COMPUTER-ASSISTED DISCOVERY AND CHARACTERIZATION OF RICE TRANSPOSABLE ELEMENTS INCLUDING THE FIRST ACTIVE MINIATURE-INVERTED REPEAT TRANSPOSABLE ELEMENT (MITE)

by

NING JIANG

(Under the direction of Susan R. Wessler)

ABSTRACT

The availability of draft sequences for the two subspecies of rice (*Oryza sativa*, *japonica* cv. Nipponbare and *indica* cv. 93-11) has significantly accelerated our understanding of transposable elements in the rice genome. The research described in this dissertation was performed with the expanding rice genomic database. First, 30 Mb of rice genomic sequence was analyzed to study the insertion preference of rice miniature inverted-repeat transposable elements (MITEs), numerically the most abundant elements in rice. Among the 6600 MITEs identified, > 10% were present as nested insertions (multimers) with the proportion of multimers differing among MITE families. The data suggest possible mechanisms underlying the formation of MITE multimers.

The second part of this dissertation concerns *Dasheng*, a novel non-autonomous long terminal repeat (LTR) element with over 1000 copies. Two hundred and fifteen elements were mapped to all twelve rice chromosomes, where more than half of the elements were located in the heterochromatic regions around centromeres. By searching 100 Mb rice genomic sequences including the almost completely assembled chromosome 1, *Dasheng* elements were found to have inserted five times more frequently into

pericentromeric regions than other regions. These features suggest *Dasheng* may serve as molecular markers for this marker-poor region of the genome.

Finally, 187 Mb of genomic sequence was analyzed in a computational approach to isolate the first active DNA transposons from rice and the first active MITE from any organism. The 430 bp *mPing,* a *Tourist*-like MITE, was shown to be actively transposing in a cell culture line. Database searches identified a family of related transposase-encoding elements (called *Pong*) that were also activated to transpose in the same cells. Virtually all new insertions of *mPing* and *Pong* elements were into low copy regions of the genome. Intriguingly, the *mPing* MITEs have preferentially amplified since domestication in cultivars adapted to environmental extremes, a situation reminiscent of McClintock's genomic shock theory for transposon activation. The isolation of an active MITE family and putative autonomous elements may provide a valuable tagging population for gene discovery and allow us to address long-standing questions about the mechanisms underlying the birth, spread and death of MITEs.

INDEX WORDS: MITE, LTR retrotransposon, activity, insertion preference, transpson display

COMPUTER-ASSISTED DISCOVERY AND CHARACTERIZATION OF RICE TRANSPOSABLE ELEMENTS INCLUDING THE FIRST ACTIVE MINIATURE-INVERTED REPEAT TRANSPOSABLE ELEMENT (MITE)

By

NING JIANG

B.S., Najing University, China, 1983

M.S., Jiangsu Agricultural College, China, 1986

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2002

COMPUTER-ASSISTED DISCOVERY AND CHARACTERIZATION OF RICE TRANSPOSABLE ELEMENTS INCLUDING THE FIRST ACTIVE MINIATURE-INVERTED REPEAT TRANSPOSABLE ELEMENT (MITE)

By

NING JIANG

Approved:

Major Professor: Susan R. Wessler

Committee: Kelly Dawe
Mike Scanlon
Zheng-Hua Ye
John McDonald

Electric Version Approved:

Gordhan L. Patel
Dean of the Graduate School
The University of Georgia
August 2002

DEDICATION

To my parents,

for the 2000 days and nights that I couldn't be around you.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my major professor, Sue Wessler, for her guidance, advice, support, and patience. I am especially grateful to her for the suggestion that I switch my research focus from maize to rice, and for her tremendous help with my writing and presentation. I am also indebted to my committee members: Kelly Dawe, John McDonald, Michael Scanlon, and Zheng-Hua Ye for their valuable advice and help. I thank the past and present members of the Wessler lab: Lane Arthur, Amy Bouck, Alexandra Casa, Bo Edwards, Cedric Feschotte, Dawn Holligan, Jianping Hu, Yun Hu, Ed Kentner, Yanhong Liu, Zenaida Magbanua, Alex Nagel, Mark Osterlund, Ryan Peeler, Ellen Pritham, Tesheka Stevenson, Lakshmi Swamy, Liangjiang Wang, Qiang Zhang, and Xiaoyu Zhang for their help and friendship. Especially, I thank Dawn for her effort in maintaining the lab, Cedric for helping me with advice on literature and writing, Liangjiang and Qiang for technical assistance, and Xiaoyu for his tremendous help in almost every aspect of my research. Special thanks also to some past and current members in Dawe lab, McDonald lab, and Scanlon lab: Nathan Bowen, Suneng Fu, Evelyn Hiatt, Jiabing Ji, King Jordan, Caroline Lawrence, Gene McCarthy, Nina Schubert, Hongguo Yu, and Cathy Zhong.

I am very grateful to our collaborators: Zhirong Bao and Sean Eddy (Washington University), Susan McCouch and Svieta Temnykn (Cornell University), Zhukang Cheng and Jiming Jiang (University of Wisconsin), Rod Wing (Clemson University), Jingdong

Liu and Wei Wu (Monsanto), and Hirohiko Hirochika (National Institute of Agrobiological Resources, Japan). Susan, Svieta, and Jiming are particularly helpful in sharing their knowledge, resources and technique, whereas Zhirong's input has dramatically changed the way that I design experiments and address questions. This dissertation would be totally different without his help.

I thank Orville Lindstrom and Michael Dirr for their instruction and support for my initial study in University of Georgia.

Finally, I thank my husband, Ping Wang, and my son, Weitian Wang, for their love and understanding. Thanks also to my friends: David Erikson, Cong Li, Song Lin, Angie Vineyard, Yuefang Wang, Donglin Zhang and Yingxia Zhang for their help and support.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Transposable elements are major components of eukaryotic genomes

Transposable elements (TEs) are DNA sequences capable of moving from one locus to another within their host genomes. Originally discovered by Barbara McClintock in the 1940s (McClintock, 1948; McClintock, 1949), TEs have been found in the genomes of all organisms tested (Capy et al., 1998). On the basis of their transposition mechanisms, TEs fall into two classes. Class I elements are characterized by their movement via an RNA intermediate during transposition. A DNA copy of the RNA intermediate is produced by reverse transcriptase and is integrated into a new locus in the host genome using other element-encoded proteins. Based on their structural features, class I TEs are further divided into two subclasses: LTR (long terminal repeat) elements and non-LTR elements. LTR elements, as exemplified by Ty elements in yeast (Boeke et al., 1985), have properties similar to retroviruses, including the LTR. Non-LTR elements, on the other hand, are not structurally related to retroviruses. These elements, including LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements), are the most abundant class I elements in the mammalian genomes(Smit, 1996). Since class I elements transpose without excision, their mobility will always increase their copy number and thereby the genome size. Consistent with this notion, recent data showed that in many species, the majority of repetitive DNA is composed of class I elements. In maize, LTR elements account for at least 50% of the genome (SanMiguel et al., 1998; SanMiguel et al., 1996). Over 40% of the human genome is represented by class I elements, most of which are LINEs and SINEs (Lander et al., 2001), suggesting that class I elements can contribute significantly to the expansion of genome size .

In contrast with class I elements, class II or DNA elements are characterized by short terminal inverted repeats (TIRs) and transposition via a DNA intermediate (reviewed in Kunze et al., 1997). Plant DNA elements (such as *Ac*/*Ds*, *Spm*/*dSpm* and *Mutator*) generally excise from one site and re-insert elsewhere in the genome. Class 2 elements can be further divided into two groups. Autonomous elements like *Ac* and *Spm* from maize encode the protein product(s) (transposase) necessary for their transposition (Baker et al., 1986; Kunze et al., 1997; Yoder, 1990). Non-autonomous elements, like *Ds* and *dSpm*, are usually internally deleted versions of autonomous elements. As a result, they require the presence of the autonomous elements *Ac* and *Spm*, respectively, for their transposition.

Due to their conservative mechanism of transposition, the copy number of class 2 element families is usually less than 100 per haploid genome. One exception to this generalization is miniature inverted repeat transposable elements (MITEs), a special category of non-autonomous elements that display very high copy number (in the thousands) and are uniformly short (usually < 600 bp). In addition, plant MITEs frequently insert into the TA dinucleotide or into a 3-bp trinucleotide (mostly TAA or TTA) (Bureau and Wessler, 1992; Mao et al., 2000; Zhang et al., 2000). Based on their TIRs and the target site sequence, most plant MITEs were categorized into two superfamilies: *Tourist*-like MITEs and *Stowaway*-like MITEs (Feschotte et al., 2002a).

Although first identified in plants including maize (Bureau and Wessler, 1992; Bureau and Wessler, 1994a; Bureau and Wessler, 1994b), rice (Bureau et al., 1996; Bureau and Wessler, 1994a; Bureau and Wessler, 1994b), green pepper (Pozueta-Romero et al., 1996) and *Arabidopsis* (Casacuberta et al., 1998; Le et al., 2000), MITEs are also abundant in several animal genomes including *C.elegans*, mosquitoes, fish and humans

(reviewed in Feschotte et al., 2002b). Despite their abundance, the transposition mechanism of MITEs remained a mystery for years, because no active MITE families were known and candidate transposase had not been identified. Recently, evidence has been accumulated linking *Tourist* and *Stowaway* MITEs with two superfamilies of transposases, *PIF/Harbinger* and *Tc1/mariner*, respectively (Feschotte and Wessler, 2002; Turcotte et al., 2001; Zhang et al., 2001). However, without activity, these associations remained circumstantial and they provide little information about how MITEs arise from autonomous elements and how they have spread so successfully throughout the genomes of plants and animals.

The targeting specificity and distribution of TEs

Although TEs can be inserted into many different sites in the genome, many elements demonstrate a certain degree of target site preference (Craig, 1997). While diverse patterns of target site selection are observed for different TEs, the evolutionary significance of target site selection might be the same: to facilitate element propagation and to optimize the element-host relationships (Brookfield, 1995).

Sequence specificity

Target site preference for insertion into specific DNA sequences is more common among class II elements than class I element. Such a preference can be revealed by the target site duplication (TSD), which is created upon the transposition of most TEs. As mentioned earlier, the *Stowaway*-like MITEs integrate specifically into the dinuleotide TA, as do their putative autonomous partners the *Tc1/mariner* elements. The target site of class I elements is generally not sequence-specific, with the exception of ZAM elements

in *Drosophila* (Leblanc et al., 1999), which always insert into the sequence GCGCGC. Interestingly, the yeast Ty elements show a partial sequence specificity, i. e., most 5 bp TSDs are composed of two random nucleotides flanking 3bp AT rich cores (Kim et al., 1998).

Structural specificity and preference for different chromosomal locations

Some elements demonstrated preference for sites with certain structural features. For instance, the target choice of *Tc1* is not only determined by the primary DNA sequence, but also by the topological structure of target site. "Hot" sites (which are targeted frequently) are only formed at supercoiled regions (Ketting et al., 1997). In vitro, HIV retrovirus inserted more frequently into sites in nucleosome than those in linear DNA (Pruss et al., 1994). Further experiments show that the severe DNA bending within the nucleosome core creates favored sites for retroviral integration (Muller and Varmus, 1994).

In addition to the structural specificity, many class I elements exhibit an uneven distribution in the genome. In *Drosophila*, two non-LTR elements, *TART* and *HeT-A* are telomere specific (Sheen and Levis, 1994). In fact, the two elements serve as telomeres of chromosomes in *Drosophila*. Likewise*, Zepp*, a LINE-like element in *Chlorella*, accumulates primarily in the telemetric region (Higashiyama et al., 1997). In yeast, the Ty5 elements preferentially insert into regions of silent chromatin at the telomeres and at the mating type loci (Zou et al., 1996), whereas the majority of Ty1 and Ty3 elements integrate near genes transcribed by RNA polymerase III. Such a preference was believed to arise from the interactions between their integration complexes and the Pol III transcription machinery (Devine and Boeke, 1996).

The uneven distribution of class I elements has also been observed in plants. In the grass family, a Ty3/*gypsy* type retrotransposon accumulates primarily in centromeric regions and is suspected to be involved in the formation and function of centromeres. Likewise, class I elements are concentrated in centromeric and pericentromeric regions in *Arabidopsis*, the first plant with a completely sequenced genome.

Self-targeting

Self-targeting refers to the preference of TEs to insert into sequences of other members of the same or a related family. In maize, double *Ds* and *Ac* elements (formed by one *Ds* or *Ac* element inserted into another) have been shown to be responsible for chromosome breakage and more complex chromosomal rearrangements (Doring et al., 1989; Doring and Starlinger, 1984; McClintock, 1949; Michel et al., 1994; Weck et al., 1984). In the slime mold, the *DIRS-1* element was found to insert into other *DIRS-1* sequences in five out of six cases analyzed (Cappello et al., 1984), and *Tp1* elements can form scrambled clusters up to 50 kb long by inserting into each other (Rothnie et al., 1994). From an evolutionary point of view, the self-insertion may benefit both the element and the host. On the one hand, the pre-existing element provides a "safe haven" for the incoming element. On the other hand, the insertion may limit the activity of the pre-existing element, which could otherwise contribute to deleterious mutation in the host genome.

Rice and its TEs

Rice belongs to the grass family, a diverse, successful family that includes many important crops such as maize, wheat, barley, sorghum and sugarcane. Among them, rice

provides the staple food for more than half of the world population. Unlike most grasses, rice has a relatively small genome of about 430 bp (Burr, 2002). It is the first higher organism for which draft sequences are publicly available for two subspecies, *japonica* and *indica* (Goff et al., 2002; Yu et al., 2002). In addition, due to the extensive synteny among the grasses, the identification and characterization of rice genes will provide valuable information for those studying related traits or regions in the genomes of other grasses (Bennetzen, 2002). These features, together with its excellent genetic map and well-established transformation system, reveals the great value of rice as a model plant.

Despite its small genome, rice is still a good model organism for the study of TEs. The genome of *Oryza sativa* contains all of the major types of elements found in the larger grass genomes including retrotransposons, MITEs and other DNA elements (Bureau et al., 1996; Mao et al., 2000; Tarchini et al., 2000; Turcotte et al., 2001). It is estimated that over 40% of the rice genome is composed of repetitive DNAs, most of this being derived from TEs (Yu et al., 2002). Furthermore, the availability of well-characterized wild relatives provides the material necessary to analyze the impact of TEs on genome evolution and speciation. The genus *Oryza* has more than twenty species whose evolutionary relationships have been the subject of several phylogenetic analyses (Ge et al., 1999; Uozo et al., 1997).

To understand the impact of TEs on genome evolution, it is very important to isolate active elements and study their transposition mechanism, because it is the ultimate force by which TEs amplify themselves and change their host genomes. Although TEs are abundant in the rice genome, few elements appear to be active. The first active elements reported in rice were the three LTR elements, *Tos10*, *Tos17*, and *Tos19*. These

elements were isolated with a simple reverse transcription PCR protocol that was devised to amplify a conserved reverse transcriptase domain in mRNAs that were isolated during cell culture (Hirochika et al., 1996). To date, these elements remained to be the only active elements that have been identified in rice, partly because that RT-PCR approach is not applicable to TEs other than LTR elements. Consequently, it demands the development of new approaches to isolate other type of active TEs in rice.

Outline for this dissertation

In this dissertation, rice TEs were searched and analyzed using the expanding rice genome database. First, a comprehensive analysis of the MITE multimers in rice was performed in order to understand an unique insertion pattern of some MITEs that has been observed in maize. The analysis led to several interesting conclusions in regard to the abundance and insertion patterns of rice MITEs, although the original questions about maize MITEs were not fully resolved. Second, we characterized one of the highest copy number and most recently amplified LTR elements in rice, *Dasheng*. As a successful nonautomous LTR element, *Dasheng* is particularly useful for studying the interaction between non-autonomous LTR elements and their autonomous partners. In addition, the clustering of *Dasheng* elements in pericentromeric regions can serve as a good marker system for these marker-poor regions. Finally, we developed a new approach for isolating active TEs from genomes where large amount of genomic sequences are available. Using this approach, we isolated the first active MITE and its putative autonomous elements, which represent a new family of transposases in plants and animals.

References

Baker, B., Schell, J., Lorz, H., and Fedoroff, N. (1986). Transposition of the maize controlling element *Activator* in tobacco, PNAS *83*, 4844-4848.

Bennetzen, J. (2002). The rice genome. Opening the door to comparative plant biology, S *296*, 60-3.

Boeke, J. D., Garfinkel, D. J., Styles, C. A., and Fink, G. R. (1985). Ty elements transpose through an RNA intermediate, C *40*, 491-500.

Brookfield, J. F. Y. (1995). Transposable elements as selfish DNA. In Mobile Genetic Elements, D. J. Sherratt, ed. (New-York, Oxford University Press), pp. 130-153.

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, PNAS *93*, 8524-8529.

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

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

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

Burr, B. (2002). Mapping and sequencing the rice genome, Plant Cell *14*, 521-3.

Cappello, J., Cohen, S. M., and Modish, H. F. (1984). Dictyostelium transposable element DIRS-1 preferentially inserts into DIRS-1 sequences, Mol Cell Biol *4*, 2207-13.

Capy, P., Bazin, C., Higuet, D., and Langin, T. (1998). Dynamics and evolution of transposable elements (Austin, Texas, Springer-Verlag).

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.

Craig, N. L. (1997). Target site selection in transposition, Annu Rev Biochem *66*, 437-74.

Devine, S. E., and Boeke, J. D. (1996). Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III, Genes Dev *10*, 620-33.

Doring, H. P., Nelsensalz, B., Garber, R., and Tillmann, E. (1989). Double *Ds* elements are involved in specific chromosome breakage, MGG *219*, 299-305.

Doring, H. P., and Starlinger, P. (1984). Barbara McClintock's controlling elements: now at the DNA level, C *39*, 253-260.

Feschotte, C., Jiang, N., and Wessler, S. R. (2002a). Plant transposable elements: where genetics meets genomics, Nat Rev Genet *3*, 329-41.

Feschotte, C., Zhang, X., and Wessler, S. (2002b). Miniature Inverted-repeat Transposable Elements (MITEs) and their relationship with established DNA transposons. In Mobile DNA II, N. Craig, R. Craigie, M. Gellert, and A. Lambowitz, eds. (Washington D.C., American Society of Microbiology Press.), pp. 1147-1158.

Feschotte, C., and Wessler, S. R. (2002). *Mariner*-like transposases are widespread and diverse in flowering plants, PNAS *99*, 280-285.

Ge, S., Sang, T., Lu, B.-R., and Hong, D.-Y. (1999). Phylogeny of rice genomes with emphasis on origins of allotetroaploid species, PNAS *96*, 14400-14405.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H.*, et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica), S *296*, 92-100.

Higashiyama, T., Noutoshi, Y., Fujie, M., and Yamada, T. (1997). Zepp, a LINE-like retrotransposon accumulated in the Chlorella telomeric region, Embo J *16*, 3715-23.

Hirochika, H., Sugimoto, K., Otsuki, Y., and Kanda, M. (1996). Retrotransposons of rice involved in mutations induced by tissue culture, Proc Natl Acad Sci *93*, 7783-7788.

Ketting, R. F., Fischer, S. E., and Plasterk, R. H. (1997). Target choice determinants of the Tc1 transposon of Caenorhabditis elegans, Nucleic Acids Res *25*, 4041-7.

Kim, J. M., Vanfuri, J. D., Boeke, J., Gabrien, A., and Voytas, D. F. (1998). Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence, Genome Res *8*, 464-478.

Kunze, R., Saedler, H., and W.E., L. (1997). Plant transposable elements, ABR *27*, 331-470.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome, Nat *409*, 860-921.

Le, Q. H., Wright, S., Yu, Z., and Bureau, T. (2000). Transposon diversity in *Arabidopsis thaliana*, PNAS *97*, 7376-7381.

Leblanc, P., Dastugue, B., and Vaury, C. (1999). The integration machinery of *ZAM*, a retroelement from Drosophila melanogaster, acts as a sequence-specific endonuclease, J Virol *73*, 7061-4.

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

McClintock, B. (1948). Mutable loci in maize., car *47*, 155-169.

McClintock, B. (1949). Mutable loci in maize, car *48*, 142-154.

Michel, D., Salamini, F., Motto, M., and Doring, H. P. (1994). An unstable allele at the maize *opaque2* locus is caused by the insertion of a double *Ac* element, MGG *243*, 334-342.

Muller, H. P., and Varmus, H. E. (1994). DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes, Embo J *13*, 4704-14.

Pozueta-Romero, J., Houlné, 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.

Pruss, D., Bushman, F. D., and Wolffe, A. P. (1994). Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core, Proc Natl Acad Sci U S A *91*, 5913-7.

Rothnie, H. M., McCurrach, K.J., Glover, L.A., and Hardman, N. (1991). Retrotransposon-like nature of *Tp1* elements: implications for the organisation of highly

repetitive, hypermethylated DNA in the genome of *Physarum polycephalum*. Nucleic Acids Res. *19*, 279-286.

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize, Nat Genet *20*, 43-5.

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, S *274*, 765-768.

Sheen, F. M., and Levis, R. W. (1994). Transposition of the LINE-like retrotransposon TART to *Drosophila* chromosome termini, PNAS *91*, 12510-12514.

Smit, A. F. (1996). The origin of interspersed repeats in the human genome, Curr Opin Genet Develop *6*, 743-748.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., and Rafalski, A. (2000). The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4, Plant Cell *12*, 381-391.

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

Uozo, S., Ikehashi, H., Ohmido, N., Ohtsubo, H., Ohtsubo, E., and Fukui, K. (1997). Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus *Oryza*, Plant Mol. Biol. *35*, 791-799.

Weck, E., Courage, U., Doring, H.-P., Fedoroff, N., and Starlinger, P. (1984). Analysis of *sh-m6233*, a mutation induced by the transposable element *Ds* in the sucrose synthase gene of *Zea mays*, EMBO *3*, 1713-1716.

Yoder, J. I. (1990). Rapid proliferation of the maize transposable element *Activator* in transgenic tomato, Plant Cell *2*, 723-730.

Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X.*, et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica), Science *296*, 79-92.

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 (Hbr)* into genic regions of maize, PNAS *97*, 1160-1165.

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, PNAS *98*, 12572-12577.

Zou, S., Ke, N., Kim, J. M., and Voytas, D. F. (1996). The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci, Genes Dev *10*, 634-45.

CHAPTER 2

INSERTION PREFERENCE OF MAIZE AND RICE MINIATURE-INVERTED REPEAT TRANSPOSABLE ELEMENTS (MITES) AS REVEALED BY THE ANALYSIS OF NESTED ELEMENTS[a]

Introduction

Transposable elements are usually divided into two classes. Class 1, or retroelements, including the long terminal repeat (LTR) retrotransposons, make up the largest fraction of most plant genomes (reviewed in Kumar and Bennetzen, 1999). Retroelements are capable of attaining very high copy numbers in a relatively short timeframe because the element-encoded mRNA, and not the element itself, forms the transposition intermediate. Class 2 or DNA elements are characterized by short terminal inverted repeats (TIRs) and transposition via a DNA intermediate (reviewed in Kunze et al., 1997). Plant DNA elements (such as *Ac*/*Ds*, *Spm*/*dSpm* and *Mutator*) generally excise from one site and re-insert elsewhere in the genome. Class 2 elements can be further divided into two groups. Autonomous elements like *Ac* and *Spm* from maize encode the products (transposase) necessary for their transposition (Baker et al., 1986; Yoder, 1990; Kunze et al., 1997). Nonautonomous elements, like *Ds* and *dSpm*, are usually internally deleted versions of autonomous elements. As a result, they require the presence of the autonomous elements *Ac* and *Spm*, respectively, for their transposition.

Due to their conservative mechanism of transposition, the copy number of class 2 element families is usually less than 100 per haploid genome. One exception to this generalization is miniature inverted repeat transposable elements (MITEs), a special category of nonautonomous elements that display very high copy number (in the thousands) and are uniformly short (usually < 500 bp). In addition, most MITEs in plants have TIRs and insert into the TA dinucleotide or into a 3-bp trinucleotide (Bureau and Wessler, 1992; 1994; Mao et al., 2000; Zhang et al., 2000). Although first identified in several plant species including maize (Bureau and Wessler, 1992; 1994a; 1994b), rice (Bureau and Wessler, 1994a; 1994b; Bureau et al., 1996), green pepper (Pozueta-Romero et al., 1996) and *Arabidopsis* (Casacuberta et al.,

1998; Le et al., 2000), MITEs are also abundant in several animal genomes including *C.elegans* (Oosumi et al., 1995; Surzycki and Belknap, 2000), mosquitoes (Tu, 1997; 2001; Feschotte and Mouchès, 2000), fish (Izsvák et al., 1999) and humans (Morgan, 1995; Smit and Riggs, 1996).

Another important feature of MITEs is their preference for insertion into low copy or genic regions (Tikhonov et al., 1999; Zhang et al, 2000; Mao et al, 2000). In addition, MITEs were also found, in several cases, to insert into each other. For instance, the first *Stowaway* element was found as an insertion in a sorghum *Tourist* element (Bureau and Wessler, 1994b), whereas in another case a *Tourist* dimer was found in the same organism (Tikhonov et al., 1999). Such MITE multimers were also reported in other organisms, including rice (Tarchini et al., 2000) and mosquitoes (Tu, 1997; Feschotte and Mouchès, 2000). Therefore, it was proposed that MITEs could be preferential targets for other MITEs (Feschotte and Mouchès, 2000).

Given the previously identified target site preference of MITEs and the frequent detection of MITE multimers, we wondered about the propensity of MITE insertion into other MITEs. Such a determination is only possible with a systematic comparison between the insertion frequency of MITEs into MITEs and the frequency of MITEs into other sequences. In this study, we report a detailed characterization of maize *Tourist* multimers, and a comprehensive analysis of MITE multimers in rice, a species known to be particularly rich in MITEs (Bureau et al, 1996; Mao et al., 2000; Tarchini et al., 2000). The availability of 30.2 Mb of rice genomic sequences has enabled us to address questions about MITE multimers that could not be answered with the available maize sequence. The analysis of rice genomic sequence not only allows us to evaluate the prevalence of MITE multimers, but may also provide new insight into the temporal order of amplification of different transposable elements in the rice genome.

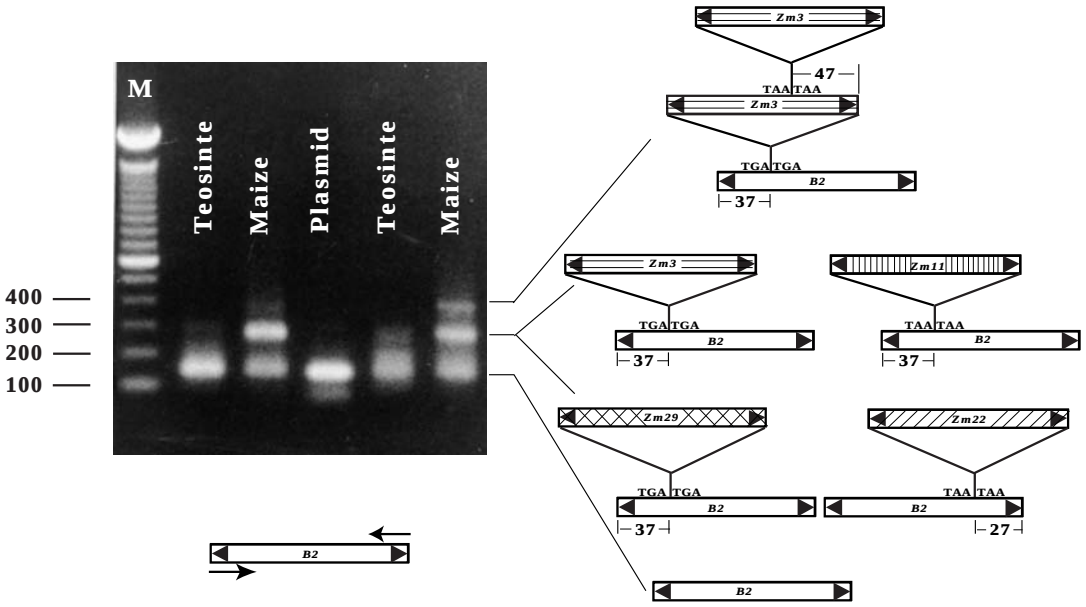
Results

Tourist multimers in maize

The first reported MITE was the *B2* element, found as a 128-bp insertion into the maize *waxy* (*wx*) gene in the mutant *wxB2* allele (Wessler and Varagona, 1985; Bureau and Wessler, 1992). Subsequent database searches revealed that this element belongs to a large family of related elements, called *Tourist*, whose members are associated with the non-coding regions of genes from maize, sorghum, barley and rice (Bureau and Wessler, 1992; 1994a). *Tourist* multimers were initially discovered in a PCR assay that was intended to identify additional *B2*-like (*Tourist*) elements in maize. Genomic DNA from maize inbred B79 was amplified with primers derived from the *B2* TIR (14 bp) and 11 bp of internal element sequence. In addition to a product corresponding to the size of the *B2* element (128 bp), larger fragments that varied in size depending on the annealing temperature were observed (Figure 2.1). PCR products from all size classes were cloned and several sequenced, revealing monomers, dimers and a trimer (Figure 2.1). Four different multimers were found among the six dimer-sized clones that were sequenced. In contrast, the largest PCR fragment corresponded to a single trimer. All multimers contained a variety of elements that, like *B2*, are all members of *Tourist* subfamily A (Bureau and Wessler, 1994a).

To rule out the possibility that the *Tourist* multimers were artifacts of PCR amplification, dimer and trimer products were used to probe a small insert library derived from B79 genomic DNA. Three of eleven sequenced clones contained *Tourist* multimers thus confirming the presence of *Tourist* multimers in the genome.

Figure 2.1: The identity of multimers resulting from PCR amplification of *B2-Tourist* elements in maize line B79. PCR products resolved by agarose gel electrophoresis and visualized by ethidium bromide staining were purified, cloned and sequenced (see Materials and Methods). The annealing temperature for PCR was 60°C for the two samples on the left and 55°C for the last 3 samples. The identity of PCR bands is diagrammed on the right. The positions of the insertion sites of the various *Tourist* subfamily members (*Zm3*, *11*, *22*, *29*) (Bureau and Wessler, 1992; Bureau and Wessler, 1994a) (N. Jiang and S. Wessler, unpublished data) into *B2* elements are shown along with the TSD.


M
Teosinte
Maize
Plasmid
Teosinte
Maize
400
300
200
100
Zm3
47
TAA TAA
Zm3
TGA TGA
B2
37
Zm3
TGA TGA
B2
37
Zm11
TAA TAA
B2
37
Zm29
TGA TGA
B2
37
Zm22
TAA TAA
B2
27
B2
B2

Insertion into preexisting MITEs

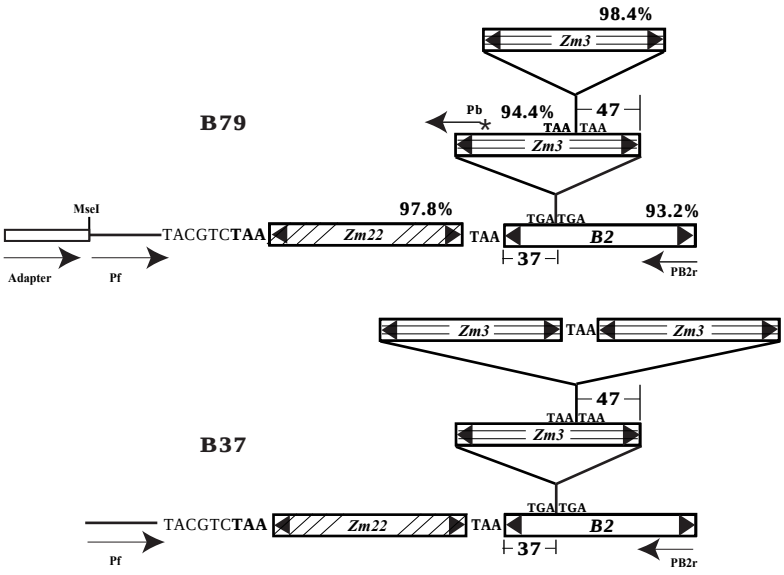
Among maize inbred lines, the insertion sites of MITEs were frequently polymorphic with respect to the presence or absence of an element at a particular locus (Casa et al., 2000; Zhang et al., 2000). Polymorphism of this type is usually associated with the recent spread of transposon families through the genome. In light of these findings, we designed a PCR assay to detect insertion site polymorphism within *Tourist* multimers. In this way evidence might be obtained for the sequential insertion of one element into another.

