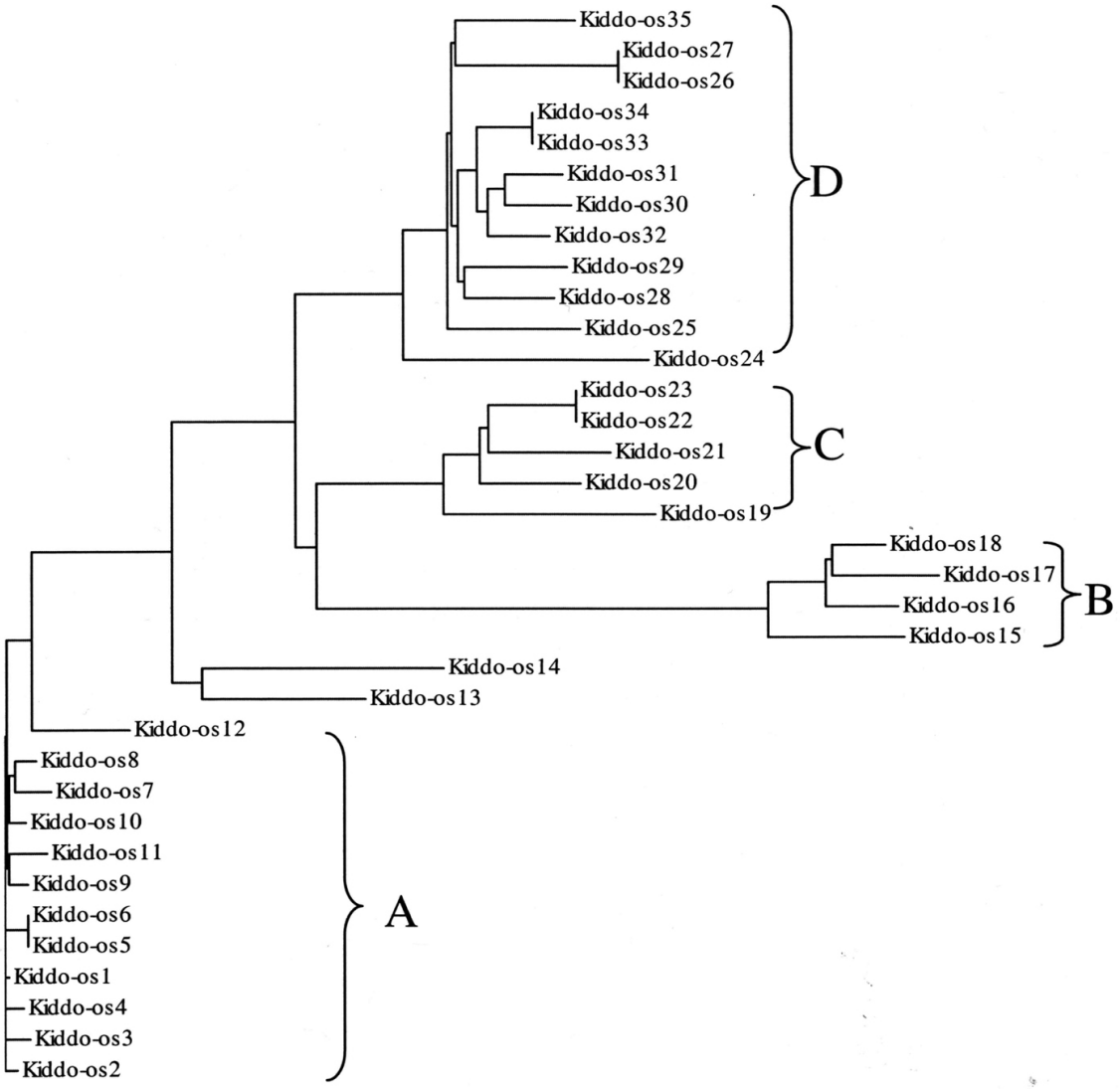

Figure 2.4. Phylogenetic tree of *Kiddo* sequences. The tree was generated using Vector NTI AlignX. Numbering for family members corresponds with that in Table 2.1. Groups A and D contain members of high (>90%) sequence similarity.

to the rice genome. One sequence from wheat (*Kiddo-ta1*) had an overall similarity (~72%) to group A *Kiddo* elements, with ~56% similarity within a region flanked by TTTTTTGA and TCAAATTT sequences and ~83% outside this region. That *Kiddo* is essentially confined to the rice genome was supported by genomic DNA blot analysis. When genomic DNAs from *Arabidopsis*, tobacco, maize, wheat, a tree (*C. acuminata*), rice IR24 and rice T309 were probed with the *Kiddo-os11* PCR product, only rice DNAs hybridized (Fig. 2.5A). This was true at both low and moderate stringency wash conditions.

The question arises as to why the *Kiddo* element has left the *rubq2* promoter in T309 but not IR24. It has been proposed that certain genomic environments provide more favorable

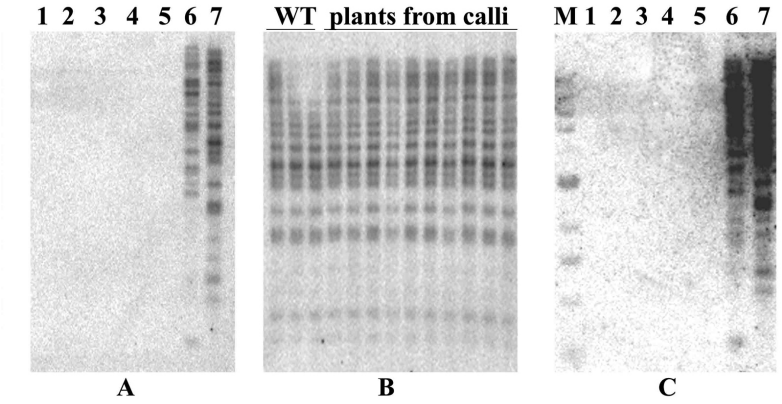

Figure 2.5. *Kiddo* is prevalent in the rice genome. Various genomic DNAs probed with the *Kiddo*-os11 sequence (**A**). Genomic DNAs from three T309 plants (WT) and ten T309 plants regenerated from independent calli, probed with the *Kiddo-os11* sequence (**B**). Various plant genomic DNAs probed with a combination of *Kiddo* group A, B, C and D sequences (**C**). Lanes: 1, *Camptotheca acuminata*; 2, tobacco; 3, *Arabidopsis*; 4, wheat; 5, maize; 6, rice IR24; 7, rice T309.

environments than others for transposition (163). Comparison of the hybridization profiles for T309 and IR24 probed with *Kiddo-os11* shows much greater complexity for T309 (Fig. 2.5A), possibly suggesting a more favorable climate for transposition exists in this line.

However, no transposition was detected when 10 plants were regenerated from independent T309 calli (Fig. 2.5B), suggesting that tissue culture intervention does not stimulate migration of *Kiddo*. A mixture of group A, B, C and D fragments were also used to probe the various genomic DNAs. That only rice genomic DNAs showed hybridization further supports the possibility that *Kiddo* is a rice-specific MITE (Fig. 2.5C). The intense banding profile seen with these combined probes reveal that the *Kiddo* family exists at a high copy number.

***Kiddo* elements are closely associated with genes**

Table 2.1 lists 39 members of the rice *Kiddo* family and one putative member from wheat. An additional 19 candidate members with incomplete TIRs were found in non-annotated sequences (data not shown). Of the 59 *Kiddo*-like sequences identified from public databases, 18 were in annotated sequences. The other 41 sequences were not annotated, and it is possible that some or all of these are present in genic regions. The distribution of the *Kiddo* elements in genomic DNA was examined with reference to coding sequences (CDS). The most distant, *Kiddo-os37* and *Kiddo-os25*, were 4168 bp and 2789 bp, respectively, from a CDS. *Kiddo-os32* and *Kiddo-os38* were at about 1 kb from a CDS. Six *Kiddo* elements resided <530 bp 5' or 3' of a CDS and seven were in intron regions. The related sequence from wheat was an EST sequence. The fact that 14 of the 18 (~80%) sequences were within 530 bp of CDSs (Table 2.1), or in introns, underscores *Kiddo*'s close association

Table 2.1. Genomic distribution of *Kiddo*

Kiddo	Accession	Genic	Position
os1	ap001550(58275-58007)	Y	523bp from CDS 3'
os2	ap001366(121108-121378)	Y	intron
os3†	ac084296(41225-40957)	?	?
os4†	ac079038(125587-125319)	?	?
os5†	ac079356(2685-2417)	?	?
os6†	ac079022(85225-84957)	?	?
os7	az134261(740-206)	?	?
os8†	ac027038(143796-144064)	?	?
os9	ap003047(63787-64055)	?	?
os10†	ac051632(168882-169150)	?	?
os11	af184280(1233-1501)	Y	210bp from TATA 5'
os12†	ac025906(138037-137770)	?	?
os13	aj245900(62716-62965)	Y	intron
os14	aq157268(196-389)	?	?
os15†	ac083943(13936-13662)	?	?
os16	ap003048(121712-121988)	?	?
os17	aq157046(280-4)	?	?
os18	ap002743(32729-33008)	Y	379bp from CDS 5'
os19†	ac084282(118802-119079)	?	?
os20	aq795953(109-386)	?	?
os21	aq689856(45-319)	?	?
os22†	ac027037(118552-118828)	?	?
os23†	ac018929(9993-10269)	?	?
os24	ap002816(24995-24724)	Y	intron
os25	ac082644(108393-108633)	N	2789bp from CDS 3'
os26	ap002864(2288-2559)	?	?
os27	ab023482(148318-148589)	Y	232bp from CDS 3'
os28	ap000836(148097-147854)	Y	intron
os29	al442114(44482-44203)	Y	intron
os30	aq273730(524-254)	?	?
os31†	ac079852(205709-205988)	?	?
os32	ac007858(28501-28777)	Y	1023bp from CDS5'
os33†	ac079029(125597-125876)	?	?
os34	ac051634(134585-134306)	Y	126bp from CDS 5'
os35	ab026295(26954-27202)	Y	367bp from CDS 5'
ta1	be517419(325-49)	Y	cDNA
incomplete members found in annotated sequences			
os36	af128457(66418-66286)	Y	intron
os37	ac069145(33249-32753)	N	4168bp from CDS 3'
os38	ap002522(113573-113331)	Y	1203bp from CDS 3'
os39	af119222(73595-73723)	Y	intron

os, *Oryza sativa*; ta, *Triticum aestivum*. Numbers in parenthesis indicate the element's position in the cited accession; where known, distances from the TATA box or CDS are noted; <2 kb from a CDS is considered genic (Y, yes); >2 kb is not genic (N, no); os11 is from IR24 *rubq2*; ?, unknown; genes have not yet been identified for the other annotated sequences; †, HTGS sequences, positions in the accessions are subject to change. with genes.

This property suggests that *Kiddo* family members are suitable new molecular markers for genic regions in rice.

DISCUSSION

Identification of TEs by database searching

Presently, classification of transposable element families and their members depends largely on sequence similarity, mostly by searching databases with a known sequence as the query sequence. The retrieved sequences are then used again to retrieve more sequences. This method is efficient, but also has its disadvantages: (1) There could be very low, possibly undetectable, sequence similarity between the start query sequences and the retrieved sequences after several rounds of search. This leads to reservations about classifying them into the same family. (2) Some sequences with same TIR may not show significant similarity in the internal sequence and it may be difficult to detect all sequences having the same TIR, even if the whole genome is sequenced. Because both DNA and retroelements are known for their repeated sequences at their ends, database search tools that are able to detect TIR or terminal direct repeated sequences in a reasonable length range will be helpful to identify transposable elements. This is especially advantageous considering the high percentage of mobile element DNA in higher eukaryotic genomes. Database searches designed to identify repeat elements may also be helpful in detecting other functional motifs in a genome.

Origin and amplification of *Kiddo*

MITEs occasionally share similar TIRs with autonomous DNA transposons and thus are thought to have derived from DNA transposons (132,164-166). However, MITEs usually

have a very high copy number (138,155,164,166), which can not be achieved by classical DNA transposition. An alternative origin of MITEs has been suggested that involves aberrant DNA replication events when DNA polymerases encounter palindromic sequences as templates. In this model (128), the 3' region of a nascent DNA strand may fold back, allowing DNA synthesis to reinitiate using the nascent DNA strand as template. A stem loop byproduct (the *Angel* MITE) may result from this aberrant replication. After its excision from genomic DNA, the stem loop can then integrate into other genomic DNA locations with the help of recombinases, providing new sites for amplification of the MITE.

The distribution of *Kiddo* members with respect to genic regions may provide insight to the amplification of the family. Of 18 *Kiddo* members in annotated sequences, 7 (40%) are in introns, 6 (30%) are <530 bp from a CDS and 4 (20%) lie between 1 and 5 kb from a CDS. Although this distribution shows clustering around genic regions, it is interesting that no MITEs are in coding sequences. This organization may indicate that the origin of the groups within the *Kiddo* family resides within introns. Such an arrangement could yield a very large number of *Kiddo* copies as a result of transcription and subsequent excision by splicing. Reverse-transcription of a proportion of the excised *Kiddo* RNA elements into DNA might confer transposition capability, permitting integration into nearby genomic DNA locations. Transposition into introns of other genes could lead to further propagation of *Kiddo*. This scenario implicates the possibility of RNA intermediates in addition to, or instead of, DNA intermediates in transposition of *Kiddo*.

Excision and transposition of *Kiddo-os11*

Since the widespread distribution of MITEs implies an ability to transpose, it is curious that movement has not been observed. If autonomous elements capable of supporting MITE transposition exist, it is likely that they are present on elements separate from the MITEs since MITEs (including *Kiddo*) do not have extensive open reading frames (144). It is conceivable that a transposase is provided *in trans* from other genes resident in the organism.

The excision of *Kiddo* from the rice T309 *rubq2* promoter appears to be a net excision, with modest damage at the first nucleotide (T) of the right side TSD. Degradation of the overhanging 5'-ends of the TSD has been observed for other DNA transposons (73). Previous evidence of excision footprints for *Hbr* (138), together with the present evidence for *Kiddo* (Fig. 2.3), suggests that MITEs can excise from genomic DNA. This supports the concept that MITEs are similar to DNA elements (146). However, caution is advisable in accepting footprints as evidence for MITE transposition as retrotransposons can also excise from genomic DNA at a very low frequency (88). Investigation of the ability of MITEs to transpose has been hindered by the fact that families with low sequence similarity may contain decayed MITE members that have lost their mobility. In contrast, highly homogeneous MITE families such as *Hbr* and *Kiddo* may well still be functional and may, therefore, be useful candidates for unraveling the mystery of MITE transposition.

Significance of the TIR sequence

TIRs are very important for transposons and are thought to serve as transposase binding sites (167), perhaps accounting for their high conservation in a given transposon family. The same TIR can be found in various MITE families from different organisms. The TIR sequence

GGGGNTGTTTGGTT present in *Tourist-D, Hbr* (138,145) is also found in *Kiddo* (Fig. 2.2). There are two explanations for the presence of this TIR in rice, maize and wheat. One is that MITEs containing this TIR are derived from a common ancestor, and that the internal sequences have undergone massive mutation so that they no longer bear recognizable similarity. This explanation assumes that the internal sequences are not functionally necessary. An alternative explanation is that sequences bearing this TIR are especially subject to aberrant replication events (128), leading to the *de novo* creation of MITE families bearing the same TIR, but with no detectable similarity in their internal sequences. In either explanation, the TIR is considered to be very important for MITE propagation. Studies on protein factor binding interactions are needed to reveal the role of this TIR in the transposition of *Tourist-D, Hbr* and *Kiddo*.

Involvement in the regulation of gene expression

Close association of MITEs with plant genes may indicate their involvement in the evolution of these genes (123,155). The close proximity of *Kiddo* members to CDSs (<530bp) suggests that the insertion of these elements could probably modify transcriptional, splicing or translational regulation of the genes. The IR24 *rubq2* promoter that contains the *Kiddo* insertion has been shown to drive high levels of reporter gene expression in transient assays (159), and it is possible that the G-box present in *Kiddo-os11* augments transcriptional activity. Therefore, it will be informative to compare the activity of *rubq2* promoters from IR24 and T309 in stably transformed rice plants to determine if the presence of the *Kiddo* insertion increases or decreases promoter strength.

CHAPTER III

IDENTIFICATION OF TWO *MUTATOR* DERIVED MITE FAMILIES*

INTRODUCTION

Transposable elements (TEs) in eukaryotes are generally classified into class I elements (retroelements) and class II (DNA transposons). Recently, a variety of miniature inverted repeat transposable element (MITE) families were identified from various organisms, including maize(138,144), rice (123,137,142), *Arabidopsis* (125,129), *Medicago* (127), *C. elegans* (54,156,168,169), mosquito (124,131,139,148), fish (128,170), *Xenopus laevis* (135), sea squirt (171) and human (172). These MITEs have structural features of DNA transposons such as target site duplications (TSDs) and terminal inverted repeats (TIRs). However, they usually have high copy numbers that are reminiscent of retrotransposons. Since MITEs are small in size (100-500 bp), they are thought to be unable to encode proteins for their amplification and are thus non-autonomous elements. Establishment of a relationship between a non-autonomous TE family and an autonomous superfamily may be achieved by demonstrating similarities in DNA sequence, structural organization, or putative transposase.

Progress is being made towards revealing the origins of several MITE families. For example, human *MER*s were found to have arisen from *Tigger* autonomous elements

---

*Reprinted with permission from "*MDM*-1 and *MDM*-2: Two *Mutator*-Derived MITE Families in Rice" by Guojun Yang and Timothy C. Hall, (2003), *J Mol Evol*,, 56(3), 255-264.

(172,173). *Emigrant* (125) and *MathE2* (129) in *Arabidopsis*, as well as *Mimo* (131), *Nemo*, and *Wujin* (124) in mosquito, have been suggested to have arisen from *Pogo* elements (132). *Tigger* and *Pogo* belong to the *Tc1/Mariner* superfamily. The putative transposase for *Tourist*-like MITE families (MITE X and MITE XI) was shown to share similarity with bacterial insertion sequence transposases such as IS5S, IS493 and IS903 (174). A relationship between rice *Stowaway* elements and *Tc1/Mariner* elements has also been proposed (175). Very recently, *Tc8* and *mPIF* were classified into the *IS5/Harbinger PIF* superfamily (53-55). Thus, these MITE families appear to be derived from class II transposons. It is intriguing, however, that no MITE family identified to date has been shown to have arisen from a well-characterized plant transposon, e.g. *Ac/Ds*, *En/Spm*, or *Mutator* (176).

In this report, we describe the identification, sequence analysis, and classification of two novel MITE families in the rice genome. These MITE families share structural and terminal sequence similarities to maize *Mutator* elements and thus appear to be descended from a *Mutator*-related ancestor. Insertion site duplications (ISDs) were found to be prevalent at gaps in the alignments of the two MITE families. As will be described, one of these ISD sites provided evidence for the transposition mechanism of *Kiddo*, a MITE element we previously described in rice (142).

MATERIALS AND METHODS

Transposable element nesting analysis

We used transposable element nesting analysis (TENA) as a genome wide TE identification method based on the TE nesting occurrences. It includes searching for nesting events with a starter TE family, identification of nesting events, identification and characterization of the new TEs and their use in subsequent rounds of analysis.

Each element of the MITE family *Kiddo* (142) was used as a starter TE to search for potential nesting events. To discover new elements lying outside this TE, folding studies (using M-fold at http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi) were conducted using the *Kiddo* element plus 2 kb of flanking sequence (1 kb 5' and 1 kb 3'). Additionally, BLAST searches were conducted using *Kiddo* members together with their flanking sequences and with the flanking sequences alone. To search for potential TEs nested within *Kiddo* members, gap sequences of >50 nucleotides in *Kiddo* alignments were used for BLAST searches and folding studies. BLAST search results were screened for new TEs, especially MITEs, based on their structural characteristics. Potential stem-loop or hairpin structures of $\geq$100 nucleotides were checked for putative TSDs since the presence of a putative TSD is indicative of a TE. Hairpin structures bearing a putative TSD were used for BLAST searches to determine the abundance of the newly identified TE. Members of these new MITE families were aligned using VNTI 6.0 AlignX (InforMax, Inc., Bethesda, MD).

Database search and alignment method

Public databases (October 22nd, 2001 update) at NCBI were used to search for rice (*Oryza sativa* cv. Nipponbare) repeat sequences using BLASTN and TBLASTX. After elimination

of duplicate sequences, they were aligned with AlignX in the VNTI 6.0 package. The default values (gap open penalty: 15, gap extension penalty: 6.66) of the program were used for alignment. A manual editing of the alignment was carried out to correct evident misalignments. Alignments were exported as MSF files and then visualized with Boxshade (http://www.ch.embnet.org/software/BOX_form.html).

RESULTS

Discovery of MITE family *MDM*-1

When a DNA sequence fragment (AF184280) containing *Kiddo*-os11 was folded with M-fold, in addition to the *Kiddo*-os11 secondary structure, another significant secondary structure was identified (inset in Fig. 3.1. 10) that contained a putative TSD sequence of TAAAAAAAA. A family of 17 additional structurally complete homologous sequences, ranging in size from 339 bp to 392 bp, was retrieved from the databases and aligned using VNTI 6.0 AlignX (Fig. 3.1). This family has a strong potential to form hairpin structures (inset in Fig. 3.1) and possesses putative TSDs of 9 bp that are almost exclusively composed of A/T nucleotides. These characteristics denote these elements as MITEs and the new family is designated *MDM*-1. No substantial DNA sequence similarity to other TEs was found in the internal regions. In fact, compared to some MITE families (such as *Hbr*, *Kiddo*, *mPIF* and *Micron*), the internal regions in this family were substantially decayed although they still retain extensive sequence similarity within the family. A total of 42 *MDM*-1 elements (including those in the alignment presented in Fig. 3.1) were retrieved; since only about 12% of the total rice genome sequence was present in the database at the time this

article was written (see Materials and Methods), we estimate the actual copy number in the rice genome to be about 400. Given the non-random distribution of the sequence data in GenBank and the presently unknown distribution of *MDM*-1 elements in rice genome, the final copy number may be much higher and can be determined once the entire sequence of the rice genome is available. For presently identified, complete (those containing TIRs on both ends) members of the *MDM*-1 family, the size ranged from 303 bp to 397 bp, with an average length of 358 bp. The A/T content ranged from 58.04% to 64.85%, with an average A/T content of 62.74%. Eleven of the 42 *MDM*-1 elements were from annotated sequences; of these, five (45%) are in putative introns and all eleven are within 2 kb of a putatively transcribed region. Based on this sample, it is tempting to speculate that many *MDM*-1 elements may have a close association with rice genes, especially with introns.

