Miniature Inverted–Repeat Transposable Elements (MITEs) Have Been Accumulated through Amplification Bursts and Play Important Roles in Gene Expression and Species Diversity in *Oryza sativa*

Chen Lu,† Jiongjiong Chen,† Yu Zhang, Qun Hu, Wenqing Su, and Hanhui Kuang*

Key Laboratory of Horticulture Biology, Ministry of Education and Department of Vegetable Crops, College of Horticulture and Forestry, Huazhong Agricultural University, Wuhan, People's Republic of China

†These authors contributed equally to this study.

***Corresponding author:** E-mail: kuangfile@gmail.com.

Associate editor: Naruya Saitou

Research article

Abstract

Miniature inverted–repeat transposable elements (MITEs) are predicted to play important roles on genome evolution. We developed a BLASTN-based approach for de novo identification of MITEs and systematically analyzed MITEs in rice genome. The genome of rice cultivar Nipponbare (*Oryza sativa* ssp. *japonica*) harbors 178,533 MITE-related sequences classified into 338 families. Pairwise nucleotide diversity and phylogenetic tree analysis indicated that individual MITE families were resulted from one or multiple rounds of amplification bursts. The timing of amplification burst varied considerably between different MITE families or subfamilies. MITEs are associated with 23,623 (58.2%) genes in rice genome. At least 7,887 MITEs are transcribed and more than 3,463 were transcribed with rice genes. The MITE sequences transcribed with rice coding genes form 1,130 pairs of potential natural sense/antisense transcripts. MITEs generate 23.5% (183,837 of 781,885) of all small RNAs identified from rice. Some MITE families generated small RNAs mainly from the terminals, while other families generated small RNAs predominantly from the central region. More than half (51.8%) of the MITE-derived small RNAs were generated exclusively by MITEs located away from genes. Genome-wide analysis showed that genes associated with MITEs have significantly lower expression than genes away from MITEs. Approximately 14.8% of loci with full-length MITEs have presence/absence polymorphism between rice cultivars 93-11 (*O. sativa* ssp. *indica*) and Nipponbare. Considering that different sets of genes may be regulated by MITE-derived small RNAs in different genotypes, MITEs provide considerable diversity for *O. sativa*.

Key words: MITEs, amplification, gene expression, evolution, small RNA.

Introduction

Miniature inverted–repeat transposable elements (MITEs) are considered as truncated derivatives of autonomous DNA transposons (Feschotte and Mouches 2000; Zhang et al. 2000; Yang et al. 2001; Feschotte et al. 2003; Yang and Hall 2003a). MITEs exhibit the structural features of DNA transposons, containing terminal inverted repeats (TIRs) flanked by small direct repeats (target site duplication, TSD). Unlike autonomous DNA transposons, however, the internal sequences of MITEs are short and devoid of open reading frame (ORF). As nonautonomous elements, MITEs transpose through transposases encoded by autonomous DNA transposons (Jiang et al. 2003; Yang et al. 2009).

MITEs were most extensively studied in rice (*Oryza sativa*), and hundreds of MITE families were discovered in the rice genome (Jiang et al. 2004; Oki et al. 2008; Han and Wessler 2010). Several methods have been developed to identify novel MITEs through searching sequences with TIR and TSD features, either directly from a database or from the groups of repetitive sequences identified using programs such as RepeatMasker (Tu 2001; Yang and Hall 2003b; Chen et al. 2009). Among them, MITE-hunter is the most successful tool in de novo identification of MITEs (Han and Wessler 2010). However, all previous programs may miss MITEs with short and/or poor-matching TIR sequences.