The locus harboring the *Tourist* trimer (Figure 2.1) was investigated for possible insertion polymorphism among different maize lines. Following methodology described in the Material and Methods, B79 genomic sequence adjacent to one end of the trimer was obtained, revealing that another *Tourist* element (*Tourist-Zm22*) had inserted adjacent to the trimer with only an intervening target site duplication (TSD)(Figure 2.2). A locus specific primer was designed from the sequence flanking *Tourist-Zm22* and used to amplify B37 genomic DNA together with a *B2* terminal primer (PB2r). The resulting PCR product, which harbored an additional *Tourist* element (*Tourist-Zm3*)(Figure 2.2), provided evidence for the progressive formation of multimers (tetramers from trimers).

Non-random insertion sites

Insertion sites within the sequenced multimers were clearly non-random. For ease of comparison, insertion sites have been calculated as the number of base pairs from the closest end of the target element to the first nucleotide of the TIR of the insertion element. For all insertions examined this value corresponded to 27 bp, 37 bp or 47 bp (Figure 2.1). To test whether this periodicity was representative of the multimers in the

Figure 2.2: Diagram of a MITE trimer and tetramer found at the same locus in different maize lines. As in Figure 2.2.1, the positions of the insertion sites of the various *Tourist* subfamily members into *Tourist-B2* and other *Tourist-Zm* elements are shown along with the TSD. The sequence flanking the trimer in B79 was amplified with primer Pb and an adapter primer (see Materials and Methods for details). For B37, the tetramer was amplified using a locus specific primer, Pf and the *B2* primer PB2r. The number above each element in B79 indicates the similarity between the element in B79 and its counterpart in B37. *: The primer was labeled with P^{33}.


98.4%
Zm3
B79
Pb
94.4%
47
TAA
TAA
Zm3
MseI
97.8%
TGA
TGA
93.2%
TACGTCTAA
Zm22
TAA
B2
37
Adapter
Pf
PB2r
Zm3
TAA
Zm3
47
TAA
TAA
Zm3
B37
TGA
TGA
TACGTCTAA
Zm22
TAA
B2
37
Pf
PB2r

maize genome, a two-step PCR assay was employed to isolate additional multimers. In this assay (see Materials and Methods), the length of the PCR products reflects the position of the insertion sites within the multimers. That is, if the insertion sites are 10 bp apart, the PCR products will appear, more or less, as a 10 bp "ladder" on the gel. Such a ladder was indeed observed (Figure 2.3). Furthermore, sequencing of selected PCR products revealed that all contained a *Tourist-Zm3* element inserted into another *Tourist* element at roughly 10 bp intervals. The composition of some of the multimers is diagrammed in Figure 2.3. In addition, these data, and the data from all prior multimer sequences, are summarized in Table 2.1.

MITE multimers in rice

In the absence of a significant amount of maize genomic sequence, analysis of maize multimers is restricted to a description of the phenomenon and the characterization of a small fraction of the existing elements. A more thorough survey, including an estimate of the proportion of the multimers present, is possible for rice because a large amount of rice genomic sequence is publicly available (Yuan et al., 2001), and the rice genome contains thousands of MITEs (Mao et al., 2000, Tarchini et al., 2000).

Prevalence of multimers

Computer searches were restricted to 30.2 Mb of complete BAC and PAC sequences, of which 6.6 Mb were derived from pericentromeric regions (based on the most recent data on the location of rice centromeres, Harushima et al., 1998; Cheng et al., 2001). No significant differences were observed in the insertion pattern of MITEs between the sequences from chromosomal arms and those from pericentromeric regions (Table 2.2 and Table 2.3). Of the rice sequences queried, a total of 6641 MITEs were

Figure 2.3: Autoradiograph of PCR products resolved on an acrylamide gel showing an approximate 10 bp "ladder". The position of primers used to obtain the PCR products from maize lines or plasmid DNA are indicated by horizontal arrows; the composition of multimers represented by the indicated PCR products is also diagrammed. Vertical arrows over some multimers represent *Zm3* insertions. Lane 1: *Zm3-B2* dimer containing plasmid; lane 2: B79 genomic DNA; lane 3: B73 genomic DNA; lane 4: recombinant inbred DNA from a B73 X Mo17 mapping population. *: The primer was labeled with P^{33}.

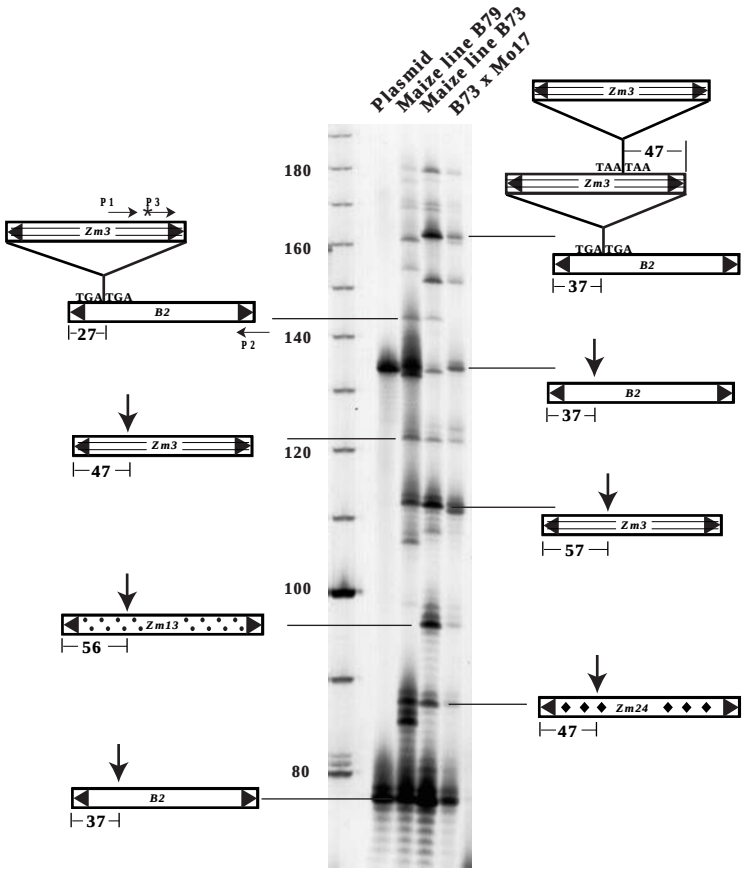

Plasmid
Maize line B79
Maize line B73
B73 x Mo17
180
160
140
120
100
80
P 1
P 3
P 2
Zm3
TGATGA
B2
27
47
TAATAA
37
57
Zm13
56
Zm24

Table 2.1. The insertion sites within *Tourist* dimers and trimers

Maize line	Preexisting element	Insertion element	Insertion site[a] (bp)	TSD	Source
B79	*B2*	*Zm3*	37	TGA	B2PCR
B79	*B2*	*Zm11*	37	TAA	
B79	*B2*	*Zm29*	37	TGA	
B79	*B2*	*Zm22*	27	TTA	
B79	*B2*[c] *(1)*	*Zm3*[c] *(2)*	37	TGA	
B79	*Zm*[c] *(2)*	*Zm3*[c] *(3)*	47	TTA	
B73	*B2*	*Zm3*	27	TTA	Two-step PCR
RIL22[b]	*Zm13*	*Zm3*	46	TAA	
B73	*B2*	*Zm3*	37	TTA	
B73	*Zm3*	*Zm3*	47	TAA	
B73	*Zm3*	*Zm3*	57	TTA	
B73	*Zm13*	*Zm3*	67	TTA	
B73	*Zm24*	*Zm3*	47	TCA	
B79	*B2*	*Zm3*	37	TGA	
RIL7	*Zm24*	*Zm3*	37	TGA	
Spanco	*Zm13*	*Zm22*	67	TTA	
Spanco	*Zm24*	*Zm22*	59	TAA	
B79	*B2*	*Zm28*	58	TAA	Genomic Library
B79	*Zm13*	*Zm25*	46	TGA	
B79	*Zm27*[c] *(1)*	*Zm3*[c] *(2)*	47	TGA	
B79	*Zm3*[c] *(2)*	Unknown element[c] (3)	47	TAA	

[a] Insertion site is defined as the number of base pairs from the closest end of the target element to the first nucloetide of the TIR of the insertion element

[b]RIL represents recombinant inbreed lines from B73 x M017 cross.

[c] Elements that form trimers. The numbers in brackets indicate the order of insertion. Element "1" is the first preexisting element; element "2" is the element which inserted into element "1"; similarly, element "3" is the element which inserted into element "2".

detected (Table 2.2) with RepeatMasker program (see Materials and Methods for details). This corresponds to 0.22 MITEs/kb of genomic DNA or 1 MITE/4.5 kb. MITEs account for 1.54 Mb of DNA or 5.1% of the genomic sequence analyzed. These values are very close to those found in a previous study of MITEs from a 350kb contig (Tarchini et al., 2000). MITEs grouped into 41 different families of which 26 were previously reported while 15 were identified in this study (Bureau and Wessler, 1994a; 1994b; Bureau et al., 1996; Song et al, 1998; Zhang and Kochert, 1998; Tarchini et al., 2000; Turcotte et al., 2001)(see supplemental data for sequence of the new identified MITE families).

Of the 6641 MITEs, 732 (or about 11%), are part of 340 multimers. This includes 293 dimers, 35 trimers, 9 tetramers, and 3 petamers (the trimers and tetramers also contain non-MITE elements). These 387 MITEs inserted into other MITEs corresponds to 387 MITEs/1540 kb of MITEs or an insertion frequency of MITEs into MITEs of 0.25 per kb or 1 MITE per 4 kb (Table 2). In contrast, there are very few insertions of MITEs into class 1 elements or into other class 2 elements despite the fact that these elements comprise a much larger fraction of the genome. While there is one MITE inserted per 4 kb of MITE DNA, there is only one MITE inserted per 330 kb of LTR retrotransposons and per 127 kb of other class 2 elements. These data indicate either a target site preference of MITEs for other MITEs or that MITE amplification preceded the amplification of the other elements in the genome. In this latter scenario, it is envisioned that the bulk of the class 1 and non-MITE class 2 elements were not in the genome when most of the MITE families were undergoing amplification. In contrast, non-MITE elements show no discrimination for insertion into MITEs (Table 2.2): while the

Table 2.2. Multimers containing rice MITEs

	Insertion of MITEs into				Insertion of other elements into	
	MITEs	LTR elements	Other DNA elements	All genomic sequences analyzed	MITEs	All genomic sequences analyzed
Number of elements	6641	1171	646		6641	
kb of DNA	1540	4650	1780	30183	1540	30183
Insertion events	387	14	14	6641	93	2185
Insertion/ kb	0.251* (1/4.0kb)	0.003** (1/330 kb)	0.008** (1/127 kb)	0.220 (1/4.5 kb)	0.060 (1/17 kb)	0.072 (1/14 kb)

*P<0.05, **P<0.01 (compared to that for all genomic sequences by $\chi 2$ test)

Table 2.3. Multimers containing rice MITEs in pericentromeric regions

	Insertion of MITEs into				Insertion of other elements into	
	MITEs	LTR elements	Other DNA elements	All genomic sequences analyzed	MITEs	All genomic sequences analyzed
Number of elements	706	366	368		706	
kb of DNA	168	1250	580	3602	168	3602
Insertion events	43	4	3	706	25	734
Insertion/ kb	0.256 (1/4.0kb)	0.003** (1/330 kb)	0.005** (1/193 kb)	0.196 (1/5.0 kb)	0.149 (1/6.7 kb)	0.204 (1/4.9 kb)

*P<0.05, **P<0.01 (compared to that for all genomic sequences by $\chi 2$ test)

frequency of insertion into MITEs is 1 per 17 kb of MITEs, the insertion frequency into all genomic DNA is slightly higher at 1 per 14 kb.

Self-insertions

The data presented in Table 2.2 reveals a slight preference for insertion of MITEs into MITEs. However, analysis of these data for individual MITE families indicates that this preference is not displayed by all families and is largely due to self-insertions. Four MITE families were analyzed in detail (Table 2.4). These families were chosen because they are abundant and they represent different groups of MITEs. Among the four families analyzed, *Castaway*, *Gaijin* and *Ditto* are related to *Tourist* elements in maize, whereas *Stowaway* elements belong to another superfamily (see Discussion). As shown in Table 2.4, all the *Tourist*-related elements have sustained more insertions per kb of DNA than the genome as a whole (insertion frequencies of .38, .32 and .60, respectively vs. .22 for all DNA, Tables 2,3). In contrast, *Stowaway* has sustained insertions at roughly the same frequency as the rest of the genome. While the increased insertions into *Castaway* and *Gaijin* elements can be completely accounted for by self-insertions, *Ditto* elements appear to attract a variety of MITEs (see Discussion). The cluster of six elements from chromosome 1 (Figure 2.4) illustrates the propensity for self-insertion among *Castaway* family members.

One could argue that the observed higher self–insertion frequency of MITEs reflects a preference of MITEs for particular regions of the genome rather than a preference for other members of the same family. If this is the case, for a certain family of MITEs there would be a comparable number of insertions into sequences flanking MITEs as there are into MITEs. Fortunately, the availability of 30.2 Mb of rice contigs

Table 2.4. The self-insertion preference of some MITEs in rice

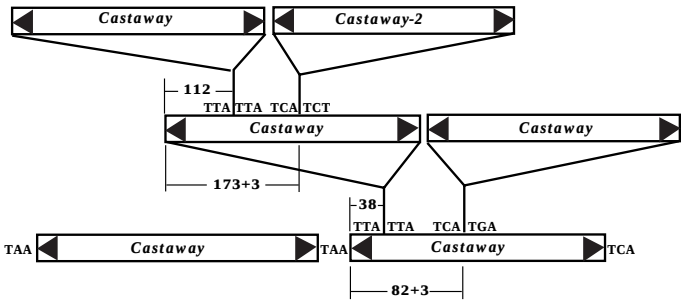
Elements	*Castaway*	*Gaijin*	*Ditto*	*Stowaway*
Copies of the element in 30.2 Mb genomic sequence	225	486	345	2690
Total size (kb)	76	71	89	561
Insertions by MITEs	29	23	53	104
Self-insertions	20(69%[b])	16(70%[b])	13(25%[b])	77 (71%[b])
Insertion frequency by all MITEs [a]	0.38**	0.32	0.60**	0.19
Self-insertion frequency [a]	0.26	0.23	0.15	0.14

[a] Insertion frequency equals insertions into the element divided by the total size (kbs) of the element

[b] Percent of self-insertions

** $p<0.01$ (compared to the average insertion frequency (0.223) in the genome by $\chi 2$ test)

Figure 2.4: A cluster of six *Castaway* elements on a rice PAC clone (accession No. ap002844) from chromosome 1. The distance from the end of the element to the insertion site is shown. *Castaway*-2 is a member of a subfamily of *Castaway*. The +3 indicates the inclusion of the 3 bp TSD that was generated upon insertion.


Castaway
Castaway-2
112
TTA TTA TCA TCT
Castaway
Castaway
173+3
38
TTA TTA TCA TGA
TAA
Castaway
TAA
Castaway
TCA
82+3

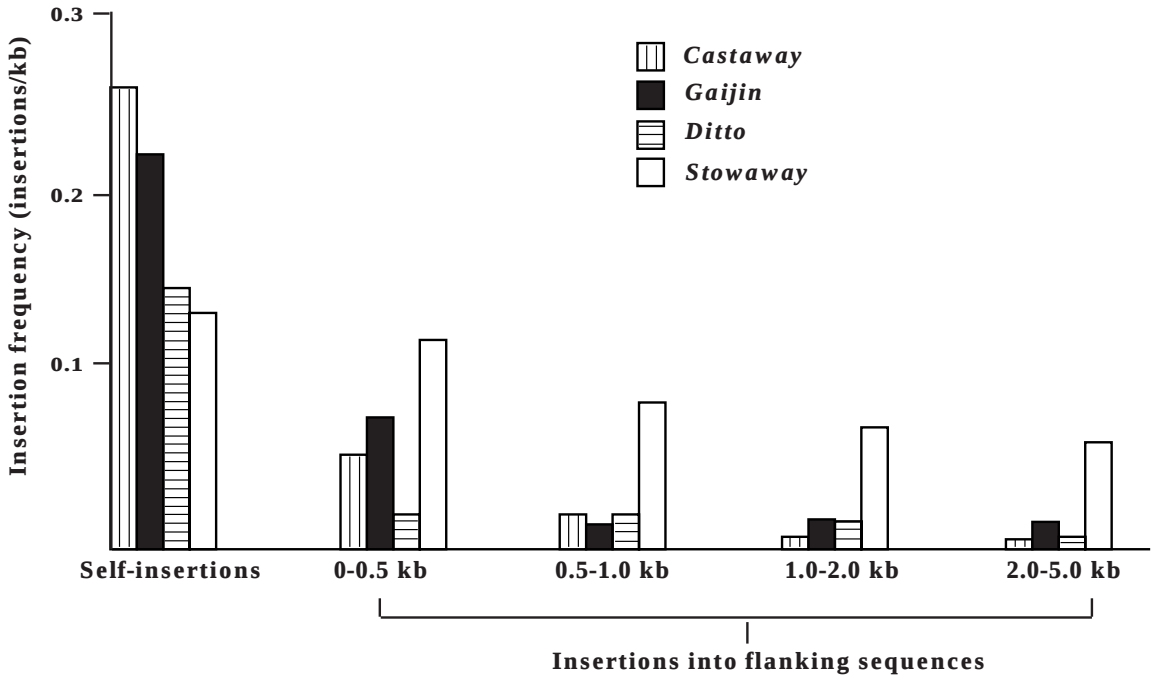
permits an analysis of the insertions into MITEs and their flanking sequences. Based on the data presented in Figure 2.5, it is evident that for all families examined, except *Stowaway*, the self-insertion frequency of MITEs is significantly higher than the insertion frequencies into their flanking sequences ($p < 0.01$ by $\chi 2$ test).

MITE multimers cannot transpose

A MITE multimer can arise in at least two ways. The first is by the insertion of a MITE into another MITE, and the second is by amplification of a multimer. If a multimer is capable of transposition, several copies of the same multimer should be detected, and the multimers should evolve similarly as single elements. Furthermore, these copies should be composed of the same elements in the same relative orientation and with the same insertion site and TSD. Among the 340 MITE multimers identified in this study, only three pairs of dimers share these structural features. However, the sequence similarity between the members of each dimer pair ranges from 65%- 72%, whereas at least one of the insertion elements in each dimer pair has homologs with over 90% similarities in the same database. The striking discrepancy suggests that these dimer pairs resulted from independent insertions instead of amplification of dimers.

There was one exception involving a dimer composed of a MITE and a DNA element. This element, (called *Midway*) initially found as an 850 bp insertion in a *Stowaway-Os1* element, has 11 bp TIRs and an 8 bp TSD. A closer examination indicates that *Midway* harbors another *Stowaway* element (*Stowaway-Os25*). That there are three *Midway*/*Stowaway* composite elements in the database sharing 93-96% overall DNA sequence identity, suggests that *Midway* can still transpose despite (or due to) the *Stowaway-Os25* insertion.

Figure 2.5: The frequencies of self-insertions and insertions into flanking sequences for four MITE families in rice. Frequencies were calculated in the same way as in Table 2.2 (see Materials and Methods for details).


0.3
0.2
0.1
Insertion frequency (insertions/kb)
Castaway
Gaijin
Ditto
Stowaway
Self-insertions
0-0.5 kb
0.5-1.0 kb
1.0-2.0 kb
2.0-5.0 kb
Insertions into flanking sequences

Discussion

Here we report the characterization and quantification of MITE multimers in maize and rice. Although MITEs multimers were first discovered in maize, limited genomic sequence precluded further analysis of these multimers. However, the high density of MITEs in the rice genome (Bureau et al., 1996; Mao et al., 2000; Tarchini et al., 2000) coupled with the availability of large amounts of genomic sequence facilitated a more comprehensive analysis of multimers in rice and has led to the following conclusions: (1) MITEs are numerically the most abundant transposable elements in the rice genome (one MITE per 4.5 kb), (2) over 10% of rice MITEs are part of multimers thus suggesting a preference for MITE insertion into MITEs, (3) an insertion preference is displayed by some, but not all MITE families, (4) for the *Castaway* and *Gaijin* families, this preference is due to a high frequency of self-insertions. In contrast, *Ditto* elements are targeted by many element families, (5) the frequency of MITE insertions into class 1 or other class 2 elements is surprisingly low, and (6) based on our analysis of 30.2 Mb rice sequences, nested MITE multimers arise from independent insertion events.

Self-insertion preference for some MITE families:

As calculated in Table 2.2, the insertion frequency of all MITEs into other MITEs is slightly higher than the average value into the whole genome. However, there is a 3-fold variation in the frequency of MITE insertions into MITEs when individual families are examined (Table 2.4). More significantly, self-insertions comprised a major part of the multimers for several families. For *Castaway*, *Gaijin* and *Stowaway*, self-insertions account for two thirds of all insertions. These data indicate that the preferential insertion of MITEs into MITEs that is displayed by some families, can, to a great extent, be

attributed to self-insertions. One exception is the *Ditto* element. Among the rice MITEs, *Ditto* elements are frequently targeted by various types of elements, including other *Ditto* elements. In addition to being targeted 53 times by 12 families of MITEs, we detected five cases of insertions by four different LTR retrotransposons and 22 examples of MITEs inserted in adjacent (with an intervening TSD) sequences.

Composite elements, arising from self-insertion, have been previously reported in maize where double *Ds* and *Ac* elements were shown to be responsible for chromosome breakage and more complex rearrangements (McClintock 1949; Courage-Tebbe et al., 1983; D**ö**ring and Starlinger, 1984; Weck et al., 1984; D**ö**ring et al., 1989; Michel et al., 1994). It was later hypothesized that chromosome breakage resulted from aberrant transposition of composite or adjacent *Ds* elements (English et al., 1993; Weil and Wessler, 1993). In contrast to the composite *Ds* elements which are still capable of transposition, the uniqueness of each MITE multimer suggests that self-insertion creates an inactive composite element. Inactivating self-insertions of the *Tp1* element of *Physarum polycephalum* have been observed previously (Rothnie et al., 1990). It has been proposed that a preference for inactivating self-insertions minimizes deleterious effects on the host by providing a safe haven for insertion while simultaneously limiting the overall transposition frequency (Rothnie et al., 1990).

Regional vs. self-insertion preference

Previous studies indicate that some MITE families preferentially insert into genic regions (Mao et al., 2000; Zhang et al., 2000). A preference for genic regions has also been observed for the maize class 2 families *Ac*/*Ds* and *Mutator* (Chen et al., 1992; Cresse et al., 1995). Regional preferences have been demonstrated for many elements in

a wide variety of species. For example, yeast *Ty5* elements integrate preferentially into regions of silent chromatin at the telomeres and the mating loci (Zou et al., 1996), and for P elements, euchromatic sites, especially 5' regions of genes, are more often targeted than heterochromatin (Berg and Spradling, 1991; Liao et al., 2000).

Irrespective of the mechanism responsible, an element with a regional preference is more likely to have a higher frequency of self-insertion than an element with no such preference. If the regional preference is the major factor leading to a high self-insertion frequency, comparable insertion frequencies are expected into elements and into their flanking genomic sequences. The availability of 30.2 Mb of rice sequence allowed us to test this assumption (Figure 2.5). For *Castaway*, *Gaijin* and *Ditto*, the self-insertion preference is more likely to be caused by the targeting of preexisting elements than by a regional preference. In contrast, *Stowaway* elements show no significant difference between insertion into preexisting elements or into flanking DNA, thus suggesting that the high ratio of self-insertions result from a regional preference. Alternatively, the presence of one *Stowaway* element may alter the flanking DNA in some manner, thereby creating a better target for future insertions. A similar effect was observed for the *in vitro* transposition of the *C. elegans Tc1* element (Ketting et al., 1997). Interestingly, *Stowaway* elements, like *Tc1*, use TA dinucleotide targets.

The difference between *Stowaway* and the three other MITE families may indicate distinct integration mechanisms for different MITE families. Like the *Tourist* elements in maize, *Castaway*, *Ditto*, and *Gaijin* all create a 3-bp target site duplication (TSD) upon insertion (Bureau et al., 1996). More importantly, the TIR of *Castaway*, *Ditto*, and *Gaijin* are all related to that of *Tourist* elements in maize, suggesting they may

belong to the same superfamily. In contrast, *Stowaway* elements appear to belong to another superfamily based on their TIR and TSD (Bureau and Wessler, 1994b). Therefore, it is likely that these two superfamilies rely on distinct sources of transposases.

Target site preference in maize multimers

The potential to form secondary structures has been noted for several MITE families since the discovery of the *Tourist* family in maize (Bureau and Wessler, 1992; Izsvák et al., 1999). Given the occurrence of multimers among maize *Tourist-B2* elements, we hypothesized that secondary structure might play a role in targeting. Consistent with this notion is the inability to detect MITE multimers involving two other maize MITE families (*Hbr* and *mPIF*) lacking significant secondary structure (N. Jiang, Q. Zhang, X. Zhang and S.R. Wessler, unpublished data). However, in the rice genome multimer formation does not correlate with the potential to form significant secondary structure. In rice, the MITEs that sustained most insertions, *Castaway* and *Ditto*, are those without significant secondary structures (Bureau et al., 1996). In contrast, *Stowaway* elements usually have significant secondary structure but do not show a targeting bias. However, these data cannot rule out the possibility that small, local stem loops, like the14 bp palindrome targeted by P elements (Liao et al., 2000) may influence targeting of MITEs.

The analysis of MITE multimers in rice was also prompted by the discovery of non-random insertion sites among *Tourist* multimers in maize (Figures 1, 2, Table 2.2.1). The 10 bp periodicity observed for *Tourist* multimers is reminiscent of the integration of human immunodeficiency virus (HIV). Integration of HIV *in vitro* occurs preferentially into bent DNA where the major groove is on the exposed face of the nucleosome

(Pryciak and Varmus, 1992; Pruss et al., 1994). The 10 bp periodicity for *Tourist* multimers could be produced in a similar pattern, i.e., the transposition machinery only attacks major or minor grooves of the DNA double helix.

In rice, some "hot" spots for insertion were observed inside the sequence of some MITEs, and some of the insertion sites are roughly 10 bp apart. However, insertions that are not 10 bp apart are also observed. Due to the fact that rice MITEs that sustained most insertions are much bigger than maize *Tourist* elements (maize *Tourist* - 130 bp, *Ditto* - 244 bp, *Castaway* - 364 bp), the distribution of insertion sites appears to be sporadic within rice MITEs. Thus, more rice multimers need to be examined in order to determine whether or not the 10-bp pattern is statistically significant. Alternatively, this feature may only belong to *Tourist* elements in maize.

To date, no autonomous element responsible for the transposition of MITEs has been available. The isolation of such elements and their associated protein(s) will ultimately facilitate biochemical analysis of the various levels of targeting exhibited by MITE families.

A deficiency of MITE insertions into non-MITE elements: targeting preference or temporal differences in amplification?

A surprising and dramatic conclusion of the data presented in Table 2.2 is that MITEs have inserted into MITEs 80 times more often than they have inserted into LTR retrotransposons and 32 times more often than they have inserted into other DNA elements (1 MITE insertion vs. 4 kb, 330 kb, 127 kb, respectively). In contrast, the frequency of insertion of LTR retrotransposons and other DNA elements into MITEs is

only slightly lower than the overall frequency of insertion of these elements into rice genomic DNA (1 insertion/17 kb of MITEs vs. 1 insertion/14 kb of genomic DNA).

Previous studies have noted a genic preference for maize class 2 elements including members of the *Ac*/*Ds* and *Mutator* families (Chen et al., 1992; Cresse et al., 1995). Differences in chromatin density and/or the extent of DNA methylation between gene rich and other regions of the genome have been proposed as possible target recognition mechanisms (Chen et al., 1992). A similar preference for genic regions has been demonstrated for members of the MITE families *Hbr* and *mPIF* (Casa et al., 2000; Zhang et al., 2000)(X. Zhang, N. Jiang and S.R. Wessler, unpublished data). In contrast, MITEs appear to be underrepresented in regions of the maize and barley genomes containing nested or clustered LTR-retrotransposons (Tikhonov et al., 1999; Dubcovsky et al., 2001). While MITEs may target gene rich regions by the same or similar mechanisms as other class 2 elements, the analysis of MITE multimers in rice provides at least two alternative explanations for the observed (skewed) distribution. Enrichment for MITEs in genic regions and their apparent absence from retrotransposon clusters or domains could reflect a self-insertion preference coupled with avoidance of retrotransposon targets. Alternatively, a dearth of MITE insertions into non-MITE transposons would also result if the bulk of MITE amplification occurred prior to the amplification of LTR retrotransposons and other class 2 elements. To unambiguously distinguish between these seemingly mutually exclusive hypothesis, it will be necessary to identify an active MITE system that can be exploited to experimentally determine MITE target preference(s). In the meantime we must rely on the comparative analysis of

related genomes to provide clues to the mechanisms underlying the observed distributions of TEs.

Materials and Methods

Plant material, DNA extraction and library construction

Maize lines B79 and B37 were obtained from the U.S. Department of Agriculture/ Agricultural Research Service Plant Introduction Station at Ames, Iowa. Maize line B73 and RILs from a cross between B73 × Mo17 were provided by Michael Lee (Iowa State University, Ames, IA). Maize line Spanco was provided by Andy Tull (University of Georgia, Athens, GA). Plant DNA was extracted as described (McCouch et al., 1988). The small insert genomic library from B79 genomic DNA was constructed as described (Zhang et al 2000).

PCR and gel electrophoresis

PCR was carried out as described (Bureau and Wessler, 1992) with annealing temperature ranging from 55˚C to 60˚C, depending on the primers. Sequences of primers are available on request.

To clone the flanking sequence of the *Tourist* trimer in Figure 2.2.1, B79 genomic DNA was digested with *Mse* I and ligated with adapters. The DNA was then amplified with a primer complementary to the adapter and primer Pb, which contains the sequence at the junction of (*Tourist*) *Zm3* and the *B2*-like element (Figure 2.2). To separate PCR products that resulted only from adapters and PCR products from the two primers, primer Pb was labeled with ^{33}P, and the PCR products were loaded on 6% denaturing

acrylamide-bisacrylamide gels and electrophoresed as described previously (Casa et al., 2000).

The two-step PCR assay described in Figure 2.3 involved amplification of genomic DNA with primer P1 and P2, followed by amplification of the PCR products with P2 and P3 (P3 was labeled with P^{33}). PCR products were resolved by PAGE, as described above.

Recovery of gel bands

DNA fragments were excised from radioactive gels by scratching the dried gel with yellow tips (Stumm et al., 1997, Elsevier Trends Journals Technical Tips online, http://tto.biomednet.com/cgi-bin/tto/pr), placing the tip in 20 ul PCR reaction mix with relevant primers for 1 min before discarding and reamplifying with the same cycling parameters as that of the original reaction. PCR products were resolved in 0.8% agarose gels, fragments were excised, purified (QIAquick, Qiagen, Chatsworth, CA) and cloned (TA cloning kit, Invitrogen). DNA templates were sequenced by the Molecular Genetics Instrumentation Facility (University of Georgia).

DNA sequences analysis

DNA sequence analysis (pairwise comparisons, multiple sequence alignments, sequence assembling and formatting) was performed with programs in the University of Wisconsin Genetics Computer Group (UWGCG) program suite (version 10.1) accessed through Research Computing Resources, University of Georgia.

<u>Retrieving sequences</u>

Completely sequenced rice BACs and PACs were retrieved from websites of different rice genomic projects including groups in the U.S.

(http://www.usricegenome.org/), Japan (http://rgp.dna.affrc.go.jp/), Korea (http://bioserve.myongji.ac.kr/ricemac.html), P.R China (http://www.ncgr.ac.cn/Ls/index.html) and Taiwan (http://genome.sinica.edu.tw/).

Screening for transposable elements

Transposable elements in rice sequences were searched with RepeatMasker (Smit and Green, http://ftp.genome.washington.edu/RM/webrepeatmaskerhelp.html). The grass repeats database in RepeatMasker was modified by adding sequences of other previously characterized transposable elements in maize and rice (References not listed in the results section: Hirochika et al., 1992, 1996; SanMiguel et al., 1996**;** Kumekawa et al., 1999; Ohtsubo et al., 1999) and new transposable elements identified in this study. New elements were found either by their similarity to known elements or by insertion into known elements. The rice genome sequences described above were used as query sequences in analysis with RepeatMasker using the modified grass repeats database at default settings. In the output of RepeatMasker, the annotation files display all of the matches and the position of matches between the query sequences and any of the sequences in the repeats database.