Discovery of MITE family *MDM*-2

A decayed copy of *Kiddo* was identified in a rice BAC clone (accession no. AC069145) at positions 32958 to 33248. When the sequence flanking this *Kiddo* was folded, the DNA between 33927 and 34239 fell into a strong hairpin structure with a putative TSD of TTAATTTAA (inset in Fig. 3.2). BLAST searches retrieved 21 additional complete homologous sequences from GenBank databases (Fig. 3.2). Except for the element in AC083943, which had a 7 bp putative TSD, all copies have putative TSDs of 9 bp that are almost exclusively composed of A/T nucleotides. All the members in this family can potentially form a strong hairpin structure similar to that of the founding member in AC069145 (Fig. 3.2). Internal sequences shared no significant sequence similarity to any other reported TEs, nor did they share internal sequence similarity to that of the *MDM*-1

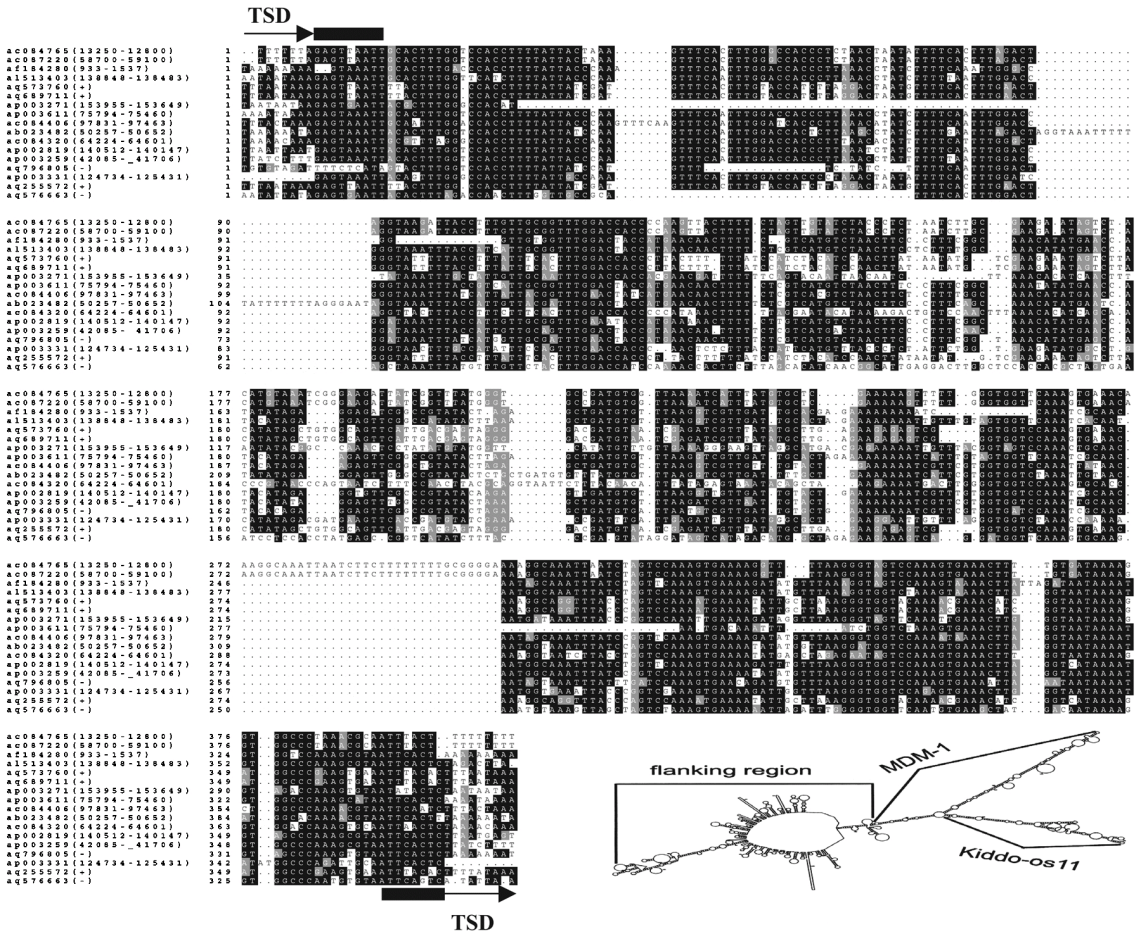

Figure 3.1. Sequence alignment of *MDM-1* (see Appendix C). Seventeen complete *MDM*-1 sequences of similar length were aligned. Putative TSDs (Arrows) of 9 bp are shown at the 5' and 3' ends of the alignment. The filled bars denote 5' and 3' ends similar to *Mutator* TIR ends. The *Kiddo-os11* element within AF184280 was removed for the purpose of alignment. An additional complete *MDM-1* sequence (AC079888 from 94077 to 947888) was not aligned because of its long sequence length. Inset: A nesting event involving *Kiddo-os11* and a *MDM-1* member as illustrated by DNA molecule folding is shown. The sequence from GenBank accession AF184280 is shown from positions 1 to 1747. According to the M-fold prediction, *Kiddo-os11* can be seen as a branch within the secondary structure of *MDM-1*, both elements lying within the *rubq2* promoter.

family. This MITE family, designated as *MDM*-2, showed mainly single nucleotide substitution mutations (see below). A total of 59 *MDM*-2 elements (including those presented in the alignment) were retrieved from genome data available at the time of writing,

and the estimated total copy number in the rice genome is about 600. As for *MDM*-1, the distribution of *MDM*-2 sequences is not known and a final estimate of the copy number will await completion of the rice genome. For presently known, complete, *MDM*-2 elements, the size ranged from 216 bp to 317 bp with an average size of 300 bp. The A/T content ranged from 56.59% to 65.63%, with an average A/T content of 59.95%. Seventeen of the 182 retrieved elements were from annotated sequences; six of these 17 *MDM*-2 elements (35%) are in putative introns and another six are within 1 kb of a putative coding region. Based on

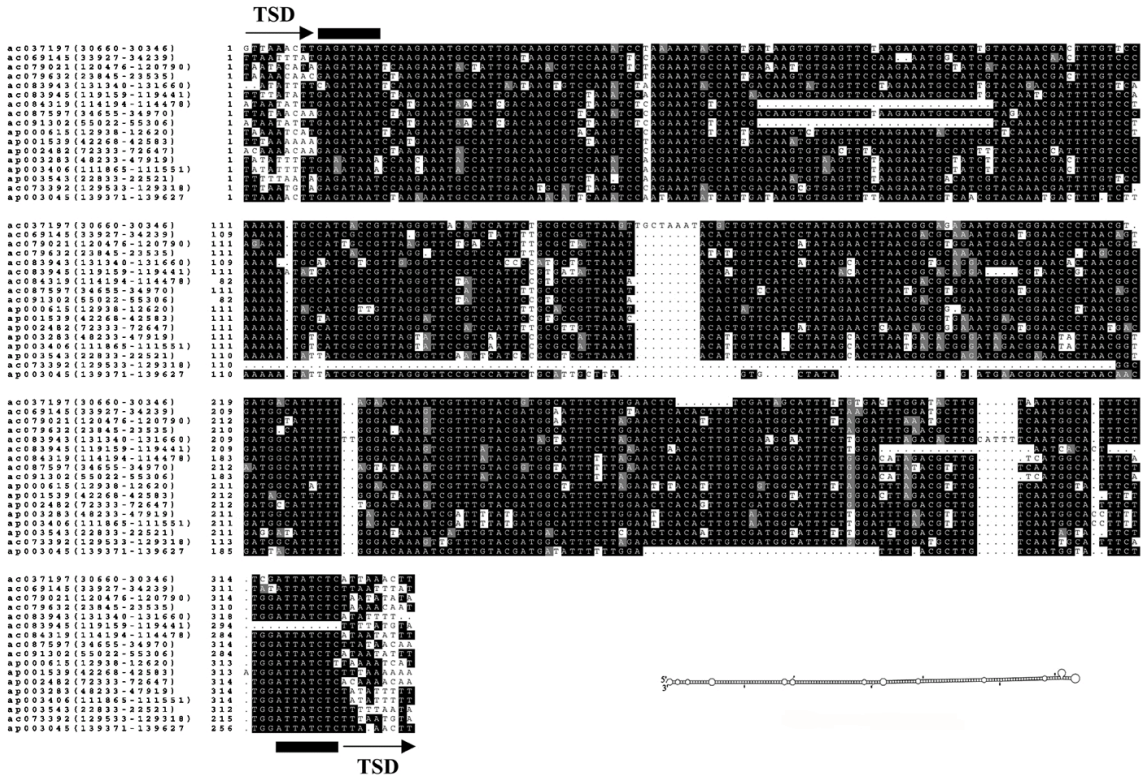

Figure 3.2. Sequence alignment of *MDM-2* (see Appendix C). Seventeen *MDM-2* sequences of similar length were aligned. Five additional complete *MDM-2* members bearing undetermined nucleotides (AP003529 from 63448 to 63046; AP003447 from 23193 to 22831; AP003445 from 19469 to 20240; AP003206 from 103554 to 104654) or long insertions (AP003412 from 40820 to 42117) were not aligned. Putative TSDs of 9 bp are shown at the 5' and 3' ends of the alignment. 5' and 3' ends similar to *Mutator* TIR ends are shown as filled bars. Inset: Hairpin structure of the *MDM-2* family. The M-fold prediction shown for the *MDM-2* element in GenBank accession AC069145 is typical of that for all members of this family.

these, admittedly small, numbers it is tempting to speculate that *MDM*-2, like *MDM*-1 may have a close association with rice genes.

Mutator-derived MITEs (*MDM*s)

MDM-1 and *MDM*-2 showed typical features of MITEs such as TIR, short length, extensive secondary structure and high copy number. However, whereas MITEs were originally considered to have putative TSDs of only two or three nucleotides (175), *MDM*-1 and *MDM*-2 have putative TSDs of 9 bp. Nevertheless, since TSD lengths up to 9 nt have been reported for other MITEs (127,139), the *MDM*s were originally classified as two new MITE families. Interestingly, the *MDM* families display features characteristic of *Mutator* elements (60,177,178) such as 9 bp TSDs and a long TIR region that could potentially form a hairpin structure (Figs. 3.1 and 3.2). In addition, putative MURA transposase binding sites of ATATGACAATATAGAGGAGTT and GCGGAATGGACGGAA were found in the middle of *MDM*1 and *MDM*-2 consensus sequences, with 75% in 21 bp and 80% in 15 bp identical to the MURA binding site (179), respectively. Importantly, 8 bp at the 5' and 3' ends of *MDM*-2 are identical to those of the maize transposon *MuDR* and 9 bp of *MDM*-1 termini are identical to that of the rice *MULE*-9 (Fig. 3.3). Given that transposon terminal sequences are extremely important for transposition, the conservation of the terminal motifs in *MDM*-1 and *MDM*-2 is striking and in strong support of the concept that these *MDM*s are derived from the same ancestor as are maize Mutator transposable elements.

Further studies on *MDM*-2 provided evidence that the *MDM*-2 family is indeed derived from an ancient *Mutator* transposon (Fig. 3.4A). An iterated database search, using the *MDM*-2 consensus sequence as a query, yielded *MDM*-2L (for *MDM*-2 long) on

chromosome one. This element (accession: AP004320 from 3568 to 9016) has a length of 5449 bp. It bears TIRs that are 82% identical to those of the 112 bp terminal consensus in *MDM*-2. *MDM*-2L shares 90% overall internal sequence identity with two high throughput genomic sequences (accessions AP004025 and AP004071) that are on chromosome two; all three elements have different putative TSDs and flanking sequences (Fig. 3.4A). When these three long elements were used for a BLASTX search of the GenBank database, putative MURA protein sequences were retrieved. An internal region of 2598 bp was found to

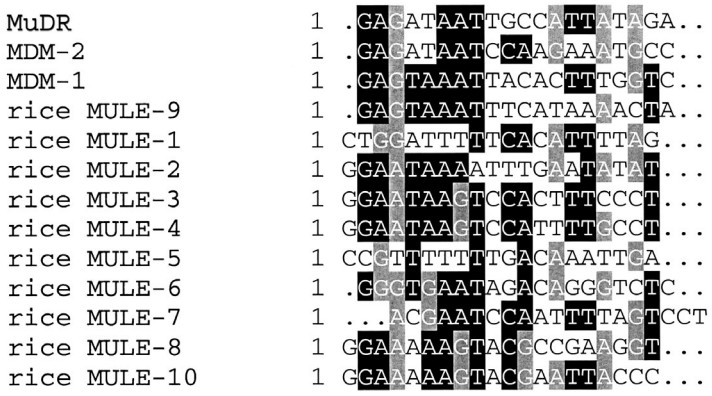

Figure 3.3. Alignment of 5' ends of *MuDR*, *MDM*s, and rice *MULE*s. The 5' terminal 20 nucleotides from maize *MuDR*, *MDM*-1, *MDM*-2, and 10 reported rice *MULE*s were aligned. *MDM*-2 showed an identical 5' terminus of eight nucleotides to that of *MuDR*. Except for rice *MULE*-1, *MULE*-5 and *MULE*-7, all elements share significant similarity in their eight terminal nucleotides. Rice *MULE*-9 is placed beneath *MDM*-1 as they have identical 5' termini.

correspond to the 722 amino acid residues of a putative maize MURA protein (accession: AAK63886), with overall 36% identity and 53% similarity. A highly conserved region of 253 amino acid residues was aligned with the corresponding region of the maize Mutator transposase MURA protein (accession: AAK13094; 44% identity) (Fig. 3.4B). In addition, a MURA binding site (179) core sequence of TTCGACGAAA was also found on the long

elements with 90% identity in 10 bp (e.g. from 2341 to 2350 on *MDM*-2L). These findings indicate that *MDM*-2 elements are indeed derived from an ancient Mutator transposon (Fig. 3.4A). We have not been able to identify a putative transposase for *MDM*-1 from the currently available databases.

Although 17 reported rice MULEs were classified into 10 families based on structural similarity to Mutator elements, none of the 10 families was a typical MITE family (175). Thus, *MDM*s are the first MITE families identified that appear to have descended from a

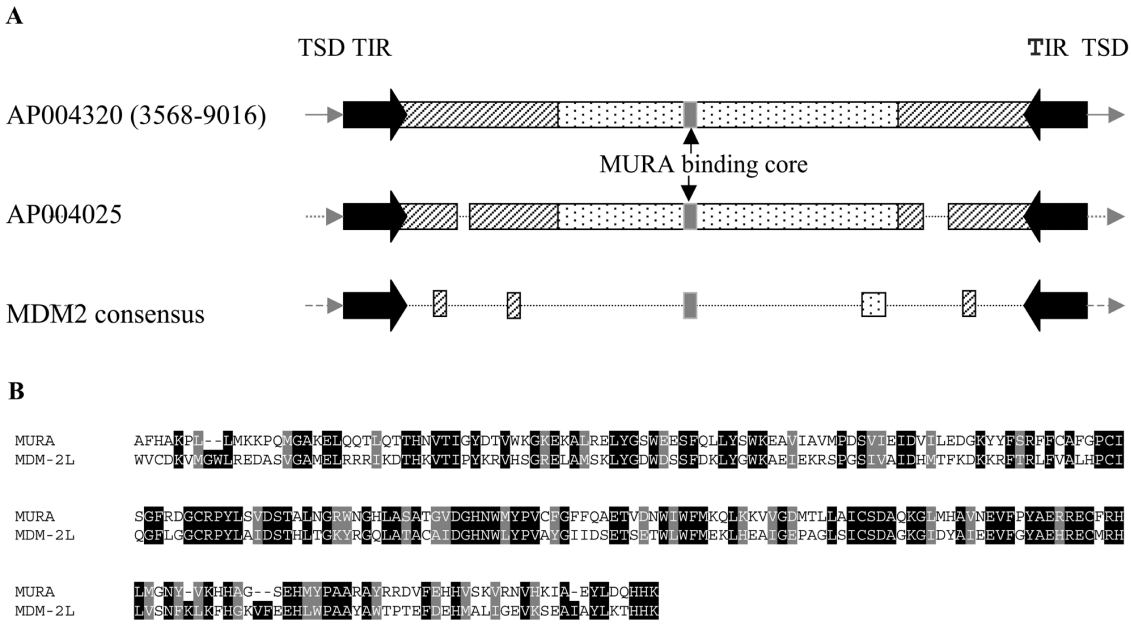

Figure 3.4. *MDM*-2 elements appear to be derived from *Mutator* transposons (see Appendix C). (**A**) Stippled bars indicate regions corresponding to *MURA* homologs. The gray regions indicate putative *MURA* binding core sequences. The different thin arrows denoting the putative TSD regions represent different flanking sequences. Dotted lines denote deletions. The positions on AP004025 are subject to change with database updates and the *MDM*-2 on AP00407 (see text) is not shown because it has identical flanking sequences to those of AP004025, an indication of a non-transposition event. The regions are not accurately proportional to the actual sequences. (**B**) Protein sequence alignment of the highly conserved region of the putative *MDM*-2 transposase and maize *MuDR* transposase *MURA* (positions 220 to 446). Regions in black represent identity and regions in gray indicate conservative changes.

maize Mutator-related origin. In contrast to the low copy number and high sequence length heterogeneity reported for MULEs (175,178), *MDM* families have high copy number and low sequence length heterogeneity.

Evidence for historical nesting events in *MDM*s

Alignments of *MDM*-1 and *MDM*-2 sequences revealed significant sequence decay (Figs. 10 and 11). Point mutations, deletion mutations and insertion mutations in these two families were studied in detail (Table 3.1). The point mutations included A/G transitions,

Table 3.1. Mutations in *MDM* alignments

Mutation		*MDM*-1		*MDM*-2	
		number	percentage	number	percentage
Point mutation	A/G	204	50.1	152	65.2
	A/T	95	23.3	25	10.7
	A/C	69	17	28	12
	Deletion & insertion	17	4.2	14	6
Deletion (≥2bp)		15	3.7	11	4.7
Insertion (≥2bp)		7	1.7	3	1.3
Sequence decay degree		high		low	

Numbers of mutations are summarized for *MDM*-1 and *MDM*-2 alignments (Figs. 3.1 and 3.2). The percentage values represent the proportion of each type of mutation in the total population. Sequence decay degrees are inversely related to the A/G transition percentage values because A/G transitions occur most frequently. The higher the percentage of A/G transition, the lower the sequence decay degree. Deletions and insertions were distinguished (1) by comparing the individual sites with the corresponding sites on the consensus sequence and (2) by studying the individual sites in respect to the alignment context in Figs. 3.1 and 3.2.

A/T transversions, A/C transversions and single nucleotide additions or deletions. Most point mutations were A/G transitions, especially in early stages of sequence decay, as shown in *MDM*-2. There are 10 CG or CNG sites in the *MDM*-1 consensus sequence; 10% of the

transitions occur at these sites and each site has at least one transition. There are 20 CG or CNG sites in the *MDM*-2 consensus sequence; 26% of the transitions occur at these sites and each site has at least one transition. The ratio of transition to transversion in many species has been shown to be approximately 1:1 (47). The unusually high transition to transversion

Table 3.2. Insertion mutations in *MDM*-1 and *MDM*-2 alignments

Accession	Insertion	Putative historical TE	ISD
MDM-1 family			
ac084406	TACCCAA**GTTTCAA**GTTTCAATTTGG	*Ac* or *Mutator*	7
ab023482	GACT**AGGTAAATTTTTTATTTTTTTAGG**	*Mutator*, *Cin3*,*Ac/Ds*	10
	GAATAGGTAAATTTACC		
ab023482	TCTGCTGATGTGTTTATGTGTTTAAG	*Ac/Ds*	8
ac084320	CGCAGGTAATTC**TTTACAACA**TTTATAG	*Tam1*	3
ac084765	AC**AAAGGCAAATTAATCTTCTTTTTTTTGCGGGG**	*Cin*4, noname *SINE*-	16
ac087220	**A**AAAGGCAAATTAATCTA	like	
af184280	AAACTTA**TTA**GATAAT	*Kiddo-os*11	3
(▵*Kiddo*-os11)			
ap003331	TAAAATAT**AT**GGCCCAG	?	2
MDM-2 family			
ac037197	GCCGTTAAGT**TGCTAAAT**ACGCTGT	retroelement, *Tat*1	5
ac083943	CATTTTT**TT**GGGACAAAA	?	2
ac083943	ACTTG**CATTTT**CAAT	*En/Spm* superfamily	3

Regions containing apparent insertion mutations in *MDM* alignments (see Figs. 3.1 and 3.2) are shown. Bold letters represent insertions in the alignments and underlined regions indicate ISDs. In the insertion mutation in *MDM*-2 of AC037197, the 'G' in the insertion site sequence TAAGT was originally 'A' because all the other members have 'A' at this position. The duplication event occurred before this A/G substitution. In the insertion mutation in *MDM*-2 of AC083943, an imperfect duplication of TTG was found, possibly resulting from a base substitution after duplication.

ratio (3:1) in *MDM*-2 may partially result from mutations at CG or CNG sites and may indicate a role for deoxycytidine 5-methylation in DNA sequence evolution. Deletion

mutations rarely occurred between G/C and G/C. Strikingly, 100% of the inserted fragments (≥ 2 bp) contain an ISD (Table 3.2). The length of this duplication varies from 2 to 16 bp and longer insertions contain longer stretches of ISD sequences. The frequent occurrence of ISDs in the EMBL DNA sequence alignment database (http://www.ebi.ac.uk/embl/Submission/alignment.html) indicates that duplication is an important aspect of insertion mutation during DNA sequence evolution (data not shown). This phenomenon is possibly caused by a DNA break repair event or by a historical transposition event with complete or incomplete excision.

Some of the insertion fragments showed significant similarity to the ends of known transposable elements. For example, the insertion of six nucleotides (ACAACA) on the *MDM*-1 member in accession AC084320 created a putative TSD of three nucleotides and the insertion itself is identical to the TIR core of Tam1, a member of the En/Spm super-family (180,181). At least one such insertion mutation in the *MDM*-1 alignment (Fig. 3.1) evidently resulted from a historical transposition event of *Kiddo*-os11 (Fig. 3.5). *Kiddo*-os11 is present inside the *MDM*-1 member in the rubq2 promoter of rice line IR24 (accession no. AF184280), but is absent at the same locus of rice line T309. The insertion of *Kiddo*-os11 into *MDM*-1 on IR24 rubq2 promoter resulted in a TSD of TTA. However, the absence of *Kiddo*-os11 from the rubq2 promoter region of rice line T309 indicates a historical excision event that left behind a footprint of two nucleotides (TA), probably due to a non-net cleavage. These observations suggest that *Kiddo* uses a 'cut-and-paste' strategy to transpose.

DISCUSSION

Internal sequence variation and insertion specificity of Mutator elements

```
  ac084765(13250-12800)  (348) GTCCAAAGTGAAACTTT----------------TGTGATAAAAGGT
  ac087220(58700-59100)  (348) GTCCAAAGTGAAACTTT----------------TGTGATAAAAGGT
al513403(138848-138483)  (324) GTCTAAAGTAAAACTTA----------------GGTAATAAAATGT
           aq573760(+)   (321) GTACAAAACGAAACATC----------------GGTAATAAAAGAT
           aq689711(+)   (321) GTACAAAACGAAACATC----------------GGTAATAAAAGAT
ap003271(153955-153649)  (262) GTTCAAATTGAAACTTG----------------GGTAATAAAAGGT
  ap003611(75794-75460)  (294) GTCTAAAGTGAAACTTA----------------GGTAATAAAATGT
  ac084406(97831-97463)  (326) GTCCAAAATAAAACTTA----------------GGTAATAAAATCT
  ab023482(50257-50652)  (356) GTCCAAAGTGAAACTTA----------------GGTAATAAAATAT
  ac084320(64224-64601)  (335) GTCCAAAGTGAAACTTT----------------GGTAATAAAAGGT
ap002819(140512-140147)  (321) GTCCAAAGTGAAACTTA----------------GGTCATAAAATGT
 ap003259(42085- 41706)  (320) GTCCAAAGTGAAACTTA----------------GGTAATAAAATGT
           aq796805(-)   (303) GTCCAAAGTGAAAATTA----------------AATAATAAAATGT
ap003331(124734-125431)  (314) GTCCAGAGTGAAACTTG----------------GGTAATAAAATAT
           aq255572(+)   (321) GTACAAAACGAAACATC----------------GGTCATAAAAGAT
           aq576663(-)   (297) GTTCAATGTGAAGCTAT----------------GACAATAAAAGGT
  af184280: rubq2(IR24)  (293) GTCCAAAGTAAAAC TTA -Kiddo(os11)- TTA GATAATAAAATGT
           rubq2(T309)         GTCCAAAGTAAAAC TTA ------------- TA GATAATAAAATGT
```

Figure 3.5. Evidence that *Kiddo* uses a 'cut-and-paste' strategy. A portion of *MDM*-1 alignment (see Fig. 3.1) is presented to demonstrate the region in which *Kiddo-os11* inserted in accession AF184280. A polymorphic locus from rice line T309 is also aligned to indicate the footprint of *Kiddo-os11* transposition. Accession numbers are shown on the left of the alignment and the numbers in parentheses indicate the sequence starting positions from the alignment in Fig. 3.1.

Mutator elements, initially discovered in Maize, are classified into six subfamilies (182). Internal sequence similarity exists among members belonging to a single subfamily, but is not detectable between members from different subfamilies (183). Among the nine MULE groups identified in arabidopsis, six contain members bearing long TIRs while the other three are non-TIR-MULEs (178). Ten groups of MULEs were identified in rice; all of them are TIR-MULEs (175). Within each TIR-MULE group, elements share similarity only at the TIRs. TIRs from different groups share no significant similarity except at the terminal 6-10 nucleotides. The internal sequences are highly variable. In the case of both *MDM*-1 and *MDM*-2, TIRs and TSDs are the only recognizable structures and thus represent extreme

forms of truncated Mutator elements. The differences in TIRs between *MDM*-1 and *MDM*-2 may indicate descent from different ancestors. Based on these extreme forms of Mutator elements, we speculate that a long TIR and the conserved termini of 8-10 nucleotides are sufficient to mediate the mobility of Mutator elements. This may explain the exceptional characteristics of internal variability for the Mutator system (183). Internal variability may be produced at the same frequency for all transposable elements, including Ac/Ds, En/Spm and Mutator. If mobility is partly dependent on internal sequence for Ac/Ds and En/Spm but not for Mutator elements, then any internally variable Mutator element may be able to propagate through duplicative transposition and thus yield the highly variable elements we see today, while movement of the other elements may be more constrained.

Interestingly, there is a consensus target site of T-T/A-T-A-T/A-T-A-A-A for *MDM*-1 and T-T-A-A-A-T/A/C-T/A/C-T-T for *MDM*-2. Compared to the consensus target site for Maize Mutator elements (G-T-T-G-G/C-A-G-G/A-G) (183), *MDM* target sites are extremely A/T rich, with 96% and 93% A/T content for *MDM*-1 and *MDM*-2 TSDs, respectively. As suggested by Benetzen (1996), different subfamilies may target different DNA sequence regions. According to the consensus target sites of *MDM*s, these elements are not likely to target protein coding sequences because sequences containing the consensus target site are highly likely to contain a stop codon (TAA, or TTA on the bottom strand). In fact, none of the *MDM* elements we have identified to date falls in a protein coding region although several are within introns or adjacent to coding regions.

The potential use of TENA for TE mining

Nesting events have been reported for retrotransposons (184,185), transposons (186), and MITEs (145,187). These findings suggest that the occurrence of nested TEs is relatively common. Since MITEs have been shown to be the most abundant type of TE in the rice genome (154,175, a high frequency of nesting can be expected. In fact, MITEs have been shown to nest frequently with other TEs, especially with other MITEs [Jiang, 2001 #1090). Traditionally, identification of MITEs in a genome was based on the observation of sequence polymorphisms. For example, polymorphism observed in the maize *wxB*2 gene led to the discovery of the Tourist MITE family (144). While the discovery of MITEs based on experimental evidence has the advantage of being direct (i.e. in providing putative evidence for transposition), it is of low efficiency and is inadequate for genome-wide MITE annotation. Recently, several indirect computer-based methods have proven to be powerful for the identification of new MITE families (123). For example, a redundance-based method using intergenic regions on 17.2 Mb of the Arabidopsis genome (174), or 4 kb windows in 910 kb of rice genomic sequence (175), as queries to search for repeat sequences was employed for the identification of TEs. Tu (2001) reported the discovery of eight novel MITE families in Anopheles gambiae using 'FINDMITE'. This program makes use of structural characteristics of known MITEs, such as TSD, TIR, short length, and the potential to form secondary structures, to determine MITE candidates.

