Plant Mol Biol (2009) 71:99–114
DOI 10.1007/s11103-009-9511-0

Contrasting evolutionary patterns and target specificities among three *Tourist*-like MITE families in the maize genome

Tatiana Zerjal · Johann Joets · Karine Alix · Marie-Angèle Grandbastien · Maud I. Tenaillon

Received: 29 October 2008 / Accepted: 31 May 2009 / Published online: 17 June 2009

Abstract Miniature inverted-repeat transposable elements (MITEs) are short, non autonomous DNA elements that are widespread and abundant in plant genomes. The high sequence and size conservation observed in many MITE families suggest that they have spread recently throughout their respective host genomes. Here we present a maize genome wide analysis of three *Tourist*-like MITE families, *mPIF*, and two previously uncharacterized families, *ZmV1* and *Zead8*. We undertook a bioinformatic analysis of MITE insertion sites, developed methyl-sensitive transposon display (M-STD) assays to estimate the associated level of CpG methylation at MITE flanking regions, and conducted a population genetics approach to investigate MITE patterns of expansion. Our results reveal that the three MITE families insert into genomic regions that present specific molecular features: they are preferentially AT rich, present low level of cytosine methylation as compared to the LTR retrotransposon *Grande*, and target site duplications are flanked by large and conserved palindromic sequences. Moreover, the analysis of MITE distances from predicted genes shows that 73% of 263 copies are inserted at less than 5 kb from the nearest predicted gene, and copies from *Zead8* family are significantly more abundant upstream of genes. By employing a population genetic approach we identified contrasting patterns of expansion among the three MITE families. All elements seem to have inserted roughly 1 million years ago but *ZmV1* and *Zead8* families present evidences for activity of several master copies within the last 0.4 Mya.

Keywords Maize · MITE · *Tourist*-like · M-STD · Methylation · Evolution

Electronic supplementary material The online version of this article (doi:10.1007/s11103-009-9511-0) contains supplementary material, which is available to authorized users.

T. Zerjal · M. I. Tenaillon
CNRS, UMR 0320/UMR 8120, Génétique Végétale, F-91190 Gif-sur-Yvette, France

J. Joets
INRA, UMR 0320/UMR 8120, Génétique Végétale, F-91190 Gif-sur-Yvette, France

K. Alix
AgroParisTech, UMR 0320/UMR 8120, Génétique Végétale, F-91190 Gif-sur-Yvette, France

M.-A. Grandbastien
INRA, Laboratoire De Biologie Cellulaire, Institut Jean-Pierre Bourgin, 78026 Versailles cedex, France

T. Zerjal (✉)
UMR de Génétique Végétale, INRA/Univ Paris-Sud/CNRS/AgroParisTech, Ferme du Moulon, F-91190 Gif-sur-Yvette, France
e-mail: zerjal@moulon.inra.fr

Introduction

Transposable elements (TEs) are discrete DNA segments found in nearly all living organisms examined, and are particularly abundant in eukaryote genomes (Craig et al. 2002). TEs are traditionally divided into two classes (Wicker et al. 2007). Class I elements, known as retrotransposons, transpose via a RNA intermediate by a "copy and paste" mechanism and are distinguished in five orders on the basis of their internal genomic organization and mechanistic features. Class II elements, known as DNA transposons, transpose via a DNA intermediate and are divided into two subclasses, distinguishable by the number of DNA strands that are cut during transposition: both strands in subclass 1 and one strand in subclass 2. Subclass 1

TEs belonging to the TIR order comprise classical "cut-and-paste" TEs that present short Terminal Inverted Repeats (TIR) at their extremities and form, as most TEs, target site duplication (TSD) upon insertion. Both TIR and TSD sequences are used to distinguish nine different superfamilies, six of which are present in plant genomes (Wicker et al. 2007).

Class I elements of the LTR (Long Terminal Repeat) order are by far the most common elements in plant genomes and therefore greatly contribute to their evolution. Hence, genome size expansions have often happened through burst of LTR retrotransposons, and differences in genome size observed among closely related species are associated to variations in LTR retrotransposon content (Ma et al. 2004; Piegu et al. 2006). In maize, the genomic expansion due to LTR amplification has been particularly dramatic leading the genome to double in size within the last three million years (SanMiguel et al. 1998; Brunner et al. 2005). As a result, the genomic content in TEs greatly differs among maize inbred lines creating extended regions of non-homology (Brunner et al. 2005).

Intense TE activity in grass genomes is, however, not exclusive to LTR retrotransposons but also involved a particular group of Class II elements (subclass 1) called Miniature Inverted Repeat Transposable Elements (MITE). MITEs were originally discovered in plants and soon identified in many different organisms including fungi (Yeadon and Catcheside 1995), insects (Braquart et al. 1999; Tu 2000), nematodes (Surzycki and Belknap 2000), fishes (Izsvak et al. 1999) and humans (Smit and Riggs 1996). MITE elements are typically small in size (<600 bp), A/T-rich and have the potentiality to form single stranded secondary structures. These small non-autonomous Class II elements are internal-deletion derivatives of autonomous elements but usually exhibit a high copy number, which distinguish them from most previously described non-autonomous TIR elements (Feschotte and Mouches 2000). Because they have lost the gene encoding the protein transposase and hence their capacity to move, they rely on the transposase encoded by autonomous elements to successfully transpose. While sequence homology with the related active DNA transposons is loose, homology of the TIRs is an essential requirement for MITE transposition (Feschotte et al. 2005; Loot et al. 2006). A high number of MITEs identified in plant genomes belong to the *Tourist*-like and *Stowaway*-like superfamilies, based on TIR and TSD sequences (Bureau and Wessler 1992; Bureau and Wessler 1994; Feschotte et al. 2002). Because of their small size, the contribution of MITEs to genome size enlargement is less dramatic than LTR retrotransposons but they form the largest TE group in terms of copy numbers in grass genomes (Feschotte et al. 2002). For example, in rice although MITEs constitute only 6% of the genome, they form the most numerous group with over 100,000 elements belonging to hundreds of different families identified by sequence homology (>80%) (Jiang et al. 2003) and hundreds of thousands of elements are also present in the maize genome (Feschotte et al. 2002).

MITE elements do not insert randomly into the genome but have target site preferences. They prefer single copy regions (Bureau et al. 1996; Naito et al. 2006) to highly repeated ones, and *Tourist*-like MITEs present TSDs formed preferentially by the trinucleotide TAA (and its complement TTA), which is duplicated upon insertion (Bureau and Wessler 1992; Zhang et al. 2001; Yang et al. 2007). Moreover, two *Tourist*-like MITE families, *mPIF* in maize and *mPing* in rice, showed a significant insertion preference for a 9-bp palindromic extended target site centered on the TSD, indicating a more complex target site specificity (Zhang et al. 2001; Naito et al. 2006).

To explain the high sequence and size conservation identified in most MITE families, it has been proposed that MITEs have amplified recently from a very limited number of master copies following a "strict master model" of amplification (Feschotte et al. 2002; Jiang et al. 2004), which must have been very successful given the high copy numbers present in the host genomes. The characterization of the rice *Tourist*-like *mPing*, which is the first active MITE identified so far, has proved not only that MITEs are capable of transposition by a 'cut and paste' mechanism (Nakazaki et al. 2003; Yang et al. 2007) typical of Class II elements, but also that the rate of accumulation per generation is indeed very high (Naito et al. 2006).

Despite this abundance of MITEs in the genome of many organisms, mechanisms that govern MITE transposition and their evolutionary genomic history within their host genomes remain poorly understood. Insights on the evolution and the tempo of amplification of MITE families within their host genome have mainly come from genome-wide studies of rice and *Arabidopsis* genomes (Santiago et al. 2002; Jiang et al. 2004; Naito et al. 2006), however, no such studies have been performed so far on maize.

What we report here is the first genome-wide analysis of three *Tourist*-like MITEs in the maize genome using the 2,500 Mb of sequence currently available. We performed a structural characterization of target sites and a bioinformatic analysis of regions flanking MITE insertions, and developed methyl-sensitive transposon display (M-STD) assays to quantify CpG methylation level of MITE flanking regions. For comparison, M-STD was also performed on the LTR retrotransposon *Grande* (Garcia-Martinez and Martinez-Izquierdo 2003). Analyses of MITE flanking regions revealed that they target preferentially AT rich and low methylated genomic regions, and a large proportion of MITEs are inserted at less than 5 kb from the nearest

predicted gene. Moreover, in two of the MITE families, we also identified palindromic sequences flanking the TSD, which are the largest extended target sites so far identified for MITEs. We used a population genetic approach to search for signs of recent activities by investigating MITE patterns of expansion. We found contrasting patterns of expansion: while the 3 families originated around 1 million years ago, we detected recent activity of several master copies within the last 0.4 Mya in 2 out of 3 families.

Materials and methods

Mining of MITE copies

The BLAST program (Altschul et al. 1990) was used to search a total of 15,521 BAC sequences representing almost entirely the maize genome (2,580 Mb, release 1a.49). This sequence dataset was downloaded from the Maize Sequencing Project web server (http://ftp.maizesequence.org/current/). Three previously characterized MITE sequences found in the *Vgt1* region (Salvi et al. 2007), in the *Dwarf8* gene (Thornsberry et al. 2001), and the *mPIF*259 element (AF416307.1) from the *mPIF* family (Zhang et al. 2001), representative of three distinct *Tourist*-like MITE families, were used as queries. BLAST results were filtered using a perl script according to the following parameters: an expected value <1e-10, a minimum nucleotide identity rate of 85% and query coverage of 90%. MITE sequences as well as 50 flanking nucleotides on both sides of the MITE insertions were subsequently extracted from BAC sequences. All copies of a given family were aligned using the Clustal algorithm implemented in the MEGA 4.0 program (Tamura et al. 2007) and alignments were refined manually. Alignments are available upon request. Consensus sequences were constructed for each MITE family after removing sequences shared by more than 3 copies using the program Bioedit (Hall 1999) with a threshold frequency set to 60%.