Identification of multimers

Potential multimers were first selected from the query sequences based on the distance between two elements in the annotation files. For example, if one element is flanked by another element on both sides, the two elements probably form a dimer. The sequence of potential multimers was further analyzed manually with programs in UWGCG. If the ends of one element were located inside the sequence of another

element, the elements were deduced to comprise a multimer; otherwise elements were deduced to be monomeric.

Calculations

The insertion frequency was calculated by dividing the number of insertions by the kb of available sequences. For individual MITE families, the amount of DNA equals the size of the consensus element multiplied by the number of elements. If the length of match was less than half of the consensus element, it was considered as half an element in calculating the amount of DNA. If a match was less than 30 bp, it was eliminated from consideration. The total amount of LTR retrotransposons was approximated by multiplying the number of elements and solo-LTRs by their average length, which are 6.9 kb and 1.8 kb, respectively. The total amount of DNA representing DNA elements was estimated similarly with an average size of 1.9 kb. The average size of LTR elements and other DNA elements was obtained by sampling an 880 kb region in chromosome 1 (71.8-73.5 cm).

In Figure 2.5, the length of flanking sequences was estimated by the number of elements $\times$ 2 $\times$ the range of flanking sequences, where 2 represents that, for each element, there are flanking sequences on both sides. For example, 2690 *Stowaway* elements were detected in the 30.2 Mb rice genomic sequence, and 359 *Stowaway* insertions were observed in the range of 1.0-2.0 kb from another *Stowaway* element. In this case, the total length of available sequences = $2690 \times 2 \times (2.0\text{-}1.0) = 5380$ kb, and the insertion frequency in this range of flanking sequences = $359 \div 5380 = 0.067$ insertion per kb. Since the purpose of the analysis is to test whether the high self-insertion frequency for some MITE families is caused by the targeting of preexisting elements or by a regional

preference, adjacent insertions (only one TSD between two elements) were not included. This type of insertions were not considered because it is not clear whether they are due to the targeting for preexisting elements or for flanking sequences.

Acknowledgments

We thank Arian Smit (The Institute for Systematic Biology, Seattle) and Phil Green (Washington University, St. Louis, MO) for providing RepeatMasker and cross_match programs, Zhirong Bao (Washington University) for valuable suggestions and discussions, Cedric Feschotte and Xiaoyu Zhang for critical reading of the manuscript, Alexander Nagel for communicating unpublished data, and Qiang Zhang and Liangjiang Wang for technical assistance. This study was supported by grants from National Institutes of Health, U. S. Department of Energy and the National Science Foundation to S.R.W.

[a] The major part of this chapter was published in Jiang, N and S. R. Wessler. 2001. *The Plant Cell.* 13:2553-2564. Reprinted here with permission of publisher.

References

Baker, B., Schell, J., Lorz, H., and Fedoroff, N. (1986). Transposition of the maize controlling element "*Activator*" in tobacco. Proc. Natl. Acad. Sci. USA **83,** 4844-4848.

Berg, C.A. and Spradling, A.C. (1991). Studies on the rate and site-specificity of P-element transposition. Genetics. **127**, 515-524.

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.

Bureau, T.E., and Wessler, S.R. (1994a). 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.

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

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

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

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*: characterization of the *Emigrant* family of elements. Plant J. **16,** 79-85.

Chen, J., Greenblatt, I.M., and Dellaporta, S.L. (1992). Molecular analysis of *Ac* transposition and DNA replication. Genetics **130,** 665-676.

Cheng, Z, Presting, G.G., Buell, C.R., Wing, R.A., Jiang, J. (2001). High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. Genetics **157**,1749-57.

Courage-Tebbe, U., Döring, H.P., Fedoroff, N.V., and Starlinger, P. (1983). The controlling element *Ds* at the *Shrunken* locus in *Zea mays*: structure of the unstable *sh-m5933* allele and several revertants. Cell **34,** 383-393.

Cresse, A.D., Hulbert, S.H., Brown, W.E., Lucas, J.R., and Bennetzen, J.L. (1995). *Mu1*-related transposable elements of maize preferentially insert into low copy number DNA. Genetics **140,** 315-324.

Döring, H.P., Nelsensalz, B., Garber, R., and Tillmann, E. (1989). Double *Ds* elements are involved in specific chromosome breakage. Mol.Gen.Genet. **219,** 299-305.

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

Dubcovsky, J., Ramakrishna, W., SanMiguel, P.J., Busso, C.S., Yan, L., Shiloff, B.A. and Bennetzen, J.L. (2001). Comparative sequence Analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol. **125**, 1342-1353.

English, J., Harrison, K., and Jones, J. (1993). A genetic analysis of DNA sequence requirements for *Dissociation* State I activity in tobacco cells. Plant Cell **5,** 501-514.

Feschotte, C., and Mouchès, 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.

Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., Yamamoto, T, Lin, S.Y., Antonio, B.A., Parco, A., Kajiya, H., Huang, N., Yamamoto, K., Nagamura, Y., Kurata, N., Khush, G.S., Sasaki, T. (1998). A high-density rice genetic linkage map with 2275 markers using a single F2 population. Genetics **148**, 479-94.

Hirochika, H., Fukuchi, A. and Kikuchi, F. (1992). Retrotransposon families in rice. Mol. Gen. Genet. **233**, 209-216.

Hirochika, H., Sugimito, K., Otsuki, Y., Tsugawa, H. and Kanda, M. (1996). Retrotransposon of rice involved in mutations induced by tissue culture. Proc. Natl. Acad. Sci. USA **93**, 7783-7788.

Izsvák, 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.

Ketting, R.F., Fischer, S.E.J., and Plasterk, R.A. (1997). Target choice determinants of the *Tc1* transposon of *Caenorhabditis elegans*. Nucl. Acids Res. **25,** 4041-4047.

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

Kumekawa, N., Ohtsubo, H., Horiuchi, T. and Ohtsubo, E. (1999). Identification and characterization of novel retrotransposons of the *gypsy* type in rice. Mol. Gen. Genet. **260**, 593-602.

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

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

Liao, G.C., Rehm, E.J., and Rubin, G.M. (2000). Insertion site preference of the P transposable element in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **97,** 3347-3351.

Mao, L., Wood, T.C., Yeisoo, 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.

McClintock, B. (1949). Mutable loci in maize. Carnegie Inst. YRBK. **48,** 142-154.

McCouch, S.R., Kochert, G., Yu, Z.H., Khush, G.S., Coffman, W.R., and Tanksley, S.D. (1988). Molecular mapping of rice chromosomes. Theor. Appl. Genet. **76,** 815-829.

Michel, D., Salamini, F., Motto, M., and Döring, H.P. (1994). An unstable allele at the maize *opaque2* locus is caused by the insertion of a double *Ac* element. Mol.Gen.Genet. **243,** 334-342.

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

Ohtsubo, H., Kumekawa, N. and Ohtsubo, E. (1999). *RIRE2*, a novel *gypsy*-type retrotransposon from rice. Genes Genet. Syst. **74**,83-91.

Oosumi, T., Belknap, W.R., and Garlick, B. (1995). *Mariner* transposons in humans. Nature **378,** 672.

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.

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

Pruss, D., Reeves, R., Bushman, F.D., and Wolfe, A.P. (1994). The influence of DNA and nucleosome structure on integration events directed by HIV integrase. J. Biochem. **269,** 25031-25041.

Pryciak, P.M., and Varmus, H.E. (1992). Nucleosomes, DNA-binding proteins and DNA sequence modulate retroviral integration target site selection. Cell **69,** 769-780.

Rothnie, H. M., McCurrach, K.J., Glover, L.A., and Hardman, N. (1991). Retrotransposon-like nature of *Tp1* elements: implications for the organisation of highly repetitive, hypermethylated DNA in the genome of *Physarum polycephalum*. Nucleic Acids Res. **19**, 279-286.

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

Smit, A.F.A. and Riggs, A.D. (1995) *Tiggers* and other transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA **93**, 1443-1448.

Song, W.Y., Pi, L.Y., Bureau, T.E. and Ronald, P.C. (1998). Identification and characterization of 14 transposon-like elements in the non-coding regions of the members of the Xa21 family of disease resistance genes in rice. Mol. Gen. Genet. **258**, 449-56.

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.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., and Rafalski, A. (2000). The complete sequence of 340 kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell **12,** 381-391.

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.

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.

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

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

Weck, E., Courage, U., Doring, H.-P., Fedoroff, N., and Starlinger, P. (1984). Analysis of *sh-m6233*, a mutation induced by the transposable element *Ds* in the sucrose synthase gene of *Zea mays*. EMBO J. **3,** 1713-1716.

Weil, C.F., and Wessler, S.R. (1993). Molecular evidence that chromosome breakage by *Ds* elements is caused by aberrant transposition. Plant Cell **5,** 515-522.

Wessler, S.R., and Varagona, M. (1985). Molecular basis of mutations at the *waxy* locus of maize: Correlation with the fine structure genetic map. Proc. Natl. Acad. Sci. USA **82,** 4177-4181.

Yoder, J.I. (1990). Rapid proliferation of the maize transposable element *Activator* in transgenic tomato. Plant Cell **2,** 723-730.

Yuan, Q., Quackenbush, J., Sultana, R., Pertea, M., Salzberg, S.L. and Buell, C.R. (2001). Rice bioinformatics. Analysis of rice sequences data and leveraging the data to other plant species. Plant Physiol. **125**, 1166-1174.

Zhang, Q., and Kochert, G. (1998). Independent amplification of two classes of *Tourists* in some *Oryza* species. Genetica **101,** 145-152.

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 (Hbr)* into genic regions of maize. Proc. Natl. Acad. Sci. USA **97,** 1160-1165.

Zou, S., Ke, N., Kim, J.M., and Voytas, D.F. (1996). The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev. **10,** 634-645.

Supplemental data: sequence of the new MITEs identified in this study

Each element is described in the order of name, target site duplication (TSD), and sequence.

BUHUI TNA
CGGCTTGTTTGGTAACTCGAGGGAAGGGGATTGGGAGTTTACACGAGGGAATTGATGATGAG
ATTAAGATAGGAATTTGATTTTTAATCCCAATATCTTGTTTGGTAAAGGTGATGGAAATTAAT
AGGGAGTTTAGGAGATTAGGTCATTAATAAAAAAAGGATGGCTGAGATTCGTTTAAACAACG
TAAAGGAGGAATTCCCTCCCAATTACCTGTCCCTACTCAAGTATTGAAAAGGAGGGAGTTCCT
TCCTCAATTCTCCATCCCCATCCCACTAAAATTCTCATGCCTCCCATCCAAACAAAAACACTTA
ATAGCATCATCTCTCTAAACTCCCAATCCCTGCTAAAAACTCCCTCCAACCAAACAGGCCG
CASIN TNA
GGTCTTGTTTGGATCATTTAGTCTCTAAACTAAAAACTTTAAACTTTAGTCTCCTACTATTTAG
ATCCAGAGACTAAAAGGGACTAAAGGTATGTGGATAGAGGGAGAGAGAAGAGAGAGGCTGC
CACTTTCTTCCTTACAAATATCTCCATAAGGGATTAATGGATTTTAATCCTAACATACACTTTT
AGTTAATTTTAGTCATCTTGTTTAAATTCTTAGAGACTAAAACAGATTAAAATGAAGGGACTA
ATGGATTAGACCCATGAACGAAACAGGACC
CENTRE NNN
GGGGCTGTTTGGTTGATGGCTATAAGTTGCCACACTTTGCCATACTTGAGGTTAGACAAGTTT
GACCATGTTAGGTGAGTGTTTGGTTGGCTTCCACAAGTGTGGCAAGATTTCTCTTTTGTAATCT
AGGTCCCACATGTAAATGTCTCTAAAAAGTGTGGCATGGTTTCATTTTGTTGGCTAGAGTGTG
GCTTGGATTTTGATGGCCTCACCTTTGACAAGTTTGGCTAAGAAAGCATGGTAAACCTAGGCA
ATGTTAGTATGTGAACCAAACAGCCCC
DELAY NNN
AGATTGCGCGTAGCGATGCAAGTGGACCGACCCGCAAGTTTACTTATAGATCAAATATATGAT
AATTCATGAATTTTTCGTATCGTGGCCCGGTTTGCCGTGTCTAGATATACAAATGGATCGATCT
ATAGACCCATTTATAAGTCAAATTAATAGATAACTTGCTAATCAAATCCACGGGTTACGCACT
AATTTTTCCTATAAACGGGTTCATGGATAAGCCTACTTGCAAATCAACCTC
ECR NNN
GAGTAAATTTCACAAAACTACATATTTTGTGGTCCAAGTTGTAGAAAACCACACGCACTTTGA
CACTTGGCACTTGAGTACATATATTTTGGTCGTGTAGTTTCACAAAACTACACTATCGATGTAC
AGATTCACCCGCGAGATGATGTGGCTATTTCACCTAGGACGAGGACATGGCATCCTATTACAG
CCATCGCACGAAATTAATATGATGCCACGTCCTCATACTAGATGAAACAACCACGTCGTCATA
TCACAAGTGAATCCATTCATCGATAGTGTGGTTTTGTGAAACTAAACTACCAAAATATATGTA
CTTAAGTGCCAAGTATCAAAATATGTGTGGTTTTCTGCAACTTGAACCATAAAATATGTAGTTT
TGTGAAATTTACTC
HELIA TNA
GGGCATCCACAATGTGCATTTAAAGTGGGTAGTAAGCAATAAAATACTACTACTAACAGGTTG
TATATATTGTAAGAGTAGTAAGTAACTAATACCAGGTAATATGCTTACTACCTAGTCATAAGG
AATAAACAAATAATTTCTCTCATCTTTCTCTCATGGCTAGTGATGGTGGGGTCCAGCTTACTAC
CCACTTTTGTTCTTTCACCATTGTGGCTACAACAACAGCTAATACCCACTGATGCAGGGCTCAC
TCTTATACACAAAGTAGGTAGTAAACATTGGTGATGCCC
ID-2 TNA
GAGCAAGGTCAATAGTGTAGCTAGCAGTAGGCTCTAAAATTTTGTCATGTCTCAGGTAGAGCT
AACTTAAAGAGCAATTTATATAATAAACTAACTACAAGTTAGTTATAAGCTACATTGGTTCAA
ACGAATTAACAAGCGGTAGGACCCACAAACGAATTTTTTTTATTGTATCTCCTAGCCGATACA
GCGTGAGGCGCTGTGCAAGCACCGGCTTCTTTCTCAATATGTGATTTTCTTTCCTTAATCTCGT
TTTAGGATGATTGATAAGTTTTAGAGCCTTCCTAATAGCACTATTGTACTTGCTC
ID-3 TNA
GGGCAAGTTTTATGATAGAGGTAGTGGTCACATCTAATTTTATCCAAATCATCTACATATTAAT
TTAGAGGTAATACATACAATAGATTGCATCTACCTATACTACACCCAATACATTTACTACCAA
ACAACTTCTCTCTTTCTCTCTTATTTCTTGAAATTGGTGTGGAGGTTGCCTCTTGCATGAGGGG
TGGCTGCTCTCCTCTCTCTTCTTTTCTCTCCTCCACATCATCATATGACTCTACATGGCATCTTA
GAATTAGTGATAAAGGGCTCATAGTACTTGCCC
ID-4 TNA

GGGCATCCACAATGTAGTGTAAAAGTAGTCCATAAGCAATAGAATATTCCATTCAGCCACCTA
TTACCATTGTGCAACTAGTGCATATATAGAAAATAGTAAAAGATATCCACATGCATATGGACT
ACCTATATGGAATAAATTATATTATCTATCTCTACTCCTCTCTCTTGTCCTAATGATACCTTGTA
TCCATGTACATTGTGGAGATTGCTATAGGAAAAAACTATTTATGTGGGTCCTATCTATATAGA
TCACTTTACCCATATATATTGATGATGCCC
LIER TNA
GGGCATCCACAATGAGGCCATAAGAATTAGCCTTATGACTTTTTTTTTATGCTATATACATTGTG
GATTTTTTTATTCTAAATACATTGTGGACATGGACATAAGACATTAAATGCTTAAAAACAACT
CTAAGCTTATGGTTGAACTTAAGGGATTAAAAAATGATATCTACTAGGGCCTATCTGCATATG
GTTGCGGATTGGTGGTATATTTCTTCTTTGATGGGACCTGGCTTATTTCTCACCCAAAGCCATA
ACCATTGTAGGGATAGATATAGTTAAAGCCATCTGCACATGGGCCCACCCTATTACTCATTCA
GCAACCAACATTGAGGTTGCCC
STOLA TNA
GGGCAGTGCCAACCCATAATGTCTATAAGTAGTGTCTATGGTGCCATGTCAATAAGACATCAC
AATAGAAACTACACTCTACAACCTATGGTTTCTTAAAGTGGGCCATTAATAAATACATCATCT
CTCTTCTCTACCAATCATATTTATTCTTCATCTATTATGAAGACACTATTCTCTCCTAATGCAAA
ACTTGATAGTGTCTAATGCATAGGTTCTCGCGTTGAAGCTGTGTCTTGCATGAGACCCAATTTC
TTCCTCTCTTCACTCTCTCTCTTAATTAATATAGTGCCACATAAGCTAAAAGTCTTACATGGCA
ATGTAGTTAATGTTATAGACACCATCCTAGATGAAGGGTTGGGACTGCCC
STONE NNN
GGGTGCGTTCAGAAGAGCTGGCTGTGACAGCAGCGGCGCTGCAGCTACATTACTCAACAAGC
AAGATTATGCTAGGAGAGCTGCAGCGCTGTAGCTGTGCAGCCACAGTCCAGCTGAACCGAAC
AGACCC
SUSU TNA
GGCCCTCTTTATTTAGGTTTATAAGCTAATTTATAAGTCGAAAAGTCTAAGCCTAAACAAACA
AGCAGCTTTTCCGTTTGGCTTTTTTAAAGCCATAAGCCACTCTAACACTATTAAGCTAAAAGCT
AGGTTGGAGAAGCTTTTTTGGCTTATGTGAGGTAGATGTATGACTCAACTACTAAACTTAGGA
CATTAAATCCACCGGCTTATAAATCATATAAGCCAATAAGCTGACTTAAAAATCTAGGCCAAT
AAGCCTAAGCCTAAACAAAGAGGGCC
WUJI TNA
GCCATATACATTGTGGAAAGTAGTATTAGCTTAAGGTGCACCTTATGCTTAAGGTCTACCACA
AGCAAAAAGAGTATATTTTTCTCTCTCTCATTTACTTCGTGAACAAGAGATGAAACAATTCCTT
TTCAATTCTATGTGGGCCCATCTTATGACTAAATTCTAATCATAACATTGTGAGTACAGTTTTA
ACAATGGTTCACTATTCATAGGGCCCACCTTAAGATTACACTACACCACAAATATTGAGGATG
CCC
YOUREN TNA
GGCCATTCCCAACCCAATGACTAGGATGGTGTCCATAGCATTAAATAAGCTGCCACCTAGGAT
GAAAAATGATGTGGCAAGTGAATAAATGAGGAAAGAGAAGGAAACCATGTCTTGCATGAGA
CATGGTTTCTACACAACATCCAAGACATCATGTGAGATAAGTAGCATTAAATTTAAGTATGGA
ATAGTGGTGTTTGCATTGGAAGAGTAGTGTCTAGTACTAGTTTCTTGATGATGTGGAGTTTATG
GAAACTATGTCTAGTGTCTTGGGTTGGGAATGGCC

CHAPTER 3

DASHENG: A RECENTLY AMPLIFIED NONAUTONOMOUS LTR ELEMENT THAT IS A MAJOR COMPONENT OF PERICENTROMERIC REGIONS IN RICE[1]

[1] Jiang, N., Bao, Z., Temnykh, S., Cheng, Z., Jiang, J., Wing, R.A., McCouch, S. R.and S.R. Wessler. 2002. *Genetics*. 161:1293-1305. Reprinted here with permission of publisher

Abstract

A new and unusual family of LTR elements, *Dasheng*, has been discovered in the genome of *Oryza sativa* following database searches of approximately 100 MB of rice genomic sequence and 78 Mb of BAC-end sequence information. With all of the cis-elements but none of the coding domains normally associated with retrotransposons (e.g. *gag*, *pol*), *Dasheng* is a novel nonautonomous LTR element with high copy number. Over half of the ~1000 *Dasheng* elements in the rice genome are full-length (5.6 – 8.6 kb), and 60% are estimated to have amplified in the past 500,000 years. Using a modified AFLP technique called transposon-display, 215 elements were mapped to all twelve rice chromosomes. Interestingly, more than half of the mapped elements are clustered in the heterochromatic regions around centromeres. The distribution pattern was further confirmed by FISH analysis. Despite clustering in heterochromatin, *Dasheng* elements are not nested, suggesting their potential value as molecular markers for these marker-poor regions. Taken together, *Dasheng* is one of the highest copy number LTR elements and one of the most recent elements to amplify in the rice genome.

Introduction

Transposable elements (TEs) have been divided into two classes, class 1 or RNA elements and class 2 or DNA elements. An RNA intermediate and a replicative mechanism of transposition are involved in the transposition of class 1 elements (LEWIN 1997). RNA elements can be further divided into several groups, including long terminal repeat (LTR) retrotransposons, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). RNA elements are capable of attaining very high copy numbers because the element-encoded mRNA, and not the element itself, forms the transposition intermediate.

LTR retrotransposons make up the largest fraction of most plant genomes (KUMAR and BENNETZEN 1999). The LTRs usually contain the initiation and termination sites of a transcript that encodes at least two genes, *gag* and *pol*. The products of these genes are involved in the different steps of retrotransposition, including reverse transcription and integration (LEWIN 1997; Figure 3.3.1). Immediately internal to the LTR is the primer binding site (PBS) and the polypurine tract (PPT). Both are important cis-elements that are necessary for the initiation of synthesis of element DNA from the RNA intermediate. LTR elements are classified into two types based on the order of their encoded genes: there are *Ty1/copia* and *Ty3/gypsy* elements (XIONG and EICKBUSH 1990). Both are prevalent in plant genomes (VOYTAS *et al*. 1992; SUONIEMI *et al*. 1998).

Differential amplification of LTR retrotransposons has been shown to be largely responsible for the C-value paradox in members of the grass clade (CHEN *et al*. 1997; SANMIGUEL and BENNETZEN 1998; DUBCOVSKY *et al*. 2001). The C-value paradox refers

to the lack of correlation between the genome size and the biological complexity of an organism (THOMAS 1971). For example, rice (*Oryza sativa*) and barley (*Hordeum vulgare*) have roughly the same number of genes and largely conserved gene order (MOORE *et al*. 1995; DUBCOVSKY *et al*. 2001). The 11-fold difference in the size of their genomes (430 vs. 4800 Mb) is due, in part, to the fact that retrotransposons comprise more than half of the barley genome and only 14% of the rice genome (VICIENT *et al*. 1999; TARCHINI *et al*. 2000). With the International Rice Genome Sequencing Program (IRGSP) scheduled for completion in less than a year (MYERS 2001), new insights about the identity and frequency of different TE families will emerge.

Despite its small genome, rice is still a model organism for the study of transposable elements. The genome of *Oryza sativa* contains all of the major types of elements found in the larger grass genomes including retrotransposons, MITEs and other DNA elements (BUREAU *et al*. 1996; MAO *et al*. 2000; TARCHINI *et al*. 2000; TURCOTTE *et al*. 2001). Furthermore, the availability of several well-characterized wild relatives provides the material necessary to analyze the impact of TEs on genome evolution and speciation. *Oryza sativa* is comprised of two cultivated subspecies (*indica* and *japonica*) with thousands of diverse cultivars distributed worldwide. The genus *Oryza* has more than twenty species whose evolutionary relationships have been the subject of several phylogenetic analyses (UOZO *et al*. 1997; GE *et al*. 1999; SHARMA *et al*. 2000).

In this study, database searches of about 100 Mb of rice genomic sequence and 78 Mb of BAC end sequence led to the identification of a new and unusual family of LTR elements called *Dasheng*. *Dasheng* is a very recently amplified family of 800 to 1300 nonautonomous elements, making it one of the most recently amplified and highest copy

number families in rice. The family also includes about 16% solo LTRs. Like many other high copy number LTR elements, *Dasheng* elements are concentrated in the gene-poor pericentromeric regions of the chromosomes, which might be the reason for its success in the small genome of rice. The availability of large amounts of genomic sequence and an almost completely assembled chromosome 1 has allowed us to address questions regarding the distribution and timing of insertion events and to test models that explain the formation of solo LTRs.

Materials and Methods

Plant material and DNA extraction

A doubled haploid (DH) mapping population (GUIDERDONI *et al*. 1992; HUANG and KOCHERT 1994) was used in conjunction with an existing SSR mapping dataset (TEMNYKH *et al*., 2001) to map *Dasheng* elements. This population consisted of a subset of 96 doubled haploid lines derived via anther culture from the inter-subspecific cross between IR64 (*O.sativa* ssp. *indica*) and Azucena (tropical *japonica*). Other rice cultivars and wild species were obtained from the McCouch lab (Cornell University) and Gary Kochert (University of Georgia). Plant DNA was extracted as described (MCCOUCH *et al*. 1988).

Genetic mapping

Transposon display was performed as described in CASA *et al*. (2000) to generate segregation patterns in the DH population with the following modifications. The element-specific primers were derived from the LTR sequence of *Dasheng* and the reaction was performed with rice DNAs. The final annealing temperature for selective amplification was 58°C with ^{33}P-labled *Dasheng* primer. Sequences of primers are available upon request. DNA fragments from transposon display were excised and cloned as described (CASA *et al*. 2000). DNA templates were sequenced by the Molecular Genetics Instrumentation Facility (University of Georgia).

The gel images of transposon display with DNAs from the DH mapping population were scored manually for presence/absence of polymorphic bands corresponding to *Dasheng* elements. The *Dasheng* markers were integrated into the SSR framework map using the Kosambi mapping function and MapMaker 3.0 software (LANDER *et al*., 1987). Markers with a ripple of LOD > 2.0 were integrated into the framework maps and those mapping with LOD < 2.0 were assigned to the most likely intervals.

FISH analysis

FISH analysis was performed as previously described (JIANG *et al*. 1995) using Nipponbare and *indica* cultivar Zhongxian 3037. The *Dasheng* probe (Figure 3.3.1) was labeled with Biotin-16-UTP and detected using a fluorescein isothiocyanate (FITC)-conjugated anti-biotin antibody (Vector Laboratories, Burlingame, CA). Propidium iodide in an antifade solution was used to counterstain the chromosomes. Chromosome

and FISH signal images were captured using a SenSys CCD (charge-coupled device) camera (Photometrics, USA) and analyzed using IPLab Spectrum software (Signal Analytics, USA).

DNA sequence analysis

DNA sequence analyses (pairwise comparisons, multiple sequence alignments, sequence assembling and formatting) were performed with programs in the University of Wisconsin Genetics Computer Group (UWGCG) program suite (version 10.1) accessed through Research Computing Resources, University of Georgia.

Identification of repetitive sequences from BAC ends

All sequences in the rice BAC end database (O.sativa cv Nipponbare) were downloaded from the website of Clemson University Genome Institute (CUGI) (http://www.genome.clemson.edu) for the initial analysis (Aug. 1999). An all versus all comparison was performed with the sequences using WUBLASTN (http://blast.wustl.edu) with parameters M=5 N=-11 Q=22 R=11 –kap E=0.001 -hspmax 5000). Groups with highest intra-group similarities (> 95%) were further characterized with BLAST search in the NCBI server (http://www.ncbi.nlm.nih.gov).

Identifying transposable elements in genomic sequence

The sequences of rice BACs and PACs were downloaded from the websites of different rice genome projects including groups in the U.S. (http://www.usricegenome.org/), Japan (http://rgp.dna.affrc.go.jp/), Korea (http://bioserve.myongji.ac.kr/ricemac.html), P.R. China (http://www.ncgr.ac.cn/Ls/index.html) and Taiwan (http://genome.sinica.edu.tw/).

Completely sequenced PACs or BACs and those in annotation and finishing (phase 2) were used as query sequences to search for transposable elements with RepeatMasker (A. Smit and P. Green, http://ftp.genome.washington.edu/RM/webrepeatmaskerhelp.html) as described (Jiang and Wessler 2001).

Copy number determination

The copy number of *Dasheng* was estimated in three ways: (1) by blasting BAC ends using LTR sequence as a query. Using this method, copy number = matches in BAC ends × 430 Mb (rice genome size) ÷ the size of the BAC ends database (in Mb); (2) by probing a rice BAC library [derived from Nipponbare (Mao et al., 2000)] with a 500 bp fragment located between the third tract of direct repeats and the PPT (see Figure 3.3.1). Using this method, copy number of *Dasheng* elements = (# positive clones ÷ number of BACs screened) × 430 Mb ÷ average size of BACs (in Mb). The raw value, estimated to be 700 elements per haploid genome, was corrected for the number of solo LTRs (16%), BACs and PACs containing two or more elements (12% of the positive clones) and truncated elements (30%). The corrected copy number was 900 – 1300 (depending on the percentage of truncated elements detected); (3) by screening the genomic sequence with RepeatMasker followed by manual examination. The copy number = # elements in genomic sequences × 430 Mb ÷ total size of the genomic sequence screened.

The copy number of other rice LTR elements (elements reported previously and those identified in this study) was estimated by blasting the BAC end database and GenBank (NCBI BLAST server) with LTR sequences. Low score matches ($e > 10^{-30}$) from GenBank were checked manually to determine whether the matches represented the

element. The copy number for each element = matches for this element × 430 Mb ÷ total size of the rice genomic sequence in GenBank.

Phylogenetic analysis and aging of elements

LTR nucleotide sequences homologous to *Dasheng* and *RIRE2* were aligned using GCG (see above). Tree production and bootstrap analyses were performed using PAUP version 4.0. Sequence similarities and standard error were calculated with MEGA program (KUMAR *et al*., 2001). Full-length elements were aged (as in SANMIGUEL et al. 1998) by comparing their 5' and 3' LTR sequences. Kimura-2 parameter distances (K) between 5' and 3' LTRs of individual elements were calculated using MEGA. An average substitution rate (r) of 6.5×10^{-9} substitutions per synonymous site per year for grasses (GAUT *et al*. 1996) was used to calibrate the ages of the elements. The time (T) since element insertion was estimated using the following formula: T = K ÷ 2r. Fifty-percent consensus sequences were determined from group-specific alignments using the EMBL consensus sequence server (http://www.bork.embl-heidelberg.de/Alignment/consensus.html).

The distribution of *Dasheng* elements in genomic sequences

The distribution of *Dasheng* on chromosome 1 of Nipponbare was constructed according to the positions of PACs and BACs that contained *Dasheng* elements (http://rgp.dna.affrc.go.jp/). Estimates of physical: genetic distance and insertion frequency was based on the data provided by the RGP (Rice Genome research Program, Japan) (http://rgp.dna.affrc.go.jp/) at the time of analysis. DNA density for chromosomal

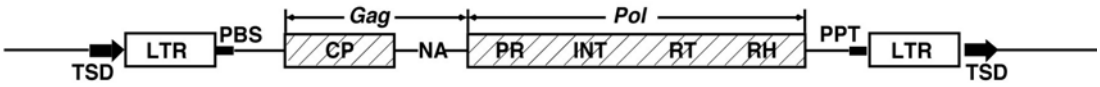
arms and pericentromeric regions was calculated from the total DNA of three contigs (20.2 – 34.5 cM and 40-50 cM in arms, 60-70 cM in pericentromeric regions) on chromosome 1. The borders of pericentromeric regions were defined as 15 cM from the center of the centromere on each arm. The position of the centromere was according to HARUSHIMA *et al*. (1998) and CHENG *et al*. (2001a). The remainder of the chromosome was defined as arms. Physical: genetic distance equals the physical length of DNA in base pairs divided by the map units covered. Insertion frequency equals the number of elements found in a certain region divided by the physical length of DNA in that region. The total amount of DNA was the size of all the clones minus overlap.