While it is likely that many more MITE families in rice will be identified with these methods, we demonstrate here the value of TE nesting analysis (TENA) as an alternative approach for MITE discovery. In this approach, a known MITE is used as the starting point

to search for MITE nesting events using BLAST and M-FOLD programs. Only regions containing MITEs are interrogated and novel MITEs that do not fit narrowly defined input characters (e.g. TIR and TSD length) can be discovered. Moreover, polymorphisms are readily apparent from the TENA approach.

MITE and TE evolution

Retrotransposons are usually thought to have very high copy numbers while DNA transposons are present in relatively low copy numbers. The popular explanation for this difference is that retrotransposons use a 'copy-and-paste' strategy instead of the 'cut-and-paste' method used by DNA transposons. However, compensatory strategies are employed by DNA transposons to overcome the disadvantage of the 'cut-and-paste' method. For example, transposon copy number can be doubled by transposing from behind to in front of a replication fork (88). Additionally, the number of copies of a transposable element in a genome is dependent not only on its ability to transpose, but also on its ability to escape from strict host surveillance processes such as gene silencing (188-190). Indeed, the high propagation rate of retrotransposons may induce earlier constraint by the host genome than is the case for DNA transposons (191). When this, and the likely positive selection for transposable elements in a genome (4), is taken into account, the generation of an equivalent number of retrotransposons and DNA transposons should be feasible.

With the discovery of multiple MITE families in a genome from one known transposon family such as Mutator (see Results) and *PIF* (55), it is therefore possible that a presently unidentified strategy exists for proliferation of DNA transposons. A component of this strategy may be reduction in size or diversification of internal sequences of the TEs

so that they are less disruptive to host functions and less subject to silencing. In the case of MITEs, they appear to have shortened their element lengths such that, while they remain transposition-competent, they can escape from genome surveillance systems because of their small size and low similarity to existing sequences in the genome. Thus, a newly formed MITE may amplify rapidly until its copy number reaches a threshold level and is recognized by genomic surveillance processes. As a result, very high copy numbers of DNA transposon descendants can be generated, a number that may exceed that for retrotransposons.

CHAPTER IV

AUTOMATED MITE ANALYSIS

INTRODUCTION

Higher eukaryotic genomes are rich in transposable elements. Two distinct types of transposable elements have been identified in higher eukaryotes: Type I elements (retrotransposable elements) use a copy-paste approach to transpose, yielding a large copy number; type II elements (DNA elements) use a cut-paste-repair approach to transpose. However, numerous families of highly repetitive (hundreds or thousands), short (100-500 bp), elements that do not seem to belong to either type of element have been reported in plants and animals over the past decade (54,55,123-131,137-139,142-144,146,149,156,169,175,187,192-195). Because these families typically bear terminal inverted repeats (TIRs) and have target site duplications (TSDs) in their flanking sequences, they were given a collective name of miniature inverted repeat transposable elements (MITEs). Since MITEs apparently do not encode proteins, perhaps because of their small size, their amplification requires the involvement of factors supplied *in trans*. Although several MITE families are thought to be related to ancestral elements that bear similar TIRs and subterminal regions and are (or were) capable of coding for transposase-like proteins (54,55,132,143), the majority of MITE families lack such links. Since different MITE families may be derived from different founder elements, a link to the ancestral element needs to be established for each individual MITE family.

The analysis of a MITE family usually involves retrieving and aligning members in a given family, searching for its origin (or putative ancestor element) and studying its association with genes in a genome. These analysis steps are laborious, especially when multiple MITE families are involved in the analysis. To retrieve members of a family from the databases or to check the association of members with genes in a genome, a BLASTN is usually carried out, each high scoring pair (HSP) of the BLAST results is manually checked, and the desired sequence or positional information is then extracted from various accessions. This process is time-consuming and error prone because: (1) the copy number for MITE families is usually large; (2) for purposes of alignment it is necessary to reverse the sequences of those hits that are on the complementary strand of the sequence, and (3) for unfinished genomes, cited positions of elements in high throughput sequences are subject to change until the genome is completely sequenced. Even for the announced genomes, updates are released frequently and, hence, the copy number, positions and annotation is subject to change. To search for the putative anchor element (that retains both TIRs and coding regions reminiscent of a transposase) for a MITE family *in silica*, a BLASTN is carried out and long elements containing similarity to both ends of the MITE are checked and are then used to do BLASTX. BLASTX is used to screen for similarity to known transposases. This process is also time-consuming because, in addition to the difficulties mentioned above, complications arise from the facts that: (1) the anchor element usually does not share internal sequence similarity to the query MITE element, thus the identification of long elements requires manual inspection and recording of the short BLAST

HSPs; (2) BLASTX searches usually take longer than BLASTN searches, and (3) long sequences dramatically delay results from BLASTX.

Here, we describe MITE analysis kit (MAK), a collection of programs designed to automate MITE analysis (http://perl.idmb.tamu.edu/mak.htm). Given the sequence of a MITE element, MAK can retrieve and orient sequences of other members of the family, identify genes closest to the MITE elements, and can predict the anchor element for the MITE family. Using MAK, we have identified two novel TE families named Math and Kid and provided evidence that they belong to the recently identified (53,55) IS5/Harbinger/PIF superfamily.

MATERIALS AND METHODS

Programming language and modules

Practical extraction and report language (Perl) (196) was used to write the programs for MAK. Transformation of sequence formats was carried out with Bioperl modules Bio::Seq and Bio::SeqIO. The module Bio::Tools::Run::RemoteBLAST was used to do remote BLAST searches and the modules Bio::Search and Bio::SearchIO were used to parse the BLAST search results. Bio::DB::GenBank was used to retrieve MITE elements and their flanking sequences. The Bio::SeqFeature module (197) was used to identify genes closest to the MITEs. Common gateway interface (CGI) programming (198) was used to set up the MAK web-based query service (http://perl.idmb.tamu.edu/mak.htm).

Computing resources

Database searches were executed in the queuing system for BLAST (QBLAST) (151) at NCBI using a Uniform Resource Locator (URL) standardized application program interface (API) (http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html#blastq). MAK was tested extensively with a UNIX system on a 48-processor SGI Origin 3800 (k2) supercomputer at the Texas A&M University supercomputing facility (http://sc.tamu.edu). It was also tested using either Linux or Win32 systems on a PC with 2 GB RAM at the Texas A&M University Institute of Developmental and Molecular Biology (IDMB). Manual BLAST was carried out at the NCBI BLAST website (http://www.ncbi.nlm.nih.gov/BLAST/) to confirm the results from MAK. AlignX in the VNTI7 package (InforMax, Bethesda, MD) was used for the alignments of DNA and protein sequences.

Data sets

Two sets of MITE sequences were used for this study. The families in the first group have reported links to known transposons and were used to test MAK anchor element prediction function. This group includes the families Emigrant/MathE2 (125,129), Tc8 (54), mPIF (55) and *MDM*-2 (143). The families in the second group did not have any reported link to a known transposon family at the time our study was carried out. This group includes families MathE1 (129) and *Kiddo*. The sequence of the dataset is supplied as Appendix A and the information about the MITE families is summarized in Table 4.1. BLAST parameters and other criteria used for analysis are provided on the drop down menu of the MAK web page.

Table 4.1. **Summary of information for MITE families used to test MAK**

Family	Organism	Anchor element	Related Transposase
*Emigrant/MathE*2	*A. thaliana*	AC006161(85200-87313)	*Pogo*
mPIF	*Z. mays*	AF412282 (1 - 3725)	*PIFa*
Tc8	*C. elegans*	AF040643 (24047-31614)	*IS5*
MDM-2	*O. sativa*	AP004320(3568- 9016)	*MURA*
*MathE*1	*A. thaliana*	unknown	*unknown*
Kiddo	*O. sativa*	unknown	*unknown*

RESULTS

Genetic principles and program pipelines

1. Member retriever

MITEs in a family share DNA sequence similarities that are readily detectable using BLAST searches. To illustrate the TIR conservation and relationship among members, an alignment is needed. Sequences of the members can be retrieved from BLAST search results. Since a BLAST hit can be on the top or bottom strand, all of the hit sequences to be used for alignment need to be in the same orientation; therefore, in MAK, the hits on the bottom (minus) strand are reversed. Since sequences adjacent to the TSDs of MITEs are often of interest to researchers, the program was designed to allow the retrieval of flanking sequences. In the Member retriever program, BLASTN searches are initiated against NCBI "nt" and "htgs" databases using a given MITE sequence. The search results are automatically retrieved and the high scoring pairs (HSPs) are parsed. If the query sequence part in an HSP is the full query MITE length, the hit sequence in the HSP is retrieved as a complete element.

If the hit part is in an opposite orientation (minus strand), the reverse sequence of the hit part is retrieved. Then, flanking sequences of user defined length are retrieved (Fig. 4.1A). In addition, long elements that do not show strong similarity along the total length of the query but do at both terminal regions can also be retrieved. using the Long element function (Fig. 4.1B).

Figure 4.1. Diagram of pipelines for MAK.

2. Anchor

MITEs are likely to be derived from various autonomous or receptor transposons. Recently, *MDM*-2, mPIF, Tc8, and Emigrant/MathE2 have been identified to be the derivatives of known transposons (54,55,132,143). The most conserved parts of a DNA transposon family lie in the TIRs because they represent the major transposase recognition sites. Since MITEs are usually so abbreviated that they retain no trace of the transposase coding region, identification of the original transposon relies heavily on their TIRs or subterminal regions. Since the elements from which MITEs are directly derived may not necessarily be the master elements responsible for their transposition, we have denoted these elements as MITE anchors. While anchor elements may not necessarily be the ancestors of the anchored elements, an evolutionary relationship is likely to exist between the anchor elements and the anchored elements. In the automated anchor finding process, a BLASTN is carried out for short matches and the HSPs are parsed for DNA fragments that are at 100 bp longer than query sequence but that do not exceed a total of the specified anchor size limit and match the query element at both ends. These long elements are possible ancestors. To determine if these elements have the potential to encode transposase, a BLASTX is carried out for each of these elements and the hits that contain the word “transposon”, “transposase” or “transposable element” in their titles are retrieved (Fig. 4.1C). False predictions usually result from transposon nesting events and thus can be identified with BLASTN searches. If only the predicted transposase-like regions inside the predicted element are repetitive at the DNA sequence level, such entries are discarded.

3. Associator

Because of their short size, MITE families are potentially less disruptive than classic transposons and they may even contribute beneficially to gene regulation (140). Nevertheless, their large copy numbers suggest that they are potentially disruptive and very few MITEs have been found in coding sequences. It is often desirable to know how closely members in a family are associated with genes and which genes have closest proximity to MITE elements. In Associator (Fig. 4.1D), a BLASTN is carried out and the accessions and positions of significant hits with lengths longer than one fourth of the query are recorded. For each of these significant hits, the name and position of the annotated gene that has the closest proximity to the center of a given MITE element is retrieved. The results for all the significant hits can be exported as a table.

Implementation of MAK

MAK runs on UNIX, Linux and Win32 platforms on which Perl 5.6.1 (http://www.perl.com) and Bioperl 1.0.2 releases (http://www.bioperl.org) are installed. The web based software (http://perl.idmb.tamu.edu/mak.htm) starts with the input of user name, email address, sequence file name and sequence(s). Then the desired function (Member retriever, Long elements, Associator, or Anchor) needs to be selected. The parameters to run Member retriever include the length of sequence flanking the MITE members, the organism in which the MITE family is present and the terminal inverted repeat (TIR) tolerance. For Long element and Anchor functions, a size limit can be selected from 2000, 5000, 10000 or 20000 bp. All the retrieved long elements are at least 100 bp longer than the query sequences. Chosen E_value and organism parameters apply for all MAK functions. Upon initiation of

the program, the user will be notified of the status of the process. The results will be sent to the specified email. While the format for the input sequence is flexible if the analysis is for a single MITE family, FASTA format is highly recommended if the analysis involves multiple MITE families. The MAK program can also be run as a queue job on a supercomputer in which multiple functions of MAK can be used to analyze several MITE families simultaneously.

When the dataset for MITEs (Table 4.1) was used to run MAK, updated information for these MITE families was obtained. The chart in Fig. 4.2 demonstrates the distance of MITEs (in completed genomes) from *MathE*1, *MathE*2 and *Tc*8 relative to their closest genes. When the dataset for MITEs with known relationships was used to run Anchor function, anchor elements for *Emigrant*/*MathE*2, *Tc*8, *mPIF*, *MDM*2 predicted by the MAK were consistent with previous reports (54,55,132,143) (see Table 4.1 and Appendix B).

Anchoring MathE1 and *Kiddo* MITE families

When the MathE1 family was used to run the Anchor function of MAK in the *Arabidopsis* genome, two identical long elements (AC007123, from 6918 to 2690; AF007271, from 16996 to 21224) with identical TIRs to the *MathE*1 element on accession AB010073 were identified. They were predicted to be a transposase gene. The sequence of these two long elements comes from an overlap region of accessions AC007123 and AF007271 on chromosome 5. Thus they represent only one element, which we named as A-MathE1 (anchor of MathE1). It shares 77% sequence identity to one terminus of MathE1 in 35 bp and 80% sequence identity to the other terminus of MathE1 in 82 bp. It has identical 12 bp TIRs to those of *MathE*1 elements. Interestingly, the internal sequences of *MathE*1 elements seem

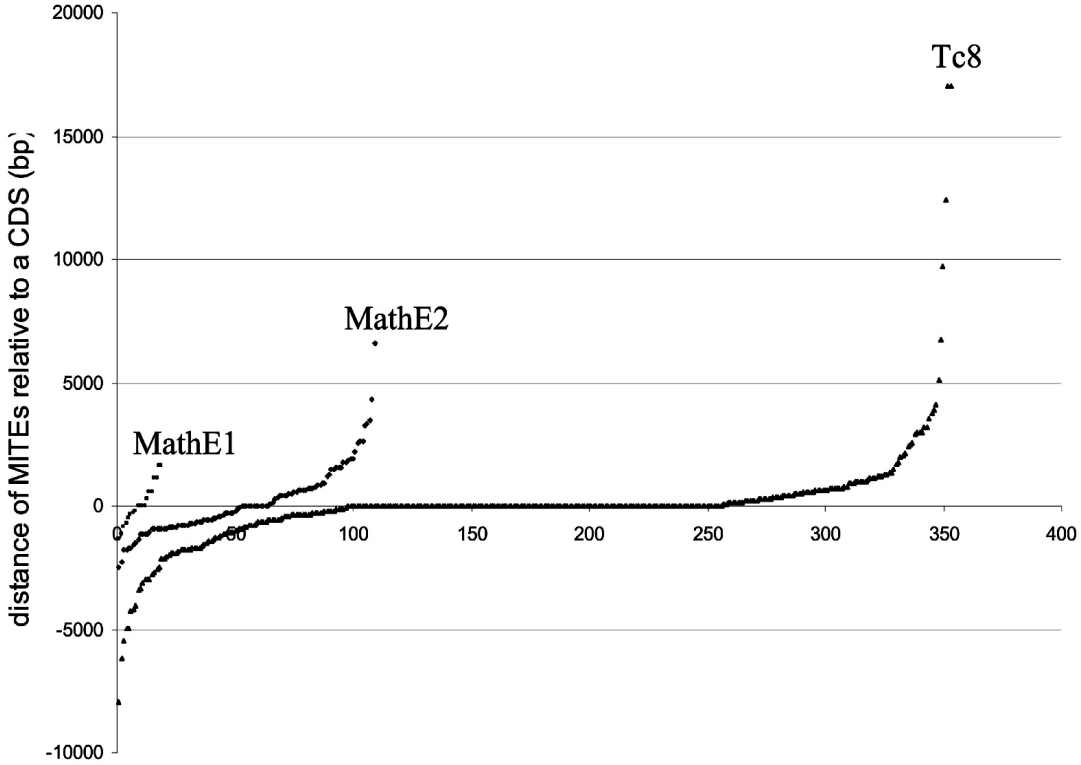

Figure 4.2. Distance of members in MITE families *MathE*1, *MathE*2 and *Tc*8 to their closest genes. The names and positions of the genes closest to the MITEs retrieved with MAK are sorted ascendingly with Microsoft Excel. The distance of a MITE inside a coding sequence (CDS) to the gene is considered 0 and the distance of a MITE at the 5' end of a CDS to the CDS is changed into a negative value. The sorted elements are numbered consecutively, starting from 1. The distance values of MITEs to a CDS are plotted against their numbering. Each unit of x-axis on the chart represents a MITE element and the distance of that MITE to a CDS is shown as the value on the y-axis.

to be derived from 29 blocks of 10 - 30 bp on AC007123 with very little divergence (>90% identity in each block). BLAST searches with the long element sequences resulted in truncated or disarmed elements. An additional long element (AB025602, from 7658 to 11849) showed an overall 98% DNA sequence identity to A-*MathE*1. Since the element on AB025602 is situated on a different locus of chromosome 5 from A-*MathE*1 on AC007123 and they share no flanking sequence similarity, it apparently results from a transposition

event. Six bp missing from the TIR at the 5' end were found to be present on the 3' end flanking sequence. These long elements and MITE family *MathE*1 converge into one transposable element family we have named Math (Fig. 4.3A). This family showed a TSD exclusively of "TTA" and has a TIR of 13 bp.

When the *Kiddo* family (142) was used to run the Anchor function of MAK in the rice genome, three long elements with typical TE characteristics were predicted to be transposase genes or pseudogenes. They are within AP004087 (gi:15281366) from 74902 to 78476 on chromosome 2, AC118347 (gi:20153328) from 20719 to 17088 on chromosome 11, and AP005461 (gi:21624013) from 78912 to 82646 on chromosome 6. These elements showed an overall sequence identity of >92%. When the long element of *Kiddo* on AP004087 was used to do a BLAST search, an additional complete element was found on AF114171 (gi:4680196) from 39268 to 43050 from *Sorghum bicolor* chromosome F. It has an overall sequence identity of 66% to that of AC118347. We name the long element on AC118347 as A-*Kiddo* (anchor of *Kiddo*). The internal DNA sequences of these four long elements are highly conserved in two regions (from ~800 bp to ~2050 bp and from ~2100 bp to ~3200 bp on AC118347; Fig. 4.4B). Additionally, 16 complete (i.e. having TIRs at both ends) elements with sizes ranging from 714 bp to 2538 bp were identified. Like the long elements, they have a consensus TSD of TAA and show high (>85%) similarity in ~250 bp and ~110 bp terminal regions, but their internal sequences do not show similarity to known transposases. They form an intermediate group (*Kiddo*E) between *Kiddo* and A-*Kiddo*. Together, these elements represent a novel transposable element family (Fig. 4.4B), which we named Kid. These have not previously been annotated in the rice genome.

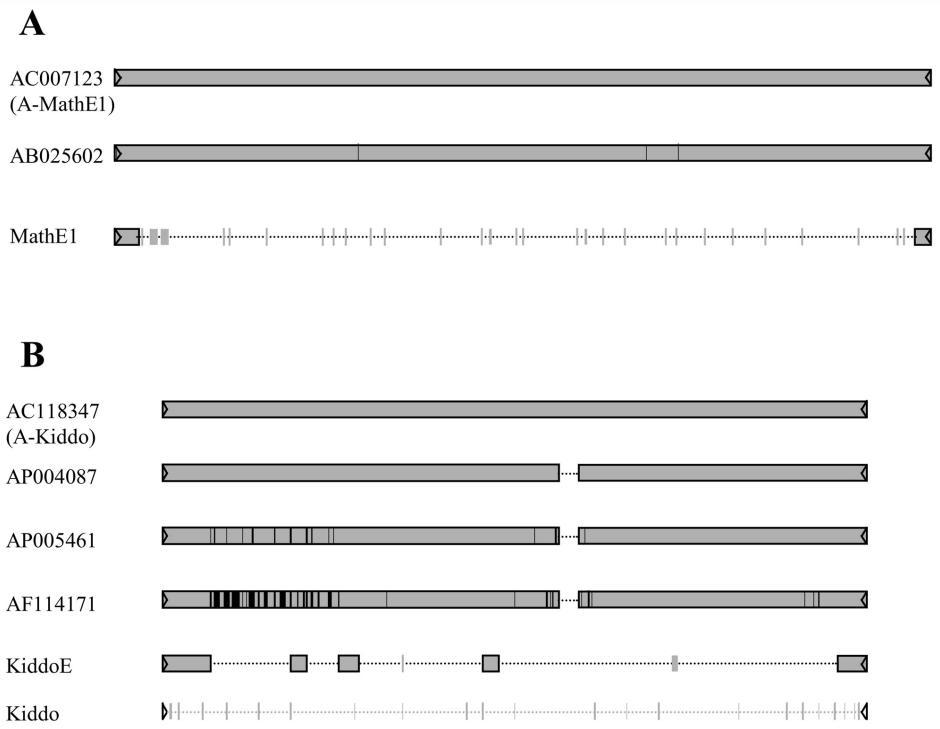

Figure 4.3. Schematic presentation of TE family *Math* (**A**) and *Kid* (**B**). Anchor elements are aligned with similar long elements and corresponding MITE families. Vertical lines in internal regions of long elements (long gray bars) indicate dissimilar regions and vertical lines connected by dotted lines in MITE elements indicate similar sequence blocks on MITEs to the anchor elements. Dotted lines indicate deletion regions (blank regions). The elements are drawn to scale. The triangles at the ends represent TIRs. The accession number on the left of the elements indicate the accession on which the elements are located and the positions for these elements on the accessions are described in Results.

Math and Kid belong to IS5/Harbinger/PIF superfamily

The anchor elements of A-*MathE*1 and A-*Kiddo*E do not share significant DNA sequence similarity with each other. Their internal sequences do not contain repetitive sequences as revealed by BLASTN searches. However, as predicted by MAK, both of them share strong similarity to putative transposase-like proteins. One of the BLASTX hits was from the

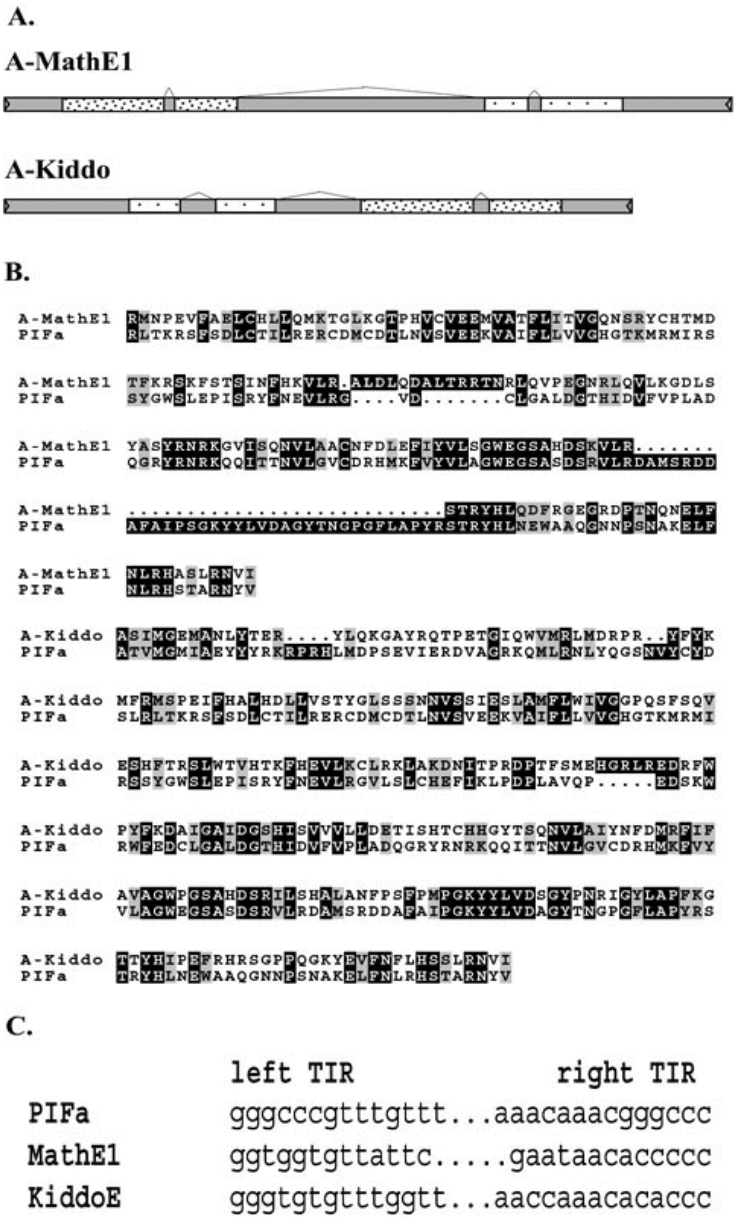

Figure 4.4. Putative gene structure for A-*MathE*1 and A-*Kiddo* (see Appendix C). (**A**) Dotted regions indicate putative coding exons. Exons showing similarity to putative *PIFa* transposase are indicated in densely dotted regions. Bridged regions indicate putative introns. Sequence alignment between A-*MathE*1 putative translation product (from 808-1642 on the DNA sequence) and maize *PIFa* putative transposase (from 70 to 296 on AF412282 protein sequence) (upper panel), and sequence alignment between A-*Kiddo* putative translation product (2044-2956 on the DNA sequence) and *PIFa* putative transposase (from 18 to 296 on AF412282) (**B**). Letters in black indicate identical residues and letters in gray indicate similar residues. Alignment of left TIRs and right TIRs from *PIFa*, *MathE*1 and *KiddoE* (**C**). Dotted lines denote omitted internal sequences.

putative transposase for maize PIFa elements (AF412282). The predicted gene structure for A-*MathE*1 and A-*Kiddo*E is shown in Fig. 4.4A and their putative translated products were aligned with putative PIFa protein, as shown in Fig. 4.4B. PIFa shares a similarity of 46% and 50% in a region of 834 bp (808-1642) and 912bp (2044-2956) to the putative A-*MathE*1 and A-*Kiddo*E proteins, respectively. In addition, they have the same TSD size, and the TIR sequences of these two elements are very similar to those of PIFa (Fig. 4.4C). Indeed, the A-*MathE*1 element on AF007271 was proposed to be a *IS5/Harbinger/PIF* member named *At*-PIF2 (55). These pieces of evidence strongly suggest that the TE families *Math* and *Kid* belong to the *IS5/Harbinger/PIF* superfamily.