MITEs were shown to be mainly distributed on chromosome arms and highly associated with genes (Wessler 1998; Mao et al. 2000; Feng et al. 2002; Santiago et al. 2002; Wright et al. 2003; Oki et al. 2008). MITE sequences are frequently transcribed with plant genes (Oki et al. 2008; Kuang et al. 2009). The observed close physical association between MITEs and plant genes has provoked the hypothesis that MITEs play important roles in gene regulation and genome evolution. This hypothesis is

partially supported by the findings that MITEs may provide coding sequences or poly(A) signals for genes (Oki et al. 2008; Kuang et al. 2009) and regulate the expressions of host genes in which they reside (El Amrani et al. 2002; Yang et al. 2005; Naito et al. 2009). MITE, if containing regulatory motifs, may upregulate gene expression (Yang et al. 2005; Naito et al. 2009). Alternatively, MITE may downregulate gene expression through MITE-derived small RNAs (Kuang et al. 2009). The MITE-derived small RNAs might be generated via the microRNA (miRNA) pathway since some MITEs have similar structure to that of miRNA genes (Piriyapongsa et al. 2007). However, the MITE-derived small RNAs in Solanaceae were most likely generated by the pathway of small interfering RNA (siRNA) biogenesis, involving genes such as *DCL3*, *DCL4*, and *RDR2* (Kuang et al. 2009). To better understand the role of MITEs on genome evolution, it is critical to systemically investigate MITE-derived small RNAs and analyze the effects of MITE on gene expression.

MITE may transpose into different sites in different genotypes, forming presence/absence (insertional) polymorphism (Casa et al. 2000; Lyons et al. 2008). When a MITE transposes into a new site, the corresponding site with no MITE is called related empty site (RESite; Le et al. 2000). The presence/absence polymorphism can be also caused by excision of MITE from a locus. When a MITE transposes away, the empty donor site left behind has an extra TSD sequence compared with the locus prior to MITE insertion. However, excision often generates footprints, including short deletions and/or insertion of unrelated sequence (Kikuchi et al. 2003; Nakazaki et al. 2003). It is hypothesized that excision with footprints is genotype dependent (Shan et al. 2005). It remains unknown how many MITE loci exhibit presence/absence polymorphism in a species and what proportion of the polymorphism was generated by insertions or excisions.

The diversity and evolutionary patterns may vary considerably in different MITE families. Some MITE families, such as the *mPing* family in rice and the *Heartbreaker* in maize, are highly conserved in a genome (Zhang et al. 2000; Jiang et al. 2003). Their high conservation indicates that they were amplified recently. In contrast, other MITE families resulted from ancient amplifications (Zerjal et al. 2009). The amplification of different MITE families is hypothesized to be activated by distinct autonomous elements (Feschotte et al. 2002; Feschotte and Wessler 2002). Though there has been considerable progress on molecular mechanisms of MITE transposition, the genetic and evolutionary mechanisms of MITE amplification and accumulation in a species remain poorly understood.

In this study, we developed a BLASTN-based approach for de novo identification of MITEs in a genome. The method was applied to genome-wide identification of MITE elements in rice cultivar Nipponbare. The distribution of MITEs, their amplification patterns, and MITE-derived small RNAs were systematically investigated. The effects of MITEs on gene expression and their contribution to the diversity of *O. sativa* were analyzed.

Materials and Methods

Development of a BLASTN-Based Approach for MITE Identification

The ends of elements from the same MITE family are usually highly conserved while their flanking sequences in a genome are unrelated. To identify MITEs, a group of Perl scripts (repetitive sequence with precise boundaries, RSPB) were written to identify repetitive sequences that have precise boundaries. If five or more repetitive sequences share high similarity (BLASTN *e* value $< 10^{-15}$), are shorter than 1,500 bp, and have precise boundaries (the ends of homologous sequences vary no more than 5 bp and the similarity of their 100-bp flanking sequences is less than 50%), they are retrieved from a database as a "group" of repetitive sequences. To save running time, no more than 20 elements are kept in each group. Notice that different groups may share certain level of sequence similarities since a high threshold was used in this step. Then each group of repetitive sequences was checked manually for the presence of TIR and TSD sequences. Only groups with precise boundaries and obvious TIR and TSD sequences were considered as MITEs. The pipeline of this procedure is described in figure 1. The program and a readme file with detailed instructions are available upon request.