Results and Discussion

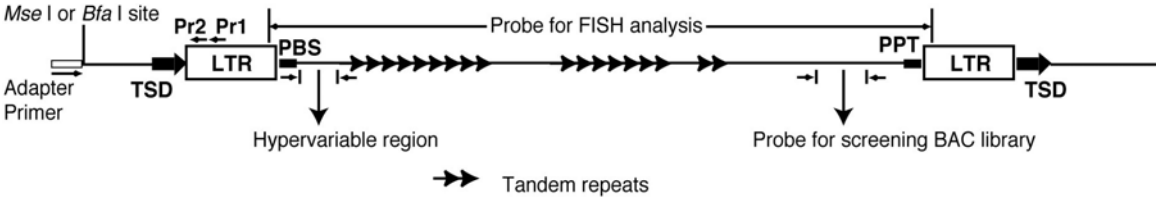
A nonautonomous LTR element with very high copy number

To identify repeat sequences that might be novel transposable elements, we performed an *all versus all* comparison with BAC end sequences of rice (*O. sativa* ssp. *japonica* cv. Nipponbare) (see Materials and Methods for details). Several groups of BAC ends were distinguished by their high within group sequence similarity (~95%). The sequence of each group was then used as a query to perform further searches in GenBank. Significant matches for all groups were found in an 8.6 kb segment of a PAC clone from rice chromosome 6 (GenBank accession no. AB023482). This region has the structural features of an LTR retroelement including a long terminal repeat (441 bp with 99.5% sequence similarity), an adjacent putative primer binding site (PBS) and

Figure 3.1. Comparison of *Dasheng* with a typical autonomous LTR retrotransposon. Coding regions are shown as slashed boxes. CP, capsid-like proteins; PR, protease; RT, reverse transcriptase; INT, integrase; RH, RNase H. The relative order of RT, RH and INT varies with different types of elements (see text). Other sequences indicated are: LTR, long terminal repeat; PBS, primer binding site; PPT, polypurine tract; NA, nucleic acid binding moiety. Arrows above the LTR of *Dasheng* indicate the positions of transposon-specific primers (Pr1 and Pr2) for transposon display. These would be used for PCR with the adapter primer shown (see Figure 3.2 and text).



A typical LTR element



Dasheng (5.6-8.6 kb)

polypurine tract (PPT), and a 5 bp target site duplication (TSD) flanking the LTR (Figure 3.1). The 441 bp LTR is related (65%-70% sequence similarity) to the LTR of *RIRE2*, a previously described *Ty3*/*gypsy* type LTR element in rice (OHTSUBO *et al*. 1999). In addition, the two elements also have similar PBSs and PPTs that only differ at 1 or 2 out of 15 nucleotides.

Despite having structural features of LTR retrotransposons, the 7.8 kb region between the LTR contains only very short ORFs with no similarity to known proteins. Instead, about 4 kb of this region is comprised of tandem repeats of an 89-90 bp unit (Figure 3.1). The other part of the internal region includes a hypervariable domain (of zero to 1.2 kb) located between the PBS and the first tract of tandem repeats (Figure 3.1). The lack of coding capacity suggests that this element is most likely nonautonomous .

The copy number of this element family (named *Dasheng*) was estimated in three ways (see below, Materials and Methods for details). Based on the prevalence of the LTR sequence in BAC ends [150 hits in 78 Mb of *Hind* III and *Eco*RI digested sequences ($e \leq 10^{-15}$)], we estimate that there are about 800 copies of *Dasheng* in the genome of cv. Nipponbare. To test whether the prevalence of *Dasheng* in BAC ends is representative of the rest of the genome, a BAC library of the cv. Nipponbare genome was screened with a *Dasheng* probe. This experiment led to a copy number estimate of 900 to 1300. In contrast, a search of about 100 Mb of publicly available assembled genomic sequence led to a copy number determination of 470 per haploid genome or approximately one element per Mb. The two to three fold difference in the values obtained from BAC screening and BAC end sequences versus genomic sequence may be due to the fact that the latter is biased toward gene-rich regions, whereas several LTR retrotransposon

families are enriched in pericentromeric regions of the genome (MILLER *et al*. 1998; LANGDON *et al*. 2000; NONOMURA and KURATA 2001; also see below).

The copy number of *Dasheng* was also compared with that of other LTR elements in rice. This was done by querying the BAC end and genomic sequence databases with LTRs from several high copy number rice elements previously described and elements identified in this study (see Materials and Methods). As with the searches using *Dasheng* sequences as queries, the results were inconsistent from one database to the other. The average values obtained (from BAC ends and from genomic sequence) were, in descending order of copy number, *Retrosat2* (1080) (GenBank accession no. AF111709), *Bajie* (identified in this study, 730), *RIRE4* (730) (KUMEKAWA et al. 1999), *SZ-19* (725) (identified in this study), *Dasheng* (635), *RIRE8* (620) (KUMEKAWA et al. 1999), *RIRE3* (510) (KUMEKAWA et al. 1999), *RIRE2* (420) (OHTSUBO *et al*. 1999), *RIRE9* (115) (LI *et al*. 2000; HAN *et al*. 2000) and *RCS1* (90) (DONG *et al*. 1998).

In a prior study, dot blot hybridization led to a copy number determination for the *RIRE2* family of 10,000 in IR36 (OHTSUBO *et al*. 1999). In contrast, we found that the number of hits using *RIRE2* sequences was no higher than that found for *Dasheng*. The striking discrepancy may be due to the presence of distantly related families, a frequent cause of copy number overestimation when employing hybridization methods (MEYERS *et al*. 2001).

The chromosomal location of *Dasheng* elements

Genetic mapping of *Dasheng* elements

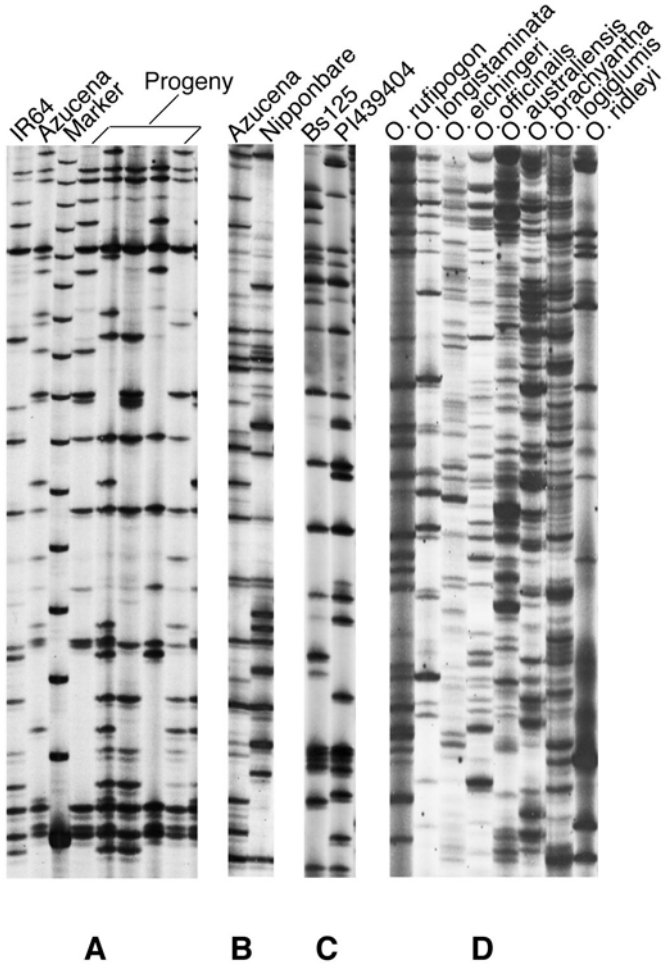
To determine the chromosomal distribution of *Dasheng*, family members were mapped using a technique called transposon display, which is a modification of the AFLP procedure that generates PCR products anchored in a transposable element and a flanking restriction site (WAUGH *et al.* 1997; VAN DEN BROECK *et al*. 1998; CASA *et al*. 2000). The number of fragments amplified in one reaction can be adjusted by adding extra bases to the adapter primer (so-called selective bases); fewer fragments will be detected with more selective bases (VOS *et al*. 1995). Transposon display has the added advantage of detecting solo LTRs since the transposon-specific primers are located within the LTR (Figure 3.1, Pr1 and Pr2).

Dasheng primers were designed so as not to recognize the related *RIRE2* elements. Insertion site polymorphism, as defined by the presence of a PCR product in one parent but not in the other, was high for the parents (IR64 and Azucena), varying from 76.3% to 84.3% for different adapter primer/enzyme combinations in this inter sub-specific cross (Table 3.1; Figure 3.2A). High levels of poly morphism were also detected within *indica* and *japonica* subspecies (Figure 3.2B and C), indicating that *Dasheng* elements can serve as a valuable marker system. Several wild species of rice were also tested with the same enzyme but they needed more selective bases (*Mse*I + T for *O. sativa*; *Mse*I + TG for wild species; Figure 3.2, A and D). The multiple fragments detected in the wild species indicate that *Dasheng* is also abundant in these genomes (Figure 3.2D).

Table 3.1. Polymorphism detected in the IR64 × Azucena mapping population

Primer/enzyme combination	Number of amplified fragments			% polymorphic
	Monomorphic	Polymorphic	Total	
Dasheng-Bfa I + A	8	43	51	84.3
Dasheng-Bfa I + C	14	45	59	76.3
Dasheng-Bfa I + G	9	47	56	83.9
Dasheng-Mse I + A	18	58	76	76.3
Dasheng-Mse I + G	9	36	45	80.0
Dasheng-Mse I + T	15	53	68	77.9
Total	73	282	355	79.4

Figure 3.2. Autoradiograph of *Dasheng* display with DNAs from *O. sativa* and other *Oryza* species. For all reactions, the transposon-specific primer was Pr2. (A) *Dasheng* display with DNAs from the IR64 × Azucena doubled haploid mapping population using adapter primer *Mse*I + T; (B) *Dasheng* display with DNAs from Azucena and Nipponbare, two *japonica* cultivars, using adapter primer *Bfa*I + C; (C) *Dasheng* display with DNAs from Bs125 and PI1439404, two *indica* cultivars, using adapter primer *Mse*I + A; (D) *Dasheng* display with DNAs from eight other *Oryza* species using adapter primer *Mse*I + TG.


IR64
Azucena
Marker
Progeny
Azucena
Nipponbare
Bs125
PI439404
O. rufipogon
O. longistaminata
O. eichingeri
O. officinalis
O. australiensis
O. brachyantha
O. coarctata
O. ridleyi
A
B
C
D

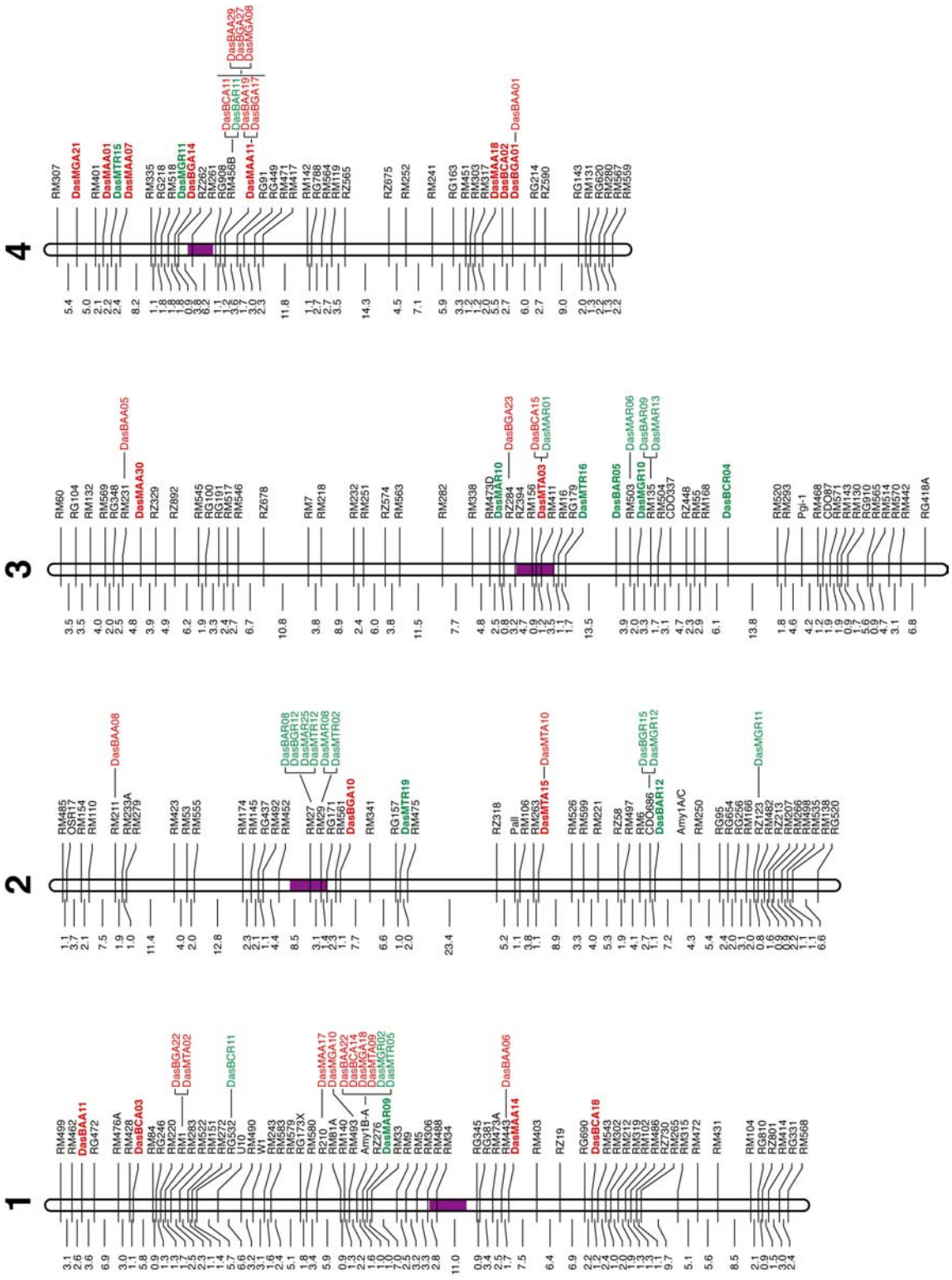
The chromosomal location of polymorphic bands was determined by integrating segregation patterns into a previously constructed framework map consisting of 432 SSR's (TEMNYKH *et al*., 2001). The map was based on a population of doubled haploid lines, derived from a cross between IR64 (*indica*) and Azucena (*japonica*). In this study, a total of 215 bands (128 from Azucena and 87 from IR64) from six primer-enzyme combinations were assigned to all 12 chromosomes (Figure 3.3). Cloning and sequencing of 20 bands provided confirmation that all fragments were amplified from element-containing loci. For this reason, the mapped bands will be referred as "*Dasheng* markers".

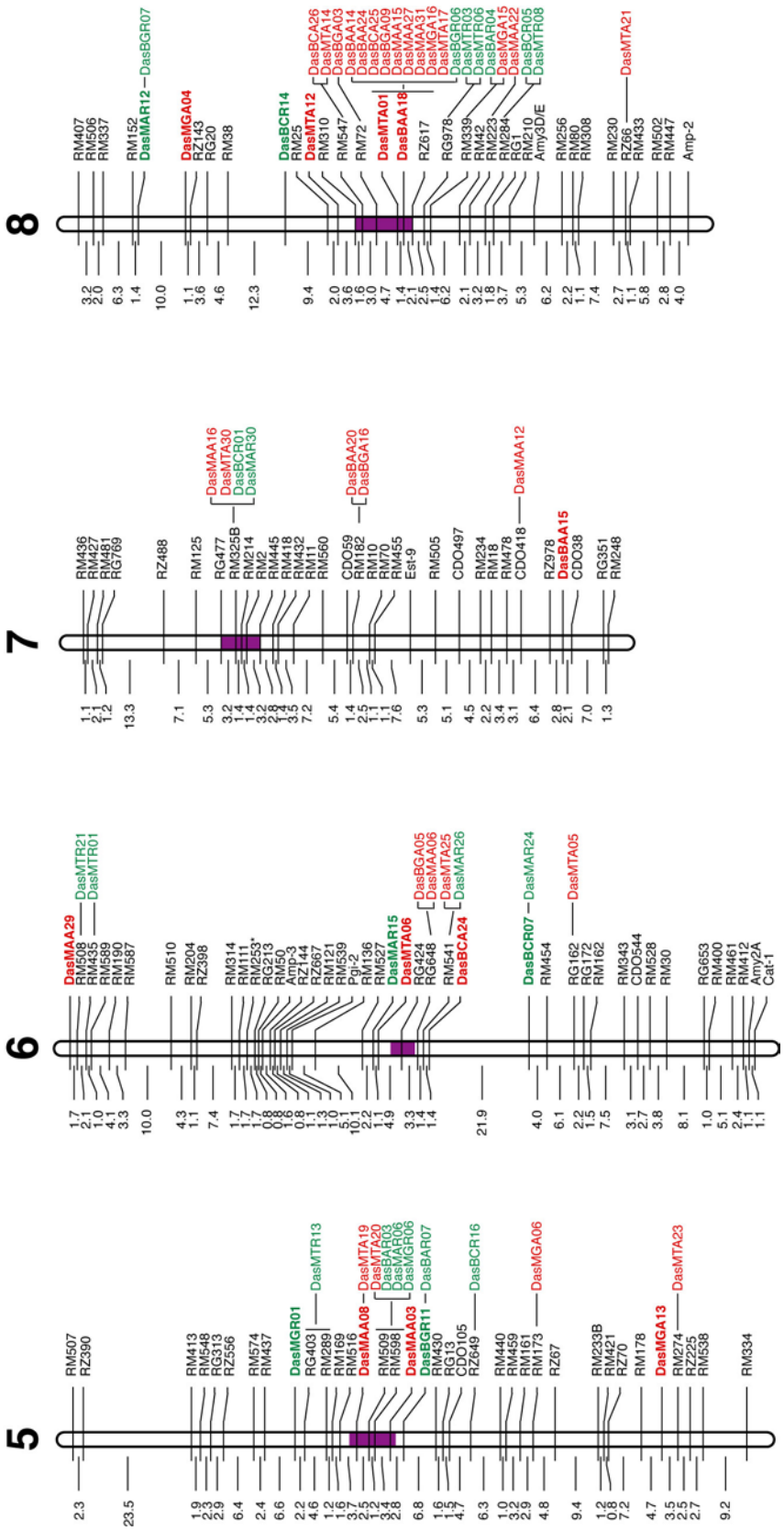
Dasheng markers cluster around all centromeres and on the long arm of chromosome 11. We define a cluster as three or more elements mapping to the same site or to adjacent loci with an average distance of less than 1 cM. Based on this definition, over 50% (120) of the elements were clustered in regions that account for only 3% of the total map distance. The largest clusters of markers were on chromosomes 4, 8, 9 and 12, which all correspond to small chromosomes containing prominent blocks of highly condensed chromatin (FUKUI and IIJIMA 1991). The correlation of *Dasheng* clusters and the distribution of heterochromatin is best seen in chromosome 11, where significantly more elements were observed in the distal region on the long arm than in the pericentromeric region (Figure 3.3). The distal region of chromosome 11 is one of the most heterochromatic regions in the rice genome (CHENG *et al*. 2001b).

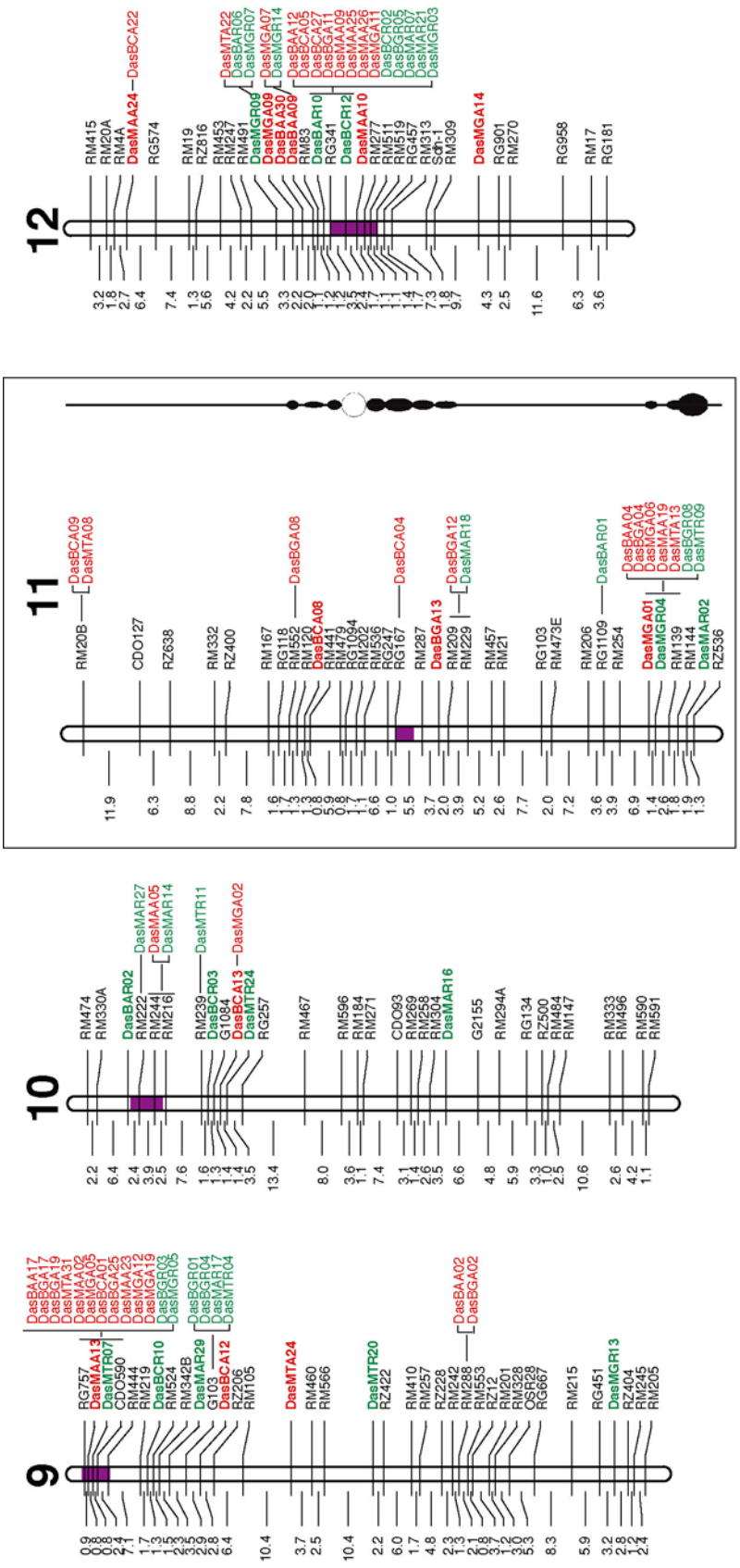
FISH analysis

The mapped elements represent only about 20% of the *Dasheng* family. FISH analysis was performed to ascertain whether the entire family shows similar clustering.

Figure 3.3. Genetic map of rice (*Oryza sativa*) with *Dasheng* and framework markers. *Dasheng* markers from Azucena and IR64 are in red and green, respectively. *Dasheng* markers with a ripple of LOD > 2.0 were integrated into the framework map (in bold). *Dasheng* markers that cosegregate with a framework marker with absolute linkage are connected to this framework marker by a horizontal or slanted line. Vertical lines indicate possible intervals for *Dasheng* markers that are mapped with low LOD scores. Centromeres are indicated by purple boxes. The position of centromeres in this map is based on TEMNYKH *et al*. (2001), except for that in chromosome 10, which is based on CHENG *et al*. (2001a). Also shown for chromosome 11 is a diagram of the distribution of heterochromatic regions (indicated by filled ovals; the open circle in the middle represents the centromere) (CHENG *et al*. 2001b).


1
2
3
4


5
RM507
RZ390
RM413
RM548
RG313
RZ556
RM574
RM437
DasMGR01
RG403
DasMTR13
RM289
RM169
RM516
DasMAA08
DasMTA19
DasMAA20
DasMAR03
DasMAR06
DasMGR06
RM509
RM598
DasMAA03
DasBGR11
DasBAR07
RM430
RG13
CDO105
RZ649
DasBCR16
RM440
RM459
RM161
RM173
DasMGA06
RZ67
RM233B
RM421
RZ70
RM178
DasMGA13
RM274
DasMTA23
RZ225
RM538
RM334
6
DasMAA29
RM508
DasMTR21
RM435
DasMTR01
RM589
RM190
RM587
RM510
RM204
RZ398
RM314
RM111
RM253*
RG213
RM50
Amp-3
RZ144
RZ667
RM121
RM539
Pgi-2
RM136
RM527
DasMAR15
DasMTA06
RG424
DasBGA05
RG648
DasMAA06
RM541
DasMTA25
DasBCA24
DasMAR26
DasBCR07
DasMAR24
RM454
RG162
DasMTA05
RG172
RM162
RM343
CDO544
RM528
RM30
RG653
RM400
RM461
RM412
Amy2A
Cat-1
7
RM436
RM427
RM481
RG769
RZ488
RM125
RG477
DasMAA16
RM325B
DasMTA30
DasBCR01
RM214
DasMAR30
RM2
RM445
RM418
RM432
RM11
RM560
CDO59
DasBAA20
RM182
DasBGA16
RM10
RM70
RM455
Est-9
RM505
CDO497
RM234
RM18
RM478
DasMAA12
CDO418
RZ978
DasBAA15
CDO38
RG351
RM248
8
RM407
RM506
RM337
RM152
DasMAR12
DasBGR07
DasMGA04
RZ143
RG20
RM38
DasBCR14
RM25
DasMTA12
RM310
RM547
RM72
DasMTA01
DasBAA18
RZ617
RG978
RM339
RM42
RM223
RM284
RG1
RM210
Amy3D/E
RM256
RM80
RM308
RM230
RZ66
RM433
RM502
RM447
Amp-2
DasBCA26
DasMTA14
DasBGA03
DasBAA14
DasBAA24
DasBGA25
DasBGA09
DasMAA15
DasMAA27
DasMAA31
DasMGA16
DasMTA17
DasBGR06
DasMTR03
DasMTR06
DasBAR04
DasMAA15
DasMAA22
DasBCR05
DasMTR08
DasMTA21


9
10
11
12

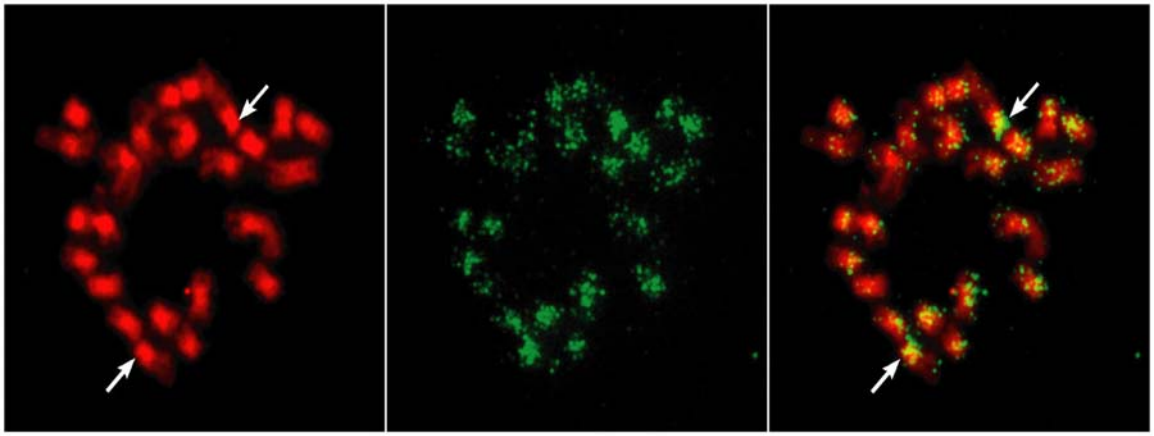
To this end, an internal fragment of *Dasheng* (Figure 3.1) was used as a FISH probe with chromosomes prepared from cv. Nipponbare and Zhongxian 3037, an *indica* cultivar (Figure 3.4). In agreement with the mapping results, the majority of the FISH signal concentrated in pericentromeric regions with the most intense signals located on several small chromosomes. One of the small chromosomes with an intense signal was unambiguously identified as chromosome 4 based on its distinctive arm ratio. This chromosome was previously found to contain one of the most heterochromatic regions in the rice genome (CHENG *et al.* 2001b). The absence of an exceptionally large cluster of elements on our genetic map could be due to the use of different strains for the genetic mapping vs. the cytogenetic analysis. Alternatively, since the genetic map reflects only polymorphic insertion sites, most of the elements on chromosome 4 may not be polymorphic in this mapping population.

Distribution of *Dasheng* on chromosome 1

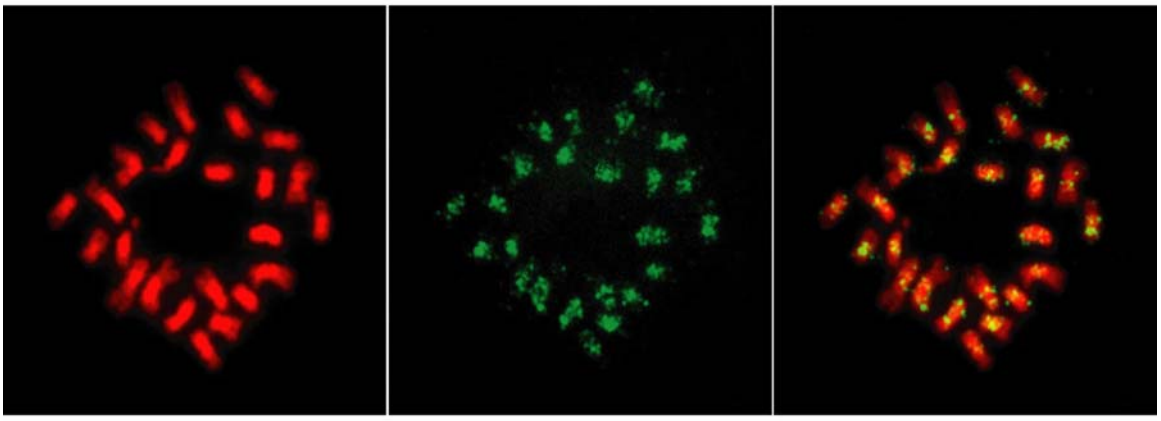
At the time of this study, about 30% of rice genomic sequence was publicly available, including almost the entire chromosome 1. To provide a direct physical measure of how densely clustered the elements are on chromosome 1, the positions of all *Dasheng* were determined from the genomic sequence. The actual distribution of *Dasheng* elements on a single chromosome permits a determination of whether the apparent clustering of *Dasheng* in pericentromeric regions on the genetic map might instead be an artifact of the lower recombination rate in these regions (MOORE and SHERMAN 1975). In other words, 1 cM may contain far more DNA around the centromeres, and this would give the appearance of clustering on a genetic map even if the insertion frequency is the same, in physical terms, as that in the gene-rich

Figure 3.4. FISH analysis of *Dasheng* distribution in rice mitotic chromosomes. *Dasheng* probes were detected by fluorescein isothiocyanate-conjugated anti-biotin antibody (green); chromosomes were stained with propidium iodide (red). Arrows point to the strong signal of *Dasheng* on the short arm of chromosome 4 in Nipponbare (see text for details).

Nipponbare (*japonica*)



Chromosome *Dasheng* Merged



Zhongxian 3037 (*indica*)

chromosome arms. To test this notion, we dissected chromosome 1 into pericentromeric regions and chromosome arms, and calculated both physical:genetic distance and insertion frequency of *Dasheng* elements in different regions (see Materials and Methods for details). Consistent with the low recombination ratio in pericentromeric regions, the ratio of physical: genetic distance was roughly three times higher (660 kb DNA per cM) in pericentromeric regions compared to chromosome arms (206 kb DNA per cM) ($p < 0.05$). However, more significant is the variation in the insertion frequency, which was about five times higher in the pericentromeric regions than in the arms (1.9 vs. 0.4 elements per Mb DNA, $p < 0.01$). These data confirm the higher density of *Dasheng* elements in pericentromeric regions.