DISCUSSION

Advantages and limitations of the computing approach

Using the automated processes in the MAK, we have successfully run a set of MITE families overnight. The output files are in standard format (e.g. FASTA) and thus can be used directly for downstream processes such as alignments and making tables. As noted in the Introduction, conventional MITE analysis is laborious and needs to be repeated each time the database is updated. Clearly, new analyses are appropriate as databases are updated, but this is relatively facile using MAK. However, since the process is heavily dependent on remote BLAST analysis on the NCBI QBLAST server, the program may encounter internal server errors and hence be terminated (if this occurs, an error message will be generated). To lower the chance of encountering an internal server error at NCBI, we usually avoid running the program at peak times (usually daytime on workdays). In the program, we allow

the retrieval of request ID (RID) for 5 times with an interval of 100 seconds before the process is allowed to die. Another alternative is to run stand alone BLAST on a local system, but this approach requires downloading a huge database from NCBI. For the Associator function output, modest manual inspection to remove duplicate entries is necessary because BLAST searches will yield two HSPs at the same DNA locus if the MITE has a typical inverted repeat structure. Further improvement of the program to remove such entries is underway.

Misannotation of PIF-like elements in GenBank

When we used A-MathE1 and A-*Kiddo*E to do BLASTX elements, several hits were titled En/spm-like transposon protein (accessions: NM_148036, NM_104832, NM_128220, NM_148535, AP003450, AB016878, AP000606, NM_148229). These En/spm-like hits were further analyzed using PSI-BLAST and iterations were carried out until no more new hits were found. Unfortunately, we were unable to find detectable peptide sequence similarity between any of these *En/spm*-like transposon proteins and the putative *En/spm* proteins TNPD-TNPA in maize (64), putative *Tam*1 proteins TNP1-TNP2 in *Antirrhinum majus* (180,181), or the putative open reading frame of *Tgm* in soybean (199,200). On the contrary, all of these hits showed strong similarity to PIFa putative transposase protein (AF412282). Together with the fact that A-MathE1 and A-*Kiddo* showed similar TIR sequence to that of the *PIFa*. We believe that these elements were misannotated in the database although it is still possible that *PIFa* and En/spm superfamiles are remotely related because they both have 3 bp TSDs and ~13 bp TIRs.

CHAPTER V

TRANSPOSITION ACTIVITY OF *KIDDO*

INTRODUCTION

Miniature inverted repeat transposable elements are thought to be derived from certain autonomous or receptor transposable elements (54,55,132,143,201). They are usually too short to encode autonomous transposases and they are even lack relic DNA sequences reminiscent of transposase coding regions. However, the high copy numbers usually observed for MITEs indicate that they have been amplified during evolution. How these high copy numbers are achieved is still a mystery.

Despite the lack of any direct evidence for transposition, MITEs are thought to amplify by means of transposition because of the structural characteristics of MITEs, such as target site duplications (TSD) and terminal inverted repeats (TIRs). Classical transposable elements (TEs) jump in two ways: copy-paste and cut-paste, with the retrotransposons (class I) using the former and transposons (class II) using the later approach. Both class I and class II TEs have TSD structures (although the sizes are different), but only class II TEs have TIRs. Class I TEs in donor sites do not leave in a transposition event; instead, new copies are made followed by their insertion into new loci. However, class II elements in donor sites are cleaved out, typically leaving footprints of two copies of the TSDs or incomplete copies of the two TSDs. The presence of TSD footprints is important evidence for distinguishing between class I and class II TEs.

We have previously shown that a *Kiddo* element is present in the *rubq2* promoter of rice line IR24 but is absent in line T309 (142). Sequence comparison of these two promoters revealed the presence of an incomplete TSD footprint of TTATA on the *rubq2* promoter of rice T309. Following our finding that a *MDM1* MITE flanks the *Kiddo* element, we have obtained more *in silica* evidence suggesting that the *Kiddo* element was inserted in the *MDM1* of the common ancestor of rice IR24 and T309 (143). After the divergence of IR24 and T309, the *Kiddo* on the T309 *rubq2* promoter excised, leaving a footprint. Based on the high similarity of *Kiddo* family elements and the footprint evidence, we postulate that the *Kiddo* family was active very recently (2 ~ 3 million years) and may still be able to transpose in contemporary rice plants.

TEs in eukaryotic genomes are usually silenced transcriptionally or post-transcriptionally (191,202-208). Even if a MITE family is able to transpose, it may not be actively transposing because the genes coding for the putative transposase may be in a silenced status, either transcriptionally or post-transcriptionally. Studies have shown that environmental stimuli (such as UV light, heat shock or cold shock) or chemicals (such as 5-azacytidine) are able to reactivate retrotransposons or transposons as a result of the release of silencing of their transposase genes (68,155,204,209,210).

In this study, we used cold shock and 5-azacytidine (5-azC) to treat wild type plants or transgenic plants containing transposition reporter genes. PCR analysis of the *rubq2* promoter using DNA obtained from 500 rice IR24 plants treated with 5-azC did not reveal excision of *Kiddo*. No excision event was detected using either transgenic *Arabidopsis* or rice plants containing transposition reporter gene. These evidence results suggest that *Kiddo*

may not be able to transpose in modern rice. Sequence analysis of the putative historical transposase genes for *Kiddo* indicated that several point mutation may have occurred, resulting in deactivation of the *Kiddo* family.

MATERIALS AND METHODS

Plant transformation

Arabidopsis thaliana ecotype Columbia was transformed using an infiltration protocol (http://www.bch.msu.edu/pamgreen/vac.htm) similar to that of Bechtold and Pelletier (211). Rice transformation was as described by Hiei *et al.* (212-213) and Dong *et al.* (214).

The 5-azC treatment

Seeds from *Arabidopsis* transgenic for pKJ were germinated on MS selection medium (215) containing 50 mg/L kanamycin. After about 14 days, when the seedlings were about to commence inflorescence formation, the plants were transferred to soil in individual 3.5 inch diameter pots and watered every day for about 1 week. One day before treatment, no water was supplied to the plants. The plants were exposed to 5-azC by application of 1 ml of solution containing 100 μg 5-azC, the pots were placed in a covered tray overnight. The next day, regular watering was resumed. Rice IR24 seeds were germinated on MS medium containing 50 mg/L 5-azC for two weeks, after which the seedlings were transferred to soil to obtain seeds.

Cold shock treatment

Seeds of *Arabidopsis* transgenic for pKJ were germinated on Petri plates containing MS selection (50 mg/L kanamycin) medium and exposed to 4°C for 1 hr three times over three days. The plants were transferred to soil and grown to maturity.

PCR amplification

DNA was extracted (158) from 500 progeny seedlings of 5-azC treated IR24 plants (in batches of 10 seedlings) and 500ng from each batch was used for PCR amplification. A PCR Master mix (Promega, Madison, WI) was used for PCR. The cycling parameters were: 94°C for 3 min, 30 X [94°C 1min, 50°C 45 sec, 72°C 2min], and 72°C for 10 min. The primers used for PCR were: 5' aagcttacggaaggaaacaaattcgg 3' and 5' tctagaatgcgaggagaggagatgag 3'.

Selection method

For *Arabidopsis* seeds obtained from 5-azC or cold treatment, MS medium containing 50 mg/L kanamycin and 10 mg/L bialaphos was used to select for *Kiddo* excision events. Rice seeds obtained from transgenic rice containing pUbi-GFP::*Kiddo* T-DNA were screened for GFP expression under a handheld Solarc light source LB-24 (Welch Allyn Inc., Skaneateles Falls, NY).

Sequence analysis

Sequences of anchor elements were retrieved using MAK (see Chapter IV). These anchor sequences were used as queries to find additional sequences that are highly similar to the queries. The identified sequences were aligned using VNTI 7.0 (InforMax, Frederick, MD)

AlignX function with manual adjustment. The mutations were manually counted and a phylogenetic tree was constructed by AlignX.

RESULTS

PCR amplification of the *rubq2* promoter from 5-azC treated rice IR24

Given the evidence for excision of *Kiddo* from the *rubq2* promoter of the common ancestor of rice lines T309 and IR24, it was of interest to attempt to repeat this process in the laboratory. A majority of the transposase genes in modern plants are silent because of homology dependent gene silencing (HDGS) (191,202-208).

5-azC, known to be potent in gene reactivation (216-220) was used to treat IR24 rice seeds (see Materials and Methods) that were subsequently grown to obtain genomic DNA

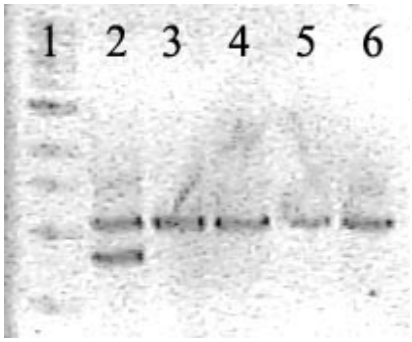

Figure 5.1. PCR amplification of a *rubq2* promoter fragment from seedlings of 5'-azC-treated rice (IR24). Lane 1: size marker. Lane 2: PCR reaction using a mixed template DNA of IR24 and T309 (19:1) used as a positive control. Lanes 3-6 are from PCR reactions using template DNA extracted from progenies of 5'-azC- treated plants.

samples. A pool of ten plants was analyzed by PCR amplification using primers flanking the *Kiddo* element in the *rubq2* promoter (Fig. 5.1). A total of 500 seeds from 20 independent

5-azC treated plants were analyzed and all of them showed the same fragment size as that of the original IR24 plants. This result indicates the transposase for *Kiddo* was not active in the IR24 rice genome.

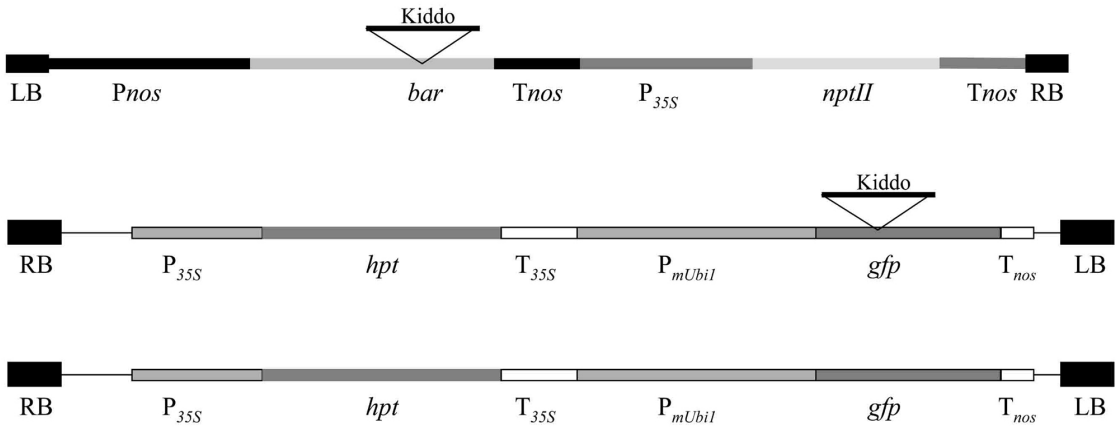

Figure 5.2. T-DNA regions of constructs used for experiments to detect *Kiddo* transposition. (**A**) T-DNA structure of construct pKJ used for *Arabidopsis* transformation, permitting seeds to be selected against kanamycin (100 mg/L). The constructs (**B**) and (**C**) were used for *Agrobacterium*-mediated rice transformation.

Selection for *Kiddo* transformation in transgenic *Arabidopsis*

Arabidopsis plants were transformed with the T-DNA binary vector shown in Fig. 5.2A. Original transformants were selected on MS medium containing 50 mg/L kanamycin. Ten young plantlets from the seeds of the original transformants were treated with 5-azC, ten were treated with cold, and a third group of ten young plantlets were treated with both cold and 5-azC (see: Materials and Methods). Seeds obtained from the treated plants were germinated on MS medium containing 10 mg/L bialaphos and 50 mg/L kanamycin to select

for any *Kiddo* excision events. More than 500 seeds from each treated plant were used for selection but no bialaphos resistant plants were obtained, suggesting that *Arabidopsis* may not contain a transposase gene capable of supporting *Kiddo* transposition.

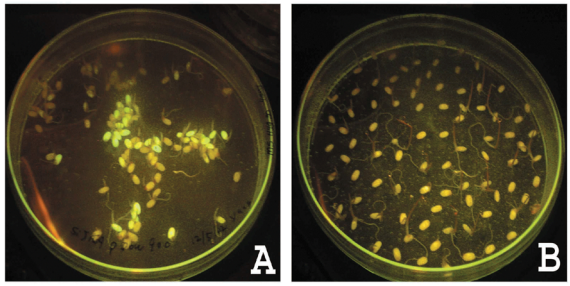

Figure 5.3. No transposition for *Kiddo* was identified in transgenic rice. (**A**) seeds from rice (T309) transgenic for the construct shown in Fig. 5.2C. Green fluorescence denotes seeds expressing GFP driven by the *mUbi1* promoter. (**B**) Seeds from rice (T309) transgenic for the construct shown in Fig. 5.2B. No GFP- expressing seeds wer found among 500 seeds. An intense blue light excitation source (see Selection Method) was used together with a 500nm filter.

Selection for *Kiddo* transformation in transgenic rice

Calli induced from rice T309 seeds were transformed with the T-DNA binary plasmid shown in Fig. 5.2B. A total of 10 transformants were obtained, grown to maturity, and 500 seeds were screened for GFP expression. Ten transgenic rice plants containing the T-DNA construct shown in Fig. 5.2C were used as a positive controls; eight of the ten yielded GFP-expressing seeds (Fig. 5.3A). Since none of the seeds from the test plants expressed GFP (Fig. 5.3B), it appears that the rice T309 genome does not encode a transposase for *Kiddo*, or that the frequency of transposition is unusually low.

***In silico* analysis of putative *Kiddo* transposase genes**

The experiments described above suggest that *Kiddo* is not active in modern *Arabidopsis*, *Japonica* rice T309, or *Indica* rice IR24. However, the large population of *Kiddo* elements in rice clearly indicates its historical activity. Since the divergence of *Indica* and *Japonica* rice occurred some 2 to 3 million years ago (162), it is possible that *Kiddo* lost its functional transposase within this time frame. In this case, it is likely that ancestral elements exist that bear limited mutations responsible for its functional inactivation.

Using the MAK program described in Chapter IV, we were able to retrieve ancestral elements that share high similarity (>90%) to each other in their DNA sequences. An alignment of the sequences revealed that each of these elements may contain mutations in the protein coding sequences.

The phylogenetic tree revealed that the closest element to the consensus sequence of these elements is the element from AC118347 (Fig. 5.4), named as A-*Kiddo* in Chapter IV. A comparison of the mutation frequency between the closest two elements (from AC118347

Figure 5.4. Phylogenetic tree derived from alignment of ancestral elements for *Kiddo*. Accession numbers represent the corresponding ancestral elements.

and AP004087) revealed a lower degree of sequence divergence for the element from AC118347. These results indicate that the element from AC118347 is most likely to be the historical transposase-coding element since this sequence contained the fewest mutations. The TIR of this element is identical to that of the *Kiddo* family (see Fig. 4.4C in Chapter

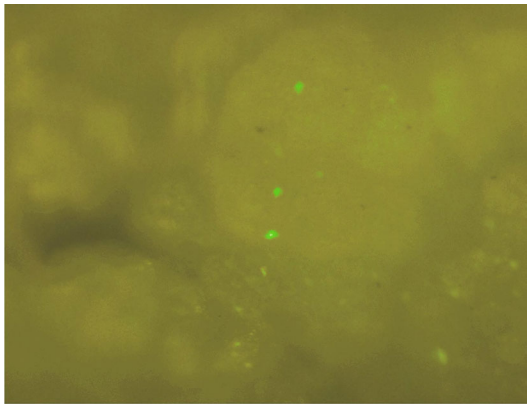

Figure 5.5. Transient GFP expression driven by the putative promoter for A-*Kiddo*. Green spots represent cells transformed with a pBJ81 (Battraw, M. J. and Hall, T. C., unpublished)-derived plasmid containing *mgfp5-er* fused to a 726 bp fragment at the 5' end of A-*Kiddo*. The 3' end of the 726 bp fragment is immediately upstream of the translation start.

IV). The predicted gene structure contains three exon regions and three intron regions. The promoter region was used to drive transcription of *gfp* coding sequence and the construct was used to do a particle bombardment experiment. The presence of green spots on rice calli after bombardment (Fig. 5.5) confirm that the predicted promoter region can drive the transcription of *gfp*. A rough estimate of activity, carried out by counting the GFP intensity

of the bombardment spots, suggested an activity equivalent to ~27% of that of the 35S promoter.

DISCUSSION

Excision vs. replication-coupled transposition

No excision of *Kiddo* was observed in our experiments, the lack of functional transposase being a likely reason. However, transposons are able to transpose without necessarily leaving an excised original copy (74,88,221-226). Additionally, the large copy numbers observed for MITEs indicate a duplicative mechanism of MITE amplification. The strict cut-and-paste mechanism does not result in an increase in the copy number. The amplification of transposons is often the result of transposition accompanied by replication and DNA repair. When the element on one daughter molecule of two newly synthesized molecules is excised and inserted in another location, the new daughter molecule may obtain a new copy of the transposon from the other daughter DNA molecule through homology dependent recombination. This mechanism may explain the large copy number for many MITE families. From copy number and footprint evidence, it is possible that MITEs transpose both through a conservative cut-and-paste approach and a coupled replication-repair mechanism.

It is also possible that the sample size we used was too small to detect any *Kiddo* excision event. We believe that this is less likely because, although the reported transposition efficiency for transposons is generally low (several out of a thousand) (223,225), the short

element length is likely to increase transposition frequency dramatically (227). Additionally, during the preparation of this dissertation, a transposition frequency of 10-30% during meiosis was reported for *mPing*, a newly identified MITE family in the rice genome (228-230).

Evolution of the *IS5/Harbinger/PIFa* superfamily

Very recently, a new transposon superfamily in plants named *IS5/Harbinger/PIFa* was discovered (53,55). The founding member of this superfamily is the P instability factor (PIF). Miniature PIF (*mPIF*) elements were found to have an insertion site preference of 5' CWCTTAGWG 3' and to produce target site duplication sequences of TAA or TTA. Interestingly, more MITE families, including Tc8 (54), MathE1, *Kiddo* (56) and mPing (228-230), have recently been identified as belonging to this superfamily. Except for the *Mutator* superfamily (143), no MITEs have been identified for the widely known TE superfamilies in plants such as *Ac/Ds* and *En/spm*. In fact, we have obtained additional evidence suggesting that several other MITE families in rice should also be classified into *IS5/Harbinger/PIFa* superfamily (data not shown). One reason for the large population of MITEs belonging to the *IS5/Harbinger/PIFa* superfamily may be that the transposases in this superfamily are much less dependent on the internal regions of the elements than are those of *Ac/Ds* and *En/spm*. This hypothesis is supported by the observation that these MITEs are conserved in the TIR region but share no detectable similarity in the internal regions. If this is an inherent attribute of this superfamily, it will be a very useful genetics

tool for mutagenesis or gene transformation because the only region required for transposability is the TIR region, which is usually around 15 bp.

CHAPTER VI

A TWO EDGED ROLE FOR *KIDDO* IN THE *RUBQ2* PROMOTER

INTRODUCTION

MITEs are numerous in higher eukaryotic genomes, including human, *Drosophila*, *C. elegans, Arabidopsis*, rice and maize. Because of the difficulty in distinguishing between genic and non-genic regions, no definitive distribution pattern has been discerned for MITEs relative to genes. However, it is evident that many MITEs are very close to or inside coding sequences. Since they are located in promoters, 5' untranslated regions (5' UTRs), introns, and 3' untranslated regions (3' UTRs), their potential effects on gene regulation are of great interest. Insertion of a MITE into a promoter may be disruptive because it may disturb promoter cis-elements, DNA topology or interactions with enhancers. Alternatively, MITEs in promoters could be constructive because they typically include *cis* elements capable of affecting transcription (142,144). Although a large number of MITE families have been identified in rice (137,142,154,175,176,187), their roles in gene regulation are not known.

As a potentially strong, ubiquitously expressed promoter for monocots, the *rubq2* promoter was recently cloned and sequenced from rice line IR24 (159). However, our discovery of two nested MITEs raises concerns about its use for agronomic purposes. A MITE named *MDM1* (143) was found upstream of the transcription start site (defined as +1) between base positions -823 and -201bp. This MITE family contains about 400 members in

the rice genome and the elements share modest similarity. Another MITE, named *Kiddo*, was found nested inside *MDM1* and it is located between -515 bp and -244 bp on the *rubq2* promoter in rice line IR24 (accession: AF184280) (142). This MITE family has about 1200 members in the rice genome and the sequences in a subgroup share very high similarity (>90%). Together, the two MITEs comprise more than 75% of the 823 bp promoter defined by Wang *et al.* (159) and it is evident that the presence of these MITEs could affect transcriptional activity of the promoter in several ways. For example, integration of these MITEs into the *rubq2* promoter may decrease its activity. Indeed, since MITEs are highly repetitive sequences that are subject to homology dependent gene silencing, their presence could result in complete silencing of the *rubq2* promoter.

To explore the effects of the MITEs on expression from the *rubq2* promoter, we introduced several truncations and ligated them to the m*GFP*5-ER reporter (231). Using a novel quantitation method for assessment of transient expression following particle bombardment of rice calli, we determined that the insertion of *Kiddo* and *MDM1* into the *rubq2* promoter stimulated, rather than disrupted, its activity. Analysis of stably transformed rice calli yielded similar results. Nevertheless, we also observed that the presence of the *Kiddo* element also exerted repressive effects through gene silencing.

MATERIALS AND METHODS

Plasmid construction

Rice *rubq*2 promoters with or without *Kiddo* were amplified from rice lines IR24 and T309, respectively, using *pfu* DNA polymerase (Stratagene). The PCR products were cloned into

plasmid pBJ81 (Battraw, M. J. and Hall, T.C., unpublished) at the *Hin*d III and *Xba* I sites and a m*GFP*5-ER sequence (231) was used to replace the *uidA* reporter gene from pBin19 (232). Truncation of the two *rubq2* promoters was carried out at both the 5' and the 3' ends. For stable transformation of rice calli, binary vectors pBRubq2-953(dIn-d*Kiddo*) and pBRubq2-953(dIn) were constructed by cloning DNA fragments from pRubq2-953(dIn-d*Kiddo*) and pRubq2-953(dIn) into pJD7 (233) between *Hin*dIII and *Nsi*I sites.

Particle bombardment

Dehusked mature seeds from rice line T309 were rinsed with 70% ethanol for 1 min, then incubated in 50% (v/v) bleach for 45 min with shaking at 120 rpm. The seeds were then washed five times with sterile distilled water prior to plating on N6 medium (234), embryo face-up, for two weeks at 28 °C. Induced calli were subcultured on N6 medium. After 10 to 14 days culture, actively growing calli were selected and precultured on high osmolarity N6 medium supplemented with mannitol and sorbitol (0.3 M each) for 4 h prior to bombardment using a Biolistic Particle Delivery System model PDS-1000 (E. I. du Pont de Nemours & Co., Wilmington, DE). For each experiment, the calli were bombarded twice with gold particles (1 mg; 1.0 μm diameter.) coated with pRubq2-953, pRubq2-953(d*Kiddo*), pRubq2-953(dIn), pRubq2-953(dIn-d*Kiddo*), pRubq2-812(dIn), or pRubq2-812(dIn-d*Kiddo*) DNA (1 μg).

Fluorescence intensity measurement and data processing

Bombarded calli were incubated at 26°C in the dark for 48 hr. For each plasmid, 100 randomly selected green fluorescent spots were imaged at resolution 1030X1030 using a

Zeiss SV11 microscope fitted with a Zeiss AxioCam HRc and AxioVision software. The exposure time was adjusted so that the pixels in spot images were not saturated. The mean fluorescent value of each individual spot was measured using the interactive measurement function of AxioVision. Each reading was normalized by subtracting the background reading. For each plasmid bombarded, an average value was taken for the normalized readings for 100 GFP spots. The same procedure was used for the quantitation of stably transformed rice calli, except that the normalized readings were corrected by subtraction of normalized readings for non-bombarded calli.