MITE target sites and sliding window analysis

Ten nucleotides upstream and downstream the MITE TSDs were analyzed to identify extended target sites, which were visualized using the program Pictograms (http://genes.mit.edu/pictogram.html). The remaining flanking nucleotides were used to estimate the nucleotide content of MITE insertion flanking regions and used to test for biases in the nucleotide composition in extended target sites using a χ^2 test. MITE nucleotide content and sequence variability expressed by Nei's measure (Nei 1987) of nucleotide diversity, π, and its standard deviation, were calculated using the program DnaSP version 3.51 (Rozas and Rozas 1999). The same program was used to define conserved and variable portions of the MITE sequences by sliding window analysis using window size and step size of 20 nucleotides. Windows with diversity equal or higher than the average sequence diversity (π) plus 2 standard deviations (SD) were defined as variable. Those with diversity equal or less the average minus 2 SD were considered as conserved.

Estimation of MITE distances to the closest predicted genes

We downloaded the BAC annotation from genomesequence.org (http://www.maizesequence.org/index.html) and used as probes each MITE sequence including 40 nucleotides upstream and downstream the insertion, to precisely predict the MITE location within each BAC. Then we identified the contig harbouring the MITE and its coordinates within each BAC and extracted from the filtered gene set release 3a.50 (a subset of the Working Gene Set that has been filtered to remove pseudogene, TE-encoded genes, and partial models) the annotation of the corresponding contig. We then filtered the annotation to retain MITE proximal genes only. It is nevertheless important to bear in mind that this analysis is highly dependent on the maize annotation quality and was performed on unordered contigs, therefore the distance of MITEs to any predicted gene is biased by the contig length.

Methyl-sensitive transposon display assay

To investigate methylation profiles associated with MITE insertions, we employed the Methyl-Sensitive Transposon Display (M-STD) method, which combines features of the Methyl-Sensitive Amplification Polymorphism (M-SAP) technique (Xiong et al. 1999) with those of the Sequence-Specific Amplification Polymorphism (known also as Transposon Display) (Waugh et al. 1997; Casa et al. 2000) both modified versions of the amplified fragment length polymorphism (AFLP) (Vos et al. 1995; Xu et al. 2000). In brief, the M-STD reveals, by TE-anchored PCR of digested genomic DNA, polymorphisms generated by TE insertions. It also provides information on the methylation status of TE flanking regions by using methyl-sensitive isoschizomeric enzymes. *Eco*RI was employed as rare cutter enzyme. The isoschizomers *Hpa*II and *Msp*I, both recognizing the 5′-CCGG-3′ restriction site, were used alternatively in combination with *Eco*R1. *Hpa*II and *Msp*I differ in their sensitivity to methylation (Xiong et al. 1999): *Hpa*II activity is blocked when either cytosine (internal or external) is fully methylated, *Msp*I is blocked when the external cytosine is fully methylated or when it is hemimethylated (which is rarely encountered) and both

enzymes are blocked by full methylation of both cytosines. Different methylation states of the internal cytosine at the 5′-CCGG-3′ restriction sites will result in different cleavage by the isoschizomers, generating variable PCR band profiles between the two digests. For clarity, we will name M-STD(*Hpa*II) and M-STD(*Msp*I) the M-STD band profile obtained for either digestions.

The M-STD display was performed in four steps: digestion, ligation of adaptors, pre-amplification and selective-amplification. Genomic DNA (250 ng) was digested with 1 U of *Eco*RI and either 2 U of *Msp*I or 2 U of *Hpa*II for 3 h at 37°C. Digestion was followed by inactivation (15 min at 70°C). *Msp*I/*Hpa*II and *Eco*RI adaptor preparation (sequences in Table 1) and ligation were performed according to Xu et al. (2000).

MITE preamplifications were performed using a primer complementary to the *Msp*I/*Hpa*II adaptor and a primer complementary to a MITE-specific sequence (the "int" primers in Table 1). Reactions were done in 20 µl containing 1× PCR buffer, 4 µl of the diluted digested-ligated DNA (1:4 in water), 0.4 µM of each primer, 0.2 mM dNTPs, 1.5 mM $MgCl_2$, and 1 unit of *Taq* DNA polymerase. Cycling parameters were: 94°C/2 min followed by 26 cycles of 94°C/30 s, 58°C/1 min (56°C for *mPIF*), 72°C/1 min and a final cycle of 72°C/3 min. The preamplification of the retrotransposon *Grande* was performed using the *Msp*I/*Hpa*II adapter primer containing at the 3' end an extra C and the *Eco*RI adapter primer containing an extra A. PCR conditions were as described in Garcia-Martinez et al. (2003).

Selective amplifications were performed using MITE nested internal primers (the "ext" primers in Table 1), 5′ labelled with IRD-700 fluorescent dye (Sigma-Genosys) and the *Msp*I/*Hpa*II primer 3'-overhung by one selective base (a C for *ZmV1* and *Zead8* and a G for *mPIF*). Amplification mixture was performed as described above from 5 µl of the diluted preamplification products (1:10 in water) and 0.25 µM of primers. The cycle conditions were 94°C/2 min, 12 touchdown cycles of 94°C/30 s, 68°C/1 min (−0.7°C each cycle), 72°C/1 min, followed by 23 cycles of 94°C/30 s, 58°C/1 min, and 72°C/1 min, and final extension at 72°C/3 min. For the LTR retrotransposon *Grande* the amplification mix was as described above using the *Msp*I/*Hpa*II primer 3'-overhung with three selective bases (CGT), and the other primer complementary to the *Grande* LTR (Table 1). Cycle conditions were identical to Garcia-Martinez and Martinez-Izquierdo (Garcia-Martinez and Martinez-Izquierdo 2003).

Amplification products were diluted (1:20) in loading buffer (94% formamide, 0.5 mg/ml bromophenol blue), and migrated, after denaturation (5 min at 95°C) in a 40 cm 5% denaturing (6 M urea) long range acrylamide gel (BMA, Rockland, ME, USA) in 1× TBE. The electrophoresis was performed in a LI-COR DNA analyzer (LI-COR, Lincoln, NE, USA) at 2,000 V for 6 h at 50°C, using the LI-COR 50–700 bp size standard as internal ladder. In order to test the reproducibility of the M-STD method and to confirm band polymorphisms between the M-STD(*Msp*I) and M-STD(*Hpa*II), each experiment was repeated three times independently and the products loaded on the same sequencing gel.

Only clear and reproducible bands were scored. For the three MITE families *Msp*I/*Hpa*II adapter primers containing all four selective bases (A, C, G, T) were used in the selective amplification for M-STD(*Msp*I). Then the selective base that gave the highest number of bands was used for the M-STD analysis (C for *ZmV1* and *Zead8* and G for *mPIF*). For the retrotransposon *Grande* we tested in M-STD(*Msp*I) 4 combinations of 3 selective nucleotides (CGG, CGT, CGC and CTA) and retained the selective bases that gave the highest number of bands for the M-STD analysis (CGT).

Table 1 Primer sequences

Primer name	Primer sequence 5′–3′[a]
*Msp*I/*Hpa*II adapter1[b]	GACGATGAGTCTAGAA
*Msp*I/*Hpa*II adapter2[b]	CGTTCTAGACTCATC
*Msp*I/*Hpa*II Primer[b]	GATGAGTCTAGAACGG
*Eco*RI adapter 1[c]	CTCGTAGACTGCGTACC
*Eco*RI adapter 2[c]	AATTGGTACGCAGTCTAC
*Eco*RI Primer[c]	GACTGCGTACCAATTC
*mPIF*_Int	CATTARTAAGATTYYAATTCCT
*mPIF*_Ext	AATTCCTCAAAATGAAAGGAAACA
*ZmV1*_Int	CRATCCCRCTCAATCCAC
*ZmV1*_Ext	TCCACATGGATTGAGAGCTAA
*Zead8*_Int	CCCCATGAAACTCCATGAAA
*Zead8*_Ext	AGCTGATGTGGCAGGCTAAT
LTR-G1[d]	CTTGGGCCTTTCGTGAG

[a] 3′-overhung selective bases of the *Msp*I/*Hpa*II and *Eco*RI primers are not included in the primer sequence and details about the selective bases used are reported in Material and Methods. R = A/G and Y = C/T

[b] As in (Xu et al. 2000)

[c] As in (Vos et al. 1995)

[d] As in (Garcia-Martinez and Martinez-Izquierdo 2003)

M-STD data analysis

M-STD was applied to a set of 10 maize inbred lines (CL187-2, F2834T, HP301, Ky21, LAN496, Mo22, N25, NY302, W85, ZN6), which DNAs were extracted as described in (Tai and Tanksley 1991) with minor modifications. Only clearly identifiable and reproducible bands were manually scored as methylated or unmethylated.

