CHARACTERIZATION OF THE *PIF*/*PONG* SUPERFAMILY OF DNA TRANSPOSONS AND THEIR RELATIONSHIP WITH *TOURIST*-LIKE MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS (MITES)

by

XIAOYU ZHANG

(Under the direction of Susan R. Wessler)

ABSTRACT

Transposable elements (TEs) are ubiquitous in the genomes of all organisms characterized to date and are major components of plant genomes. The two classes of TEs, class 1 (RNA) and class 2 (DNA) elements, are distinguished by their transposition intermediate. DNA elements are further classified into superfamilies based on homology in the transposases catalyzing their transposition. Miniature inverted-repeat transposable elements (MITEs) are short, nonautonomous DNA elements widespread and abundant in plants and animals. Because MITEs lack coding capacity, their origin and transposition mechanism remained largely unknown. The researches described in this dissertation were directed to understand the origin and transposition mechanism of *Tourist*-like MITEs.

Chapter 2 describes the identification of a *Tourist*-like MITE family (*mPIF*) and the characterization of an active DNA element family, *P Instability Factor* (*PIF*) in maize (*Zea mays*). Significant similarities shared by *PIF* and *mPIF* elements indicate *mPIF* MITEs are nonautonomous members of the *PIF* family. Identification of a transposase-encoding *PIF* element (*PIFa*) through genetic analysis led to the discovery of a new superfamily of DNA elements (called *PIF*/*Pong*) widespread in eukaryotes. These results indicate that members of the *PIF*/*Pong* superfamily are responsible for the origin and amplification of *Tourist*-like MITEs.

Chapter 3 characterizes the *PIF*/*Pong* superfamily with regard to its distribution, evolution and relationship with *Tourist*-like MITEs. A comprehensive survey identified hundreds of *PIF*/*Pong*-like transposases from plants, animals and fungi, and the evolutionary relationships of these transposases were examined through phylogenetic analyses. Relationships between *PIF*/*Pong*-like elements and *Tourist*-like MITEs were investigated in rice (*Oryza sativa*), where the vast majority of *Tourist*-like MITEs was found related to *PIF*/*Pong*-like elements.

Chapter 4 was directed to understand the contribution of TE proliferation to the genome size expansion of *Brassica oleracea* and the cause for the low TE content of the *Arabidopsis* genome. The TE diversity, abundance and phylogeny of *Arabidopsis* and *B. oleracea* were compared. The results indicate that the amplification of both class 1 and class 2 TEs contributed significantly to the *B. oleracea* genome size expansion and that reduced TE proliferation is largely responsible for the low TE content and small genome size of *Arabidopsis*.

INDEX WORDS: Transposable elements, *PIF*, *Pong*, *Tourist*-like MITEs, Genome evolution, Genome size expansion, *Brassica oleracea*, *Arabidopsis*

CHARACTERIZATION OF THE *PIF*/*PONG* SUPERFAMILY OF DNA TRANSPOSONS AND THEIR RELATIONSHIP WITH *TOURIST*-LIKE MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS (MITES)

by

XIAOYU ZHANG

B.S., University of Science and Technology of China, China, 1997

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

2003

CHARACTERIZATION OF THE *PIF*/*PONG* SUPERFAMILY OF DNA TRANSPOSONS AND THEIR RELATIONSHIP WITH *TOURIST*-LIKE MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS (MITES)

by

XIAOYU ZHANG

Approved:

Major Professor: Susan R. Wessler

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

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2003

DEDICATION

To My Family.

ACKNOWLEDGEMENTS

I am deeply indebted to my major professor and mentor, Sue Wessler, for her guidance, support and patience. I am also grateful 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: Amy Bouck, Alexandra Casa, Bo Edwards, Cedric Feschotte, Dawn Holligan, Yun Hu, Ning Jiang, Ed Kentner, Zenaida Magbanua, Alex Nagel, Mark Osterlund, Ryan Peeler, Ellen Pritham, Liangjiang Wang and Qiang Zhang, for their help and friendship. Finally, I thank my parents, Jinchi Zhang and Shanyuan Yu, my sister, Chu Zhang, and my wife, Rong Wu, for their unconditional love and support, and my son, Kevin Zhang, for filling my life with joy.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Classification of Transposable Elements.

Transposable elements (TEs) are mobile genetic entities that are able to transpose from one chromosomal locus to another in their host genome. Since their initial discovery by Barbara McClintock in the 1940s (McClintock 1948; McClintock 1949), TEs have been found in the genomes of all organisms characterized (Capy et al. 1998). TEs are divided into two classes based on their transposition intermediates. Transposition of class 1 (RNA) elements occurs via an RNA intermediate and involves a reverse transcription process (Boeke and Corces 1989; Bingham and Zachar 1989). There are two types of class 1 elements: those flanked by long terminal direct repeats (LTRs) are called LTR-retrotransposons, while those that do not possess a LTR but terminate at their 3' end with a poly A tract are referred to as non-LTR retrotransposons (Doolittle et al. 1989; Xiong and Eickbush 1990). LTR retrotransposons have been further divided into *Ty1*/*copia*-like and *Ty3*/*gypsy*-like groups that differ from each other in both the order and the level of sequence homology of their encoded proteins (Xiong and Eickbush 1990). Non-LTR retrotransposons include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (Hutchison et al. 1989; Deininger 1989). LINEs encode proteins that are necessary for their transposition, whereas SINEs do not possess protein-coding capacity and their transposition is catalyzed by the proteins encoded by LINEs.

Class 2 elements transpose via a DNA intermediate and usually have short terminal inverted repeats (TIRs). They are organized into families based on their encoded transposase (TPase), such as the *Ac*/*Ds* and *Spm*/*dSpm* families

in maize and the *P* element family in *Drosophila*. Each family consists of autonomous and nonautonomous members. Autonomous elements encode TPase that binds to *cis*-acting sequences residing in the terminal regions of both autonomous and nonautonomous elements to catalyze their transposition. Nonautonomous elements do not encode functional TPase and usually arise from autonomous elements by point mutations and/or internal deletion(s). Integration of 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 conserved sequence (Capy et al. 1998).

Miniature Inverted-repeat Transposable Elements (MITEs).

Miniature inverted-repeat transposable elements (MITEs) are short, nonautonomous DNA elements. Different MITE families share structural but not sequence similarity, including small size (~100 – 500 bp), intrafamily homogeneity in size as well as sequence and significant target site preference (TAA for *Tourist* and TA for *Stowaway*) (Feschotte et al. 2002a; Feschotte et al. 2002b). Since their initial discovery in grasses, MITEs have been found in a wide range of plant, animal and fungal genomes where they have accumulated to very high copy number (thousands or tens of thousands per family) (Bureau and Wessler 1992; Morgan 1995; Bureau et al. 1996; Oosumi et al. 1996; Tu 1997; Casacuberta et al. 1998; Izsvak et al. 1999; Feschotte and Mouchès 2000; Tu 2001; Feschotte et al. 2002b; Feschotte et al. 2002a). In addition, many MITE families were preferentially found in the proximity of genes (located in introns or

5’ and 3’ untranslated regions) (Bureau and Wessler 1992; Bureau and Wessler 1994; Feschotte et al. 2002b; Feschotte et al. 2002a). The abundance of MITEs coupled with their genic association suggests that they may have played an important role in generating allelic diversity by affecting gene expression (Wessler et al. 1995; Wessler et al. 2001).

The basis for the classification of DNA elements is their TPase. MITEs cannot be classified in this manner as they do not possess any coding capacity. In addition, until very recently, no MITE had been found to be actively transposing (Jiang et al. 2003). Furthermore, none of available MITE sequence revealed clear-cut relationship with TPase-encoding DNA elements (e.g., as their deletion derivative). Instead, most of the tens of thousands of MITEs identified from plant and animal genomes have been broadly divided into two groups based on sequence similarities in their TIRs and TSDs: *Tourist*-like (TAA TSDs) and *Stowaway*-like (TA TSDs). It has been recently shown that *Stowaway*-like MITEs are associated with members of the *Tc1*/*mariner* family. The origin of *Tourist*-like MITEs is the main focus of the research described in this dissertation.

DNA Element Superfamilies.

Some DNA element families share significant sequence homology in their TPases, which is often accompanied by similarities in their TIR sequence as well as the length and sometimes the sequence of their TSDs. Accordingly, DNA element families are grouped into superfamilies. Examples are the *hAT* superfamily (Kunze and Weil 2002), *Mutator*-like elements (MULEs) (Yu et al.

2000; Mao et al. 2000; Lisch 2002), the CACTA superfamily (Rubin et al. 2001; Kunze and Weil 2002) and the *Tc1*/*mariner* superfamily (Plasterk and van Luenen 2002; Feschotte and Wessler 2002).

The *hAT* superfamily. This superfamily of DNA transposons was named after *hobo* from Drosophila, *Ac* from maize and *Tam3* from snapdragon (Calvi et al. 1991). Related elements have been identified from a wide range of plants, animals and fungi (Rubin et al. 2001; Kunze and Weil 2002). Structural features shared by *hAT*-like elements include short TIRs (12, 11 and 12 bp for *hobo*, *Ac* and *Tam3*, respectively) and 8-bp TSDs (Streck et al. 1986; Muller and Clarkson 1980; Hehl et al. 1991). *hAT*-like elements encode a single TPase that is necessary and sufficient for their transposition. The most conserved region in *hAT*-like TPases is a ~50-aa. domain located at the C terminus that has recently been shown to be involved in the dimerization of TPase proteins (Calvi et al. 1991; Rubin et al. 2001). The TIRs as well as subterminal repetitive motifs of *Ac* element are involved in TPase binding (Kunze and Starlinger 1989; Becker and Kunze 1997; Weil and Kunze 2000), whereas for another active element, *Hermes* (from housefly), subterminal repetitive motifs are absent and its TIRs serves as the TPase-binding substrate (Warren et al. 1994). Phylogenetic analyses have shown that *hAT*-like elements in plants, animals and fungi group into three distinct clades, indicating that they were derived from a common ancestor that radiated into modern species via vertical inheritance (Rubin et al. 2001).

***Mutator*-like elements (MULEs).** The maize *Mutator* elements have relatively long TIRs (220 bp) and generate 9-bp TSDs upon integration (Bennetzen 1996). The autonomous *Mutator* element, *MuDR*, encodes two genes (*mudrA* and *mudrB*) that are convergently transcribed (Hershberger et al. 1995). The *mudrA* gene encodes the TPase (MURA), whereas the function of *mudrB* coding product (MURB), if any, has not been clearly demonstrated (Lisch et al. 1999). A number of elements related to the maize *Mutator* have been identified from plants and recently from fungi, together comprising a large and diverse superfamily called *Mutator*-like elements (MULEs). Interestingly, unlike *MuDR*, none of the MULEs encodes a MURB homolog (Yu et al. 2000; Singer et al. 2001). The length and sequence of MULE TIRs vary significantly. Most MULEs are flanked by 9-bp TSDs although slight variation in the length of TSDs exists (8-10 bp) (Yu et al. 2000). The most conserved region in the TPases of MULEs is a 130-aa. segment that also shares weak similarity with the TPases of some bacterial insertion sequences (Eisen et al. 1994). Phylogenetic analysis of MULEs indicated that the evolution of MULEs is characterized by vertical transmission of multiple distinct lineages co-existing in grass genomes (Lisch et al. 2001).

CACTA-like elements. The nucleotide sequence of the termini of the maize *En*/*Spm* family, the *Tam1* family from snapdragon and a large number of related families is 5'-CACTA-3'. This group of DNA element families are collectively referred to as the CACTA superfamily (Kunze and Weil 2002). CACTA-like elements are flanked by 3-bp TSDs and share several structural features,

including short TIRs (13 bp for *En*/*Spm* and *Tam1*) and highly structured subterminal regions containing repetitive motifs (Pereira et al. 1986; Bonas et al. 1984). One unique feature of CACTA-like elements is the ability to encode two proteins that are produced by the alternative splicing of the same mRNA precursor (Masson et al. 1989; Nacken et al. 1991). One protein, called TNP2-like, is well-conserved among all CACTA-like elements, whereas the second protein (TNP1-like) is only weakly similar in different CACTA-like elements. The proposed model for the transposition of CACTA-like element is that TNP1-like protein binds to repetitive motifs residing in the subterminal region of CACTA-like elements and recruits TPN2-like protein, the TPase, via protein-protein interaction to catalyze subsequent biochemical reactions (Gierl et al. 1988; Raina et al. 1998). To date, members of the CACTA superfamily have only been found in plants and a comprehensive phylogenetic analysis has yet to be performed.

The *Tc1*/*mariner* superfamily. The *Tc1* family from *C. elegans* and the *mariner* family from Drosophila define a large superfamily of DNA elements that are widespread in plants, animals and fungi (Robertson et al. 1998; Plasterk et al. 1999; Plasterk and van Luenen 2002). Three clades have been resolved for the *Tc1*/*mariner* superfamily, *Tc1*-like, *mariner*-like and *pogo*-like (Plasterk et al. 1999). Members of this superfamily are relatively short (1.3-2.4 kb in length), encode a single TPase gene, and always insert into and duplicate the dinucleotide 5'-TA-3' (Jacobson et al. 1986; Medhora et al. 1991; Capy et al. 1992; Ivics et al. 1997; Plasterk et al. 1999; Plasterk and van Luenen 2002). In

animals, *Tc1*/*mariner*-like elements appear to have undergone frequent horizontal transfer from one host genome to another, as elements isolated from species that diverged over 100 million years ago (Mya) can be virtually identical (Robertson 1993). A recent survey and phylogenetic analysis of plant *mariner*-like elements indicated that they are widespread in flowering plants and that their evolution is characterized by vertical transmission of multiple ancient lineages (Feschotte and Wessler 2002). In addition, a detailed characterization of *mariner*-like elements in rice has led to the discovery that this superfamily of DNA elements is associated with *Stowaway*-like MITEs (Feschotte et al. 2003).

TEs Are Major Component of Plant Genomes.

TEs are widespread in plants and are major components of the genomes they reside in (Kumar 1996; Pearce et al. 1996; Wright et al. 1996; SanMiguel et al. 1996; SanMiguel et al. 1998; Feschotte et al. 2002a). In fact, the differential proliferation can largely account for the over 1,000 fold variations of plant genome size, ranging from ~50 Mb (*Cardamine amara*, family Brassicaceae) to over 87,000 Mb (*fritillaria assyriaca*, family Liliaceae) (Bennett and Leitch 1995). Approximately 10% of the small genome of *Arabidopsis* (~120 Mb) consists of TEs (Leutwiler et al. 1984), whereas over 80% of the nuclear DNA in plants with large genomes (e.g., maize 2,500 Mb) is comprised of TEs. Class 1 elements are more abundant than class 2 elements, likely because of their replicative transposition mechanisms. LTR-retroelements are known to make up a large fraction of some plant genomes such as maize (57%) (Meyers et al. 2001). In

particular, a few families of LTR-retroelements can amplify to extremely high copy numbers. For example, four families of LTR-retroelements (*Ji*, *Opie*, *Huck* and *Zeon-1*) account for 32% of the maize genome (~2,500 Mb) (Meyers et al. 2001) and one family of LTR-retroelements (*IRRE*) account for ~10% of the *Iris brevicaulis* genome (~10,000 Mb) (Kentner et al. 2003). Although LINEs are widespread in plants, their abundance appears to vary significantly. For example, only ~0.1% of the maize genome (~2,500 Mb) consists of LINEs (Meyers et al. 2001), whereas a single family of LINEs (called *del2*) is present at 250,000 copies in lily (90,000 Mb) and make up ~4% of its genome (Leeton and Smyth 1993). It is generally thought that DNA elements do not contribute significantly to plant genome size variations as they typically do not accumulate to high copy number. However, exceptions have been documented recently. Two families of CACTA-like elements, the *Tpo1* family from *Lolium perenne* (5,000 Mb) and the *Caspar* family from wheat (*Triticum monococcum*, ~5,000 Mb) are present at ~5,000 and ~3,000 copies, respectively (Langdon et al. 2003; Wicker et al. 2003). In addition, MITEs, with their high copy number, can account for a significant fraction of the genome they reside in (e.g., ~6% of the ~450 Mb rice genome) (N. Jiang and S.R. Wessler, unpublished data).

Dissertation Outline.

The research described in this dissertation was directed to address two questions. The first question concerns the origin and transposition mechanism of *Tourist*-like MITEs. In Chapter 2, a high copy number *Tourist*-like MITE family

(called *miniature PIF*, *mPIF*) from maize was identified and characterized. Several significant similarities shared by *mPIF* elements and an active maize transposon family called *P Instability Factor* (*PIF*) (Walker et al. 1997) led to the conclusion that *mPIF* elements are nonautonomous members of the *PIF* family. This represented the first case where a MITE family was related to an active TPase-encoding DNA family. The *PIF* family was further characterized and additional evidence supporting the relationship between *mPIF* and *PIF* was presented. To classify the *PIF* family with regards to known DNA families, a TPase-encoding *PIF* element (*PIFa*) was isolated through genetic and molecular approaches. Sequence and phylogenetic analyses of *PIFa* coding sequence uncovered the existence of a new superfamily of related DNA transposons (called *PIF*/IS*5*) that are widespread in eukaryotes and distantly related to some bacterial insertion sequences. Several *PIF*-like elements were found to be associated with *Tourist*-like MITE families in their respective genomes. In addition, a second active member of the *PIF*/IS*5* superfamily (*Pong*) was recently identified in rice where it is associated with an active family of MITEs (*mPing*) (Jiang et al. 2003). Taken together, these results indicate that members of the *PIF*/IS*5* superfamily are responsible for the origin and amplification of *Tourist*-like MITEs.

The distribution and evolution of *PIF*/*Pong*-like elements and their relationship with *Tourist*-like MITEs is the focus of Chapter 3. Computational and molecular approaches were combined in a comprehensive survey of *PIF*/*Pong*-like elements, which led to the identification of hundreds of related TPase

sequences from plants, animals and fungi. Subsequent phylogenetic analyses resolved the structure of the *PIF*/*Pong* superfamily and defined three clades. Examination of plant *PIF*/*Pong*-like elements indicated that their evolution is characterized by the vertical transmission of multiple ancient and distinct lineages. The structure of *PIF*/*Pong*-like elements and their relationship with *Tourist*-like MITEs were determined in rice, where it was found that most *Tourist*-like MITEs are associated with either *PIF*- or *Pong*-like elements. Furthermore, for the first time, a high copy number MITE family (*Castaway*, >3,000 copies) was found to be a direct deletion derivative of a TPase-encoding element (*OsPIF6*).

The second question addressed in this dissertation concerns the basis for the recent genome size expansion of *Brassica oleracea* since its divergence from *Arabidopsis* ~15-20 Mya and the underlying reason for the low TE content of the *Arabidopsis* genome. Chapter 4 describes the comparison of TE diversity, abundance and phylogeny in these two species. The results showed that amplification of both class1 and 2 elements in *B. oleracea* contributed significantly to its recent genome size increase, and provided evidence that the low TE content of the *Arabidopsis* genome is largely due to reduced TE proliferation and not due to the massive loss of TE sequences.

References.

Becker, H.A. and R. Kunze. 1997. Maize Activator transposase has a bipartite DNA binding domain that recognizes subterminal sequences and the terminal inverted repeats. *Mol. Gen. Genet.* **254**: 219-30.

Bennett, M.D. and I.J. Leitch. 1995. Nuclear DNA amount in angiosperms. *Annals of Botany* **76**: 113-176.

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

Bingham, P.M. and Z. Zachar. 1989. Retrotransposons and the FB transposon from *Drosophila melanogaster*. In *Mobile DNA* (ed. D.E. Berg and M.M. Howe), pp. 485-502. American Society for Microgiology, Washington, D.C.

Boeke, J.D. and V.G. Corces. 1989. Transcription and reverse transcription of retrotransposons. *Annu Rev Microbiol* **43**: 403-34.

Bonas, U., H. Sommer, and H. Saedler. 1984. The 17-kb *Tam1* element of *Antirrhinum majus* induces a 3-bp duplication upon integration into the chalcone synthase gene. *EMBO J.* **5**: 1015-1019.

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 wild-type rice genes. *Proc. Natl. Acad. Sci. USA* **93**: 8524-8529.

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

-. 1994. *Stowaway*: a new family of inverted-repeat elements associated with genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6**: 907-916.

Calvi, B.R., T.J. Hong, S.D. Findley, and W.M. Gelbart. 1991. Evidence for a common evolutionary origin of inverted repeat transposons in *Drosophila* and plants - *hobo*, *Activato*r, and *Tam3*. *Cell* **66**: 465-471.

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

Capy, P., A. Koga, J.R. David, and D.L. Hartl. 1992. Sequence analysis of active mariner elements in natural populations of Drosophila simulans. *Genetics* **130**: 499-506.

Casacuberta, E., J.M. Casacuberta, P. Puigdomenech, and A. Monfort. 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.

Deininger, P.L. 1989. SINEs: Short Interspersed Repeated DNA Elements in higher eukaryotes. In *Mobile DNA* (ed. D.E. Berg and M.M. Howe), pp. 619-636. American Society for Microbiology, Washington, DC.

Doolittle, R.F., D.-F. Feng, M.s. Johnson, and M.A. McClure. 1989. Origins and evolutionary relationships of retroviruses. *Q.Rev.Biol.* **64**: 1-30.

Eisen, J.A., M.I. Benito, and V. Walbot. 1994. Sequence similarity of putative transposases links the maize Mutator autonomous element and a group of bacterial insertion sequences. *Nucleic Acids. Res.* **22**: 2634-6.

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

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.

Feschotte, C., L. Swamy, and S.R. Wessler. 2003. Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with *Stowaway* MITEs. *Genetics **163***: 747-758.

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

Feschotte, C., X. Zhang, and S. Wessler. 2002b. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In *Mobile DNA II* (ed. N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz), pp. 1147-1158. American Society for Microbiology Press, Washington, DC.

Gierl, A., S. Lutticke, and H. Saedler. 1988. *TnpA* product encoded by the transposable element *En-1* of *Zea mays* is a DNA binding protein. *EMBO J.* **7**: 4045-4053.

Hehl, R., W.K.F. Nacken, A. Krause, H. Saedler, and H. Sommer. 1991. Structural analysis of *Tam3*, a transposable element from *Antirrhinum majus,* reveals homologies to the *Ac* element from maize. *Plant Mol. Biol.* **16**: 369-371.

Hershberger, R.J., M.-I. Benito, K.J. Hardemen, C. Warren, V.L. Chandler, and V. Walbot. 1995. Characterization of the major transcripts encoded by the

regulatory *MuDR* transposable element of maize. *Genetics* **140**: 1087-1098.

Hutchison, C.A., S.C. Hardies, D.D. Loeb, W.R. Shehee, and M.H. Edgell. 1989. LINES and related retrotransposons: long interspersed repeated sequences in the eukaryotic genome. In *Mobile DNA* (ed. D.E. Berg and M.M. Howe), pp. 593-617. American Society of Microbiology, Washington, D.C.

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

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

Jacobson, J.W., M.M. Medhora, and D.L. Hartl. 1986. Molecular structure of a somatically unstable transposable element in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **83**: 8684-8688.

Jiang, N., Z. Bao, X. Zhang, H. Hirochika, S.R. Eddy, S.R. McCouch, and S.R. Wessler. 2003. An active DNA transposon family in rice. *Nature* **421**: 163-7.

Kentner, E.K., M.L. Arnold, and S.R. Wessler. 2003. Characterization of high-copy-number retrotransposons from the large genomes of the louisiana iris species and their use as molecular markers. *Genetics* **164**: 685-97.

Kumar, A. 1996. The adventures of the Ty1-*copia* group of retrotransposons in plants. *Trends Genet.* **12**: 41-43.

Kunze, R. and P. Starlinger. 1989. The putative transposase of transposable element Ac from *Zea mays* L. interacts with subterminal sequences of Ac. *EMBO J.* **8**: 3177-3185.

Kunze, R. and C.F. Weil. 2002. The hAT and CACTA superfamilies of plant transposons. In *Mobile DNA II* (ed. N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz), pp. 565-610. American Society for Microbiology Press, Washington DC.

Langdon, T., G. Jenkins, R. Hasterok, R.N. Jones, and I.P. King. 2003. A High-Copy-Number CACTA Family Transposon in Temperate Grasses and Cereals. *Genetics* **163**: 1097-108.

Leeton, P.R.J. and D.R. Smyth. 1993. An abundant LINE-like element amplified in the genome of *Lilium speciosum*. *Mol.Gen.Genet.* **237**: 97-104.

Leutwiler, L., B. Hough-Evans, and E. Meyerowitz. 1984. The DNA of *Arabidopsis thaliana*. *Mol. Gen. Genet.* **219**: 225-234.

Lisch, D. 2002. Mutator transposons. *Trends Plant Sci.* **7**: 498-504.

Lisch, D., L. Girard, M. Donlin, and M. Freeling. 1999. Functional analysis of deletion derivatives of the maize transposon MuDR delineates roles for the MURA and MURB proteins. *Genetics* **151**: 331-41.

Lisch, D.R., M. Freeling, R.J. Langham, and M.Y. Choy. 2001. Mutator transposase is widespread in the grasses. *Plant Physiol.* **125**: 1293-1303.

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

Masson, P., G. Rutherford, J.A. Banks, and N. Fedoroff. 1989. Essential large transcripts of the maize *Spm* transposable element are generated by alternative splicing. *Cell* **58**: 755-765.

McClintock, B. 1948. Mutable loci in maize. *Carnegie Inst.Washington Yearb.* **47**: 155-169.

-. 1949. Mutable loci in maize. *Carnegie Inst.Washington Yearb.* **48**: 142-154.

Medhora, M., K. Maruyama, and D.L. Hartl. 1991. Molecular and functional analysis of the *mariner* mutator element *Mos1* in *Drosophila*. *Genetics* **128**: 311-318.

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**: 1660-76.

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

Muller, F. and S.G. Clarkson. 1980. Nucleotide sequence of genes coding for tRNAPhe and tRNATyr from a repeating unit of X. laevis DNA. *Cell* **19**: 345-53.

Nacken, W.K.F., R. Piotrowiak, H. Saedler, and H. Sommer. 1991. The transposable element Tam1 from *Antirrhinum majus* shows structural

homology to the maize transposon *en*/*Spm* and gas no sequence specificity of insertion. *Mol.Gen.Genet.* **228**: 201-208.

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

Pearce, S.R., G. Harrison, D. Li, J.S. Heslop-Harrison, A. Kumar, and A.J. Flavell. 1996. The Ty1-*copia* group retrotransposons in *Vicia* species: copy number, sequence heterogeneity and chromosomal localisation. *Mol. Gen. Genet.* **250**: 305-315.

Pereira, A., H. Cuypers, A. Gierl, Z. Schwarz-Sommer, and H. Saedler. 1986. Molecular analysis of the *En*/*Spm* transposable element system of *Zea mays*. *EMBO J.* **5**: 835-841.

Plasterk, R.H.A., Z. Izsvák, and Z. Ivics. 1999. Resident aliens: the Tc1/*mariner* superfamily of transposable elements. *Trends Genet.* **15**: 326-332.

Plasterk, R.H.A. and H.G. van Luenen. 2002. The Tc1/*mariner* family of transposable elements. In *Mobile DNA II* (ed. N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz), pp. 519-532. American Society for Microbiology Press, Washington D.C.

Raina, R., M. Schlappi, B. Karunanandaa, A. Elhofy, and N. Fedoroff. 1998. Concerted formation of macromolecular Suppressor-mutator transposition complexes. *Proc. Natl. Acad. Sci. USA* **95**: 8526-8531.

Robertson, H.M. 1993. The *mariner* transposable element is widespread in insects . *Nature* **362**: 241-245.

Robertson, H.M., F.N. Soto-Adames, K.O. Walden, R.M. Avancini, and D.J. Lampe. 1998. The mariner transposons of animals: horizontally jumping genes. In *Horizontal gene transfer* (ed. M. Syvanen and C.I. Kido), pp. 268-284. Chapman & Hall, London.

Rubin, E., G. Lithwick, and A.A. Levy. 2001. Structure and evolution of the hAT transposon superfamily. *Genetics* **158**: 949-57.

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.

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

Singer, T., C. Yordan, and R.A. Martienssen. 2001. Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). *Genes Dev.* **15**: 591-602.

Streck, R.D., J.E. MacGaffey, and S.K. Beckendorf. 1986. The structure of hobo transposable elements and their insertion sites. *EMBO J.* **5**: 3615-3623.

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.

-. 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.

Walker, E.L., W.B. Eggleston, D. Demopulos, J. Kermicle, and S.L. Dellaporta. 1997. Insertions of a novel class of transposable elements with a strong target site preference at the *r* locus of maize. *Genetics* **146**: 681-693.

Warren, W.D., P.W. Atkinson, and D.A. Obrochta. 1994. The hermes transposable element from the house fly, *Musca domestica*, is a short inverted repeat-type element of the *hobo*, *Ac*, and *Tam3* (hAT) element family. *Genet Res.* **64**: 87-97.

Weil, C.F. and R. Kunze. 2000. Transposition of maize Ac/Ds transposable elements in the yeast saccharomyces cerevisiae [In Process Citation]. *Nat. Genet.* **26**: 187-90.

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

Wessler, S.R., A. Nagel, and A.M. Casa. 2001. Miniature inverted repeat transposable elements help create genomic diversity in maize and rice. In *Proceedings of the Fourth International Rice Genetics Symposium* (ed. G.S. Khush, D.S. Brar, and B. Hardy), pp. 107-116. International Rice Research Institute, Los Banos, Philippines.

Wicker, T., R. Guyot, N. Yahiaoui, and B. Keller. 2003. CACTA Transposons in Triticeae. A Diverse Family of High-Copy Repetitive Elements. *Plant Physiol.* **132**: 52-63.

Wright, D., N. Ke, J. Smalle, B.M. Hauge, H.M. Goodman, and D.F. Voytas. 1996. Multiple non-LTR retrotransposons in the genome of Arabidopsis thaliana. *Genetics* **142**: 569-578.

Xiong, Y. and T.H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**: 3353-62.

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

CHAPTER 2

P INSTABILITY FACTOR: AN ACTIVE MAIZE TRANSPOSON SYSTEM ASSOCIATED WITH THE AMPLIFICATION OF *TOURIST*-LIKE MITES AND A NEW SUPERFAMILY OF TRANSPOSASES[1]

[1] 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. 2.1a) (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.

Materials and Methods

Genetic stocks, DNA extraction and library construction.

All strains were derived from the maize inbred W22.

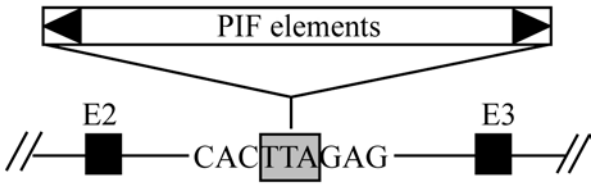
r-sc:124Y2902: A derivative of *R-sc:124* (*R* allele conferring pigmentation of aleurone, embryo and coleoptile) with a 2.3 kb *PIF* insertion in the second intron of the *Sc* component (16) causing loss of kernel pigmentation. Excision restores kernel pigmentation.

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

Fig. 2.1: 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.

a



b

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

c

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

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. 2.2). 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. 2.2) 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. 2.2) 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.2.2). 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 x 10^5 screened (representing ~1.6 x 10^5 kb or ~6 % of the genome) provided an estimate of ~6 x 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. 2.6) (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.

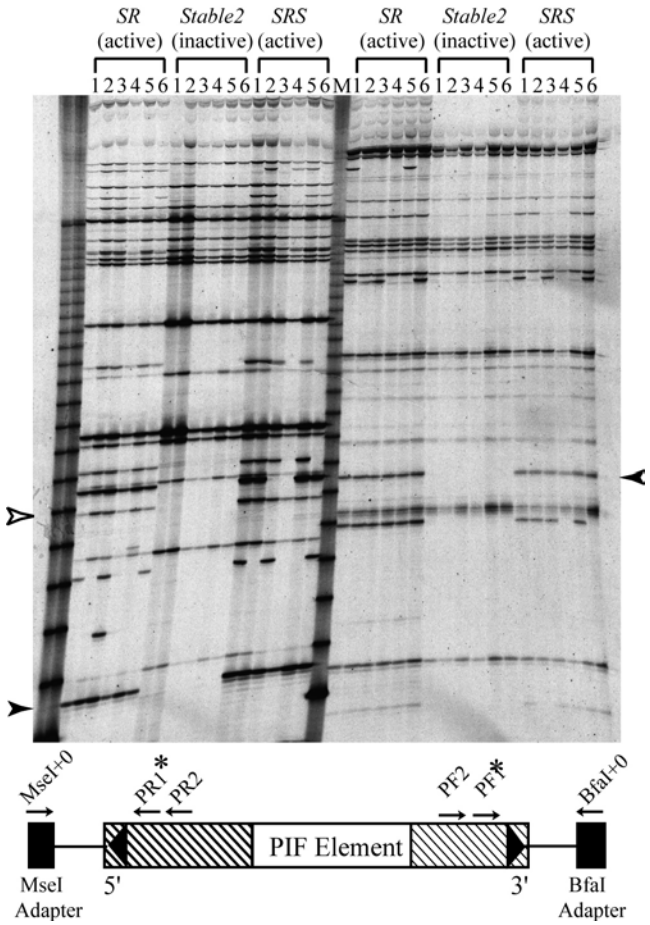
Identical extended target site preference for *mPIF* and *PIF* elements.

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. 2.1). 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. 2.2). 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. 2.1c).

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

Fig. 2.2: 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
PF1
BfaI+0
PIF Element
MseI Adapter
5'
3'
BfaI Adapter

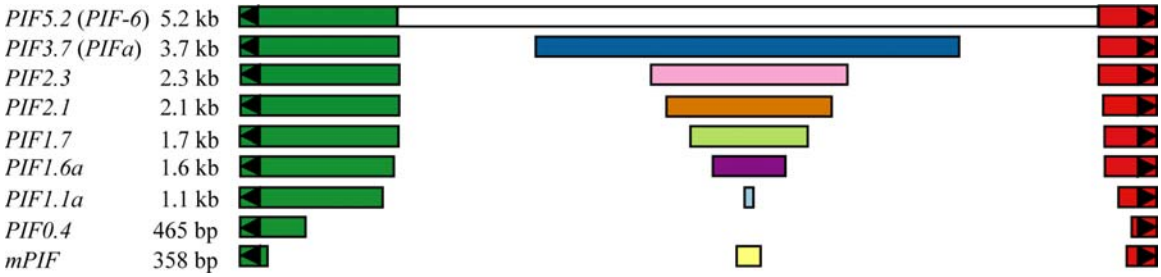
preference ever documented among eukaryotic class 2 TEs. Additional support for the existence of a specific 9-bp insertion site comes from the fact that the sequences flanking *mPIF* elements judged to have inserted most recently (based on highest sequence identity to the *mPIF* consensus and insertion 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.

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. 2.3)

Fig. 2.3: 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

(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. 2.3). Unfortunately, none of these elements were considered autonomous since computer analysis failed to detect significant coding capacity or any similarity to known transposases.

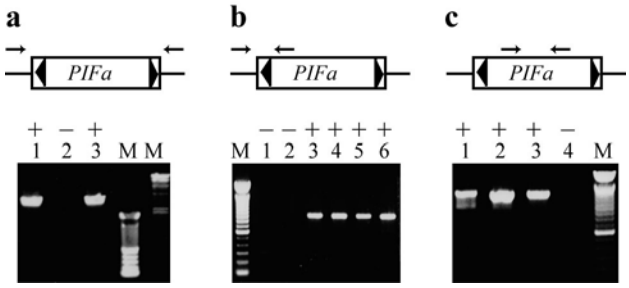
Isolation of the *PIFa* element.

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. 2.2). 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. 2.4a). 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. 2.3, 2.4a).

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. 2.4b, 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. 2.4b, 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. 2.4c). This result also demonstrated that the loss of *PIFa* correlated with the loss of *PIF* activity.