Dasheng **elements are not nested**

LTR retrotransposons are commonly found in large clusters in the genomes of grasses. In many instances, these clusters are comprised of LTR retrotransposons inserted into other members of the same family (like *BARE-1*, SHIRASU *et al*. 2000) or into elements of other families (SANMIGUEL *et al*. 1998). Two rice LTR elements, *RIRE3* and *RIRE8* were previously found to be nested (KUMEKAWA et al. 1999). Although the density of *Dasheng* in the rice genome (~1000 copies / 430 Mb, this study) is comparable to that of *BARE-1* in barley (14000 copies / 4800 Mb) (VICIENT *et al*. 1999), nested insertions of *Dasheng* elements were not observed. Only six out of 109 *Dasheng* elements are located within 10 kb of another *Dasheng* element, and the shortest

distance between two *Dasheng* elements was 1.6 kb. As such, the clustering of *Dasheng* is unlikely to be due to a self-insertion preference, as has been observed for some retrotransposons and MITEs (HIGASHIYAMA et al. 1997; JIANG and WESSLER 2001).

Since the pericentromeric regions are enriched in repetitive sequences, including transposable elements (DONG *et al*. 1998; LANGDON *et al*. 2000; NONOMURA and KURATA 2001), the clustering of *Dasheng* could also be attributed to an insertion preference for other repetitive DNA, such as microsatellites or other transposable elements (CHRISTENSEN *et al*. 2000). To address this question, sequences flanking all *Dasheng* elements in the database were used as queries in computer-assisted searches. Of the 109 elements, 19 were found within an identifiable transposable element, 24 were located within 100 bp of an element and about half were associated with low copy number sequences. Among this latter group, none showed significant similarity with a comprehensive database of rice TEs (N. JIANG and S. WESSLER, unpublished data).

In addition, unlike *RCS1*, *RIRE* 3 and *RIRE8*, three other high copy number LTR retrotransposons in rice (DONG *et al*. 1998; KUMEKAWA et al. 1999; LANGDON *et al*. 2000; NONOMURA and KURATA 2001), *Dasheng* elements were not flanked by the *RCS2* centromere repeat (DONG *et al*. 1998), indicating that *Dasheng* is not a centromeric component. However, over half of the 215 *Dasheng* markers described in this study are located in pericentromeric regions. Since *Dasheng* elements do not specifically insert into other repetitive sequences, these markers may prove useful in the construction of fine structure maps of rice pericentromeric regions and isolation of genes buried in heterochromatic regions. Other cloning strategies frequently miss such genes.

Recent amplification of *Dasheng*

Evidence from LTR similarity

Since the LTR of a single retrotransposon is identical upon insertion (LEWIN 1997), sequence divergence between LTRs provides a measure of the time of insertion when an estimate of the nucleotide substitution rate is available (SANMIGUEL *et al*. 1998, BOWEN and MCDONALD 2001). The average substitution rate in the *adh1* and *adh2* loci of grasses ($6.5 \times 10\text{-}9$ substitutions per synonymous site per year) has been used to estimate the time of insertion of maize retrotransposons (GAUT *et al*. 1996; SANMIGUEL *et al*. 1998). In this study, a search of the 100 MB of publicly available rice sequence led to the identification of 109 *Dasheng* elements of which 60 were full-length (56%), 32 were truncated (28%) and 17 were solo LTRs (16%). Among the 60 full-length elements, 35 (58%) have more than 99.5% LTR similarity, with 15 being identical. In the discussions that follow, these elements are referred to as "recent". LTR sequence similarity of the other 25 elements varies from 92.7% to 99.1%. By using the same base substitution rate as SANMIGUEL *et al*. (1998), we estimate that the *Dasheng* elements with more than 99.5% LTR identity (58% of the available full-length elements) inserted within the last 500,000 years. This is a conservative estimate since LTRs evolve more rapidly than coding regions like *adh1* and *adh2* (SANMIGUEL *et al*. 1998) and because reverse transcription is known to be an error-prone process. Based on a comparison with other high-copy number LTR elements in rice, the *Dasheng* family has the highest ratio of elements with identical LTRs (15 out of 109) (N. JIANG and S. WESSLER, unpublished data). Thus, *Dasheng* may have amplified more recently than all other high copy number elements in the rice genome.

A hypervariable region and tandem repeats

A phylogenetic tree was constructed based on the LTR sequences of *Dasheng* elements and some *RIRE2* elements (Figure 3.5) and used to evaluate whether other structural features of *Dasheng* correlated with recently amplified elements. Of particular interest were a hypervariable region, the tandem repeats and the solo LTRs.

As mentioned above, the hypervariable region is located between the PBS and the first tract of tandem repeats (Figure 3.3.1). This region consists of common sequence shared by many or a few elements (no sequence is shared by all elements) and unique sequence. A similarly organized region of sequence heterogeneity was reported for the *Stonor* elements of maize (MARILLONNET and WESSLER 1998). Interestingly, more than half of the recent elements (19 out of 35) have the same sequence in this region (see Figure 3.5 for branch lengths of elements labeled with an asterisk), suggesting that the recent amplification of *Dasheng* could be due largely to the transposition of elements in this subgroup. In addition, the average length of the tandem repeat region in this group is significantly longer than that of other elements (3.7 vs. 2.2 kb, $t<0.001$). These data do not permit a determination of whether younger elements have more repeats or if repeats are gradually deleting from the older elements.

Evolution of *Dasheng* elements

Targeted insertion versus negative selection

Having both complete sequences of the element and precise chromosomal locations permits a preliminary determination of whether the clustering of *Dasheng* elements in the pericentromeric region is due to targeted insertion into the gene-rich arms

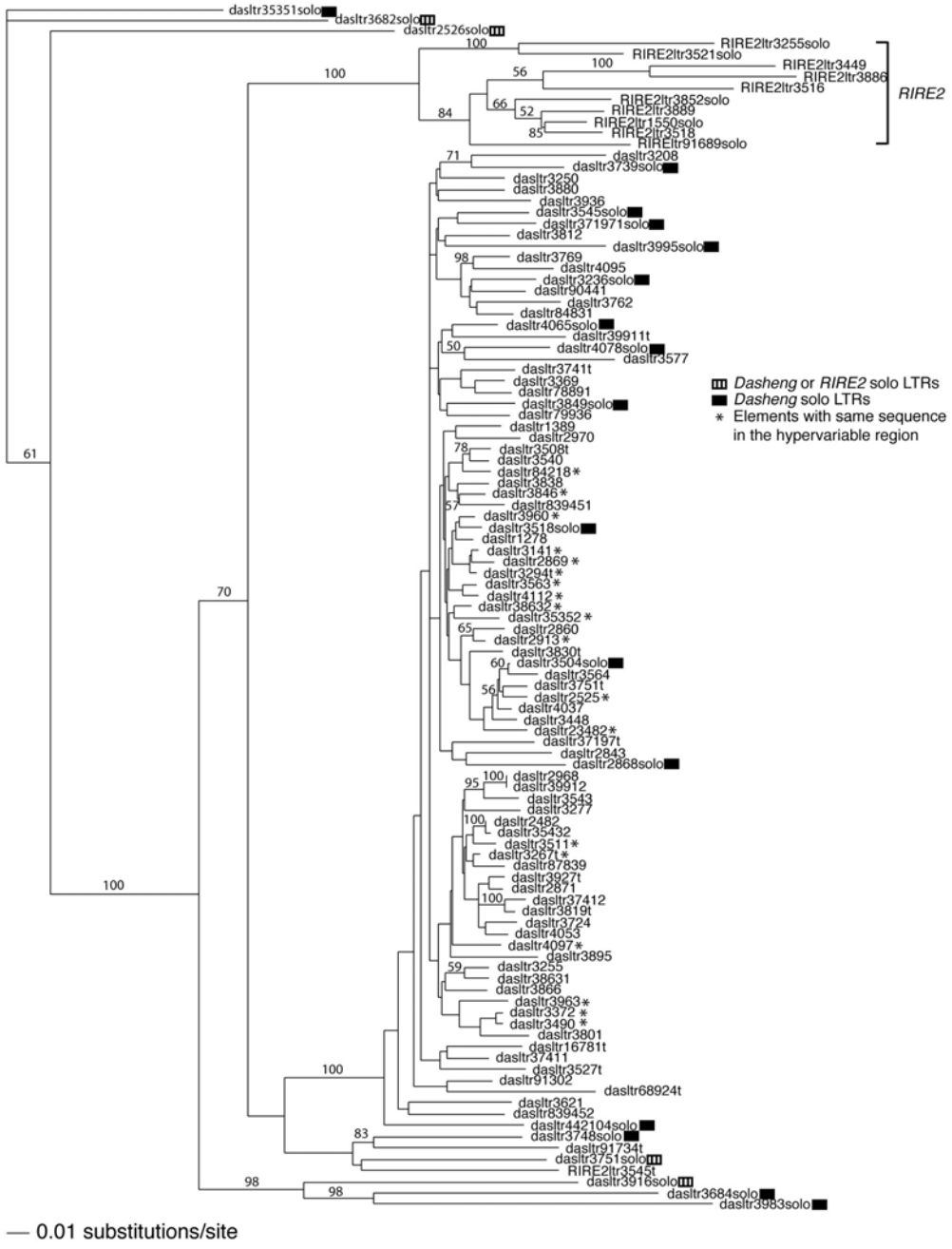

dasltr35351solo
dasltr3682solo
dasltr2526solo
RIRE2ltr3255solo
RIRE2ltr3521solo
RIRE2ltr3449
RIRE2ltr3886
RIRE2ltr3516
RIRE2ltr3852solo
RIRE2ltr3889
RIRE2ltr1550solo
RIRE2ltr3518
RIREltr91689solo
RIRE2
dasltr3208
dasltr3739solo
dasltr3250
dasltr3880
dasltr3936
dasltr3545solo
dasltr371971solo
dasltr3812
dasltr3995solo
dasltr3769
dasltr4095
dasltr3236solo
dasltr90441
dasltr3762
dasltr84831
dasltr4065solo
dasltr39911t
dasltr4078solo
dasltr3577
dasltr3741t
dasltr3369
dasltr78891
dasltr3849solo
dasltr79936
dasltr1389
dasltr2970
dasltr3508t
dasltr3540
dasltr84218*
dasltr3838
dasltr3846*
dasltr839451
dasltr3960*
dasltr3518solo
dasltr1278
dasltr3141*
dasltr2869*
dasltr3294t*
dasltr3563*
dasltr4112*
dasltr38632*
dasltr35352*
dasltr2860
dasltr2913*
dasltr3830t
dasltr3504solo
dasltr3564
dasltr3751t
dasltr2525*
dasltr4037
dasltr3448
dasltr23482*
dasltr37197t
dasltr2843
dasltr2868solo
dasltr2968
dasltr39912
dasltr3543
dasltr3277
dasltr2482
dasltr35432
dasltr3511*
dasltr3267t*
dasltr87839
dasltr3927t
dasltr2871
dasltr37412
dasltr3819t
dasltr3724
dasltr4053
dasltr4097*
dasltr3895
dasltr3255
dasltr38631
dasltr3866
dasltr3963*
dasltr3372*
dasltr3490*
dasltr3801
dasltr16781t
dasltr37411
dasltr3527t
dasltr91302
dasltr68924t
dasltr3621
dasltr839452
dasltr442104solo
dasltr3748solo
dasltr91734t
dasltr3751solo
RIRE2ltr3545t
dasltr3916solo
dasltr3684solo
dasltr3983solo
Dasheng or RIRE2 solo LTRs
Dasheng solo LTRs
* Elements with same sequence in the hypervariable region
— 0.01 substitutions/site

Figure 3.5. Phylogenetic analysis of LTR sequences of *Dasheng* and some *RIRE2* elements using the neighbor-joining algorithm from distance matrices. Branch length is proportional to genetic distance. Bootstrap values > 50 are indicated as a percentage of 1000 replicates. "*Dasheng* or *RIRE2* solo LTRs" indicates that the sequence similarity of these solo LTRs to the *Dasheng* consensus is not significantly different from that to *RIRE2* consensus by t test.

or post-insertion selection. If *Dasheng* shows no target site preference, but elements are lost over time from the arms, the arms should contain more recent insertions than the pericentromeric regions. However, no significant difference is seen in the number of recent insertions in arms vs. pericentromeric regions. In chromosomal arms, six out of 27 (22%) full-length elements have identical LTRs, whereas in pericentromeric regions, nine out of 33 (27%) have identical LTRs ($\chi 2$ test; $p > 0.50$).

Origin of solo LTRs

Solo LTRs are believed to arise from intra-element recombination between transiently paired LTRs (PARKET *et al*. 1995). Recently, the formation of solo LTRs in barley has been proposed as a mechanism that can reverse genome expansion (SHIRASU *et al*. 2000). With 16% of the sequenced elements present as solo LTRs, the *Dasheng* family provides a unique opportunity to address questions about the formation of solo LTRs by analyzing both their age and their distribution relative to full-length family members.

There are at least two models that can account for the formation of solo LTRs. If it is a stochastic process, older insertions are more likely to have undergone recombination and the distribution of solo LTRs should be no different than the distribution of full-length elements. On the other hand, since solo LTRs arise from intra-element recombination, the frequency of solo LTR formation might correlate with regional recombination frequencies. In this case, there would be relatively fewer solo LTRs in the centromeric and pericentromeric regions where recombination rates are much lower than those in the gene-rich chromosomal arms (MOORE and SHERMAN 1975). As can be seen in Figure 3.5, solo LTRs are associated more often with the longer

branches indicating that they are older, on average, than the full-length elements (0.0437 vs. 0.0187, $t < 0.001$). In addition, the ratio of solo LTRs to full-length elements in the arms (8 : 27) is only slightly higher than that in pericentromeric regions (6 : 33), and the difference is not significant ($p > 0.10$ by $\chi 2$ test). Taken together, these data suggest that solo LTR formation in the *Dasheng* family is probably a stochastic process.

Concluding remarks

In this study, we characterized an unusual LTR element in rice. As a special category of LTR elements, *Dasheng* is distinguished by its (1) lack of coding capacity, (2) presence of long tracts of tandem repeats, (3) clustering in heterochromatic regions, (4) high copy number, and (5) recent amplification.

Since *Dasheng* is among the rice elements of highest copy number and most recent amplification, it is of great interest to know if members of the *Dasheng* family are still capable of retrotransposition. To date, activity has not been demonstrated for any of the high copy number LTR retrotransposons in rice. The only active rice elements (such as *Tos17*) are present in less than five copies and are activated to retrotranspose by tissue culture (HIROCHIKA *et al*. 1996; AGRAWAL *et al*. 2001; YAMAZAKI *et al*. 2001). In fact, although LTR retrotransposons are the major component of most plant genomes, the only high copy number LTR retrotransposon with demonstrated activity is *BARE-1* from barley (SUONIEMI *et al*. 1996; JAASKELAINEN *et al*. 1999). As such, it is the only genomic component known to be capable of contributing significantly to genome size variation between populations and related species in plants (KALENDAR *et al*. 2000). Like *BARE-1* in barley, *Dasheng* is a major component of the genome of cultivated rice, *Oryza sativa.*

In addition, a preliminary survey indicates that *Dasheng* is probably abundant in all species of the genus *Oryza* (Figure 3.3.2). For these reasons it will be important for future studies to determine whether *Dasheng* elements are still capable of transposition.

The origin of the *Dasheng* is also of interest since it is a nonautonomous class 1 element. Nonautonomous transposable elements are widespread in eukaryotic organisms. For DNA elements and non-LTR retrotransposons, the copy number of nonautonomous elements is usually much higher than that of the corresponding autonomous element (KAPITONOV and JURKA 1999; FESCHOTTE and MOUCHÈS 2000; INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM 2001). Unlike other classes of nonautonomous elements, only a few high copy number LTR elements have been characterized. The only other plant element is the maize *Zeon-1*, which has a copy number of 6,000 to 32,000 (MEYERS *et al*. 2001). However, unlike *Dasheng*, *Zeon-1* is one of the oldest elements in the maize genome (HU *et al*. 1995, SANMIGUEL *et al*. 1998). As such, it will be difficult, if not impossible, to deduce what autonomous element could be responsible for the amplification of *Zeon-1*. In contrast, the availability of most of the rice genome sequence coupled with the recent amplification of *Dasheng* facilitates a comprehensive analysis of autonomous elements that could have given rise to *Dasheng*. At this time, the *RIRE2* family with its related LTRs, as well as the presence of some recently amplified members, is the best candidate. Further studies are underway to establish additional connections between these two LTR element families.

Acknowledgments

We thank I. King Jordan (National Institutes of Health and National Library of Medicine, NCBI, Bethesda, MD) and Nathan J. Bowen (National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD) for assistance in sequence and phylogenetic analysis and Xiaoyu Zhang for critical reading of the manuscript. This study was supported by grants from the U. S. Department of Energy (DEFG02-94ER20135) and the National Science Foundation (DBI-0077709) to S.R.W.

Literature Cited

AGRAWAL, G. K., M. YAMAZAKI, M. KOBAYASHI, R. HIROCHIKA, A. MIYAO et al., 2001 Screening of the rice viviparous mutants generated by endogenous retrotransposon *Tos17* insertion. Tagging of a zeaxanthin epoxidase gene and a novel OsTATC gene. Plant Physiol **125:** 1248-1257.

BOWEN, N. J., and J. F. MCDONALD, 2001 *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. **11:** 1527-40.

BUREAU, T. E., P. C. RONALD and S. R. WESSLER, 1996 A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wildtype rice genes. Proc. Natl. Acad. Sci. USA **93:** 8524-8529.

CASA, A. M., C. BROUWER, A. NAGEL, L. WANG, Q. ZHANG et al., 2000 The MITE family heartbreaker (*Hbr*): molecular markers in maize. Proc. Natl. Acad. Sci. USA **97:** 10083-10089.

CHEN, M. P., P. SANMIGUEL, A. C. DE OLIVEIRA, S. S. WOO, H. ZHANG et al., 1997 Microcolinearity in *sh2*-homologous regions of the maize, rice and sorghum genomes. Proc. Natl. Acad. Sci. USA **94:** 3431-3455.

CHENG, Z., G. G. PRESTING, C. R. BUELL, R. A. WING and J. JIANG, 2001a High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. Genetics **157:** 1749-1757.

CHENG, Z., R. BUELL, R. A.WING, M. GU and J. JIANG 2001b Towards a cytological characterization of the rice genome. Genome Res. **11:**2133-2141.

CHRISTENSEN, S., G. PONT-KINGDON and D. CARROLL, 2000 Target specificity of the endonuclease from the *Xenopus laevis* non-long terminal repeat retrotransposon, *Tx1L*. Mol Cell Biol **20:** 1219-1226.

DONG, F., J. T. MILLER, S. A. JACKSON, G. L. WANG, P. C. RONALD et al., 1998 Rice (*Oryza sativa*) centromeric regions consist of complex DNA. Proc. Natl. Acad. Sci. USA **95:** 8135-8140.

DUBCOVSKY, J., W. RAMAKRISHNA, P. J. SANMIGUEL, C. S. BUSSO, L. YAN et al., 2001 Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol **125:** 1342-1353.

FESCHOTTE, C., and C. MOUCHÈS, 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.

FUKUI, K., and K. IIJIMA, 1991 Somatic chromosome map of rice by imaging methods. Theor. Appl. Genet. **81:** 589-596.

GE, S., T. SANG, B.-R. LU and D.-Y. HONG, 1999 Phylogeny of rice genomes with emphasis on origins of allotetroaploid species. Proc. Natl. Acad. Sci. USA **96:** 14400-14405.

GAUT, B. S., B. R. MORTON, B. C. MCCAIG and M. T. CLEGG, 1996 Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene rbcL. Proc. Natl. Acad. Sci. USA **93:** 10274-9.

GUIDERDONI, E., E. GALINATO, J. LUISTRA and G. VERGARRO, 1992 Anther culture of tropical *japonica* x *indica* hybrids of rice (*Oryza sativa* L.). Euphytica **62:** 219-224.

HAN, C. G., M. J. FRANK, H. OHTSUBO and E. OHTSUBO, 2000 New transposable elements identified as insertions in rice transposon *Tnr1*. Genes Genet. Syst. **75:** 69-77.

HARUSHIMA, Y., M. JANO, A. SHOMURA, M. SATO, T. SHIMANO et al., 1998 A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics **148:** 479–494.

HIGASHIYAMA, T., Y. NOUTOSHI, M. FUJIE and T. YAMADA, 1997 *Zepp*, a LINE-like retrotransposon accumulated in the *Chlorella* telemetric region. EMBO J. **16:** 3715-23.

HIROCHIKA, H., K. SUGIMOTO, Y. OTSUKI and M. KANDA, 1996 Retrotransposons of rice involved in mutations induced by tissue culture. Proc. Natl. Acad. Sci. USA **93:** 7783-7788.

HU, W. M., O. P. DAS and J. MESSING, 1995 *Zeon-1*, a member of a new maize retrotransposon family. Mol. Gen. Genet. **248:** 471-480.

HUANG, H., and G. KOCHERT, 1994 Comparative RFLP mapping of an allotetraploid wild rice species (*Oryza latifolia*) and cultivated rice (*O. sativa*). Plant Mol. Biol. **25:** 633-648.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

JAASKELAINEN, M., A. H. MYKKANEN, T. ARNA, C. M. VICIENT, A. SUONIEMI et al., 1999 Retrotransposon *BARE-1:* expression of encoded proteins and formation of virus-like particles in barley cells. Plant J. **20:** 413-422.

JIANG, J., B. S. GILL, G. L. WANG, P. C. RONALD and D. C. WARD, 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.

JIANG, N., and S. R. WESSLER, 2001 Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. Plant Cell **13:** 2553-2564.

KALENDAR, R., J. TANSKANEN, S. IMMONEN, E. NEVO and A. H. SCHULMAN, 2000 Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. USA **97:** 6603-6607.

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

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

KUMAR, S., K. TAMURA, I. B. JAKOBSEN and M. NEI, 2001 MEGA2: Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Arizona, USA.

KUMEKAWA, N., H. OHTSUBO, T. HORIUCHI and E. OHTSUBO, 1999 Identification and characterization of novel retrotransposons of the *gypsy* type in rice. Mol. Gen. Genet. **260:** 593-602.

LANDER, E. S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. J. DALY et al., 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174-181.

LANGDON, T., C. SEAGO, M. MENDE, M. LEGGETT, H. THOMAS et al., 2000 Retrotransposon evolution in diverse plant genomes. Genetics **156:** 313-325.

LEWIN, B. 1997. *Genes VI.* Oxford University Press, New York.

LI, Z. Y., S. Y. CHEN, X. W. ZHENG and L. H. ZHU, 2000 Identification and chromosomal localization of a transcriptionally active retrotransposon of *Ty3-gypsy* type in rice. Genome **43:** 404-408.

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

MARILLONNET, S., and S. R. WESSLER, 1998 Extreme structural heterogeneity among the members of a maize retrotransposon family. Genetics **150:** 1245-56.

MCCOUCH, S. R., G. KOCHERT, Z. H. YU, G. S. KHUSH, W. R. COFFMAN et al., 1988 Molecular mapping of rice chromosomes. Theor. Appl. Genet. **76:** 815-829.

MEYERS, B. C., S. V. TINGEY and M. MORGANTE, 2001 Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. **11:** 1660-1676.

MILLER, J. T., F. DONG, S. A. JACKSON, J. SONG and J. JIANG, 1998 Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. Genetics **150:** 1615-1623.

MOORE, C. W., and F. SHERMAN, 1975 Role of DNA sequences in genetic recombination in the iso-1-cytochrome c gene of yeast. I. Discrepancies between physical distances and genetic distances determined by five mapping procedures. Genetics **79:** 397-418.

MOORE, G., T. FOOTE, T. HELENTJARIS, K. DEVOS, N. KURATA et al., 1995 Was there a single ancestral cereal chromosome? Trends in Genetics **11:** 81-82.

MYERS, T. 2001 Rice genome consortium will finish ahead of schedule. Nature **409** :752.

NONOMURA, K., and N. KURATA, 2001 The centromere composition of multiple repetitive sequences on rice chromosome 5. Chromosoma **110:** 284-291.

OHTSUBO, H., N. KUMEKAWA and E. OHTSUBO, 1999 *RIRE2*, a novel *gypsy*-type retrotransposon from rice. Genes Genet. Syst. **74:** 83-91.

PARKET, A., O. INBAR and M. KUPIEC, 1995 Recombination of *Ty* elements in yeast can be induced by a double-strand break. Genetics **140:** 67-77.

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

SANMIGUEL, P., B. S. GAUT, A. TIKHONOV, Y. NAKAJIMA and J. L. BENNETZEN, 1998 The paleontology of intergene retrotransposons of maize. Nat. Genet. **20:** 43-5.

SHARMA, S. D., S. R. DHU and P. K. AGARWAL, 2000 Species of genus *Oryza* and their interrelationships. In *Rice breeding and genetics* (ed. Nanda, J.S.), pp311-346. Science Publisher, Enfield, New Hampshire.

SHIRASU, K., A. H. SCHULMAN, T. LAHAYE and P. SCHULZE-LEFERT, 2000 A contiguous 66 kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. **10:** 908-915.

SUONIEMI, A., A. NARVANTO and A. H. SCHULMAN, 1996 The *BARE-1* retrotransposon is transcribed in barley from an LTR promoter active in transient assays. Plant Mol. Biol. **31:** 295-306.

SUONIEMI, A., J. TANSKANEN and A. H. SCHULMAN, 1998 *Gypsy*-like retrotransposons are widespread in the plant kingdom. The Plant J. **13:** 699-705.

TARCHINI, R., P. BIDDLE, R. WINELAND, S. TINGEY and A. RAFALSKI, 2000 The complete sequence of 340 kb of DNA around the rice *Adh1-adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell **12:** 381-391.

TEMNYKH, S., G. DECLERCK, A. LUKASHOVA, L. LIPOVICH, S. CARTINHOUR et al., 2001 Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. **11:** 1441-1452.

THOMAS, C. A., 1971 The genetic organization of chromosomes. Annu. Rev. Genet. **5:** 237-256.

TURCOTTE, K., S. SRINIVASAN and T. BUREAU, 2001 Survey of transposable elements from rice genomic sequences. Plant J. **25:** 169-179.

UOZO, S., H. IKEHASHI, N. OHMIDO, H. OHTSUBO, E. OHTSUBO et al., 1997 Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus *Oryza*. Plant Mol. Biol. **35:** 791-799.

VAN DEN BROECK, D., T. MAES, M. SAUER, J. ZETHOF, P. DE KEUKELEIRE et al., 1998 Transposon display identifies individual transposable elements in high copy number lines. The Plant J. **13:** 121-129.

VICIENT, C. M., A. SUONIEMI, K. ANAMTHAWAT-JONSSON, J. TANSKANEN, A. BEHARAV et al., 1999 Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. Plant Cell **11:** 1769-1784.

VOS, P., R. HOGERS, M. BLEEKER, M. REIJANS, T. VAN DE LEE et al., 1995 AFLP, a new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407-4414.

VOYTAS, D. F., M. P. CUMMINGS, A. KONIECZNY, F. M. AUSUBEL and S. R. RODERMEL, 1992 *copia*-like retrotransposons are ubiquitous among plants. Proc. Natl. Acad. Sci. USA **89:** 7124-7128.

WAUGH, R., K. MCLEAN, A.J. FLAVELL, S.R. PEARCE, A. KUMAR, B.B.T. THOMAS, and W. POWELL, 1997 Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol. Gen. Genet. **253**: 687-694.

XIONG, Y., and T. H. EICKBUSH, 1990 Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:** 3353-3362.

YAMAZAKI, M., H. TSUGAWA, A. MIYAO, M. YANO, J. WU et al., 2001 The rice retrotransposon *Tos17* prefers low-copy-number sequences as integration targets. Mol. Gen. Genet. **265:** 336-344.

CHAPTER 4

AN ACTIVE DNA TRANSPOSON FAMILY IN RICE[2]

[2] Jiang, N., Bao, Z., Zhang, X., Eddy, S. R., McCouch, S. R. and S.R. Wessler. Submitted to *Nature*.

Abstract

The availability of draft sequences for the two subspecies of rice (*Oryza sativa)*, *japonica* (cv. Nipponbare) and *indica* (cv. 93-11), has been exploited in a computational approach to identify the first active DNA transposons from rice and the first active miniature inverted-repeat transposable element (MITE) from any organism. A sequence classified as a *Tourist*-like MITE of 430 bp (called *miniature Ping* or *mPing*) was present in about 70 copies in Nipponbare and 12 copies in 93-11. *mPing* elements, which are all nearly identical, were found to actively transpose in an *indica* cell culture line. Database searches identified a family of related transposase-encoding elements (called *Pong*) that were also activated to transpose in the same cells. Virtually all new insertions of *mPing* and *Pong* elements were into low copy regions of the rice genome. Intriguingly, the *mPing* MITEs have preferentially amplified since domestication in cultivars adapted to environmental extremes; a situation reminiscent of McClintock's genomic shock theory for transposon activation.

Introduction

Rice is the most important crop for human nutrition in the world. At 430 Mb, it also has the smallest genome among the agriculturally important cereals (including maize, sorghum, barley and wheat)[1]. For these reasons rice is the focus of several genome sequencing projects in both the public and private sectors. It is the first higher organism for which draft sequences are publicly available for two subspecies, *japonica* and *indica*[2-4].

Computer-assisted analyses of rice genomic sequence has revealed that despite its small size, over 40% is composed of repetitive DNA, most of this being derived from transposable elements (TEs)[3,4]. Although the largest component of TEs in the rice genome is class 1 LTR retrotransposons (14%), numerically the largest group (with 100,000 elements divided into hundreds of families) is miniature inverted-repeat transposable elements (MITEs), comprising about 6% of the genome[5,6]. Since their discovery in maize over a decade ago[7], MITEs have been found to be the predominant TE associated with the noncoding regions of the genes of flowering plants, especially the grasses[8-11]. After their discovery in plants, MITEs were found in several animal genomes including *C.elegans*, mosquitoes, fish, and human (reviewed in[12]).

MITEs are structurally reminiscent of nonautonomous DNA (class 2) elements with their small size (less than 600 bp) and short (10 to 30 bp) terminal inverted repeat (TIR). However, their high copy number (up to 10,000 copies/family) and target-site preference (for TA or TAA) distinguish them from most previously described nonautonomous DNA elements[12].

DNA transposable element families contain: (1) nonautonomous elements that do not encode transposase but retain the *cis*-sequences necessary for transposition, and (2) autonomous elements that encode the transposase required for their mobility and for the mobility of nonautonomous family members[13]. Nonautonomous elements, which make up a significant fraction of eukaryotic genomes, have been classified into families based on the transposase responsible for their mobility. Classifying MITEs in this way has been problematic since no actively transposing MITEs had been reported in any organism. Instead, the tens of thousands of plant MITEs have been classified into two superfamilies based on the similarity of their TIRs and their target site duplication (TSD): *Tourist*-like MITEs and *Stowaway*-like MITEs[6,14,15]. Recently, evidence has accumulated linking *Tourist* and *Stowaway* MITEs with two superfamilies of transposases, *PIF*/*Harbinger* and *Tc1*/*mariner*, respectively[14-16]. However, without activity, these associations remained circumstantial and they provide little information about how MITEs arise from autonomous elements and how they have spread so successfully throughout the genomes of plants and animals.

Results

Computer-assisted identification of a new MITE family

No active transposition has been demonstrated for any of the hundreds of MITEs families previously identified in the rice genome or in any other organism[13]. Most MITE families are characterized by their high copy number (hundreds to thousands per haploid genome) and intra-family sequence identity that is rarely over 95%. Since newly

amplified elements should be identical, these families have most likely been inactive for hundreds of thousands or even millions of years.

We hypothesized that an active MITE family would be characterized by extremely low intra-family sequence divergence. The availability of almost half of the Nipponbare genome (187 Mb by Dec. 24, 2001) in public databases (http://rgp.dna.affrc.go.jp) provided the possibility of searching for repeat families with the structural features of MITEs and with very low intra-family sequence divergence. To this end, a two step protocol was employed involving the use of an algorithm to identify repeat families followed by manual screening of the output for a MITE family with virtually identical members. Twelve hundred seventy five repeats were identified with RECON[17], a program for *de novo* repeat family identification. Manual inspection of these sequences indicated that the repeat we now call *mPing* (for *miniature Ping*) was a candidate for an active family.

The TSDs (the trinucleotide TAA or TTA) and TIRs of *mPing* indicate that it is a *Tourist*-like MITE of 430 bp (Figure 4.1). Of the thirty six copies mined from 270 Mb of Nipponbare sequence (downloaded from RGP on April 25, 2002, including overlaps), twenty six were identical while nine differed at only a single position. One has 5 C to T transitions in a 10 bp region. All copies are classified as subtype A (Figure 4.1). In contrast, eight complete copies of *mPing* were found in the 361 Mb of contig sequence of the *indica* cultivar 93-11, and one almost complete copy of *mPing* (one TIR missing due to sequence truncation) was found in the unassembled sequence[4]. The 93-11(*indica*) elements represent three subtypes (designated *mPing*A, B, C; Figure 4.1). Based on these

values, the entire genome of Nipponbare and 93-11 are estimated to contain 70 and 12 copies of *mPing*, respectively.