Three classes of bands were identified. The first class included bands present in both M-STD(*Msp*I) and M-STD(*Hpa*II) profiles and were scored as unmethylated. The second included bands present in the M-STD(*Msp*I) but missing from the M-STD(*Hpa*II) and were considered methylated. The third class of rarely appearing bands corresponded to those present in M-STD(*Hpa*II) but missing from the M-STD(*Msp*I). These bands have multiple possible origins. They are either due to the hemimethylation of the external cytosine (that blocks the activity of *Msp*I but not of *Hpa*II) or may result from length polymorphism when two neighbouring CCGG sites harbour different methylation status. For instance, when the closest CCGG site to a TE insertion is methylated and the neighbouring one is not, one expect to obtain a longer fragment for M-STD(*Hpa*II) than for M-STD(*Msp*I). Because of the ambiguity on the origin of these bands, we did not take them into account in the estimation of methylation levels. M-STD band profiles were analyzed separately for each TE (*mPIF, ZmV1, Zead8* and *Grande*). Differences in proportion of methylated and unmethylated bands were tested by chi-square tests.

MITE intrafamily diversity analysis

To investigate MITE intrafamily diversity we constructed median-joining (MJ) networks using the network algorithm (Bandelt et al. 1999) implemented in the program Network 4.5 (www.fluxus-engineering.com/sharenet.htm), which was originally designed to deal with recurrent mutation. As input data we used aligned sequences of MITE copies of a given family assuming no recombination among TEs. The network method allows combining all minimum-spanning trees (where the total branch length is the minimum necessary to connect all copies) from a given data set, within a single network, which represents the minimum-spanning network of the data set. A median-joining network consists of nodes and links connecting the nodes. The nodes are MITE sequences from the data set; the links connecting the nodes represent genetic distances and are calculated by the number of nucleotide differences (characters) between sequences. To account for differences in informativeness of characters (i.e. frequencies of changes for each nucleotide position) the program allows applying a nucleotide specific weighting scheme. We used a weighting scheme ranging from 2 to 20 accordingly to each nucleotide frequency of change, assigning higher weights to the informative sites and lower weights to the highly variable sites. To simplify the final network output we used the maximum parsimony (MP) option.

Within each MITE family, the frequency distribution of the number of differences between pairs of sequences (pairwise differences), and raggedness index, *r*, were computed in Arlequin version 3.11 (Schneider et al. 2000). In order to compare frequency distributions of pairwise differences among MITE families characterized by various lengths, we normalized the number of pairwise differences by the length of the consensus sequence in bp and expressed it as a percentage. We therefore obtained comparable frequency distributions of the percentage of pairwise differences (called hereafter mismatch distributions) for each MITE family.

Age estimates of MITE families

To estimate the age of MITE families, we first calculated the average level of nucleotide substitutions (K) existing between each MITE element and the family consensus sequence using Kimura 2-parameter distance (Kimura 1980) with a transition/transversion ratio of 2. The average MITE family age was estimated using the formula $T = K/\mu$ (Kapitonov and Jurka 1996; Jiang et al. 2002), assuming $\mu = 3.3 \times 10^{-8}$ (upper limit $\mu = 5.1 \times 10^{-8}$ and lower limit $\mu = 2.0 \times 10^{-8}$) and corresponding to the maize intergenic-region substitution rate per site per year (Clark et al. 2005). For *ZmV1*, we estimated the age of each cluster identified by the network analysis, considering the most frequent sequence within each cluster as the consensus sequence.

Results

Description of three MITE families

mPIF, *ZmV1* and *Zead8* are the three MITE families analyzed in this study and were chosen because they are potentially active (Tables 2 and 3). The *mPIF* family has been previously described as containing potentially active MITEs that share several common features with *P Instability Factor* (*PIF*) elements, including identical 14-bp TIRs, similar subterminal sequences and a strong preference for insertion into a 9-bp palindromic extended target site surrounding the TSD (Walker et al. 1997; Zhang et al. 2001). The second family, *ZmV1*, whose first element was discovered in the *Vgt*1 region (Salvi et al. 2007), is characterized by having short length (<140 bp) and low intrafamily nucleotide diversity ($\pi = 0.049$), and exhibits almost identical 14-bp TIRs and 50% internal sequence identity with some *Tourist-Zm* MITE elements (Bureau and Wessler 1992). The third family, *Zead8*, whose first element was discovered in the *Dwarf*8 gene, contains all characteristics of the *Tourist*-like superfamily including 14-bp TIRs that are 90% identical to the rice actively transposing *Tourist*-like MITE *mPing* and the TTA/TAA TSD. However, computer-based sequence similarity with

Blastn in Plant Repeat Databases (http://plantrepeats.plantbiology.msu.edu/index.html) (Ouyang and Buell 2004) did not reveal any sequence homology to characterized MITE family.

Structural characterization of three *Tourist*-like MITE families

Using sequence similarity search (BLAST) we identified 246 *mPIF*, 250 *ZmV1* and 109 *Zead8* sequences within the public genome sequences of *Zea mays*. When several MITE insertions were homologous in their flanking regions but mapped to different positions, we decided to keep only one of them in the alignment and discarded the others. Such homology may be caused by insertion within other TEs that may have transposed. We also discarded the insertions containing large internal gaps. In total, 209, 211 and 79 copies were aligned for *mPIF*, *ZmV1* and *Zead8* respectively (Table 2).

To confirm the presence of the previously described *mPIF* 9-bp palindromic extended insertion site (Zhang et al. 2001; Naito et al. 2006) in our *mPIF* dataset and to investigate the existence of palindromic insertion sites associated with the two other MITE families, we extended the analysis to 10 nucleotides upstream and downstream MITE TSDs. After retrieving only copies with intact TSDs, we analyzed a total 200 *mPIF,* 211 *ZmV1* and 79 *Zead8* insertion sites. For *mPIF* we recognized the 9-bp insertion site previously described but found that the palindromic motif could be extended to 19 bp centered on the TSD (Fig. 1a). In *ZmV1* we identified a palindromic insertion site of 17 nucleotides centered on the TSD (Fig. 1b). For *Zead8* no palindromic sequence was identified beyond the TSD when the full dataset was analyzed, but when recently inserted copies (see below, cluster 1 of the network analysis) were analyzed separately, a 15 nucleotide palindromic sequence centered on the TSD (Fig. 1c) was identified. In this sequence the 3 bases upstream and downstream the TSD corresponded to the target-site sequence identified for the *mPing* MITE in rice (Naito et al. 2006).

We further analyzed base composition in the flanking regions and extended target sites for *mPIF* and *ZmV1*. Both are AT-rich regions: flanking regions contain 63 and 55% of AT and palindromic sequences contain 70 and 54% of AT, for *mPIF* and *ZmV1* respectively. We referred to the average base composition of the flanking region (excluding the extended palindromic target site) to test for bias in nucleotide frequency at each site of the palindromic sequence. Of the 19 bp *mPIF* and the 17 bp *ZmV1* extended insertion sites, 19 and 15 bp, respectively, were significantly biased as estimated by a χ^2 test ($P < 0.01$). In contrast, we did not detect any nucleotide frequency bias at the nucleotide positions beyond the palindromic site (data not shown). Specific positions showed a strong bias towards C or G (Fig. 1b). For example, *mPIF* exhibited a C at position -3 and a G at position +3 in 73 and 68% of the sequences, respectively. Similarly, *ZmV1* exhibited a strong bias toward C at position −4 and −5 (69 and 56% of the sequences, respectively) and toward G at position +4 and +5 (58 and 51% of the sequences respectively). This bias was even stronger when only flanking regions of recently inserted copies were analyzed (i.e. MITE copies within clusters, see network analysis results). Among cluster 2 sequences, a C was present at position −4 and −5 in 76% (22 out of 34) and 65% (26 out of 34), respectively, and in cluster 3, 77% (14 out of 22) and 64% (17 out of 22) respectively.

Table 3 Nucleotide diversity and age estimates

Name[a]	π	Kimura 2 distance[b]	Average age estimate[c]
mPIF	0.067	0.038	1.10 (0.7–1.9)
ZmV1	0.049	0.031	0.95 (0.5–1.5)
ZmV1 cluster 1	0.017	0.0094	0.29 (0.17–0.5)
ZmV1 cluster 2	0.017	0.0095	0.29 (0.17–0.5)
ZmV1 cluster 3	0.022	0.011	0.32 (0.19–0.5)
ZmV1 cluster 4	0.024	0.013	0.40 (0.25–0.6)
Zead8	0.065	0.040	1.20 (0.7–2.0)
Zead8 cluster 1	0.021	0.012	0.37 (0.2–0.6)

[a] *mPIF*, *ZmV1* and *Zead8* refer to the full dataset of each MITE family. The four *ZmV1* clusters and *Zead8* cluster 1 include exclusively copies falling within the corresponding clusters identified by network analysis

[b] Average Kimura 2-parameter distance

[c] Average age estimate using Kimura 2-parameter distance expressed in million of years

Table 2 MITE families' features

Name	Length	Nb. of elements	AT content[a]	AT content[b]	TIR sequences 5′-3′	TSD
mPIF	358 ± 3 bp	209	72%	63%	GGCCCCATTTGTTT	TAA/TTA
ZmV1	137 ± 3 bp	211	57%	55%	GGGCTTGTTCGGTT	TAA/TTA
Zead8	365 ± 6 bp	79	64%	56%	GGCTATTCACAATG	TAA/TTA

[a] Average AT content of MITE sequences

[b] Average AT content of MITE insertion surrounding regions calculated on 74 flanking nucleotides (excluding the extended target sites)