Agrobacterium-mediated stable transformation

Embryogenic callus was selected from calli induced from mature seeds of rice line T309 as described for bombardment. Transformation was based on the method of Hiei *et al*. (212-213) and Dong *et al*. (214). The selected calli were co-cultivated with *Agrobacterium* containing binary plasmids pBRubq2-953(dIn-d*Kiddo*) or pBRubq2-953(dIn) in the dark at 21°C for three days. After co-cultivation, calli were rinsed with sterile distilled water containing cefotaxime (250 mg/l) and transferred to N6 selection medium supplemented with 2, 4-D (2 mg/l), hygromycin (50 mg/l), and cefotaxime (250 mg/l). Calli were transferred to fresh selection medium every two weeks. After four weeks culture, calli exhibiting GFP expression were selected under the fluorescence microscope and transferred to fresh N6 selection medium. GFP intensity measurements for these calli were obtained after a further week of culture.

5-azC treatment of the GFP-silenced calli

Calli showing strong GFP expression in the quantitation assays were subcultured monthly onto on fresh selection medium. After 8 weeks, calli that no longer expressed GFP were counted and transferred onto N6 medium containing 50 mg/L 5-azC. GFP expression was examined after one and two weeks of culture in 5-azC.

RESULTS

A novel method to measure GFP expression in bombarded callus cells

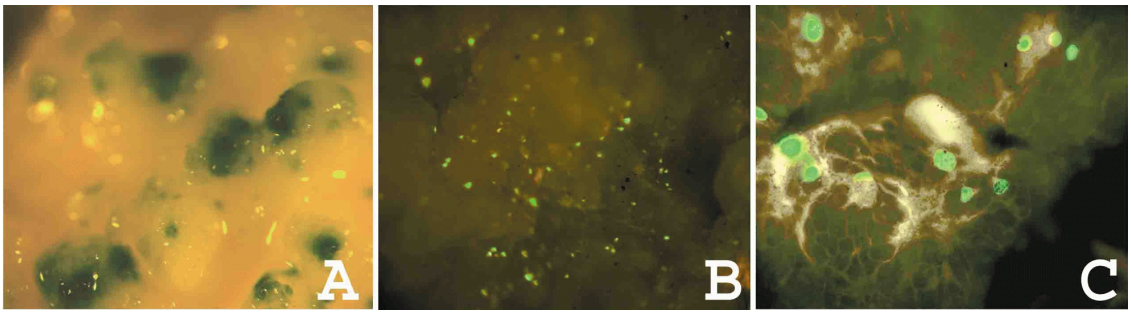

Figure 6.1. Rice calli bombarded with constructs containing *uidA* or *mgfp5-er* reporter genes. GUS staining (blue areas) of rice calli bombarded with a *uidA* reporter gene (32x magnification) (**A**); GFP expression (green spots) of rice calli bombarded with *mgfp5-er* (32x magnification) (**B**); GFP expressing cells from (**B**) under higher magnification (660x) (**C**).

Although *uidA* has been used extensively as a reporter gene for callus bombardment experiments, the reaction leading to blue histochemical staining involves soluble intermediates that diffuse from the cell(s) of origin (Fig. 6.1A). Since GFP expressed from *mgfp5-ER* targets the ER membrane, it is essentially cell-autonomous. Bombardment

experiments yielded discrete fluorescent spots of similar size (Fig. 6.1B) that, under higher magnification, appeared to be single cells (Fig. 6.1C). Each of these single cells represents an independent transformation event, permitting the transient GFP expression level to be measured as the average GFP intensity per cell.

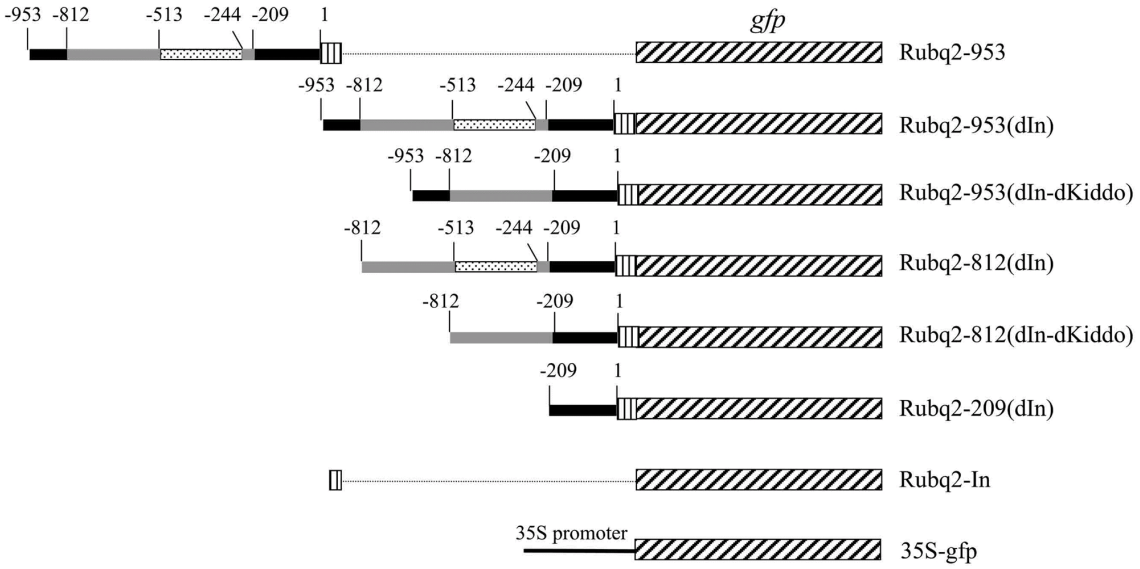

Figure 6.2. Diagram of truncated *rubq2* promoters fused to *mgfp5-er*. The numbered positions are relative to the transcriptional start site (+1). Dotted bars represent MITE *Kiddo*. Gray bars indicate MITE *MDM1*. Dotted lines denote *rubq2* intron 1 and vertically stripped bars indicate 5' UTR of *rubq2*.

Qualitative analysis of *rubq2* promoter truncations

We wished to measure the contribution of MITE insertions to *rubq2* promoter activity. Several truncations of the *rubq2* promoter are shown in Fig. 6.2, some of which include deletion of the intron. All of the truncated promoters were functional in bombarded rice calli. Plasmids pRubq2-953(dIn-d*Kiddo*) and pRubq2-953(dIn) were used for quantitation in bombarded calli and their corresponding binary vectors, pBRubq2-953(dIn-d*Kiddo*) and

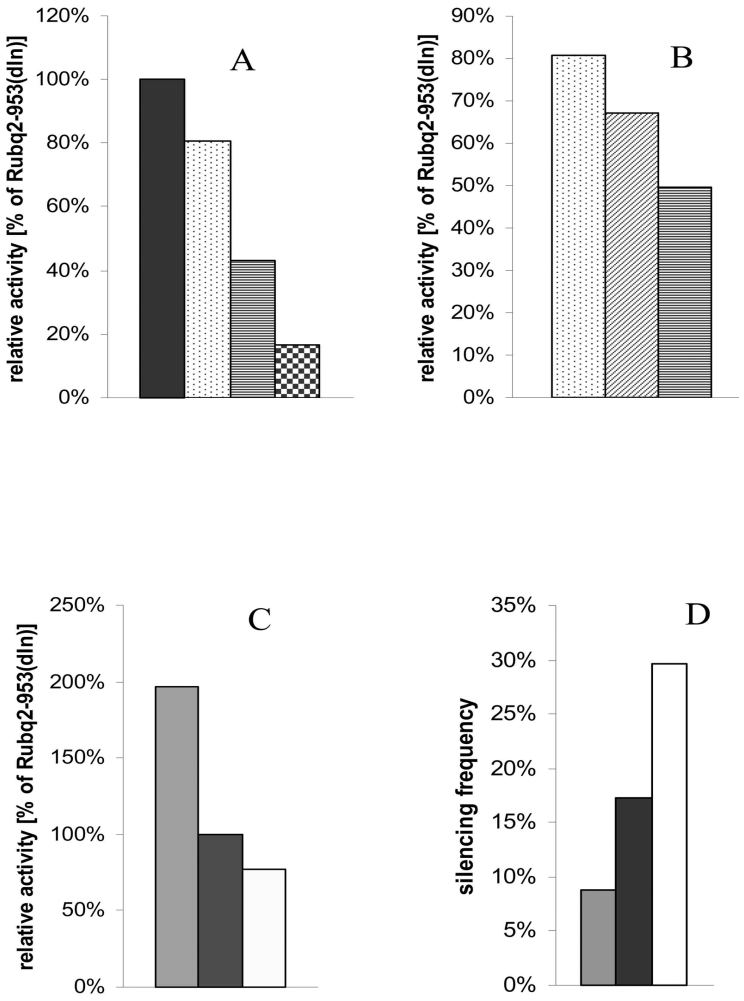

Figure 6.3. Quantitation of the effects of *Kiddo* and *MDM*1 on the *rubq2* promoter. (**A**) Transient GFP expression from pRubq2-953(dIn) (black bar), pRubq2-812(dIn) (dotted bar), pRubq2-209(dIn) (patterned bar) and pRubq2-In (checkerboard bar). (**B**) Transient GFP expresssion from pRubq2-812(dIn) (dotted bar), pRubq2-812(dIn-d*Kiddo*) (diagonal stripes) and pRubq2-209(dIn) (patterned bar). The activity of pRubq2-953(dIn) was designated as 100%. (**C**) GFP expression from stably transformed rice calli containing p35S-gfp (gray bar), pRubq2-953(dIn) (black bar) and pRubq2-953(dIn-d*Kiddo*) (white bar). (**D**) Silencing frequency of 3 month old calli transgenic for p35S-gfp (gray bar), pRubq2-953(dIn-d*Kiddo*) (black bar) and pRubq2-953(dIn) (white bar). A total of 150 calli was used. (The standard errors are less than 7% and 2% of the mean values for transient analysis and stable transformation analysis respectively).

pBRubq2-953(dIn), were used for quantitation in stably transformed calli. Rice calli (T309) bombarded with pRubq2-In (containing only a portion of the 5' UTR and the intron in front

of the coding region) showed GFP expression, as has previously been observed for wheat calli bombarded with constructs containing the *mUbi1* intron 1 ligated to the GUS reporter (235). To simplify our analysis, we used constructs that do not contain the intron sequence for further study.

Contribution of *Kiddo* to *rubq2* promoter activity

Measurement of transient expression was made by determining the average green fluorescence intensity at 500 nm emission wavelength for 100 fluorescent cells of bombarded rice calli. The activity obtained using pRubq2-953(dIn) was defined as 100%. When pRubq2-953(dIn), pRubq2-812(dIn) or pRubq2-209(dIn) were used for bombardment, a 40% decrease in activity was seen for pRubq2-209(dIn), compared with that for the other two plasmids, presumably reflecting the deletion of the two MITEs (Fig. 6.3A). Comparison of GFP expression levels for calli bombarded with pRubq2-812(dIn), pRubq2-812(dIn-d*Kiddo*) and pRubq2-209-In revealed that *Kiddo* and *MDM1* contributed ~15% and ~20%, respectively, to *rubq2* promoter activity (Fig. 6.3B). In addition, the sequence between -812 and -953 contributes about 20% to the *rubq2* core promoter activity.

Plasmids pBRubq2-953(dIn-d*Kiddo*) and pBRubq2-953(dIn) were used for *Agrobacterium*-mediated rice callus transformation. GFP expression was determined for independent, stably transformed calli (150 for each construct) five weeks after transformation. A contribution of ~20% to *rubq2* core promoter activity was found for the *Kiddo* element (Fig. 6.3C), which is in close agreement with the result obtained from

bombardment. Taken together, these data provide evidence that, rather than being disruptive, *Kiddo* and *MDM1* play a beneficial role in *rubq2* promoter activity.

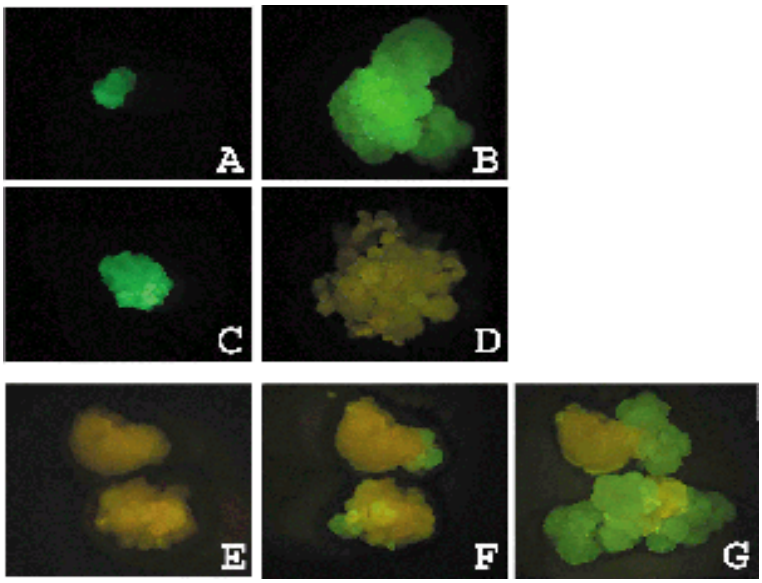

Figure 6.4. Silencing and reactivation of GFP expression driven by *rubq2* promoters in transgenic calli. (**A**) and (**C**): one month old calli expressing GFP. (**B**) Three month old callus from (**A**), still expressing GFP. (**D**) Three month old callus from (**B**) that has lost GFP expression. (**E**) Image taken immediately after GFP non-expressing calli were subcultured onto selection medium containing 5'-azC; the same calli are shown after (**F**) one week and (**G**) two weeks of incubation.

***Kiddo*-induced silencing of the *rubq2* promoter**

Calli transformed with pBRubq2-953(dIn-d*Kiddo*) and pBRubq2-953(dIn) that showed strong GFP expression were propagated on N6 selection medium containing 50 mg/L hygromycin. After 13 weeks, GFP expression was seen to be silenced in ~18% of calli transformed with pBRubq2-953(dIn-d*Kiddo*) and in ~30% of the calli transformed with pBRubq2-953(dIn). However, only ~9% of the calli transformed with a control plasmid *35S-gfp-nos* in pJD7 backbone (233) were silenced (Fig. 6.3D and Fig. 6.4A-D). The higher

silencing frequency observed for the *Rubq2*/*gfp* calli may be the result of homology-dependent gene silencing induced by the endogenous *rubq2* gene or the MITEs. Since the only difference between the two *rubq2* constructs is the presence or absence of *Kiddo*, the higher silencing frequency for pRubq2-953(dIn) over that for pRubq2-953(dIn-d*Kiddo*) can be attributed to the presence of *Kiddo*.

To test if GFP expression could be recovered in the silenced calli, they were transferred onto fresh Petri plates containing an N6 medium with 5-azC (50 mg/L). After two weeks, all cells of new calli growing out from the silenced calli were found to express GFP (Fig. 6.4E-G). Thus, silencing of GFP expression in the calli appears to be accompanied by increased methylation of the transgene(s).

DISCUSSION

Evolutionary implications for the two-edged role of MITE

In this study, a two-edged role was found for *Kiddo* and *MDM1* since these elements contributed positively to promoter activity but also induced silencing of the promoter. For endogenous genes, these two effects may act simultaneously as a fine-tuning mechanism to keep the expression level of a gene at a certain level to avoid over expression. The fact that MITEs have been found in the promoters of some constitutive genes including *Adh1*, *amylase*, *actin*, *NMDP II* and *ubiquitin* (145) may reflect the importance of MITEs in controlling the expression level of these genes. It is also possible that the two opposite effects may function differentially in certain tissues and at various developmental stages.

Interestingly, neither of the positive or negative effects of either MITE element is so dramatic that the expression pattern of the promoter is significantly changed. Considering the high copy numbers of MITEs in the genome and their close association with genes, a "minor" effect of MITEs on gene expression may be beneficial to the genome. Since minor effects probably better tolerated than are dramatic changes, such MITE insertions have a better chance to persist through future generations. These tolerable changes may provide an approach for optimization of genome structure to dynamically adapt to the constantly changing environments of higher eukaryotic organisms.

GFP as a reporter in callus bombardment and its quantitation

To accurately evaluate the effects of MITE insertions on gene expression by particle bombardment, it was necessary to establish a quantitative analytical system. While fluorimetric analysis of GUS expression (236,237), normalized for protein content, is a widely-used approach it is of necessity an approximation of activity because of cell-to-cell diffusion of β-glucuronidase. In contrast, GFP expression from the *mgfp5-er* is cell-autonomous (Fig. 6.1C), permitting analysis of expression levels/cell. In this work, we took advantage of this characteristic of GFP and used fluorescence imaging to quantitate GFP expression levels.

Application of the *rubq2* promoter in research and industry

Combining the results obtained from the bombarded calli and the stably transformed rice calli, we estimate that the contribution of *Kiddo* to *rubq2* core promoter activity is ~20%. Although the present experiments showed that a minimal rice *rubq2* promoter element

together with intron 1 drove substantial reporter gene expression (Fig. 6.1), we did not further study promoter-intron interactions nor such interactions with the effect of MITE insertions and it is possible that such interactions could act positively or negatively. More studies, including examination of promoter activity in progeny generations is merited. However, for driving expression in transgenic plants, the induction of higher silencing frequencies engendered by the presence of the MITE insertions suggests that the *rubq2* promoter without the *Kiddo* insertion is advisable. This precaution would be even more advisable if conditions are found that stimulate movement of the MITE element(s).

CHAPTER VII

SUMMARY

Repetitive sequences are found in both prokaryotic and eukaryotic genomes and are especially prevalent in eukaryotes. Transposable elements (TEs), a group of moderately repetitive sequences, constitute a major portion of many eukaryotic genomes. Recently identified miniature inverted repeat transposable elements (MITEs) have large copy numbers and are of interest for studies in genetics, genome evolution and gene regulation. In this study, the novel MITE families *Kiddo*, *MDM1*and *MDM2* were identified in the rice genome. The *MDM1* and *MDM2* families were characterized as being derived from elements belonging to the *Mutator* TE superfamily.

In a search for autonomous elements for *Kiddo* (i.e. sequences containing a functional transposase), we carried out experiments to detect transposition activity of *Kiddo* in rice and *Arabidopsis*. No transposition was seen for *Kiddo*, even under conditions imposing genome stresses.

To facilitate the *in silico* analysis of MITEs, we established a web-based computer program named MAK. In addition to identification of ancestral elements for MITEs, MAK is also capable of retrieving MITE sequences and the genes associated with MITEs. Using MAK, ancestral elements were identified for *Kiddo* in the rice genome that carry nonsense and frame-shift mutations when compared to a consensus sequence. Together with the fact

that we were unable to stimulate experimental transposition of *Kiddo*, this evidence indicates that *Kiddo* and its autonomous elements are not active in rice.

The presence of the *MDM1* and *Kiddo* in the *rubq2* promoter provided a good opportunity to study the effects of MITE insertions in promoters. Quantitation of transient and stable expression levels for GFP driven by truncated *rubq2* promoters revealed that *Kiddo* and *MDM1* contributed to promoter activity. However, a silencing effect on the promoter was also seen for these MITEs. Neither the contribution nor the repression effect on expression mediated by *Kiddo* and *MDM1* was dramatic. Given the large copy numbers for MITEs and their association with genes, it appears possible that the presence of MITEs may be beneficial to the host genome. If it is found that MITEs can be mobilized in response to environmental signals, it is likely that they play a role in genome regulation and evolution.

REFERENCES

1. Krakauer, D.C. and Nowak, M.A. (1999) Evolutionary preservation of redundant duplicated genes. *Semin Cell Dev Biol*, **10,** 555-559.

2. Tartof, K.D. (1975) Redundant genes. *Annu Rev Genet*, **9,** 355-385.

3. Elder, J.F., Jr. and Turner, B.J. (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. *Q Rev Biol*, **70,** 297-320.

4. Kidwell, M.G. and Lisch, D.R. (2000) Transposable elements and host genome evolution. *Trends Ecol. Evol*, **15,** 95-99.

5. Bancroft, I. (2002) Insights into cereal genomes from two draft genome sequences of rice. *Genome Biol*, **3**, reviews 1015.1- reviews 1015.3.

6. Bennett, P. (2000) Demystified ... microsatellites. *Mol Pathol*, **53,** 177-183.

7. Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature*, **314,** 67-73.

8. Schlotterer, C. (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma*, **109,** 365-371.

9. Britten, R.J. and Kohne, D.E. (1968) Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science*, **161,** 529-540.

10. Singer, M.F. (1982) Highly repeated sequences in mammalian genomes. *Int Rev Cytol*, **76,** 67-112.

11. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. and Yurov, Y. (2001) Alpha-satellite DNA of primates: old and new families. *Chromosoma*, **110,** 253-266.

12. Jenuwein, T. and Allis, C.D. (2001) Translating the histone code. *Science*, **293,** 1074-1080.

13. Turner, B.M. (2002) Cellular memory and the histone code. *Cell*, **111,** 285-291.

14. Ahmad, K. and Henikoff, S. (2002) Histone H3 variants specify modes of chromatin assembly. *Proc Natl Acad Sci USA*, **99 Suppl 4,** 16477-16484.

15. Doudna, J.A. and Rath, V.L. (2002) Structure and function of the eukaryotic ribosome: the next frontier. *Cell*, **109,** 153-156.

16. Spirin, A.S. (2002) Ribosome as a molecular machine. *FEBS Lett*, **514,** 2-10.

17. Ramakrishnan, V. (2002) Ribosome structure and the mechanism of translation. *Cell*, **108,** 557-572.

18. Beckmann, R., Spahn, C.M., Eswar, N., Helmers, J., Penczek, P.A., Sali, A., Frank, J. and Blobel, G. (2001) Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell*, **107,** 361-372.

19. Henikoff, S., Ahmad, K. and Malik, H.S. (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293,** 1098-1102.

20. Meluh, P.B., Yang, P., Glowczewski, L., Koshland, D. and Smith, M.M. (1998) Cse4p is a component of the core centromere of *Saccharomyces cerevisiae*. *Cell*, **94,** 607-613.

21. Takahashi, K., Chen, E.S. and Yanagida, M. (2000) Requirement of Mis6 centromere connector for localizing a CENP-A-like protein in fission yeast. *Science*, **288,** 2215-2219.

22. Buchwitz, B.J., Ahmad, K., Moore, L.L., Roth, M.B. and Henikoff, S. (1999) A histone-H3-like protein in C. elegans. *Nature*, **401,** 547-548.

23. Henikoff, S., Ahmad, K., Platero, J.S. and van Steensel, B. (2000) Heterochromatic deposition of centromeric histone H3-like proteins. *Proc Natl Acad Sci USA*, **97,** 716-721.

24. Henikoff, S. (2000) Heterochromatin function in complex genomes. *Biochim Biophys Acta*, **1470,** O1-8.

25. Hendriks, R.W., Hinds, H., Chen, Z.Y. and Craig, I.W. (1992) The hypervariable DXS255 locus contains a LINE-1 repetitive element with a CpG island that is extensively methylated only on the active X chromosome. *Genomics*, **14,** 598-603.

26. Vafa, O. and Sullivan, K.F. (1997) Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. *Curr Biol*, **7,** 897-900.

27. Ono, S. (1972) So much "junk" DNA in our genome. *Brookhaven Symp Biol*, **23,** 366-370.

28. Zuckerkandl, E. (1992) Revisiting junk DNA. *J Mol Evol*, **34,** 259-271.

29. Hurst, G.D. and Werren, J.H. (2001) The role of selfish genetic elements in eukaryotic evolution. *Nat Rev Genet*, **2,** 597-606.

30. Charlesworth, B., Sniegowski, P. and Stephan, W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, **371,** 215-220.

31. Doolittle, W.F., Kirkwood, T.B. and Dempster, M.A. (1984) Selfish DNAs with self-restraint. *Nature*, **307,** 501-502.

32. Doolittle, W.F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284,** 601-603.

33. Dover, G.A., Flavell, R.B. and Systematics Association. (1982) *Genome evolution*. The Systematics Association special volume, no. 20, Published for the Systematics Association by Academic Press, New York.

34. Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes. *Annu Rev Genet*, **32,** 339-377.

35. van Belkum, A., Scherer, S., van Alphen, L. and Verbrugh, H. (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev*, **62,** 275-293.

36. Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J.L. and Marques, S. (2002) Species-specific repetitive extragenic palindromic (REP) sequences in *Pseudomonas putida*. *Nucleic Acids Res*, **30,** 1826-1833.

37. Stern, M.J., Ames, G.F., Smith, N.H., Robinson, E.C. and Higgins, C.F. (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell*, **37,** 1015-1026.

38. Hulton, C.S., Higgins, C.F. and Sharp, P.M. (1991) ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. *Mol Microbiol*, **5,** 825-834.

39. Gilson, E., Clement, J.M., Brutlag, D. and Hofnung, M. (1984) A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. *EMBO J*, **3,** 1417-1421.

40. Higgins, C.F., Ames, G.F., Barnes, W.M., Clement, J.M. and Hofnung, M. (1982) A novel intercistronic regulatory element of prokaryotic operons. *Nature*, **298,** 760-762.