Transposition of *mPing* in cell culture

The only rice transposable elements previously shown to be active were three families of LTR retrotransposons that transposed in both *japonica* (Nipponbare) and *indica* (C5924) during cell culture[18]. Transposition of one of these elements, *Tos17*, was associated with its transcriptional activation in culture[18]. To assess whether *mPing* elements were also activated in the same cell lines, a technique called transposon display (TD) was employed to detect new *mPing* insertions that may have occurred in culture. TD is a modification of the AFLP procedure that generates PCR products that are anchored in a transposable element and in a flanking restriction site[19,20]. Since all of the *mPing* elements are virtually identical except for the central region (see Figure 4.1), element-specific primers located in subterminal sequence (see Methods) were designed in order to amplify all family members and flanking host sequence.

Comparison of the number of PCR products amplified from DNAs (referred to as amplicon from here on) isolated from Nipponbare (*japonica*) and C5924 (*indica*) plants before culture are consistent with the copy number estimates of ~70 and ~12, respectively (Figure 4.2A). However, a differential response to cell culture was observed. Whereas the Nipponbare amplicon pattern is the same before and after culture, the C5924 culture line has undergone a dramatic increase in the number of products. To determine whether the difference was due to nonspecific genomic rearrangements in this cell line, TD was

Figure 4. 1. Comparison of *mPing*, *Ping* and *Pong* elements. Black triangles represent TIRs (nucleotide sequences of the TIRs and TSDs of the rice *Pong*, the maize *PIF*[16] and the bacterial IS*1031c* and IS*112* elements are shown). The thick vertical black line in *mPing* stands for internal sequences that differ among the four subtypes derived from *Ping*. An alignment of the sequences around the breakpoint of each subtype is at the top. The arrowhead indicates the breakpoint in *Ping* where 4923 bp of its internal sequence is not shown in the alignment.

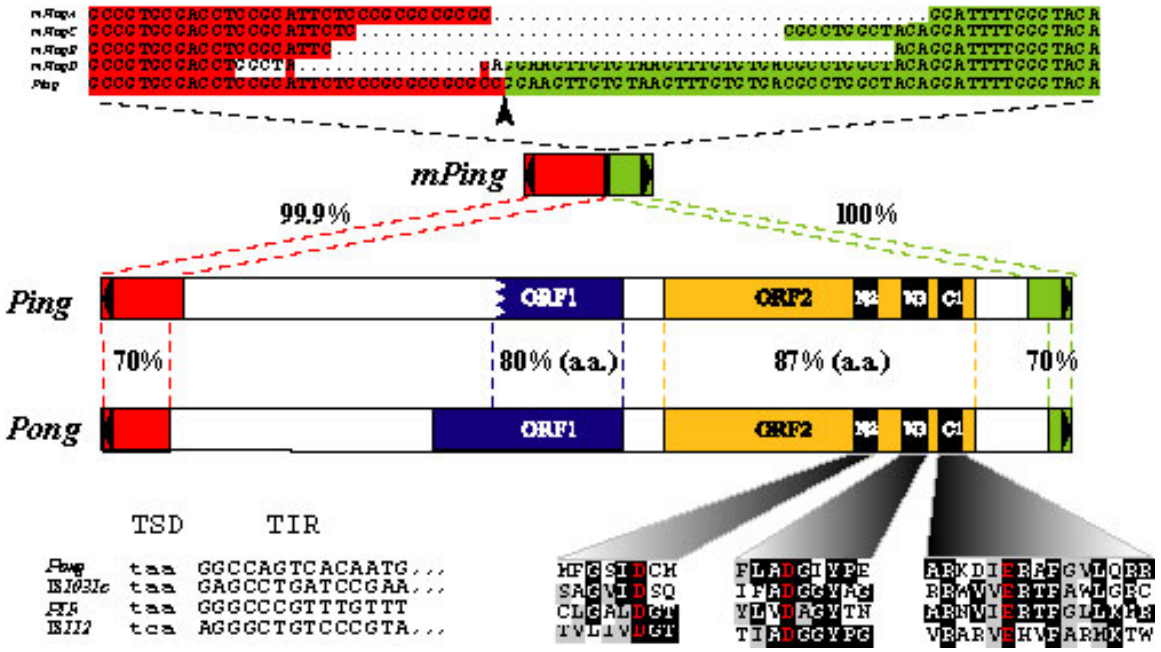
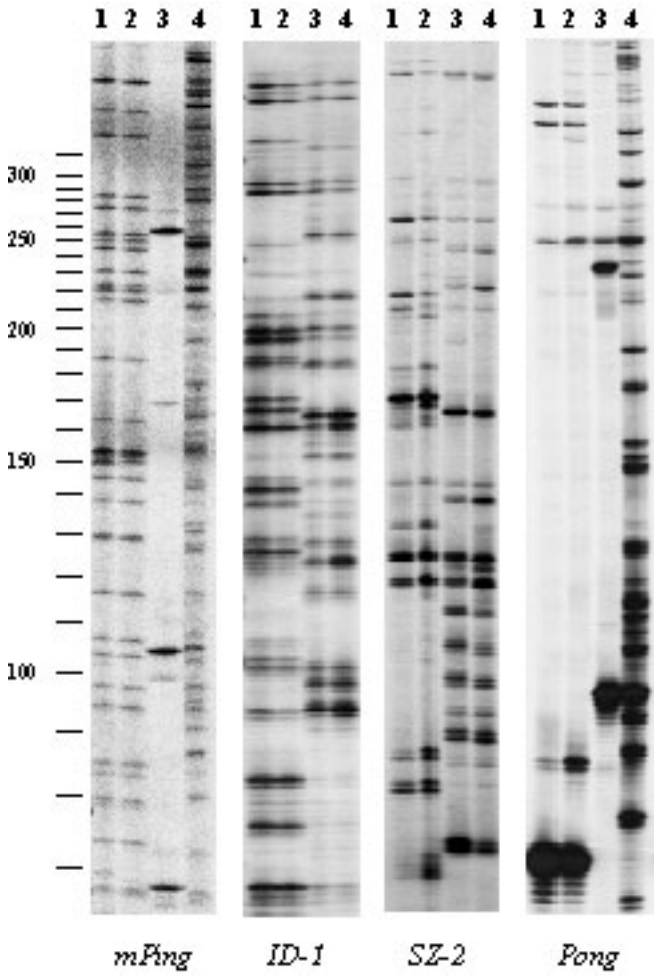

mPing
99.9%
100%
Ping
ORF1
ORF2
70%
80% (a.a.)
87% (a.a.)
70%
Pong
ORF1
ORF2
TSD
TIR
Pong taa GGCCAGTCACAATG...
IS1031c taa GAGCCTGATCCGAA...
MITE taa GGGCCCGTTTGTTT
IS112 tca AGGGCTGTCCCGTA...

Figure 4. 2. Autoradiograph of transposon display gels of *mPing* and *Pong* amplicons with rice genomic DNAs isolated before and after cell culture. The same genomic DNAs (digested and ligated with adapters) were used for each set of primers: 1, Nipponbare; 2, calli of Nipponbare; 3, C5924; 4, Oc cell lines derived from C5924[42]. The migration of DNA markers is on the left in base pairs.


1 2 3 4
1 2 3 4
1 2 3 4
1 2 3 4
300
250
200
150
100
mPing
ID-1
SZ-2
Pong

repeated using the same template DNAs but this time, the *mPing* primers were replaced with either primers derived from the consensus sequence of a *gypsy* type LTR retrotransposon (*SZ-2,* Jiang and Wessler, unpublished data) or from another rice MITE (*ID-1*[6]). In contrast with the results when *mPing* primers were used, the amplicons with these primers were essentially identical before and after cell culture. Taken together these results suggest that transposition and amplification were responsible for the different patterns of *mPing* amplicons. It is important to note that the extent of amplification of *mPing* elements cannot be quantified because the cell line may contain a mixture of sub-populations, each with a few different insertions.

mPing targets low copy (genic) sequences

In several studies, MITEs have been found predominantly in noncoding genic regions[10,11,21]. However, without actively transposing MITEs, it has not been possible to determine whether this distribution reflects preferential targeting or selection against insertion into other regions of the genome. To address this question, 42 PCR products from cell line C5924 were recovered from the TD gel, reamplified, subcloned and sequenced. Insertion sites of newly transposed elements were deduced by using sequences flanking the TIR (37 to 268 bp in length) to query the newly released cv. 93-11 (*indica*) draft sequence[4] and Nipponbare sequence in public databases (http://rgp.dna.affrc.go.jp). Thirty four of 42 sequences matched entries from *indica* contigs while one of the sequences was found only in *japonica* (cv. Nipponbare). Thirty two of the 35 matches were single copy sequences, and one was in a two-copy sequence (see Supplemental Table 1). The remaining two insertion sites were in or next to other

MITEs that were themselves in single copy sequences. Thus, 34 of 35 new insertions were in single copy regions of the genome. In addition, two of the six insertion sites located in fully annotated PACs or BACs were predicted exons (http://rgp.dna.affrc.go.jp), indicating that the previously noted absence of MITEs in exons is most likely due to negative selection against mutagenic insertions.

The amplicons recovered from TD gels only contain the end of a transposon. Complete elements were identified in two ways. Host sequences flanking both ends of the transposon (obtained from the BLAST searches described above) were employed in the design of PCR primers that were used with template DNA to recover the entire intervening transposon. In this way, insertions before cell culture were first identified. Among the seven *mPing* insertions discovered in the original C5924 line before culturing, four were of *mPingB*, two of *mPingC*, and one of a new subtype called *mPingD* (Fig.1). Recall that *mPingA*, *B* and *C* were originally found in the genomic DNA of the sequenced *indica* cultivar, 93-11. Cultivar C5924 is a related but different *indica* variety, thus it may have a slightly different ensemble of *mPing* elements.

The approach used to amplify the family of complete elements from plant DNA could not be used to amplify elements from the cell culture lines. As mentioned already, cultured cells are a mixed population containing cells with or without insertions at a particular locus and PCR will preferentially amplify the smaller product. Instead, primers designed for each *mPing* subtype were used with a specific primer located in the flanking sequence of each new insertion site of the C5924 cell line. Of 34 insertion sites tested in this way, twenty-three were found to contain *mPingB*, eight, *mPingC* and three, *mPingD*. Thus, all subtypes found in the plant before culturing were capable of transposition.

Two candidate autonomous elements

Like other MITE families, *mPing* elements have no coding capacity and as such, are incapable of catalyzing their own transposition. Thus, movement of *mPing* must be catalyzed by a transposase encoded *in trans*. To identify putative autonomous elements, the *mPing* sequence was used to query all available rice genomic sequence for related, but longer elements. A single element called *Ping*, is 5,353 bp and shares 253 bp and 177 bp, respectively, of its terminal sequences with *mPing*. That 429 of 430 bp are identical in the two elements suggests that *mPing* arose very recently from the larger element by internal deletion. Further blast searches using *Ping* as query led to the discovery of *Pong* (5,166 bp), with identical 15 bp TIRs and similar subterminal regions as *mPing* and *Ping* (Figure 4.1). Both *Ping* and *Pong* are, like *mPing*, flanked by a 3 bp TSD of the trinucleotide TTA. The only copy of *Ping* was found in Nipponbare (there are none in the available 93-11 sequence), but at least five copies of *Pong* were found in Nipponbare and at least three copies in 93-11. Six of these eight copies of *Pong* are complete and nearly identical (>99% identity) while two are incomplete due to sequence truncation. Finally, deletion derivatives of *Pong* (such as *mPong* MITEs) could not be found either by database searches or by PCR using primers located in the subterminal regions of *Pong* with a variety of rice genomic DNAs as template.

A new family of transposases in plants and animals

In addition to their termini, *Ping* and *Pong* also share sequence similarity in two blocks of internal sequence corresponding to the two major ORFs of each element (Figure 4.1). When used as queries in tBlastN searches against GenBank databases, both

ORFs yield numerous hits (E value < 1e-10) from a wide range of plants as well as animals and fungi (see supplemental data, Table 2). ORF2 homologs are abundant in plants, and most frequently found in organisms with large amounts of genomic sequence in databases: 82 hits (E value < e-46) from rice, 56 hits (E value <e-23) from *Arabidopsis* and over 100 hits (E value< e-36) from *Brassica oleracea*. ORF1 homologs are more divergent and less abundant. Most ORF1 homologs are located within 2 kb of ORF2 homologs where they are arranged in the same order and orientation as they are in *Ping* and *Pong*. Furthermore, several pairs of ORF1 and ORF2 homologs are flanked by TIRs and TSDs that are similar to those of *Ping* and *Pong* (not shown). It is therefore likely that each "pair" of ORF1 and ORF2 homologs belong to the same element.

The function of ORF1 is unclear. It has only very weak sequence similarity to Myb DNA binding domains (Pfam 7.3, E-value 0.002). ORF2, on the other hand, shares significant similarity with the transposase of the maize *PIF* element[16]. Significantly, *PIF* is also associated with a *Tourist*-like MITE family in maize (called *mPIF*)[16]. *PIF* belongs to a superfamily of transposons related to the bacterial IS*5* element (the *PIF*/IS*5* superfamily) with members identified in all three eukaryotic kingdoms. Examination of ORF2 revealed a putative DDE motif (Figure 4.1) containing a triad of acidic amino acids also found at the catalytic core of the transposases of bacterial insertion sequences as well as some eukaryotic transposons. The amino acid residues surrounding and including the DDE motif (referred to as N2, N3 and C1 domains by Rezsohazy et al[22]) are highly conserved among *Ping*/*Pong*, *PIF* and several IS*5*-like elements (Fig.1).

Transposition of *Pong*

Although *mPing* elements are clearly derived from *Ping*, the evidence suggests that *Ping* is not the autonomous element that mobilizes *mPing*. *Ping* was only detected as a single copy in Nipponbare: it is absent in the draft sequence of 93-11 (~84% of the genome) and from 20 of 24 rice cultivars tested by PCR (only four temperate *japonicas* harbor *Ping*: Nipponbare, Gihobyeo, JX 17 and Koshikari, data not shown). The apparent absence of *Ping* from all *indica* cultivars tested provides strong evidence that it could not be responsible for the movement of *mPing* elements in the *indica* cell line. *Pong,* in contrast, is present in multiple near-identical copies in both *indica* and *japonica*. In addition, the ORF1 in *Ping* appears to be truncated at the 5' end compared to its homologs, lacking at least 60 well conserved amino acids. Therefore, *Pong* seems to be a more likely candidate for an autonomous element that mobilizes *mPing*.

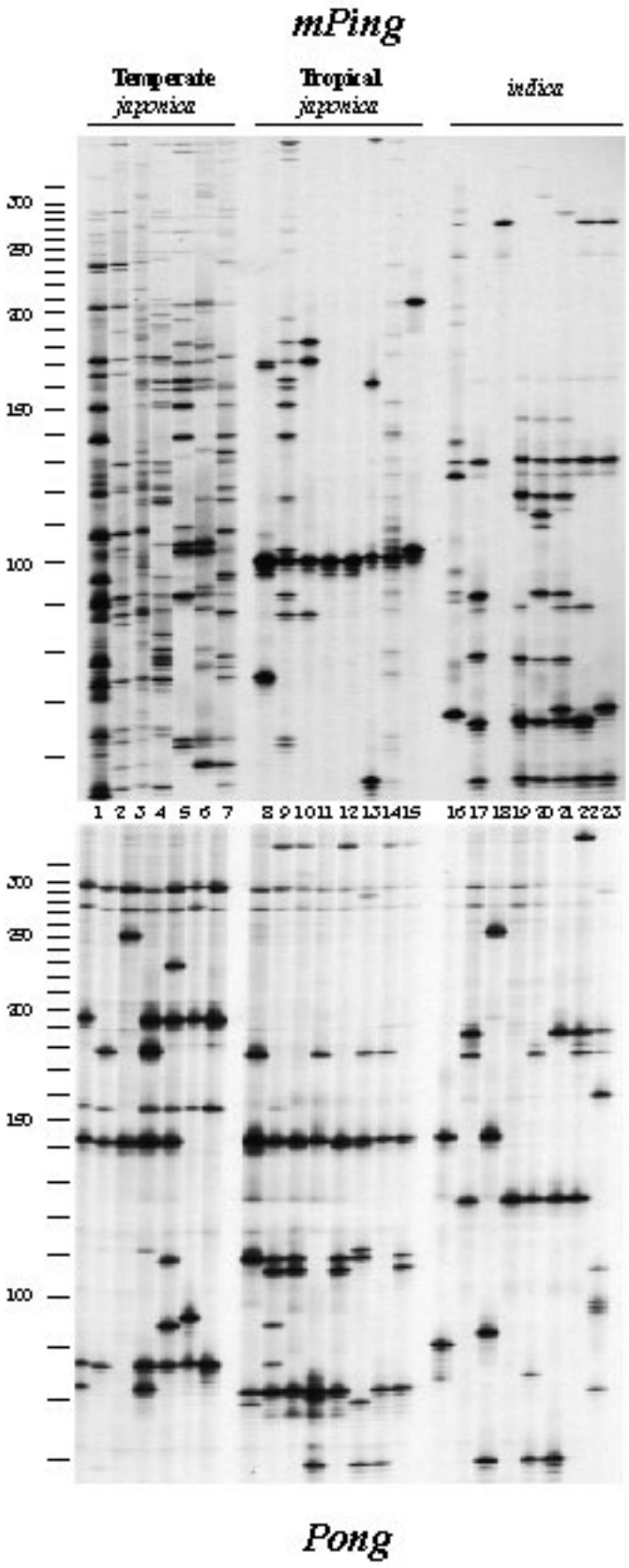
If *Pong* is the autonomous element responsible for the transposition of *mPing* elements, it should also be capable of transposition. By exploiting the sequence differences between *Pong* and *mPing*, PCR primers were designed to amplify *Pong* elements but not *mPing* in a TD assay. As can be seen in Figure 4.2, the results with the *Pong* primers mirror the *mPing* results. That is, the number of amplicons increased dramatically in the *indica* cell line but remained virtually the same in Nipponbare. The nature of the insertion sites and the inserted elements were determined in the same way as was done for *mPing*. Nine out of ten insertion sites were located in single copy sequences (see Supplemental Table 1). Eight newly inserted elements were successfully amplified by PCR and all were indistinguishable in size from *Pong*.

Amplification since domestication

The difference in the estimated copy number of *mPing* elements in a *japonica* (Nipponbare) and an *indica* (93-11) genome (70 vs. 12) suggested recent amplification of this MITE family, perhaps since domestication. To assess the timing of amplification, TD was undertaken with a panel of *O. sativa* DNAs and those from wild rice species to determine the approximate copy number and distribution of *mPing* and *Pong* elements (Figure 4.3). There is a recognizably different pattern of *mPing* products in *indica*, tropical *japonica* and temperate *japonica* cultivars, suggesting that the mobility of this element is diagnostic of the genetic and ecological differentiation that has come to characterize these groups since domestication (see Discussion)[23-25]. As can be seen in Figure 4.3, the temperate *japonicas* contain the largest number of different *mPing*-anchored amplicons while the tropical *japonicas* contain the fewest. This dramatic difference in *mPing* copy number between the two sub-groups of *japonica* is significant in light of evidence that the temperate and tropical cultivars are more closely related to each other than either is to *indica*[26-28].

On the other hand, the patterns of amplicons derived from *Pong* show less difference between the three groups of *O. sativa* cultivars indicating that *Pong* elements have not amplified significantly since domestication. As such, this data suggests that *Pong* elements may not be responsible for the activation of *mPing*. An alternative interpretation is that transposition of the larger *Pong* elements into genic regions would be more mutagenic than amplification of *mPing*. In this case, the amplification of an autonomous element would not necessarily correlate with the amplification of its nonautomous counterparts. In contrast, amplification of *Pong* may occur in cultured cells

Figure 4. 3. Autoradiograph of transposon display gels of *mPing* and *Pong*. Genomic DNAs: 1, Nipponbare; 2, Gihobyeo; 3, JX 17; 4, Koshikari; 5, Calrose; 6, Early Wataribune; 7, Shinriki; 8, Azucena; 9, Lemont; 10, Jefferson; 11, Moroberekan; 12, Rexoro; 13, Wab56-104; 14, Carolina Gold; 15, Kaybonnet; 16, C5924; 17, IR64; 18, Kasalath; 19, GuangLuAi4; 20, 93-11; 21, Tequing; 22, IR36; 23, Bs125. The migration of DNA markers is on the left in base pairs.


mPing
Temperate
japonica
Tropical
japonica
indica
300
250
200
150
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Pong

because they do not go through meiosis and self-fertilization. Thus all new insertion sites in cultured cells would be heterozygous with a wild type allele.

Discussion

A computational approach to identify active TEs

The recent availability of large quantities of genomic sequence has led to a shift from the genetic characterization of single elements to genome-wide analysis of transposable element populations[13]. It is now known that active elements comprise a tiny fraction of the TE complement of most multicellular organisms. In the absence of a mutation caused by a TE insertion, the task of determining whether a particular TE is active, inactive or epigenetically silenced presents a new challenge to those studying eukaryotic TEs. One response to this challenge has been to reconstruct active transposases based on phylogenetic analysis of a subset of inactive TE copies. This approach has been used successfully to reconstruct autonomous elements of the *Tc1*/*mariner* superfamily[30].

Here we report the first computational approach to isolate active TEs from a host genome. Although this methodology requires access to a large fraction of the genomic sequence of an organism, this need not be assembled sequence; unassembled, draft should suffice. To date, only a single active DNA transposon has been isolated from a vertebrate (*Tol2*[31]), and none from mammals. The protocol described in this report could

be applied to the available draft sequences of many organisms including human, mouse and zebrafish.

The first active DNA transposon from rice

The grasses, including rice, maize, sorghum, barley and wheat, have emerged in recent years as a collective model genetic system with rice and its two sequenced genomes as the anchor species[32]. To fully capitalize on this collective model, it will be necessary to hasten the speed of gene discovery in rice through the development of efficient tagging and knockout populations. Although maize TEs have been very useful in plant gene discovery[33], heterologous tagging systems in rice are still under development[34]. Similarly, the rice *Tos17* LTR retrotransposon appears to be of limited utility for these purposes since its induction has been restricted to cell culture which seems to induce a variety of other lesions (H. Sakai, DuPont/Pioneer, unpublished observation). Although the activity of *Ping-Pong* family is also induced in cell culture, the amplification of *mPing* elements in some rice cultivars suggests the existence of mechanism(s) that induce transposition during plant growth (see below). In addition to identifying these mechanisms, the ease of rice transformation should allow introduction of active *Pong* elements and permit mobilization of endogenous *Ping* and *Pong* members thus exploiting their targeted insertion into genic regions.

The first active MITE

Despite the widespread occurrence of MITEs in eukaryotes[13], this is the first report of their active transposition. The fact that *Tourist*-like MITEs account for about

60% of the rice MITEs (60,000 elements in over 50 families, N, Jiang and S. R. Wessler, unpublished data) motivated our search for potentially active families among rice genomic sequence. The discovery of an active MITE family and associated full-length elements marks the beginning of a new era of understanding these abundant and widespread components of eukaryotic genomes. Long-standing questions about the mechanisms underlying the birth, spread and death of MITEs can now be addressed experimentally. Some questions have already been answered such as discovering that the prevalence of MITEs in single copy regions primarily reflects targeting rather than selection. With the observation that thirty four out of thirty five newly transposed *mPing* elements and nine of ten newly transposed *Pong* elements inserted into single copy regions following cell culture, it is clear that at least some MITE families have an overwhelming insertion preference for genic regions. The results reported here also lead to some unexpected preliminary conclusions. One is that *Pong*, which is related but clearly distinct from *mPing*, is the most likely source of transposase for this family. The simultaneous transposition of *mPing* and *Pong* elements in cell culture coupled with failure to detect a more likely transposase source in the two draft sequences provides strong evidence that *Pong* is the autonomous element. These data suggest that one possible reason for the success of MITEs is an ability to be crossmobilized by related transposases.

Activation of *mPing* and *Pong* in cell culture

Both the *mPing* MITEs and the *Pong* elements were activated during cell culture as evidenced by the recovery of new insertions. Activation of *Tos17* retrotransposition in

the same cell line had been attributed to the induction of element transcription by some aspect of culturing[18]. Although the activation of TEs in cell culture appears to be a widespread phenomenon[35], to our knowledge this is the first report of the co-activation of retro and DNA elements. Since these element classes transpose by dramatically different mechanisms, their co-activation implies that culture conditions alter the global regulation of TE activity, perhaps through genome-wide changes in chromatin structure. These changes may, for example, allow the transcriptional machinery to access the regulatory sequences of both retro and DNA elements, and this may ultimately result in activation.

Transposition since domestication

In addition to their activation in cell culture, *mPing* elements have transposed more extensively in temperate vs. tropical *japonica* cultivars (Figure 4.3). The distinct patterns of TE amplicons seen with TD for the temperate cultivars suggest, but do not prove, that amplification occurred independently in the different cultivars. Temperate and tropical *japonicas* are believed to have diverged from a common ancestor since domestication - between 5000 and 7000 years ago[26,27,29]. These two varietal groups are adapted to radically different temperature and water regimes[24,25,36]. Tropical *japonica* cultivars (previously known as *javonica*) are broadly adapted to tropical and sub-tropical environments while temperate *japonicas* represent an evolutionary extreme, having been selected for productivity in cool, temperate zones with very short growing seasons (N. China, Korea, Japan, California).

Recently, the differential copy number of the retrotransposon *BARE1* in populations of wild barley was attributed to the ecologically distinct habitats of the host[37].

Although additional investigation is required to determine whether activation of *mPing* is associated with exposure to environmental stress, the implications of stress activation during the domestication of rice is intriguing and potentially profound. One potential source of activation common to both growth in temperate climates and cell culture is reduced temperatures, shown previously to increase the activity of plant and animal transposases up to 1000 fold[38]. Thus, in a scenario reminiscent of McClintock's genome shock theory[39], stress activation of *mPing* elements during the domestication of temperate *japonicas* followed by their preferential insertion into genic regions may have diversified these cultivars and hastened their domestication by creating new allelic combinations that might be favored by human selection. In this regard it is intriguing to note that Yano et al[40] speculated that an apparent gamma-ray induced insertion in an intron of the rice homolog of the *Constans* gene (*Hd1*) was responsible for a quantitative change in flowering time. Upon close inspection, this insertion is identifiable as an *mPing* element.

Methods

Plant material

DNA from cv Nipponbare (*japonica*), C5924 (*indica*), and cell cultures derived from these cultivars was provided by H. Hirochika (National Institute of Agrobiological Resources, Tsukuba, Ibaraki 305, Japan). DNA from other rice cultivars and wild species were from the McCouch lab (Cornell University) or from Gary Kochert (University of Georgia). Plant DNA was extracted as described[41].

Computer-assisted identification of repeat families

Nipponbare sequence (total 187 Mb) was downloaded from http://rgp.dna.affrc.go.jp on Dec 24, 2001, and 93-11 sequence (361 Mb) was downloaded from http://210.83.138.53/rice/index.php on Feb.25, 2002. Nipponbare sequence was used for a systematic identification of repeat families using an all-vs.-all comparison with WU-BLASTN2.0 (http://blast.wustl.edu) (options M=5 N=-11 Q=22 R=11 –kap E=0.0001 –hspmax 5000 wordmask=dust wordmask=seq maskextra=50). Alignments were then clustered into repeat families using RECON (http://www.genetics.wustl.edu/eddy/recon) with default options and further examined individually [using programs in the University of Wisconsin Genetics Computer Group program suite (GCG, version 10.1), accessed through Research Computing Resources (University of Georgia)].

The output of 1275 repeat families was searched manually for similarity with features of known MITEs including: short terminal inverted repeat (TIR), target site duplication (TSD), and size (N. Jiang, Z. Bao, S. R. Eddy and S. R. Wessler, manuscript in preparation). Repeat #1031 (called *mPing*) was identified as a *Tourist*-like MITE because of (1) size (430bp); (2) TIR similarity to other known *Tourist* MITEs (Figure 4.1) [6,10,14], and (3) TSD of TTA or TAA.

Elements related to *mPing* in Nipponbare (updated on Feb. 25, 2002) and 93-11 were identified with BLAST search (WU-BLASTN 2.0) using a derived consensus sequence of *mPing*. Two classes of potentially autonomous elements were recovered and called either *Ping* or *Pong* (Figure 4.1).

Transposon display (TD) and recovery of new insertion sites

TD was performed as described[20] with the following modifications. Element-specific primers were designed based on the sub-terminal sequences of *mPing*, *ID-1*, *SZ-2* and *Pong* and used in PCR with a variety of rice genomic DNAs from different cultivars, wild species, calli and cell lines. For Figure 4.2, the adapter primers are: *Mse*I + 0 for *mPing* display, *Mse*I + AT for *ID-1* display, *Bfa*I + C for *SZ-2* display, and *Bfa*I + 0 for *Pong* display. For both displays in Figure 4.3, *Mse*I + 0 was used as the adapter primer. The other primer was specific either for *mPing* (Fig.3 top) or *Pong* (Fig.3 bottom). PCR fragments from TD gels were recovered, reamplified and sequenced as described[6]. Sequences of primers and parameters for PCR are available on request.

Sequence analysis

Homologues of ORF1 and ORF2 (in *Ping* and *Pong*) were identified following database searches (tBlastn) using the blast server from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) against the NR, EST and GSS databases.

References

1. Arumuganathan, K. & Earle, E. D. Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* **9**, 208-218 (1991).
2. Burr, B. Mapping and sequencing the rice genome. *Plant Cell* **14**, 521-3. (2002).
3. Goff, S. A. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92-100. (2002).
4. Yu, J. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79-92. (2002).
5. Tarchini, R., Biddle, P., Wineland, R., Tingey, S. & Rafalski, A. The complete sequence of 340 kb of DNA around the rice *Adh1-adh2* region reveals interrupted colinearity with maize chromosome 4. *Plant Cell* **12**, 381-391 (2000).
6. Jiang, N. & Wessler, S. R. Insertion preference of maize and rice miniature inverted-repeat transposable elements as revealed by the analysis of nested elements. *Plant Cell* **13**, 2553-2564 (2001).
7. Bureau, T. E. & Wessler, S. R. *Tourist*: a large family of inverted-repeat element frequently associated with maize genes. *Plant Cell* **4**, 1283-1294 (1992).
8. Bureau, T. E. & Wessler, S. R. 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 (1994).
9. Bureau, T. E. & Wessler, S. R. *Stowaway*: a new family of inverted-repeat elements associated with genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6**, 907-916 (1994).

10. Bureau, T. E., Ronald, P. C. & Wessler, S. R. 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 (1996).

11. Mao, L. et al. Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. *Genome Res.* **10**, 982-990 (2000).

12. Feschotte, C., Zhang, X. & Wessler, S. in *Mobile DNA II* (eds. Craig, N., Craigie, R., Gellert, M. & Lambowitz, A.) 1147-1158 (American Society of Microbiology Press., Washington D.C., 2002).

13. Feschotte, C., Jiang, N. & Wessler, S. R. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329-41. (2002).

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

15. Feschotte, C. & Wessler, S. R. *Mariner*-like transposases are widespread and diverse in flowering plants. *Proc. Natl. Acad. Sci. USA* **99**, 280-285 (2002).

16. Zhang, X. et al. *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 (2001).

17. Bao, Z. & Eddy, S. R. Automated de novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res.* **In Press** (2002).

18. Hirochika, H., Sugimoto, K., Otsuki, Y. & Kanda, M. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci. USA* **93**, 7783-7788 (1996).

19. Van den Broeck, D. et al. *Transposon Display* identifies individual transposable elements in high copy number lines. *The Plant J.* **13**, 121-129 (1998).

20. Casa, A. M. et al. The MITE family *heartbreaker* (*Hbr*): molecular markers in maize. *Proc. Natl. Acad. Sci. USA* **97**, 10083-9 (2000).

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

22. Rezsohazy, R., Hallet, B., Delcour, J. & Mahillon, J. The IS*4* family of insertion sequences: evidence for a conserved transposase motif. *Mol Microbiol* **9**, 1283-95. (1993).

23. Chang, T. T. The origin, evolution, dissemination and diversification of Asian and African rice. *Euphytica* **25**, 425-441 (1976).