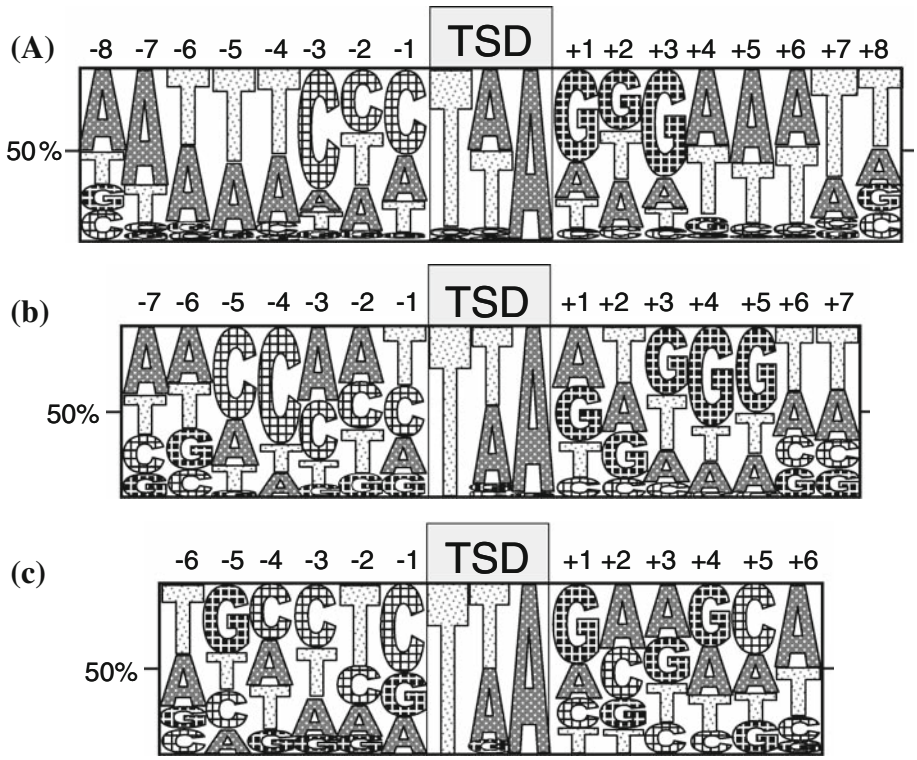

Fig. 1 Pictograms of the extended target sites as obtained from 200 *mPIF* elements (**a**), 211 *ZmV1* elements (**b**) and 31 *Zead8* elements belonging to cluster 1 (**c**). Numbers on top of each figure indicate the nucleotide position from the TA/TA TSD, indicating with "−"and "+"the positions at the 5′ and 3′ end of the TSD, respectively. The height of each letter is proportional to the relative frequency of each nucleotide at that position and frequencies decrease from the top to the bottom of the pictogram

To define variable and conserved regions among MITE copies within the 3 families, we performed a sliding window (Fig. 2) analysis using π, the average number of pairwise differences, as a nucleotide diversity estimate. In *ZmV1* the two windows containing 5′ and 3′ TIRs are the most conserved, with values below the average π value minus 2 SD and the rest of the windows are equal or above the average π value (Fig. 2a). In *mPIF* the windows containing the TIRs have π values very close to the average, while the central part of the sequence is more variable and characterized by higher π values, and the final 3′ end more conserved (Fig. 2b). In *Zead8* the most variable regions contain the 5′ and 3′ TIRs respectively, while the rest of the regions are overall conserved (Fig. 2c). A similar result was also obtained when MITE copies belonging to *Zead8* cluster 1 were analyzed separately (figure not shown). This analysis pinpointed conserved regions within families whereas no common pattern in the distribution of nucleotide diversity among the three MITE families emerged.

Positioning of MITEs with respect to predicted genes

We calculated the distance separating each MITE insertion from the nearest predicted gene by analysing MITE insertion flanking regions, for the full contig size. In total, we found that 263 MITE copies harbour a gene in their flanking regions (Table 4). This number must be considered as a lower bound first because of the small size of a large number of contigs (i.e. 117 of the 232 MITEs for which we did not find any gene in the flanking regions are in contigs smaller than 20 kb); second because of the "edge effect" due to insertion of MITEs in the vicinity of a contig end (i.e. 151 of the 232 MITEs were located at less than 5 kb from either the 5′ or 3′ contig end). Indeed, for the 263 MITEs for which we could estimate the distance from the nearest predicted gene, the average contig size was of 47.5 kb and the MITE average distance from either the 5′ or 3′ contig end was of 24 kb. Among these 263 MITE copies, 44 were located within a gene, most of which lied in introns (84.4%) as compared to exons (15.6%) and 73% were found at less than 5 kb to the nearest predicted gene (Table 4). Interestingly the distribution of MITE distances to the nearest predicted gene varied among families. For example in *mPIF*, we were able to estimate the distance to the nearest gene for 88 copies (42% of all *mPIF* copies analyzed) and found that 21.6% of these lied within a gene, while 15.9% were found at less than 500 bp and similar frequencies were found also for *ZmV1*. For *Zead8*, we identified genes in the flanking sequences of 55 copies (70% of all *Zead8* copies analyzed) of which only 5.7% were within a gene, while 27.3% were found at less than 500 bp (see Table 4). The distribution of MITE insertions in upstream and downstream regions of genes also varied among MITE families. For *mPIF* and *ZmV1* copies inserted at less than 500 bp from a gene, we did not detect a significant difference between the number of copies located upstream or downstream of genes (6 and 8 copies respectively for *mPIF* and 9 and 8 copies respectively for *ZmV1*). In contrast, *Zead8* copies were preferentially inserted upstream

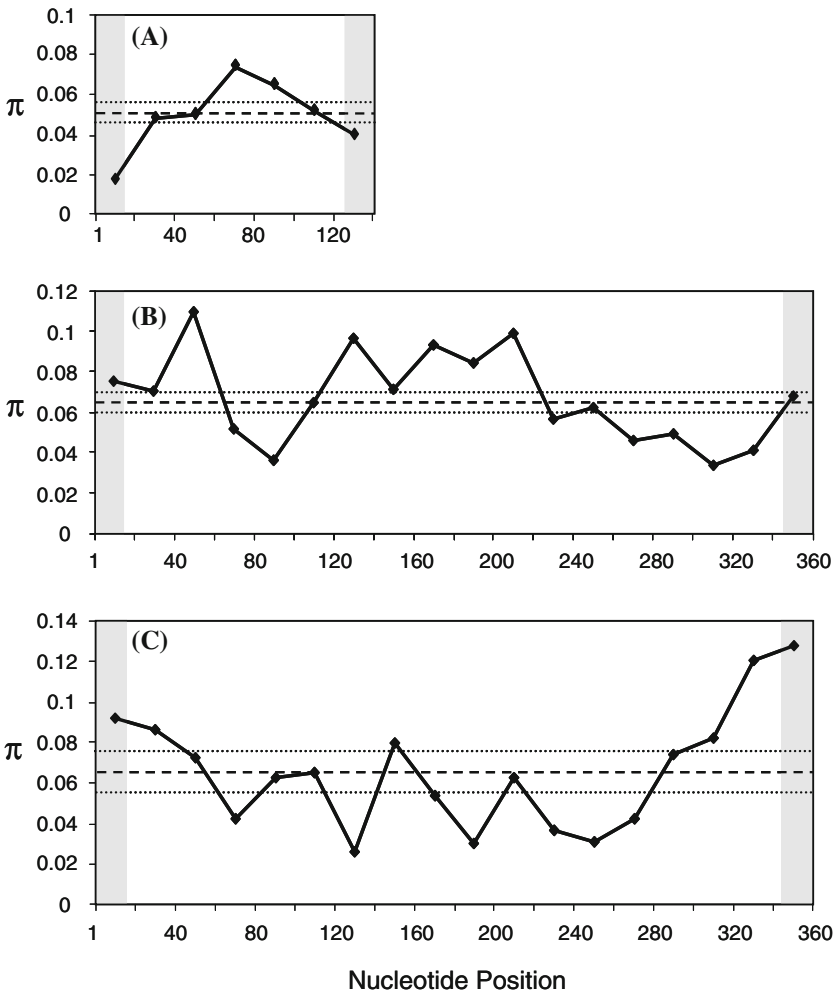

Fig. 2 Sliding window analysis of *ZmV1* (**a**), *mPIF* (**b**) and *Zead8* (**c**) sequences. Average nucleotide diversity (π) is indicated with a *dashed line*; while the average diversity ± 2 SD is indicated with a *dotted line*. Regions containing the TIRs are indicated in *grey*. Conservation of TIR sequences is evident only in the *ZmV1* MITE family (**a**), while they stand as the most variable portions in the *Zead8* MITE family (**c**)

of genes (12 out of 15 copies, χ^2 test $P < 0.03$). A similar trend was also observed when considering MITE copies located within 5 kb from the nearest gene.