Fig. 2.4: *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
b
c
PIFa
PIFa
PIFa
+ − +
1 2 3 M M
− − + + + +
M 1 2 3 4 5 6
+ + + −
1 2 3 4 M

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

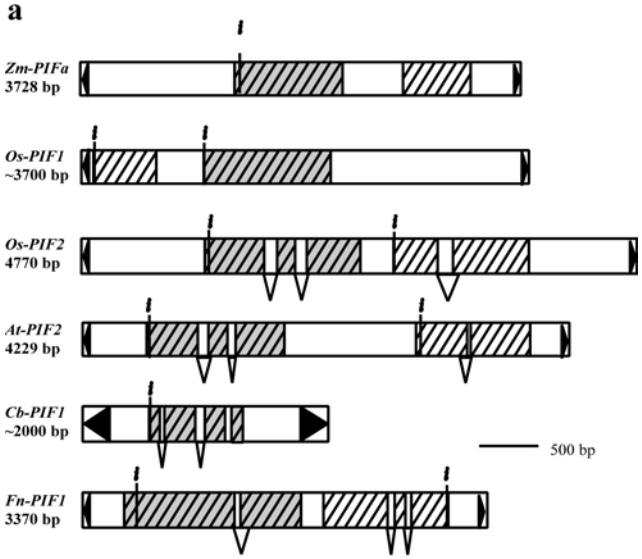
The sequence of *PIFa* revealed a 3,728 bp element that, like all other *PIF* elements, contains the conserved terminal regions (Fig. 2.2). The central 2.5 kb, not found in other *PIF* elements, contains two ORFs longer than 100 a.a. (Fig. 2.5a). Only the first ORF (313 a.a.) produced significant hits (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. 2.5 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. 2.7) (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 proteins identified by database searches and generating phylogenetic trees. A CLUSTALW multiple alignment (Supplemental Fig. 2.7) 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. 2.5c). Bacterial transposases and eukaryotic

Fig. 2.5: 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* AF412282; *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. 2.7). 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.

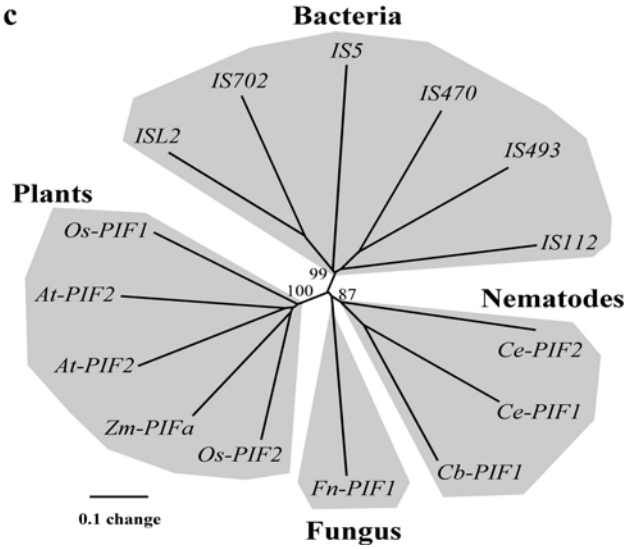
a



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



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. 2.5c).

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. 2.5b). 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. 2.8). 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.

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. 2.5b). 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.

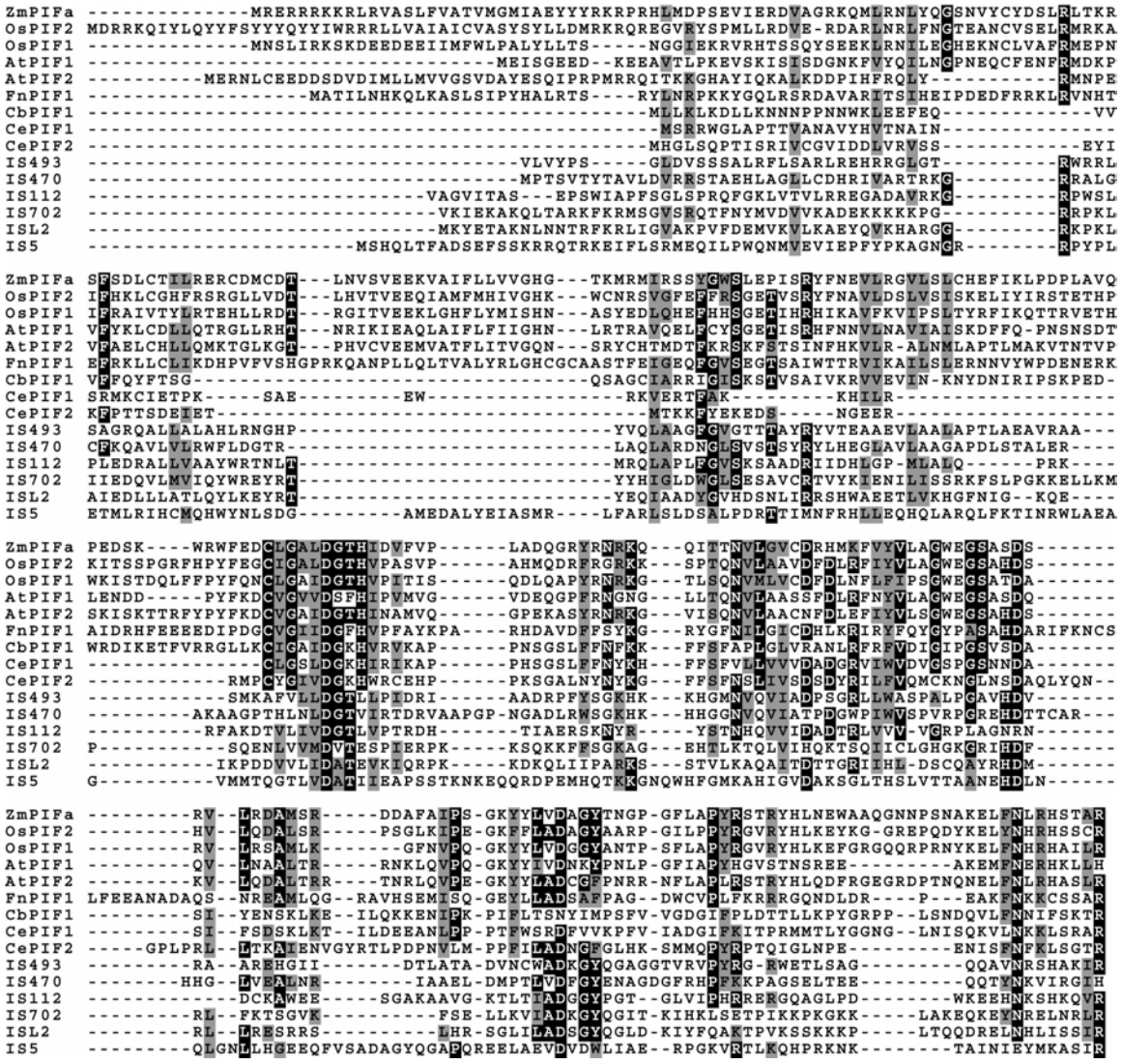
Supplemental data

Supplemental Fig. 2.6: Similarities in terminal inverted-repeats (TIRs) and target site duplications (tsd) among some *Tourist*-like MITEs.

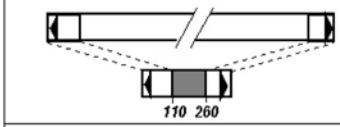
species	family	tsd	TIRs bp	sequence	reference
Z. mays	*mPIF*	TTA	14	GGGCCCGTTTGTTT	(*)
Z. mays	*TouristA*	TTA	14	GGCCTTGTTCGGTT	(26)
Z. mays	*TouristB*	TTA	14	GGCCTTGTTTAGTT	(26)
Z. mays	*TouristC*	TTA	14	GGGCCTGTTTAGAT	(26)
Z. mays	*TouristD*	TTA	14	GGGGGTGTTTGGTT	(26)
Z. mays	*Hbr*	NNN	14	GGGCCTGTTTGGTT	(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. 2.7: 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. 2.5 for accession numbers.


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

Supplemental Fig. 2.8: *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. 2.5, 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

CHAPTER 3

PIF- AND *PONG*-LIKE ELEMENTS: DISTRIBUTION, EVOLUTION AND RELATIOHSHIP WITH *TOURIST*-LIKE MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS[1]

[1] Zhang, X., Jiang, N., Feschotte, C. and Wessler, S.R. Accepted for publication by *Genetics*. 7/14/2002

ABSTRACT

Miniature inverted-repeat transposable elements (MITEs) are short, nonautonomous DNA elements that are widespread and abundant in plant genomes. Most of the hundreds of thousands of MITEs identified to date have been divided into two major groups based on shared structural and sequence characteristics: *Tourist*-like and *Stowaway*-like. Since MITEs have no coding capacity, they must rely on transposases encoded by other elements. Two active transposons, the maize *P Instability Factor* (*PIF*) and the rice *Pong* element, have recently been implicated as sources of transposase for *Tourist*-like MITEs. Here we report that *PIF*- and *Pong*-like elements are widespread, diverse and abundant in eukaryotes with hundreds of element-associated transposases found in a variety of plant, animal and fungal genomes. The availability of virtually the entire rice genome sequence facilitated the identification of all the *PIF*/*Pong*-like elements in this organism and permitted a comprehensive analysis of their relationship with *Tourist*-like MITEs. Taken together, our results indicate that *PIF* and *Pong* are founding members of a large eukaryotic transposon superfamily, and that members of this superfamily are responsible for the origin and amplification of *Tourist*-like MITEs.

INTRODUCTION

Transposable elements (TEs), which are a major component of all characterized eukaryotic genomes, have been divided into two classes according to their transposition intermediate. Class 1 (RNA) elements transpose via an RNA intermediate and most either have long terminal repeats (LTR-retrotransposons) or terminate at one end with a polyA tract [long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)]. Class 2 (DNA) elements transpose via a DNA intermediate and usually have short terminal inverted repeats (TIRs). DNA elements can be further classified into families based on the transposase (TPase) that catalyzes their movement. A TE family is comprised of one or more TPase-encoding autonomous elements and up to several thousand nonautonomous elements that do not encode functional TPases but retain the *cis*-sequences necessary to be mobilized by the cognate TPase (for reviews CAPY et al. 1998; FESCHOTTE et al. 2002a).

Miniature inverted-repeat transposable elements (MITEs) were first discovered in the grasses and later found in other flowering plant as well as in animal genomes (for review FESCHOTTE et al. 2002b). Structurally, MITEs are reminiscent of nonautonomous DNA elements, but their high copy number, intrafamily homogeneity in size and sequence, and strong target site preference (most insert into and duplicate the dinucleotide TA or trinucleotide TTA/TAA) distinguish them from most previously described nonautonomous elements (WESSLER et al. 1995). Since MITEs lack coding sequences, their classification

has been based on the sequence similarity of TIR and target site duplication (TSD). Using these criteria, most of the tens of thousands of plant MITEs have been divided into two groups: *Tourist*-like (3-bp TSDs, usually TTA/TAA) and *Stowaway*-like (2-bp TSD, usually TA) (FESCHOTTE et al. 2002b).

Two distantly related families of active DNA transposons have recently been associated with *Tourist*-like MITEs. The maize *P Instability Factor* (*PIF*) and a *Tourist*-like MITE family called *miniature PIF* (*mPIF*) share identical TIRs, similar subterminal sequences, and a strong preference for insertion into the 9-bp palindrome CWCTTAGWG with duplication of the central TTA (WALKER et al. 1997; ZHANG et al. 2001). An even closer relationship was found in rice, where the 430 bp *Tourist*-like MITE called *mPing* was shown to be a deletion derivative of a 5.2 kb transposase-encoding element called *Ping* (JIANG et al. 2003; KIKUCHI et al. 2003; NAKAZAKI et al. 2003). Several lines of evidence, however, led to the conclusion that a related element in the rice genome, called *Pong*, was the most likely source of TPase mobilizing *mPing* elements (JIANG et al. 2003). Indeed, *Pong* elements and *mPing* MITEs were found to be actively transposing in the same cell culture line. *Pong* is an unusual eukaryotic transposon because it contains two ORFs. ORF1 may be involved in DNA binding as it includes a domain with weak similarity to the DNA-binding domain of *myb* transcription factors (JIANG et al. 2003). ORF2 most likely encodes the TPase, as it contains an apparent DD35E motif, a signature consisting of three acidic residues found in the catalytic domains of some eukaryotic and prokaryotic TPases (JIANG et al. 2003; MAHILLON and CHANDLER 1998; REZSOHAZY et al. 1993).

PIF and *Pong* elements share several features, including amino acid sequence conservation in their catalytic domain, nucleotide sequence similarity in their TIRs and identical TSDs (TTA) (JIANG et al. 2003). The putative TPases of both *PIF* and *Pong* have a large number of homologs in GenBank that have been annotated as unknown/hypothetical proteins. Examination of several *PIF* TPase homologs from plants, nematodes and a fungus showed that they resided in *PIF*-like transposable elements (ZHANG et al. 2001). Furthermore, *PIF*-like and *Pong* TPases were shown to be distantly related to the TPases of bacterial insertion sequences of the IS*5* group (JIANG et al. 2003; KAPITONOV and JURKA 1999; LE et al. 2001; ZHANG et al. 2001). Thus, *PIF* and *Pon*g are the founding members of what appears to be a new and widespread superfamily of DNA transposons called *PIF*/IS*5*. Significantly, some of these *PIF*-like elements were found to be associated with *Tourist*-like MITEs in their respective genomes (APARICIO et al. 2002; FESCHOTTE et al. 2002b; LE et al. 2001; ZHANG et al. 2001).

In this study we look at the distribution and evolution of the *PIF/*IS*5* superfamily of transposases and characterize their relationship with *Tourist*-like MITEs. To this end we conducted a systematic survey (database searches and PCR assays) of putative *PIF*- and *Pong*-like TPases in plants and animals. Phylogenetic analyses of over 600 TPase fragments from 56 species define three major groups, each represented by multiple ancient and distinct lineages. The availability of virtually the entire sequence of rice (GOFF et al. 2002 ; YU et al. 2002) permitted the identification and characterization of all *PIF*- and *Pong*-like elements in a single genome. Furthermore, the association between rice *PIF*-

and *Pong*-like elements and *Tourist*-like MITEs was explored by performing a genome-wide comparison of these elements. This represents the first comprehensive analysis of the origin of *Tourist*-like MITEs in any organism.

RESULTS

Distribution and Abundance of *PIF*/*Pong*-like TPases

Identification of *PIF*/*Pong*-like TPases through database searches. A systematic survey was carried out using the TPases of *PIF* and *Pong* as queries in tBlastn searches against several public databases. Significant similarity was detected in over 1,000 entries from a wide range of eukaryotic species, including 21 plants, 19 animals and two fungi (listed in Table 1). Six hundred and seventy three hits (574 were unique) with significant homology to the catalytic domains of *PIF* or *Pong* were selected for further analysis.

One striking result from database searches was the abundance of *PIF*/*Pong*-like TPases in some organisms, especially plants. This is most apparent in genomes with large amounts of sequence information: for the catalytic domains alone, there were ~80 hits (E value < 10^{-23}) in *Arabidopsis thaliana*, ~350 hits (E value < 10^{-10}) in rice (*Oryza sativa* c.v. Nipponbare), and ~170 hits (E value < 10^{-30}) in *Brassica oleracea* [~ 30% of its 600 Mb genome available for blast at TIGR (http://www.tigr.org/tdb/e2k1/bog1/)]. In animals, the number of *PIF*/*Pong*-like TPases varies from over 100 in the African malaria mosquito (*Anopheles gambiae*) and ~300-400 in zebrafish (*Danio rerio*) to a few

Table 3.1. Species with *PIF*- or *Pong*-like TPases.

Plants	Angiosperms	Monocot	Asparagus, Barley (Hv), Bamboo (Aa), Coix (Ca), *D. ensifolia*, Ehrharta (Ec), Foxtail millet, *Gongora ilense* (Gi), Johnson grass (Sh), *J. ascendens* (Ja), Maize (Zm)[a], Oat (As), *P. latifolius* (Pl), Red Millet (Pm), Rice (Os)[a], Sorghum (Sb)[a], Sugarcane (Shc), *T. pilosum* (Tp), Teosinte (Zp, Zh), Wheat (Ta),
		Dicot	*Arabidopsis* (At)[a], *Brassica oleracea* (Bo), Ice plant, Lettuce, *L. japonicus* (Lj), Medicago (Mt), Peppermint, Potato, Soybean, Sweet leaf, Suger beet, Tomato (Le),
	Gymnosperms		Pine
	Algae		*Physcomitrella patens*, *Porphyra yezoensis*
Animals	Invertebrate	Nematode	*C. briggase*[a], *C. elegans*[a]
		Insect	*Drosophila*, African malaria mosquito, Silkworm
		Echinoderm	Sea urchin
		Ascidian	Sea squirt
	Vertebrate	Fish	Medaka fish, Takifugu[a], Trout, Zebrafish
		Amphibian	*Xenopus*
		Bird	Chicken
		Mammal	Chimpanzee, Cow, Human, Mouse, Pig, Rat
Fungi			*F. neoformans* (Fn)*, N. crassa*

For sequences used in generating the phylogenetic tree in Fig. 3.3, initials of species names are shown.

[a]*PIF*-like TPases were identified in these species by previous studies (APARICIO et al. 2002; JIANG et al. 2003; KAPITONOV and JURKA 1999; LE et al. 2001; ZHANG et al. 2001)

(less than three) in *Drosophila melanogaster*, *Caenorhabditis elegans* and human.

Nucleotide sequences of the 574 unique *PIF*- and *Pong*-like TPases were conceptually translated into amino acid sequences after removal of introns (see below) and judicious correction of frameshifts caused by small (1-2 bp) insertions or deletions. The resulting amino acid sequences were compared in order to detect conserved regions that might signify functional domains. Several blocks

of highly conserved residues were identified for *PIF*-like TPases (Fig. 3.1a). Block H corresponds to a predicted helix-turn-helix (HTH) domain that may be involved in DNA binding. Blocks N2, N3 and C1 most likely comprise the catalytic domain as they contain an apparent DDE motif with the three acidic residues centered in blocks N2, N3 and C1, respectively. A DDE motif is also present in *Pong*-like TPases (Fig. 3.1b), but unlike *PIF*, no HTH domain was predicted.

***Pong*-like TPases are usually adjacent to ORF1 homologs.** tBlastn searches using the ORF1 of *Pong* as query also yielded a large number of hits from plants. When located on long contigs, these ORF1 homologs were usually found within 1-2 kb of *Pong*-like TPases, indicating that each "pair" of ORF1 and TPase was encoded by the same element. In fact, when the termini of *Pong*-like elements were defined in *O. sativa* (see below) and *A. thaliana* (X. Zhang and S. R. Wessler, unpublished data), nearly all elements were found to encode both ORFs. ORF1s are significantly more divergent than the TPases of *Pong*-like elements. Three blocks of conserved residues were found when ORF1s were compared (A-C, Fig. 3.1b), with the most conserved block (block A) centered in a ~100-a.a. region that displays weak homology to the DNA binding domains of *myb* transcription factors from some plants and animals (Jiang et al. 2003).

Additional *PIF*-like TPases from grasses. The majority of *PIF*-like sequences (~80%) were from only a few species (rice, *Arabidopsis*, *B. oleracea* and *A. gambiae*) since this survey was limited by the availability of DNA sequences in

Figure 3.1. Structure and conserved coding regions of *PIF*- and *Pong*-like elements. Triangles represent element TIRs. Coding regions are represented by dark gray boxes. Regions that are unrelated between different *PIF*- or *Pong*-like families are represented by light gray boxes. DDE motifs in TPases are indicated by asterisks, and open boxes localize regions that were used to deduce the phylogeny of *PIF*- or *Pong*-like elements (Fig. 3.2, 3.3, 3.4, 3.5 and 3.7). (a) *PIF*-like elements. A multiple alignment of the conserved regions of select *PIF*-like TPases is shown. Horizontal arrows indicate the position of dPCR primers, and the open and filled arrow heads denote the position of intron 1 and intron 2, respectively. Horizontal lines indicate the predicted HTH domain (H) and three blocks of conserved residues (N2, N3 and C1) that likely comprise the catalytic domain (see text). (b) *Pong*-like elements. Multiple alignments of the conserved regions from ORF1 and TPase of *Pong*-like elements are shown. Horizontal lines in the ORF1 alignment indicate three blocks of conserved residues (A, B and C). *Os-PIF2*, *Os-PIF1*, *At-PIF2*, *At-Harb* and *Pong* were previously reported (JIANG et al. 2003; KAPITONOV and JURKA 1999; ZHANG et al. 2001). The *PIF* TPase shown in (a) is a full-length TPase isolated from maize (X. Zhang and S. R. Wessler, unpublished data). Other sequences were named according to the species initials (see Table 1), followed by their GenBank accession number.

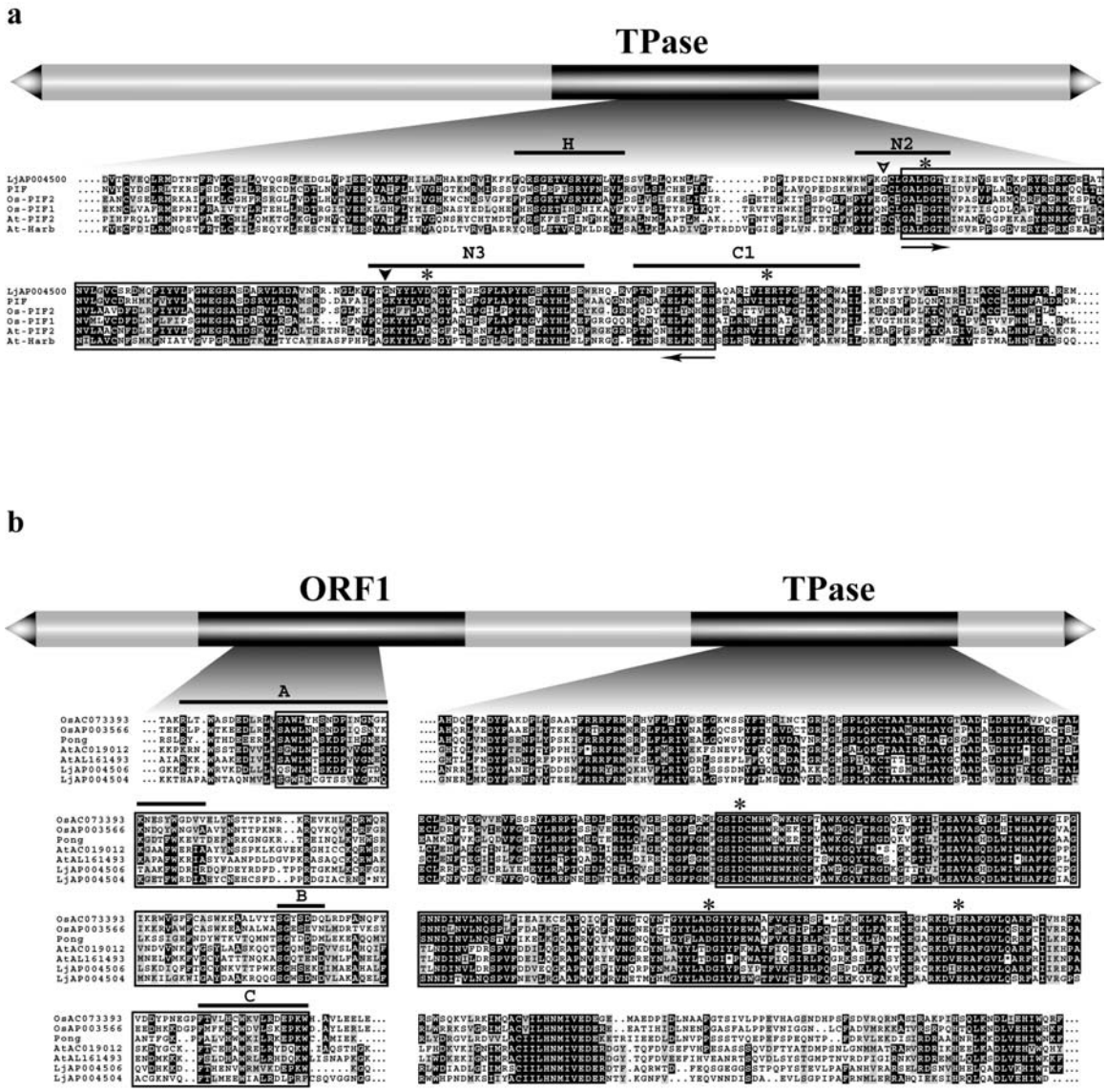


Figure 1

databases. To better resolve the phylogeny of *PIF* elements, additional TPase sequences were isolated from species with established evolutionary relationships but limited sequence information. To this end, a degenerate PCR (dPCR) procedure was employed to amplify *PIF*-like TPase fragments from selected grass species. Grasses were chosen for this analysis because their phylogeny is well characterized (KELLOGG 2001) and they harbor the only known active *PIF* and *Pong* elements (JIANG et al. 2003; WALKER et al. 1997; ZHANG et al. 2001).

dPCR primers were derived from the conserved blocks N2 and C1 in *PIF*-like TPases (see Fig. 3.1a for positions and *Materials and Methods* for sequences) and used to amplify a ~120 a.a. region from 20 grass species as well as several basal monocots (listed in Table 1). The amplified region included the majority of the catalytic domain in *PIF*-like TPases, extending from three a.a. upstream of the first Asp to eight a.a. upstream of the Glu of the DDE motif (Fig. 3.1a, boxed region). PCR products of the expected size (~360 bp) were successfully amplified from all 20 grasses tested and their close relative *Joinvillea*, as well as several Asparagales (e.g., *G. ilense*) (data not shown). In addition to the ~360-bp fragments, most species yielded larger PCR products (~450 bp) that, when sequenced, were found to contain an intron (see below).

Forty five fragments from fifteen species were sequenced; all were unique, indicating that there are multiple distinct TPases in each species and that only a small fraction had been sampled. No product was amplified from the more basal monocots such as Zamia, Ginkgo or Gnetum. Failure to amplify TPase fragments

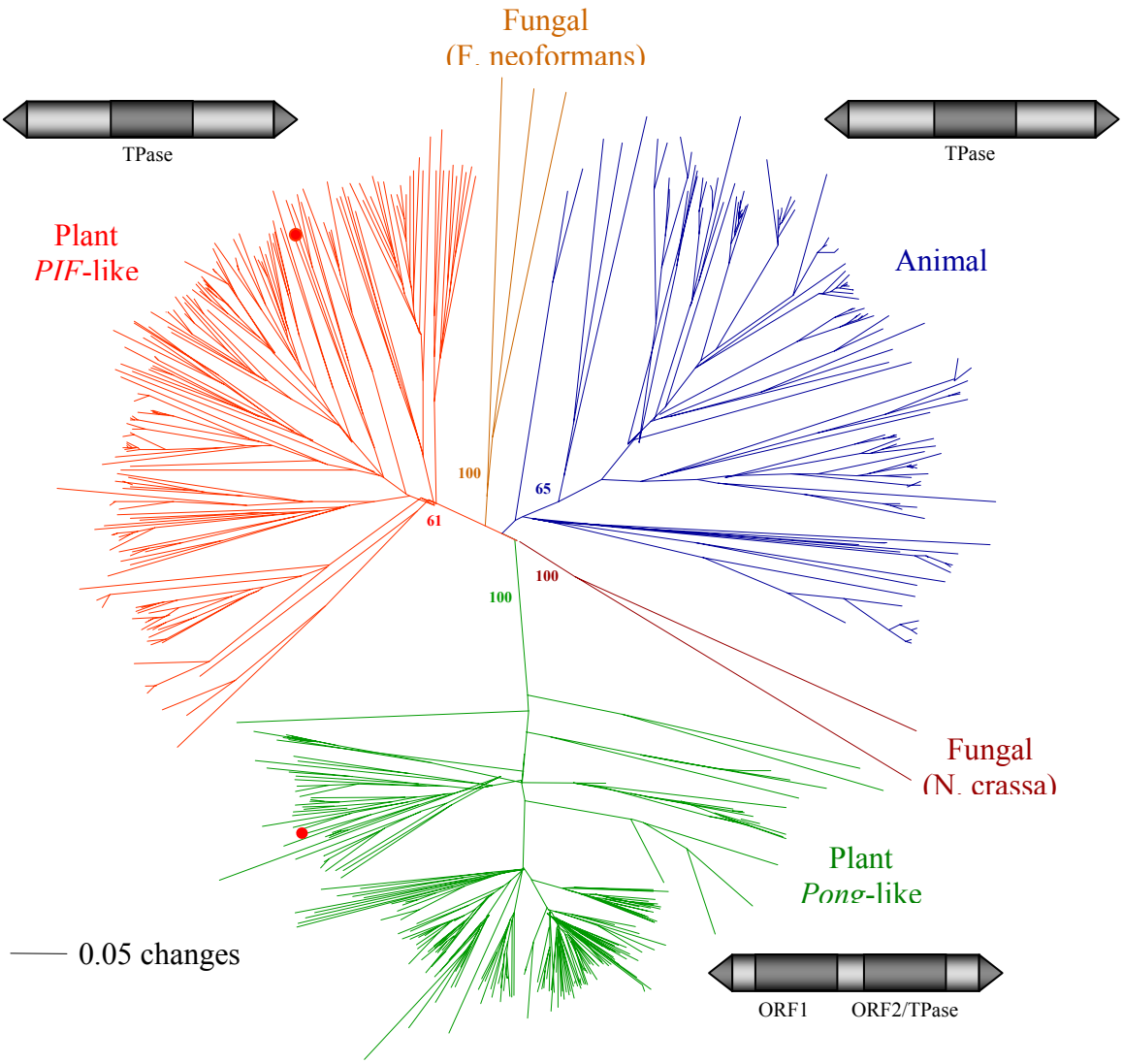
by dPCR from these species may be due to nucleotide variation in the primer recognition region or to the absence of *PIF*-like TPases.

Phylogeny of *PIF*-like and *Pong*-like TPases

Three major clusters of *PIF*- and *Pong*-like TPases. The TPase fragments identified by database mining and those isolated by dPCR were pooled and their evolutionary relationships examined. A multiple alignment was constructed from the 45 dPCR products and 574 unique database hits and used to generate an unrooted phylogenetic tree (Fig. 3.2). The majority of the sequences clustered into three groups: the plant *PIF*-like group, the plant *Pong*-like group and the animal group. In addition, the five fungal sequences clustered into two small, species-specific groups.

Clustering of the two plant groups was supported by bootstrap values as well as by several features that were shared within each group but not between groups. First, the spacing (i.e. numbers of residues) between the second Asp and the Glu of the DDE motif differed between *PIF*-like and *Pong*-like groups but was consistent within each group. *PIF*-like TPases exhibit DD47E or DD48E spacing whereas *Pong*-like TPases exhibit an invariant DD35E spacing. Second, where sufficient sequence information was available, nearly all *Pong*-like TPases were adjacent to an ORF1 homolog. No similar association with another conserved ORF was seen for the *PIF*-like TPases. Based on the comparison of TPase sequences, the animal group was equally related to both plant groups.

Figure 3.2. Phylogeny of *PIF* and *Pong*-like TPases. The unrooted tree was generated using the neighbor-joining method from a CLUSTALW multiple alignment of the catalytic domain from 619 *PIF*- or *Pong*-like TPases identified by database searches or isolated by dPCR (see text). The maize *PIF* TPase and rice *Pong* TPase are represented by the filled red circle in their respective group. Element structure of the plant *PIF*-like, plant *Pong*-like and the animal groups are shown. Triangles represent TIRs, light gray boxes represent non-coding regions, and dark gray boxes represent coding regions. Bootstrap values were calculated from 500 replicates.

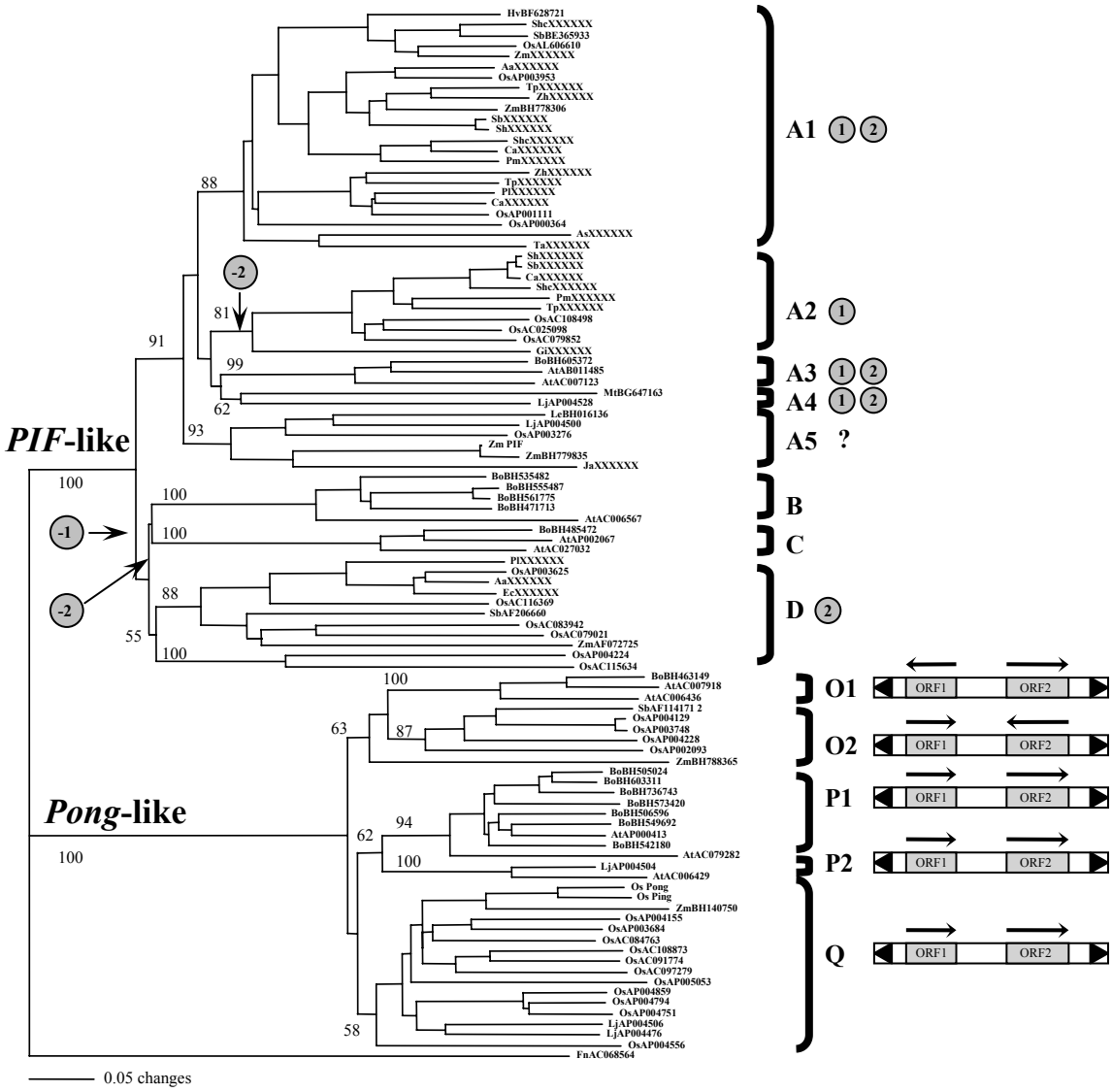

Fungal
(F. neoformans)
TPase
TPase
Plant
PIF-like
Animal
100
65
61
100
100
Fungal
(N. crassa)
Plant
Pong-like
0.05 changes
ORF1
ORF2/TPase

However, like the plant *PIF* group, the animal group was not associated with additional ORFs.