Genome-Wide De Novo Identification of MITEs in Rice

The RSPB program was applied to de novo identification of MITEs in the pseudomoleculars of MSU annotation 6.1 (http://rice.plantbiology.msu.edu/) of rice cultivar Nipponbare (*O. sativa* ssp. *japonica*). Groups (BLASTN *e* value $< 10^{-15}$) of repetitive sequences with precise boundaries were identified using RSPB. Then the repetitive sequences as well as 100-bp flanking sequences were retrieved and aligned. The two terminals of the conserved sequences (i.e., the repetitive sequences) were checked for at least 6-bp inverted repeats. Up to 15 bp immediately flanking the repetitive sequences were checked for potential TSD sequences.

Classification and Characterization of MITEs in Rice

Different MITE groups with significant similarity (BLASTN *e* value $< 10^{-10}$) were combined into a MITE family. In other words, a MITE family is defined as all elements with significant similarities with each other (*e* value $< 10^{-10}$) but no similarity with elements from other families (*e* value $> 10^{-10}$). The confirmed MITE families were assigned into superfamilies based on their TIR and TSD sequences. Each MITE family was named as *OsX#*, where *X* is a letter representing a superfamily and # is a number representing the family. Letter *T* stands for *Tc1/Mariner*, *M* for *Mutator*, *h* for *h*AT, *C* for *CACTA*, *P* for *PIF/Harbinger*, and *Mi* for *Micron*. For example, *OsT5* refers to the 5th MITE family of the *Tc1/Mariner* superfamily in *O. sativa*.

An element with typical length and structure was chosen from each family as the "seed" element. To identify diverse and/or partial elements, seed elements from each

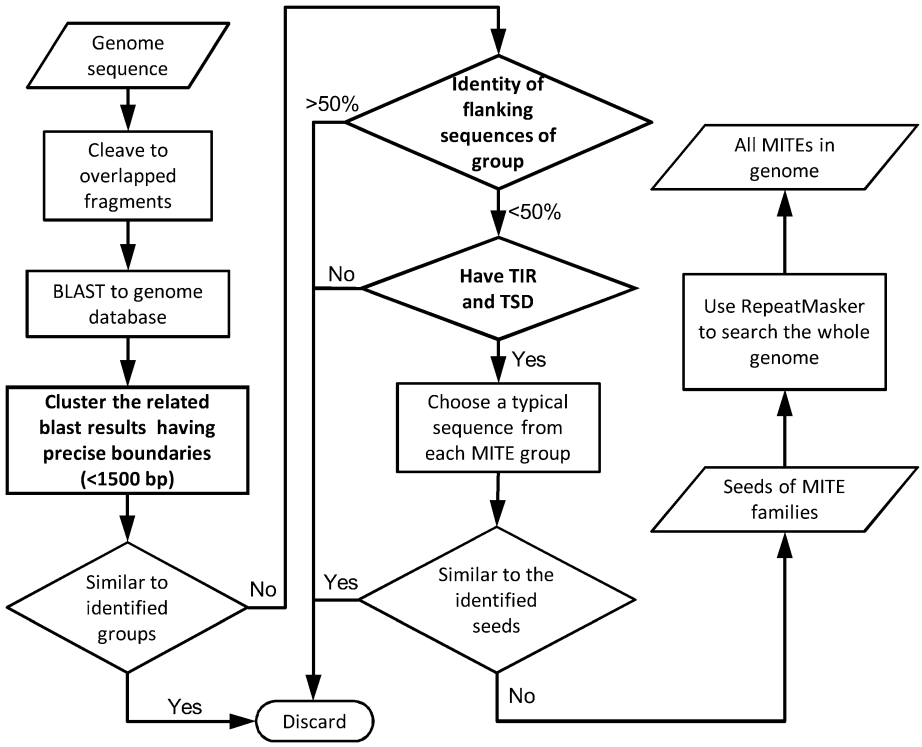

Fig. 1. Pipeline of RSPB.

family were used as a reference library by RepeatMasker v3.2.9 (with cross_match as search engine) in search of the entire genome sequence. The partial elements having significant similarities with previously reported long (>1 kb) DNA transposons were excluded.

Association of MITEs and Genes in Rice Genome

The distribution of each MITE family on chromosomes was plotted out using a window of 100 kb. To test if MITEs are preferentially associated with genes, noncoding sequences are divided into two groups: genes' noncoding sequences (GNSs) and intergenic sequences (ISs). GNS is defined as all intron sequences and 500-bp sequences upstream of start codon and 500-bp sequences downstream of stop codon of a gene. All other noncoding sequences are considered as IS. The distance of a MITE to the coding sequence of its closest gene was used as its distance from gene, and the average numbers of MITEs per gene located at different distance (a window of 100 bp) from genes were calculated.