24. Morishima, H. & Oka, H. I. Phylogenetic differentiation of cultivated rice. XXVII. Numerical evaluation of the *indica-japonica* differentiation. *Japan. J. Breed.* **31**, 402-413 (1981).

25. Ueno, K. Differentiation of ecotypes in *Oryza sativa* L. 2. Characteristics of ecotypes: Japanese lowland and upland rice. *Bull. Inst. Agri. Res. Tohoku Univ.* **39**, 43-49 (1988).

26. Ting, Y. The origin and evolution of cultivated rice in China. *Acta Agron. Sinica* **8**, 243-260 (1957).

27. Glaszmann, J. C. Isozymes and classification of Asian rice varieties. *Theor. Appl. Genet.* **74**, 21-30 (1987).

28. Wang, Z. Y., Second, G. & Tanksley, S. D. Polymorphism and phylogenetic relationships among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. *Theor. Appl. Genet.* **83**, 565-581 (1992).

29. Matsuo, T., Futsuhara, Y., Kikuchi, F. & Yamaguchi, H. *Science of the Rice Plant* (Ministry of Agriculture, Forestry and Fisheries, Tokyo, Japan, 1997).

30. Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a *Tc1*-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501-10. (1997).

31. Kawakami, K., Shima, A. & Kawakami, N. Identification of a functional transposase of the *tol2* element, an *Ac*-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. *Proc. Natl. Acad. Sci. USA* **97**, 11403-8 (2000).

32. Bennetzen, J. The rice genome. Opening the door to comparative plant biology. *Science* **296**, 60-3. (2002).

33. Walbot, V. Strategies for mutagenesis and gene cloning using transposon tagging and T-DNA insertional mutagenesis. *Annu. Rev. Plant Physiol.* **43**, 49-82 (1992).

34. Greco, R. et al. Transposon insertional mutagenesis in rice. *Plant Physiol* **125**, 1175-7. (2001).

35. Grandbastien, M.-A. Activation of plant retrotransposons under stress conditions. *Trends in Plant Sciences* **3**, 181-189 (1998).

36. Glaszmann, J. C. & Arraudeau, M. Rice plant type variation: "*Japonica*" - "*Javanica*" relationships. *Rice Genet. Newslett.* **3**, 41-43 (1986).

37. Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. & Schulman, A. H. 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 (2000).

38. Carpenter, R., Martin, C. & Coen, E. S. Comparison of genetic behavior of the transposable element *Tam3* at two unlinked pigment loci in *Antirrhinum majus*. *Mol.Gen.Genet.* **207**, 82-89 (1987).

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

40. Yano, M. et al. *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene *CONSTANS. Plant Cell* **12**, 2473-2484 (2000).

41. McCouch, S. R. et al. Molecular mapping of rice chromosomes. *Theor. Appl. Genet.* **76**, 815-829 (1988).

42. Baba, A., Hasezawa, S. & Syono, K. Cultivation of rice protoplasts and their transformation mediated by *Agrobacterium* Spheroplasts. *Plant Cell Physiol.* **27**, 463-471 (1986).

Acknowledgements

We thank H. Hirochika (Molecular Genetics Department, National Institute of Agrobiological Sciences, Tsukuba, Japan) for the genomic DNAs from cv. C5924 and from C5924 and Nipponbare cell culture lines, Jeremy Edwards (Department of Plant Breeding, Cornell University) for help with database searches, Cédric Feschotte and

Ellen Pritham for critical reading of the manuscript, and Dawn Holligan for technical assistance. This study was supported by a grant from NSF to S.R.E., S.R.M. and S.R.W.

Supplemental Table 1. New insertion sites of *mPing* and *Pong* in C5924 cell line

Element	Adapter primer	Size of the fragment from TD	Hit in database	Position of insertion site[1]
mPing	*Mse* I + A	89	Contig482	21140
		101	Contig7079	5186
		119	Contig10017	7552
		129	Contig16063	4588
		137	Contig21518	1963
		148	Contig24556	2216
		165	Contig40966	517
		173	Contig4310	4497
		195	Contig391	157
		237	Contig34913	276
		247	Contig65368	166
		320	Contig23309	338
	Mse I + C	126	Contig909	21055
		152	Contig17922	3757
		172	Contig17222	4466
		187	Contig4708	2198
		238	Contig18737	4241
		244	Contig42209	1029
		256	Contig25343	2750
	Mse I + G	93	Contig17946	4149
		124	Contig17419	2467
		136	Contig23646	1322
		147	Contig1408	5777
		153	Contig2742	15355
		189	Contig17506	1995
		217	Contig15096	1964
		239	Contig63623	327
		262	Contig5138	10139
		302	Contig7692	1306
			Contig22012	962
	Mse I + T	143	AL731884[2]	10135[2]
		162	Contig13315	1527
		164	Contig2742	15355
		196	Contig63668	72
		230	Contig10585	7471
		267	Contig10749	6329
Pong	*Mse* I + 0	116	Contig7304	8347
		141	Contig2744	3314
		169	Contig1031	18894
		181	Contig18	48390
		203	Contig35873	1
		224	Contig3321	8382
			Contig16311	5969
			Contig24541	2774
		257	Contig494	22438
		318	Contig18755	2443
		340	Contig51	16732
		377	Contig6747	193

[1]The position of the first nucleotide of the trinucleotide TSD is listed
[2]GenBank accession number for Nipponbare. All others are from 93-11

Supplemental Table 2. Organisms with *Pong*-like transposases

Plant	Monocots	Rice (AP003986), Sorghum (AF114171), Barley (AJ001317), Wheat (AF459639), Maize (BH140750)
	Dicots	*Arabidopsis* (AC018660), Soybean (AF271796), *L. japonica* (AP004506), Suger beet (BI643302), Medicago (BG585958), Tomato (AW616734), Stevia (BG525000), Peppermint (AW255120), Brassica (BH493441)
	Algae	*Physcomitrella patens* (BJ164583), *Porphyra yezoensis* (AV436370)
Animal	Invertebrates	*C. elegans* (AF040643), *C. briggsae* (AC090524), *Drosophila* (AE003496), Silkworm (AV404936), Mosquito (AAAB01008967), Ciona (AV996094)
	Vertebrates	Zebrafish (AL591210), Mouse (BI247185), Pig (BF191773), Cow (BE668489), Human (AK057237)
Fungus		*F. neoformans* (AC068564), *N. crassa* (NC93G11)

.Sequences were found by tBlastn searches (one GenBank accession number from each species is shown as an example).

CHAPTER 5

CONCLUDING REMARKS (CONCLUSIONS)

Study TEs in the genomic era

In this dissertation, a combination of computational and experimental approaches were taken to study rice MITEs and LTR elements, including their target site preference, activity, and amplification. It is our belief that this approach is applicable to other TEs in rice and other organisms for which sufficient sequence data exist.

The abundance of MITEs in the rice genome

The relative abundance of different TEs was evaluated by analyzing 30 Mb rice sequence (7% of the rice genome). This is the first comprehensive analysis in rice that covers both genic regions and pericentromeric regions. It is shown that MITEs account for 5.1% of the genome, and they are the numerically the most abundant elements in the rice genome. Interestingly, there is no significant variation in the density of MITEs between pericentromeric regions and genic regions. This is in striking contrast to the fact that LTR elements and other DNA elements are much more (two to three fold) abundant in the pericentromeric regions than in the genic regions.

Consistent with previous studies, most of the rice MITEs fall into two groups based on the TSDs and TIRs: *Tourist*-like elements (about 60% of all MITEs) and *Stowaway*-like elements (about 40% of all MITEs). It is estimated there are a total of 60, 000 copies of *Tourist*-like elements accounting for 3% of the rice genome.

The first active MITE

The 430 bp *mPing* element, a *Tourist*-like element that originally identified by RECON, a program for *de novo* repeat family identification. *mPing* was found to be actively transposing in the cell culture line of the *indica* cultivar C5924. This is the first active MITE ever reported and will allow us to test various types of hypotheses regarding

the birth, spread, and amplification of MITEs. In addition, the differential amplification of *mPing* elements in different ecotypes of rice [~70 copies in Nipponbare (*japonica*) and ~12 copies in 93-11 (*indica*)] suggests that there might be a correlation between the amplification of elements and exposure to environment stress. Identifying the factors inducing the activity of *mPing* will further our understanding about the regulation of TE activity in rice.

A new family of transposase

Two full-length elements, *Ping* and *Pong* were found to be associated with *mPing*. Since *Ping* has almost identical TIRs and subterminal sequences as *mPing*, it is most likely the element that gave rise to *mPing*, while *Pong* shows only ~70% of sequence similarity to *mPing*. The *Ping/Pong* elements represent a new family of transposase that is distantly related to *PIF* elements in maize, which is associated with *mPIF*, a *Tourist*-like MITE with very high copy number (at least 6,000). Unlike the *PIF* elements, which contain only one ORF, the *Ping/Pong* family seems to have two ORFs. ORF2 shares significant similarity with known transposases, including *PIF* element, while ORF1codes for a protein of unknown function but has very weak similarity to Myb DNA binding proteins. Future research will test whether this ORF is dispensable and what role it may play in transposition.

Autonomous and non-autonomous elements

Recently, emerging evidence has shown that MITEs are a special group of non-autonomous DNA elements. In this study it was shown for the first time that a MITE (*mPing*) is a direct deletion derivative of a longer DNA element (*Ping*)(Figure 4.1). In this case, it is most conceivable that *Ping* is the autonomous element that mobilizes

mPing, similar to the relationship between *Ac* and *Ds*. However, the absence of *Ping* in C5924 and the co-activation of *mPing* and *Pong* in cell culture suggest that *Pong* is most likely the source of transposase responsible for the transposition of *mPing* in tissue culture.

Nevertheless, not all the evidence favored this hypothesis. For example, the amplification of *mPing* is not accompanied by the amplification of *Pong* in temperate *japonica*. In addition, no transposition of *mPing* was detected in the tissue culture of Nipponbare, whose genome harbored several copies of *Pong* elements that appear to be functional.

As discussed in chapter 4, the lack of co-amplification of *mPing* and *Pong* in temperate *japonica* may suggest that *Pong* is not the autonomous element. Alternatively, the lack of correlation is due to the highly mutagenic effect created by *Pong* insertion in genic regions, whereas *mPing* insertions might not be so deleterious because of its small size. On the other hand, the failure to detect new *mPing* insertions in Nipponbare callus can be explained if we assume reduced temperature is the key factor that activates the *Ping/Pong* family. Considering that Nipponbare (or its progenitor) has been exposed to reduced temperature for thousands of years, the callus of Nipponbare may not respond to the reduced temperature anymore. In addition, it is possible that, since *mPing* already amplified in Nipponbare to a certain extent, it was silenced by the host genome. Fortunately, some of these uncertainties or questions can be dissected and addressed experimentally. For example, we can test whether *mPing* can be cross-mobilized by *Pong* in vitro and whether an even lower temperature can activate the *Ping/Pong* family in Nipponbare in vivo.

Another interesting feature about the *Ping/Pong* family is that *Ping*, the putative progenitor of *mPing*, is only present in four out of the twenty-four rice cultivars tested. In Nipponbare, only a sole copy of *Ping* was detected. On the contrary, *Pong* is present in all the cultivars with several copies, yet no MITE was found to be associated with it. Based on this fact, it is tempting to hypothesize that the progenitors of MITEs may only be present in the genome transiently (in an evolutionary scale) and cannot be amplified. That might explain why for most of the MITEs in rice, no corresponding full-length element can be found (N. Jiang, X. Zhang, unpublished data). Again the hypothesis is testable: in rice there are 82 *Ping/Pong* related elements (chapter 4). A comparison of the copy number of elements with or without MITEs will answer the question whether the progenitors can be amplified or not. More importantly, the comparison of these two types of elements may provide some information about what features (for a full-length element) are important for the formation of MITEs.

Compared to non-autonomous DNA elements, non-autonomous LTR elements are relatively rare. However, the fact that *Dasheng*, a non-autonomous LTR element, is one of the highest copy number LTR elements in rice indicates that they have the potential to amplify and contribute to genome expansion. Since *Dasheng* is also one of the most recent elements in the rice genome, it is feasible to activate it in certain backgrounds and to study whether *RIRE2*, a *gypsy*-type element that related *Dasheng*, is the autonomous element responsible for the amplification of *Dasheng*.

Target site preference of TEs in the rice genome

MITEs

The association of MITEs with plant genes was noticed as early as the discovery of MITEs themselves, although one may argue that the early results were biased due to the gene-rich database (reviewed in Feschotte et al., 2002). Later, the study of *Hbr* elements in maize clearly demonstrated that *Hbr* elements are located more frequently in low copy sequences than in repetitive sequences (Zhang et al., 2000). However, this result alone does not necessarily prove that MITEs target low copy or genic regions since in some case, current distribution does not reflect the target site preference of TEs. In human genome, most *Alu* elements are found to be located in GC-rich regions. Yet most of the recent *Alu* insertions are located in AT-rich regions, just as their autonomous element, the *L1* element (Lander et al., 2001). Hence it is speculated that there is some selective advantage for the insertions in GC-rich regions (Lander et al., 2001). In this study, we studied the target site preference of *mPing* by characterizing the new insertions induced in cell culture. Cell culture is a system with little selective pressure because they do not go through meiosis and self-fertilization; thus there is a wild type allele for every new insertion. In addition, not all the genes are required to be functional in cell culture. As such, the new insertions isolated from the *indica* cell line have been subject to little selection and should largely reflect the targeting specificity of the element of interest. Based on our study on *mPing* and *Pong*, it appears that the *Ping/Pong* family does have an strong insertion preference for low copy sequences. Furthermore, the fact that two out of six *mPing* insertions are located in the putative exons of genes suggests that the absence of MITEs in exons of genes in the databases is a result of selection. In other

words, the 100, 000 MITEs present in the current rice genome represent only a fraction of the MITEs ever inserted in the genome.

One question that arises from this conclusion is that if MITEs (at least some of them) have such as a strong preference into genic regions, how may they survive and thrive? One possibility, as discussed above, is that the effect of MITE insertions might be not as deleterious as the insertions of their autonomous elements, probably due to the size effect. The second possibility, as shown in this dissertation, is that some MITEs, especially some *Tourist*-like elements, have developed self-insertion preference, which may minimize deleterious effects on the host.

Dasheng

Unlike MITEs, which are associated with genes, the majority of *Dasheng* elements are located in the pericentromeric regions but are not associated with centromeres. Circumstantial evidence suggests that the biased distribution is a result of targeting heterchromatin rather than a preference for repetitive sequences. Since there are twenty wild rice species with genomes varying 3-fold in size, an interesting question is whether the copy number of *Dasheng* is correlated with genome size and if it is, whether the distribution pattern is the same as that in rice.

Targeting preference is family-specific

An important conclusion from the analysis of nested elements is that the targeting preference of MITEs varies from family to family. For example, *Stowaway*-like elements show a very different pattern from that of *Tourist*-like elements in terms of self-targeting. Even among the *Tourist*-like elements significant variation was found among different families. As a result, the strong target specificity for single copy regions by *mPing* does

not necessarily imply all the MITE families have an equally strong target specificity, although this is likely to be true for most MITEs. Consistent with this notion, Mao et al (2000) reported that only about 10% of the *Explorer* (a *Tourist*-like element) insertions are associated with putative rice genes while the average ratio for all MITEs is over 50%. Moreover, the relatively even distribution of MITEs along chromosomes compared to other TEs (Table 2.3) suggests either that some MITEs do not have a similar preference as that shown by *mPing* or that the pericentromeric region was initially gene-rich. A comparison of MITEs composition between genic regions and pericentromeric regions may answer the question whether different MITEs target different regions.

Like the situation with MITEs, different LTR elements have different target site preference. For instance, *Tos17*, a low copy number LTR element in rice, preferentially inserts into low copy number regions, which is in contrast to the distribution of *Dasheng*. In this regard, it is intriguing to notice that in rice MITEs are the only known elements that target genic regions and still achieve high copy numbers. Given both the high copy number and the target specificity, the impact of MITEs on the rice genome, and probably on the genomes of other plants, is beyond our imagination.

References

Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics, Nat Rev Genet *3*, 329-41.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome, Nat *409*, 860-921.

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

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 (Hbr)* into genic regions of maize, PNAS *97*, 1160-1165.

APPENDIX

P INSTABILITY FACTOR: AN ACTIVE MAIZE TRANSPOSON SYSTEM ASSOCIATED WITH THE AMPLIFICATION OF *TOURIST*-LIKE MITES AND A NEW SUPERFAMILY OF TRANSPOSASES[3]

[3] Published in Zhang, X., Feshotte, C., Zhang, Q., Jiang, N., Eggelston, W.B. and S.R. Wessler 2001 *Proc. Natl. Acad. Sci. USA*. 98:12572-12577. Reprinted here with permission of publisher

Abstract

Miniature inverted-repeat transposable elements (MITEs) are widespread and abundant in both plant and animal genomes. Despite the discovery and characterization of many MITE families, their origin and transposition mechanism are still poorly understood largely because MITEs are nonautonomous elements with no coding capacity. The starting point for this study was *P instability factor* (*PIF*), an active DNA transposable element family from maize that was first identified following multiple mutagenic insertions into exactly the same site in intron 2 of the maize anthocyanin regulatory gene *R*. In this study we report the isolation of a maize *Tourist*-like MITE family called *miniature PIF* (*mPIF*) that shares several features with *PIF* elements including identical terminal inverted repeats, similar subterminal sequences and an unusual but striking preference for an extended 9 bp target site. These shared features indicate that *mPIF* and *PIF* elements were amplified by the same or a closely related transposase. This transposase was identified through the isolation of several *PIF* elements and the identification of one element (called *PIFa*) that co-segregated with *PIF* activity. *PIFa* encodes a putative protein with homologs in *Arabidopsis*, rice, sorghum, nematodes and a fungus. Our data suggest that *PIFa* and these *PIF*-like elements belong to a new eukaryotic DNA transposon superfamily that is distantly related to the bacterial IS*5* group and are responsible for the origin and spread of *Tourist*-like MITEs.

Introduction

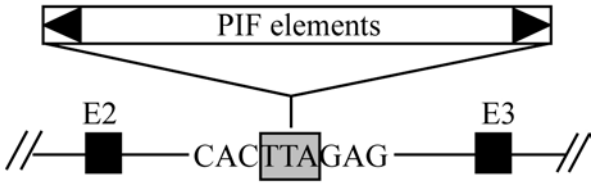
Transposable elements (TEs) have been divided into two classes according to their transposition intermediate. Class 1 (RNA) elements transpose via an RNA intermediate and most have either long terminal repeats (LTR-retrotransposons) or terminate at one end with a poly A tract (LINEs and SINEs). Class 2 (DNA) elements transpose via a DNA intermediate and usually have terminal inverted repeats (TIRs). In eukaryotes, class 2 families, such as the maize *Ac*/*Ds* or the *Drosophila P* elements, consist of autonomous and nonautonomous members. Autonomous elements encode transposase that binds to *cis*-acting sequences residing in the terminal regions of both autonomous and nonautonomous elements to catalyze their transposition [for review (1)]. Nonautonomous elements usually arise from autonomous elements by point mutations and/or internal deletion(s). Integration of most TEs results in a duplication of the target site, so that each element is flanked by a target site duplication (TSD) of conserved length and sometimes sequence (1).

Miniature inverted-repeat transposable elements (MITEs) are a recently described group of TEs that have been found in a wide range of plants and animals (2-10). In plants, the majority of characterized MITE families can be divided into two groups based on similarity of their TIRs and TSDs: there are *Tourist*-like MITEs and *Stowaway*-like MITEs. Despite the abundance of MITEs in many genomes (~2% of *C.elegans* and ~6% of rice), their origin and transposition mechanism remains poorly understood (11-13). All MITE families have a suite of common structural features including high copy number (~500 to 10,000 per haploid genome), conserved within-family length (<500-bp)

and sequence and target site preference. The fact that many MITE families share their TIRs, TSDs and, in one case, even internal sequences with larger TEs encoding transposases has been interpreted to mean that MITEs originated from autonomous DNA elements (6, 9, 10, 14, 15).

To date, no MITE family has been shown to be actively transposing. In the absence of activity, it has been difficult to determine how MITEs are generated and how they attain such high copy numbers. For this reason the focus of this study is an actively transposing family of class 2 elements from maize called *P instability factor* (*PIF*). *PIF* elements were first discovered as six independent insertions into exactly the same site in intron 2 of the maize *R* gene (Fig. 1a)(16). These six elements inserted in both orientations and fell into two structural classes, referred to as *PIF-6* (5.2 kb) and *PIF-12* (2.3 kb). Of particular interest was the finding that *PIF* was related to a 364 bp MITE-like sequence that appeared to have inserted into another maize TE (16). In this study we demonstrate that this 364 bp sequence is the founding member of a *Tourist*-like MITE family called *miniature PIF* (*mPIF*). In addition to their sequence similarities, *mPIF* and *PIF* elements insert into a sequence-specific 9-bp palindrome. The structure of the *PIF* family was further investigated through the isolation of several family members including the putative autonomous *PIF* element (*PIFa*). *PIFa*-like elements were identified by database searches in rice, *Arabidopsis* as well as nematodes and a fungus. These data provide evidence for a superfamily of elements that may be responsible for the amplification of *Tourist*-like MITEs in the genomes of plants and animals.

a



b

mPIF: $C_{77}W_{58}C_{77}$ [T_{97} $T_{91}A_{97}$] $G_{52}W_{68}$ G_{61}

c

PIF: $C_{69}W_{92}C_{53}$ [$T_{100}T_{95}A_{95}$] $G_{92}W_{100}G_{85}$

Fig. 1: Target site preference of *PIF* and *mPIF* elements**. (a)** Six *PIF* elements inserted independently into exactly the same position in the second intron of the maize *R* gene (16). Triangles represent *PIF* TIRs and black rectangles represent exons 2 and 3 of *R*. **(b)** Consensus extended target site derived from a comparison of the sequences flanking 30 *mPIF* elements. **(c)** Consensus extended target site derived from a comparison of the sequences flanking 14 *PIF* elements. Gray rectangles indicate the trinucleotide duplicated upon element insertion (the TSD). Numbers represent the percentage of times that a nucleotide appeared at that position.

Materials and Methods

Genetic stocks, DNA extraction and library construction.

All strains were derived from the maize inbred W22.

r-sc:124Y2902: A derivative of *R-sc:124* (*R* allele conferring pigmentation of aleurone, embryo and coleoptile) with a 2.3 kb *PIF* insertion in the second intron of the *Sc* component (16) causing loss of kernel pigmentation. Excision restores kernel pigmentation.

r-g:14qs131: A derivative of *R-r:standard* that contains only the *P* component (*R* gene that confers pigmentation of roots, coleoptiles, seedling leaf tips and anthers). Insertion of a 5.2 kb *PIF* into intron 2 (16) eliminates pigmentation in these tissues while element excision restores color.

Stable 2*:* A *PIF*-inactive strain homozygous for *r-sc:124Y2902* (provided by J. Kermicle, University of Wisconsin), derived as following: a *PIF*-active strain homozygous for *r-sc:124Y2902* was crossed to a *PIF*-inactive strain homozygous for *r-r* (*R* allele conditioning colorless kernels and colored plants) and several resulting ears were found to have very few or no spotted or solidly pigmented kernels, indicating low or no *PIF* activity. Seeds from each ear were grown, self-pollinated and *PIF*-inactive strains homozygous for the *r-sc:124Y2902* chromosomes obtained. Stable 2 is one such strain that lost *PIF* activity, as no spotted kernels were observed above background when it was self-pollinated. However, spotted kernels were readily observed at normal frequency when Stable 2 was crossed to strains with *PIF* activity.

Strain R: a *PIF*-active strain homozygous for the *r-g:14qs131* allele.

Plant DNA was extracted from young leaves as described (17). The small insert genomic library was constructed from strain R as described (18).

Generation of a population segregating for *PIF* activity.

Stable 2 (*r-sc:124Y2902*, *PIF*-inactive, see above) was crossed with strain R (homozygous for *r-g:14qs131*, *PIF*-active) to produce a population of plants called SR (*PIF*-active) (Fig. 2). Spotted kernels from this population (due to somatic excision of the *PIF* element from *r-sc:124Y2902*) were grown and crossed to Stable 2 to obtain a population (called SRS, Fig. 2) segregating for *PIF* activity. 15 SR and 28 SRS plants were generated from spotted kernels and 13 Stable 2 plants were generated from unpigmented kernels. DNA was extracted from young leaves and analyzed by transposon display (see below).

Transposon display and recovery of gel bands.

Transposon display (TD) was performed as described (19) with the following modifications. *PIF*-specific PCR primers (PR1, PR2, PF1 and PF2, see Fig. 2) were derived from the *PIF* subterminal sequences to specifically amplify the flanking sequences of *PIF* but not *mPIF* elements (primer sequences available upon request). The primer combinations used were: PR2 and *Mse*I+0 for 5' end pre-selective amplification, PR1 (labeled with ^{33}P) and *Mse*I+0 for 5' end selective amplification, PF2 and *Bfa*I+0 for 3' end pre-selective amplification, PF1 (labeled with ^{33}P) and *Bfa*I+0 for 3' end selective amplification. The final annealing temperature was 55^{o}C (PCR cycle parameters available upon request). Radioactive PCR products were recovered from polyacrylamide

gels as described (http://tto.biomednet.com/cgi-bin/tto/pr) and amplified by PCR with the same primers and under the same conditions as those used for the respective TD selective amplifications.

PCR amplification and sequencing of *PIF* elements.

PIF0.4, PIF1.1, PIF1.6 and *PIF1.7* were amplified from total genomic DNA by PCR using *Taq* DNA polymerase (Perkin Elmer) with primers derived from the *PIF* subterminal sequences such that they would not amplify *mPIF* elements. Longer *PIF* elements were amplified using Elongase (GIBCOBRL) under conditions that favor the production of long products (20) with primers derived from *PIF* sequences internal to the *PIF0.4* deletion breakpoints (Fig.2). Amplification of the *PIFa* element employed primers derived from flanking genomic sequences (PCR cycle parameters and primer sequences available upon request).

PCR products were cloned using the TA Cloning Kit from Invitrogen (Invitrogen Corporation, Carlsbad, CA) according to manufacturer's instructions. All sequencing reactions were performed by the Molecular Genetics Instrumentation Facility of the University of Georgia. The consensus sequence for 32 *mPIFs* and the sequence of *PIFa* were deposited in GenBank under accession numbers AF416298-AF416329 and AF412282, respectively.

Computational analysis.

GenBank database searches were performed with the various BLAST servers available via the National Center for Biotechnology Information

(http://www.ncbi.nlm.nih.gov). The gene structure of *PIFa* and *PIF*-like elements was predicted by the NetGene2 (http://www.cbs.dtu.dk) (21); NetStart 1.0 (http://www.cbs.dtu.dk); (22); and FGENESH (http://genomic.sanger.ac.uk/gf/gf.html) (23) programs. Protein sequences were obtained from GenBank or by conceptual translation of predicted genes and aligned using CLUSTALX (24). Phylogenetic analysis was carried out with PAUP* version 4.0b8 (25), using both the neighbor-joining and maximum parsimony methods with default parameters.

Results and Discussion

mPIF is a MITE family.

Several features of the previously identified 364-bp *PIF*-related sequence including short TIRs and a 3-bp TSD rich in A and T residues were reminiscent of MITEs. Southern blot analysis confirmed that this sequence was highly repetitive in maize but not in sorghum or rice (data not shown). To estimate the copy number of related elements in the maize genome and to isolate more copies for analysis, a genomic library (average insert size 1.5 kb) was prepared from maize inbred line W22 and screened with the 364 bp sequence. The hybridization of 369 plaques out of 1.1×10^5 screened (representing $\sim 1.6 \times 10^5$ kb or ~6 % of the genome) provided an estimate of $\sim 6 \times 10^3$ copies of this sequence per haploid genome (369 / 6 % = 6,150). In contrast, the copy number of the larger *PIF* elements was estimated by Southern blot analysis to be ~25 (WB. Eggleston, unpublished data).

Thirty-two of the 369 positive plaques were randomly chosen for further analysis. Thirty of the 32 contained complete elements that were, on average, 358 bp, had perfect 14-bp TIRs and displayed over 90% sequence identity. All elements were rich in A and T residues (71%) and had no significant coding capacity. Twenty-eight of the thirty full-length elements were flanked by a conserved 3-bp TSD (TTA/TAA). We named this new MITE family *miniature PIF* (*mPIF*). A consensus sequence was derived from 32 *mPIFs* and was deposited in GenBank (accession nos. AF16298-AF416329). Based on the TSD and TIR sequences, *mPIF* can be classified as a typical *Tourist*-like MITE family (Supplemental Fig. 1) (26) . Comparison between the consensus *mPIF* sequence and previously characterized *PIF* elements (16) reveals identical TIRs and similar subterminal sequences extending for ~100 bp from the termini (overall similarity of ~70%). The most internal 150 bp of *mPIF* elements was not related to *PIF* elements.

Identical extended target site preference for *mPIF* and *PIF* elements.

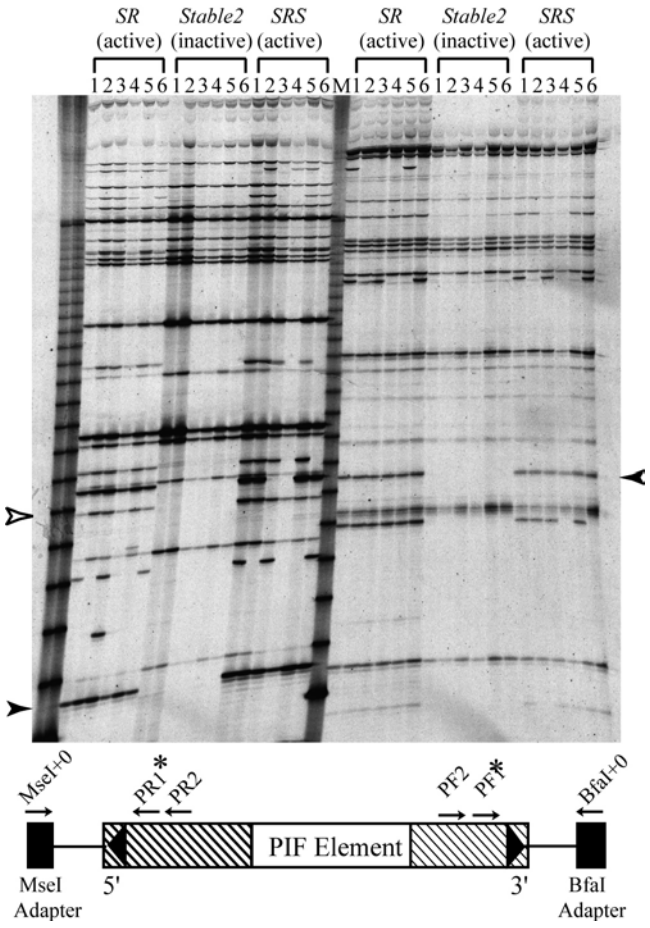
The insertion of six of the larger *PIF* elements into exactly the same site in the *R* gene prompted us to examine whether *mPIF* and *PIF* insertion sites were conserved beyond the TSD. Comparison of the sequences flanking the 30 full-length *mPIF* elements revealed remarkable conservation of an extended 9-bp target site centered on the TSD (Fig. 1). Significantly, this sequence matches the insertion site in the *R* gene.

To determine whether the larger *PIF* elements have the same target site preference, sequences flanking some of the other ~25 *PIF* elements in the genome were recovered by using the transposon display (TD) procedure. TD is a modification of the AFLP procedure (19, 27) that generates PCR products anchored in a transposon and a

flanking restriction site (see *Materials and Methods*). To this end, PCR primers were designed to amplify genomic sequences flanking *PIF* (and not *mPIF*) termini. Approximately 50 PCR products, 25 from each end, were displayed after gel electrophoresis (Fig. 2). This corresponds to about 25 *PIF* elements which is in agreement with the prior copy number determination (WB Eggleston, unpublished data). A total of 14 PCR products were recovered, sequenced and used to derive a consensus target site that was found to be identical for both *mPIF* and *PIF* elements (Fig. 1c).