Methylation status of MITE flanking regions

Another property of MITE flanking regions that we were interested in investigating was the level of cytosine methylation. We optimized the Methyl-Sensitive Transposon Display (M-STD) technique for the three MITE families and the LTR retrotransposon *Grande* (Garcia-Martinez and Martinez-Izquierdo 2003). The isoschizomers *Hpa*II and *Msp*I that display differential sensitivity to cytosine methylation were used to detect differences in methylation patterns. *Hpa*II activity is inhibited if either cytosine is fully methylated, while *Msp*I is inhibited by methylation at the external cytosine only. The M-STD method, even if assessing the CpG methylation status of internal cytosines

Table 4 MITE distance to the nearest predicted gene

MITE family	N[a]	Distances in bp					
		0 (within gene)[b]	<500	500–1,000	1,000–5,000	>5,000	n.d.[c]
mPIF	88	21.6% (19)	15.9% (14)	5.7% (5)	26.1% (23)	30.7% (27)	120
ZmV1	120	18.3% (22)	14.2% (17)	10% (12)	30% (36)	27.5% (33)	90
Zead8	55	5.4% (3)	27.3% (15)	16.4% (9)	31% (17)	20% (11)	22
Total	263	16.7% (44)	17.5% (46)	9.9% (26)	28.9% (76)	27% (71)	232

[a] Number of MITEs for which we could determine the distance to the closest gene

[b] Percentage (number) of MITEs for each class of distance

[c] n.d. = not determined. It refers to the number of MITEs for which we could not determine the distance to the nearest gene

at 5′CCGG 3′ sites only, has the advantage to be highly efficient for large-scale detection of methylation levels at TE flanking sequences. Indeed, changes in the PCR band profiles between the two enzymatic digestions of the same DNA sample reflect different CpG methylation state of the restriction sites that are in the proximity of TEs (example in Fig. 3).

In total we scored 98 bands for *mPIF*, 191 for *ZmV1*, 50 for *Zead8*, and 117 for *Grande*. In Fig. 4 are shown the proportion of methylated and unmethylated bands obtained for each TE. We observed significant differences between the number of methylated and unmethylated bands for each MITE family and for the LTR retrotransposon *Grande* (*mPIF*: $\chi^2 = 4.1$, $P = 0.04$; *ZmV1*: $\chi^2 = 110.6$ $P = 7.0\text{E}^{-26}$; *Zead8*: $\chi^2 = 25.92$, $P = 3.5\text{E}^{-07}$; *Grande*: $\chi^2 = 30$ $P = 4\text{E}^{-08}$), and a striking difference in the proportion of methylated bands between MITE families and the LTR retrotransposon. Yet, the number of unmethylated bands in the three MITE families ranged from 60 and 88%, while only 25% of the bands were unmethylated in LTR retrotransposon *Grande* (Fig. 4). Among the MITE families, *mPIF* exhibited relatively high methylation level (40% of methylated bands) while *ZmV1* and *Zead8* (12 and 14% of methylated bands respectively) exhibited level of methylation comparable to the average maize genomic methylation level (14.4% of the sites) as estimated by M-SAP (Lu et al. 2008).

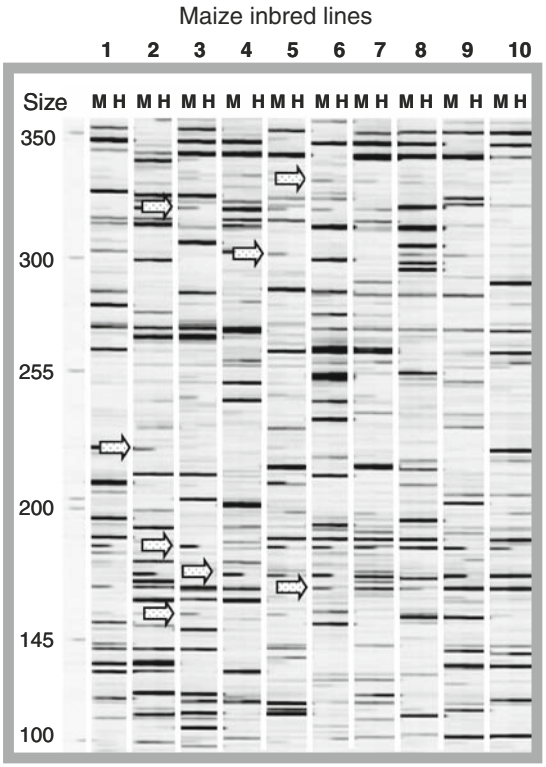

Fig. 3 M-STD gel for *ZmV1* using 10 maize inbred lines (1 = CL187-2, 2 = F2834T, 3 = HP301, 4 = Ky21, 5 = LAN496, 6 = Mo22, 7 = N25, 8 = NY302, 9 = W85, 10 = ZN6). For each maize inbred line the M-STD(*Msp*I) and the M-STD(*Hpa*II) (noted as M and H respectively) were loaded side by side to compare M-STD band profiles. Methylated fragments, indicated with *arrows*, are those that are present in the M-STD(*Msp*I) but missing from M-STD(*Hpa*II). Only one *arrow* per locus is indicated but several inbreds may exhibit a similar pattern. The 50–700 bp sizing standard is indicated on the *left*

Diversity patterns and time estimates within MITE families

We used a population genetic framework in order to get insights into the evolutionary history of the three MITE families. Phylogenetic analyses of TE DNA sequences are generally conducted using hierarchical bifurcating trees. Because our MITE datasets are characterized by: (i) few phylogenetically informative characters, (ii) the presence of recurrent mutations, (iii) the probable persistence of ancestral copies and of multiple descendents from single ancestors, such analyzes are not recommended (Posada and Crandall 2001). Instead, we opted for using networks to study MITE intrafamily diversity. Networks are connecting nodes with branches whose length is related to the amount of evolutionary changes among MITE sequences (copies). The shape of the network, determined by the spatial distribution of the nodes and the length of the branches is an indicator of past demographic events (Saillard et al. 2000; Jobling et al. 2004). For instance, an expanded star network, characterized by numerous nodes distributed around its centre and separated by long branches, most likely results from an old population expansion. Similarly, within a network, a strict star-shaped cluster defined by a high frequency central node (i.e. a single sequence shared by many copies) surrounded by many, almost identical, low frequency nodes, is indicative of a very recent expansion

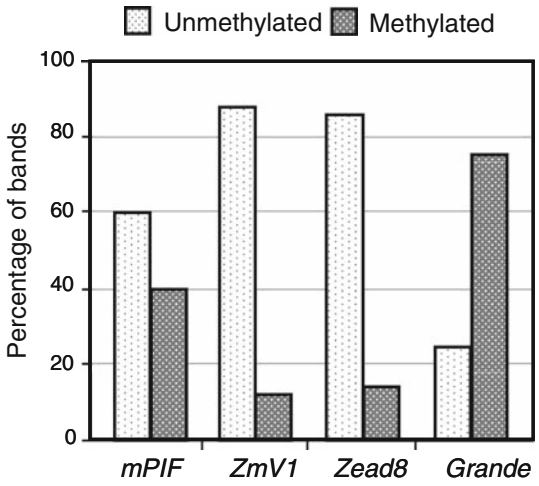

Fig. 4 Comparison of the fraction (%) of methylated and unmethylated bands obtained for the four TE families

from an ancestral element, i.e. the central node (Jobling et al. 2004).

The network analysis revealed very different network shapes for the three MITE families (Figs. 5, 6, 7). For *mPIF* (Fig. 5) the network had an expanded star shape. No nodes were present at the centre of the network and all of them were distributed around it, separated by long branches of similar length (indicating similar amount of nucleotide differences). In contrast, for *ZmV1* (Fig. 6) we obtained an expanded star shape network, but it differed from *mPIF* network by the presence, in the centre of it, of four strict star-shaped clusters of copies referred as cluster 1, cluster 2, cluster 3 and 4 (Fig. 6). These are distinguishable by the presence of a central high frequency element sequence (represented by the large size central circle) surrounded by many unique or almost unique sequences (small circles) separated by fewer nucleotidic changes. The *Zead8* network is instead divided in two portions, probably representing two distinct subpopulations of copies separated by several informative sites. The majority of the network is characterized by distantly related sequences connected by long branches and representing the older copies of this family. The portion of the network included in the dashed oval is instead characterized by the presence of closely related, but not identical, sequences, consistent with a moderate recent activity probably of several master copies (Fig. 7).

To further investigate the diversity patterns within MITE families, we analyzed the pairwise distribution for each MITE family. Studies on pairwise distribution have shown that the shape of the distribution is influenced by episodes of population expansion (Slatkin and Hudson 1991; Rogers and Harpending 1992). A smooth bell-shaped pairwise distribution indicates a rapid population expansion from a single master element, whereas a ragged, multimodal distribution indicates a population whose size has been constant over a long period of time. For *mPIF* and *ZmV1* the observed distributions are bell-shaped as shown also by the very low raggedness values (0.002 and 0.0122, respectively) (Fig. 8). The two curves, however, differ by having a different distribution mode. For *ZmV1* the mode is comprised between 2 and 3% of pairwise differences with over 32% of the observed values falling in it. For *mPIF*, instead, the mode is comprised between 6 and 7% of pairwise differences comprising almost 30% of the observed values. Compared to *ZmV1*, the distribution of *mPIF* is therefore skewed towards elevated values of the percentage of pairwise differences as further evidenced by the absence of observations in the lowest class (0–1). A different distribution was obtained for *Zead8*. It did not exhibit a smooth unimodal pattern but rather a bimodal distribution characterized by two picks, one centered between the 3 and 4% of pairwise differences and the other between the 8 and 9% (Fig. 8).