Phylogeny of plant *PIF*- and *Pong*-like TPases. Phylogenetic relationships among plant *PIF*- and *Pong*-like elements were determined by analyzing a subset of 99 sequences (63 *PIF*-like, 36 *Pong*-like) that were selected to represent the different lineages within each group. A CLUSTALW multiple alignment was constructed from the catalytic domains of these sequences and used to generate a phylogenetic tree (Fig. 3.3). Both plant groups are monophyletic and heterogeneous. In each group, amino acid identity between sequences from distantly related species can be higher than that between two sequences from the same or from a closely related species, suggesting the presence of multiple ancient lineages of both *PIF*- and *Pong*-like elements.

The plant *PIF*-like group is composed of four major lineages (A-D). Lineage A is the largest and most complex with members from both monocots and dicots. It can be further divided into five sublineages (A1-A5). A1 includes five grass subfamilies (Panicoideae, Ehrhartoideae, Bambusoideae, Pooideae and the ancestral Pharoideae), indicating that this sublineage was present before the diversification of the grasses approximately 70 Mya. Although only two grass subfamilies (Panicoideae and Ehrhartoideae) contributed sequences to A2, this lineage may be even more ancient than A1 as it is also found in the orchid *Gongora ilense* (order Asparagales). A3 and A4 are each found in a single dicot family (A3 in Brassicaceae and A4 in Fabaceae). A5 is present in both monocots

Figure 3.3. Phylogeny of plant *PIF*- and *Pong*-like TPases. The phylogenetic tree was generated using the neighbor-joining method from a CLUSTALW multiple alignment of 99 catalytic domains of *PIF*- and *Pong*-like TPases and rooted with the catalytic domain of a *PIF*-like TPase from *F. neoformans*. Presence of intron 1 and/or intron 2 in a certain *PIF*-like lineage is shown as a "1" or "2" in a gray circle. The intron content of lineage A5 is variable (see text). Loss of intron 1 and/or intron 2 from *PIF*-like TPases (a "-1" or "-2" in a gray circle) is indicated by arrows. The relative organizations of ORF1/TPase for *Pong*-like element lineages are shown (arrows indicate direction of transcription). Sequences are named according to the species initial (see Table 3.1) followed by the GenBank accession number. Bootstrap values were calculated from 1,000 replicates.


HvBF628721
ShcXXXXXX
SbBE365933
OsAL606610
ZmXXXXXX
AaXXXXXX
OsAP003953
TpXXXXXX
ZbXXXXXX
ZmBH778306
SbXXXXXX
ShXXXXXX
ShcXXXXXX
CaXXXXXX
PmXXXXXX
ZbXXXXXX
TpXXXXXX
PlXXXXXX
CaXXXXXX
OsAP001111
OsAP000364
AsXXXXXX
TaXXXXXX
ShXXXXXX
ShXXXXXX
CaXXXXXX
ShcXXXXXX
PmXXXXXX
TpXXXXXX
OsAC108498
OsAC025098
OsAC079852
GiXXXXXX
BoBH605372
AtAB011485
AtAC007123
MtBG647163
LjAP004528
LeBH016136
LjAP004500
OsAP003276
Zm PIF
ZmBH779835
JaXXXXXX
BoBH535482
BoBH555487
BoBH561775
BoBH471713
AtAC006567
BoBH485472
AtAP002067
AtAC027032
PlXXXXXX
OsAP003625
AaXXXXXX
EcXXXXXX
OsAC116369
SbAF206660
OsAC083942
OsAC079021
ZmAF072725
OsAP004224
OsAC115634
BoBH463149
AtAC007918
AtAC006436
SbAF114171 2
OsAP004129
OsAP003748
OsAP004228
OsAP002093
ZmBH788365
BoBH505024
BoBH603311
BoBH736743
BoBH573420
BoBH506596
BoBH549692
AtAP000413
BoBH542180
AtAC079282
LjAP004504
AtAC006429
Os Pong
Os Ping
ZmBH140750
OsAP004155
OsAP003684
OsAC084763
OsAC108873
OsAC091774
OsAC097279
OsAP005053
OsAP004859
OsAP004794
OsAP004751
LjAP004506
LjAP004476
OsAP004556
FnAC068564
PIF-like
Pong-like
-1
-2
-2
0.05 changes
A1 1 2
A2 1
A3 1 2
A4 1 2
A5 ?
B
C
D 2
O1
O2
P1
P2
Q
ORF1
ORF2

and dicots and includes the only known active *PIF*-like element, the maize *PIF*. B and C are two small lineages from dicots, both restricted to the Brassicaceae family. Lineage D is another monocot-specific lineage found in four grass subfamilies (Panicoideae, Ehrhartoideae, Bambusoideae, and Pharoideae).

Pong-like TPases clustered into three major lineages (O-P). Lineage O included two sublineages, the dicot-specific O1 and the monocot-specific O2. Lineage P is dicot-specific, suggesting that it emerged in dicots after their separation from monocots. P could also be divided into two sublineages (P1 and P2). P1 was only found in the Brassicaceae family and included the majority of *Pong*-like TPases from *A. thaliana* (71%) and nearly all from *B. oleracea* (137 of 139). The P2 sublineage is probably older than P1 as it is also present in the Fabaceae family. Most TPase sequences in lineages Q were from *O. sativa*. However, the presence of one sequence from *Z. mays* and two from *L. japonicus* in lineage Q suggests that it is also an ancient lineage.

Introns in plant *PIF*-like elements. Although the original maize *PIF* element lacks introns (ZHANG et al. 2001), many plant *PIF*-like TPases contain one or two introns in their catalytic domains. The boundaries of these introns (donor/acceptor sites) were predicted with very high confidence (90-100%), and the coding sequences were restored (compared to intron-less TPases) after their removal. Introns in *PIF*-like TPases can be classified into two classes based on their position (intron 1 and intron 2, Fig. 3.1a). Intron 1 is located six a. a. upstream of the first Asp residue of the DDE motif and intron 2 is located six a. a.

upstream of the second Asp residue. Both introns are short (83 bp on average) and A/T rich (71% on average), with little conservation in length or sequence either within or among species. Significantly, the intron number and position from *PIF*-like TPases were consistent with the lineage designations. Two introns were present in three sublineages of A (A1, A3 and A4), only intron 1 was present in A2, only intron 2 was present lineages D, and no introns were found in lineages B and C. Sublineage A5 was an exception: some TPases did not contain any intron while others contained two.

Two models have been proposed to explain the diversity of introns associated with related coding sequences: the “intron-early” model (loss of introns from an intron-rich ancestor) (GILBERT et al. 1997) or the “intron-late” model (addition of introns to an intron-less ancestor) (LOGSDON 1998). It is unlikely that the intron-late model explains the distribution of *PIF* introns as it would require multiple and independent intron acquisitions at identical positions. The intron-early explanation is more parsimonious since the data can be most easily interpreted by hypothesizing the existence of an ancestral *PIF* TPase with both introns and that multiple independent loss events occurred during evolution (Fig. 3.3). According to this model, both introns were retained in the ancestor of lineage A, but intron 2 was lost in sublineage A4. Intron 1 was lost in the common ancestor of lineages B through D so that none of these lineages contains intron 1. Intron 2 was subsequently lost in the common ancestor of lineages B and C. The predicted step-wise loss of introns from *PIF*-like TPase genes contrasts with

the plant *mariner*-like TPases, where the data is more consistent with the acquisition of introns during evolution.

ORF1/TPase arrangements. Three different arrangements of ORF1 and TPase were observed for *Pong*-like elements. All elements within a lineage or sublineage exhibit the same organization. Specifically, ORF1 and TPase were organized in a "head-to-head" alignment for O1, a "tail-to-tail" alignment for O2, and a "tail-to-head" alignment for P and Q (see Fig. 3.3).

PIF- and *Pong*-like Elements in Rice

The availability of virtually the entire genomic sequence of *O. sativa* (GOFF et al. 2002; YU et al. 2002) made it possible to conduct a comprehensive analysis of the relationships between *PIF*- and *Pong*-like elements and *Tourist*-like MITEs. To do this, *PIF*- and *Pong*-like TPase were first identified by computer-assisted analysis and then the sequences flanking these hits were searched to define full-length *PIF* and *Pong*-like elements.

***PIF* and *Pong* TPases.** tBlastn searches using as query the TPases of *PIF* and *Pong* led to the identification of 205 and 145 hits (E value < -10), respectively, from the TIGR rice database (ssp. *japonica*, cv. Nipponbare). Duplicate hits located on overlapping regions of BACs were excluded as were severely truncated TPases (containing less than 50% of the complete coding region). The remaining 116 *PIF*-like TPases and 80 *Pong*-like TPases were relatively full-

length and contained the entire catalytic domain. After removal of introns and correction of frameshifts caused by small insertion/deletions (1-2 bp), full-length *PIF*-like TPases were found to range in size from 392 to 432 a.a., while full-length *Pong*-like TPases were from 416 to 549 a.a..

The evolutionary relationships of rice *PIF*- and *Pong*-like TPases were determined by generating phylogenetic trees from CLUSTALW multiple alignments of their catalytic domains (Fig. 3.4 and 3.5). Two major lineages were resolved for *PIF*-like TPases that correspond to lineages A (including sublineages A1, A2 and A5) and D as shown in Fig. 3.3. Correlation between intron content of a *PIF*-like TPase and the TPase lineage is similar to that described in the broader plant survey (Fig. 3.3) except for two additional intron loss events: *OsPIF3* lost intron 1 and *OsPIF18* lost intron 2. *Pong*-like TPases also clustered into two major groups, corresponding to the sublineage O2 and lineage Q in Fig. 3.3. The ORF1/TPase alignment in lineage Q was found to be "tail-to-head" while that in lineage O2 is "head-to-head".

Characterizing full-length elements. *PIF*- and *Pong*-like TPases were grouped into families based on TPase sequence identities, with members of the same family being over 90% identical. In this way, 27 *PIF*-like families and 26 *Pong*-like families (including the *Pong* family) were defined. These families were designated *OsPIF* (for *O. sativa PIF*) and *OsPong* (for *O. sativa Pong*), followed by the number of the family. Elements of the same family were further designated with a letter (e.g., *OsPIF1a* and *OsPIF1b*, see Fig. 3.4 and 3.5).

Figure 3.4. Phylogeny of *OsPIF* TPases and the structure of the encoding element. The neighbor-joining tree was constructed from a CLUSTALW multiple alignment of the catalytic domains (boxed region in Fig. 3.1a) of 116 *OsPIF* TPases and rooted with the catalytic domain of *OsPong*. Bootstrap values were calculated from 1,000 replicates. Structure of *OsPIF* elements is depicted at the right. Black triangles represent element TIRs, colored lines represent non-coding regions, black boxes represent TPase genes, and open triangles indicate truncation in TPase genes. The positions of intron 1 (pink box) and intron 2 (blue box) are shown. The positions of insertions by other TEs are indicated by gray triangles above the elements (the identity and length of these insertions are described to the right). Dotted lines represent missing regions from elements that are incomplete because of gaps in genomic sequences or rearrangements after insertion (such as deletions or large insertions). The length of elements and TPase genes is drawn to scale. *OsPIF3d* and *OsPIF7b* were previously reported as *Os-PIF1* and *OsPIF-2*, respectively (ZHANG et al. 2001).

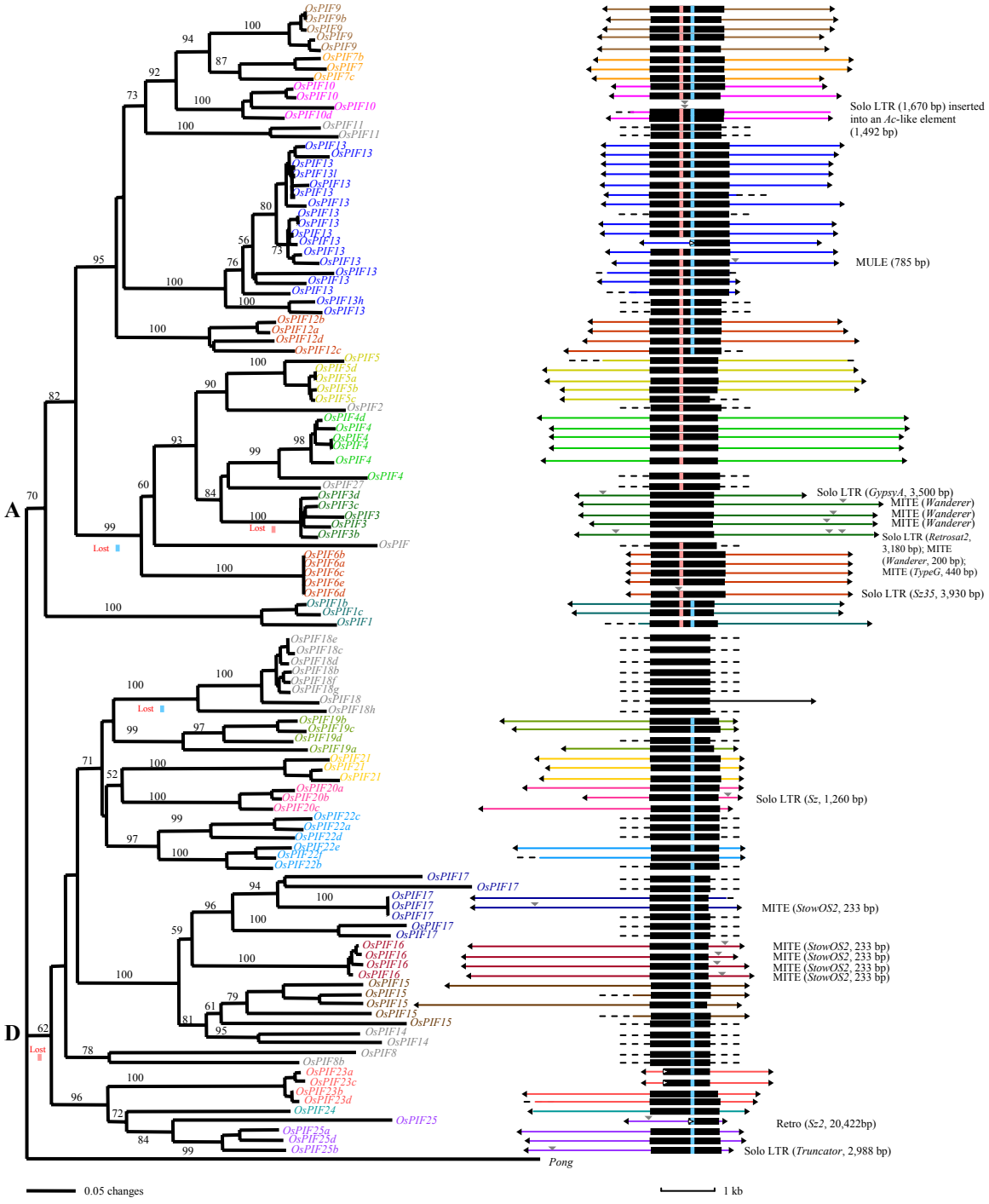
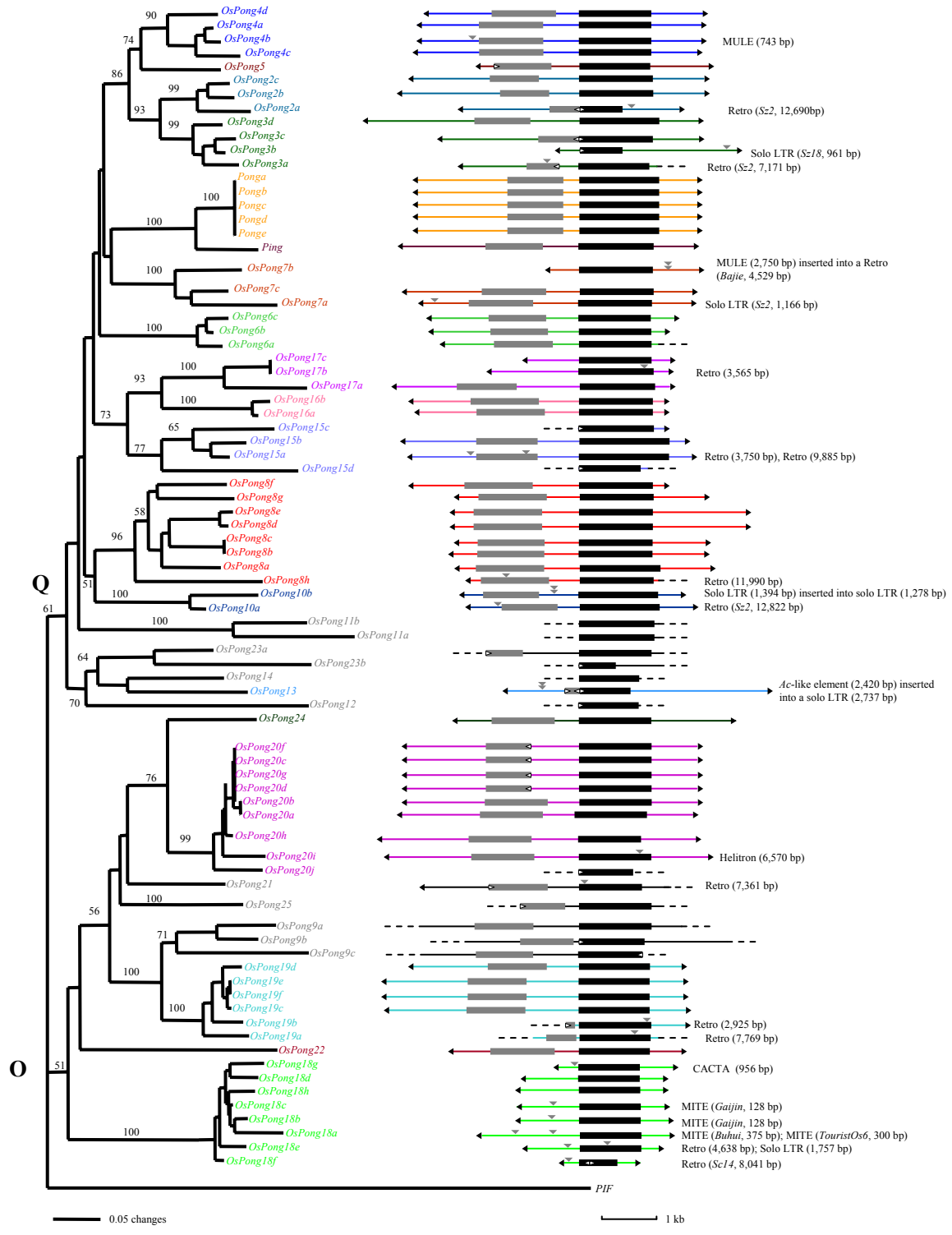

A
D
OsPIF9
OsPIF9b
OsPIF9
OsPIF9
OsPIF9
OsPIF7b
OsPIF7
OsPIF7c
OsPIF10
OsPIF10
OsPIF10
OsPIF10a
OsPIF11
OsPIF11
OsPIF13
OsPIF13
OsPIF13
OsPIF13l
OsPIF13
OsPIF13
OsPIF13
OsPIF13
OsPIF13
OsPIF13
OsPIF13
OsPIF13
OsPIF13
OsPIF13
OsPIF13
OsPIF13h
OsPIF13
OsPIF12b
OsPIF12a
OsPIF12d
OsPIF12c
OsPIF5
OsPIF5d
OsPIF5a
OsPIF5b
OsPIF5c
OsPIF2
OsPIF4d
OsPIF4
OsPIF4
OsPIF4
OsPIF4
OsPIF4
OsPIF27
OsPIF3d
OsPIF3c
OsPIF3
OsPIF3
OsPIF3b
OsPIF
OsPIF6b
OsPIF6a
OsPIF6c
OsPIF6e
OsPIF6d
OsPIF1b
OsPIF1c
OsPIF1
OsPIF18e
OsPIF18c
OsPIF18d
OsPIF18b
OsPIF18f
OsPIF18g
OsPIF18
OsPIF18h
OsPIF19b
OsPIF19c
OsPIF19d
OsPIF19a
OsPIF21
OsPIF21
OsPIF21
OsPIF20a
OsPIF20b
OsPIF20c
OsPIF22c
OsPIF22a
OsPIF22d
OsPIF22e
OsPIF22f
OsPIF22b
OsPIF17
OsPIF17
OsPIF17
OsPIF17
OsPIF17
OsPIF17
OsPIF16
OsPIF16
OsPIF16
OsPIF16
OsPIF15
OsPIF15
OsPIF15
OsPIF15
OsPIF14
OsPIF14
OsPIF8
OsPIF8b
OsPIF23a
OsPIF23c
OsPIF23b
OsPIF23d
OsPIF24
OsPIF25
OsPIF25a
OsPIF23d
OsPIF23b
Pong
Lost
Lost
Lost
Lost
0.05 changes
1 kb
Solo LTR (1,670 bp) inserted into an Ac-like element (1,492 bp)
MULE (785 bp)
Solo LTR (Gypsy4, 3,500 bp)
MITE (Wanderer)
MITE (Wanderer)
MITE (Wanderer)
Solo LTR (Retrosat2, 3,180 bp); MITE (Wanderer, 200 bp); MITE (TypeG, 440 bp)
Solo LTR (Sz35, 3,930 bp)
Solo LTR (Sz, 1,260 bp)
MITE (StowOS2, 233 bp)
MITE (StowOS2, 233 bp)
MITE (StowOS2, 233 bp)
MITE (StowOS2, 233 bp)
MITE (StowOS2, 233 bp)
Retro (Sz2, 20,422bp)
Solo LTR (Truncator, 2,988 bp)

Figure 3.5. Phylogeny of *OsPong* TPases and the structure of the encoding element. The neighbor-joining tree on the left was constructed from a CLUSTALW multiple alignment of the catalytic domains (boxed region in Fig. 3.1c) of 80 *OsPong* TPases and rooted with the catalytic domain of the maize *PIF*. Bootstrap values were calculated from 1,000 replicates. Structure of *OsPong* elements is shown at the right. Black triangles represent element TIRs, colored lines represent non-coding regions, gray boxes represent ORF1s, black boxes represent ORF2s, and open triangles indicate truncations in ORF1s or ORF2s. The positions of insertions by other TEs are indicated by gray triangles above the elements (the identity and length of these insertions are described to the right). Dotted lines represent missing regions from elements that are incomplete because of gaps in genomic sequences or rearrangements after insertion (such as deletions or large insertions). The length of elements as well as ORF1 and TPase genes is drawn to scale.


OsPong4d
OsPong4a
OsPong4b
OsPong4c
OsPong5
OsPong2c
OsPong2b
OsPong2a
OsPong3d
OsPong3c
OsPong3b
OsPong3a
Ponga
Pongb
Pongc
Pongd
Ponge
Ping
OsPong7b
OsPong7c
OsPong7a
OsPong6c
OsPong6b
OsPong6a
OsPong17c
OsPong17b
OsPong17a
OsPong16b
OsPong16a
OsPong15c
OsPong15b
OsPong15a
OsPong15d
OsPong8f
OsPong8g
OsPong8e
OsPong8d
OsPong8c
OsPong8b
OsPong8a
OsPong8h
OsPong10b
OsPong10a
OsPong11b
OsPong11a
OsPong23a
OsPong23b
OsPong14
OsPong13
OsPong12
OsPong24
OsPong20f
OsPong20c
OsPong20g
OsPong20d
OsPong20b
OsPong20a
OsPong20h
OsPong20i
OsPong20j
OsPong21
OsPong25
OsPong9a
OsPong9b
OsPong9c
OsPong19d
OsPong19e
OsPong19f
OsPong19c
OsPong19b
OsPong19a
OsPong22
OsPong18g
OsPong18d
OsPong18h
OsPong18c
OsPong18b
OsPong18a
OsPong18e
OsPong18f
PIF
Q
O
MULE (743 bp)
Retro (Sz2, 12,690bp)
Solo LTR (Sz18, 961 bp)
Retro (Sz2, 7,171 bp)
MULE (2,750 bp) inserted into a Retro (Bajie, 4,529 bp)
Solo LTR (Sz2, 1,166 bp)
Retro (3,565 bp)
Retro (3,750 bp), Retro (9,885 bp)
Retro (11,990 bp)
Solo LTR (1,394 bp) inserted into solo LTR (1,278 bp)
Retro (Sz2, 12,822 bp)
Ac-like element (2,420 bp) inserted into a solo LTR (2,737 bp)
Helitron (6,570 bp)
Retro (7,361 bp)
Retro (2,925 bp)
Retro (7,769 bp)
CACTA (956 bp)
MITE (Gaijin, 128 bp)
MITE (Gaijin, 128 bp)
MITE (Buhui, 375 bp); MITE (TouristOs6, 300 bp)
Retro (4,638 bp); Solo LTR (1,757 bp)
Retro (Sc14, 8,041 bp)
0.05 changes
1 kb

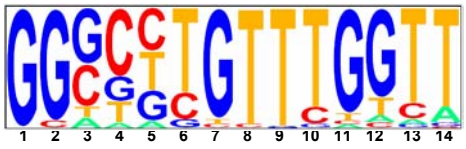
The identification of complete *OsPIF* and *OsPong* elements was complicated by the fact that inter-family comparisons indicated that sequence similarity was restricted to the known ORFs. For this reason, full-length elements were identified by comparison of sequences flanking the TPases within the same family where high sequence similarity extended into sequences flanking the ORFs. Sequences marking the boundary of similarity between elements of the same family were then searched for TIRs related to those of *PIF* or *Pong* and the flanking 3-bp TSDs characteristic of *PIF* and *Pong* elements (TTA/TAA). In this way, TIRs of 21 of the 27 *OsPIF* families (71 elements) and 20 of the 26 *OsPong* families (61 elements) were identified (see supplemental data for accession numbers and positions).

The TIRs of *OsPIFs* were of variable length, ranging from 10 bp (*OsPIF4*) to 45 bp (*OsPIF20*). In contrast, *OsPong* TIRs were more uniform: all were 14 to 18 bp long except for one family (represented by a single element *OsPong5*, 66-bp TIRs). Comparison of the TIRs of *OsPIFs* and *OsPongs* showed similarities (most began with 5'-GGSC-3', where S represents G or C) as well as differences (the fifth nucleotide was usually A in *OsPongs* but was rarely an A in *OsPIFs*) (Fig. 3.6a and 3.6b). The inner TIRs contained *PIF*-specific and *Pong*-specific motifs [*OsPIF*: 5'-TGTTTGGTT-3' (positions six to 14); *OsPong*: 5'-STMCAA-3' (positions seven to 12), where M stands for A or C].

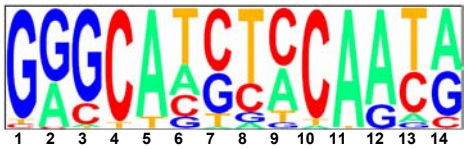
Full-length *OsPIFs* ranged from 2,305 bp (*OsPIF23a*) to 22,169 bp (*OsPIF25c*) and *OsPongs* ranged from 2,612 bp (*OsPong18d*) to 18,753 bp (*OsPong15a*). Most of this variation was due to the insertion of other TEs. These

Figure 3.6. Comparison of the TIRs of rice *PIF*-like elements, *Pong*-like elements and *Tourist*-like MITEs, shown in the form of pictograms. The terminal 14 bases from both ends are compared. The nucleotide A is shown in green, C in red, G in blue, and T in yellow. Numbers indicate the position of a nucleotide from the element end (e.g., "1" indicates the terminal nucleotide). The height of each nucleotide represents the relative frequency of that nucleotide at that position. (a) Pictogram of the TIRs from 21 *OsPIF* families (71 elements). (b) Pictogram of the TIRs from 20 *Pong*-like families (61 elements). (c) Pictogram of the TIRs from 20 previously described rice *Tourist*-like MITE families [*Tourist_Ia*, *Ib*, *Ic*, *III* (*Gaijin*), *IV* (*Castaway*), *V* (*Wanderer*), *VII*, *VIII*, *IX*, *XI*, *XII*, *XIV*, *XV*, *XVI*, *Type C*, *Buhui*, *Casin*, *Centre*, *Stone*, *Susu*]. (d) Pictogram of the TIRs from eight previously described rice *Tourist*-like MITE families (*Tourist_VI*, *Helia*, *Qiqi*, *ID-2*, *ID-3*, *ID-4*, *Lier*, *Stola*, *Youren*) (BUREAU et al. 1996; JIANG and WESSLER 2001; TARCHINI et al. 2000; TURCOTTE et al. 2001).

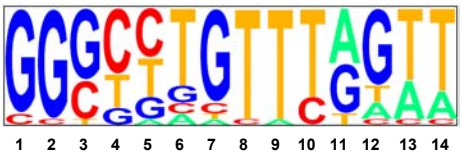
a ***OsPIF*** **TIRs**



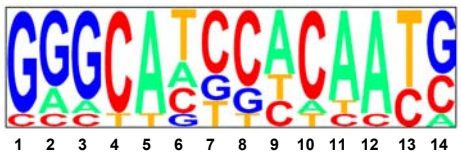
b ***OsPong*** **TIRs**



c ***PIF*****-like** ***Tourist*** **TIRs**



d ***Pong*****-like** ***Tourist*** **TIRs**



secondary TE insertions were located by searching full-length elements with RepeatMasker and Blastn. Sixteen *OsPIFs* and 24 *OsPongs* were found to contain a variety of TEs insertions (see Fig. 3.4 and 3.5 for their positions and identities), including other DNA elements [*Ac*-like, *Mutator*-like (MULEs) and CACTA-like), MITEs (*Tourist*-like and *Stowaway*-like), LTR retroelements, solo LTRs (*Copia*-like and *Gypsy*-like), non-LTR retroelements (LINEs) and in one case, a *Helitron* element. In a few instances, members of *OsPIF* and *OsPong* families (e.g., *OsPIF3*, *OsPIF16* and *OsPong18*) harbored the same MITE insertion at the same position, indicating that the MITE insertion did not prevent further transposition of these elements. When TE insertions were excluded, the length of most *OsPIFs* (50 of 71) and *OsPongs* (45 of 61) was found to be in the range of 4 to 6 kb.

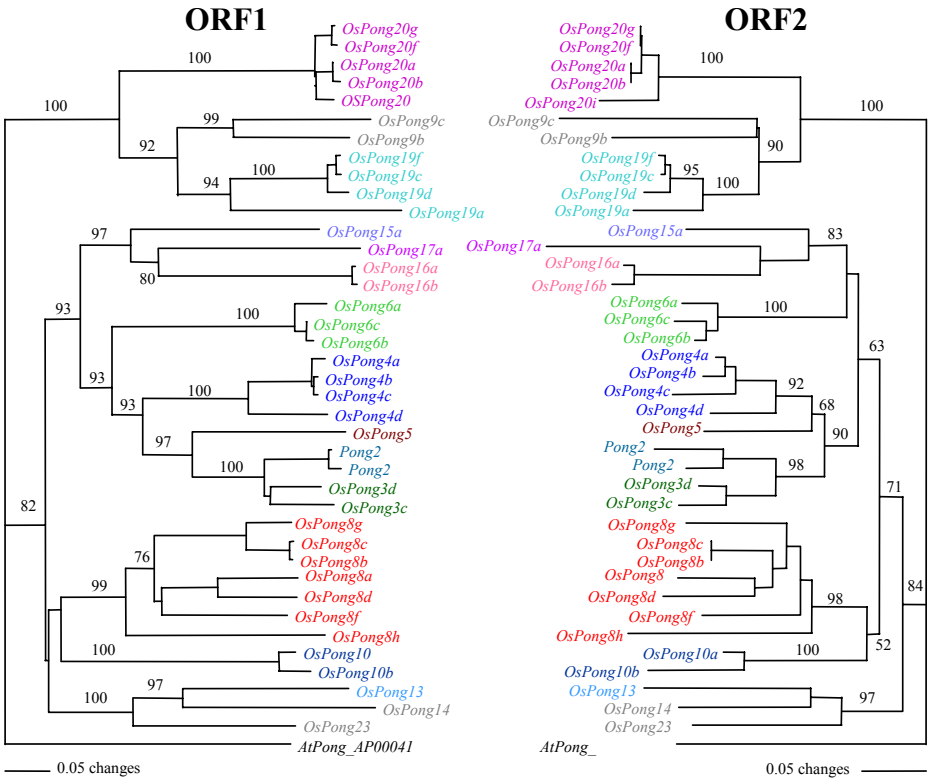
In several instances, *OsPIF* and *OsPong* families include elements that are nearly identical, suggesting that they transposed recently and may still be capable of further transposition. For example, *OsPIF6* includes five complete elements (~4.1 kb) located on four different chromosomes (Chr3, 7, 9 and 10) that are, on average, ~99.6% identical over their entire length. In addition, their coding sequences are not interrupted by stop codon or frameshifts. Similarly, *OsPong20* includes four elements (~5.3 kb) that are on average over 99.7% identical.

In contrast, interfamily sequence conservation, even between closely related families, was restricted to coding regions and TIRs. For example, the nucleotide sequences of the TPase genes of *OsPIF9* and *OsPIF7* were 60%

identical over ~1.2 kb, but these two families did not share additional sequence similarity aside from their TIRs. Similarly, the *OsPong2* and *OsPong3* families share 81% and 85% nucleotide identity in their ORF1 (~900 bp) and TPase (~1.3 kb) coding regions, respectively, but have completely diverged non-coding regions.

Co-evolution of ORF1 and TPase in *OsPongs*. Nineteen of the 20 *OsPong* families with defined termini encoded both ORF1 and TPase. As mentioned above, two alignments of ORF1 and TPase were observed for *OsPongs*: those in lineage Q have the "head-to-tail" alignment and those in lineage O2 have the "head-to-head" alignment. Inter-element recombination has been shown to be a significant force in the evolution of mobile elements (e.g. ADEY et al. 1994; JORDAN and MCDONALD 1998; LERAT et al. 1999; MCCLURE 1996). The presence of two ORFs whose organization varies in different *Pong*-like elements prompted us to investigate whether inter-element recombination had contributed to the evolution of *Pong*-like elements. To address this question, the phylogenies of ORF1 and the TPase of *OsPong* elements were compared to determine whether ORFs in the same element were co-evolving, or whether discrepancies existed that might suggest independent evolution. Two phylogenetic trees were generated for these elements, one based on a ~110 a.a. region in their ORF1s including all three conserved blocks (boxed region in Fig. 3.1b) and one based on the catalytic domains of their TPases (Fig. 3.7). Comparison of the two trees showed that the phylogenies determined from the two coding sequences were

Figure 3.7. Phylogenetic trees of (a) *OsPong* ORF1 and (b) *OsPong* ORF2. Trees were generated using the neighbor-joining method from CLUSTALW multiple alignments of the conserved regions in *OsPong* ORF1 (boxed in Fig. 3.1b) and the catalytic domain of ORF2 (boxed in Fig. 3.1c), respectively, and rooted with the corresponding regions of ORF1 and ORF2 from an *Arabidopsis Pong*-like element *AtPong_AP000413* (acc. no. AP000413; ORF1: 34,047-34412, ORF2: 32091-32453). Where multiple *OsPong* elements are identical in these regions, only one is shown. The name and color of *OsPongs* are the same as in Fig. 3.6. Bootstrap values were calculated from 1,000 replicates.


ORF1
OsPong20g
OsPong20f
OsPong20a
OsPong20b
OSPong20
OsPong9c
OsPong9b
OsPong19f
OsPong19c
OsPong19d
OsPong19a
OsPong15a
OsPong17a
OsPong16a
OsPong16b
OsPong6a
OsPong6c
OsPong6b
OsPong4a
OsPong4b
OsPong4c
OsPong4d
OsPong5
Pong2
Pong2
OsPong3d
OsPong3c
OsPong8g
OsPong8c
OsPong8b
OsPong8a
OsPong8d
OsPong8f
OsPong8h
OsPong10
OsPong10b
OsPong13
OsPong14
OsPong23
AtPong_AP00041
0.05 changes
ORF2
OsPong20g
OsPong20f
OsPong20a
OsPong20b
OsPong20i
OsPong9c
OsPong9b
OsPong19f
OsPong19c
OsPong19d
OsPong19a
OsPong15a
OsPong17a
OsPong16a
OsPong16b
OsPong6a
OsPong6c
OsPong6b
OsPong4a
OsPong4b
OsPong4c
OsPong4d
OsPong5
Pong2
Pong2
OsPong3d
OsPong3c
OsPong8g
OsPong8c
OsPong8b
OsPong8
OsPong8d
OsPong8f
OsPong8h
OsPong10a
OsPong10b
OsPong13
OsPong14
OsPong23
AtPong_
0.05 changes

consistent, indicating that inter-element recombination had probably not occurred.