Natural Sense/Antisense Transcript

Full-length cDNA sequences from rice cultivar Nipponbare were downloaded from KOME (http://cdna01.dna.affrc.go.jp/cDNA/) and NCBI and mapped to genome sequences by BLAT (Kent 2002). A MITE sequence is claimed in a cDNA if its genomic position overlaps the mapped cDNA region. If the MITE sequences in two cDNAs are in opposite orientations and can form at least 100-bp double strand RNAs with >90% nucleotide identity (Wang et al. 2006), the two transcripts are considered as MITE-derived sense/antisense transcripts. The presence of MITE-related sequences in RNAs was also investigated in rice transcriptome sequences (Lu et al. 2010). However, the orientations of these transcripts were unclear, and therefore they were not used for the analysis of sense/antisense transcripts.

Sequence Analysis

If a MITE element is at most 3 bp short at its terminals when compared with the seed element, it is considered as a full-length element. Sequences of full-length elements of the same MITE family were aligned using MUSCLE v3.8 (Edgar 2004). Neighbor-joining trees (pairwise deletion for gaps and Kimura 2-parameter substitution model) for MITE families were constructed using MEGA 4 (Tamura et al. 2007). Pairwise diversity was defined as the number of mismatched sites divided by alignment length. Each gap was considered as one single mismatch. A Perl script was written to calculate pairwise nucleotide diversity among members of a MITE family and to construct histogram of pairwise nucleotide diversity. Sequence exchanges were analyzed using four-gamete method implemented in DnaSP v5 (Librado and Rozas 2009) and Geneconv v1.81a (Sawyer 1989) using the default settings. Substitution rate of 1.3×10^{-8} base substitutions per site per year was used to estimate the divergence time between two sequences (Ma and Bennetzen 2004).

Diversity Generated by MITE Insertions

The MITE sequences from rice cultivar 93-11 genome (http://rice.genomics.org.cn/rice/index2.jsp) were retrieved by RepeatMasker using the seed element from each MITE

family in Nipponbare as query sequences. A MITE locus is defined based on the 1,000-bp sequences flanking a MITE element. The flanking sequences were used to identify the allele of a MITE locus in the other cultivar. To be allelic, the flanking sequences must be the best hits in BLAST and with an *e* value $<10^{-50}$.

MITE-Derived Small RNAs

The 781,885 rice small RNAs (including miRNAs) were downloaded from http://csrdb.ucdavis.edu/smrnas/ and miRbase version 16. All MITE elements identified from the rice genome were BLASTed with these small RNA sequences. If a small RNA has perfect match with a MITE sequence, it is considered as MITE derived. Small RNAs derived from full-length MITEs were analyzed to map them in MITE sequences. Due to variations of MITE length, relative positions in MITE sequence were used to map small RNAs. For example, if a full-length MITE is 200 bp in length and has perfect match with a small RNA between the 20th and the 40th nucleotide, the small RNA is mapped to the 10–20% region of the MITE. The number of small RNAs at each relative position (1–100%) was counted and a curve showing the number of distinct small RNAs at each relative position was drawn.

The positions of small RNA were also investigated for individual MITE families. In this case, only elements with identical or similar length in a family were analyzed. To limit the bias caused by small sample size, only curves with more than 100 small RNAs were reported.

Analysis of Gene Expression

To investigate the effects of MITE insertion on gene expression, the paired-end RNA-Seq data sets of 2-week-old shoot of rice cultivars Nipponbare and 93-11 were downloaded from the EMBL Sequence Read Archive (SRA) under accession number ERA000212 (Lu et al. 2010). The RNA sequences were aligned to their corresponding genomic sequences: MSU RGAP v6.1 pseudomolecules of Nipponbare and BGI draft chromosomal sequences of 93-11, using SOAP v2.21 (Li et al. 2009). Paired-end reads that were mapped uniquely to genome sequence were further aligned to genes annotated by MSU RGAP v6.1 (Nipponbare) and BGI glean (93-11) using R package girafe (Toedling et al. 2010). The expression level was normalized through reads per kilobase of exon model per million mapped reads (RPKM; Mortazavi et al. 2008), and the approach described by Ramsköld et al. (2009) was used to estimate the background.