41. Gilson, E., Perrin, D., Saurin, W. and Hofnung, M. (1987) Species specificity of bacterial palindromic units. *J Mol Evol*, **25,** 371-373.

42. Bachellier, S., Clement, J.M. and Hofnung, M. (1999) Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res Microbiol*, **150,** 627-639.

43. Frank, A.C., Amiri, H. and Andersson, S.G. (2002) Genome deterioration: loss of repeated sequences and accumulation of junk DNA. *Genetica*, **115,** 1-12.

44. Makarova, K.S., Aravind, L., Wolf, Y.I., Tatusov, R.L., Minton, K.W., Koonin, E.V. and Daly, M.J. (2001) Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol Mol Biol Rev*, **65,** 44-79.

45. McClintock, B. (1947) Cytogenetic studies of maize and *Neurospora*. *Carnegie Institution of Washington Year Book*, **46,** 146-152.

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

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

48. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408,** 796-815.

49. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B.,

Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409,** 860-921.

50. Ohshima, K., Hamada, M., Terai, Y. and Okada, N. (1996) The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. *Mol Cell Biol*, **16,** 3756-3764.

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

52. Shedlock, A.M. and Okada, N. (2000) SINE insertions: powerful tools for molecular systematics. *Bioessays*, **22,** 148-160.

53. Jurka, J. and Kapitonov, V.V. (2001) PIFs meet Tourists and Harbingers: a superfamily reunion. *Proc Natl Acad Sci USA*, **98,** 12315-12316.

54. Le, Q.H., Turcotte, K. and Bureau, T. (2001) Tc8, a Tourist-like transposon in *Caenorhabditis elegans*. *Genetics*, **158,** 1081-1088.

55. Zhang, X., Feschotte, C., Zhang, Q., Jiang, N., Eggleston, W.B. and Wessler, S.R. (2001) P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc Natl Acad Sci USA*, **98,** 12572-12577.

56. Yang, G. and Hall, T.C. (2003) MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Res*. (accepted).

57. Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G. and Chovnick, A. (1990) Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics*, **124,** 339-355.

58. Capy, P., Anxolabehere, D. and Langin, T. (1994) The strange phylogenies of transposable elements: are horizontal transfers the only explantation? *Trends Genet*, **10,** 7-12.

59. Bennetzen, J.L. (1996) The Mutator transposable element system of maize. *Curr Top Microbiol Immunol*, **204,** 195-229.

60. Walbot, V. (1991) The Mutator transposable element family of maize. *Genet Eng*, **13,** 1-37.

61. Gierl, A. (1996) The En/Spm transposable element of maize. *Curr Top Microbiol Immunol*, **204,** 145-159.

62. Gierl, A. and Saedler, H. (1989) The En/Spm transposable element of *Zea mays*. *Plant Mol Biol*, **13,** 261-266.

63. Masson, P., Banks, J.A. and Fedoroff, N. (1991) Structure and function of the maize Spm transposable element. *Biochimie*, **73,** 5-8.

64. Pereira, A., Cuypers, H., Gierl, A., Schwarz-Sommer, Z.S. and Saedler, H. (1986) Molecular analysis of the En/Spm transposable element system of *Zea mays*. *EMBO J*,**,** 835-841.

65. Christensen, S., Pont-Kingdon, G. and Carroll, D. (2000) Target specificity of the endonuclease from the *Xenopus laevis* non-long terminal repeat retrotransposon, Tx1L. *Mol Cell Biol*, **20,** 1219-1226.

66. Christensen, S., Pont-Kingdon, G. and Carroll, D. (2000) Comparative studies of the endonucleases from two related *Xenopus laevis* retrotransposons, Tx1L and Tx2L: target site specificity and evolutionary implications. *Genetica*, **110,** 245-256.

67. Garrett, J.E., Knutzon, D.S. and Carroll, D. (1989) Composite transposable elements in the *Xenopus laevis* genome. *Mol Cell Biol*, **9,** 3018-3027.

68. Weiner, A.M. (2002) SINEs and LINEs: the art of biting the hand that feeds you. *Curr Opin Cell Biol*, **14,** 343-350.

69. Kumar, A. and Bennetzen, J.L. (1999) Plant retrotransposons. *Annu Rev Genet*, **33,** 479-532.

70. Noma, K., Ohtsubo, E. and Ohtsubo, H. (1999) Non-LTR retrotransposons (LINEs) as ubiquitous components of plant genomes. *Mol Gen Genet*, **261,** 71-79.

71. Malik, H.S., Burke, W.D. and Eickbush, T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol*, **16,** 793-805.

72. Luan, D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72,** 595-605.

73. Saedler, H. and Nevers, P. (1985) Transposition in plants: a molecular model. *EMBO J*, **4,** 585-590.

74. Craig, N.L. (2002) *Mobile DNA II*, ASM Press, Washington DC.

75. Engels, W.R., Johnson-Schlitz, D.M., Eggleston, W.B. and Sved, J. (1990) High-frequency P element loss in *Drosophila* is homolog dependent. *Cell*, **62,** 515-525.

76. Doak, T.G., Doerder, F.P., Jahn, C.L. and Herrick, G. (1994) A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. *Proc Natl Acad Sci USA*, **91,** 942-946.

77. Capy, P. (1997) *Evolution and impact of transposable elements*, Contemporary issues in genetics and evolution, No.6, Kluwer Academic Publishers, Boston, MA.

78. Capy, P. (1998) *Dynamics and evolution of transposable elements*. Landes Bioscience, Austin, TX.

79. Becker, Y. (1995) A short introduction to the origin and molecular evolution of viruses. *Virus Genes*, **11,** 73-77.

80. Hurst, L.D. (1995) Selfish genetic elements and their role in evolution: the evolution of sex and some of what that entails. *Philos Trans R Soc Lond B Biol Sci*, **349,** 321-332.

81. Orgel, L.E. and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature*, **284,** 604-607.

82. Edgell, D.R., Fast, N.M. and Doolittle, W.F. (1996) Selfish DNA: the best defense is a good offense. *Curr. Biol.*, **6,** 385-388.

83. Dimitri, P. and Junakovic, N. (1999) Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Trends Genet*, **15,** 123-124.

84. McDonald, J.F. (1993) Evolution and consequences of transposable elements. *Curr Opin Genet Dev*, **3,** 855-864.

85. Nevers, P. and Saedler, H. (1977) Transposable genetic elements as agents of gene instability and chromosomal rearrangements. *Nature*, **268,** 109-115.

86. McClintock, B. (1984) The significance of responses of the genome to challenge. *Science*, **226,** 792-801.

87. Wilke, C.M., Maimer, E. and Adams, J. (1992) The population biology and evolutionary significance of Ty elements in *Saccharomyces cerevisiae*. *Genetica*, **86,** 155-173.

88. Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol*, **42,** 251-269.

89. Fedoroff, N. (2000) Transposons and genome evolution in plants. *Proc. Natl. Acad. Sci. USA*, **97,** 7002-7007.

90. Finnegan, D.J. (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet*, **5,** 103-107.

91. Wright, S. and Finnegan, D. (2001) Genome evolution: sex and the transposable element. *Curr Biol*, **11,** R296-299.

92. Girard, L. and Freeling, M. (1999) Regulatory changes as a consequence of transposon insertion. *Dev Genet*, **25,** 291-296.

93. Errede, B., Company, M. and Hutchison, C.A., 3rd. (1987) Ty1 sequence with enhancer and mating-type-dependent regulatory activities. *Mol Cell Biol*, **7,** 258-265.

94. Roelants, F., Potier, S., Souciet, J.L. and de Montigny, J. (1997) Delta sequence of Ty1 transposon can initiate transcription of the distal part of the URA2 gene complex in *Saccharomyces cerevisiae*. *FEMS Microbiol Lett*, **148,** 69-74.

95. Herr, W. and Clarke, J. (1986) The SV40 enhancer is composed of multiple functional elements that can compensate for one another. *Cell*, **45,** 461-470.

96. Siliciano, P.G. and Tatchell, K. (1984) Transcription and regulatory signals at the mating type locus in yeast. *Cell*, **37,** 969-978.

97. Tanda, S. and Corces, V.G. (1991) Retrotransposon-induced overexpression of a homeobox gene causes defects in eye morphogenesis in *Drosophila*. *Embo J*, **10,** 407-417.

98. Bradley, D., Carpenter, R., Sommer, H., Hartley, N. and Coen, E. (1993) Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the plena locus of *Antirrhinum*. *Cell*, **72,** 85-95.

99. Corces, V.G. and Geyer, P.K. (1991) Interactions of retrotransposons with the host genome: the case of the gypsy element of *Drosophila*. *Trends Genet*, **7,** 86-90.

100. Dorsett, D. (1990) Potentiation of a polyadenylylation site by a downstream protein-DNA interaction. *Proc Natl Acad Sci USA*, **87,** 4373-4377.

101. Holdridge, C. and Dorsett, D. (1991) Repression of hsp70 heat shock gene transcription by the suppressor of hairy-wing protein of *Drosophila melanogaster*. *Mol Cell Biol*, **11,** 1894-1900.

102. Jack, J., Dorsett, D., Delotto, Y. and Liu, S. (1991) Expression of the cut locus in the *Drosophila* wing margin is required for cell type specification and is regulated by a distant enhancer. *Development*, **113,** 735-747.

103. Geyer, P.K. and Corces, V.G. (1992) DNA position-specific repression of transcription by a *Drosophila* zinc finger protein. *Genes Dev*, **6,** 1865-1873.

104. Spana, C., Harrison, D.A. and Corces, V.G. (1988) The *Drosophila melanogaster* suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon. *Genes Dev*, **2,** 1414-1423.

105. Mazo, A.M., Mizrokhi, L.J., Karavanov, A.A., Sedkov, Y.A., Krichevskaja, A.A. and Ilyin, Y.V. (1989) Suppression in *Drosophila*: su(Hw) and su(f) gene products interact with a region of gypsy (mdg4) regulating its transcriptional activity. *Embo J*, **8,** 903-911.

106. Greene, B., Walko, R. and Hake, S. (1994) Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations. *Genetics*, **138,** 1275-1285.

107. Kim, H.Y., Schiefelbein, J.W., Raboy, V., Furtek, D.B. and Nelson, O.E., Jr. (1987) RNA splicing permits expression of a maize gene with a defective Suppressor-

mutator transposable element insertion in an exon. *Proc Natl Acad Sci USA*, **84,** 5863-5867.

108. Raina, R., Cook, D. and Fedoroff, N. (1993) Maize Spm transposable element has an enhancer-insensitive promoter. *Proc Natl Acad Sci USA*, **90,** 6355-6359.

109. Barkan, A. and Martienssen, R.A. (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.

110. Chatterjee, M. and Martin, C. (1997) Tam3 produces a suppressible allele of the DAG locus of *Antirrhinum majus* similar to Mu-suppressible alleles of maize. *Plant J*, **11,** 759-771.

111. Michaud, E.J., van Vugt, M.J., Bultman, S.J., Sweet, H.O., Davisson, M.T. and Woychik, R.P. (1994) Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage. *Genes Dev*, **8,** 1463-1472.

112. McClintock, B. (1987) *The discovery and characterization of transposable elements: the collected papers of Barbara McClintock*. Genes, cells, and organisms, No.17, Garland Pub., New York.

113. Sutton, W.D., Gerlach, W., Schwartz, D. and Peacock, W. (1984) Molecular analysis of *Ds* controlling element mutations at the *Adh1* locus of maize. *Science*, **223,** 1265-1268.

114. Gorbunova, V., Ramos, C., Hohn, B. and Levy, A.A. (2000) A nuclear protein that binds specifically to several maize transposons is not essential for Ds1 excision. *Mol Gen Genet*, **263,** 492-497.

115. Shen, W.H., Ramos, C. and Hohn, B. (1998) Excision of Ds1 from the genome of maize streak virus in response to different transposase-encoding genes. *Plant Mol Biol*, **36,** 387-392.

116. Bravo-Angel, A.M., Becker, H.A., Kunze, R., Hohn, B. and Shen, W.H. (1995) The binding motifs for Ac transposase are absolutely required for excision of Ds1 in maize. *Mol Gen Genet*, **248,** 527-534.

117. Shen, W.H., Das, S. and Hohn, B. (1992) Mechanism of Ds1 excision from the genome of maize streak virus. *Mol Gen Genet*, **233,** 388-394.

118. Wessler, S.R. (1991) The maize transposable Ds1 element is alternatively spliced from exon sequences. *Mol Cell Biol*, **11,** 6192-6196.

119. Sullivan, T.D., Schiefelbein, J.W., Jr. and Nelson, O.E., Jr. (1989) Tissue-specific effects of maize bronze gene promoter mutations induced by Ds1 insertion and excision. *Dev Genet*, **10,** 412-424.

120. Dennis, E.S., Sachs, M.M., Gerlach, W.L., Beach, L. and Peacock, W.J. (1988) The Ds1 transposable element acts as an intron in the mutant allele Adh1-Fm335 and is spliced from the message. *Nucleic Acids Res*, **16,** 3815-3828.

121. Pisabarro, A.G., Martin, W.F., Peterson, P.A., Saedler, H. and Gierl, A. (1991) Molecular analysis of the Ubiquitous (Uq) transposable element system of *Zea mays*. *Mol Gen Genet*, **230,** 201-208.

122. Caldwell, E.E. and Peterson, P.A. (1992) The Ac and Uq transposable element systems in maize: interactions among components. *Genetics*, **131,** 723-731.

123. Bureau, T.E., Ronald, P.C. and Wessler, S.R. (1996) A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl. Acad. Sci. USA*, **93,** 8524-8529.

124. Tu, Z. (1997) Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. *Proc. Natl. Acad. Sci. USA*, **94,** 7475-7480.

125. Casacuberta, E., Casacuberta, J.M., Puigdomenech, P. and Monfort, A. (1998) Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the *Emigrant* family of elements. *Plant J.*, **16,** 79-85.

126. Braquart, C., Royer, V. and Bouhin, H. (1999) DEC: a new miniature inverted-repeat transposable element from the genome of the beetle *Tenebrio molitor*. *Insect. Mol. Biol.*, **8,** 571-574.

127. Charrier, B., Foucher, F., Kondorosi, E., d'Aubenton-Carafa, Y., Thermes, C., Kondorosi, A. and Ratet, P. (1999) Bigfoot. A new family of MITE elements characterized from the *Medicago genus*. *Plant J.*, **18,** 431-441.

128. Izsvak, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H. and Hackett, P.B. (1999) Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J. Mol. Evol.*, **48,** 13-21.

129. Surzycki, S.A. and Belknap, W.R. (1999) Characterization of repetitive DNA elements in *Arabidopsis*. *J Mol Evol*, **48,** 684-691.

130. Casa, A.M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S. and Wessler, S.R. (2000) Inaugural article: the MITE family *heartbreaker* (*Hbr*): molecular markers in maize. *Proc. Natl. Acad. Sci. USA*, **97,** 10083-10089.

131. Feschotte, C. and Mouches, C. (2000) Recent amplification of miniature inverted-repeat transposable elements in the vector mosquito *Culex pipiens*: characterization of the *Mimo* family. *Gene*, **250,** 109-116.

132. Feschotte, C. and Mouches, C. (2000) Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. *Mol. Biol. Evol.*, **17,** 730-737.

133. Hikosaka, A., Yokouchi, E. and Kawahara, A. (2000) Extensive amplification and transposition of a novel repetitive element, *xstir*, together with its terminal inverted repeat in the evolution of *Xenopus*. *J Mol Evol*, **51,** 554-564.

134. Hu, J., Reddy, V.S. and Wessler, S.R. (2000) The rice R gene family: two distinct subfamilies containing several miniature inverted-repeat transposable elements. *Plant Mol. Biol.*, **42,** 667-678.

135. Lepetit, D., Pasquet, S., Olive, M., Theze, N. and Thiebaud, P. (2000) *Glider* and *Vision*: two new families of miniature inverted-repeat transposable elements in *Xenopus laevis* genome. *Genetica*, **108,** 163-169.

136. Miller, W.J., Nagel, A., Bachmann, J. and Bachmann, L. (2000) Evolutionary dynamics of the *SGM* transposon family in the *Drosophila obscura* species group. *Mol. Biol. Evol.*, **17,** 1597-1609.

137. Akagi, H., Yokozeki, Y., Inagaki, A., Mori, K. and Fujimura, T. (2001) Micron, a microsatellite-targeting transposable element in the rice genome. *Mol Genet Genomics*, **266,** 471-480.

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

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

140. El Amrani, A., Marie, L., Ainouche, A., Nicolas, J. and Couee, I. (2002) Genome-wide distribution and potential regulatory functions of *AtATE*, a novel family of miniature inverted-repeat transposable elements in *Arabidopsis thaliana*. *Mol. Genet. Genomics*, **267,** 459-471.

141. Santiago, N., Herraiz, C., Goni, J.R., Messeguer, X. and Casacuberta, J.M. (2002) Genome-wide analysis of the emigrant family of MITEs of *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **19,** 2285-2293.

142. Yang, G., Dong, J., Chandrasekharan, M.B. and Hall, T.C. (2001) *Kiddo*, a new transposable element family closely associated with rice genes. *Mol. Gen. Genomics*, **266,** 417-424.

143. Yang, G. and Hall, T.C. (2003) *MDM*-1 and *MDM*-2, two *Mutator*-derived MITE families in rice. *J. Mol. Evol.*, **56,** 255-264.

144. Bureau, T.E. and Wessler, S.R. (1992) *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell*, **4,** 1283-1294.

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

146. Bureau, T.E. and Wessler, S.R. (1994) *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*, **6,** 907-916.

147. Tu, Z. (1999) Genomic and evolutionary analysis of *Feilai*, a diverse family of highly reiterated SINEs in the yellow fever mosquito, *Aedes aegypti*. *Mol. Biol. Evol.*, **16,** 760-772.

148. Tu, Z. (2000) Molecular and evolutionary analysis of two divergent subfamilies of a novel miniature inverted repeat transposable element in the yellow fever mosquito, *Aedes aegypti*. *Mol. Biol. Evol.*, **17,** 1313-1325.

149. Tu, Z. (2001) Maque, a family of extremely short interspersed repetitive elements: characterization, possible mechanism of transposition, and evolutionary implications. *Gene*, **263,** 247-253.

150. Tu, Z. and Orphanidis, S.P. (2001) Microuli, a family of miniature subterminal inverted-repeat transposable elements (MSITEs): transposition without terminal inverted repeats. *Mol Biol Evol*, **18,** 893-895.

151. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215,** 403-410.

152. Finnegan, D.J. (1992) Transposable elements. *Curr Opin Genet Dev*, **2,** 861-867.

153. Pozueta-Romero, J., Houlne, G. and Schantz, R. (1996) Nonautonomous inverted repeat *Alien* transposable elements are associated with genes of both monocotyledonous and dicotyledonous plants. *Gene*, **171,** 147-153.

154. Mao, L., Wood, T.C., Yu, Y., Budiman, M.A., Tomkins, J., Woo, S., Sasinowski, M., Presting, G., Frisch, D., Goff, S., Dean, R.A. and Wing, R.A. (2000) Rice transposable elements: a survey of 73,000 sequence-tagged- connectors. *Genome Res.*, **10,** 982-990.

155. Wessler, S.R., Bureau, T.E. and White, S.E. (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.*, **5,** 814-821.

156. Oosumi, T., Garlick, B. and Belknap, W.R. (1995) Identification and characterization of putative transposable DNA elements in solanaceous plants and *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA*, **92,** 8886-8890.

157. Taylor, B. and Powell, A. (1982) Isolation of plant DNA and RNA. *Focus*, **4,** 4-6.

158. Buchholz, W.G., Connell, J.P., Kumpatla, S.P. and Hall, T.C. (1998) Molecular analysis of transgenic rice. *Methods Mol. Biol.*, **81,** 397-415.

159. Wang, J., Jiang, J. and Oard, J.H. (2000) Structure, expression and promoter activity of two polyubiquitin genes from rice (*Oryza sativa* L.). *Plant Science*, **156,** 201-211.

160. Wadkins, R.M. (2000) Targeting DNA secondary structures. *Curr. Med. Chem.*, **7,** 1-15.

161. Jiang, J., Gill, B.S., Wang, G.L., Ronald, P.C. and Ward, D.C. (1995) Metaphase and interphase fluorescence in situ hybridization mapping of the rice genome with bacterial artificial chromosomes. *Proc. Natl. Acad. Sci. USA*, **92,** 4487-4491.

162. Huke, R.E. and Huke, E.H. (1990) *Rice: then & now*, International Rice Research Institute, Manila, Philippines.

163. Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A.H. (2000) From the cover: 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.

164. Unsal, K. and Morgan, G.T. (1995) A novel group of families of short interspersed repetitive elements (SINEs) in *Xenopus*: evidence of a specific target site for DNA-mediated transposition of inverted-repeat SINEs. *J Mol Biol*, **248,** 812-823.

165. Yeadon, P.J. and Catcheside, D.E. (1995) Guest: a 98 bp inverted repeat transposable element in *Neurospora crassa*. *Mol. Gen. Genet.*, **247,** 105-109.

166. Morgan, G.T. (1995) Identification in the human genome of mobile elements spread by DNA- mediated transposition. *J. Mol. Biol.*, **254,** 1-5.

167. Becker, H.A. and Kunze, R. (1997) Maize *Activator* transposase has a bipartite DNA binding domain that recognizes subterminal sequences and the terminal inverted repeats. *Mol. Gen. Genet.*, **254,** 219-230.

168. Oosumi, T., Garlick, B. and Belknap, W.R. (1996) Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. *J Mol Evol*, **43,** 11-18.

169. Surzycki, S.A. and Belknap, W.R. (2000) Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc Natl Acad Sci USA*, **97,** 245-249.

170. Izsvak, Z., Ivics, Z. and Hackett, P.B. (1997) Repetitive elements and their genetic applications in zebrafish. *Biochem Cell Biol*, **75,** 507-523.

171. Simmen, M.W. and Bird, A. (2000) Sequence analysis of transposable elements in the sea squirt, *Ciona intestinalis*. *Mol. Biol. Evol.*, **17,** 1685-1694.

172. Smit, A.F. and Riggs, A.D. (1996) Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci USA*, **93,** 1443-1448.

173. Robertson, H.M. (1996) Members of the pogo superfamily of DNA-mediated transposons in the human genome. *Mol Gen Genet*, **252,** 761-766.

174. Le, Q.H., Wright, S., Yu, Z. and Bureau, T. (2000) Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA*, **97,** 7376-7381.

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

176. Wessler, S.R., Nagel, A. and Casa, A. (2001) Miniature inverted repeat transposable elements help create genomic diversity in maize and rice, In Khush, G. S., Brar, D. S., and Hardy, B. (eds.), *Rice Genetics IV*. Science Publishers, Inc., Manila, Philippines, pp. 107-116.

177. Chandler, V.L. and Hardeman, K.J. (1992) The Mu elements of *Zea mays*. *Adv Genet*, **30,** 77-122.

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

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

180. Bonas, U., Sommer, H. and Saedler, H. (1984) The 17-kb *Tam1* element of *Antirrhinum majus* induces a 3-bp duplication upon integration into the chalcone synthase gene. *EMBO J*, **3,** 1015-1019.

181. Nacken, W.K., Piotrowiak, R., Saedler, H. and Sommer, H. (1991) The transposable element Tam1 from *Antirrhinum majus* shows structural homology to the maize transposon En/Spm and has no sequence specificity of insertion. *Mol Gen Genet*, **228,** 201-208.

182. Bennetzen, J.L., Springer, P.S., Cresse, A.D. and Hendrickx, M. (1993) Specificity and regulatin of the *Mutator* transposable element system of maize. *Crit Rev Plant Sci*, **12,** 57-95.

183. Bennetzen, J.L. (1996) The Mutator transposable element system of maize, In Saedler, H., and Gierl, A. (eds.), *Transposable elements*. Springer, Berlin, pp. 195-229.

184. SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274,** 765-768.

185. Shirasu, K., Schulman, A.H., Lahaye, T. and Schulze-Lefert, P. (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.*, **10,** 908-915.

186. Hagemann, S., Miller, W.J., Haring, E. and Pinsker, W. (1998) Nested insertions of short mobile sequences in *Drosophila* P elements. *Chromosoma*, **107,** 6-16.

187. Jiang, N. and Wessler, S.R. (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.

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

189. Lindroth, A.M., Cao, X., Jackson, J.P., Zilberman, D., McCallum, C.M., Henikoff, S. and Jacobsen, S.E. (2001) Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science*, **292,** 2077-2080.

190. Steimer, A., Amedeo, P., Afsar, K., Fransz, P., Scheid, O.M. and Paszkowski, J. (2000) Endogenous targets of transcriptional gene silencing in *Arabidopsis*. *Plant Cell*, **12,** 1165-1178.

191. Jensen, S., Gassama, M.P. and Heidmann, T. (1999) Taming of transposable elements by homology-dependent gene silencing. *Nat Genet*, **21,** 209-212.

192. Song, W.Y., Pi, L.Y., Bureau, T.E. and Ronald, P.C. (1998) Identification and characterization of 14 transposon-like elements in the noncoding regions of members of the *Xa21* family of disease resistance genes in rice. *Mol. Gen. Genet.*, **258,** 449-456.

193. Iwamoto, M., Nagashima, H., Nagamine, T., Higo, H. and Higo, K. (1999) A *Tourist* element in the 5'-flanking region of the catalase gene *CatA* reveals evolutionary

relationships among *Oryza* species with various genome types. *Mol. Gen. Genet.*, **262,** 493-500.

194. Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L. and Avramova, Z. (1999) Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc. Natl. Acad. Sci. USA*, **96,** 7409-7414.

195. Elrouby, N. and Bureau, T.E. (2000) Molecular characterization of the Abp1 5'-flanking region in maize and the *Teosintes*. *Plant Physiol.*, **124,** 369-378.

196. Schwartz, R.L. and Christianson, T. (1997) *Learning Perl*, O'Reilly, Sebastopol, CA.

197. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D. and Birney, E. (2002) The bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12,** 1611-1618.

198. Gundavaram, S. (1996) *CGI Programming on the World Wide Web*, O'Reilly & Associates, Inc., Sebastopol, CA.

199. Rhodes, P. and Vodkin, L. (1985) Highly structured sequence homology between an insertion element and the gene inwhich it resides. *Proc. Natl. Acad. Sci. USA*, **82,** 493-497.

200. Rhodes, P. and Vodkin, L. (1988) Organization of the *Tgm* family of transposable elements in soybean. *Genetics*, **120,** 597-604.

201. Turcotte, K. and Bureau, T. (2002) Phylogenetic analysis reveals stowaway-like elements may represent a fourth family of the IS630-Tc1-mariner superfamily. *Genome*, **45,** 82-90.

202. von Sternberg, R.M., Novick, G.E., Gao, G.P. and Herrera, R.J. (1992) Genome canalization: the coevolution of transposable and interspersed repetitive elements with single copy DNA. *Genetica*, **86,** 215-246.

203. Bowen, N.J. and Jordan, I.K. (2002) Transposable elements and the evolution of eukaryotic complexity. *Curr Issues Mol Biol*, **4,** 65-76.

204. Capy, P., Gasperi, G., Biemont, C. and Bazin, C. (2000) Stress and transposable elements: co-evolution or useful parasites? *Heredity*, **85,** 101-106.

205. Wessler, S.R. (1998) Transposable elements and the evolution of gene expression. *Symp Soc Exp Biol*, **51,** 115-122.

206. Romero, D.A. and Klaenhammer, T.R. (1993) Transposable elements in Lactococci: a review. *J Dairy Sci*, **76,** 1-19.

207. Charlesworth, B. (1988) The maintenance of transposable elements in natural populations. *Basic Life Sci*, **47,** 189-212.

208. Lonnig, W.E. and Saedler, H. (1997) Plant transposons: contributors to evolution? *Gene*, **205,** 245-253.

209. Arnault, C. and Dufournel, I. (1994) Genome and stresses: reactions against aggressions, behavior of transposable elements. *Genetica*, **93,** 149-160.

210. Deragon, J.M. and Capy, P. (2000) Impact of transposable elements on the human genome. *Ann Med*, **32,** 264-273.

211. Bechtold, N. and Pelletier, G. (1998) In planta *Agrobacterium*-mediated transformation of adult *Arabidopsis thaliana* plants by vacuum infiltration. *Methods Mol Biol*, **82,** 259-266.

212. Hiei, Y., Ohta, S., Komari, T. and Kumashiro, T. (1994) Efficient transformation of rice (*Oryza sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. *Plant J*, **6,** 271-282.

213. Hiei, Y., Komari, T. and Kubo, T. (1997) Transformation of rice mediated by *Agrobacterium tumefaciens*. *Plant Mol Biol*, **35,** 205-218.

214. Dong, J., Kharb, P., Teng, W. and Hall, T.C. (2001) Characterization of rice transformed via an *Agrobacterium*-mediated inflorescence transformation. *Mol. Breeding*, **7,** 187-194.

215. Murashige, T. and Skoog, F. (1962) A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol. Plant*, **15,** 473-497.

216. Kumpatla, S.P. and Hall, T.C. (1998) Longevity of 5-azacytidine-mediated gene expression and re- establishment of silencing in transgenic rice. *Plant Mol Biol*, **38,** 1113-1122.

217. Christman, J.K. (2002) 5-azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene*, **21,** 5483-5495.

218. McInerney, J.M., Nawrocki, J.R. and Lowrey, C.H. (2000) Long-term silencing of retroviral vectors is resistant to reversal by trichostatin A and 5-azacytidine. *Gene Ther*, **7,** 653-663.

219. Di Ianni, M., Terenzi, A., Perruccio, K., Ciurnelli, R., Lucheroni, F., Benedetti, R., Martelli, M.F. and Tabilio, A. (1999) 5-azacytidine prevents transgene methylation *in vivo*. *Gene Ther*, **6,** 703-707.

220. Kumpatla, S.P., Teng, W., Buchholz, W.G. and Hall, T.C. (1997) Epigenetic transcriptional silencing and 5-azacytidine-mediated reactivation of a complex transgene in rice. *Plant Physiol*, **115,** 361-373.

221. Berg, D.E. and Howe, M.M. (1989) *Mobile DNA*, American Society for Microbiology, Washington DC.

222. Gierl, A., Saedler, H. and Peterson, P.A. (1989) Maize transposable elements. *Annu Rev Genet*, **23,** 71-85.

223. Grinsted, J. (1986) Evolution of transposable elements. *J Antimicrob Chemother*, **18 Suppl C,** 77-83.

224. Saedler, H. and Gierl, A. (1996) *Transposable elements*. Current Topics in Microbiology and Immunology, 204, Springer, New York.

225. Kunze, R., Saedler, H. and Lonnig, W.E. (1997) Plant transposable elements. *Adv. Botn. Res.*, **27,** 331-470.

226. Kunze, R. (1996) The maize transposable element activator (Ac). *Curr Top Microbiol Immunol*, **204,** 161-194.

227. Lampe, D.J., Grant, T.E. and Robertson, H.M. (1998) Factors affecting transposition of the Himar1 mariner transposon in vitro. *Genetics*, **149,** 179-187.

228. Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S.R., McCouch, S.R. and Wessler, S.R. (2003) An active DNA transposon family in rice. *Nature*, **421,** 163-167.

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

230. Nakazaki, T., Okumoto, Y., Horibata, A., Yamahira, S., Teraishi, M., Nishida, H., Inoue, H. and Tanisaka, T. (2003) Mobilization of a transposon in the rice genome. *Nature*, **421,** 170-172.

231. Haseloff, J. (1999) GFP variants for multispectral imaging of living cells. *Methods Cell Biol*, **58,** 139-151.

232. Bevan, M. (1984) Binary *Agrobacterium* vectors for plant transformation. *Nucleic Acids Res*, **12,** 8711-8721.

233. Hall, T.C., Kumpatla, S.P., Kharb, P., Iyer, L., Cervera, M., Jiang, Y., Wang, T., Yang, G., Teerawanichpan, P., Narangajavana, J. and Dong, J. (2001) Gene silencing and its reactivation in transgenic rice, In Khush, G. S., Brar, D. S., and Hardy, B. (eds.), *Rice Genetics IV.* Science Publishers, Inc., New Delhi (India), pp. 465-481.

234. Chu, C.C., Wang, C.C., Sun, C.S., Hsu, C., Ying, K.C., Chu, C.Y. and Bin, F.Y. (1975) Establishment of an efficient medium for another culture of rice through comparative experimentation on initrogen sources. *Scientia Sinica*, **18,** 659-668.

235. Salgueiro, S., Pignocchi, C. and Parry, M.A. (2000) Intron-mediated gusA expression in tritordeum and wheat resulting from particle bombardment. *Plant Mol Biol*, **42,** 615-622.

236. Jefferson, R.A., Kavanagh, T.A. and Bevan, M.W. (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J*, **6,** 3901-3907.

237. Jefferson, R.A. (1989) The GUS reporter gene system. *Nature*, **342,** 837-838.

APPENDIX A

>Emigrant-MathE2 (ab005244)

cagtaaaacctctataaattaataatgtcgggaccgaaaaaatttattaatttagagaggtattagtttatcgataaattaataattattaatttataaagagattttacaaattttgtaatttttacaaaaaatc
gatataaaaaatacttttttctttataatcaatatttttttctaaaactaattacaagcaaaaaacaattattaagtttaaaaaataatacaaagtttttcgacgtagtaaaaatattagaattaagctctacaa
ataattaatagaaattgaaataaataaatattagagtcatatttcagtaaatattttacatataatattgatatatatatacatacatatgtgtatatagtgataaatttacataattattaatttataatattgat
gggaccatatatttatataagattttcaaaaaaattattatcttattaatttatcgatttgtatcaatttttacactggtcccaactcgggaccggaaaaaattattaatttatagagattattaatttatcgagta
ttaatttatggaggttttactg

>mPIF (af416327)

gggcccgtttgtttcccttcattttgaggaattggaatttaattaatggagtaggctatttttagaatgtggcattccacaactttccaaagtgtatatataagtctatctcaaattcatggggtgagagatg
gaaattgattctatagatttacatgctacttttctaatgtacaatttataacacactcttctacttgcttctctataacataaatgtagtgtataactatctccctcttatggttttaggataatatacaaatataat
acatagacaaatatattaactaattagttttgtctaaattataattattagagtggaattcaattccaacgaaacaaacggggcc

>Tc8 (Z81454)

ggggttattcaagtaatgtagcaaaatgtatttaaatacatttgtgacgtcacaaatgtattaaaatacatgtttttatgtattttaatacagttgtgacgtaatttttctacacttttaattttccgacactactt
gaataacccc

>*MDM*2 (consensus)

gagataatccaagaaatgccattgacaagcgtccaagtcccagaaatgccatcgacaagtgtgagttccaagaaatgccatcgtacaaacgattttgtcccaaaaatgccatcgccgttagggttc
ctccattccgcgccgttaaatacactgttcatcctatagaacttaacggcgggaatggacggaaccctaacggcgatggcatttttgggacaaatcgtttgtacgatggcatttcttggaactcacact
tgtcgatggcatttctggatttggacgcttgtcaatggcatttctggattatctc

>MathE1 (ab010073)

gggggtgttattcgttaatggattttaatagaattgaaatccaatactaatccactgttattcaactaatgattctaaatccaactttaaaatctagtgttattggaactcatatttgtaaatgatgctgaatag
attttaaagtttggtgttattcaattaagattttaaatatttcattaaaatccaatgttattcaaaactaaaagacttgtaaaatctttatatattaattgatttgaaatgaaacctctttggagtttcatgagtaaat
cattagaatcaaaatccaaaacatattgcagagattttgaaattcatacaaacatattcttaaaaacaatatcagaaaacaaatctaaagcttaaagaattgactttaactccataattgaataacaccacc

>*Kiddo*A (concensus)

ggggctgtttggttcccagccatactttaccattacttgccaacaaaagttgccacaccttgtctaaggtgaggtgatcaaattgttagccacaacttactaagcctaagggaatcttgccacacttttt
gagccattgacacgtgggacttaatttgttagagggaaatcttgccacaactgtggctacaaccaaacacctgtcaaatttgcctaaccttaggcgtggcaaactgtggcaaagtgtggcttacaac
caaacacaccc

APPENDIX B

Anchor results for *Kiddo*A:

>AP005513, from 120586 to 126231, len=5645

Transposase: Putative tnp2 transposase [Oryza sativa (japonica cultivar-group)]

>AP004067, from 28393 to 34038, len=5645

Transposase: Putative tnp2 transposase [Oryza sativa (japonica cultivar-group)]

>AC136842, from 126407 to 130038, len=3631

Transposase: En/Spm-like transposon-like protein [Oryza sativa (japonica cultivar-group)] dbj|BAC10694.1| En/Spm-like transposon-like protein [Oryza sativa (japonica cultivar-group)]

Transposase: En/Spm-like transposon protein, putative; protein id: At1g61510.1 [Arabidopsis thaliana]

Transposase: hypothetical protein [Oryza sativa (japonica cultivar-group)] gb|AAM18154.1|AC092172_14 Putative En/Spm-like transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: contains similarity to En/Spm-like transposon protein~gene_id:MTO24.15 [Arabidopsis thaliana]

Transposase: transposase-like [Oryza sativa (japonica cultivar-group)]

Transposase: En/Spm-like transposon protein; protein id: At2g26630.1 [Arabidopsis thaliana] pir||T00996 En/Spm-like transposon protein [imported] - Arabidopsis thaliana gb|AAC14510.1| En/Spm-like transposon protein [Arabidopsis thaliana]

Transposase: similar to En/Spm-like transposon protein, putative; protein id: At1g43722.1 [Arabidopsis thaliana]

Transposase: contains similarity to En/Spm-like transposon protein~gene_id:MQP15.2 [Arabidopsis thaliana]

Transposase: putative transposase [Zea mays]

Transposase: Putative transposase [Oryza sativa (japonica cultivar-group)] gb|AAM47612.1|AC122147_1 Putative transposase [Oryza sativa (japonica cultivar-group)]

Transposase: similar to En/Spm-like transposon protein; protein id: At5g35695.1 [Arabidopsis thaliana]

Transposase: Putative transposase [Oryza sativa (japonica cultivar-group)]

Transposase: similar to En/Spm-like transposon protein; protein id: At4g04635.1 [Arabidopsis thaliana]

Transposase: Putative transposase [Oryza sativa]

>AC118347, from 17088 to 20719, len=3631

Transposase: En/Spm-like transposon-like protein [Oryza sativa (japonica cultivar-group)] dbj|BAC10694.1| En/Spm-like transposon-like protein [Oryza sativa (japonica cultivar-group)]

Transposase: En/Spm-like transposon protein, putative; protein id: At1g61510.1 [Arabidopsis thaliana]

Transposase: hypothetical protein [Oryza sativa (japonica cultivar-group)] gb|AAM18154.1|AC092172_14 Putative En/Spm-like transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: contains similarity to En/Spm-like transposon protein~gene_id:MTO24.15 [Arabidopsis thaliana]

Transposase: transposase-like [Oryza sativa (japonica cultivar-group)]

Transposase: En/Spm-like transposon protein; protein id: At2g26630.1 [Arabidopsis thaliana] pir||T00996 En/Spm-like transposon protein [imported] - Arabidopsis thaliana gb|AAC14510.1| En/Spm-like transposon protein [Arabidopsis thaliana]

Transposase: similar to En/Spm-like transposon protein, putative; protein id: At1g43722.1 [Arabidopsis thaliana]

Transposase: contains similarity to En/Spm-like transposon protein~gene_id:MQP15.2 [Arabidopsis thaliana]

Transposase: putative transposase [Zea mays]

Transposase: Putative transposase [Oryza sativa (japonica cultivar-group)] gb|AAM47612.1|AC122147_1 Putative transposase [Oryza sativa (japonica cultivar-group)]

Transposase: similar to En/Spm-like transposon protein; protein id: At5g35695.1 [Arabidopsis thaliana]

Transposase: Putative transposase [Oryza sativa (japonica cultivar-group)]

Transposase: similar to En/Spm-like transposon protein; protein id: At4g04635.1 [Arabidopsis thaliana]

Transposase: Putative transposase [Oryza sativa]

>AP005461, from 78912 to 82646, len=3734

Transposase: En/Spm-like transposon-like protein [Oryza sativa (japonica cultivar-group)] dbj|BAC10694.1| En/Spm-like transposon-like protein [Oryza sativa (japonica cultivar-group)]

Transposase: En/Spm-like transposon protein, putative; protein id: At1g61510.1 [Arabidopsis thaliana]

Transposase: hypothetical protein [Oryza sativa (japonica cultivar-group)] gb|AAM18154.1|AC092172_14 Putative En/Spm-like transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: transposase-like [Oryza sativa (japonica cultivar-group)]

Transposase: contains similarity to En/Spm-like transposon protein~gene_id:MTO24.15 [Arabidopsis thaliana]

Transposase: En/Spm-like transposon protein; protein id: At2g26630.1 [Arabidopsis thaliana] pir||T00996 En/Spm-like transposon protein [imported] - Arabidopsis thaliana gb|AAC14510.1| En/Spm-like transposon protein [Arabidopsis thaliana]

Transposase: similar to En/Spm-like transposon protein, putative; protein id: At1g43722.1 [Arabidopsis thaliana]

Transposase: contains similarity to En/Spm-like transposon protein~gene_id:MQP15.2 [Arabidopsis thaliana]

Transposase: putative transposase [Zea mays]

Transposase: Putative transposase [Oryza sativa (japonica cultivar-group)] gb|AAM47612.1|AC122147_1 Putative transposase [Oryza sativa (japonica cultivar-group)]

Transposase: similar to En/Spm-like transposon protein; protein id: At5g35695.1 [Arabidopsis thaliana]

Transposase: similar to En/Spm-like transposon protein; protein id: At4g04635.1 [Arabidopsis thaliana]

Transposase: Putative transposase [Oryza sativa]

Transposase: Putative transposase [Oryza sativa (japonica cultivar-group)]

>AP005461, from 78912 to 88775, len=9863

Transposase: Putative tnp2 transposase [Oryza sativa (japonica cultivar-group)]

Transposase: TNP2-like transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: putative transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: Putative TNP2 transposase [Oryza sativa]

Transposase: Putative transposase [Oryza sativa]

Transposase: Putative transposase [Oryza sativa] gb|AAN34959.1| Putative TNP2-like transposable element [Oryza sativa (japonica cultivar-group)]

Transposase: putative transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: putative transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: putative transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: Putative Tam1 transposon protein TNP2 [Oryza sativa]

Transposase: TNP2-like transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: Putative tnp2 transposase [Oryza sativa]

Transposase: Putative TNP2 transposon [Oryza sativa (japonica cultivar-group)]

Transposase: Putative transposase [Oryza sativa]

Transposase: Putative transposase [Oryza sativa (japonica cultivar-group)]

Transposase: similar to transposase [Oryza sativa (japonica cultivar-group)]

Transposase: putative transposon protein [Oryza sativa (japonica cultivar-group)]

Transposase: putative transposon-related TNP2 protein [Oryza sativa (japonica cultivar-group)]

Transposase: Putative transposase [Oryza sativa (japonica cultivar-group)] gb|AAM47620.1|AC122147_9 Putative transposase protein [Oryza sativa (japonica cultivar-group)]

Transposase: putative transposon protein [Oryza sativa (japonica cultivar-group)] gb|AAN16333.1| TNP2-like protein [Oryza sativa (japonica cultivar-group)]

Transposase: Putative TNP2 transposase [Oryza sativa]

Transposase: putative transposon [Oryza sativa (japonica cultivar-group)]

Anchor results for MathE1:

>AC007123, from 2690 to 6918, len=4228 query=MathE1(ab010073)

Transposase: (NM_148036) similar to En/Spm-like transposon protein; protein id:

Transposase: (AP005486) transposase-like [Oryza sativa (japonica cultivar-group)]

Transposase: (AC079852) Putative transposase [Oryza sativa]

Transposase: (AC025098) Putative transposase [Oryza sativa (japonica

Transposase: (NM_104832) En/Spm-like transposon protein, putative; protein id:

Transposase: (AC092553) Putative transposase [Oryza sativa (japonica

Transposase: (NM_128220) En/Spm-like transposon protein; protein id: At2g26630.1

Transposase: (AP000606) contains similarity to En/Spm-like transposon

Transposase: (AF412282) putative transposase [Zea mays]

Transposase: (NM_148535) similar to En/Spm-like transposon protein, putative;

Transposase: (AB016878) contains similarity to En/Spm-like transposon

Transposase: (AP003450) En/Spm-like transposon-like protein [Oryza sativa

Transposase: (NM_148229) similar to En/Spm-like transposon protein; protein id:

>AF007271, from 16996 to 21224, len=4228 query=MathE1(ab010073)

Transposase: (NM_148036) similar to En/Spm-like transposon protein; protein id:

Transposase: (AP005486) transposase-like [Oryza sativa (japonica cultivar-group)]

Transposase: (AC079852) Putative transposase [Oryza sativa]

Transposase: (AC025098) Putative transposase [Oryza sativa (japonica

Transposase: (NM_104832) En/Spm-like transposon protein, putative; protein id:

Transposase: (AC092553) Putative transposase [Oryza sativa (japonica

Transposase: (NM_128220) En/Spm-like transposon protein; protein id: At2g26630.1

Transposase: (AP000606) contains similarity to En/Spm-like transposon

Transposase: (AF412282) putative transposase [Zea mays]

Transposase: (NM_148535) similar to En/Spm-like transposon protein, putative;

Transposase: (AB016878) contains similarity to En/Spm-like transposon

Transposase: (AP003450) En/Spm-like transposon-like protein [Oryza sativa

Transposase: (NM_148229) similar to En/Spm-like transposon protein; protein id:

>AC074227, from 19780 to 29476, len=9696 query=MathE1(ab010073)

Transposase: (NM_148805) similar to putative transposon protein; protein id:

Transposase: (NM_116652) putative transposon protein; protein id: At4g04140.1

Transposase: (NM_103194) mutator-like transposase, putative; protein id:

Transposase: (AP002029) mutator-like transposase [Arabidopsis thaliana]

Transposase: (AC079280) mutator-like transposase, putative [Arabidopsis thaliana]

Transposase: (NM_127043) Mutator-like transposase; protein id: At2g14790.1

Transposase: (NM_148274) similar to mutator-like transposase, putative; protein

Transposase: (NM_102371) mutator-like transposase, putative; protein id:

Transposase: (NM_123031) similar to mutator-like transposase, putative; protein

Transposase: (AB025605) mutator-like transposase [Arabidopsis thaliana]

Transposase: (NM_147932) similar to mutator-like transposase, putative; protein

Transposase: (NM_116958) putative transposon protein; protein id: At4g08890.1

Transposase: (NM_126844) Mutator-like transposase; protein id: At2g11210.1

Transposase: (NM_103191) mutator-like transposase, putative; protein id:

Transposase: (AC018460) Similar to mutator transposase [Arabidopsis thaliana]

Transposase: (NM_126572) Mutator-like transposase; protein id: At2g05490.1

Transposase: (NM_126876) Mutator-like transposase; protein id: At2g12150.1

Transposase: (AC079280) mutator-like transposase, putative [Arabidopsis thaliana]

Transposase: (AB018112) mutator-like transposase [Arabidopsis thaliana]

Transposase: (AF177535) contains similarity to maize transposon MuDR (GB:M76978)

Transposase: (NM_126701) Mutator-like transposase; protein id: At2g07320.1

Transposase: (NM_147298) similar to mutator-like transposase, putative; protein

Transposase: (AB017065) contains similarity to En/Spm-like transposon

Transposase: (NM_148291) similar to Mutator-like transposase; protein id:

Transposase: (AB006701) contains similarity to En/Spm-like transposon

>AC008238, from 12252 to 22211, len=9959 query=MathE1(ab010073)

Transposase: monosaccharide-carrier {clone CST1} [Chenopodium rubrum=fat hen,

Transposase: monosaccharide-carrier {clone CST4} [Chenopodium rubrum=fat hen,

>AL109737, from 25285 to 32651, len=7366 query=MathE1(ab010073)

Transposase: (NM_116379) putative transposon protein; protein id: At4g01490.1

>AC090030, from 14420 to 19718, len=5298 query=MathE1(ab010073)

Transposase: (AP000364) Similar to Transposon MAGGY gag and pol gene homologues.