The only *OsPong* family that does not harbor an ORF1 is *OsPong18*, where all eight elements only contain TPase. Absence of ORF1 in these elements is likely due to internal deletion for two reasons. First, the length of these elements is unusually short, ranging from 1,365 bp to 2,745 bp. Second, Blastn searches using *OsPong18* elements as queries identified several additional family members that do not encode TPase but contain coding sequence for ORF1 (e.g., AP003799; 96,105 - 99,317). Thus, all *OsPong* families with defined ends encode both ORF1 and TPase.

Insertion sites of *OsPIFs* and *OsPongs*. Prior studies have shown that some plant DNA transposons, such as members of the *Ac*/*Ds* and *Mutator* families, have a preference for insertion into single copy regions of the genome (CHEN et al. 1987; CRESSE et al. 1995; DIETRICH et al. 2002). Similarly, in a recent study it was shown that the majority of new *Pong* insertions (nine of 10) were into single copy sequences of the rice genome (JIANG et al. 2003). To test whether this is also true for *OsPIFs* and other *OsPongs*, the immediate flanking sequences (100 bp from each end) of all elements with defined termini were searched using RepeatMasker and tBlastn. Of the 132 *OsPIFs* and *OsPongs* examined, 109 (82.6%) were in single copy sequences, 13 (9.8%) were in other DNA elements, six (5.5%) were in retro-elements and four (3.0%) were in unknown repeats.

Relationships between *OsPIFs*, *OsPongs* and *Tourist*-like MITEs.

Identification of full length *OsPIF* and *OsPong* elements permitted the first genome-wide analysis of the relationship between these TPase-encoding elements and *Tourist*-like MITEs. Sequence identities between *OsPIFs* and *OsPongs* and *Tourist*-like MITEs were determined in two ways. First, we investigated whether *Tourist*-like MITE families could be associated with *OsPIFs* or *OsPongs* based on the sequences of their TIRs. To do this, the TIRs of 31 published rice *Tourist*-like MITE families were examined (BUREAU et al. 1996; JIANG and WESSLER 2001; TARCHINI et al. 2000; TURCOTTE et al. 2001). The TIRs of the majority of MITE families (28 of 31) were of two types, one (20 families) was strikingly similar to the TIRs of *OsPIFs* (Fig. 3.6c) and the other (nine families) was more similar to *OsPongs* (Fig. 3.6d). Second, more extensive sequence similarity between individual *OsPIF* or *OsPong* families and *Tourist*-like MITEs was examined. Terminal sequences (200 bp from each end) from the *OsPIF* and *OsPong* families with defined ends were used as queries to search a rice repetitive sequence database (N. Jiang, Z. Bao, S. R. Eddy and S. R. W., manuscript in preparation). Significant nucleotide similarity that extended beyond the TIRs was taken as evidence that an *OsPIF* or *OsPong* family was associated with a *Tourist*-like MITE family. Nine Of the 21 *Tourist*-like MITE families with *OsPIF*-like TIRs were found to be associated with *OsPIF* families, while three of the 9 *Tourist*-like MITE families with *OsPong*-like TIRs were found to be associated to *OsPong* families.

In some cases a MITE family was clearly identified as a deletion derivative of a particular *OsPIF* or *OsPong* family (Fig. 3.8). For example, the high copy number *Castaway* family (~3000 copies, BUREAU et al. 1996; JIANG and WESSLER 2001) appears to be derived by a simple deletion from the *OsPIF6* family. The apparent deletion breakpoints in *OsPIF6* occur at a 4 bp direct repeat (TTCC, underlined in Fig. 3.8a) that is only present as a single copy in *Castaway*. Similar relationships between several other *OsPIF* families and *Tourist*-like MITE families are shown in Fig. 3.8b, 3.8c and 3.8d. In contrast, although sequence similarities between *OsPong* and *Tourist*-like MITEs were detected, they were not as extensive as those seen with *OsPIF* (see Fig.3.8 e, f).

DISCUSSION

PIF- and *Pong*-like Elements Are Widespread and Abundant

Here we present the first comprehensive analysis of *PIF*- and *Pong*-like elements in eukaryotes. Prior studies reported that the TPases of *PIF* and *Pong* were distantly related to the bacterial IS*5* TPases and noted the presence of *PIF*-like elements in several plant (rice, sorghum and *Arabidopsis*), animal (*C. elegans*, *C. briggsae* and fugu fish) and fungal (*F. neoformans*) genomes (APARICIO et al. 2002; JIANG et al. 2003; KAPITONOV and JURKA 1999; LE et al. 2001; ZHANG et al. 2001). In this study, over 600 *PIF*- and *Pong*-like TPases were identified or isolated from 35 plants, 19 animals and two fungi. Phylogenetic analyses of these TPases defined three major groups, each represented by multiple distinct

Figure 3.8. Sequence similarity between *OsPIF* (a-d) or *OsPong*-like elements (e, f) and *Tourist*-like MITEs in rice. Each alignment represents an example of an *OsPIF* or *OsPong* family that shares significant sequence similarity with a *Tourist*-like MITE. *TouNJ-30* and *mPile10* were identified in this study, while other MITEs were previously reported. Arrows indicate element TIRs, open boxes indicate the regions used to generate the pictograms in Figure 6, and horizontal lines denote the location of direct repeats flanking the deletion breakpoints (discussed in text).

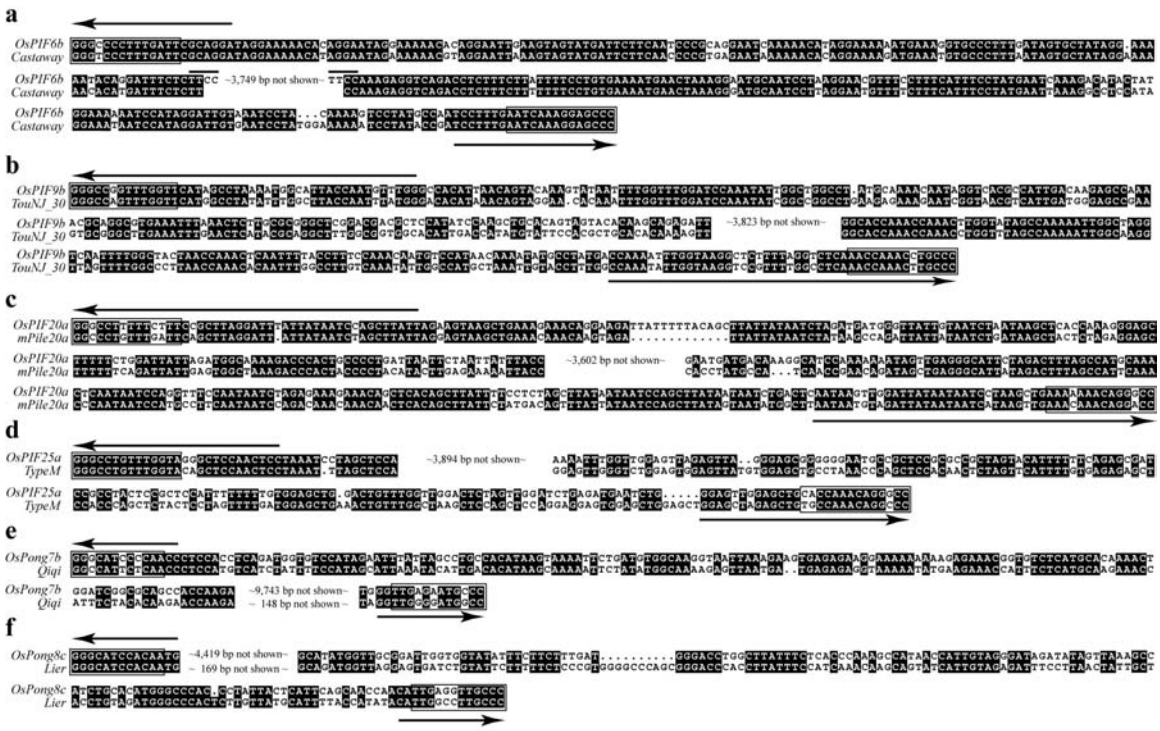

a
OsPIF6b
Castaway
~3,749 bp not shown~
b
OsPIF9b
TouNJ_30
~3,823 bp not shown~
c
OsPIF20a
mPile20a
~3,602 bp not shown~
d
OsPIF25a
TypeM
~3,894 bp not shown~
e
OsPong7b
Qiqi
~9,743 bp not shown~
~ 148 bp not shown ~
f
OsPong8c
Lier
~4,419 bp not shown~
~ 169 bp not shown ~

Figure 8

lineages. Taken together, the *PIF*/IS*5* superfamily of TPases is ancient and its members are widespread. To date, the only other TPase superfamily with such a broad distribution in eukaryotes and prokaryotes is IS630/*Tc1*/*mariner* (FESCHOTTE and WESSLER 2002; ROBERTSON 2002).*PIF*- and *Pong*-like elements are especially abundant in plants including both monocotyledons and dicotyledons. Large numbers of *PIF*- and *Pong*-like TPases were detected in three plants with relatively small genomes: ~80 copies in *Arabidopsis* (130 Mb); ~350 copies in rice (450 Mb) and over 1,000 copies in *B. oleracea* (extrapolated to ~600 Mb). Although significant sequence is not yet available for plants with large genomes, such as maize (2500 Mb) and barley (5000 Mb), the degenerate PCR assay indicates that these genomes also harbor multiple and diverse lineages (Fig. 3.3). Given that amplification of transposable elements is largely responsible for the huge differences in plant genome size (BENNETZEN 2002; FESCHOTTE et al. 2002a), it is reasonable to assume that even larger families of *PIF* and *Pong* will be found once these genomes are sequenced.

ORF1 of *Pong*

The rice *Pong* element encodes two ORFs (ORF1 and ORF2), of which ORF2 is most likely the TPase while the function of ORF1 is unknown. Database searches revealed a large number of homologs for both ORFs and where found, they were usually in "pairs" with an ORF1 homolog located within ~1-2 kb of an ORF2 homolog. All *OsPong* families with defined termini also encoded both

ORFs. The presence of a conserved ORF1 in virtually all *Pong*-like lineages suggests that it is necessary for the active transposition of *Pong*-like elements.

A requirement for a protein other than the transposase is unusual for a eukaryotic transposon, having only been described previously for members of the CACTA superfamily (KUNZE and WEIL 2002). Although the autonomous *Mutator* element *MuDR* from maize encodes two proteins (MURA and MURB), it is so far the only *Mutator* element shown to encode more than a single ORF among hundreds of MULEs examined (LISCH 2002; SINGER et al. 2001; YU et al. 2000). For CACTA-like elements, multiple proteins are encoded by alternatively spliced transcripts (e.g., TNPA and TNPD of the maize *En*/*Spm* element) (KUNZE and WEIL 2002). In contrast, our data indicates that ORF1 and ORF2 are separate transcription units. First, each ORF has a promoter that was predicted with high confidence by computer programs (data not shown). Second, elements harboring three distinct alignments of ORFs 1 and 2 were detected in plant genomes: “head-to-tail”, “head-to-head” and “tail-to-tail” (Fig. 3.3). The fact that ORFs1 and 2 would be transcribed from opposite strands in the head-to-head or tail-to-tail arrangements rules out the possibility that alternatively spliced transcripts are involved.

Several features of ORF1 provide clues to its possible function(s). Weak similarity between the most conserved region in ORF1 (Fig. 3.1, block A) and the *myb* DNA binding domain of some plant and animal transcription factors suggests that ORF1 may encode a DNA-binding protein (JIANG et al. 2003). The TPase of most transposons usually contain two functional domains, (1) a DNA-

binding domain that recognizes and binds specifically to the ends of all family members, and (2) a domain that catalyzes the transposition reactions. For many eukaryotic transposases, DNA-binding is usually associated with HTH domains that reside in the N-half of the protein [e.g., *Tc1*/*mariner*-like elements (PIETROKOVSKI and HENIKOFF 1997; PLASTERK et al. 1999) and *PIF*-like elements]. Interestingly, ORF2 of *Pong*-like elements has a recognizable catalytic domain, but has no recognizable DNA-binding domain. One can envision a model whereby the product of ORF1 binds to the ends of *Pong*-like elements and recruits ORF2 by protein-protein interactions. Alternatively, products of ORFs 1 and 2 may form heterodimer that binds to the element ends. If the products of ORFs 1 and 2 interact, it is reasonable to expect that they have been co-evolving, a feature consistent with the data presented in this study (Fig 3.7). That is, the phylogenies of ORF1 and ORF2 are very consistent and no interfamily rearrangement was found (Fig. 3.7). The requirements of *Pong* transposition as well as the possible interaction between ORF1 and ORF2 are under investigation.

OsPIF **and Os*Pong*** **Elements.**

This study identified 116 *OsPIF* and 80 *OsPong* TPases representing all of the lineages of *PIF* and *Pong* Tpases detected in monocot genomes. As such, rice is a suitable model to study the evolution of *PIF*- and *Pong*-like elements in plants as well as their relationship with *Tourist*-like MITEs.

OsPIF and *OsPong* elements were grouped into 27 and 26 families, respectively, based on sequence identity of their coding regions. These groupings received additional support when it was determined that elements of the same family share extensive sequence similarity in non-coding regions. Several *OsPIF* and *OsPong* families (such as *OsPIF4*, *5*, *6*, *9*, *12*, *13*, *23* and *OsPong8*, *17*, *19*, *20*) include members that are nearly identical. Furthermore, each family includes at least one putative autonomous member whose coding region is not interrupted by a stop codon or a frameshift mutation. These features are indicative of recent and perhaps ongoing activity of multiple *OsPIF*/*OsPong* families.

OsPIF, Os*Pong* and *Tourist*-like MITEs

In previous studies, *PIF* and *Pong* elements were isolated as the TPase sources for two families of *Tourist*-like MITEs, *mPIF* and *mPing*, respectively (JIANG et al. 2003; ZHANG et al. 2001). In this study, many other *PIF*- and *Pong*-like elements were identified, including all the families in rice. Characterization of these elements permitted a comprehensive analysis of the relationships between these TPase-encoding elements and *Tourist*-like MITEs. The major conclusion from this analysis is that most *Tourist*-like MITE families are related to either *PIF*- or *Pong*-like elements based solely on a comparison of their TIRs. Of the 31 previously described *Tourist*-like MITE families in rice, the TIRs of 20 were found to be more closely related to the consensus *OsPIF* TIR while the TIRs of nine were more closely related to the consensus *OsPong* TIR (Fig. 3.6).

Attempts to associate individual *Tourist*-like MITE families with specific *OsPIF*- or *OsPong*- families uncovered many clear-cut relationships (Fig. 3.8). For example, the MITE family *Castaway* (~3,000 copies) was found to be derived from the *OsPIF6* family by internal deletion and subsequent amplification. Relationships between *Tourist*-like MITEs and *OsPong* families are less apparent as sequence similarity is limited to the subterminal regions (as shown in Fig. 3.8e and 3.8f). However, detection of sequence identity between subterminal regions of *OsPong* families and *Tourist*-MITE families, albeit limited, is significant in light of the fact that even closely related *OsPong* families display no sequence identity in their subterminal regions.

In summary, the characterization of virtually all full length *PIF*- and *Pong*-like elements in the rice genome has permitted a determination of the extent of their relatedness with most of the 60,000 *Tourist*-like MITEs residing in this genome. Our data indicates that many *Tourist*-like MITEs originated from *OsPIF* and *OsPong* elements by internal deletion and subsequent amplification. However, 16 of the 28 *Tourist*-like MITEs examined in this study were not clearly associated with *OsPIF*/*OsPong* families. It is possible that their cognate *OsPIF*/*OsPong* families were lost from the genome. Such a scenario is not difficult to imagine considering *OsPIF*/*OsPong* elements are present at much lower copy number (several per families) than *Tourist*-like MITEs (hundreds or thousands per family). Alternatively, some *Tourist*-like MITE families may have originated by chance events, where, for example, a pair of nearby inverted repeats (and other *cis*-requirements, if any) were mobilized fortuitously by an

endogenous *PIF*- or *Pong*-like TPase and subsequently amplified to high copy numbers.

The rice genome harbors over 90,000 MITEs: 60,000 *Tourist*-like MITEs and 30,000 *Stowaway*-like MITEs (FESCHOTTE et al. 2003; N. Jiang and S. R. Wessler, unpublished data). Whereas elements of the *PIF*/*Pong*/IS5 superfamily (*OsPIF* and *OsPong*) appear to be responsible for the origin and amplification of *Tourist*-like MITEs, rice elements related to the *Tc1*/*mariner* superfamily (called *Osmars*) have been associated with *Stowaway*-like MITEs (FESCHOTTE et al. 2003). Given that plant genomes are known to harbor many other DNA transposon families (including CACTA-like, *hAT*-like and MULEs), one obvious question is, why are most MITE families associated with only these two superfamilies? A key factor may be the *cis*-requirements for transposition. Transposition of MULEs requires long TIRs (over 200 bp) (BENITO and WALBOT 1997), while that of CACTA-like and *hAT*-like elements requires both TIRs and subterminal repetitive motifs (reviewed in KUNZE and WEIL 2002). Several studies have shown that for these elements, multiple TPase binding sites reside in the subterminal repeats. In contrast, the binding sites for animal *mariner* TPases seem to be restricted to the short element TIRs (~28 to 31 bp) (AUGE-GOUILLOU et al. 2001; LAMPE et al. 2001; ZHANG et al. 2001). Indeed, artificial transposons containing just the *mariner* TIRs have been successfully mobilized by the cognate *mariner* TPase, *in vitro* and *in vivo* in bacteria (e.g. LAMPE et al. 1999; LAMPE et al. 2001; TOSI and BEVERLEY 2000). Although the *cis*-requirements for transposition of *PIF and Pong*-like elements has yet to be determined, these

elements do not contain any apparent repetitive motifs in subterminal regions, suggesting that their TPase-binding sites may also reside in their short TIRs. Thus, the minimal *cis*-requirements for transposition by members of the *mariner*-like and *PIF*/*Pong*-like families may significantly enhance the probability of generating shorter elements by deletion from larger elements or by chance, that could be mobilized by the TPases encoded by these families.

Concluding Remarks

The *PIF* and *Pong* elements are founding members of a very large and dynamic superfamily of class 2 elements that are widespread in flowering plant. The impact of these elements is significant, as *PIF*- and *Pong*-like families are capable of expansion through the amplification and diversification of both autonomous and nonautonomous members, including very high copy number MITEs. Furthermore, with a demonstrated preference for insertion into genic regions, *PIF*- and *Pong*-like elements and their associated *Tourist*-like MITEs appear to be a major force generating genetic diversity and influencing the evolution of plants.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Science Foundation and National Institute of Health to S.R.W.

MATERIALS AND METHODS

PCR amplification of *PIF*-like TPases

Degenerate primers were derived from the regions encoding amino acid residues GALDGTH (D1F1; 5' -GGIGCHHTIGATGGHACWCA- 3') and ELFNPRH (KR1; 5' -ATGICKMIRRTTRAACAAYTC- 3') (Fig. 3.1a, positions indicated by arrows). PCR amplifications were performed with 10-100 ng of genomic DNA in 50 µl reactions. Cycling parameters were: one cycle at 94 °C for 3 min, 36 cycles at 94 °C for 30s, 50 °C for 30s, 72 °C for 1min, and one cycle at 72 °C for 5 min. Forty µl of the reaction were resolved on 1% agarose gels, and desired fragments were purified from agarose using the QIAquick Gel Extraction Kit (QIAGEN) and cloned using the TOPO-TA Cloning Kit (Invitrogen) according to manufacturers' instructions. Sequencing reactions were performed by the Molecular Genetics Instrumentation Facility of the Univ. of Georgia. The sequences of 45 *PIF*-like TPase fragments were deposited in the GenBank database (accession nos. XXXXXXXX-XXXXXXXX).

Database searches, sequence and phylogenetic analyses

Database searches were performed with blast servers available from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov, databases nr, gss, est and wgs_Anopheles) against GenBank nr, est, gss and wgs_anopheles databases as well as the rice genomic database at The Institute for Genomic Research (http://tigrblast.tigr.org/euk-blast/index.cgi?project=osa1).

Nucleotide sequences obtained from database searches and dPCR amplifications were conceptually translated into amino acid sequences and aligned with CLUSTALW with default parameters. Introns were predicted with Netgene2 (available at http://www.cbs.dtu.dk) (HEBSGAARD et al. 1996)]. Putative HTH motifs were predicted using the NPS@ program (available at http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_hth.html). Multiple sequence alignments used to generate the phylogenetic tree in Fig. 3.2 were performed with the CLUSTALW server available at European Bioinformatics Institute (http://www.ebi.ac.uk/clustalw/) with default parameters. Multiple alignments used to generate other phylogenetic trees were performed using the MacVector program. Phylogenetic trees were generated based on the neighbor-joining method, using PAUP* Version 4.0b8 (SWOFFORD 1999) with default parameters. Pictograms were generated at http://genes.mit.edu/pictogram.html.

LITERATURE CITED

ADEY, N. B., S. A. SCHICHMAN, D. K. GRAHAM, S. N. PETERSON, M. H. EDGELL et al., 1994 Rodent l1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. Mol. Biol. Evol. **11:** 778-789.

APARICIO, S., J. CHAPMAN, E. STUPKA, N. PUTNAM, J. M. CHIA et al., 2002 Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science **297:** 1301-1310.

AUGE-GOUILLOU, C., M. H. HAMELIN, M. V. DEMATTEI, M. PERIQUET and Y. BIGOT, 2001 The wild-type conformation of the Mos-1 inverted terminal repeats is suboptimal for transposition in bacteria. Mol. Genet. Genomics **265:** 51-57.

BENITO, M. I., and V. WALBOT, 1997 Characterization of the maize Mutator transposable element MURA transposase as a DNA-binding protein. Mol. Cell. Biol. **17:** 5165-5175.

BENNETZEN, J. L., 2002 Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29-36.

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 wild-type rice genes. Proc. Natl. Acad. Sci. USA **93:** 8524-8529.

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

CHEN, J., I. GREENBLATT and S. DELLAPORTA, 1987 Transposition of *Ac* from the *P* locus of maize into unreplicated chromosomal sites. Genetics **117:** 109-116.

CRESSE, A. D., S. H. HULBERT, W. E. BROWN, J. R. LUCAS and J. L. BENNETZEN, 1995 Mu1-related transposable elements of maize preferentially insert into low copy number DNA. Genetics **140:** 315-324.

DIETRICH, C. R., F. CUI, M. L. PACKILA, J. LI, D. A. ASHLOCK et al., 2002 Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and

sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics **160:** 697-716.

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

FESCHOTTE, C., L. SWAMY and S. R. WESSLER, 2003 Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with *Stowaway* MITEs. Genetics ***163*:** 747-758.

FESCHOTTE, C., and S. R. WESSLER, 2002 *Mariner*-like transposases are widespread and diverse in flowering plants. Proc. Natl. Acad. Sci. USA **99:** 280-285.

FESCHOTTE, C., X. ZHANG and S. WESSLER, 2002b Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons, pp. 1147-1158 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, DC.

GILBERT, W., S. J. DE SOUZA and M. LONG, 1997 Origin of genes. Proc. Natl. Acad. Sci. USA **94:** 7698-7703.

GOFF, S. A., D. RICKE, T. H. LAN, G. PRESTING, R. WANG et al., 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92-100.

HEBSGAARD, S. M., P. G. KORNING, N. TOLSTRUP, J. ENGELBRECHT, P. ROUZE et al., 1996 Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. Nucleic Acids Res. **24:** 3439-3452.

JIANG, N., Z. BAO, X. ZHANG, H. HIROCHIKA, S. R. EDDY et al., 2003 An active DNA transposon family in rice. Nature **421:** 163-167.

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.

JORDAN, I. K., and J. F. MCDONALD, 1998 Evidence for the role of recombination in the regulatory evolution of *Saccharomyces cerevisiae Ty* elements. J. Mol. Evol. **47:** 14-20.

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

KELLOGG, E. A., 2001 Evolutionary history of the grasses. Plant Physiol. **125:** 1198-1205.

KIKUCHI, K., K. TERAUCHI, M. WADA and H. Y. HIRANO, 2003 The plant MITE mPing is mobilized in anther culture. Nature **421:** 167-170.

KUNZE, R., and C. F. WEIL, 2002 The hAT and CACTA superfamilies of plant transposons, pp. 565-610 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington DC.

LAMPE, D. J., B. J. AKERLEY, E. J. RUBIN, J. J. MEKALANOS and H. M. ROBERTSON, 1999 Hyperactive transposase mutants of the Himar1 mariner transposon. Proc. Natl. Acad. Sci. USA **96:** 11428-11433.

LAMPE, D. J., K. K. WALDEN and H. M. ROBERTSON, 2001 Loss of transposase-DNA interaction may underlie the divergence of mariner family

transposable elements and the ability of more than one mariner to occupy the same genome. Mol. Biol. Evol. **18:** 954-961.

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

LERAT, E., F. BRUNET, C. BAZIN and P. CAPY, 1999 Is the evolution of transposable elements modular? Genetica **107:** 15-25.

LISCH, D., 2002 Mutator transposons. Trends Plant Sci. **7:** 498-504.

LOGSDON, J. M., JR., 1998 The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. **8:** 637-648.

MAHILLON, J., and M. CHANDLER, 1998 Insertion sequences. Microbiol Mol Biol Rev **62:** 725-774.

MCCLURE, M. A., 1996 The complexities of viral genome analysis: the primate lentiviruses. Curr. Opin. Genet. Dev. **6:** 749-756.

NAKAZAKI, T., Y. OKUMOTO, A. HORIBATA, S. YAMAHIRA, M. TERAISHI et al., 2003 Mobilization of a transposon in the rice genome. Nature **421:** 170-172.

PIETROKOVSKI, S., and S. HENIKOFF, 1997 A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. Mol. Gen. Genet. **254:** 689-695.

PLASTERK, R. H. A., Z. IZSVÁK and Z. IVICS, 1999 Resident aliens: the Tc1/*mariner* superfamily of transposable elements. Trends Genet. **15:** 326-332.

REZSOHAZY, R., B. HALLET, J. DELCOUR and J. MAHILLON, 1993 The IS4 family of insertion sequences: evidence for a conserved transposase motif. Mol. Microbiol. **9:** 1283-1295.

ROBERTSON, H. M., 2002 Evolution of DNA transposons, pp. 1093-1110 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, DC.

SINGER, T., C. YORDAN and R. A. MARTIENSSEN, 2001 Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes Dev. **15:** 591-602.

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

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.

TOSI, L. R., and S. M. BEVERLEY, 2000 cis and trans factors affecting Mos1 mariner evolution and transposition in vitro, and its potential for functional genomics. Nucleic Acids. Res. **28:** 784-790.

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

WALKER, E. L., W. B. EGGLESTON, D. DEMOPULOS, J. KERMICLE and S. L. DELLAPORTA, 1997 Insertions of a novel class of transposable elements with a strong target site preference at the *r* locus of maize. Genetics **146:** 681-693.

WESSLER, S. R., T. E. BUREAU and S. E. WHITE, 1995 LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr. Opin. Genet. Dev. **5:** 814-821.

YU, J., S. HU, J. WANG, G. K. WONG, S. LI et al., 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79-92.

YU, Z., S. I. WRIGHT and T. E. BUREAU, 2000 Mutator-like Elements in Arabidopsis thaliana. Structure, diversity and evolution. Genetics **156:** 2019-2031.

ZHANG, L., A. DAWSON and D. J. FINNEGAN, 2001 DNA-binding activity and subunit interaction of the mariner transposase. Nucleic Acids. Res. **29:** 3566-3575.

ZHANG, X., C. FESCHOTTE, Q. ZHANG, N. JIANG, W. B. EGGLESTON et al., 2001 *P Instability Factor*: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. Proc. Natl. Acad. Sci. USA **98:** 12572-12577.

CHAPTER 4

COMPARATIVE ANALYSES OF TRANSPOSABLE ELEMENTS IN *ARABIDOPSIS THALIANA* AND *BRASSICA OLERACEA*[1]

[1] Zhang, X and Wessler, S.R. To be submitted to *Proc. Natl. Acad. Sci. USA*.

Abstract

The model dicot plant *Arabidopsis thaliana* has one of the smallest known plant genomes (~125 Mb), largely due to its low TE content. *Brassica oleracea* is closely related to *Arabidopsis* (diverged ~15-20 Mya), yet its genome (~600 Mb) is approximately five times as large as the *Arabidopsis* genome. To address the questions regarding the underlying reason for the low TE content of the *Arabidopsis* genome and the contribution of TE proliferation to the genome size expansion of *B. oleracea*, we performed a comparative analysis of all major types of TEs in these two species. A set of strategies were derived and used to identify all TEs with significant coding sequence from *Arabidopsis* and *B. oleracea*, and the abundance, diversity and phylogenies of TEs in the two species were compared. These comparisons allowed us to infer the number of TE lineages present in the last common ancestor of *Arabidopsis* and *B. oleracea* as well as how each lineage has evolved since the divergence of the species. The results indicated that the amplification of both class 1 and class 2 elements contributed significantly to the genome size expansion of *B. oleracea*, and that the low TE content of *Arabidopsis* is largely due to the lack of any significant TE amplification in its genome.

Introduction

The model dicotyledonous plant *Arabidopsis thaliana* offers many advantages for basic and applied plant research, including its small genome size (~125Mb), miniature stature, short life cycle and the ease of laboratory maintenance and propagation. The complete genome of *Arabidopsis* is also available to facilitate discovery and functional studies of genes as well as comparative studies of genome organization and evolution (The *Arabidopsis* Genome Initiative 2000).

Brassica oleracea belongs to the same taxonomic family (Brassicaceae) as *Arabidopsis*. The two species are closely related (divergence ~15-20 Mya) (Yang et al. 1999) and share ~85% nucleotide sequence identity in protein coding regions (Cavell et al. 1998). Several *B. oleracea* cultivars are of worldwide economical importance, such as cabbage, kale, broccoli, cauliflower, Brussels sprouts and kohlrabi. Comparative mapping of *B. oleracea* and *Arabidopsis* has revealed collinear organizations of genes (synteny) in smaller chromosomal intervals (10 cM or less), but has also uncovered extensive chromosomal rearrangements (Kowalski et al. 1994; Sadowski et al. 1996; Paterson et al. 1996).

The *B. oleracea* genome (~600 Mb) (Arumuganathan and Earle 1991) has expanded through triplication since its divergence from *Arabidopsis* (Cavell et al. 1998; Lagercrantz 1998; Lan et al. 2000). However, genome triplication cannot fully explain the ~5 times genome size difference between the two species. Thus,

other mechanisms have been involved in the genome size expansion of *B. oleracea* since it diverged from the last common ancestor with *Arabidopsis*. Since the proliferation of transposable elements (TEs) has been implicated in the expansion of grass genomes (SanMiguel et al. 1996; SanMiguel et al. 1998), it is possible that they are also involved in the recent genome size expansion of *B. oleracea*.

TEs are ubiquitous in all characterized eukaryotic genomes and have been divided into two classes based on their transposition intermediates (Capy et al. 1998). Class 1 (RNA) elements transpose via an RNA intermediate and most either have long terminal repeats (LTR-retrotransposons) or terminate at one end with a polyA tract [long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)] (Doolittle et al. 1989; Xiong and Eickbush 1990). LTR-retrotransposons have been further classified as either *Ty1*/*copia*-like or *Ty3*/*gypsy*-like elements based on the order of their encoded proteins (Xiong and Eickbush 1990). Class 2 (DNA) elements transpose via a DNA intermediate and usually have terminal inverted repeats (TIRs). They have been grouped into superfamilies (e.g., *Tc1*/*mariner* (Plasterk and van Luenen 2002; Feschotte and Wessler 2002), *hAT* (Kunze and Weil 2002), CACTA (Rubin et al. 2001; Kunze and Weil 2002), MULEs (Yu et al. 2000; Mao et al. 2000; Lisch 2002) and *PIF*/*Pong* superfamilies (Zhang et al. 2001; Jiang et al. 2003) based on homology in the transposases (TPases) that catalyze their movement.

TEs are major components of plant genomes and contribute significantly to the >1,000 fold genome size variation among plants. For example, over 80%

of the maize genome (2,500 Mb) is comprised of TEs (SanMiguel et al. 1996; SanMiguel et al. 1998; Meyers et al. 2001). In contrast, although *Arabidopsis* harbors all the TE types found in larger plant genomes, TE copy number is generally low; all TEs account for only ~10% of total nuclear DNA. Low TE content could be explained by one of two possible mechanisms: either lack of TE amplification or systematical removal (deletion) from the genome (Bennetzen and Kellogg 1997; Devos et al. 2002).

To address the question regarding the contribution of TE proliferation to genome size expansion of *B. oleracea* and to distinguish between the two possible causes for the low TE content of *Arabidopsis*, we compared the abundance, diversity and phylogeny of TEs in these two species. Such a comparison was possible because large amounts of sequence information are available for both species. To facilitate the annotation of the sequenced *Arabidopsis* genome, a shotgun sequence database for *B. oleracea* was recently generated by The Institute for Genomic Research (TIGR). The current *B. oleracea* database covers approximately 1/3 of the genome (~220 Mb of ~600 Mb) and consists of ~350,000 short reads (average length ~650 bp) that were sequenced from the ends of 2.5 kb genomic clones.

In this study, several strategies were deduced and used in a comparative analysis of the complete *Arabidopsis* genomic sequence and the fragmentary *B. oleracea* database. The purpose of this study was to identify all known TEs with significant coding capacities and estimate their copy number. Comparison of the diversity, abundance and phylogeny of TEs in *Arabidopsis* and *B. oleracea*

allowed us to infer the TE diversity in their last common ancestor and to determine how each type of TE has evolved since the divergence of the two species. Nearly all TE lineages were found in both species, but are generally present at higher copy number in *B. oleracea* than in *Arabidopsis.* Class 1(retro) elements were found to be the most abundant TE class in both genomes with both LTR- and non-LTR elements comprising the largest fraction of total genomic DNA. Surprisingly, several families of DNA elements have amplified to very high copy number in *B. oleracea* and have contributed significantly to genome size increase. These results indicate that, whereas the amplification of RNA elements can largely account for the genome size variations among grasses, the recent amplification of both RNA and DNA element is responsible for the genome size increase of *B. oleracea*. Furthermore, our results suggested that the low TE content in *Arabidopsis* was due to the lack of significant TE amplification in its genome and not to wholesale deletion of TEs as has been suggested by a recent study (Devos et al. 2002).

Results

Comparison of TE abundance in *Arabidopsis* and *B. oleracea*.