As defined above, genes are considered to be associated with MITEs if a MITE is inserted in its introns or 500-bp flanking regions. Accordingly, rice genes were categorized into either a group with MITE insertion or a group away from MITE. The expression of genes with MITE insertion was compared with that of genes away from MITEs in rice cultivars Nipponbare and 93-11. The expression level for genes with MITE insertions and genes away from MITEs was analyzed using *t*-test and Mann–Whitney *U* (MWU) test. To better demonstrate the effects of MITEs on expression, the distribution densities of expression levels ($\log_2$ RPKM) of genes with MITE insertion and genes away from MITEs were drawn using the *hist* function in R platform, respectively.

We also compared the expression of genes that have MITE insertion in one culitvar but not in the other. The complete genomic sequences of Nipponbare and 93-11 were aligned using MAUVE v2.3.1 (Darling et al. 2004) to identify orthologous genomic regions between the two cultivars. If a gene from orthologous genomic regions of one cultivar is the best hit of a gene in the corresponding region of the other cultivar by bidirectional BLASTN (*e* value $< 10^{-80}$ and nucleotide identity $> 95\%$), the two genes were considered to be alleles. The raw read numbers of alleles in the two cultivars were input into the R package DESeq (Anders and Huber 2010) to compare the expression of each pair of alleles, including those that have MITE insertion in one cultivar but have no MITE insertion in the other cultivar.

Results

Development of a BLASTN-Based Approach for De Novo Identification of MITEs

The full-length elements of a MITE family should have similar boundaries. In order to identify novel MITEs, a series of Perl scripts, RSPB were written to search for groups of repetitive sequences with precise boundaries (fig. 1). RSPB was applied to the genome of rice cultivar Nipponbare (*O. sativa* ssp. *japonica*), and 1,588 groups of repetitive sequences with precise boundaries were obtained. Visual investigation discovered that 1,273 (80.2%) of them have the structural features of TIR and TSD. Of them, 187 (14.7%) are solo-LTRs and were excluded from further analysis. The remaining 1086 groups have precise boundaries, obvious TIR and TSD sequences, and structural features (i.e., A/T rich TSD, the length of TSD) similar to those of known MITE superfamilies; therefore, these groups were considered MITE sequences. These newly obtained candidate MITE groups were compared with rice MITE families reported previously using BLASTN. The MITE groups identified from this study have significant similarities (BLASTN *e* value $< 10^{-10}$) with 97.8% of the 186 MITE families deposited in Repbase 15.08 (Jurka 2000), 98.9% of the 14,957 MITE sequences in MSU *Oryza* Repeat Database v3.3 (Ouyang and Buell 2004), and 89.8% of the 551 MITE families identified by MITE-hunter (Han and Wessler 2010). RSPB identified hundreds of groups of novel MITE sequences that have not been detected previously (see below). Therefore, the RSPB is an efficient and reliable method for MITE identification and can be used to identify the vast majority of the MITE families in a genome without prior information.

The Rice Genome Has 178,533 MITE-Related Sequences from 338 MITE Families

Different MITE groups identified in this study as well as from previous studies were combined into a MITE family if they have significant sequence similarities (BLASTN *e* value $< 10^{-10}$). Consequently, 343 tentative MITE

Table 1. Summary of MITE Superfamilies in Rice Cultivar Nipponbare.

Superfamily	Family Number	Total Elements	Length of All Elements (bp)	TSD	TIR (bp)
Tc1/Mariner	47	50,207	9.04×10^6	TA	4–554
PIF/Harbinger	88	59,407	1.21×10^7	TWA	4–686
CACTA	6	3,859	9.47×10^5	3 bp	9–800
hAT	81	15,299	3.64×10^6	8 bp	5–575
Mutator	115	49,126	1.11×10^7	9/10 bp	4–799
Micron	1	655	2.11×10^5	$(TA)_n$	0–142
Total	338	178,553	3.70×10^7		

families were obtained. Based on their TSD patterns and TIR sequences, the 343 families were classified into superfamilies *Tc1/Mariner*, *Mutator*, *PIF/Harbinger*, *h*AT, and *CACTA* (table 1). Besides the five superfamilies, the *Micron* MITE family, which was inserted specifically into $(TA)_n$ repeats and with specific TIR sequences, is considered as an independent type (Akagi et al. 2001).