The average age of a MITE family considered to have derived from one or very few closely related master copies can be determined from the level of sequence divergence accumulated among copies since their insertion. After their insertions, we expect most of these copies not to be functionally constrained and to accumulate mutations at a neutral rate. Thus, the average divergence between the members of a given family from their respective family consensus sequence, provide an estimate of the age of the family. The consensus sequence is considered here as the best proxy for the common ancestor of the family. The average age estimates of the three MITE families are summarized in Table 3 and they are all very similar being roughly 1 Mya. Signs of recent activity were detected in both *ZmV1* and *Zead8* (Table 3). In *ZmV1* an intense activity of several master copies seem to have occurred within the last 0.4 Mya and led to the formation of the 4 clusters of copies as revealed by the network analysis (Fig. 6). In *Zead8* the amplification of few master copies led to the formation of a subpopulation of closely related sequences within the last 0.37 Mya. We also calculated ages from the average pairwise nucleotide diversity for each MITE family, without inferring a consensus sequence and found similar age estimates (data not shown).

Discussion

In the present study we used population genetics and experimental approaches to investigate the structure and evolutionary patterns of three maize *Tourist*-like MITE families, *mPIF*, *ZmV1* and *Zead8*. Previously it has been shown that MITEs do not integrate randomly into their host genome but insert preferentially in single copy regions. In these regions they target specific short sequences (TSD) which are duplicated upon insertion (Bureau et al. 1996; Naito et al. 2006) and in few cases an extended 9 bp palindromic target site has also been identified (Zhang et al. 2001; Naito et al. 2006). Our analysis of 200 *mPIF* and 210 *ZmV1* insertion sites revealed high specificity for larger target sites (Fig. 1a, b). In *mPIF* we identified a palindromic target site twice as large as previously reported (Zhang et al. 2001), extending for 19 nucleotides (Fig. 1a). In *ZmV1* we identified a new palindromic target sequence of 17 nucleotides (Fig. 1b). Such palindromic motifs associated with a high level of nucleotide conservation at specific positions likely reflect their role in target site specificity, further suggesting that MITE target preference is more complex than previously thought.

Preference for specific target sites has already been described for few transposons (Craig 1997; Ketting et al. 1997; Mahillon and Chandler 1998; Liao et al. 2000) but they were generally small in size and poorly conserved. To

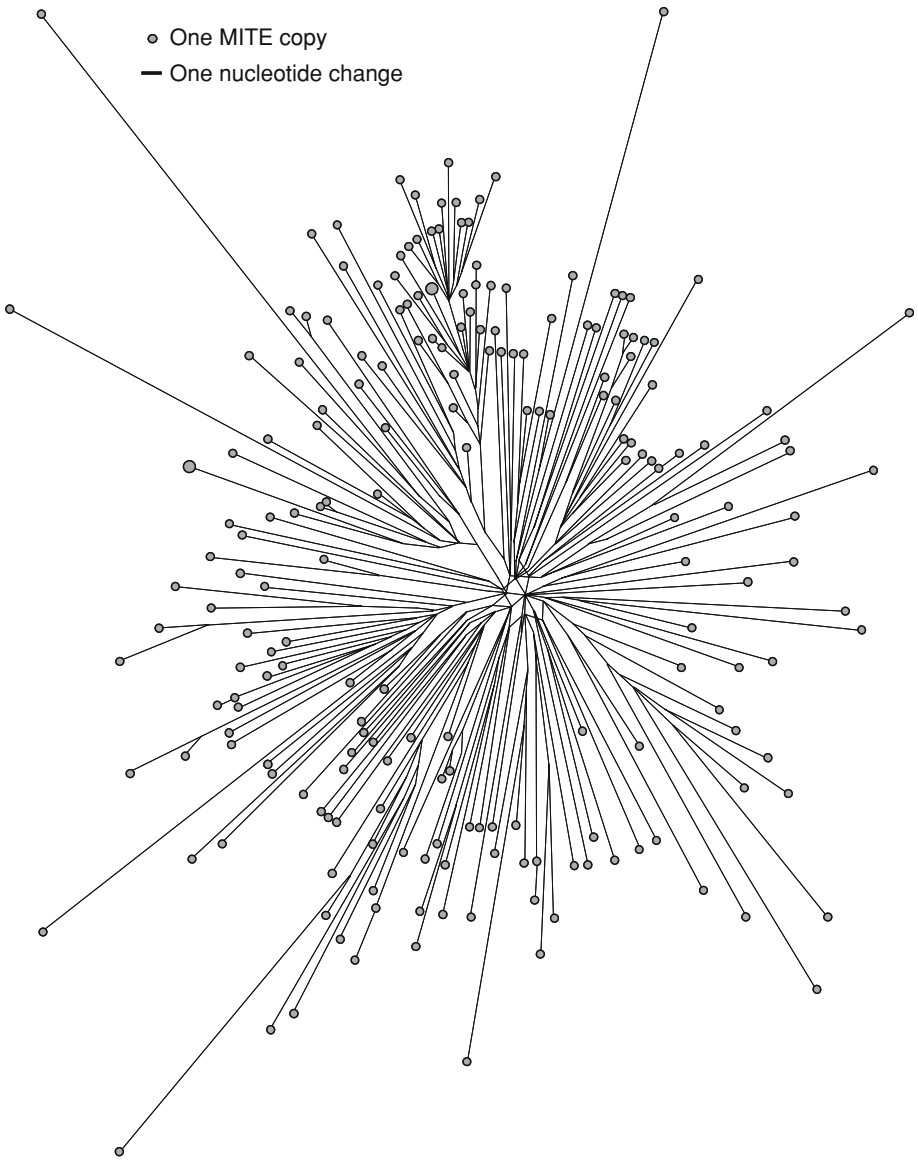

Fig. 5 Median-joining network of *mPIF* copies. Each *circle* represents a TE *sequence* and the *circle* area is proportional to the number of identical sequences found in our sample. Similarly, branch length is proportional to the number of nucleotide changes

our knowledge, there is so far only one TE, from the *PIF/Harbinger* superfamily, that presents an extended target site similar in size (17 bp) and specificity to the one we are describing here (Kapitonov and Jurka 2004). It is worth noticing that members of the *PIF/Harbinger* superfamily are responsible for the origin and amplification of *Tourist*-like MITEs (Zhang et al. 2004). Therefore, the presence of extended palindromic target sites associated with *Tourist*-like MITE families may be common, and provides indirect evidences that different transposons of the *PIF/Harbinger* superfamily are responsible for the recent amplification of these *Tourist*-like MITEs.

If target sites indeed guide MITE insertions, we would expect a higher level of target site conservation in recently inserted copies. Yet, analysis of flanking regions of recently inserted *ZmV1* copies (copies belonging to the four *ZmV1* clusters) revealed a higher degree of nucleotide conservation than the full dataset. Interestingly, while no palindromic motif was detectable in *Zead8* when analyzing the full dataset, we identified such a motif in a young subpopulation of copies (Cluster1 in Table 2 and Fig. 7). This motif was very similar to the target site identified for *mPing* MITE family in rice, which transposition is guaranteed by the active transposon *Pong,* supporting the hypothesis that the related *Zead8* transposon may be a *Pong*-like transposon.

In addition to palindromic motifs, MITE insertions are also affected by the presence of a specific genomic environment determined by the nucleotide composition and methylation level. Molecular characterization of regions surrounding MITE insertions in *Arabidopsis thaliana* and *Caenorhabditis elegans* (Surzycki and Belknap 2000;