The abundance of TEs in *Arabidopsis* and *B. oleracea* was determined as described in *Materials and Methods*. Briefly, for each type of TE (e.g., CACTA-like elements or *copia*-like LTR elements), the coding sequences from previously described elements were compared by CLUSTALW multiple alignments and the

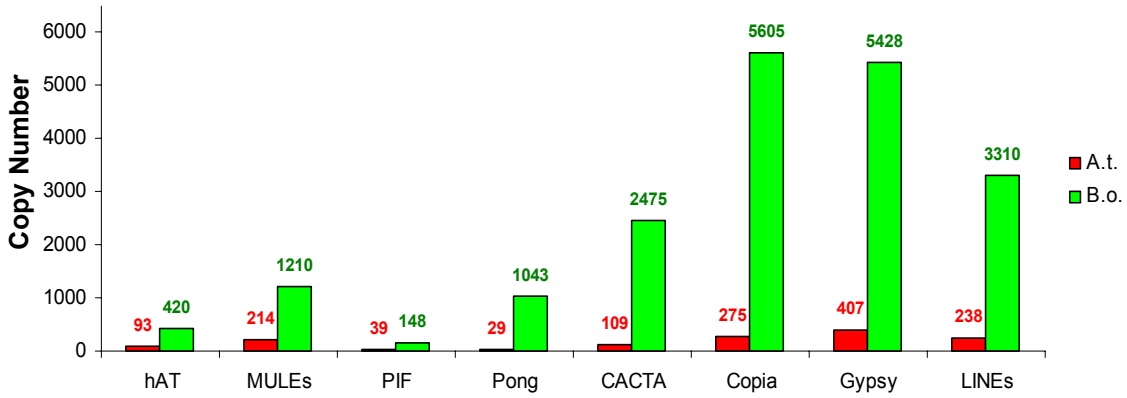
most conserved regions were identified. These regions were used as queries in tBlastn searches to identify all related *Arabidopsis* TEs. The resulting *Arabidopsis* hits were compared by CLUSTALW multiple alignments and analyzed phylogenetically to resolve lineages (defined as a monophyletic group supported by bootstrap values). A subset of *Arabidopsis* sequences representing all lineages of a certain TE type were used as queries in tBlastn searches against the *B. oleracea* database to identify their homologs in this species. The copy number of each type of TE was estimated based on the query length and number of hits (see *Materials and Methods*).

In both species class 1 elements are more abundant than class 2 elements (Figure 4.1a and Table 4.1). LTR-retrotransposons are the predominant TE type, with *gypsy*-like elements more abundant than c*opia*-like elements in *Arabidopsis* whereas *B. oleracea* contains more *copia*-like elements than *gyps* -like elements. The copy number of LINEs is comparable to that of LTR-retrotransposons. The most abundant type of class 2 TE in *B. oleracea* are CACTA-like elements, whereas MULEs are present at higher copy number in *Arabidopsis* than other class 2 elements.

Comparison of the TE abundance in these two species showed that all TE types are more numerous in *B. oleracea* (Figure 4.1a and Table 4.1). This result is not surprising considering the genome of *B. oleracea* is approximately five times larger than that of *Arabidopsis*. For this reason TE densities (copy number per Mb of genomic sequence) were calculated to identify any TEs that have copy number higher than expected for proportional increase. As shown in Figure 4.1b,

Figure 4.1. Comparison of (A) the abundance and (B) the density (copies per Mb) of different types of TEs in *Arabidopsis* and *B. oleracea*. Values from *Arabidopsis* are shown in red and those from *B. oleracea* are shown in green.

A


Copy Number
6000
5000
4000
3000
2000
1000
0
93
420
214
1210
39
148
29
1043
109
2475
275
5605
407
5428
238
3310
hAT
MULEs
PIF
Pong
CACTA
Copia
Gypsy
LINEs
A.t.
B.o.

B

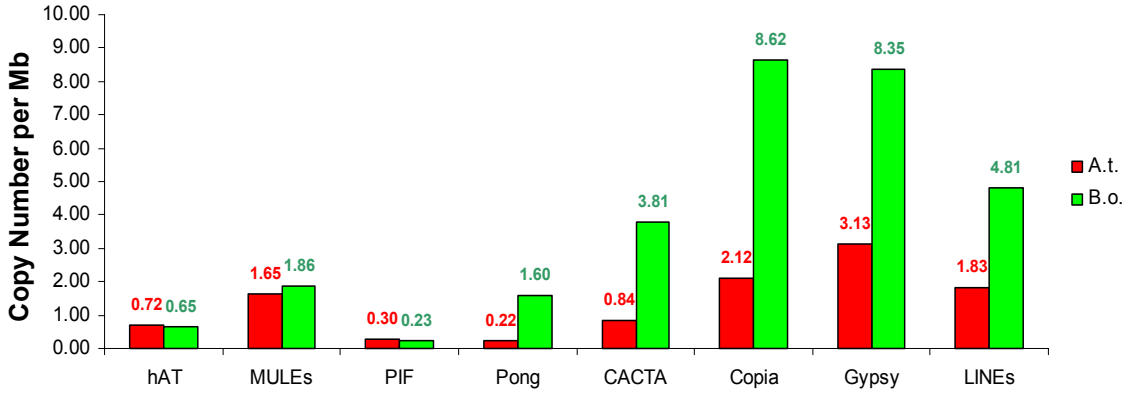

Copy Number per Mb
10.00
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
0.72
0.65
1.65
1.86
0.30
0.23
0.22
1.60
0.84
3.81
2.12
8.62
3.13
8.35
1.83
4.81
hAT
MULEs
PIF
Pong
CACTA
Copia
Gypsy
LINEs
A.t.
B.o.

Table 4.1. TE abundance in *Arabidopsis* and *B. oleracea* *

Element	Copy No. in *A.t.*	No. per Mb	Total Length (Mb)	% of Genome	Copy No. in *B.o.*	No. per Mb	Total Length (Mb)	% of Genome
hAT	93	0.7	0.5	0.4	420	0.6	2.1	0.4
MULEs	214	1.6	1.1	0.8	1210	1.9	6.1	1.0
PIF	39	0.3	0.2	0.2	148	0.2	0.7	0.1
Pong	29	0.2	0.2	0.1	1043	1.6	5.7	1.0
CACTA	109	0.8	1.0	0.8	2475	3.8	22.3	3.7
Copia	275	2.1	1.7	1.3	5605	8.6	33.6	5.6
Gypsy	407	3.1	2.4	1.9	5428	8.4	32.6	5.4
LINEs	238	1.8	1.3	1.0	3310	5.1	18.2	3.0
Total	~1,400	10.8	8.3	6.4	~19,600	30.2	121.3	20.2

* Based on elements with significant coding capacities.

the densities of *hAT*, MULEs and *PIF*-like elements are similar in both genomes, whereas the densities of *Pong*-like, CACTA-like and all class 1 elements are significantly higher in *B. oleracea*.

Comparative phylogenetic analyses of TEs in *Arabidopsis* and *B. oleracea*. The evolution of TE coding regions is analogous to that of gene families. Phylogenetic analysis of gene family members from the genome of a certain species permits one to infer the number of duplication events that have occurred and the relative timing of these events. In contrast, comparison of gene family members from different species addresses questions regarding the timing of gene duplication relative to speciation and whether some gene family members were further duplicated or lost in a certain species. The evolution of TEs, like gene families, is characterized by iterated processes of duplication, divergence, and duplication. The transposition and duplication of TEs enable them to evolve at a much higher rate than host genes, while the lack of purifying selection pressure allows them to diverge more rapidly. Nevertheless, the similarity

between TEs and gene families regarding duplication and divergence means that similar strategies can be used to study the evolution of TEs. Intraspecies and interspecies comparative analyses have recently been successfully adopted to study plant TE superfamilies. For example, phylogenetic analysis of the *PIF*/*Pong* superfamily members in rice revealed multiple lineages in a single genome. Interspecies analysis of this superfamily in plant genomes determined that multiple lineages existed prior to the divergence of monocots and dicots and revealed the extensive diversification of these lineages during plant evolution (Zhang et al. 2003).

To address the questions regarding the role of TE amplification in the recent genome size increase of *B. oleracea* the cause of the low TE content in *Arabidopsis*, the phylogenies of all major types of TE in the two species were determined. This was done by comparison of sequences from both genomes by CLUSTALW multiple alignments and generating phylogenetic trees.

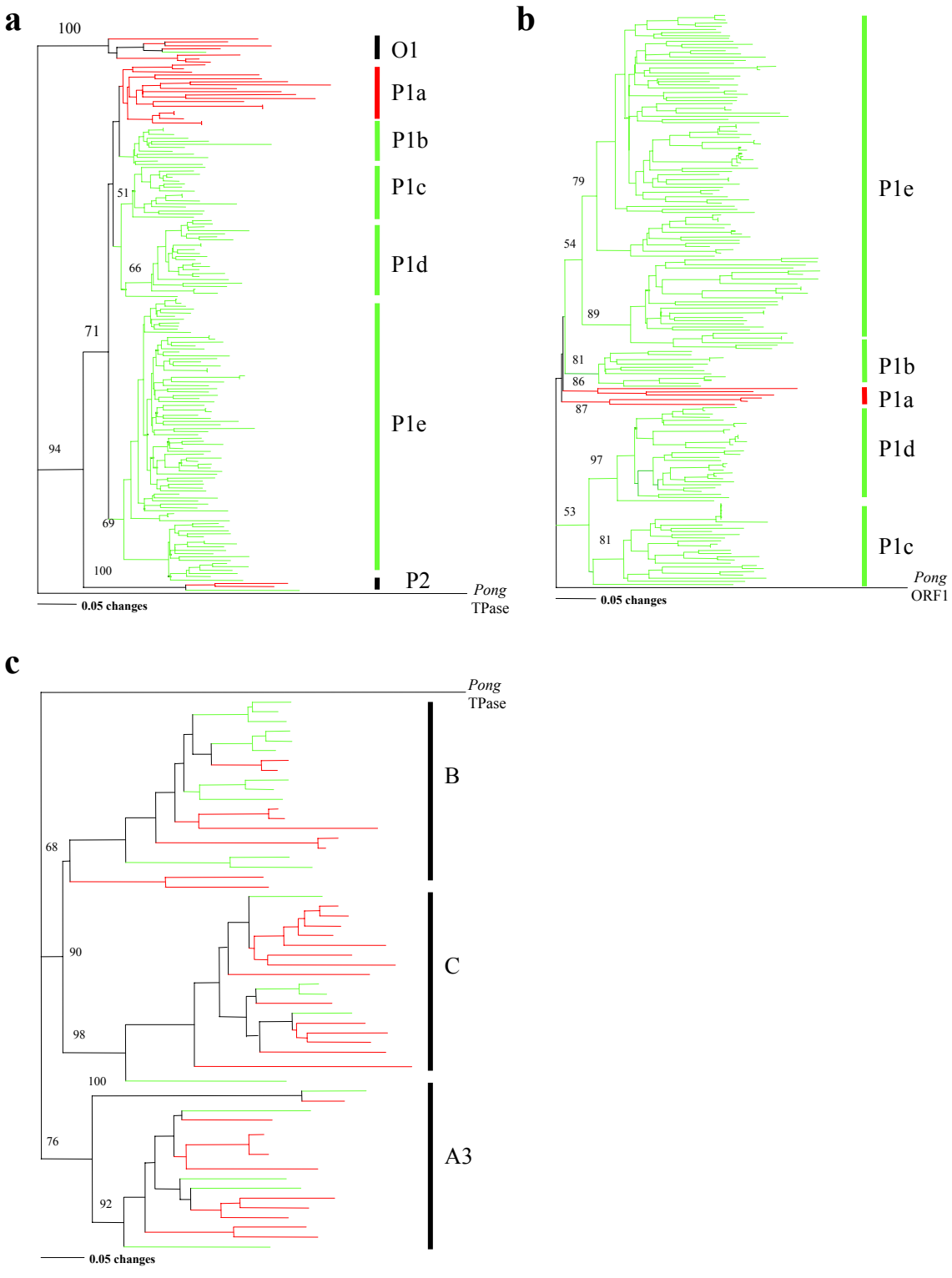
***Pong*-like and *PIF*-like elements.** Two active DNA transposons, the maize *PIF* element and the rice *Pong* element, served as the founding members of a recently discovered superfamily of DNA transposons, the *PIF*/*Pong* superfamily (Zhang et al. 2001; Jiang et al. 2003). Plant members of this superfamily group into two clades: *PIF*-like and *Pong*-like. In addition to their TPases, *Pong*-like elements encode another ORF (called ORF1) whose presence in all *Pong*-like elements suggests that it encodes a function necessary for transposition. In contrast, a second conserved ORF was not found in *PIF*-like elements. A recent

survey of TPase-encoding *PIF*- and *Pong*-like elements in rice showed that they are present at moderate copy number (~120 and ~80, respectively).

Pong-like elements are present at much higher density in *B. oleracea* (1.6 copies per Mb) than in *Arabidopsis* (0.2 copy per Mb), whereas the density of *PIF*-like elements is similar in the two species (0.2 and 0.3 copy per Mb, respectively). tBlastn searches using the catalytic domain (~120 aa.) of the rice *Pong* TPase identified 27 hits from *Arabidopsis* and 139 from *B. oleracea* that contained the entire query region. A phylogenetic tree generated from the CLUSTALW multiple alignment resolved three lineages that corresponded to three previously described dicot *Pong* lineages (O1, P1 and P2; Figure 4.2a). All three lineages were found in both species, indicating that they were present in the last common ancestor. Of the three lineages, P1 includes the majority of sequences from *Arabidopsis* (18 of 27) and nearly all from *B. oleracea* (137 of 139). Within P1, *B. oleracea* sequences clustered into four large, species-specific groups (P1b-P1e), indicating that several lineages of *Pong*-like elements have undergone explosive amplification since their last common ancestor. The average pairwise nucleotide identity among *B. oleracea* elements in P1 is 79.2%.

In addition to the catalytic domain, approximately 1,000 *B. oleracea* reads were homologous to other regions of the *Pong* TPase. It is unlikely that the large number of *Pong*-like TPase sequences identified by tBlastn searches isan artifact of database bias (e.g., a few elements were sequences multiple times) as only four pairs of sequences used to generate the phylogenetic tree in Figure 4.2a are identical. In order to obtain additional evidence supporting the abundance and

Figure 4.2. Phylogeny of (A) *Pong*-like TPase in *Arabidopsis* and *B. oleracea*, (B) *Pong*-like ORF1s in *Arabidopsis* and *B. oleracea* and (C) *PIF*-like TPases in *Arabidopsis* and *B. oleracea*. *Arabidopsis* sequences are shown in red and *B. oleracea* sequences are shown in green. These phylogenetic trees were generated using the neighbor-joining method and rooted with the TPase of the maize *PIF* elements and the ORF1 of the rice *Pong* element, respectively. Bootstrap values were calculated from 500 replicates.


a
100
51
66
71
94
69
100
O1
P1a
P1b
P1c
P1d
P1e
P2
Pong
TPase
0.05 changes
b
79
54
89
81
86
87
97
53
81
P1e
P1b
P1a
P1d
P1c
Pong
ORF1
0.05 changes
c
Pong
TPase
68
90
98
100
76
92
B
C
A3
0.05 changes

explosive amplification of *Pong*-like elements, the abundance and phylogeny of *Pong* ORF1s were also determined. Since *Pong*-like elements encode both ORF1 and TPase (Jiang et al. 2003), the abundance of the two ORFs should be similar if element transposition was responsible. tBlastn searches using the rice *Pong* ORF1 (340 aa.) as query identified over 700 ORF1 homologs in *B. oleracea* (E value < e-10), whereas only 21 (E value < e-5) were found in *Arabidopsis*.

All 21 *Pong* ORF1s in *Arabidopsis* were found to be located within 1-2 kb of a *Pong*-like TPase, indicating that each pair of ORF1 and TPase are encoded by the same element. Because the *B. oleracea* database is highly fragmented (average read length is ~650 bp), it was not possible to directly correlate the ORF1 and TPase encoded by the same element. However, since the *B. oleracea* sequences were generated from both ends of ~2.5-kb genomic clones, the presence of a pair of ORF1 and TPase on the ends of the same clone would indicate that they are encoded by the same elements. Comparison of the clone names of TPase and ORF1 hits showed that over one fourth (>200) ORF1 hits are located on the same clone as a TPase hit.

Our previous study showed that the phylogenies of *Pong* ORF1s (determined based on a 110-aa. region) and TPases from rice were highly consistent (Zhang et al. 2003). The phylogeny of *Pong* ORF1 in *Arabidopsis* and *B. oleracea* was determined based on the same region as was used in rice and compared to that of TPase (Figure 4.2b). Although ORF1s were significantly more divergent than TPases (average pairwise nucleotide identity was 59.5%

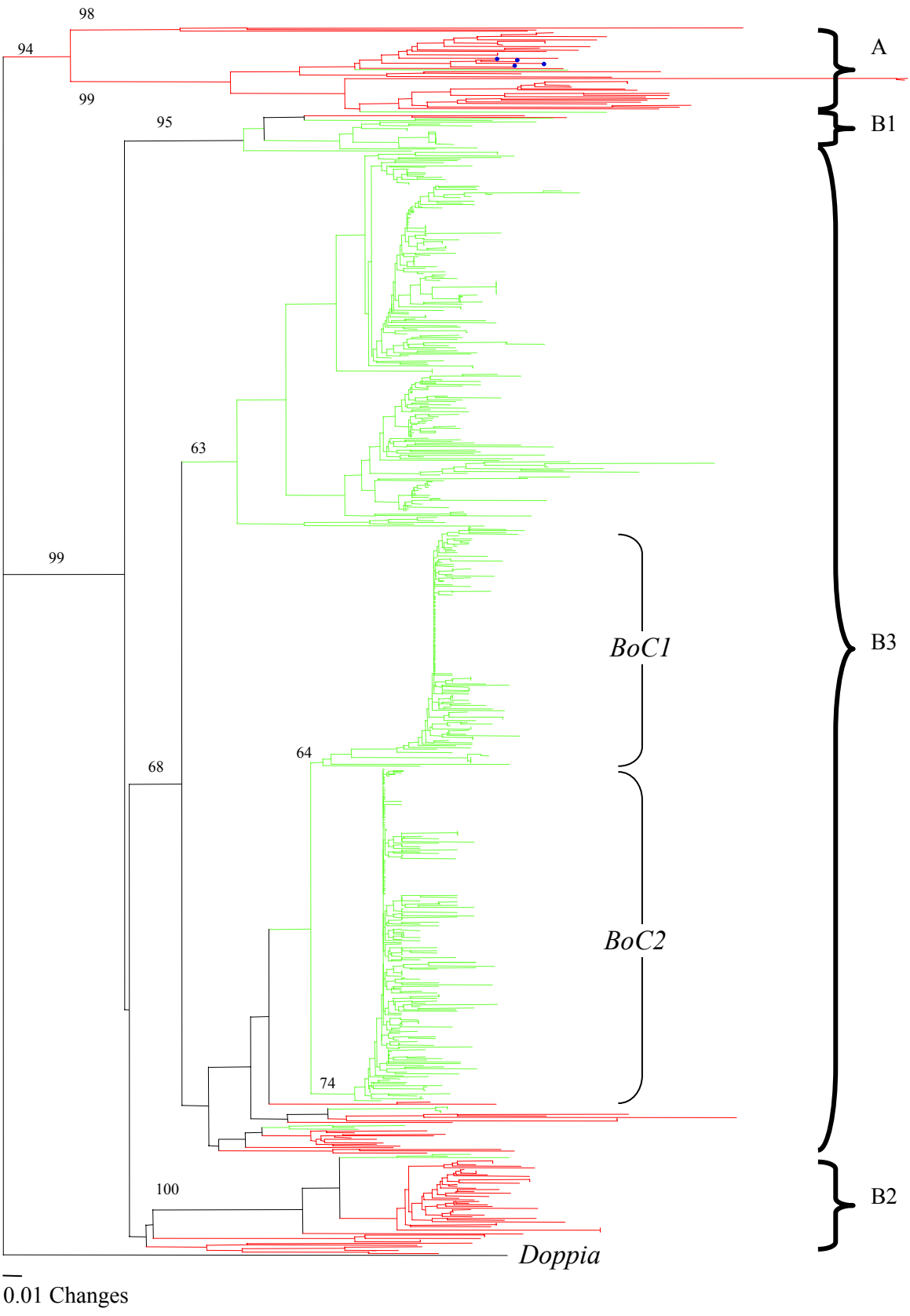
compared to 79.2% for TPases), their phylogeny was highly consistent. That is, both ORF1 and TPase sequences in *B. oleracea* clustered into four large species-specific groups. In addition, based on the pairwise association between ORF1s and TPases, each ORF1 group could be linked to a TPase group (i.e., encoded by the same group of *Pong*-like elements), and corresponding groups exhibited very similar topology as well as similar numbers of sequences.

In contrast to *Pong*-like elements, the density of *PIF*-like elements is not significantly different in *Arabidopsis* and *B. oleracea*. tBlastn searches using the catalytic domain of the maize *PIF* TPase (120 aa.) identified 35 hits from *Arabidopsis* and 21 from the current *B. oleracea* database that contained the entire query region. These sequences were compared by CLUSTALW multiple alignment and used to generate a phylogenetic tree (Figure 4.2c). *PIF*-like TPases in *Arabidopsis* and *B. oleracea* clustered into three lineages (A3, B and C); all included sequences from both species and, within each lineage, there were no smaller, species-specific clusters. In both species, *PIF*-like TPases are significantly more divergent than *Pong*-like TPases (longer branch in Figure 4.2c than Figure 4.2a). Therefore, all *PIF* lineages were present in the last common ancestor of *Arabidopsis* and *B. oleracea*, and none has significantly increased their copy number in either species.

CACTA-like elements. Previous studies have identified four families of CACTA-like elements in *Arabidopsis* (named *Atenspm1-4*) (Kapitonov and Jurka 1999), of which one family (*Atenspm1*) was found to be actively transposing in a *ddm1*

backgroup (Miura et al. 2001). A CACTA-like family has recently been described in *B. rapa* (Suzuki et al. 1999). The putative TPases of *Atenspm1* (889 aa.) and the *B. rapa* CACTA-like element (703 aa.) are ~38% identical and the highest level of homology was found in a 100-aa. region (77% identical) corresponding to positions 272-371 and 301-400 in the TPases of *Atenspm1* and the *B. rapa* element, respectively. This region was used to as query to identify additional CACTA-like TPases in *Arabidopsis* and *B. oleracea* through tBlastn searches. One hundred and twenty one and 540 hits that contained the entire query region were identified in *Arabidopsis* and *B. oleracea*, respectively. A CLUSTALW multiple alignment was constructed from these sequences and used to generated a phylogenetic tree (Figure 4.3). *Arabidopsis* and *B. oleracea* elements clustered into two major clades, A and B. Clade A contains 40 of the 121 *Arabidopsis* sequences (including members of the *Atenspm1* family) but only two of the 540 *B. oleracea* sequences. Clade B consists of three lineages (B1 through B3), all of which are present in both *Arabidopsis* and *B. oleracea*. Lineage B1 is present at low copy number in both species, while lineages B2 resembles clade A in that it is relatively abundant in *Arabidopsis* (50 sequences) but scarce in *B. oleracea* (three sequences). The vast majority of *B. oleracea* sequences (500) comprise the B3 lineage and cluster into three large, species-specific groups. Two groups, named *BoC1* and *BoC2*, include many highly similar sequences. Forty two of the 118 *BoC1* sequences were identical, and 46 of the 181 *BoC2* sequences were identical. Such high intrafamily sequence identity indicates that these two families have very recently amplified to high copy number.

Figure 4.3. Phylogeny of CACTA-like elementws in *Arabidopsis* (shown in red) and *B. oleracea* (shown in green). This phylogenetic tree was generated using the neighbor-joining method and rooted with the TPase of the maize *Doppia* element. Sequences from four described *Arabidopsis* elements (*CAC*1 though *CAC4*) are indicated by filled dots and their names shown to the right. Bootstrap values were calculated from 250 replicates.


98
94
99
95
63
99
68
64
74
100
A
B1
B3
B2
BoC1
BoC2
Doppia
0.01 Changes

Two approaches were undertaken to provide additional evidence for the abundance of *BoC1* and *BoC2* in *B. oleracea*. First, a 360-bp fragment (including the 100-aa. region used to deduce their phylogeny) was amplified from *BoC1* and *BoC2* by PCR and used as probe to screen a *B. oleracea* BAC library. Consistent with their high abundance, numerous BAC clones hybridized to the probe (not shown). Second, the ends of *BoC1* and *BoC2* elements were derived and their flanking sequences were examined. To this end, 100 sequences flanking *BoC1* elements were inspected and all were found to be unique. Taken together, these results indicate that *BoC1* and *BoC2* elements are present in *B. oleracea* at high copy number.

MULEs and *hAT*-like elements. A previous survey of *Arabidopsis* MULEs where only 17.2 Mb of genomic sequences were queried identified nine families including 72 elements, of which twelve contained TPase (Yu et al. 2000). In addition, a family (called *AtMu1*) was found to be actively transposing in a *ddm1* background (Singer et al. 2001). *hAT*-like elements have not been systematically analyzed in *Arabidopsis*. However, a previous survey for *hAT*-like elements in eukaryotes indicated that dozens of TPase-encoding elements existed in *Arabidopsis* (Rubin et al. 2001). Additional TPases of MULEs and *hAT*-like elements in *Arabidopsis* and *B. oleracea* were identified when the most conserved region of their respective TPases (the catalytic domain for MULEs and the dimerization domain for *hAT*-like elements) were used as queries in tBlastn searches. These sequences were then used to deduce the phylogenies of these

two superfamilies (Figure 4.4 and 4.5). Both MULEs and *hAT*-like elements are represented by multiple small lineages (each containing few elements) that diverged prior to the divergence of *Arabidopsis* and *B. oleracea*. No significant amplification was observed for either superfamily.

***Copia*-like LTR-retroelements.** It has been estimated that the *Arabidopsis* genome harbors approximately 300 *Arabidopsis copia*-like elements (Terol et al. 2001). A previous study of a subset of 25 elements resolved six lineages (*Copia* I – VI) (Terol et al. 2001). Comparison of the *pol* coding region of these 25 elements revealed that the most conserved region was a 156-aa. segment in the RT domain (not shown). tBlastn searches using this segment as query identified 268 and 638 hits from *Arabidopsis* and *B. oleracea*, respectively, that contained the entire query sequence. A phylogenetic tree generated from the CLUSTALW multiple alignment resolved two major clades (A and B, see Figure 4.6). Clade A included all six previously reported lineages as well as five lineages identified in this study. Clade B did not include any previously described *copia*-like elements from either species and can be divided into two lineages. Nineteen complete elements belonging to newly identified lineages were identified from *Arabidopsis* and listed in Table 4.2. Of the thirteen *copia*-like lineages, one (*Copia* XIII) was only found in *B. oleracea*. The remaining twelve lineages are present in both species and all are present at a higher density in *B. oleracea* than in *Arabidopsis*. Two lineages, *Copia* IX and *Copia* XI, showed the largest difference in density between *Arabidopsis* and *B. oleracea*. They are present in *B. oleracea* at ~2

Figure 4.4. Phylogeny of MULEs in *Arabidopsis* (shown in red) and *B. oleracea* (shown in green). This phylogenetic tree was generated using the neighbor-joining method and rooted with the TPase of the maize *Mutator* element (MURA). Sequences from described elements are indicated by filled dots and their names shown to the right. Arrows indicate two families of *B. oleracea* LINEs with multiple identical or near-identical copies (see text). Bootstrap values were calculated from 250 replicates.

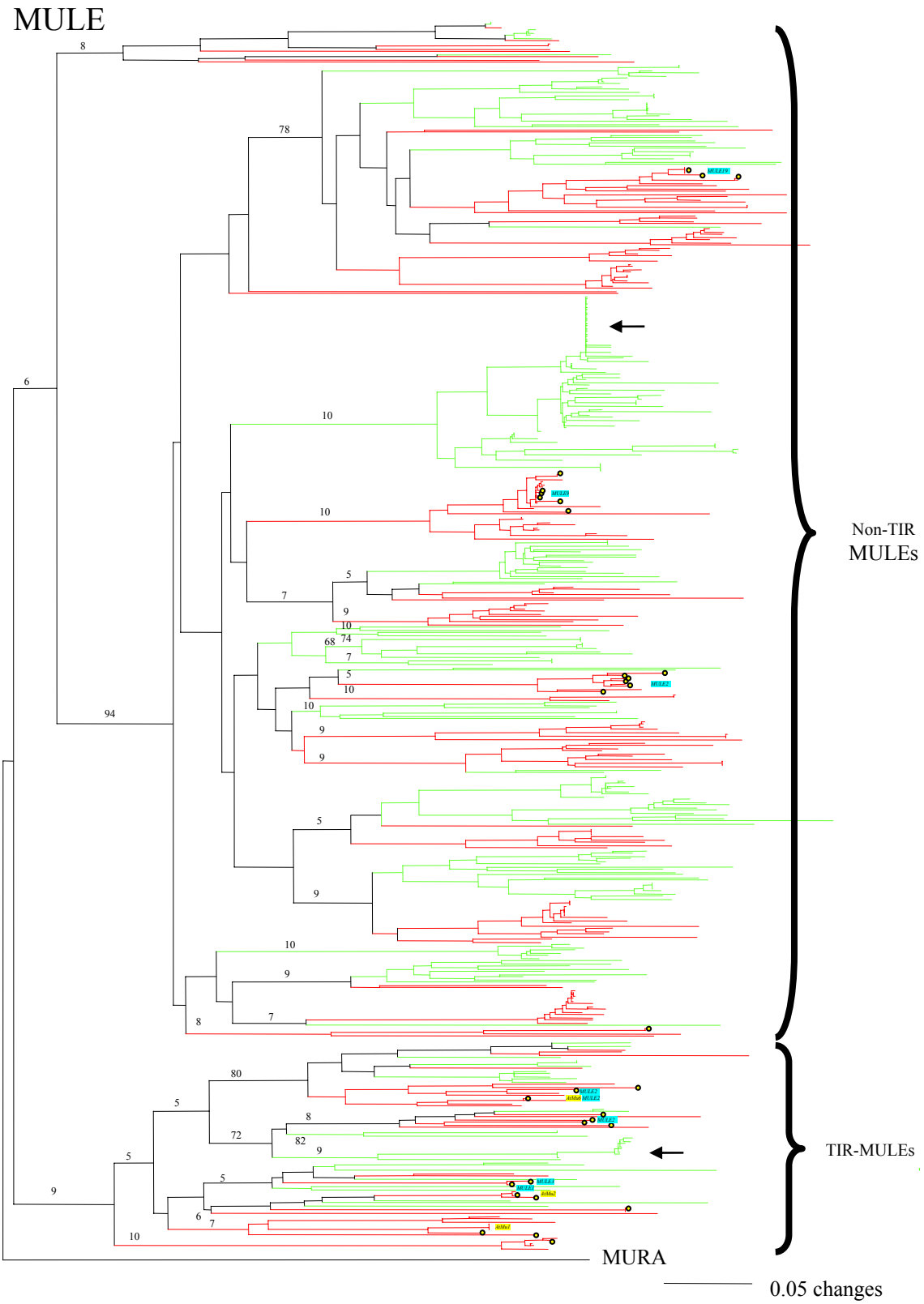

MULE
Non-TIR
MULEs
TIR-MULEs
MURA
0.05 changes

Figure 4.5. Phylogeny of *hAT*-like elementws in *Arabidopsis* (shown in red) and *B. oleracea* (shown in green). This phylogenetic tree was generated using the neighbor-joining method and rooted with the TPase of the *Hermes* element from house fly. *Tag1*, *Tag2* and other described *Arabidopsis* elements are indicated by filled dots. Bootstrap values were calculated from 250 replicates.

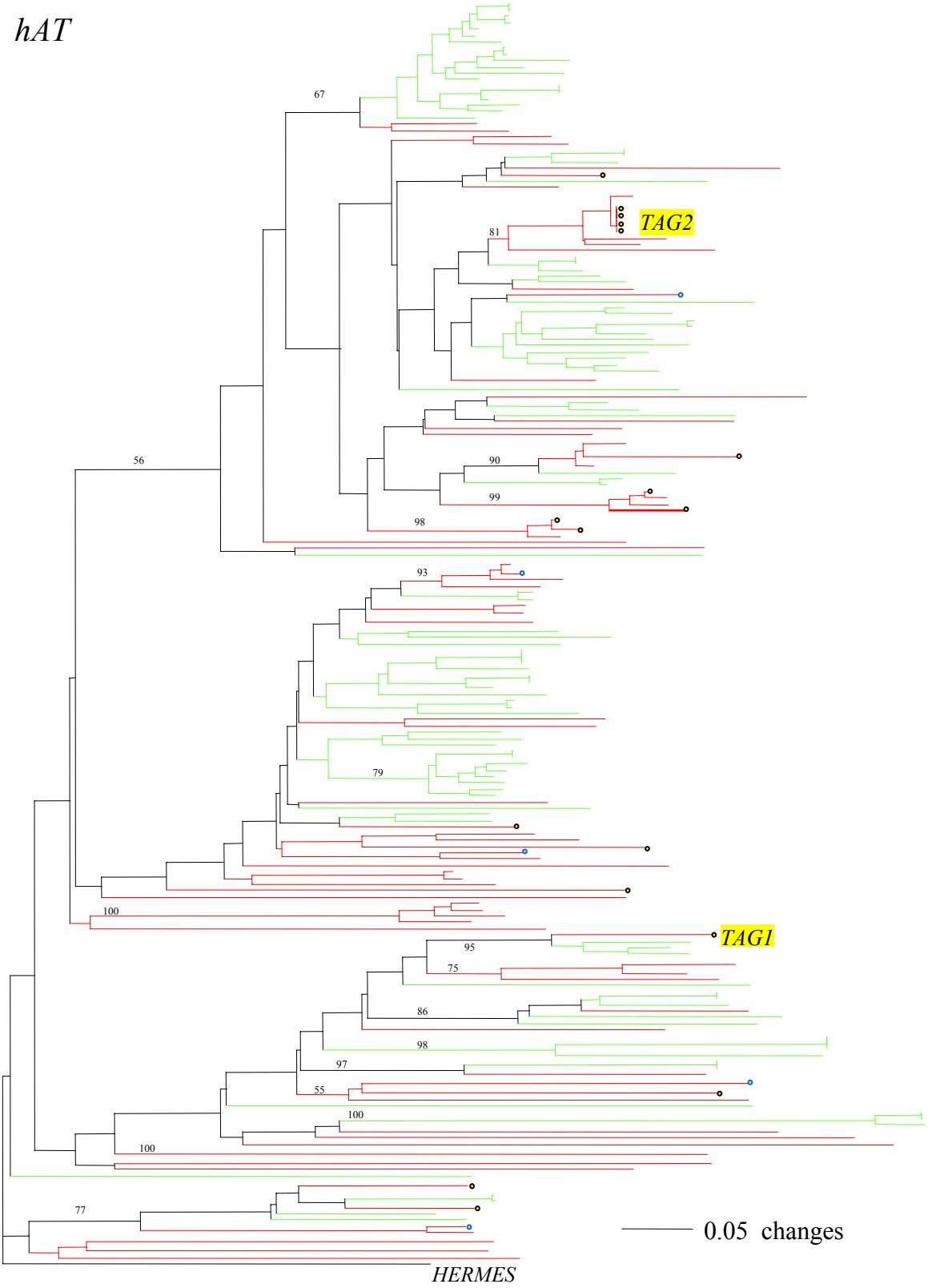

hAT
67
TAG2
81
56
90
99
98
93
79
100
TAG1
95
75
86
98
97
55
100
100
77
HERMES
0.05 changes

Figure 4.6. Phylogeny of *copia*-like LTR retroelements in *Arabidopsis* (shown in red) and *B. oleracea* (shown in green). This phylogenetic tree was generated using the neighbor-joining method and rooted with *Ty1*. Sequences from described elements are indicated by filled dots and their names shown to the right. Described lineages are labeled in blue and new lineages identified in this study are labeled in brown. Arrows indicate two recently amplified families of *B. oleracea copia*-like elements (*BoCP_IXa* and *BoCP_IXc*). Bootstrap values were calculated from 250 replicates.