Transposase: (AC080019) Similar to Transposon MAGGYgagandpolgenehomologues [Oryza

Anchor results for MathE2:

>AC006161, from 85200 to 87313, len=2113 query=MathE2(ab005244)

Transposase: hypothetical protein - fruit fly (Drosophila melanogaster) transposon

Transposase: (U49973) ORF1; MER37; putative transposase similar to pogo element

Transposase: probable transposase - human transposon MER37

Transposase: hypothetical protein Tigger 2 - human transposon MER37 (fragment)

Transposase: (AF205929) transposase [Candida albicans]

Anchor results for *MDM*2:

>AP004320, from 26330 to 31778, len=5448 query=*MDM*2(consensus)

Transposase: (AC084884) Putative mudrA protein - maize transposon MuDR [Oryza

Transposase: (AC113948) putative transposon protein [Oryza sativa (japonica

Transposase: mudrA protein - maize transposon MuDR

Transposase: (NM_111332) Mutator-like transposase; protein id: At3g04605.1,

Transposase: (AP003333) putative mutator-like transposase [Oryza sativa (japonica

Transposase: (AC078839) mudrA protein - maize transposon MuDR [Oryza sativa]

Transposase: (AC084404) putative transposon protein [Oryza sativa]

Transposase: (NM_103520) mutator-like transposase, putative; protein id:

Transposase: (AC037197) Putative mutator-like transposase [Oryza sativa]

Transposase: (AC084831) putative transposon protein [Oryza sativa]

Transposase: (AC099774) putative transposase related protein [Oryza sativa

Transposase: (AC093180) Putative mutator-like transposase [Oryza sativa (japonica

Transposase: (AB017061) mutator-like transposase-like protein [Arabidopsis

Transposase: (AF177535) contains similarity to maize transposon MuDR (GB:M76978)

Transposase: (AP003749) putative mutator-like transposase [Oryza sativa (japonica

Transposase: (AB018112) mutator-like transposase [Arabidopsis thaliana]

Transposase: (NM_106459) putative Mutator-like transposase; protein id:

Transposase: (NM_116937) putative MuDR-A-like transposon protein; protein id:

Transposase: (AC084218) similar to Oryza sativa Mutator-like transposase

Transposase: (AC074105) Putative transposon protein [Oryza sativa]

Transposase: (AC018460) Similar to mutator transposase [Arabidopsis thaliana]

Transposase: (NM_128616) putative Mutator-like transposase; protein id:

Transposase: (NM_103071) mutator transposase MUDRA, putative; protein id:

Transposase: (NM_148022) similar to Mutator-like transposase; protein id:

Transposase: (NM_101142) mutator-like transposase, putative; protein id:

Transposase: (NM_126693) Mutator-like transposase; protein id: At2g07230.1

Transposase: (NM_148291) similar to Mutator-like transposase; protein id:

Transposase: (AF177535) contains similarity to maize transposon MuDR (GB:M76978)

Transposase: (AC069324) Putative maize transposon MuDR mudrA-like protein [Oryza

Transposase: (AB022213) mutator-like transposase [Arabidopsis thaliana]

Transposase: (NM_126876) Mutator-like transposase; protein id: At2g12150.1

Transposase: (NM_102371) mutator-like transposase, putative; protein id:

Transposase: (NM_128479) Mutator-like transposase; protein id: At2g29230.1

Transposase: (AB023031) mutator-like transposase-like [Arabidopsis thaliana]

Transposase: (AC002342) putative Mutator-like transposase, 3' partial [Arabidopsis

Transposase: (NM_123031) similar to mutator-like transposase, putative; protein

Transposase: (AB025605) mutator-like transposase [Arabidopsis thaliana]

Transposase: (AP002029) mutator-like transposase [Arabidopsis thaliana]

Transposase: (NM_103305) mutator-like transposase, putative; protein id:

Transposase: (NM_101589) putative mutator-like transposon protein; protein id:

Transposase: (NM_127914) Mutator-like transposase; protein id: At2g23500.1

Transposase: (AP003315) putative mutator-like transposase [Oryza sativa (japonica

Transposase: (AC006216) Similar to gi|3047071 F7N22.10 maize transposon MuDR

Transposase: (AP003220) putative mutator-like transposase [Oryza sativa (japonica

Transposase: (AC078839) Mutator-like transposase [Oryza sativa]

Transposase: (AP003416) putative mutator-like transposase [Oryza sativa (japonica

Transposase: (AC079037) Putative mutator-like transposase [Oryza sativa]

Transposase: (AC084295) putative transposase related protein [Oryza sativa]

Transposase: (AP003273) putative mutator-like transposase [Oryza sativa (japonica

Transposase: (AP003849) similar to mutator-like transposase [Oryza sativa

Transposase: (AC068924) mutator-like transposase [Oryza sativa (japonica

Transposase: (AC091122) mutator-like transposase, 3'-partial [Oryza sativa

Transposase: (NM_126533) Mutator-like transposase; protein id: At2g05010.1

Transposase: (AC018929) mutator-like transposase [Oryza sativa]

Transposase: (AC091774) putative transposon protein [Oryza sativa]

Transposase: (NM_103191) mutator-like transposase, putative; protein id:

Transposase: (AC090485) Putative mutator-like transposase [Oryza sativa] [Oryza

Transposase: (AJ238507) transposase related protein [Zea mays]

Transposase: (NM_126975) Mutator-like transposase; protein id: At2g14030.1

Transposase: (AP000366) Similar to maize transposon MuDR mudrA-like protein.

Transposase: (AC027038) putative transposase [Oryza sativa (japonica

Transposase: (NM_103308) mutator-like transposase, putative; protein id:

Transposase: (AP004194) putative mutator-like transposase [Oryza sativa (japonica

Transposase: (NM_126572) Mutator-like transposase; protein id: At2g05490.1

Transposase: (NM_113960) putative transposase related protein; protein id:

Transposase: (NM_127043) Mutator-like transposase; protein id: At2g14790.1

Transposase: (AC027656) Strong similarity to a mutator-like transposase from

Transposase: (NM_126898) Mutator-like transposase; protein id: At2g12720.1

Transposase: (NM_113946) Mutator-like transposase; protein id: At3g30455.1

Transposase: (NM_127077) Mutator-like transposase; protein id: At2g15150.1

Transposase: (NM_114204) putative transposase; protein id: At3g43360.1

Anchor results for m*PIF*:

>AF412282, from 1 to 3725, len=3724 query=m*PIF*(af416327)

Transposase: (AF412282) putative transposase [Zea mays]

Transposase: (AP005486) transposase-like [Oryza sativa (japonica cultivar-group)]

Transposase: (AC025098) Putative transposase [Oryza sativa (japonica

Transposase: (NM_104832) En/Spm-like transposon protein, putative; protein id:

Transposase: (AC079852) Putative transposase [Oryza sativa]

Transposase: (AP003450) En/Spm-like transposon-like protein [Oryza sativa

Transposase: (NM_148535) similar to En/Spm-like transposon protein, putative;

Transposase: (AC092553) Putative transposase [Oryza sativa (japonica

Transposase: (AP000606) contains similarity to En/Spm-like transposon

Transposase: (NM_128220) En/Spm-like transposon protein; protein id: At2g26630.1

Transposase: (AB016878) contains similarity to En/Spm-like transposon

Transposase: (NM_148036) similar to En/Spm-like transposon protein; protein id:

Transposase: (NM_148229) similar to En/Spm-like transposon protein; protein id:

>AC114395, from 45020 to 53618, len=8598 query=m*PIF*(af416327)

Transposase: (AP000364) Similar to Transposon MAGGY gag and pol gene homologues.

Transposase: (AC080019) Similar to Transposon MAGGYgagandpolgenehomologues [Oryza

Anchor results for Tc8:

>AF040643, from 24108 to 31614, len=7506 query=Tc8(2815028)

Transposase: (AP000606) contains similarity to En/Spm-like transposon

Transposase: (AC092553) Putative transposase [Oryza sativa (japonica

Transposase: (AP005486) transposase-like [Oryza sativa (japonica cultivar-group)]

Transposase: (AC079852) Putative transposase [Oryza sativa]

APPENDIX C

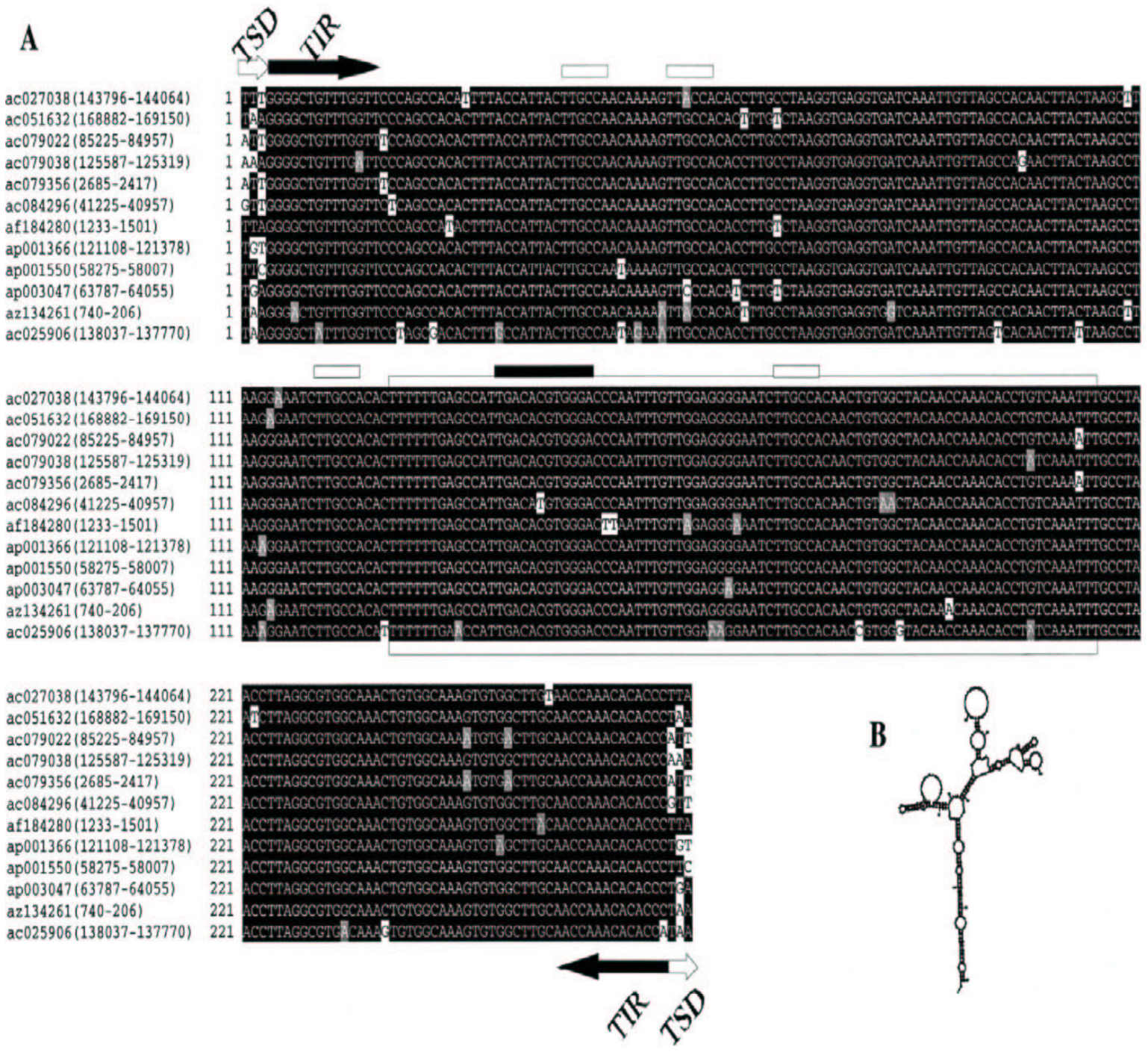

Figure 2.2.

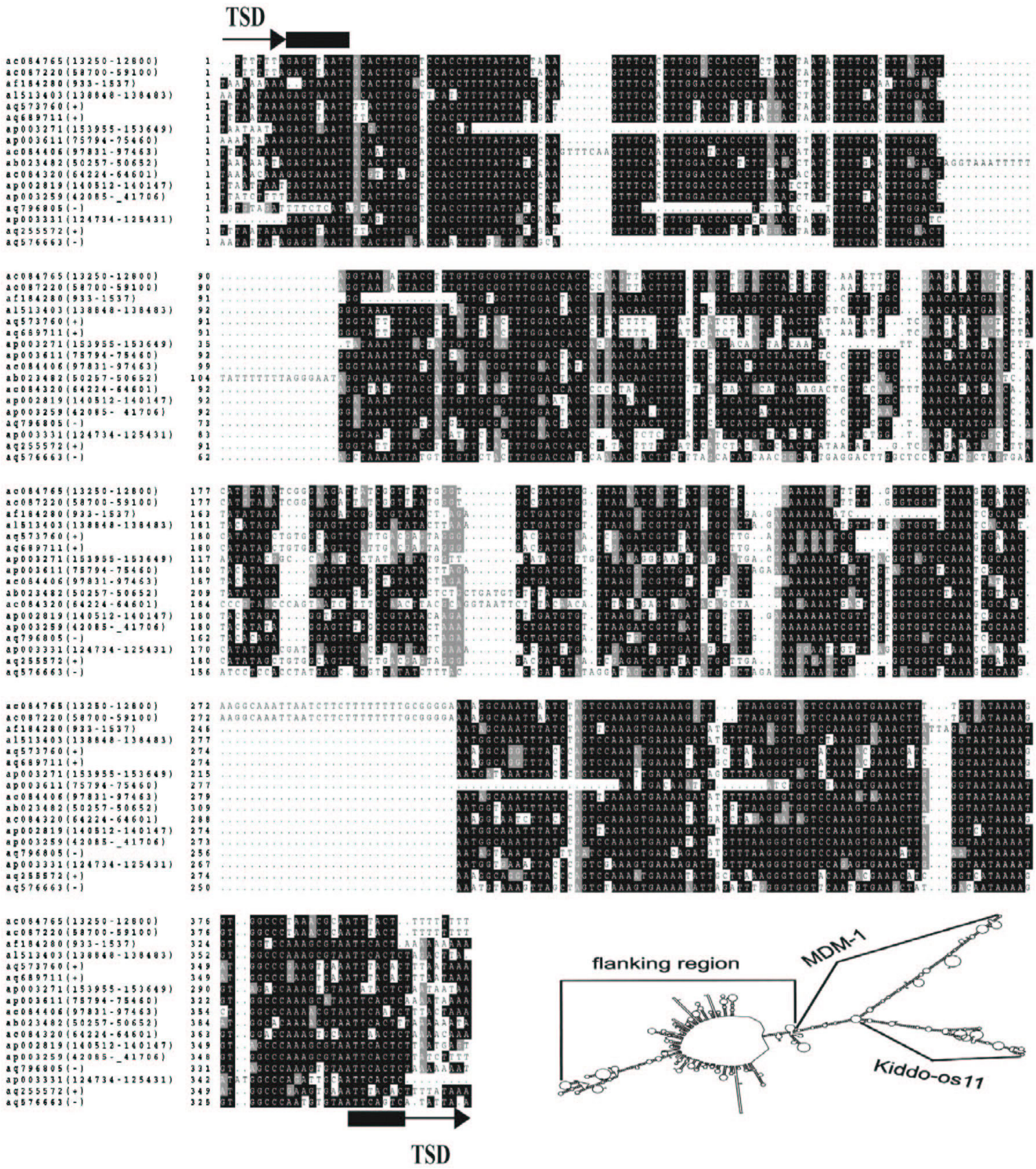

Figure 3.1.

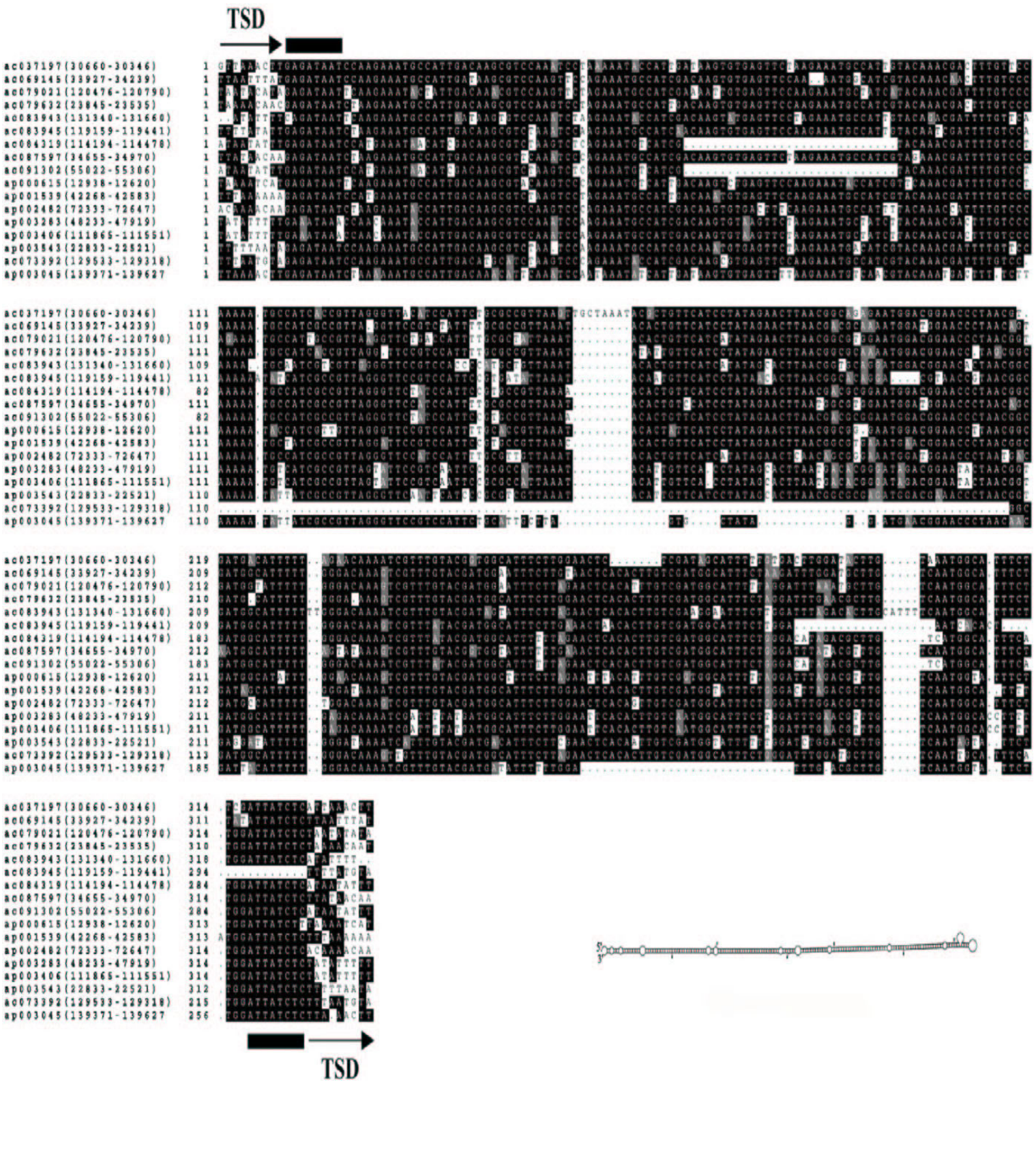

Figure 3.2.

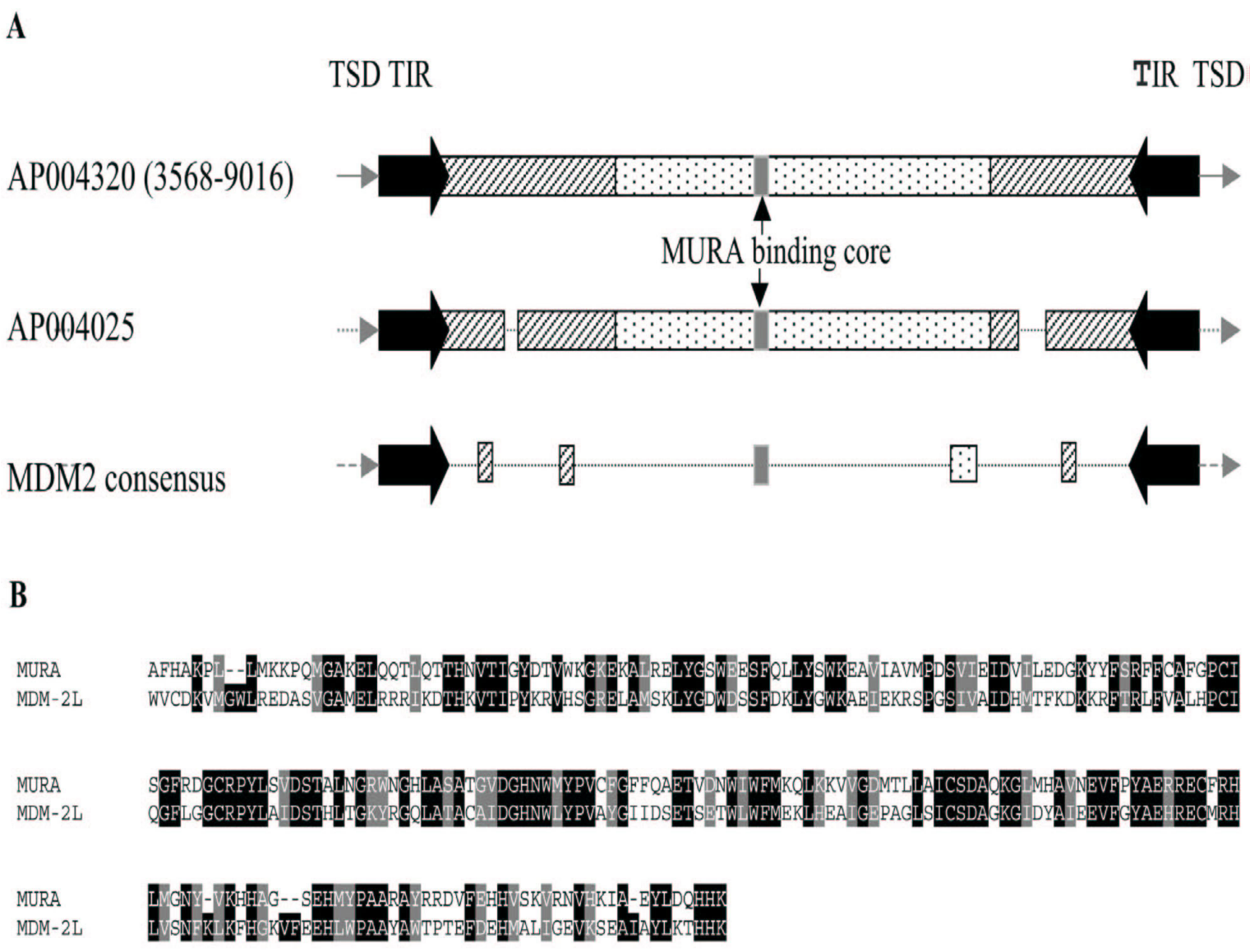
A
TSD TIR
TIR TSD
AP004320 (3568-9016)
MURA binding core
AP004025
MDM2 consensus
B
MURA AFHAKPL--LMKKPQMGAKELQQTLQTTHNVTIGYDTVWKGKEKALRELYGSWEESFQLLYSWKEAVIAVMPDSVIEIDVILEDGKYYFSRFFCAFGPCI
MDM-2L WVCDKVMGWLREDASVGAMELRRRIKDTHKVTIPYKRVHSGRELAMSKLYGDWDSSFDKLYGWKAEIEKRSPGSIVAIDHMTFKDKKRFTRLFVALHPCI
MURA SGFRDGCRPYLSVDSTALNGRWNGHLASAIGVDGHNWMYPVCFGFFQAETVDNWIWFMKQLKKVVGDMTLLAICSDAQKGLMHAVNEVFPYAERRECFRH
MDM-2L QGFLGGCRPYLAIDSTHLTGKYRGQLATACAIDGHNWLYPVAYGIIDSETSETWLWFMEKLHEAIGEPAGLSICSDAGKGIDYAIEEVFGYAEHRECMRH
MURA LMGNY-VKHHAG--SEHMYPAARAYRRDVFEHHVSKVRNVHKIA-EYLDQHHK
MDM-2L LVSNFKLKFHCKVFEEHLWPAAYAWTPTEFDEHMALIGEVKSEAIAYLKTHHK

Figure 3.4.

A.

A-MathE1

A-Kiddo

B.

```
A-MathE1  RMNPEVFAELCHLLQMKTGLKGTPHVCVEEMVATFLITVGQNSRYCHTMD
PIFa      RLTKRSFSDLCTILRERCDMCDTLNVSVEEKVAIFLLVVGHGTKMRMIRS

A-MathE1  TFKRSKFSTSINFHKVLR.ALDLQDALTRRTNRLQVPEGNRLQVLKGDLS
PIFa      SYGWSLEPISRYFNEVLRG....VD........CLGALDGTHIDVFVPLAD

A-MathE1  YASYRNRKGVISQNVLAACNFDLEFIYVLSGWEGSAHDSKVLR.......
PIFa      QGRYRNRKQQITTNVLGVCDRHMKFVYVLAGWEGSASDSRVLRDAMSRDD

A-MathE1  ............................STRYHLQDFRGEGRDPTNQNELF
PIFa      AFAIPSGKYYLVDAGYTNGPGFLAPYRSTRYHLNEWAAQGNNPSNAKELF

A-MathE1  NLRHASLRNVI
PIFa      NLRHSTARNYV

A-Kiddo   ASIMGEMANLYTER....YLQKGAYRQTPETGIQWVMRLMDRPR..YFYK
PIFa      ATVMGMIAEYYYRKRPRHLMDPSEVIERDVAGRKQMLRNLYQGSNVYCYD

A-Kiddo   MFRMSPEIFHALHDLLVSTYGLSSSNNVSSIESLAMFLWIVGGPQSFSQV
PIFa      SLRLTKRSFSDLCTILRERCDMCDTLNVSVEEKVAIFLLVVGHGTKMRMI

A-Kiddo   ESHFTRSLWTVHTKFHEVLKCLRKLAKDNITPRDPTFSMEHGRLREDRFW
PIFa      RSSYGWSLEPISRYFNEVLRGVLSLCHEFIKLPDPLAVQP.....EDSKW

A-Kiddo   PYFKDAIGAIDGSHISVVVLLDETISHTCHHGYTSQNVLAIYNFDMRFIF
PIFa      RWFEDCLGALDGTHIDVFVPLADQGRYRNRKQQITTNVLGVCDRHMKFVY

A-Kiddo   AVAGWPGSAHDSRILSHALANFPSFPMPGKYYLVDSGYPNRIGYLAPFKG
PIFa      VLAGWEGSASDSRVLRDAMSRDDAFAIPGKYYLVDAGYTNGPGFLAPYRS

A-Kiddo   TTYHIPEFRHRSGPPQGKYEVFNFLHSSLRNVI
PIFa      TRYHLNEWAAQGNNPSNAKELFNLRHSTARNYV
```

C.

```
                left TIR                    right TIR
PIFa      gggcccgtttgttt...aaacaaacgggccc
MathE1    ggtggtgttattc.....gaataacaccccc
KiddoE    gggtgtgtttggtt...aaccaaacacaccc
```

Figure 4.4.

VITA

Name: Guojun Yang

Permanent address: 1 Hensel Dr., Y2G

College Station, TX 77840

USA

Education: Sichuan Teachers' College (Normal)

Nanchong, P.R. China

B.S., 1991, Biology

Guangxi Agricultural University

Nanning, P.R. China

M.S., 1994, Microbiology

Texas A&M University

College Station, Texas, U.S.A.

Ph.D., 2003, Biology