Extended target site preference has been reported for several bacterial transposons (28, 29) and there is evidence that some eukaryotic class 2 elements may have some preference beyond the TSD (30, 31). However, to our knowledge, *PIFs* and *mPIFs* display the longest and most specific target site preference ever documented among eukaryotic class 2 TEs. Additional support for the existence of a specific 9-bp insertion site comes from the fact that the sequences flanking *mPIF* elements judged to have inserted most recently (based on highest sequence identity to the *mPIF* consensus and insertion site polymorphism among maize strains) are most similar to the consensus target sequence (data not shown). What is particularly surprising is that despite targeting such a specific insertion site, *mPIF* elements still managed to attain a higher copy number than virtually all other characterized class 2 elements. Given that a 9-bp sequence is expected to occur, on average, about once in 250 kb, ~10,000 copies of this sequence are predicted to be in the maize genome. It is remarkable to consider that most of these sites may be occupied by *mPIF* elements.

Fig. 2: Transposon display (TD) analysis of a population segregating for *PIF* activity. Only a subset of the population analyzed by TD is shown. *PIF* TD was carried out from both the 5' end (left half of gel) and the 3' end (right half of gel). Arrowheads indicate PCR products that co-segregate with activity. Open arrowhead indicates PCR products that did not co-segregate with activity in other plants (not shown). SR: plants heterozygous for the autonomous *PIF* element. Stable 2: plants without *PIF* activity. SRS: *PIF*-active plants from the cross between SR and Stable 2 (see *Materials and Methods* for details). M: 30~330 bp molecular weight marker. A schematic representation indicating the positions of the PCR primers is also shown. Arrows represent PCR primers and stars indicate primers labeled with ^{33}P, black rectangles represent *Bfa*I or *Mse*I adapters and hatched rectangles represent terminal regions conserved in all sequenced *PIF* elements.

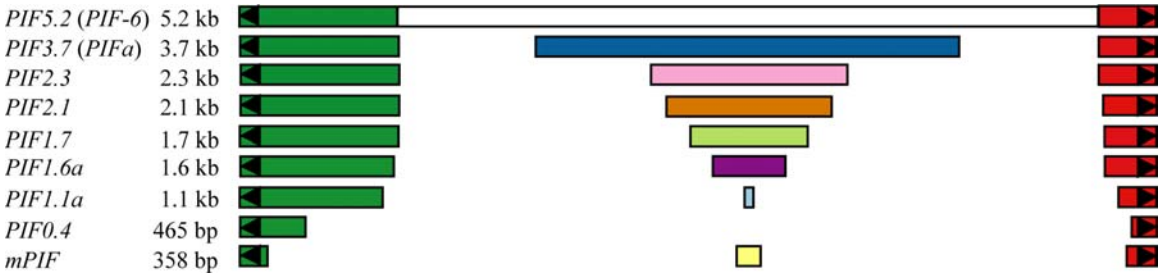

SR (active)
Stable2 (inactive)
SRS (active)
SR (active)
Stable2 (inactive)
SRS (active)
1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6 M 1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6
MseI+0
PR1
PR2
PF2
PF
BfaI+0
PIF Element
MseI Adapter
5'
3'
BfaI Adapter

Structure of *PIF* family members.

Since target site preference has been shown, in a few cases, to be a function of the transposase (28, 32), the existence of a common 9-bp target for both *mPIF* and *PIF* elements strongly suggests that their transposition reactions are catalyzed by the same or a closely related transposase. For this reason, it was thought that isolation of additional *PIF* elements might lead to the isolation of the autonomous element responsible for the origin and amplification of both *mPIF* and *PIF* elements.

The two *PIF* elements at the *R* locus (*PIF5.2* and *PIF2.3*) are nonautonomous elements that only share their terminal sequences (Fig. 3) (16). To isolate additional *PIF* family members, PCR primers derived from *PIF* sequences internal to the TIRs were used to amplify genomic DNA. Primers were designed to amplify *PIF* but not *mPIF* elements. The predominant PCR product was of 483 bp and was found to be a deletion derivative of a longer *PIF* element (*PIF0.4*). Three other products of 1.1 kb, 1.6 kb and 1.7 kb were also cloned and sequenced. To isolate longer elements that may not have competed successfully in the initial PCR reactions, primers derived from sequences internal to the deletion breakpoint of *PIF0.4* were employed, along with PCR conditions that favor the production of longer products. This procedure led to the isolation of eight additional *PIF* elements ranging from ~1.1 kb to ~5.2 kb, of which four were completely sequenced. All of the elements (except *PIF04*) are highly conserved (>90%) in their terminal regions, however, the internal sequences are dissimilar and serve to distinguish distinct subfamilies (Fig. 3). Unfortunately, none of these elements were considered autonomous since computer analysis failed to detect significant coding capacity or any similarity to known transposases.

Fig. 3: Schematic representation of the structure of the *PIF* transposon family. Elements are named according to their length and are drawn to scale. Only one element from each subfamily is shown. *PIF5.2* is previously described as *PIF-6* and *PIF2*.3 is 98% identical to *PIF-12* (16). Black triangles represent TIRs. Green and red rectangles represent the terminal sequences conserved in all elements (see text). Open rectangle indicates the fact that the internal region of *PIF5.2* was not sequenced. Dark blue, pink, brown, light green, purple and light blue rectangles represent internal regions unique to each subfamily. Yellow rectangle represents the internal region of *mPIF*.


PIF5.2 (PIF-6) 5.2 kb
PIF3.7 (PIFa) 3.7 kb
PIF2.3 2.3 kb
PIF2.1 2.1 kb
PIF1.7 1.7 kb
PIF1.6a 1.6 kb
PIF1.1a 1.1 kb
PIF0.4 465 bp
mPIF 358 bp

Isolation of the *PIFa* element.

A genetic approach to isolate an autonomous *PIF* element was employed involving the application of transposon display to a population segregating for *PIF* activity (see *Materials and Methods*). Genomic DNA from plants grown from spotted kernels (+*PIF* activity) and colorless kernels (-*PIF* activity) were analyzed using primers facing outward from the *PIF* termini (Fig. 2). Only one product from each end co-segregated with *PIF* activity. The sequences derived from these products were used to design PCR primers from the genomic sequences adjacent to the *PIF* termini (20 bp and 25 bp from the 5' and 3' termini, respectively) and used in a single reaction to amplify genomic DNA (Fig. 4a). One product of 3.7 kb was amplified from the *PIF* active but not the *PIF* inactive plants, thus confirming that the co-segregating TD products were derived from sequences flanking the same element (designated *PIFa*) (Fig. 3, 4a).

Additional evidence for the co-segregation of *PIFa* with *PIF* activity was obtained by carrying out amplification reactions with different primer pairs. Primers derived from the internal region of *PIFa* and from sequences flanking the *PIFa* insertion site should amplify a 900-bp product if *PIFa* is at the locus. A product of this size was obtained from four *PIF* active strains that had served as parents for progeny without *PIF* activity where, presumably, *PIFa* had been lost following meiosis (Fig. 4b, lanes 3-6). The absence of *PIFa* from *PIF*-inactive plants is indicated by the failure to amplify a 900-bp product from 12 plants whose DNA was grouped into two pools of six (Fig. 4b, lanes 1-2). Finally, two of the 28 *PIF*-active SRS plants did not have *PIFa* at the original locus, as determined by TD, possibly because *PIFa* had transposed to another site in the genome. To test if this was the case, PCR primers were designed from an internal region

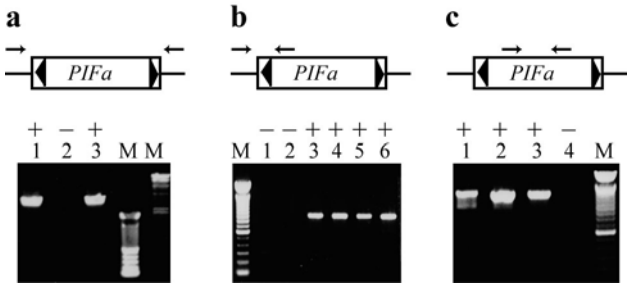
of *PIFa* that is not present in other *PIF* elements. Amplification of DNA from these two active plants along with 14 inactive plants (derived from parents heterozygous for *PIFa*, see Materials and Methods) confirmed the presence of *PIFa* in the former but not in the latter (Fig. 4c). This result also demonstrated that the loss of *PIFa* correlated with the loss of *PIF* activity.

***PIF* is member of a superfamily of DNA transposons.**

The sequence of *PIFa* revealed a 3,728 bp element that, like all other *PIF* elements, contains the conserved terminal regions (Fig.2). The central 2.5 kb, not found in other *PIF* elements, contains two ORFs longer than 100 a.a. (Fig. 5a). Only the first ORF (313 a.a.) produced significant hits (E value $>10^{-15}$, with sequences from *Arabidopsis* and rice BACs and with two *Sorghum bicolor* entries) when used as a query in TBLASTN searches with translated sequences (complete list available upon request). Amino acid identities among these sequences range from 25 to 50% (45-65% similarity) over 100-250 a.a. tracts. Further TBLASTN searches with some plant products and multiple iterations with PSI-BLAST also uncovered significant similarity with two putative proteins from *Caenorhabditis elegans*, one from its close relative *C. briggsae*, and one from the basidiomycete fungus *Filobasidiella neoformans* (see Fig. 5 legend for accession numbers). Finally, limited but significant homology was detected with several transposases encoded by bacterial insertion sequences of the IS*5* group (Supplemental Fig. 2) (29) .

The evolutionary relationship among these proteins was analyzed by aligning the translated product from the complete *PIFa* ORF (313 a.a.) with other *PIF*-like putative

Fig. 4: *PIFa* is present in *PIF*-active plants but absent from *PIF*-inactive plants. Agarose gels of PCR products are shown. A "+" or "-" indicates the presence or absence, respectively, of *PIF* activity in the strains used for genomic DNA isolation. **(a)** Amplification of the entire *PIFa* element using primers derived from flanking genomic sequences. A 3.7-kb product was obtained from *PIF*-active (SR, lane 1 and SRS, lane 3), but not from *PIF*-inactive (Stable 2, lane 2) plants (see *Materials and Methods* for strain designations). **(b)** Amplification of genomic DNA from the *PIFa* insertion site. Products of the appropriate size (~900 bp) were obtained from *PIF*-active plants that have served as the progenitors for the *PIF*-inactive Stable 2 (lanes 3~6), but not from 12 Stable 2 plants grouped into two pools of six each (lanes 1~2). **(c)** PCR amplification of an internal region of *PIFa* not present in any other sequenced *PIF* element. Products of appropriate size (~1.3 kb) were obtained from SR (lane 1), as well as two SRS plants that do not have *PIFa* at this locus (SRS15 and SRS31, lanes 2~3), suggesting that *PIFa* has transposed but may still be present in the genome. No product was obtained from a pool of 14 Stable 2 plants (lane 4). Arrows represent the positions of PCR primers, triangles represent TIRs and lines represent *PIFa* flanking sequences. M: molecular weight marker.


a
PIFa
+ − +
1 2 3 M M
b
PIFa
− − + + + +
M 1 2 3 4 5 6
c
PIFa
+ + + −
1 2 3 4 M

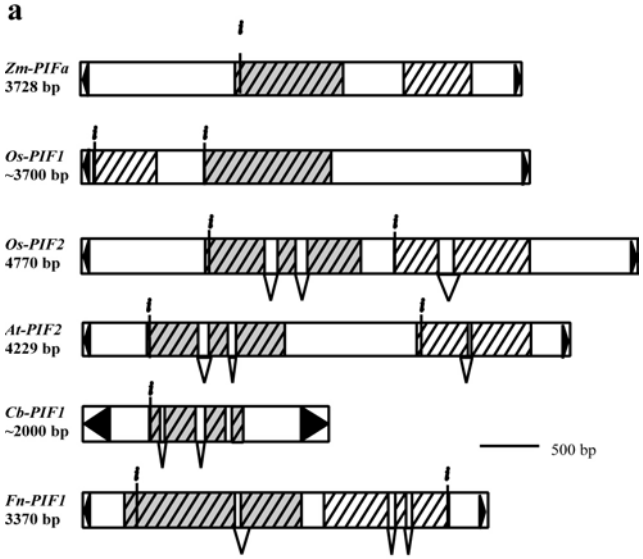
proteins identified by database searches and generating phylogenetic trees. A CLUSTALW multiple alignment (Supplemental Fig. 2) revealed several well-conserved amino acid blocks, most notably among the plant products. Both the neighbor-joining and parsimony methods produced trees with similar topologies (Fig. 5c). Bacterial transposases and eukaryotic homologs group separately while plant and nematodes products form distinct monophyletic clades within the eukaryotic sequences. Nonetheless, branch lengths between and within kingdoms indicate that there is extensive diversity in this protein superfamily (Fig. 5c).

To determine if the *PIF*-like coding sequences were part of TEs, sequences flanking these hits were searched for structural features reminiscent of transposons. Several *Arabidopsis* and rice ORFs as well as the *C. briggsae* ORF are flanked by inverted repeats (IRs) that share significant sequence similarity with the maize *PIF* TIRs (Fig. 5b). In addition, these IRs, like *PIF* TIRs, are flanked by a direct repeat of the TTA trinucleotide. Furthermore, BLAST searches reveal that each of these *PIF*-like elements belongs to a repeat family in their respective genomes (called *At-PIF, Os-PIF* and *Cb-PIF*, respectively) where they display high intra-family sequence similarities (>90%). Interestingly, many *PIF*-like family members are short internally deleted copies of homogeneous size that resemble *mPIF* and other MITEs (Supplemental Fig. 3). All of these MITEs are *Tourist*-like, as they possess TIRs similar to some of the previously described *Tourist* elements and are flanked by a 3-bp A/T-rich sequence that is probably the TSD.

Fig. 5: The *PIF*-IS*5* superfamily of transposons. **(a)** Structure and coding capacity of *PIFa* and several *PIF*-like elements. ORFs larger than 100 a.a. are schematically depicted as hatched rectangles. The predicted intron/exon structure is shown as well as the putative initiation codon (indicated by "*i*"). TIRs are represented by black triangles. Rectangles shaded in grey represent ORFs sharing significant similarity (i.e. *PIF*-like transposases). Other ORFs are not related, although the *At-PIF2* downstream gene can encode a protein that has several paralogs in the *Arabidopsis* genome. However, these paralogous sequences are not associated with a *PIF*-like transposase (data not shown). In addition, *Os-PIF1* and *Cb-PIF1* contain nested insertions of a variety of repetitive sequences, thus making it difficult to unambiguously determine element length. For this reason, the length shown for these *PIF*-like elements is approximate. Species, GenBank accession numbers and coordinates are: *Zm-PIFa*, *Z. mays* xxxxxxxx; *Os-PIF1*, *Oryza sativa* AC025098, 101769-109139; *Os-PIF2*, *O. sativa* AP01111, 2889-7665; *At-PIF2*, *A. thaliana* TM021B04, 16996-21224; *CbPIF1*, *C. briggsae* AC090524, 69398-71455; *Fn-PIF1*, *Filobasidiella neoformans* AC068564, 3620-6989. **(b)** Putative target site duplication (TSD) and terminal inverted-repeats (TIRs, size in bp) of *PIF*-like elements. **(c)** Phylogenetic relationship of putative *PIF*-like proteins and IS*5* transposases. The unrooted tree was constructed with the neighbor-joining method from a CLUSTALX alignment, which includes the complete product conceptually translated from the largest ORF of *Zm-PIFa* (313 a.a.), various eukaryotic homologs identified by database searches and several representatives of the IS*5* group of transposases (29 ; see Supplemental Fig. 2). Bootstrap values (1000 replicates) support the grouping of the plant, nematode and bacterial proteins. *Ce-PIF2* is identical to the product recently reported as the *Tc8.1*

putative transposase by Le et al. (2001). Species and GenBank accession numbers are: *At-PIF1*, *A. thaliana* AB017067; *Ce-PIF1*, *C. elegans* CEF57G4; *Ce-PIF2*, *C. elegans* CELF14D2; IS*5*, *Escherichia coli* J01735; IS*L2*, *Lactobacillus helveticus* X77332; IS*702*, *Calothrix sp.* X60384; IS*470*, *Streptomyces lividans* AB032065; IS493, *S. lividans* M28508.

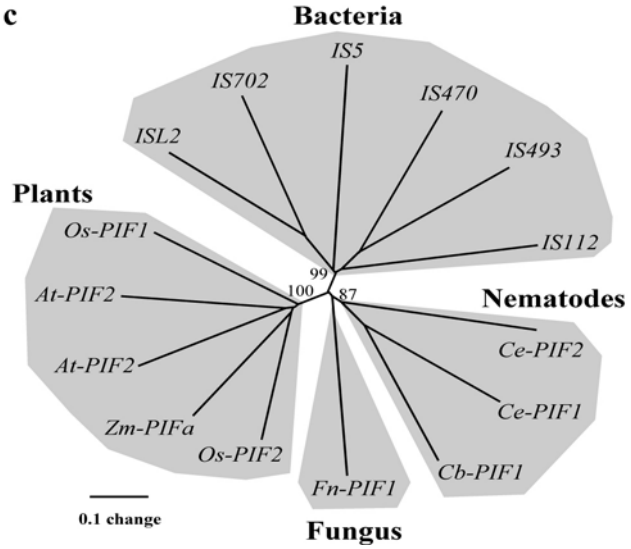
a



b

Species	Name	TSD	TIR	
Z. maize	*Zm-PIFa*	TTA	14	GGGCCCGTTTGTTT
O. sativa	*Os-PIF1*	TTA	15	GGCCTYGTTTGGCTG
O. sativa	*Os-PIF2*	ATA	14	GGGGTTGTTTGGTT
A. thaliana	*At-PIF2*	TTA	20	GGKGGTGTTATTGGTTAGTG
C. briggsae	*Cb-PIF1*	TTA	270	TGCCGCGTTCAAAG
F. neoformans	*Fn-PIF1*	ATT	18	AGGGGTAGACAAAATGCA

c



Several features shared by *PIF* and *PIF*-like elements strongly suggest that together they represent a new superfamily of eukaryotic DNA transposons that arose from a common ancestor. These features include their homologous coding regions as well as TIRs of similar length and sequence shared by all plant *PIF*-like elements. In addition, all the *PIF*-like elements identified in this study generate a 3-bp TSD and, in all but one case, the duplication is TTA (it is AAT for the *F. neoformans* element). Consensus extended target sites cannot be derived for the *PIF*-like elements due to the small number of elements identified by database search. However, since the length and sequence of the TSDs are functions of the transposase (28, 29, 33, 34), the similarities noted among the *PIF*-like elements suggest that their transposases are related not only evolutionarily, but also functionally.

As mentioned above, coding regions shared by *PIF*-like elements are also related to the transposases encoded by the IS*5* group of bacterial insertion sequences (Fig. 5b). Interestingly, many IS*5* elements also create 3-bp TSDs upon insertion [e.g. subgroup ISL*2*, IS*427* and IS*1031*; (29)] and some display a preference for TNA targets (e.g. subgroup IS*1031*). Moreover, IS*1031A* from *Acetobacter xylinum* has an extended target preference for the motif TCTNAR, with TNA being duplicated (29). This consensus matches that of *PIF* elements. Taken together, these data support the view that *PIF*-like elements belong to a new eukaryotic DNA transposon superfamily that is distantly related to the bacterial IS*5* group.

PIF-like elements belong to the same superfamily as *Harbinger*, a previously identified sequence that was discovered as part of an extensive search for repeats in the *Arabidopsis* genome (35). Our database searches indicate that *Harbinger* represents only

one of the multiple *PIF* lineages present in the *Arabidopsis* genome (unpublished data). Kapitonov and Jurka (35) also reported similarities between the putative transposase of *Harbinger* and several hypothetical proteins from rice, sorghum and *C. elegans* as well as the transposases of IS*5* elements. Based on these similarities, they proposed to classify *Harbinger* as a member of a new superfamily of DNA transposons. However, in their study, only *Harbinger* was characterized as a "bonafide" transposable element (i.e. with TIRs and other features of DNA elements). More recently, one of the putative IS*5*-related transposases identified by Kapitonov and Jurka (35) in *C. elegans* was shown to be part of a transposable element associated with *Tourist*-like MITE family members (36). Our results extend these findings by showing that IS*5*-related TE families are present in diverse eukaryotic organisms, including maize, rice, *C. briggsae* and a fungus. As the maize *PIF* was the first family identified in eukaryotes (16) and the only one with demonstrated activity, we propose to name this new superfamily of DNA transposons the *PIF*-IS*5* superfamily.

Conclusions.

The origin and spread of MITEs throughout plant and animal genomes remains largely a mystery despite the characterization of many MITE families and the availability of thousands of MITE sequences. A major reason for this is that MITEs are nonautonomous elements with no significant coding capacity. As such, associations between MITE families and potentially autonomous elements has, until this study, been restricted to computer-assisted searches for larger related elements in genomes that are completely sequenced like *A. thaliana* or *C.elegans* (15, 35-37). We call this the

"bottom-up" approach since the sequences of nonautonomous family members are utilized as queries to identify potentially autonomous family members. The major limitation of this approach is that nothing is known about the genetic activity of the larger elements and hence of the entire TE family.

In contrast, the starting point for this study was *PIF*, an active class 2 TE family. Similarity between *PIF* elements and a 364 bp sequence led to the discovery of *mPIF*, a *Tourist*-like MITE family, the discovery of an unprecedented 9-bp palindromic target sequence for *PIF* and *mPIF* elements, and the identification of the putative autonomous *PIFa*, which encodes a transposase that is related to transposases encoded by other TEs in plant, animal and bacterial genomes. We call this the "top-down" approach since a family of genetically active elements was used to identify a MITE family. The association of a MITE family with a genetically active system should ultimately furnish the biochemical tools necessary to address, experimentally, the larger questions regarding the origin and spread of MITEs.

Acknowledgements.

We thank Dr. Jerry Kermicle for providing the maize strain Stable 2 and Drs. Kelly Dawe and Michael Scanlon for helpful discussions. This work was supported by a grant from the National Institutes of Health to S.R.W.

References

1. Capy, P., Bazin, C., Higuet, D. & Langin, T. (1998) *Dynamics and Evolution of Transposable Elements* (Landes, Austin, TX).

2. Bureau, T. E. & Wessler, S. R. (1992) *Plant Cell* **4,** 1283-1294.

3. Bureau, T. E., Ronald, P. C. & Wessler, S. R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 8524-8529.

4. Casacuberta, E., Casacuberta, J. M., Puigdomenech, P. & Monfort, A. (1998) *Plant J.* **16,** 79-85.

5. Morgan, G. T. (1995) *J. Mol. Biol.* **254,** 1-5.

6. Oosumi, T., Garlick, B. & Belknap, W. R. (1996) *J. Mol. Evol.* **43,** 11-18.

7. Tu, Z. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 7475-7480.

8. Izsvak, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H. & Hackett, P. B. (1999) *J. Mol. Evol.* **48,** 13-21.

9. Feschotte, C. & Mouchès, C. (2000) *Gene* **250,** 109-116.

10. Tu, Z. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 1699-1704.

11. Surzycki, S. A. & Belknap, W. R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 245-249.

12. Tarchini, R., Biddle, P., Wineland, R., Tingey, S. & Rafalski, A. (2000) *Plant Cell* **12,** 381-391.

13. 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. & Wing, R. A. (2000) *Genome Res.* **10,** 982-990.

14. Smit, A. F. A. & Riggs, A. D. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1443-1448.

15. Feschotte, C. & Mouchès, C. (2000) *Mol. Biol. Evol.* **17,** 730-737.

16. Walker, E. L., Eggleston, W. B., Demopulos, D., Kermicle, J. & Dellaporta, S. L. (1997) *Genetics* **146,** 681-693.

17. McCouch, S. R., Kochert, G., Yu, Z. H., Khush, G. S., Coffman, W. R. & Tanksley, S. D. (1988) *Theor. Appl. Genet.* **76,** 815-829.

18. Zhang, Q., Arbuckle, J. & Wessler, S. R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 1160-1165.

19. Casa, A. M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S. & Wessler, S. R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10083-10089.

20. Marillonnet, S. & Wessler, S. R. (1998) *Genetics* **150,** 1245-1256.

21. Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouze, P. & Brunak, S. (1996) *Nucleic Acids Res.* **24,** 3439-3452.

22. Pedersen, A. C. & Nielsen, H. (1997) *Plant Mol. Biol.* **5,** 226-233.

23. Salamov, A. A. & Solovyev, V. V. (2000) *Genome Res.* **10,** 516-522.

24. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997) *Nucleic Acids Res* **25,** 4876-4882.

25. Swofford, D. L. (1999), ***PAUP*: phylogenetic analysis using parsimony and other methods.*** **Sinauer, Sunderland, MA**

26. Bureau, T. E. & Wessler, S. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1411-1415.

27. Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, B. A., Pot, J., Peleman, J., Kuiper, M. & Zabeau, M. (1995) *Nucleic Acids Res.* **23,** 4407-4414.

28. Craig, N. L. (1997) *Annu. Rev. Biochem.* **66,** 437-474.

29. Mahillon, J. & Chandler, M. (1998) *Microbiol. Mol. Biol. Rev.* **62,** 725-774.

30. Ketting, R. F., Fischer, S. E. & Plasterk, R. H. (1997) *Nucleic Acids Res.* **25,** 4041-4047.

31. Liao, G. C., Rehm, E. J. & Rubin, G. M. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 3347-3351.

32. Pribil, P. A. & Haniford, D. B. (2000) *J. Mol. Biol.* **303,** 145-159.

33. Beall, E. L. & Rio, D. C. (1997) *Genes Dev.* **11,** 2137-2151.

34. Plasterk, R. H. A., Izsvák, Z. & Ivics, Z. (1999) *Trends Genet.* **15,** 326-332.

35. Kapitonov, V. V. & Jurka, J. (1999) *Genetica* **107,** 27-37.

36. Le, Q. H., Turcotte, K. & Bureau, T. (2001) *Genetics* **158,** 1081-1088.

37. Le, Q. H., Wright, S., Yu, Z. & Bureau, T. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 7376-7381.

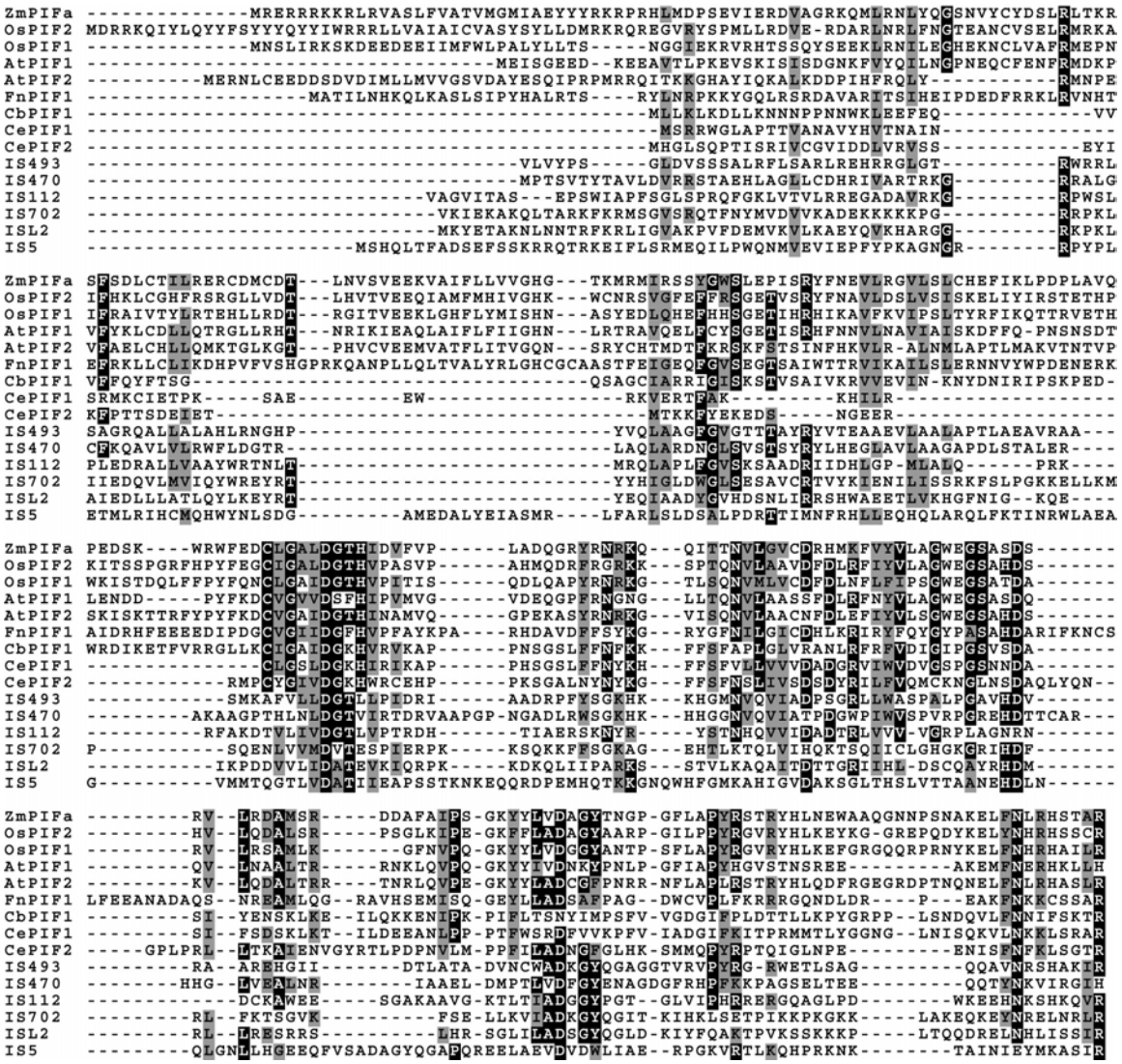
Supplemental data

Supplemental Fig. 1: Similarities in terminal inverted-repeats (TIRs) and target site duplications (tsd) among some*Tourist*-like MITEs.

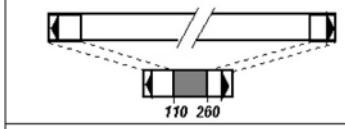
species	family	tsd	TIRs bp	sequence	reference
Z. mays	*mPIF*	TTA	14	GGGCCCGTTTGTTT	(*)
Z. mays	*TouristA*	TTA	14	GGCCTTGTTCGGTT	(26)
Z. mays	*TouristB*	TTA	14	GGCCTTGTTTAGTT	(26)
Z. mays	*TouristC*	TTA	14	GGGCCTGTTTAGAT	(26)
Z. mays	*TouristD*	TTA	14	GGGGGTGTTTGGTT	(26)
Z. mays	*Hbr*	NNN	14	GGGCCTGTTTGGTT	(18)
O. sativa	*Gaijin*	TTA	16	GGSYGTGTTTAGTTCA	(3)
O. sativa	*Wanderer*	TTA	11	GGCTGTGTTCG	(3)
O. sativa	*Castaway*	TTA	13	GGGCCCATTTGAA	(3)
O. sativa	*Ditto*	TTA	15	GAGCARGTTYAATAG	(3)
A. thaliana	*ATTIRX1A*	TTA	15	GGGGATGTATTCAATC	(35)
A. thaliana	*ATTIR16T3*	TTA	15	GGGGGTGTTATTGGTT	(35)

(*):this paper

Supplemental Fig. 2: CLUSTALW multiple alignment of putative *PIF*-like products and transposases of the IS*5* group. Shading is based on a simple majority rule. Identical residues are in white on a black background and similar residues are in black on a grey background. See legend of Fig. 5 for accession numbers.


ZmPIFa
OsPIF2
OsPIF1
AtPIF1
AtPIF2
FnPIF1
CbPIF1
CePIF1
CePIF2
IS493
IS470
IS112
IS702
ISL2
IS5

Supplemental Fig. 3: *PIF*-like elements are associated with MITE-like (*mPIFs*) derivatives in their respective genomes. Dashed lines between potentially full-length elements and their derivatives delimits regions of significant sequence similarity (70-95%). Numbers below refer to the positions where sequence homology drops off. Grey boxes in small derivatives indicate regions without obvious similarity with the larger *PIF*-like element. Black triangles represent terminal inverted-repeats. Sizes of mPIFs are from consensus sequences that were derived from alignments of multiple copies. Copy numbers were extrapolated from genomic library screening (*Zm-mPIF*) or database mining (*Os-PIF2, At-PIF2, Cb-PIF1*). Accession numbers and coordinates of *PIF*-like elements are given in the legend of Fig. 5, those of *mPIFs* are available upon request.



Name	Size (bp)	Copy number
Zm-PlFa	3728	1
Zm-mPlF	358	~6,000
Os-PlF2	4770	ND
Os-mPlF2	270	~150
At-PlF2	4229	1
At-mPlFs	410	~20
Cb-PlF1	~2000	ND
Cb-mPlF1a	244	~50
Cb-mPlF2b	60	~30