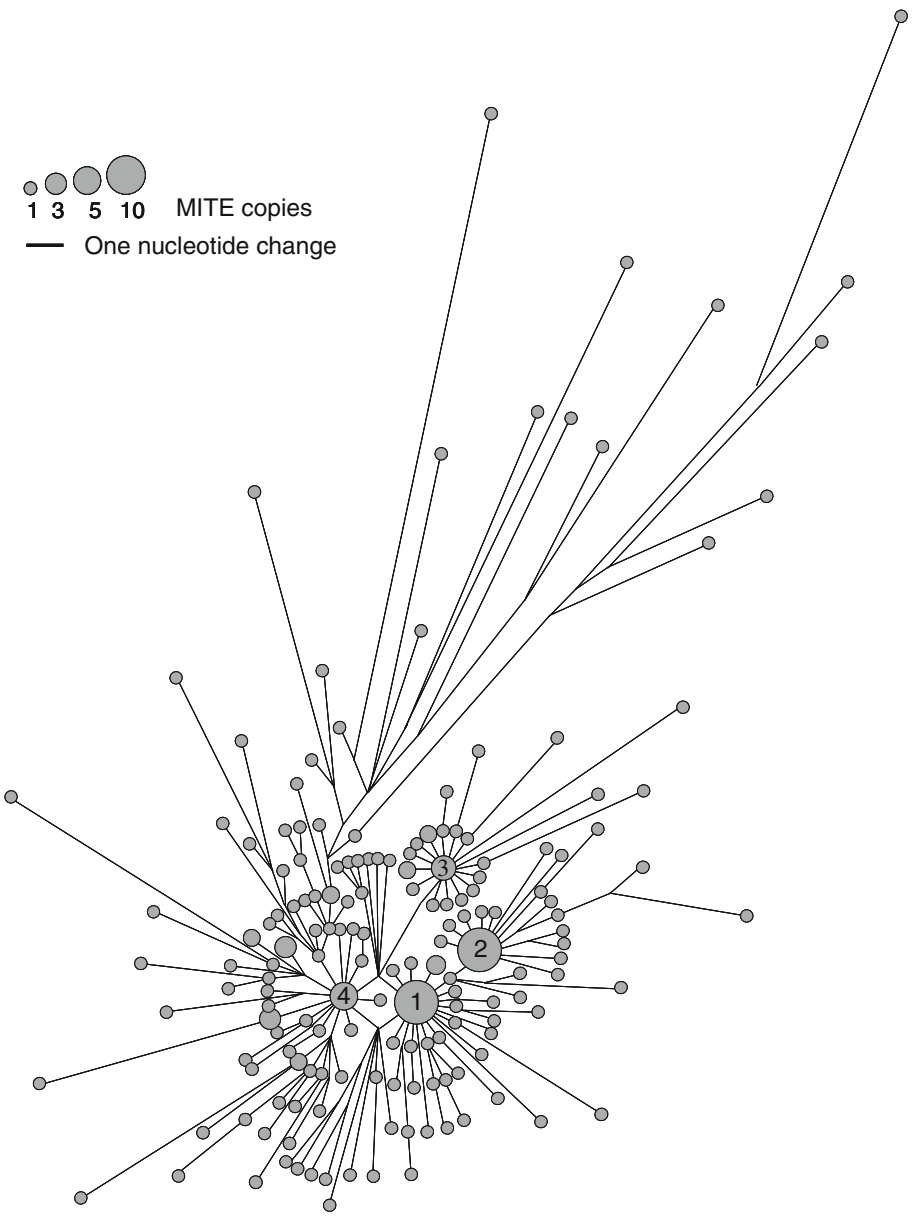

Fig. 6 Median-joining network of *ZmV1* copies. Each *circle* represents a TE *sequence* and the *circle* area is proportional to the number of identical sequences found in our data set. Branch length is proportional to the number of nucleotidic changes. The numbers 1, 2, 3 and 4 indicate the four strict star-shaped clusters grouping identical or nearly identical sequences

Santiago et al. 2002) revealed an insertion bias towards AT-rich regions. This tendency has been interpreted as a surviving strategy to escape purifying selection acting on coding regions (usually GC rich). Our results confirm a similar trend for MITE target regions in the maize genome, as evidenced by the elevated AT content found in the MITE flanking regions analysed (Table 2). Note that AT-rich regions in the maize genome often correspond to intronic and intergenic regions, which present an average AT content of 57.7 and 54% respectively, as shown by sequence analysis of 100 random BACs (Haberer et al. 2005). Our bioinformatic analysis of MITE insertion distribution with respect to predicted genes confirms MITEs' tendency to insert into intronic and intergenic regions close to genes (Table 4).

Regarding the methylation level, the quantification of cytosine methylation levels at MITE flanking regions in rice has revealed that MITEs insert primarily in poorly methylated regions (Takata et al. 2005; Takata et al. 2007). Using the M-STD analysis we report a similar pattern in the maize genome. The internal cytosine at the 5′-CCGG-3′ sites of MITE flanking regions, was, on average, threefold less methylated than the ones present in the vicinity of the LTR retrotransposon *Grande* (Fig. 4). This striking difference fits well with the previous knowledge that LTR retrotransposons tend to be embedded in hypermethylated regions (Bennetzen et al. 1994), while MITEs are preferentially inserting into single copy regions next to genes that in plants are generally poorly methylated. Consistent with these observations, we found that 73% of 263 MITE

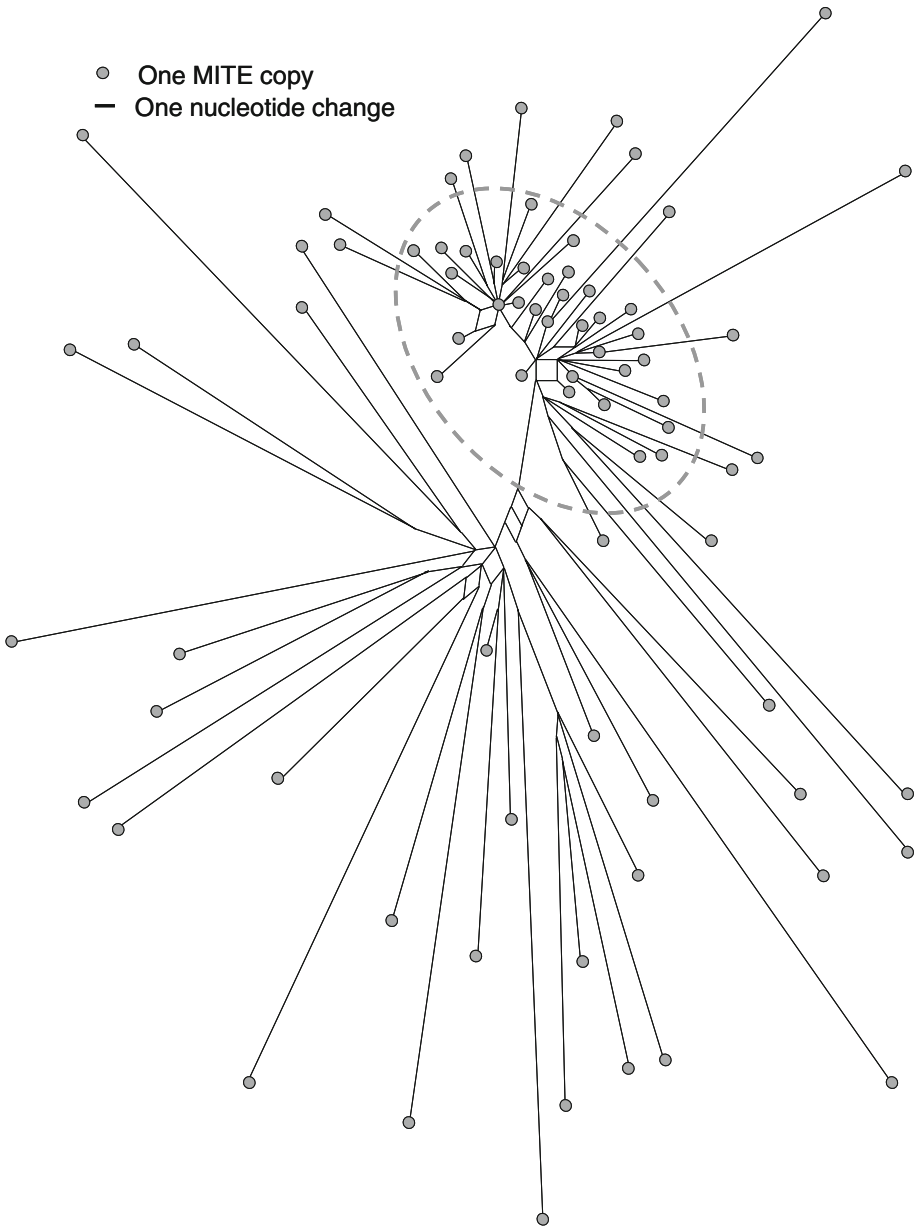

Fig. 7 Median-joining network of *Zead8* copies. Each *circle* represents a TE *sequence* and branch length is proportional to the number of nucleotidic changes. The shape of the network is consistent with the existence of two subpopulations of copies, one older with highly divergent sequences and one recent containing closely related ones (included within the dashed oval and called *Zead8* cluster 1 in Table 3)

insertions were located less than 5 kb from the nearest predicted gene. The M-STD analysis revealed different degrees of methylation among MITE families. *mPIF* family exhibited higher level of methylation (40% of bands; Fig. 4) compared to *ZmV1* and *Zead8*. This could result from preferential insertion of *mPIF* copies into more repetitive intergenic regions, consistent with the higher AT content of *mPIF* surrounding regions (Table 2), and could be the reason why a smaller proportion of *mPIF* copies, compared to the other MITE families, harboured genes in their flanking regions (Table 4). Altogether, our results suggest that MITE target preference are dictated by target motifs and DNA accessibility.

Despite the fact that MITEs are among the most numerous transposable elements in many organisms, mechanisms that govern MITE transposition remain poorly understood. What is known is that TIR sequences are required for MITE mobilization because they are recognized by the transposase proteins encoded by autonomous transposons (Feschotte et al. 2005; Loot et al. 2006). As a consequence, TIR sequences should present a higher level of conservation compared to other MITE non essential sequences. We observed TIR conservation only in *ZmV1* (Fig. 2a) and interestingly, this family is the only one exhibiting signs of intense recent expansion (Fig. 6). On the contrary, in *Zead8* no TIR conservation was seen, not even when *Zead8* cluster 1 copies were analyzed separately (data not shown).

The sliding window analysis of *Zead8* copies (Fig. 2c) revealed instead an unusually high conservation of AT-rich

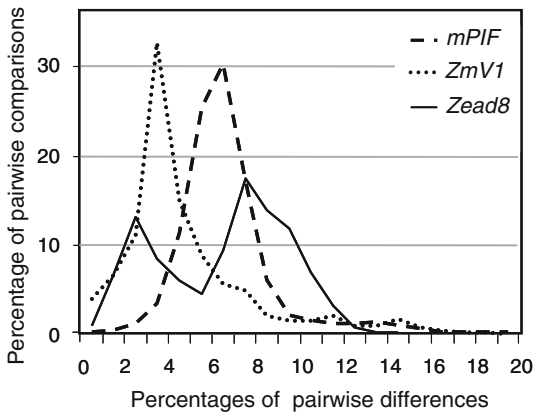

Fig. 8 Frequency distributions of pairwise differences obtained for the 3 MITE families. Pairwise differences are represented as percentages obtained by dividing the number of pairwise differences in each class by the length in bp of each MITE family

internal motives not present in the other MITE families. It is not simple to understand why such a contrasted sequence diversity exists between TIRs and internal sequences in this MITE family. However, it is tempting to postulate that while sequence variation at the TIR level for *Zead8* may indicate the loss of the active transposon necessary for its transposition, the high conservation of internal sequences may reflect a functional role of *Zead8* copies in the maize genome. Indeed, MITE insertions into promoter regions can mediate gene expression levels (Yang et al. 2005) through the regulatory motives they may harbour in their sequences (Oki et al. 2008). Interestingly the bioinformatic analysis of MITE insertions in relation to genes seems to support this hypothesis at least for one of the MITE families analyzed. *Zead8* copies were found largely upstream of genes, while *mPIF* and *ZmV1* exhibited no significant difference in the number of insertions located upstream versus downstream of genes.