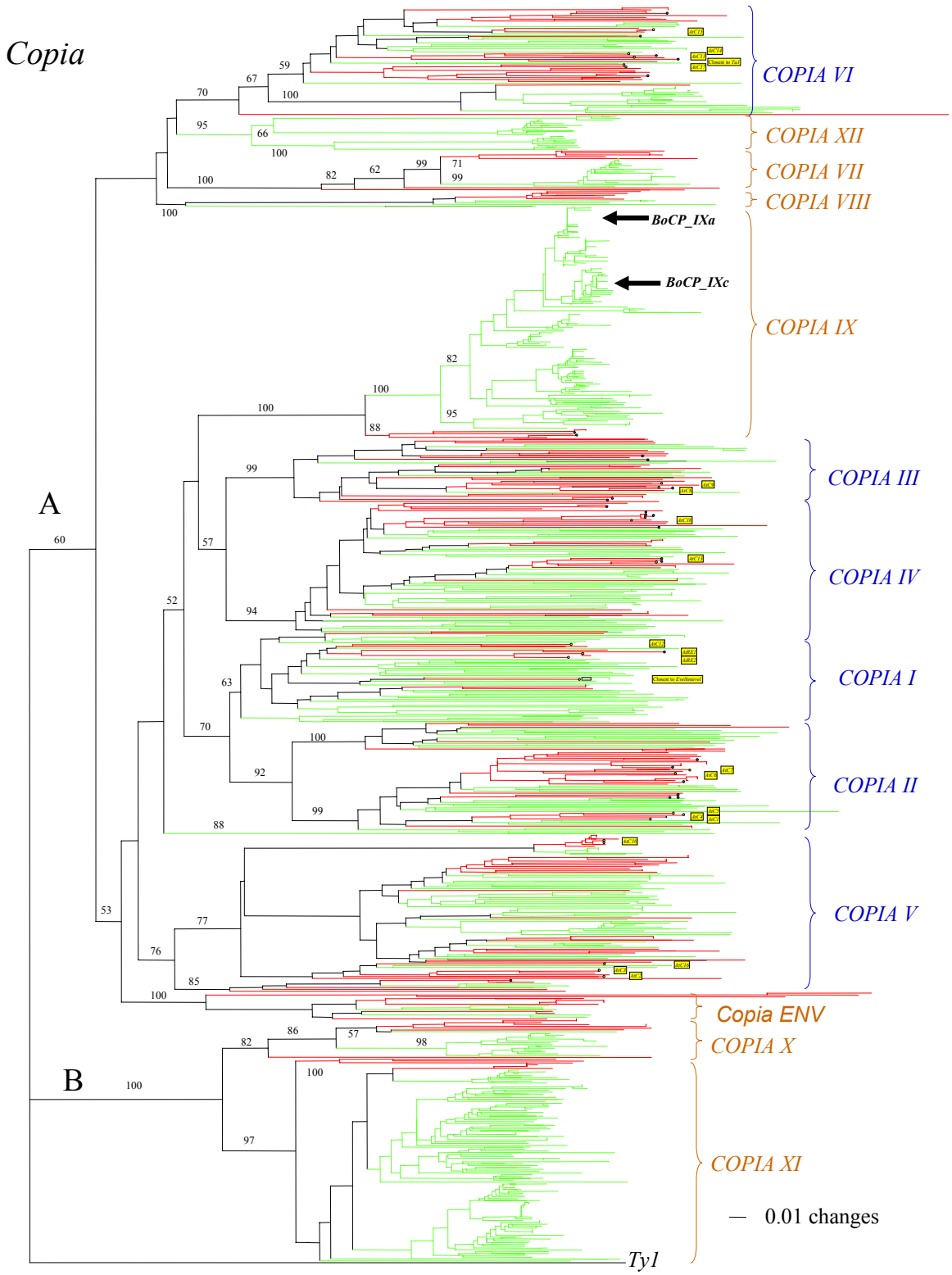

Copia
COPIA VI
COPIA XII
COPIA VII
COPIA VIII
BoCP_IXa
BoCP_IXc
COPIA IX
COPIA III
COPIA IV
COPIA I
COPIA II
COPIA V
Copia ENV
COPIA X
COPIA XI
A
B
0.01 changes
Ty1

Table 4.2. Summery of newly identified full-length *copia*-like LTR-retroelements.

Lineage	Clone Name	Acc.No	From	To	Length	LTR length	LTR identity	TSD
Copia VIII	T7M24	AF077408	2707	9244	6538	586	97%	ACAGT
	T15D2	AP002054	61202	67838	6637	589	96%	TTGGA
Copia ENV	F23N14_1	AL138638	7767	16478	8712	570/563	96%	AACAT
	F23N14_2	AL138638	58903	68055	9153	509	96%	TCTTC
	MVA11	AP001311	43913	58881	14969	558	94%	GGTTG
	T22A15	AC021666	26443	35521	9079	569	96%	ACATC
	F1D9_2	AP002460	66980	75180	8201	531	96%	GGTGG
	T2N12	AB073164	38526	46259	7734	609	96%	AGAGC
	MQD19	AB026651	51655	60737	9083	549	99%	GATAT
Copia X	T32G9	AC079605	94848	97735	2888	322/330	95%	ATTAG
	F14D7	AC021198	11188	16871	5684	378/324	93%	ATGGA
	MJI6	AP002043	25137	31083	5947	330	97%	TTATC
	T17A2	AF160183	30519	36948	6430	256/267	90%	ATAAT
Copia XI	F3L12	AC007178	10339	15408	5070	182	97%	ATAAC
	F7J8	AL137189	60189	66616	6428	337	99%	GATGC
	T13B17	AP002459	7605	13947	6343	336	99%	TAATT
	T14C9	AC006601	85787	92117	6330	336	98%	TTTCT
	T15F17	AF262042	27987	34417	6430	337	99%	ACTTT
	T17A11	AC006194	36362	41300	4938	322	96%	AGTTG

copies per Mb (~1,200- 1,400 copies in the whole genome), but each with ~0.08 copy per Mb in *Arabidopsis* (total copy number for each was fewer than ten copies). Importantly, *B. oleracea* sequences in these two lineages clustered into species-specific groups with short branch length (Figure 4.6), indicating that they have amplified in *B. oleracea* after its divergence from *Arabidopsis*. Three small clusters of *B. oleracea* sequences (~25 sequences in each cluster) are highly similar (>98% identical), indicating that three *copia*-like families may have been very recently active in *B. oleracea* (named *BoCP_IXa*, *b* and *c*; arrows in Figure 4.6).

***Gypsy*-like LTR-retroelements.** *Gypsy*-like elements have been divided into two clades: metaviruses and errantiviruses (Pringle 1998; Hull 1999). Both clades

populate plant genomes, and each clade had split into at least two subclades before the divergence of monocots and dicots ~200 Mya. Approximately 60 *gypsy*-like elements have been described in *Arabidopsis* and grouped into eight lineages (Marin and Llorens 2000): *Tma*, *Legolas*, *Gimli* and *Gloin* are metaviruses, while *Athila*, *Little Athila*, *Tat*, *Tft1* and *Tft2* are errantiviruses. The most conserved coding region in the previously described *Arabidopsis gypsy*-like elements was found to be a 156-aa. segment in the RT domain (not shown). tBlastn searches using this segment as query identified 286 and 714 sequences, respectively, from *Arabidopsis* and *B. oleracea* that contained the entire query region. The phylogeny of *gypsy*-like elements in *Arabidopsis* and *B. oleracea* was determined by constructing a CLUSTALW multiple alignment from these sequences and generating a phylogenetic tree (Figure 4.7). Two major clades were resolved (E and M) that corresponded to the errantivirus and metavirus clades of *gypsy*-like elements. Clade E can be further divided into two subclades (E1 and E2) and includes eleven lineages, four of which were previously reported. Clade B consists of three subclades (M1 through M3) and twelve lineages, of which four were previously reported. Twelve complete elements belonging to newly identified lineages were identified from *Arabidopsis* and listed in Table 4.3. Nine of the 23 *gypsy*-like lineages are present in both species, of which one lineage (*Tat*) has similar copy number in *Arabidopsis* and *B. oleracea* (~15 copies), whereas eight are present at much higher density in *B. oleracea* than in *Arabidopsis* (2-120 fold more copies per Mb). Of the remaining fourteen lineages, five included only *Arabidopsis* sequences whereas nine are specific to

Figure 4.7. Phylogeny of *gypsy*-like LTR retroelements in *Arabidopsis* (shown in red) and *B. oleracea* (shown in green). This phylogenetic tree was generated using the neighbor-joining method and rooted with *Ty3*. Sequences from described elements are indicated by filled dots. Described lineages are labeled in blue and new lineages identified in this study are labeled in brown. Bootstrap values were calculated from 250 replicates.

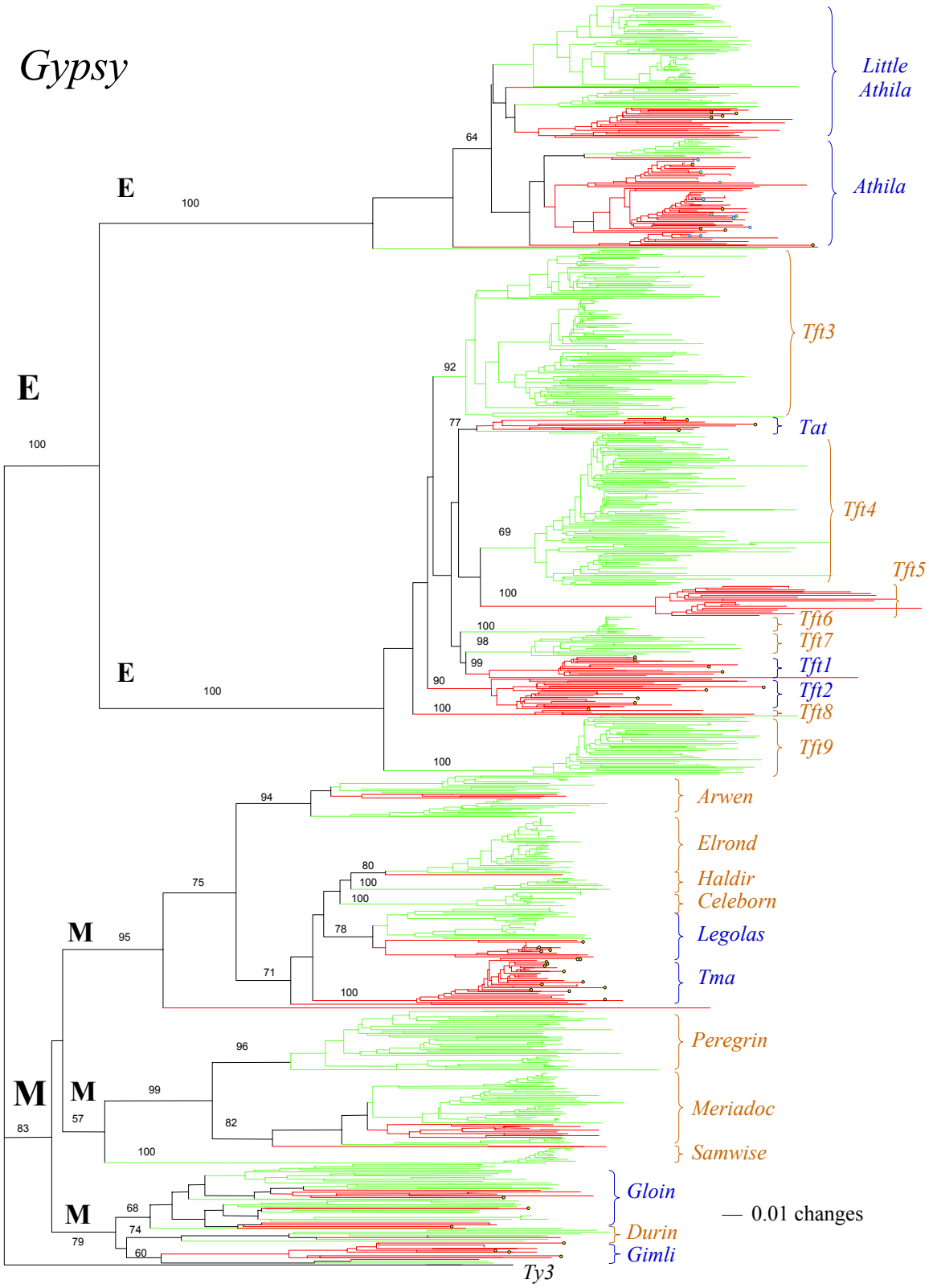

Gypsy
E
E
E
M
M
M
M
Little Athila
Athila
Tft3
Tat
Tft4
Tft5
Tft6
Tft7
Tft1
Tft2
Tft8
Tft9
Arwen
Elrond
Haldir
Celeborn
Legolas
Tma
Peregrin
Meriadoc
Samwise
Gloin
Durin
Gimli
Ty3
0.01 changes

Table 4.3. Summery of newly identified full-length *gypsy*-like LTR-retroelements.

Lineage	Clone Name	Acc. No.	From	To	Length	LTR length	LTR Identity	TSD
Tft5	F6H5_2	AP002035	79237	91078	11842	803	97%	GAGTG
Tft5	F1M23_2	AP002033	40040	45524	5485	803	97%	CATTT
Tft5	F3F24	AC018632	21921	36793	14873	762	94%	TGTTT
Tft5	F23N14_2	AL138638	42004	52703	10700	1077	98%	?
Tft8	F2G19	AC083835	87208	97212	10005	415	99%	TTAAG
Elrond	T5E15_2	AC019013	15378	23967	8590	1795/1702	97%	ATACT
Meriadoc	F6H5_1	AP002035	3309	9689	6381	702	96%	GGCTC
	MYM9	AP000377	11377	17581	6205	822	96%	TATAT
	F2H10	AC026757	53924	61592	7669	879/859	96%	TCTTA
	F19M13_1	AC006267	77876	83526	5651	861/817	96%	CAACC
	T7B9_1	AP002067	25755	32137	6383	843/826	97%	TCAGG
	T14A11_2	AC012327	49036	56075	7040	1156	94%	CCTTC

B. oleracea. Note that all fourteen species-specific lineages have shorter branch length compared to the nine lineages that are shared by the two species, indicating that they emerged in *Arabidopsis* or *B. oleracea* after the divergence from a common ancestor.

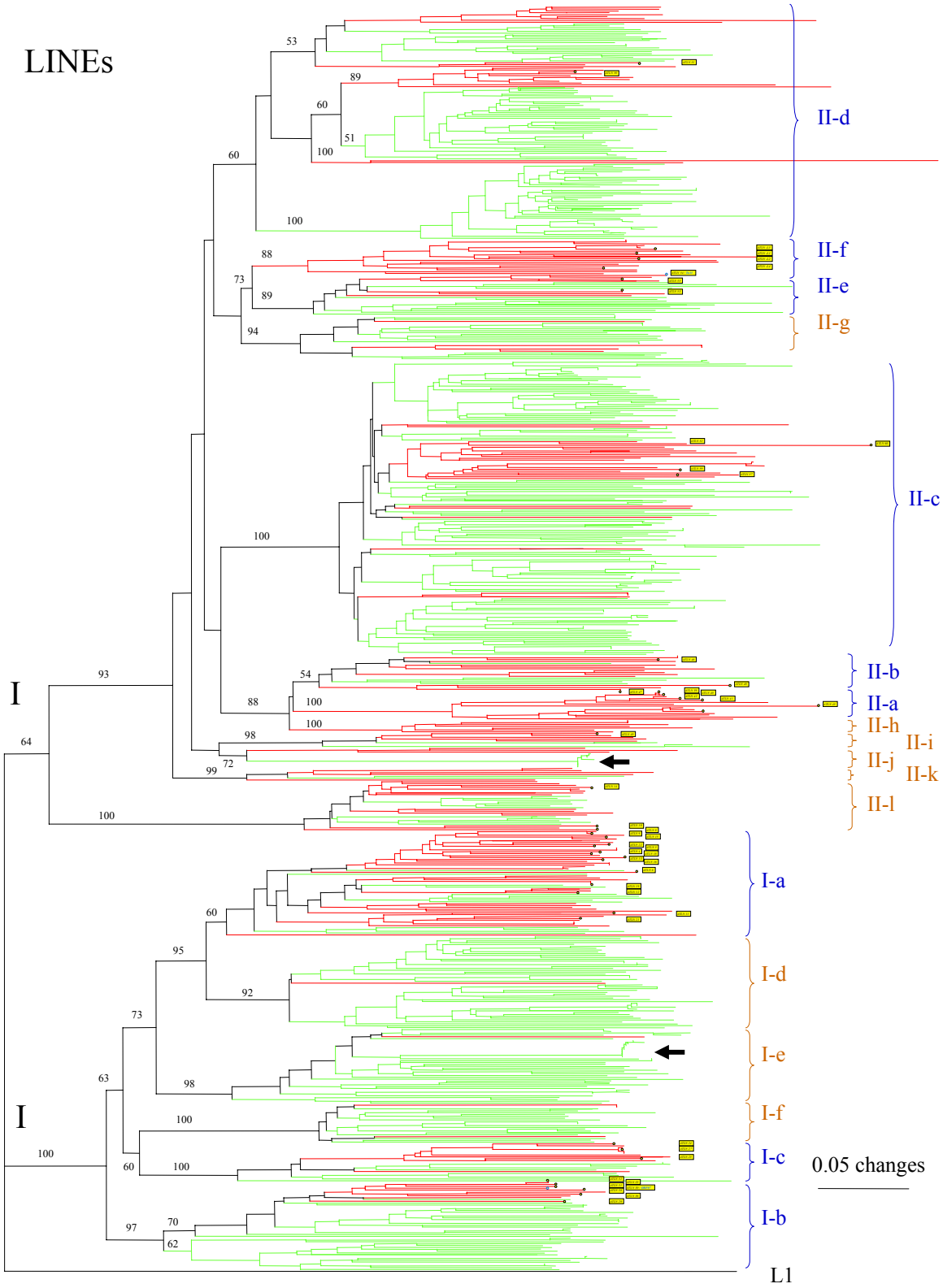
Long interspersed nuclear elements (LINEs). A recent survey of *Arabidopsis* LINEs identified 219 elements in this species, and phylogenetic analysis of a subset of 62 elements defined two clades (I and II) that were comprised of three (I-a – I-c) and six (II-a – II-f) lineages, respectively (Noma et al. 2000). The most conserved coding region in the described *Arabidopsis* LINEs was found to be a 151-aa. region in their RT domain (not shown). Using this region as query in tBlastn searches, 238 and 498 sequences containing the entire query region were identified from *Arabidopsis* and *B. oleracea*, respectively. The phylogeny of LINEs in *Arabidopsis* and *B. oleracea* was determined based on these sequences (Figure 4.8). Two clades (I and II) were resolved and further divided

into eighteen lineages, including nine newly identified lineages (I-d – I-f and II-g – II-l). Three lineages (II-a, f, h) only include *Arabidopsis* sequences. The remaining fifteen lineages were found in both *Arabidopsis* and *B. oleracea*, of which one (I-a) is present at higher density in *Arabidopsis*, nine (I-b – f, II-c – e and II-g) are present at higher density in *B. oleracea*, and five (II-b, II-i – l) are present at similar density in the two species. In general, LINE lineages in *Arabidopsis* and *B. oleracea* are characterized by very long branch lengths, suggesting that LINEs have been quiescent for a long period of time. There are, however, two notable exceptions. Lineage I-e includes ten *B. oleracea* sequences that are identical, and a group of nine *B. oleracea* sequences in lineage II-j are on average 99.5% identical (arrows in Figure 4.8).

Discussion

The *B. oleracea* genome was partially sequenced to aid the annotation of protein coding regions in the *Arabidopsis* genome. Here we have extended the comparison to proteins encoded by both class 1 and 2 TEs to address two questions: the role of TE amplification in the genome size expansion of *B. oleracea* and the underlying reason for the low TE content of the *Arabidopsis* genome. To do this, a set of strategies were derived to identify all major types of TE from these two species. For each TE type, sequences from *Arabidopsis* and *B. oleracea* were pooled into the same dataset and their phylogenies determined. These analyses allowed us to infer the number of lineages present in the last

Figure 4. 8. Phylogeny of LINEs in *Arabidopsis* (shown in red) and *B. oleracea* (shown in green). This phylogenetic tree was generated using the neighbor-joining method and rooted with L1. Sequences from described elements are indicated by filled dots and their names shown to the right. Described lineages are labeled in blue and new lineages identified in this study are labeled in brown. Arrows indicate two families of *B. oleracea* LINEs with multiple identical or near-identical copies (see text). Bootstrap values were calculated from 250 replicates.


LINEs
II-d
II-f
II-e
II-g
II-c
II-b
II-a
II-h
II-i
II-j
II-k
II-l
I
I-a
I-d
I-e
I-f
I-c
I-b
L1
0.05 changes

common ancestor of *Arabidopsis* and *B. oleracea* and to determine the success of different lineages by comparing their copy number in the two species as well as their ability to form new lineages. The results provide the first data regarding the genome wide behavior of TEs in *Arabidopsis* and *B. oleracea* since their divergence from a common ancestor. In addition, this study should serve as a model for future analyses because the strategies and methodologies described here can be used to extract information from any partial, fragmentary genomic databases.

The *Arabidopsis* genome is compact in large part due to its low TE content. The biological basis for its small genome size has been the subject of ongoing debate and remains unsolved. It is not known whether the low TE content is due to the lack of significant TE amplification or to the massive loss of TEs from its genome (Bennetzen and Kellogg 1997; Devos et al. 2002). The results of this study strongly support the former and provide little evidence form the latter. Overall, nearly all lineages of each type of TE are present in both *Arabidopsis* and *B. oleracea*. For example, the two species share all three *Pong*-like lineages (Figure 4.2a, b), all four CACTA-like lineages (Figure 4.3), twelve of the thirteen *copia*-like lineages (Figure 4.4), and fifteen of the eighteen LINE lineages. The data suggest that the shared lineages were present in their last common ancestor. In a few cases some lineages are in only one species. For example, one *copia*-like lineage (*Copia* XII, Figure 4.4) was only in *B. oleracea* while three LINE lineages (II-a, f, h, Figure 4.8) only included *Arabidopsis* sequences. These lineages have shorter branch lengths compared to lineages

that are common to both species, suggesting that they emerged in *Arabidopsis* or *B. oleracea* after their divergence. Species-specific lineages may have evolved from related elements or via horizontal transfer. Despite these few exceptions, the fact that the vast majority of TE lineages are retained in both species suggests that *Arabidopsis* has not lost a substantial amount of TEs since its divergence from *B. oleracea*. For these reasons, we concluded that it is unlikely that the low TE content of *Arabidopsis* is due to large scale elimination of TEs from its genome.

Data present in this study are more consistent with the view that the low TE content of *Arabidopsis* is largely due to the lack of significant amplification of any TE type in its genome. First, ancestral lineages retained in both species almost always have much lower copy number in *Arabidopsis* than in *B. oleracea*. For example, all twelve *copia*-like lineages shared by the two species are present at much lower density (two to over 70 fold) in *Arabidopsis* than in *B. oleracea*. Second, few lineages have amplified in *Arabidopsis* and for those that have, the extent of amplification has been small. For example, *Pong*-like lineage P1a (19 sequences) and CACTA-like lineages A (40 sequences) and B2 (50 sequences) are the only lineages that have amplified considerably in *Arabidopsis*. However, even for these lineages the extent of amplification is much smaller than that in *B. oleracea*, where *Pong*-like lineages P1b-e and CACTA-like lineage B3 have amplified to ~1,000 and ~2,300 copies (extrapolated to whole genome), respectively.

In contrast, both class 1 and class 2 elements have amplified in *B. oleracea* since its divergence from *Arabidopsis* where they have contributed significantly to the genome size increase. The total length of TEs in *B. oleracea* (~120 Mb or ~20% of its genome) is approximately fifteen times more than that in *Arabidopsis* (~8 Mb or 6% of its genome). Such a difference reflects the recent amplification of both class 1 and class 2 elements, such as *copia*-like elements (lineages IX and XI, Figure 4.4), *Pong*-like elements (lineages P1b-e, Figure 4.2) and CACTA-like elements (lineage B3, Figure 4.3). In addition, although some lineages present in both species are not significantly amplified in *B. oleracea*, they usually have higher copy number in *B. oleracea* than in *Arabidopsis* (e.g., all other *copia*-like lineages are more abundant in *B. oleracea*). As a result, both class 1 and class 2 elements are more numerous in *B. oleracea*. Class 1 elements, including both LTR- and non-LTR elements, can account for ~78 Mb of nuclear DNA in *B. oleracea* (~14% of its genome) but only 5-6 Mb in *Arabidopsis* (~4% of its genome). In addition, class 2 elements also make up a larger fraction of nuclear DNA in *B. oleracea* (~37 Mb or 6% of its genome) than in *Arabidopsis* (~3 Mb or 2-3% of its genome). Taken together, these results indicate that the proliferation of TEs in *B. oleracea* contributes significantly to the observed genome size difference with *Arabidopsis*.

Prior studies based on grass genomes have shown that the amplification of LTR-retroelements and in particular, a few predominant families of LTR-retroelements, is largely responsible for genome size variations (SanMiguel et al. 1996; SanMiguel et al. 1998). For example, four families of LTR-retroelements

(*Ji*, *Opie*, *Huck* and *Zeon-1*) account for 32% of the maize genome (Meyers et al. 2001), and one family of LTR-retroelements (*IRRE*) account for ~10% of the *Iris brevicaulis* genome (~10,000 Mb) (Kentner et al. 2003). These findings have led to the conclusion that plant genomes expand mostly through the massive amplification of a few LTR retrotransposon families. However, results from this study indicate that these conclusions cannot be generalized to all plant genomes. Massive amplification of a few LTR element families has not occurred in *B. oleracea* with the most abundant family having only ~140 copies (*BoCP_IXc*). Instead, the accumulation of numerous small LTR element families has played a significant role in the *B. oleracea* genome size expansion. A similar situation was found in rice (~450 Mb), where LTR elements are represented by numerous small families. However, this type of LTR element composition may not be common for all small plant genomes. For example, it has been recently reported that the genome of sorghum (~700 Mb) contains a high copy number LTR element family (*Retrosor6*, ~6,000-7,000 copies) that accounts for ~6% of the nuclear DNA (Peterson et al. 2002).

DNA elements have also amplified to high copy number and comprise a significant fraction of the *B. oleracea* genome (~6%). Most notably, two families of CACTA-like elements (*BoC1* and *BoC2*, Figure 4.3) have amplified to >500 and >800 copies, respectively, in *B. oleracea* and together account for over 2% of the total genomic DNA. High copy number CACTA-like families have been recently reported in grasses with large genomes. There are ~5,000 *Tpo1* elements in ryegrass (*Lolium perenne*, 5,000 Mb) (Langdon et al. 2003) and

~3,000 *Caspar* elements in wheat (*Triticum monococcum*, ~5,000 Mb) (Wicker et al. 2003). Here we show that CACTA-like families can also accumulate to high copy number in small dicot genomes such as *B. oleracea*, and that such high copy number could result from the explosive amplification of just a few elements.

The contribution of DNA elements to genome size has been generally considered insignificant. This view is based on the low copy number of active DNA element families, determined by approaches such as Southern blots (e.g., *Ac*/*Ds* and *En*/*Spm* families in maize have dozens of copies). However, it is likely that active DNA families are just the tip of the iceberg and the vast majority of DNA elements have not been identified in genomes like maize. Results from this study revealed a high level of sequence diversification among related DNA element families, where interfamily sequence identity, even in the most conserved coding region, is no more than 80%. Similarly, the active rice *Pong* family consists of only 5 copies, yet over 100 *Pong*-like elements exist in the rice genome and share less than 60% nucleotide identity with each other in their TPase coding region (Zhang et al. 2003). This level of sequence diversification would have prevented the detection of the vast majority of related families when the sequence of a particular family is used as Southern blot probe.

Although our study demonstrates that *Arabidopsis* and *B. oleracea* inherited and retained largely the same collection of TE lineages from their last common ancestor ~15-20 Mya, TEs have been able to attain much higher copy number in *B. oleracea* than in *Arabidopsis*. This difference may be due to two factors that are not mutually exclusive. First, *Arabidopsis*, with its high gene

density, may not be able to tolerate large scale TE amplification because of the mutagenic effects of TE insertions. In support of this view is the fact that a high percentage (~48%) of T-DNA insertions are knockouts (Szabados et al. 2002). In contrast, it was recently reported that the *B. oleracea* genome is the product of a triplication event that occurred after its divergence from *Arabidopsis* (Cavell et al. 1998; Lagercrantz 1998; Lan et al. 2000). Such an event could have produced numerous safe havens for TE insertions due to functional redundancy. This may be particularly relevant in the recent amplification of DNA transposons, such as *Pong*-like elements, which have been shown previously to target genic regions (Jiang et al. 2003; Zhang et al. 2003). Second, there is evidence that TE activity is under rigorous epigenetic control in *Arabidopsis*. Two DNA element families (*AtMu1* and *Atenspm1*) were found to be active in a *ddm1* background defective in epigenetic regulations where they were able to increase their respective copy numbers up to 20 fold in just a few generations (Singer et al. 2001; Miura et al. 2001). However, the activity of these elements was never observed in wildtype background. Although a connection between epigenetic regulation and genome duplication events has not been experimentally demonstrated, it is not unreasonable to suggest that a relaxation of epigenetic control precedes or accompanies increases in ploidy. One could imagine a scenario whereby genome duplication and TE activation occurred in the same individual(s). Subsequently, the availability of numerous safe havens created by genome duplication would begin an era of TE-related genome expansion.

Materials and Methods

Database search strategies, sequence and phylogenetic analysis.

For each type of TE, coding sequences of known *Arabidopsis* elements were compared (not shown) and the most conserved region was identified (see Results). These regions were used as queries in tBlastn searches against the *Arabidopsis* database (ATH1_bacs.seq) available at The Institute for Genomic Research (TIGR; http://tigrblast.tigr.org/er-blast/index.cgi?project=ath1) to identify all *Arabidopsis* homologs. These homologs were compared by CLUSTALW multiple alignments and used to generate phylogenetic trees (not shown). One sequence from each major lineage of a certain type of *Arabidopsis* TE was then used as query in tBlastn searches against the TIGR *B. oleracea* database (brassica prelim sequences; http://tigrblast.tigr.org/euk-blast/index.cgi?project=bog1). The resulting hits were compared and redundant sequences (identified by two or more *Arabidopsis* queries) were removed. Severely truncated *B. oleracea* sequences (lacking ten or more amino acid residues) were not analyzed further. The remaining *B. oleracea* sequences were compared to their *Arabidopsis* homologs by CLUSTALW multiple alignments (available upon request) and used to generate phylogenetic trees described in the text. Multiple sequence alignments were performed with the CLUSTALW server available at European Bioinformatics Institute (http://www.ebi.ac.uk/clustalw/) with default parameters. Phylogenetic trees were

generated based on the neighbor-joining method, using PAUP* Version 4.0b8 program with default parameters.

Estimation of TE copy numbers in *B. oleracea*.

The number of hits from *B. oleracea* could not be directly converted into TE copy number because the TIGR *B. oleracea* database consists of short reads (on average ~650 bp). For example, two hits from two different reads could represent different regions of the same element. Note that the length of queries used in our Blast searches was less than the average length of reads and therefore one element could not have been represented more than twice. Thus, the hits for a particular type of TE could be divided into two groups: those covering the entire query (full-length hits) and those containing only part of the query (partial hits). Each full-length hit represents one element, whereas statistically each partial hit represents half an element. Let the probability that a particular hit is full length be P_f, the probability that a hit is partial should be 1- P_f. The copy number of a certain type of element in the current *B. oleracea* database (N_{cp}) can be estimated based on the number of hits (N_{hits}) as:

$$N_{cp} = N_{hits} * \left[P_f + 1/2\left(1 - P_f\right)\right] \quad \text{(A)}$$

or

$$N_{cp} = N_{hits} * 1/2\left(1 + P_f\right) \quad \text{(B)}$$

P_f can be calculated based on the query length (L_q) and read length (L_c) as:

$$P_f = \left(L_c - L_q\right)/\left(L_c + L_q\right) \quad \text{(C)}$$

Equation C states that the probability that a hit is full length equals the total number of ways that the entire query is located on a read, divided by the total number of ways that the query and a read overlap by at least one residue. In reality, however, a read must contain a minimum number of residues homologous to the query (L_{min}) to be recognized by Blast programs. Thus, equation B should be modified as:

$$P_f = \left[L_c - \left(L_q - 2 * L_{\min}\right)\right] / \left[L_c + \left(L_q - 2 * L_{\min}\right)\right] \quad \text{(D)}$$

Combining equations B and D and considering the current *B. oleracea* database covers approximately one third of the genome, the total copy (N) in the entire genome should be:

$$N = 3 * Ncp = 3 * Nhits * 1/2\left\{1 + \left[L_c - \left(L_q - 2 * L_{\min}\right)\right] / \left[L_c + \left(L_q - 2 * L_{\min}\right)\right]\right\} \quad \text{(E)}$$

Thus, the copy number of a certain type of elements can be calculated based on the number of hits and query length using equation E. Two tests were performed on equation E and the results are provided in Figure 4.9.

References

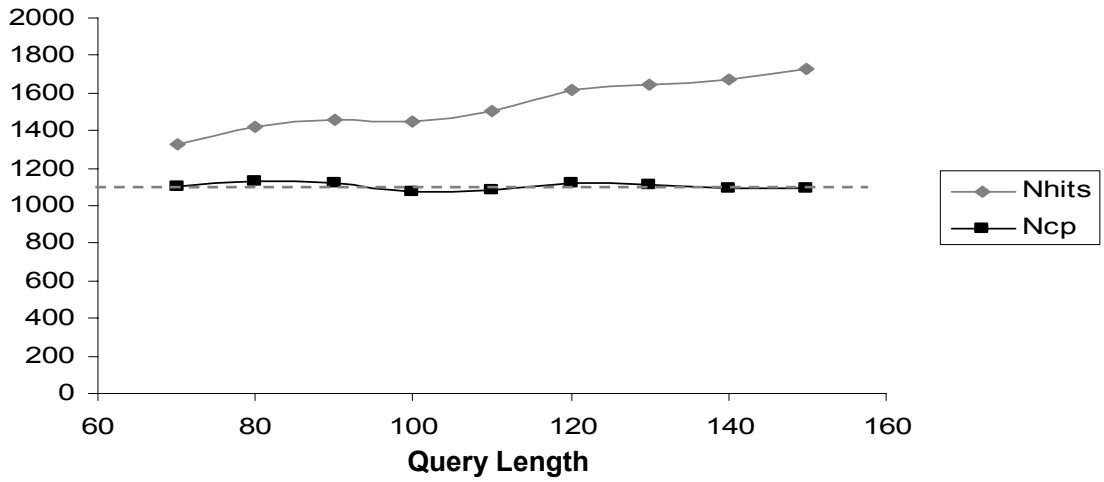
Arumuganathan, K. and E.D. Earle. 1991. Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* **9**: 208-218.

Bennetzen, J.L. and E.A. Kellogg. 1997. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**: 1509-1514.

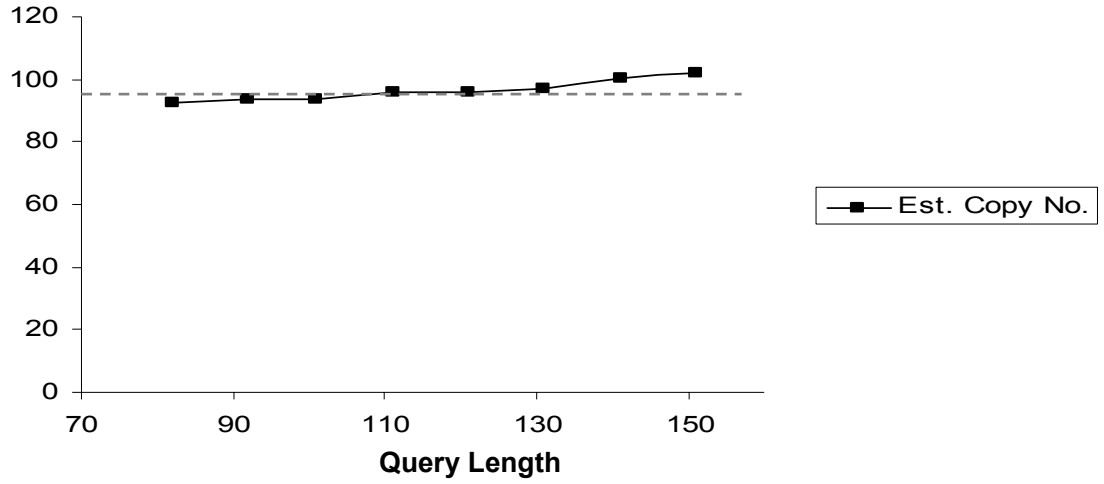
Capy, P., C. Bazin, D. Higuet, and T. Langin. 1998. *Dynamics and evolution of transposable elements*. Springer-Verlag, Austin, Texas.

Figure 4.9. Test results of euqation E. Equation E contains two correlated variables, query length (L_q) and number of hits (N_{hits}). Increase of query length resulted in more hits. However, the total copy number should not change as it is a constant. This was tested on *B. oleracea* LINEs as an example, where the most conserved RT region in *Arabidopsis* elements was used as queries. As shown in Figure 4.9a, when L_q increased from 70 a.a. to 150 a.a., N_{hits} increased from 1,329 to 1,729. However, there was very little variation in N_{cp} values (on average 1,103). Thus, the total copy number of LINEs in *B. oleracea* was estimated as 3,310. A second test of equation E was performed on a simulated database. One hundred *Arabidopsis* fragments, each 10 kb long and containing a LINE element, were combined and the resulting 1 Mb sequence was randomly divided by a computer program into 1,534 segments. Each segment was 600-700 bp long and the average length was 652 bp. The same queries as used above were used to blast this simulated database, and the results were shown in Figure 4.9b. Although there appeared to be a slight underestimate, the average value of N_{cp} (96.5) was very close to the actually copy number of LINEs in the database.

A


2000
1800
1600
1400
1200
1000
800
600
400
200
0
60
80
100
120
140
160
Query Length
Nhits
Ncp

B


120
100
80
60
40
20
0
70
90
110
130
150
Query Length
Est. Copy No.

Cavell, A.C., D.J. Lydiate, I.A. Parkin, C. Dean, and M. Trick. 1998. Collinearity between a 30-centimorgan segment of Arabidopsis thaliana chromosome 4 and duplicated regions within the Brassica napus genome. *Genome* **41**: 62-9.

Devos, K.M., J.K. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* **12**: 1075-9.

Doolittle, R.F., D.-F. Feng, M.s. Johnson, and M.A. McClure. 1989. Origins and evolutionary relationships of retroviruses. *Q.Rev.Biol.* **64**: 1-30.

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

Hull, R. 1999. Classification of reverse transcribing elements: a discussion document. *Arch Virol* **144**: 209-13; discussion 213-4.

Jiang, N., Z. Bao, X. Zhang, H. Hirochika, S.R. Eddy, S.R. McCouch, and S.R. Wessler. 2003. An active DNA transposon family in rice. *Nature* **421**: 163-7.

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

Kentner, E.K., M.L. Arnold, and S.R. Wessler. 2003. Characterization of high-copy-number retrotransposons from the large genomes of the louisiana iris species and their use as molecular markers. *Genetics* **164**: 685-97.

Kowalski, S.P., T.-H. Lan, K.A. Feldmann, and A.H. Paterson. 1994. Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea*

chromosomes reveals islands of conserved organization. *Genetics* **138**: 499-510.

Kunze, R. and C.F. Weil. 2002. The hAT and CACTA superfamilies of plant transposons. In *Mobile DNA II* (ed. N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz), pp. 565-610. American Society for Microbiology Press, Washington DC.

Lagercrantz, U. 1998. Comparative mapping between Arabidopsis thaliana and Brassica nigra indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* **150**: 1217-28.

Lan, T.H., T.A. DelMonte, K.P. Reischmann, J. Hyman, S.P. Kowalski, J. McFerson, S. Kresovich, and A.H. Paterson. 2000. An EST-enriched comparative map of Brassica oleracea and Arabidopsis thaliana. *Genome Res.* **10**: 776-88.

Langdon, T., G. Jenkins, R. Hasterok, R.N. Jones, and I.P. King. 2003. A High-Copy-Number CACTA Family Transposon in Temperate Grasses and Cereals. *Genetics* **163**: 1097-108.

Lisch, D. 2002. Mutator transposons. *Trends Plant Sci.* **7**: 498-504.

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

Marin, I. and C. Llorens. 2000. Ty3/Gypsy retrotransposons: description of new Arabidopsis thaliana elements and evolutionary perspectives derived from comparative genomic data. *Mol. Biol. Evol.* **17**: 1040-9.

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**: 1660-76.

Miura, A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani. 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature* **411**: 212-4.

Noma, K., H. Ohtsubo, and E. Ohtsubo. 2000. ATLN elements, LINEs from Arabidopsis thaliana: identification and characterization. *DNA Res* **7**: 291-303.

Paterson, A.H., T.H. Lan, K.P. Reischmann, C. Chang, Y.R. Lin, S.C. Liu, M.D. Burow, S.P. Kowalski, C.S. Katsar, T.A. DelMonte, K.A. Feldmann, K.F. Schertz, and J.F. Wendel. 1996. Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. *Nat. Genet.* **14**: 380-2.

Peterson, D.G., S.R. Schulze, E.B. Sciara, S.A. Lee, J.E. Bowers, A. Nagel, N. Jiang, D.C. Tibbitts, S.R. Wessler, and A.H. Paterson. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* **12**: 795-807.

Plasterk, R.H.A. and H.G. van Luenen. 2002. The Tc1/*mariner* family of transposable elements. In *Mobile DNA II* (ed. N.L. Craig, R. Craigie, M.

Gellert, and A.M. Lambowitz), pp. 519-532. American Society for Microbiology Press, Washington D.C.

Pringle, C.R. 1998. Virus taxonomy--San Diego 1998. *Arch Virol* **143**: 1449-59.

Rubin, E., G. Lithwick, and A.A. Levy. 2001. Structure and evolution of the hAT transposon superfamily. *Genetics* **158**: 949-57.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796-815.

Sadowski, J., P. Gaubier, M. Delseny, and C.F. Quiros. 1996. Genetic and physical mapping in Brassica diploid species of a gene cluster defined in Arabidopsis thaliana. *Mol. Gen. Genet.* **251**: 298-306.

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.

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

Singer, T., C. Yordan, and R.A. Martienssen. 2001. Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). *Genes Dev.* **15**: 591-602.

Suzuki, G., N. Kai, T. Hirose, K. Fukui, T. Nishio, S. Takayama, A. Isogai, M. Watanabe, and K. Hinata. 1999. Genomic organization of the S locus:

Identification and characterization of genes in SLG/SRK region of S(9) haplotype of Brassica campestris (syn. rapa). *Genetics* **153**: 391-400.

Szabados, L., I. Kovacs, A. Oberschall, E. Abraham, I. Kerekes, L. Zsigmond, R. Nagy, M. Alvarado, I. Krasovskaja, M. Gal, A. Berente, G.P. Redei, A.B. Haim, and C. Koncz. 2002. Distribution of 1000 sequenced T-DNA tags in the Arabidopsis genome. *Plant J.* **32**: 233-42.

Terol, J., M.C. Castillo, M. Bargues, M. Perez-Alonso, and R. de Frutos. 2001. Structural and evolutionary analysis of the copia-like elements in the Arabidopsis thaliana genome. *Mol. Biol. Evol.* **18**: 882-92.

Wicker, T., R. Guyot, N. Yahiaoui, and B. Keller. 2003. CACTA Transposons in Triticeae. A Diverse Family of High-Copy Repetitive Elements. *Plant Physiol.* **132**: 52-63.

Xiong, Y. and T.H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**: 3353-62.

Yang, Y.W., K.N. Lai, P.Y. Tai, and W.H. Li. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J. Mol. Evol.* **48**: 597-604.

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

Zhang, X., C. Feschotte, Q. Zhang, N. Jiang, W.B. Eggleston, and S.R. Wessler. 2001. *P Instability Factor*: an active maize transposon system associated

with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci. USA* **98**: 12572-12577.

Zhang, X., N. Jiang, C. Feschotte, and S. Wessler. 2003. PIF- and Pong-like Elements: Distribution, Evolution and Relationship with *Tourist*-like Miniature Inverted-repeat Transposable Elements. *Genetics* **Accepted**.

CHAPTER 5

CONCLUSIONS

The maize *Ac*/*Ds* family was discovered by Barbara McClintock over half a century ago through her skillful cytological and genetic studies (McClintock 1951). For decades after this discovery, plant TE studies focused on the genetic and molecular characterization of individual active families that could be identified when they inserted into genes and caused mutant phenotypes. However, we now know that the vast majority of the thousands or even millions of TEs colonizing plants genomes are transpositionally inactive and as such, have evaded "traditional" methods of discovery based on activity. This situation began to change with the availability of increasingly larger amounts of genomic sequences and the application of computational and phylogenetic approaches. These factors led to a shift in TE studies from the characterization of individual active families to the large scale identification and characterization of TEs based on sequence homology to previously characterized elements. More recently, sequence information and our knowledge of TEs have accumulated to such an extent that it is now possible to perform comparative analysis of all TEs in the genomes of related plant species.

The research described in this dissertation includes three parts that corresponded to three stages of TE studies. Part one (Chapter 2) describes the genetic and molecular analyses that led to the discovery that an active DNA element family (*PIF*) is associated with a *Tourist*-like MITE family (*mPIF*) in maize and the identification of a TPase-encoding *PIF* element. In part two (Chapter 3), computational and phylogenetic approaches that had been successfully applied to TE studies by others were adopted to characterize the *PIF*/*Pong* superfamily

with respect to its distribution, evolution and relationship with *Tourist*-like MITEs. Part three (Chapter 4) involves the derivation and use of a set of strategies and methodologies to analyze all major types of TEs in *Arabidopsis* and *B. oleracea* in order to address questions regarding the contribution of differential TE proliferation to the genome size difference.

The active maize *PIF* family is associated with a family of *Tourist*-like MITEs: molecular and genetic approaches.

Since the initial discovery of *Tourist* MITEs over a decade ago (Bureau and Wessler 1992; Bureau and Wessler 1994; Bureau et al. 1996), numerous *Tourist*-like MITE families have been identified from plant and animal genomes (Feschotte et al. 2002b; Feschotte et al. 2002a). Common features shared by *Tourist*-like MITE families, including similar TIRs and target site preference, suggested that there existed a large superfamily of DNA transposons that were responsible for their origin and amplification. Such a superfamily, however, had remained elusive for nearly a decade as the TIRs and TSDs of *Tourist* MITEs were not related to any transposase-encoding DNA element.

The starting point for this study was *PIF*, an active family of DNA transposons first identified by their mutagenic insertions into the maize *R* gene (Walker et al. 1997). *PIF* was of interest because it shared sequence similarity with a short segment that structurally resembled *Tourist*-like MITEs. We demonstrated that this segment was the founding members of large *Tourist*-like MITE family (called *mPIF*), and significant similarities shared by *PIF* and *mPIF*

elements led to the conclusion that *mPIF* MITEs are nonautonomous members of the *PIF* family (Zhang et al. 2001). These findings marked the first time that a MITE family was linked to a genetically active DNA element family. The activity of *PIF* elements was critical in this study as it permitted the identification and isolation of a TPase-encoding element (*PIFa*) using genetic and molecular approaches.

Distribution, diversity and evolution of *PIF*/*Pong*-like elements and their relationship with *Tourist*-like MITEs: computational and phylogenetic approaches.

The wide distribution and diversity of *Tourist*-like MITEs suggested that their TPase-encoding autonomous partners should also be widespread and diverse. A second active family associated with *Tourist*-like MITEs, the *Pong* family, was found in rice shortly after the isolation of *PIF* (Jiang et al. 2003). *PIF* and *Pong* are related as they share homology in both their TPases and TIRs. However, *PIF* and *Pong* were not related to any previously identified DNA element. In order to examine the distribution and diversity of *PIF*/*Pong*-like elements and their relationship with *Tourist*-like MITEs, it was necessary to identify additional *PIF*/*Pong*-like elements using methods independent of element activity. For this reason, computational and phylogenetic approaches that were successfully applied to analyze other TEs were adopted. Database searches led to the finding that *PIF*/*Pong*-like elements are widespread and abundant in eukaryotes, together comprising a new superfamily of eukaryotic DNA elements called the

PIF/*Pong* superfamily. Subsequent phylogenetic analysis of ~600 sequences resolved the three major groups of this superfamily, plant *PIF*, plant *Pong* and an animal group. The two plant groups were analyzed in detail, where it was found that both *PIF* and *Pong*-like elements are represented by multiple ancient and distinct lineages that co-exist in plant genomes.

The availability of the rice draft sequence made it possible to examine the relationship between *PIF*/*Pong*-like elements and *Tourist*-like MITEs on a whole genome scale. Identification of all *PIF*/*Pong*-like elements in this genome and comparison between these elements and *Tourist*-like MITEs showed that nearly all *Tourist*-like MITEs in rice are related to either *PIF*- or *Pong*-like elements. Furthermore, clear-cut relationships were found between several *PIF*/*Pong*-like families and *Tourist*-like MITEs. Taken together, these findings indicate that members of the *PIF*/*Pong* superfamily are responsible for the origin and amplification of *Tourist*-like MITEs.

Characterization of rice *PIF*/*Pong*-like elements also revealed a surprisingly high level of sequence divergence, both in the coding and non-coding regions. The average interfamily sequence identity in the TPase coding region was found to be below 60%, and there is virtually no interfamily sequence similarity in non-coding regions except for their short TIRs. This observed high level of interfamily sequence divergence has implication regarding the generally perception that DNA elements are present at low copy number and do not contribute significantly to plant genome size. This notion was based on the low copy number of active DNA element families, as determined by approaches such

as Southern blots (e.g., *Ac*/*Ds* and *En*/*Spm* families in maize have dozens of copies). However, it is most likely that active DNA families only represent the tip of the iceberg and that the vast majority of elements in plant genomes are transpositionally inactive and distantly related to the active lineages. For example, the active *Pong* family consists of only five copies, yet 25 additional *Pong*-like families exist in rice that include 80 TPase-encoding elements and an even larger number of internally deleted nonautonomous elements. Similarly, the active maize *PIF* family consists of ~25 copies. However, the dPCR assay as well as database search results from a very limited dataset revealed the presence of at least a dozen different *PIF*-like TPases in maize that are less than 65% identical to each other even in the most conserved catalytic domain. Similar situations have also been documented for other DNA element superfamilies such as plant *mariner*-like elements (Feschotte et al. 2003) and MULEs (Walbot and Rudenko 2002). High levels of interfamily sequence divergence would have prevented the detection of related but distinct families by Southern blots when the sequence of an active family is used as probe. Thus, previous conclusions regarding the abundance of DNA elements in plant genomes are most likely gross underestimates.

Comparative analysis of TEs in *Arabidopsis* and *B. oleracea*: the derivation and use of new strategies and methodologies in computational and phylogenetic analysis.

The above analyses were focused on one type of TE (*PIF*/*Pong*-like elements and *Tourist*-like MITEs). As our knowledge about TEs and the available genomic sequence information accumulates, it is now becoming possible to characterize and compare virtually all TEs from the genomes of related plant species. This type of analysis is more species-orientated than TE-orientated and can be used to address questions such as the role of TE proliferation in the evolution of plant genomes. Part three of this dissertation (Chapter 4) represents, to our knowledge, the first comparative analyses of *Arabidopsis* and *B. oleracea* TEs used to determine the mechanism(s) underlying expansion of the *B. oleracea* genome and the possible reasons for the low TE content of *Arabidopsis*. The results of this study indicate that the amplification of both retroelements and DNA elements contributed significantly to the *B. oleracea* genome expansion, and that the lack of significant amplification of any TE type, is responsible for the low TE content of *Arabidopsis*. Our results also indicate that the compact *Arabidopsis* genome has not sustained significant loss of TEs since sharing a common ancestor with *B. oleracea*.

The findings from this part of the study differ in several ways from previous conclusions based on analysis of large grass genomes. Specifically, for the grasses it was found that amplification of a few families of LTR-retroelements was largely responsible for genome size variations (SanMiguel et al. 1996;

SanMiguel et al. 1998; Meyers et al. 2001). In *B. oleracea*, however, massive amplification of a few families of LTR-retroelements did not occur. Instead, the proliferation of numerous families of LTR as well as non-LTR retroelements contributed significantly to genome expansion in this species. DNA elements have also amplified to high copy number and account for a large fraction of total genomic DNA. Thus, the previous conclusions cannot be generalized to all plant genomes.

Perspectives.

The computational analyses described in this study provided support for the conclusion that *PIF*/*Pong*-like elements are responsible for the origin and amplification of *Tourist*-like MITEs. Although strong circumstantial evidence is presented, in the absence of activity, our results cannot definitively conclude that *Tourist*-like MITEs are non-autonomous members of the *PIF*/*Pong* superfamily. Such a conclusion must await direct test using *in vivo* and *in vitro* assay systems. In addition, further molecular and biochemical analyses are required to characterize the transposition mechanisms of the active *PIF* and *Pong* elements and address questions about the relative success of MITEs.

The pioneering work of Barbara McClintock fundamentally changed our view of genomes: they are dynamic due to their TE complement. We now know that the TE component is much larger than anyone had anticipated, especially in plants. Perhaps the most interesting finding from the analysis of the TE content of *Arabidopsis* and *B. oleracea* is not *how* they are different from large grass

genomes, but that they *can* be different. That is, although TEs are major components of all plant genomes, they have contributed in different ways: large grass genomes represent one scenario, the small genomes of *Arabidopsis* and *B. oleracea* represent something different .Finding two very different scenarios in the first two groups of plants studies indicates that the evolutionary history of a group of organisms and its TE complement will probably be unique. This is really not surprising given the opportunistic nature of evolution. This also means that understanding the mechanisms underlying the different reasons for genome expansion will be an exciting area of future research. In addition, it will be interesting to know whether TE amplification coincided with major evolutionary events such as polyploization. Evidence has accumulated recently indicating that TEs are under rigorous epigenetic control by their host. In fact, it has been proposed that epigenetic mechanisms have evolved in plants largely to regulate TEs. Given that explosive amplifications of a variety of TEs have been documented repeatedly in many plant and animal genomes, it will be of great interest to determine whether one or many mechanisms have been responsible for the ability of TEs to evade, even temporarily, epigenetic regulation.

References

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 wild-type rice genes. *Proc. Natl. Acad. Sci. USA* **93**: 8524-8529.

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

-. 1994. *Stowaway*: a new family of inverted-repeat elements associated with genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6**: 907-916.

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

Feschotte, C., L. Swamy, and S.R. Wessler. 2003. Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with *Stowaway* MITEs. *Genetics **163***: 747-758.

Feschotte, C., X. Zhang, and S. Wessler. 2002b. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In *Mobile DNA II* (ed. N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz), pp. 1147-1158. American Society for Microbiology Press, Washington, DC.

Jiang, N., Z. Bao, X. Zhang, H. Hirochika, S.R. Eddy, S.R. McCouch, and S.R. Wessler. 2003. An active DNA transposon family in rice. *Nature* **421**: 163-7.

McClintock, B. 1951. Chromosome organization and genic expression. *Cold Spring Harbor Symp.Quant.Biol.* **16**: 13-47.

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**: 1660-76.

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.

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

Walbot, V. and G.N. Rudenko. 2002. *MuDR*/*Mu* Transposable Elements of Maize. In *Mobile DNA II* (ed. N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz), pp. 533-564. American Society for Microbiology Press, Washington, DC.

Walker, E.L., W.B. Eggleston, D. Demopulos, J. Kermicle, and S.L. Dellaporta. 1997. Insertions of a novel class of transposable elements with a strong target site preference at the *r* locus of maize. *Genetics* **146**: 681-693.

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

APPENDIX A

BOS: A NEW SUPERFAMILY OF SHORT INTERSPERSED NUCLEAR ELEMENTS (SINES) WIDESPREAD IN BRASSICACEAE

INTRODUCTION

Class 1 (RNA) elements transpose via an RNA intermediate that involves a reverse transcription step. There are two types of class 1 elements: those flanked by long terminal direct repeats (LTRs) are called LTR-retrotransposons, and those do not possess LTR but terminate at 3' end with poly A are referred to as non-LTR retrotransposons. LTR retrotransposons were further divided into *Ty1*/*copia*-like and *Ty3*/*gypsy*-like based on the order of proteins encoded by their *pol* gene. Non-LTR retrotransposons include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs).

LINEs are widespread in all three eukaryotic kingdoms but much mre diverse and abundant in animals than in plants and fungi. Phylogenetic analyses of LINEs have resolved five lineages: R2, L1, RTE, I and Jockey. All LINE lineages are present in animals and so far two lineages (L1 and RTE) have been found in plants. Elements in both L1 and RTE lineages encode a protein containing two domains, ADE and RT, that are involved in the nicking of target DNA and cDNA synthesis, respectively. LINEs belonging to the L1 group encode an additional ORF (*orf1*) that may be involved in RNA binding. The 5' region of LINEs contains internal promoter motifs (A box and B box) that are recognized by DNA polymerase III (Pol III), and their 3' region contains a polyadenylation signal. LINEs are flanked by target site duplication (TSD) of variable length (usually 10-20 bp) that are generated upon retrotransposition. SINEs are short (less than 500 bp), nonautonomous elements that structurally resemble LINEs with an internal

Pol III promoter near 5' end and terminate at 3' end with poly A. SINEs do not possess any protein coding capacity and their retrotransposition is catalyzed *in trans* by the protein complex encoded by LINEs. Similar to LINEs, retrotransposition of SINEs also results in the duplication of target site sequences of variable length. Some SINEs appeared to have derived from LINEs as they share significant homology in the 3' region, whereas other SINEs are more likely to have originated from small cellular RNA genes such as tRNA or 7SL RNA.

LINEs and SINEs are abundant in animals. For example, the human genome harbors more than 500,000 LINE1 (L1) elements, accounting for ~17% of the total nuclear DNA. Proteins encode by L1 elements also mobilize a large family of SINEs (called *Alu*) that are present with ~1.5 million copies and make up another ~10% of the human genome. The abundance of LINEs varies significantly from species to species. For example, ~0.1% of the maize genome (~2,500 Mb) consists of LINEs, whereas a single family of LINEs (called *del2*) is present with 250,000 copies in the lily genome (90,000 Mb) and make up ~4% of total genomic DNA. SINEs have also been described in plants, such as the *p-SINE1* family in rice (~6,500 copies), *TS* family in tobacco (50,000 copies), and the *AtSN1*, *AtSN2*/RAthE1 and RAthE2 families in *Arabidopsis* (~70, ~130-150 and ~60 copies, respectively). The most extensively studied plant SINE is the S1 family identified from *Brassica napus* (~500 copies). S1 is present in all *Brassica* species tested (e.g. *B. rapa* and *B. oleracea*), but not in *Arabidopsis*, indicating that it has emerged after the divergence of *Brassica* and *Arabidopsis*.

A recently survey of transposable elements in the *B. oleracea* genome showed that there are over 3,000 LINEs in this species, grouped into fifteen lineages that predated the divergence of *Brassica* and *Arabidopsis*. The abundance and diversity of LINEs in *B. oleracea* suggested that it may also harbor a large and diverse collection of SINEs. Here we report the identification and characterization of a new superfamily of SINEs (named *BoS*) that are present with at least 2,000 copies in *B. oleracea*, of which 436 were analyzed in detail. *BoS* elements exhibit the typical structural features of SINEs, including short size (167-229 bp), conserved motifs (A box and B box) in 5' promoter region, 3' poly A tract and flanking TSDs of variable length (10-20 bp). Based on sequence homology in their promoter regions, *BoS* elements were grouped into twelve distinct families (*BoS_a* through *BoS_k*), some of which are present in other *Brassica* species as well as *Arabidopsis*. Comparison of the sequence in the promoter region of *BoS* elements to that in the *Arabidopsis* tRNA genes suggested that most *BoS* families derived from a common tRNA ancestor that was most likely to be the glutamine tRNA gene.

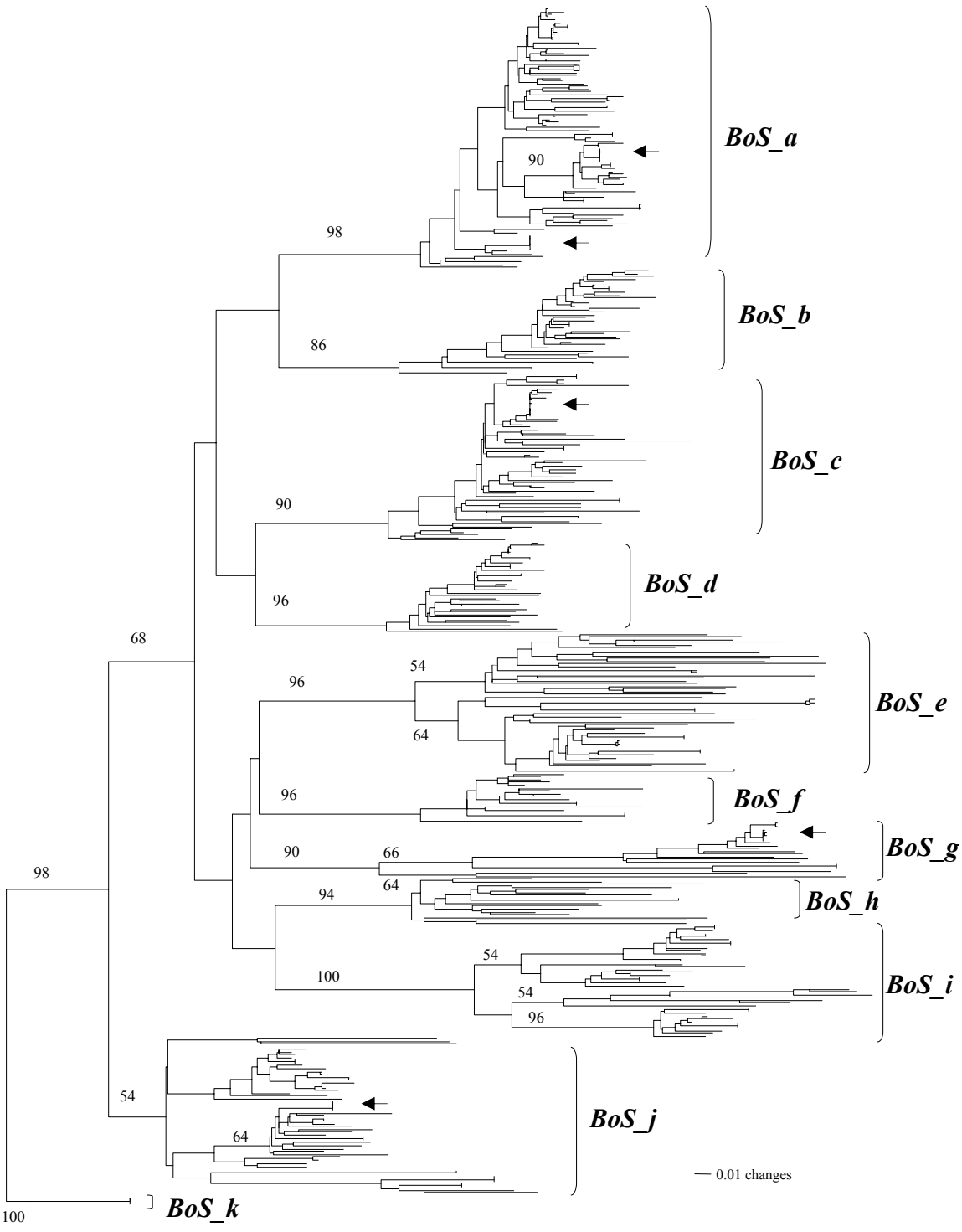
RESULTS AND DISCUSSION

Identification of a new superfamily of SINEs in *B. oleracea*. A 194-bp segment located on *B. oleracea* clone JBOGL76RB (position 165-358) was coincidentally found to exhibit structural features of a SINE, including short size, 3' poly (A) tract (22 bp) and a flanking direct repeat (12 bp with one mismatch).

This segment does not share significant sequence similarity with any known SINEs. Blastn searches using this segment as query identified over 900 hits in *B. oleracea*, of which 50 were full length (on average 188 bp), ~75-100% identical to the query, ended with 3' poly (A) tracts and most were flanked by short direct repeats. A large number of 5'-truncated elements were also found but not analyzed further. In addition, some hits appeared to be structurally complete but only shared weak sequence similarity to the query. Furthermore, some hits were distantly related to the 5' region but not the 3' region of the query sequence. Examination of the later type of hits revealed that some also ended with 3' poly (A) tracts and were flanked by short direct repeats. Thus, it appeared that several related but distinct families of SINE elements existed in *B. oleracea*. Blastn searches were performed using these hits as queries, and in some cases new families were found and used again as queries in further Blastn searches. This process was repeated 9 times and a total of 436 full-length SINEs were identified that, based on sequence similarity, can be grouped into distinct families (see below). We concluded that they represent a new superfamily of SINEs (named *BoS* for *B. oleracea* SINEs). The copy number of full-length *BoS* elements in *B. oleracea* was conservatively estimated to be at least 2,000.

Phylogenetic analysis of *BoS*. The highest sequence homology among *BoS* families was found in the 5' region (~80-90 bp). This region was used to deduce their evolutionary relationship. A phylogenetic tree was generated from a multiple alignment of this region from all full-length elements and shown in Fig. 1. Eleven

Figure 1. Phylogeny of *BoS* elements in *B. oleracea*. This is a neighbor-joining tree generated from a CLUSTALW multiple alignment of the 5' promoter region of 436 *BoS* elements. Bootstrap vales are calculated from 1,000 replicates. Arrows indicated clusters of elements with high sequence identities.


BoS_a
BoS_b
BoS_c
BoS_d
BoS_e
BoS_f
BoS_g
BoS_h
BoS_i
BoS_j
BoS_k
0.01 changes

lineages, or families, were resolved and named *BoS_a* through *BoS_k* (summarized in Table 1). The average length of a family ranged from 167 bp (*BoS_b*) to 229 bp (*BoS-f*). The most homogenous family is *BoS_c* with its 61 elements sharing on average 91% sequence identity, whereas *BoS_i* is the most heterologous family and the overall identity among its 41 members is only 61%. It should be noted that the observed low intrafamily sequence identity in *BoS_i* and several other families (e.g., *BoS_a*, *e*, *g* and *j*) was due to the existence of divergent subfamilies (see below). Several families included small numbers of elements (4-7) that were identical (arrows in Fig. 1), indicating that they may be recently active.

Table 1. Summary of 11 families of *BoS* SINEs in *B. oleracea*

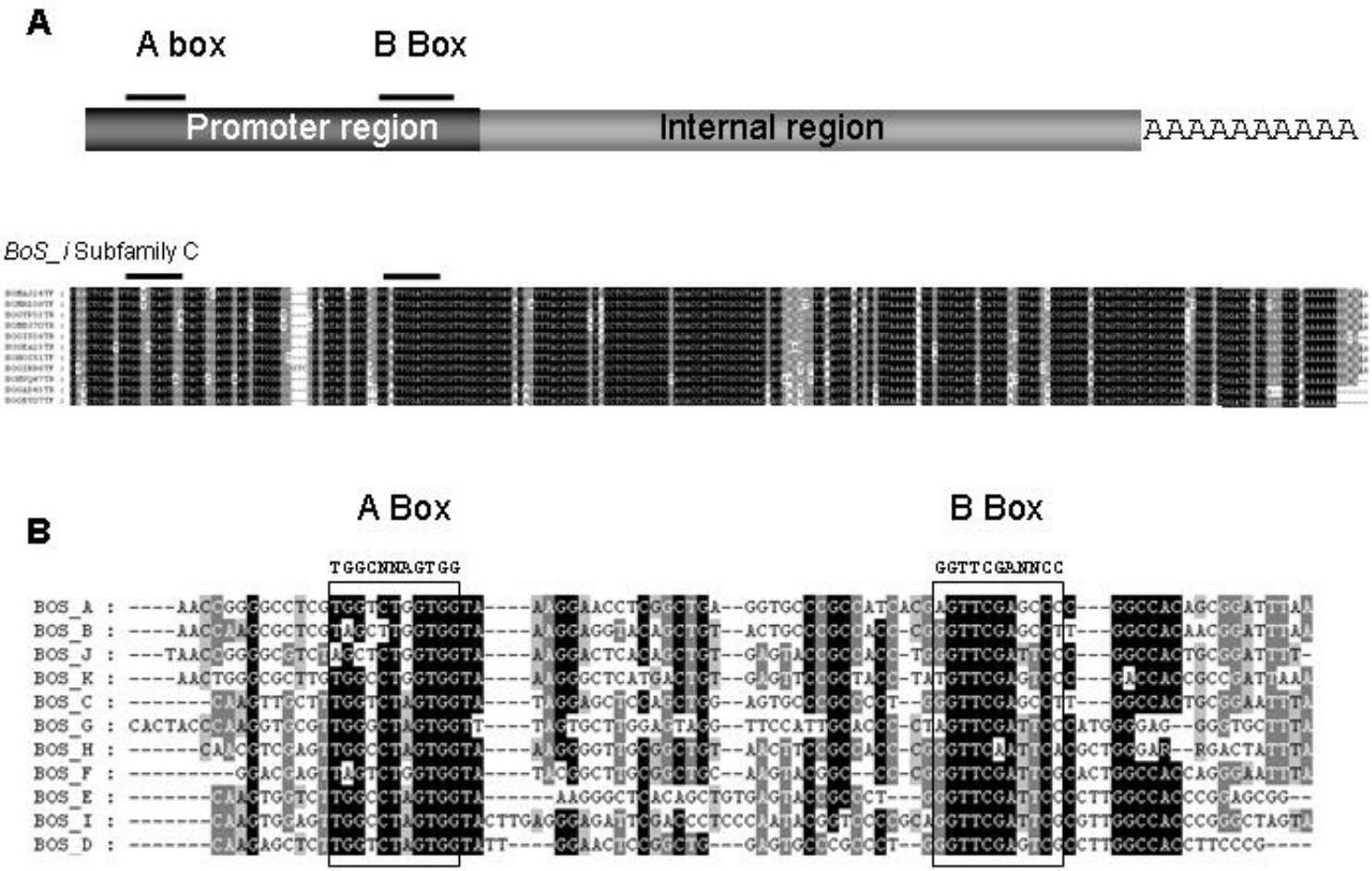
Family	No. of full-length elements	Average Length	Average identity	A Box Seq	B Box Seq
BoS_a	95	187	71%	TGGTCTGGTGG	AGTTCGAGCCC
BoS_b	38	167	89%	TAGCTTGGTGG	GGTTCGAGCCT
BoS_c	61	218	91%	TGGTCTAGTGG	GGTTCGAGCCT
BoS_d	33	172	88%	TGGTCTAGTGG	GGTTCGAGTCG
BoS_e	51	188	71%	TGGCCTAGTGG	GGTTCGATTCC
BoS_f	18	229	87%	TAGTCTGGTGG	GGTTCGATTCG
BoS_g	20	187	70%	TGGGCTAGTGG	AGTTCGATTCC
BoS_h	17	139	77%	TGGCCTAGTGG	GGTTCAATTCA
BoS_i	41	198	61%	TGGCCTAGTGG	GGTTCGATTCG
BoS_j	57	183	62%	AGCTCTGGTGG	GGTTCGATTCC
BoS_k	3	214	100%	TGGCCTGGTGG	TGTTCGAGTCC

Some described SINE families (e.g. S1) were generated from a small number of "founder" elements that proliferated into subfamilies; while other families (e.g. RAthE1) were generated from a single founder element and can not be divided into subfamilies. Both scenarios were observed for *BoS* elements. Six of the 11 families (*BoS_b*, *c*, *d*, *f*, *h* and *k*) did not exhibit subfamily structure, whereas

three (*BoS_e*, *g* and *i*) were clearly comprised of 2-3 subfamilies. Although the remaining two families (*BoS_a* and *j*) can not be divided into distinct subfamilies, they each contained a small monophyletic cluster of elements, the grouping of which was supported by bootstrap values. Interestingly, the internal regions of these two clusters of elements were different from the remaining elements in the same family.