Most MITE families described to date present a high level of MITE intrafamily homology and size similarity between copies (Bureau and Wessler 1992; Zhang et al. 2000) suggesting that they can be considered as a population that experienced several successive steps of amplifications (bursts) from a handful of master copies. Bursts are followed by periods of inactivity and drift. Identity in sequence and size will hence decrease over time due to random mutation. A large number of identical or very similar copies within a MITE family, therefore, stands as evidence for a recent burst. Our population genetic analyses reveal a consistent sign of activity dating back 1 million years ago, as well as contrasted patterns of expansion among the 3 MITE families. Hence, the network topology of *mPIF* (Fig. 5) could be the signature of an ancient population expansion of one or fewer ancestral copies, followed by independent differentiation of each new copy, accordingly to the strict master model proposed for MITE family origin (Feschotte et al. 2002). *ZmV1* network reveals instead the presence of waves of activity of several closely related master copies within the last 0.4 Mya that gave rise to strict star-shaped clusters (Fig. 6 and Table 3). In *Zead8* the network shape suggests the activity of several master copies that determined the formation of two closely related subpopulations, one of which has been moderately active within the last 0.37 Mya (Fig. 7).

The network results are supported by the pairwise distributions of MITE intra-family sequence diversity (Fig. 8). The unimodal pairwise distributions found for *mPIF* and *ZmV1* as well as the depleted raggedness values, fit well with the expectations of expanding populations. In *mPIF*, the skew of the pairwise distribution toward elevated values and the tightness of the bell shape, indicate a unique ancient expansion for these sequences. In *ZmV1*, the skew of the distribution towards low values reflects the presence of many identical or nearly identical sequences which likely result from recent episodes of expansion. Finally, the bimodal distribution of pairwise differences obtained for *Zead8* mirror the network obtained for this family, reinforcing the idea that this family derive from two waves of expansion separated in time (Table 3).

In conclusion, by combining a molecular characterization and a bioinformatic analysis of regions flanking MITE insertions, we first showed that *Tourist*-like elements target preferentially AT rich and low methylated regions, inserting preferentially into intronic and intergenic regions close to genes. Second, we showed that among the three MITE families analysed, *Zead8* was the only one having a significant large number of insertions located upstream of genes, perhaps reflecting their potential effect on gene regulation. Third, we discovered in *mPIF* and *ZmV1* palindromic target sites characterised by sequence specificity and size, which are the largest ever documented for any TE. Both of these results suggest that several factors contribute to MITE insertion preference and these factors differ among MITE families. Moreover, by employing a population genetic approach to analyse sequence divergence among MITE copies within each family, we identified successive waves of expansion among the three MITE families. The oldest insertions in all families date back roughly to 1 million years, but more recent activity of several master copies within the last 0.4 Mya was detected in *Zead8* and *ZmV1*. It would be interesting, now that the majority of the maize sequence is available, to extend this study to other MITE families in order to establish a clear picture of the evolution of MITEs as well as to identify families that may have contributed to the adaptive history of maize.

Acknowledgments We thank Catherine Damerval, Domenica Manicacci and Clementine Vitte for critical reading of the manuscript and Julie Dawson for review of the text. Corinne Mhiri provided with technical advices and support. This work was funded by the Agence Nationale de la Recherche (ANR-05-JCJC-0067-01 to M.I.T.).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48

Bennetzen JL, Schrick K, Springer PS, Brown WE, SanMiguel P (1994) Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. Genome 37:565–576

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

Brunner S, Fengler KA, Morgante M, Tingey SV, Rafalski JA (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17:343–360

Bureau TE, Wessler SR (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4:1283–1294

Bureau TE, Wessler SR (1994) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6:907–916

Bureau TE, Ronald PC, Wessler SR (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

Casa AM, Brouwer C, Nagel A, Wang L, Zhang Q, Kresovich S, Wessler SR (2000) The MITE family heartbreaker (Hbr): molecular markers in maize. Proc Natl Acad Sci USA 97:10083–10089

Clark RM, Tavaré S, Doebley JF (2005) Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Mol Biol Evol 22:2304–2312

Craig NL (1997) Target site selection in transposition. Annu Rev Biochem 66:437–474

Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) (2002) Mobile DNA II. ASM Press, Washington

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

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341

Feschotte C, Osterlund MT, Peeler R, Wessler SR (2005) DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. Nucleic Acids Res 33:2153–2165

Garcia-Martinez J, Martinez-Izquierdo JA (2003) Study of the evolution of the Grande retrotransposon in the Zea genus. Mol Biol Evol 20:831–841

Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G, Butler E, Wing RA, Rounsley S, Birren B, Nusbaum C, Mayer KFX, Messing J (2005) Structure and architecture of the maize genome. Plant Physiol 139:1612–1624

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp 41:95–98

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

Jiang N, Jordan IK, Wessler SR (2002) *Dasheng* and *RIRE*2. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. Plant Physiol 130:1697–1705

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR (2003) An active DNA transposon family in rice. Nature 421:163–167

Jiang N, Feschotte C, Zhang X, Wessler SR (2004) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol 7:115–119

Jobling MA, Hurles ME, Tyler-Smith C (2004) Human evolutionary genetics. Garland Publishing, Abingdon and New York

Kapitonov V, Jurka J (1996) The age of Alu subfamilies. J Mol Evol 42:59–65

Kapitonov VV, Jurka J (2004) Harbinger transposons and an ancient HARBI1 gene derived from a transposase. DNA Cell Biol 23:311–324

Ketting RF, Fischer SE, Plasterk RH (1997) Target choice determinants of the Tc1 transposon of Caenorhabditis elegans. Nucleic Acids Res 25:4041–4047

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 11:1–120

Liao GC, Rehm EJ, Rubin GM (2000) Insertion site preferences of the P transposable element in Drosophila melanogaster. Proc Natl Acad Sci USA 97:3347–3351

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

Lu Y, Rong T, Cao M (2008) Analysis of DNA methylation in different maize tissues. J Genet Genomics 35:41–48

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

Mahillon J, Chandler M (1998) Insertion sequences. Microbiol Mol Biol Rev 62:725–774

Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, Tanisaka T, Wessler SR (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci USA 103:17620–17625

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

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

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

Ouyang S, Buell CR (2004) The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:D360–D363

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

Posada D, Crandall K (2001) Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol 16:37–45

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552–569

Rozas J, Rozas R (1999) DnaSP version 3: AN integreted program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718–726

Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E, Li B, Hainey CF, Radovic S, Zaina G, Rafalski JA, Tingey SV, Miao GH, Phillips RL, Tuberosa R (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci USA 104:11376–11381

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Santiago N, Herraiz C, Goni JR, Messeguer X, Casacuberta JM (2002) Genome-wide analysis of the Emigrant family of MITEs of Arabidopsis thaliana. Mol Biol Evol 19:2285–2293

Schneider S, Roessli D, Excoffier L (2000) Arlequin: a software for population genetic data. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Slatkin M, Hudson RR (1991) Paiwise comparisons on mitocondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Smit AF, Riggs AD (1996) Tiggers and DNA transposon fossils in the human genome. Proc Natl Acad Sci USA 93:1443–1448

Surzycki SA, Belknap WR (2000) Repetitive-DNA elements are similarly distributed on Caenorhabditis elegans autosomes. Proc Natl Acad Sci USA 97:245–249

Tai T, Tanksley S (1991) A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue. Plant Mol Biol 8:297–303

Takata M, Kishima Y, Sano Y (2005) DNA methylation polymorphisms in rice and wild rice strains: detection of epigenetic markers. Breed Sci 51:57–63

Takata M, Kiyohara A, Takasu A, Kishima Y, Ohtsubo H, Sano Y (2007) Rice transposable elements are characterized by various methylation environments in the genome. BMC Genomics. doi: 10.1186/1471-2164-1188-1469

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28:286–289

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

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Walker EL, Eggleston WB, Demopulos D, Kermicle J, Dellaporta SL (1997) Insertion of a novel class of transposable elements with a strong target site preference at the *r* locus of maize. Genetics 146:681–693

Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BB, Powell W (1997) Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol Gen Genet 253:687–694

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Xiong LZ, Xu CG, Saghai Maroof MA, Zhang Q (1999) Patterns of cytosine methylation in an elite rice hybrid and its parental lines, detected by a methylation-sensitive amplification polymorphism technique. Mol Gen Genet 261:439–446

Xu M, Li X, Korban SS (2000) AFLP-based detection of DNA methylation. Plant Mol Biol Rep 18:361–368

Yang G, Lee YH, Jiang Y, Shi X, Kertbundit S, Hall TC (2005) A two-edged role for the transposable element *Kiddo* in the *rice ubiquitin2* promoter. Plant Cell 17:1559–1568

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

Yeadon PJ, Catcheside DE (1995) Guest: a 98 bp inverted repeat transposable element in Neurospora crassa. Mol Gen Genet 247: 105–109

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

Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR (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, Jiang N, Feschotte C, Wessler SR (2004) PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. Genetics 166:971–986