Structure of *BoS* elements. All families of *BoS* elements exhibit characteristic structural features of SINEs, including a 5' region containing the promoter motifs (A box and B box) recognized by DNA polymerase III, a 3' poly (A) end, and an internal region between them. Alignments of members of the *BoS_i* family (subfamily C) are shown in Fig. 2a as an example to illustrate the typical structure of *BoS* elements. The 5' region was found to be similar in all *BoS* families, whereas the length and sequence of the internal region varied significantly across families and sometimes between subfamilies. A consensus sequence was derived for the 5' region of each family and compared in Fig. 2b. The promoter motifs were very well conserved and had significantly homology with those in SINEs (S1, *AtSN1*, *AtSN2*/RathE1 and RathE2) and *Arabidopsis* tRNA genes (Fig. 2b). In addition, there was also sequence conservation among *BoS* families in the regions between and immediately flanking the A and B boxes. However, these regions were dissimilar between *BoS* and described SINEs or *Arabidopsis* tRNA genes (not shown). Computer programs did not predict typical tRNA secondary structure *BoS* elements.

Figure 2. Structure of *BoS* elements. (A) Structure of *BoS_i* elements (subfamily c). A schematic representation of SINE is shown above. (B) Comparison of the 5' promoter region of 11 *BoS* families.


A
A box
B Box
Promoter region
Internal region
AAAAAAAAAA
BoS_i Subfamily C
B
A Box
B Box
TGGCNNAGTGG
GGTTCGANNCC

Distribution of *Bos* elements in Brassicaceae. Sequence variation among different *BoS* families indicated that they have diverged for a relatively long period of time. Therefore, it is possible that SINEs belonging to the *BoS* superfamily existed in other species of the Brassicaceae family. Consensus sequence of the 5' region of each *BoS* family was used as query to search the genomic sequences of *Arabidopsis* as well as other Brassica species. Several SINEs related to *BoS* were identified in *B. napus*, *B. juncea* and *Arabidopsis* (Table 2). Thus, the *BoS* superfamily of SINEs appeared to be widespread in Brassicaceae. Interestingly, each of the newly identified elements was more closely related to a particular *BoS* family than to any other families in *B. oleracea*. This suggested that the divergence *BoS* families preceded that of Brassicaceae.

Table 2. *BoS*-like LINEs in other Brassicaceae species

Species	Acc. No.	From	To	Length (bp)	Most Similar to
Arabidopsis	AC007843	44622	44850	229	*BoS_a*
	AC022492	24507	24735	229	*BoS_a*
	AC006577	66970	67197	228	*BoS_a*
	AP000600	60515	60743	229	*BoS_a*
	AC002409	38122	38350	229	*BoS_a*
B. napus	AF052241	1167	1330	164	*BoS_b*
B. juncea	AF271220	436	645	210	*BoS_f*

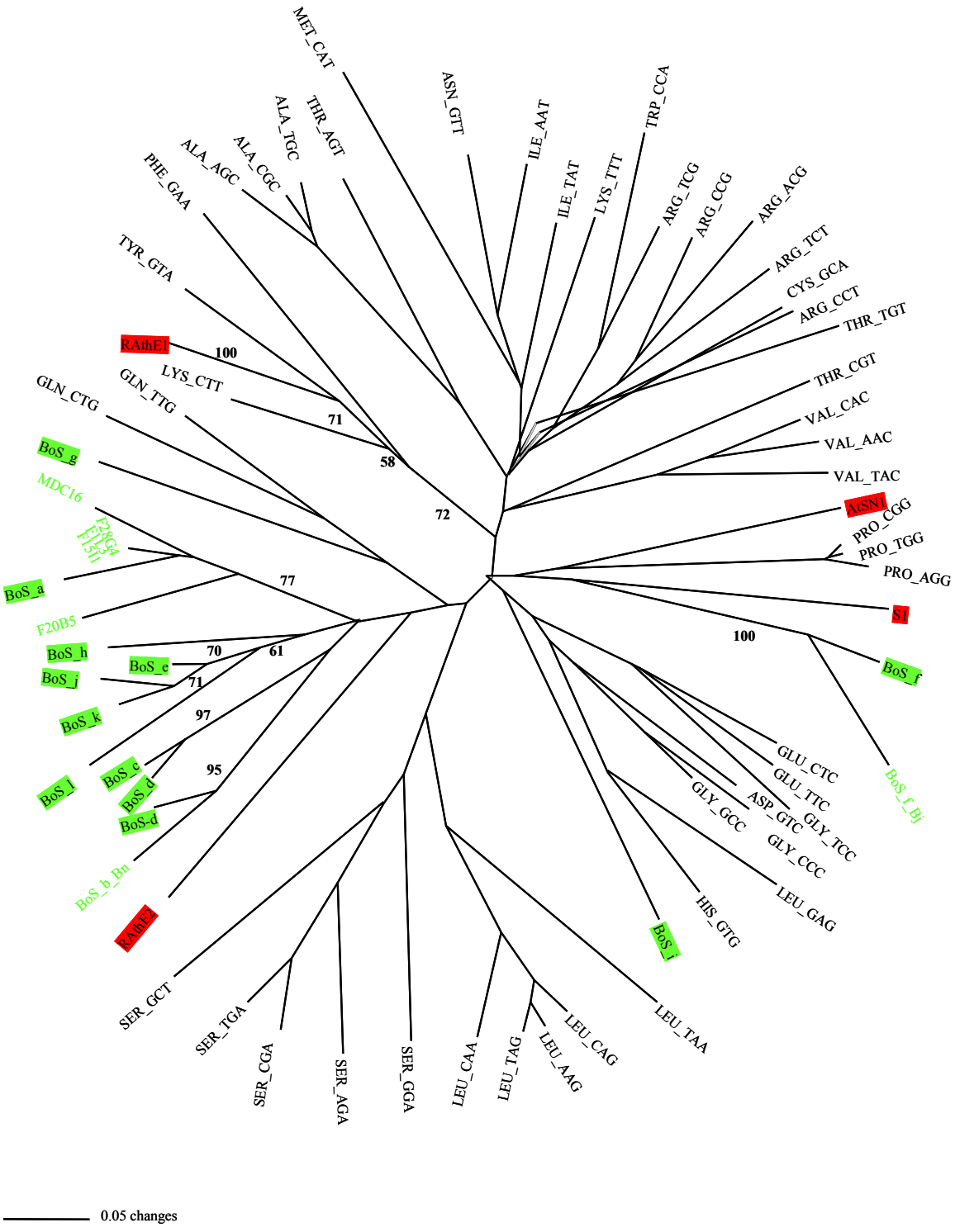
Evolutionary origin of *BoS* elements. Most plant SINEs were ancestrally derived from tRNA genes. In order to identify the possible tRNA ancestor(s) of *BoS* elements, the promoter regions (A box, B box and the intervening sequence) of *BoS* elements were compared to those of 47 *Arabidopsis* tRNA genes. The four described SINEs as well as several *BoS*-like SINEs from Brassicaceae were also included in this comparison. An unrooted neighbor-

joining tree was generated from a multiple alignment of these sequences and shown in Fig. 3. Most *BoS* families were found to be related to glutamine tRNA genes, except for *Bos_f*, which was more closely related to proline tRNA, and *BoS_i*, whose relationship with tRNA genes was unclear. Thus, most *BoS* families appeared to share a common tRNA gene ancestor that was likely the glutamine tRNA. However, it should be noted that such a relationship did not gain strong support from bootstrap values.

MATERIALS AND METHODS

Database searches, sequence and phylogenetic analyses. Blastn searches were performed using the blast server available at The Institute for Genome Research (TIGR; http://www.tigr.org) against the *B. oleracea* database (prelim Brassica sequences; http://www.tigr.org/tdb/e2k1/bog1/). The sequences of tRNA genes used in this study were obtained from The Genomic tRNA Database (GtRDB; http://rna.wustl.edu/GtRDB/). Multiple sequence alignments were performed with the CLUSTALW server available at European Bioinformatics Institute (http://www.ebi.ac.uk/clustalw/) with default parameters. Phylogenetic trees were generated based on the neighbor-joining method, using PAUP* Version 4.0b8 program with default parameters.

Figure 3. Relationship between *BoS* elements and *Arabidopsis* tRNA genes. This is a neighbor-joining tree generated from a multiple alignments of the 5' promoter region of consensus sequences of 11 *Bos* families, related *BoS* elements from other species, four previously described SINEs and 47 *Arabidopsis* tRNA genes. Bootstrap vales are calculated from 1,000 replicates.


MET_CAT
ASN_GTT
ILE_AAT
ILE_TAT
LYS_TTT
TRP_CCA
ARG_TCG
ARG_CCG
ARG_ACG
ARG_TCT
CYS_GCA
ARG_CCT
THR_TGT
THR_CGT
VAL_CAC
VAL_AAC
VAL_TAC
AtSN1
PRO_CGG
PRO_TGG
PRO_AGG
S1
BoS_f
BoS_f_Bj
GLU_CTC
GLU_TTC
GLY_TCC
ASP_GTC
GLY_GCC
GLY_CCC
LEU_GAG
HIS_GTG
BoS_i
LEU_TAA
LEU_CAG
LEU_AAG
LEU_TAG
LEU_CAA
SER_GGA
SER_AGA
SER_CGA
SER_TGA
SER_GCT
RAthE2
BoS_b_Bn
BoS-d
BoS_d
BoS_c
BoS_l
BoS_k
BoS_j
BoS_e
BoS_h
F20B5
BoS_a
F28G4
F15J5
MDC16
BoS_g
GLN_CTG
GLN_TTG
LYS_CTT
RAthE1
TYR_GTA
PHE_GAA
ALA_AGC
ALA_CGC
ALA_TGC
THR_AGT
100
71
58
72
77
70
61
71
97
95
100
0.05 changes

APPENDIX B

PIF-LIKE ELEMENTS IN ANIMAL GENOMES AND THEIR RELATIONSHIP WITH *TOURIST*-LIKE MITES

Phylogeny of animal *PIF*-like TPases.

PIF-like TPases from animals were more heterogeneous than their plant homologs. To determine their phylogenetic relationships, the catalytic domain of a representative subset of animal *PIF*-like TPases was compared by CLUSTALW multiple alignment and used to generate an unrooted phylogenetic tree (Figure 1). Three clusters (X-Z) were resolved; one (X) was seen only in zebrafish while two (Y and Z) were more widespread. Cluster Z was found in a wide range of vertebrates (fish, bird and mammal) while sequences in cluster Y were from invertebrates and *Xenopus*. Cluster Y was further divided into three lineages (Y1-Y3). The most basal, Y1, is present only in nematodes, whereas Y2 only included sequences from African malaria mosquito (*Anopheles gambiae*), and Y3 is found in insects and *Xenopus*. In both clusters Y and Z the phylogeny of TPases was fairly consistent with the phylogeny of the species.

***PIF*-like elements in *A. gambiae* and their relationship with *Tourist*-like MITEs**

The availability of a complete draft sequence of *A. gambiae* coupled with the abundance of *PIF*-like elements permitted an examination of the relationship between *PIF*-like elements and MITEs in this organism. One hundred and eleven *PIF*-like TPase fragments were identified in the *A. gambiae* sequence. As shown in Figure 2, *PIF*-like TPases clustered into a large Y2-like group (85 sequences) and a small Y3-like group (26 sequences). All TPases in the Y2 group contained the DD37E motif while those in Y3 contained either a DD35E or a DD36E motif.

Figure 1. Phylogeny of animal *PIF*-like elements. The unrooted phylogenetic tree was generated using the neighbor-joining method from a CLUSTALW multiple alignment of the catalytic domains of 18 animal *PIF*-like elements, selected to represent different lineages. Bootstrap values were calculated from 1,000 replicates.

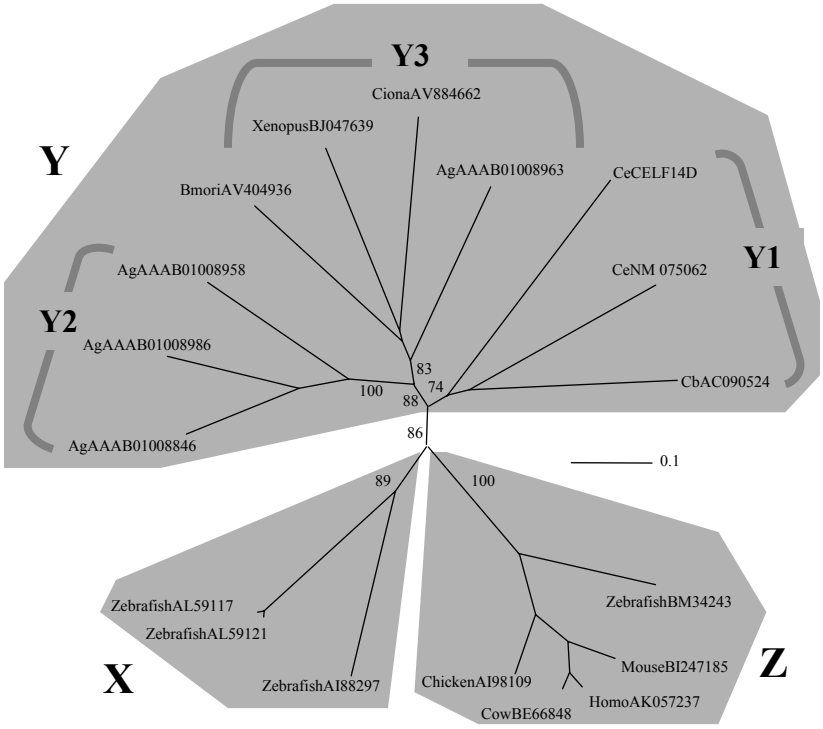
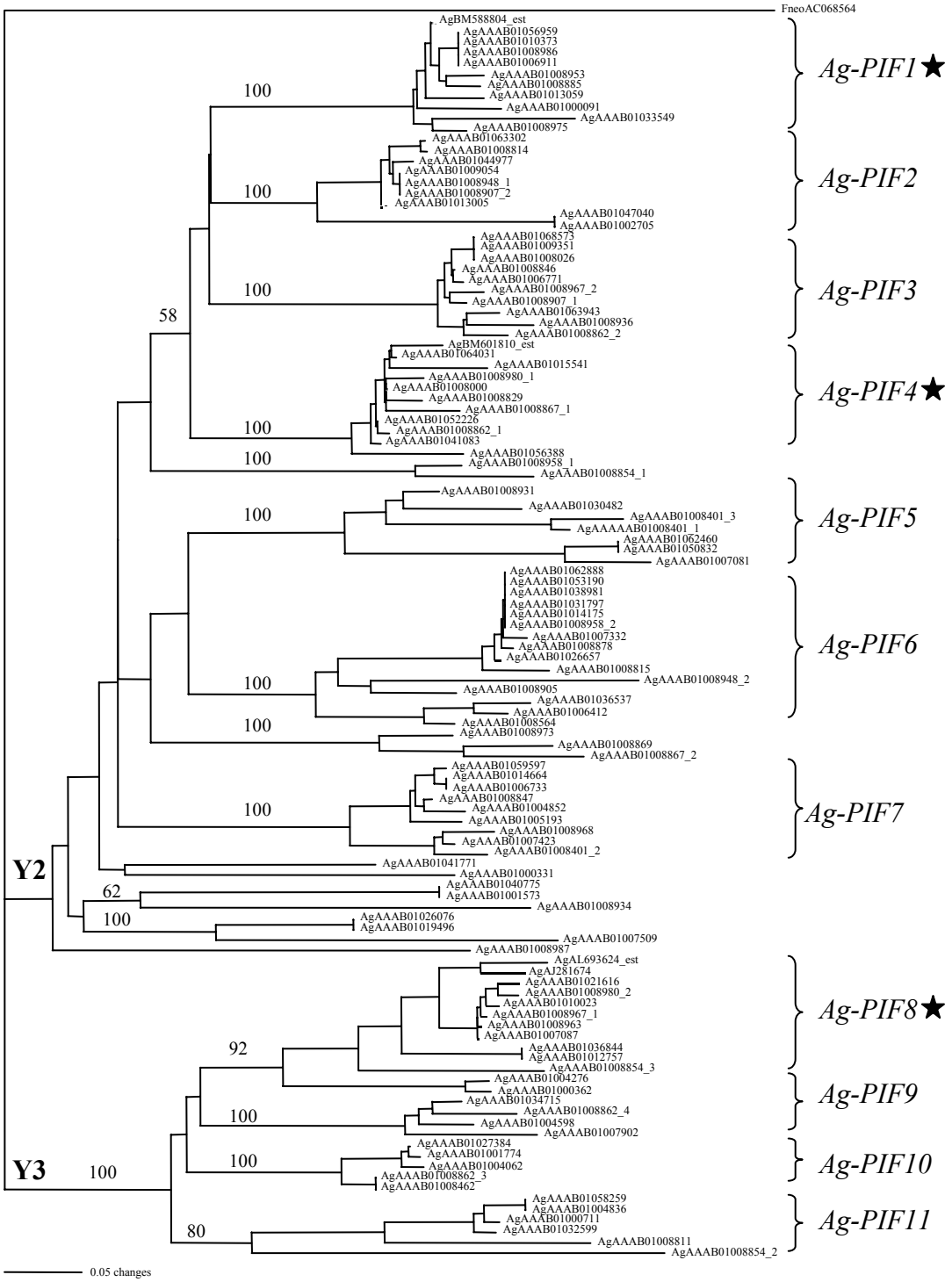

Y3
CionaAV884662
XenopusBJ047639
Y
AgAAAB01008963
CeCELF14D
BmoriAV404936
Y1
AgAAAB01008958
CeNM 075062
Y2
AgAAAB01008986
83
74
CbAC090524
100
88
AgAAAB01008846
86
0.1
89
100
ZebrafishBM34243
ZebrafishAL59117
ZebrafishAL59121
MouseBI247185
Z
X
ChickenAI98109
ZebrafishAI88297
HomoAK057237
CowBE66848

Figure 2. Phylogeny of *PIF*-like elements in *A. gambiae*. The phylogenetic tree was generated using the neighbor-joining method from a CLUSTALW multiple alignment of the catalytic domains of 114 *PIF*-like elements, of which 111 were mined from *A. gambiae* genomic sequence and 3 were ESTs. The catalytic domain of a *F. neoformans PIF*-like element was used as outgroup to root the tree. Bootstrap values were calculated from 1,000 replicates. Stars indicate that an EST sequence was identified for that family.

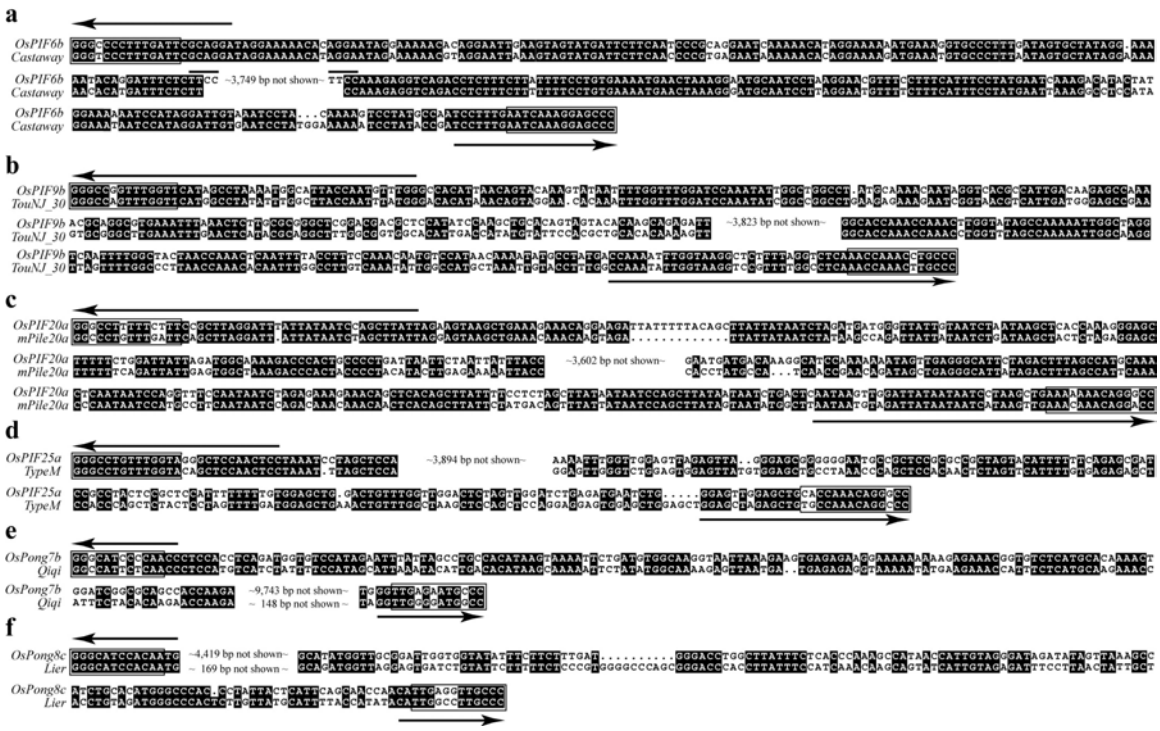

FneoAC068564
AgBM588804_est
AgAAAB01056959
AgAAAB01010373
AgAAAB01008986
AgAAAB01006911
AgAAAB01008953
AgAAAB01008885
AgAAAB01013059
AgAAAB01000091
AgAAAB01033549
AgAAAB01008975
100
Ag-PIF1★
AgAAAB01063302
AgAAAB01008814
AgAAAB01044977
AgAAAB01009054
AgAAAB01008948_1
AgAAAB01008907_2
AgAAAB01013005
100
AgAAAB01047040
AgAAAB01002705
Ag-PIF2
AgAAAB01068573
AgAAAB01009351
AgAAAB01008026
AgAAAB01008846
AgAAAB01006771
AgAAAB01008967_2
AgAAAB01008907_1
AgAAAB01063943
AgAAAB01008936
AgAAAB01008862_2
100
58
Ag-PIF3
AgBM601810_est
AgAAAB01064031
AgAAAB01015541
AgAAAB01008980_1
AgAAAB01008000
AgAAAB01008829
AgAAAB01008867_1
AgAAAB01052226
AgAAAB01008862_1
AgAAAB01041083
100
AgAAAB01056388
AgAAAB01008958_1
AgAAAB01008854_1
100
Ag-PIF4★
AgAAAB01008931
AgAAAB01030482
AgAAAB01008401_3
AgAAAAB01008401_1
AgAAAB01062460
AgAAAB01050832
AgAAAB01007081
100
Ag-PIF5
AgAAAB01062888
AgAAAB01053190
AgAAAB01038981
AgAAAB01031797
AgAAAB01014175
AgAAAB01008958_2
AgAAAB01007332
AgAAAB01008878
AgAAAB01026657
AgAAAB01008815
AgAAAB01008948_2
AgAAAB01008905
AgAAAB01036537
AgAAAB01006412
AgAAAB01008564
100
AgAAAB01008973
AgAAAB01008869
AgAAAB01008867_2
Ag-PIF6
AgAAAB01059597
AgAAAB01014664
AgAAAB01006733
AgAAAB01008847
AgAAAB01004852
AgAAAB01005193
AgAAAB01008968
AgAAAB01007423
AgAAAB01008401_2
100
Ag-PIF7
Y2
AgAAAB01041771
AgAAAB01000331
AgAAAB01040775
AgAAAB01001573
62
AgAAAB01008934
AgAAAB01026076
AgAAAB01019496
100
AgAAAB01007509
AgAAAB01008987
AgAL693624_est
AgAJ281674
AgAAAB01021616
AgAAAB01008980_2
AgAAAB01010023
AgAAAB01008967_1
AgAAAB01008963
AgAAAB01007087
AgAAAB01036844
AgAAAB01012757
AgAAAB01008854_3
92
Ag-PIF8★
AgAAAB01004276
AgAAAB01000362
AgAAAB01034715
AgAAAB01008862_4
AgAAAB01004598
AgAAAB01007902
100
Ag-PIF9
AgAAAB01027384
AgAAAB01001774
AgAAAB01004062
AgAAAB01008862_3
AgAAAB01008462
100
Y3
100
Ag-PIF10
AgAAAB01058259
AgAAAB01004836
AgAAAB01000711
AgAAAB01032599
AgAAAB01008811
AgAAAB01008854_2
80
Ag-PIF11
0.05 changes

Based on sequence identities and bootstrap values, seven families were assigned to the Y2 group (designated *Ag-PIF1* though *Ag-PIF7*), and four were assigned to the Y3 group (designated *Ag-PIF8* though *Ag-PIF11*). A number of families included members that were identical or highly similar to each other, indicating that they may still be capable of transposition. Consistent with this was the identification of three *A. gambiae* ESTs that are nearly identical to the coding regions of *Ag-PIF1*, *Ag-PIF4* and *Ag-PIF8* elements, respectively (Figure 2, indicated by stars).

Attempts to associate *PIF*-like elements with MITEs was problematic because most TPase sequences were located on short contigs and longer scaffolds often contained gaps in the vicinity of the TPases. Nevertheless, the termini of five families that could be successfully defined (*Ag-PIF1*, *2*, *3*, *6* and *8*) were compared with the terminal sequences of three previously described MITE families (*Joey*, *TAA-I-Ag* and *TAA-II-Ag*) that have the TSD TTA. *Ag-PIF1* and *Joey* have nearly identical terminal sequences (62 of 65 bp at 5'end and 35 of 35 bp at 3'end) (Figure 3a), while *Ag-PIF2* and *TAA-II-Ag* share extensive similarity over the entire length of the MITE (Figure 3b).

The terminal sequences of each *Ag-PIF* element with defined termini were also used to search the *A. gambiae* genomic sequence for related MITE families that have not been reported. In this way, two new MITE families were identified; both are flanked by TAA/TTA TSDs (Table 1), and each displays extensive sequence similarity over their entire length with one of the five *PIF*-like families

Figure 3. Relationship between *PIF*-like elements and *Tourist*-like MITEs in *A. gambiae*. (a) Comparison of the terminal sequences of *Ag-PIF1*, *Joey* (211 bp internal sequence not shown) and the entire sequence of *Ag-mPIF1*. (b) Comparison of the terminal sequences of *Ag-PIF1* and the entire sequence of *TTA-II-Ag*. (c) Comparison of the terminal sequences of *Ag-PIF6* and the entire sequence of *Ag-mPIF6*. In each case, the MITE sequence is a consensus derived from all complete copies in the *A. gambiae* genome. Arrows indicate element TIRs. Horizontal lines indicate the presence of direct repeats flanking the deletion breakpoint in MITEs.


a
OsPIF6b
Castaway
~3,749 bp not shown~
b
OsPIF9b
TouNJ_30
~3,823 bp not shown~
c
OsPIF20a
mPile20a
~3,602 bp not shown~
d
OsPIF25a
TypeM
~3,894 bp not shown~
e
OsPong7b
Qiqi
~9,743 bp not shown~
~ 148 bp not shown ~
f
OsPong8c
Lier
~4,419 bp not shown~
~ 169 bp not shown ~

with defined termini (*Ag-PIF1*, *6*) (Figure 3a 3c). For this reason, we have designated these MITEs *Ag-mPIF1* and *Ag-mPIF6*.

Table 1. Summary of four *Tourist*-like MITE families in *A. gambiae*

MITE family	Acc. No. and Position*	Average Size	Average Pairwise Identity	No. of Complete Copies	No. of Partial Copies**
Ag-mPIF1	AAAB01008964; 9,252,462-9,252,888	427 bp	93%	40	~60
Ag-mPIF6	AAAB01008987; 14,601,400-14,608,585	186 bp	93%	24	~100
TTA-II-Ag	AAAB01008839; 1,143,447-1,143,588	110 bp	93%	114	~100
Joey	AAAB01008948; 1,258,371-1,258,723	353 bp	89%	113	~250

* One element from each family is given as an example.
** Partial elements containing more than half of a complete element.

PIF-like elements in zebrafish and their relationship with *Tourist*-like MITEs

Like *A. gambiae*, zebrafish (*Danio rerio*) has a significant amount of available genomic sequence and contains a large number of *PIF*-like elements. Four fragments encoding catalytic domain of *PIF*-like TPases were identified from ~19Mb (~1%) of genomic sequence, indicating that there may be ~400 *PIF*-like TPases in the zebrafish genome. Two of the four fragments were nearly identical (99% a.a. identity), and the sequences flanking the TPases were compared to define the entire elements and identify their TIRs and flanking TAA TSDs. The two elements (AL591210, 80,099 – 84,118; AL591172, 43,142 – 47,591) were over 98% identical over their entire length (~4.2 kb) and serve to define a family designated *Dr-PIF1*. The terminal sequences of *Dr-PIF1* were used to search zebrafish genomic sequences for related MITEs. One MITE family was identified based on its sequence similarity to *Dr-PIF1* and designated *Dr-mPIF1*. The 48

complete *Dr-mPIF1* elements from 45 Mb (or 2.6%) of zebrafish genomic sequences were divided into three very similar but distinct subtypes (designated *Dr-mPIF1a*, *Dr-mPIF1b* and *Dr-mPIF1c*, **Table 2**), with elements of each subtype sharing the same diagnostic deletions/insertions. A comparison of *Dr-PIF1* to the consensus sequence of each subtype is shown in Fig. 4. At the 5' end, *Dr-mPIF1a*, *Dr-mPIF1b* and *Dr-mPIF1c* are 96%, 94% and 96% identical, respectively, to *Dr-PIF1* over 92 bp. At the 3' end, they are 95%, 91% and 94% identical, respectively, to *Dr-PIF1* over ~210 bp.

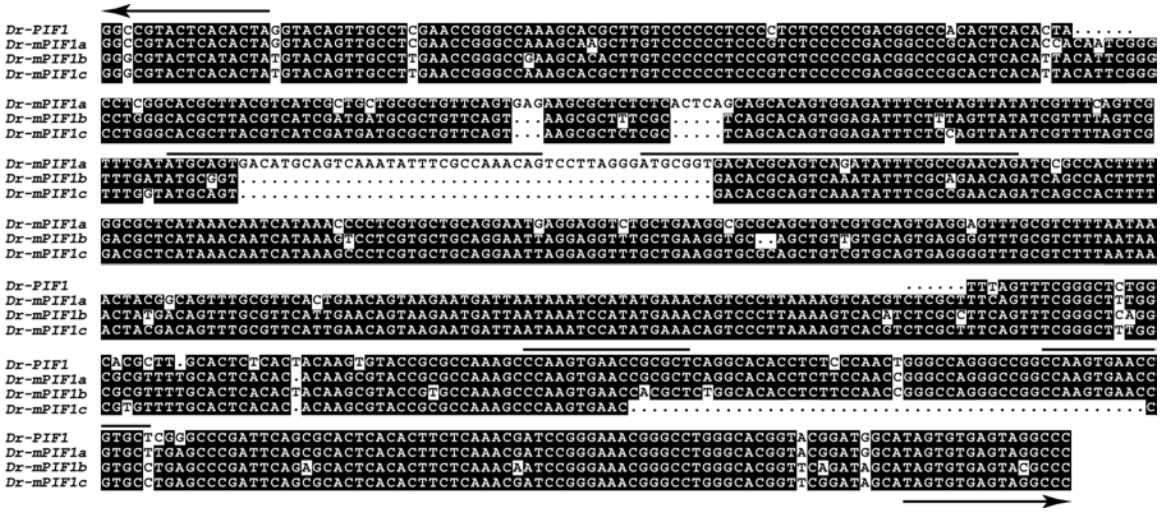
Table 2. Summary of the *Dr-mPIF1 Tourist*-like MITE families in Zebrafish.

Subtype	Acc. No. and Position*	Average Size	Average Pairwise Identity	No. of Complete Copies	Estimated Copy No. in Genome**
Dr-mPIF1a	AL808121; 85,975-86,660	673 bp	90%	22	831
Dr-mPIF1b	AL732436; 18,572-19,207	628 bp	83%	18	680
Dr-mPIF1c	AL929204; 116,392-116,987	579 bp	82%	8	302

* One element from each family is listed as an example.
** Copy numbers were estimated by dividing the numbers of complete elements by 45 (Mb of zebrafish sequences queried against) and multiply by 1,700 (zebrafish genome size in Mb).

The internal sequences are highly similar in all three *Dr-mPIF1* subtypes but are not present in *Dr-PIF1*, indicating that *Dr-mPIF1* MITEs were not derived directly from the two TPase-encoding *Dr-PIF1* elements. A search for elements that might have given rise to *Dr-mPIF1s* (e.g., by internal deletions) led to the identification of several longer elements ranging from 0.9 to 1.4 kb in length (not shown). These elements, however, all represented *Dr-mPIF1* MITEs with secondary insertions by other repetitive sequences. Thus, the cognate

Figure 4. Relationship between *PIF*-like elements and *Tourist*-like MITEs in zebrafish. A comparison between the terminal sequence of *Dr-PIF1* and the consensus sequences of three subtypes of *Dr-mPIF1* MITEs is shown. Arrows indicate element TIRs. Horizontal lines indicate the presence of direct repeats flanking the deletion breakpoint in MITEs.


Dr-PIF1
Dr-mPIF1a
Dr-mPIF1b
Dr-mPIF1c

element(s) of *Dr-mPIF1* MITEs was not in the sequenced portion of the zebrafish genome.

Comparison of the three subtypes of *Dr-mPIF1* MITEs revealed an interesting feature. As shown in Figure 4, *Dr-mPIF1a* contains a 45-bp region that is not present in the two shorter subtypes, and a 49-bp region is present in *Dr-mPIF1a* and *Dr-mPIF1b* but absent in *Dr-mPIF1c*. The presence of direct repeats (Fig. 4, indicated by horizontal lines) at the boundaries of these two regions indicates that their absence from the shorter subtypes is most likely due to deletions. Were *Dr-mPIF1b* and *Dr-mPIF1c* derived from *Dr-mPIF1a* by deletions and subsequently amplified? Should this be the case, the sequence divergence of the three subtypes should be *Dr-mPIF1a* > *Dr-mPIF1b* > *Dr-mPIF1c*. However, the exact opposite was observed: *Dr-mPIF1a* elements are most similar to each other (Table 2) and therefore appear to be more recently amplified than the other two subtypes. In addition, each subtype contains nucleotide substitutions and small insertion/deletions that are not present in other subtypes. Taken together, these data indicate that the three subtypes of *Dr-mPIF1* MITEs were derived independently from very similar yet distinct origins, and that the multiple deletion events could be involved in the formation of a MITE.

Conclusions

PIF-like elements in animals are diverse and can be divided into at least three monophyletic clusters (X, Y and Z). The abundance of *PIF*-like elements varies

significantly in animal species, from hundreds in *A. gambiae* and zebrafish to just a few (less than three) in *D. melanogaster*, *C. elegans* and human. A detailed characterization of *PIF*-like elements in *A. gambiae* defined eleven families, some of which included multiple elements that are highly similar (e.g., *Ag-PIF1*, *2*, *3* and *6*). Furthermore, TPases of *Ag-PIF1*, *4* and *8* families appear to be expressed. These results indicate that several families of *PIF*-like elements may be active in *A. gambiae*. Similarly, the high sequence identity shared by two *Dr-PIF1* elements (98%) suggests that this family of *PIF*-like elements may also be active.

Relationship between *PIF*-like elements and *Tourist*-like MITEs was analyzed in *A. gambiae* and zebrafish. Complete *PIF*-like elements were first defined and their terminal sequences were either compared to previously described *Tourist*-like MITEs or used to identified related *Tourists*-like MITEs that have not been reported. Clear-cut relationship was identified in several cases. Taken together, these results indicate that, similar to their plant homologs, *PIF*-like elements in animals are also responsible for the origin and amplification of *Tourist*-like MITEs.