The MITEs identified above are full-length elements since they were identified as groups of homologous sequences with precise boundaries. In order to retrieve partial or divergent ones, the elements with typical length and structure (called the seed elements hereafter) from each family were used as reference library to search the rice cultivar Nipponbare genome using RepeatMasker (http://repeatmasker.org). More than 187,000 MITE-related sequences, including long elements, were obtained from the rice genome.

Using the seed elements as reference, 62,100 (33.1%) elements were confirmed to be full length. It should be noted that many full-length elements might be missed due to variations at the terminals since a strict criteria was used to define full-length elements (see MM section). The vast majority (79.2%) of the verified full-length elements are 100–400 bp in length, with two distribution peaks at ~150 and 240 bp. Some MITE families have conserved length. For example, 97.5% of the 4,513 full-length elements from the *OsT38* family are 238–240 bp. In contrast, most MITE families have a large variation on length. The elements from the *OsM88* family vary from 84 to 3,300 bp. The length of all MITE elements has a continuous distribution, and there is no clear cut between short (MITE) and long transposable elements. Nevertheless, elements >800 bp have visibly fewer copies than those <800 bp. Hereafter, only elements ≤800 bp were considered MITE, consistent with previous studies (Han and Wessler 2010). In five families, all elements are larger than 800 bp and they were not analyzed further in this study. After all families and elements of >800 bp were removed, 178,533 MITE-related sequences (with a total of 37.0 Mb) from 338 families were found in the genome of rice cultivar Nipponbare. At least 4,238 (2.36%) elements had insertions of MITE sequences from other families, that is, they are part of MITE multimers (Jiang and Wessler 2001). Of the 338 MITE families, 37 were identified for the first time in this study, whereas 26 were discovered previously but missed using RSPB (supplementary table 1, Supplementary Material online).

Rapid Amplification of MITE Families at Different Times

Thirty-seven MITE families, randomly chosen to represent all superfamilies and families with different copy numbers (43–1,613), were selected to investigate the mechanisms of MITE amplifications. Pairwise nucleotide diversities of full-length elements were calculated and histograms were drawn for each of the 37 MITE families. The histograms of most MITE families are wave-like curves. Fifteen of them have unimodal distribution (fig. 2*A*), 4 have bimodal distribution (fig. 2*B*), and the other 18 have multimodal distribution or flat peaks. The wave-like histograms of pairwise diversity indicated that each family has experienced rapid population expansion (burst) in evolutionary history (Rogers and Harpending 1992).

The histograms for the 15 MITE families with unimodal distribution have their peaks at different diversities (0.00–0.25), suggesting that the amplification bursts occurred at distinct time for these MITE families (fig. 2*A*). The histogram for the *mPing* family (*OsP54*) has only the front face of a wave, centered at diversity = 0 (fig. 2*A*). The 51 *mPing* elements in Nipponbare have three distinct sequences, with at most two single nucleotide polymorphisms (SNPs) between any two elements. The low nucleotide diversity and the existence of many identical elements indicate that the *mPing* family is still under rapid amplification (Naito et al. 2006). In contrast, the amplification time for the *OsT18* family (average pairwise nucleotide diversity is 0.257) was estimated to be 19.8 million years ago (Ma) (Ma and Bennetzen 2004).

Some MITE Families Have Had Multiple Rounds of Rapid Amplifications

To investigate the evolutionary mechanisms underlying the different shapes of the histograms of pairwise nucleotide diversity, phylogenetic trees were constructed for 11 MITE families representing different histograms. The families with unimodal distribution of pairwise nucleotide diversity have phylogenetic trees of a star shape. No obvious clades are present on the tree, suggesting rapid amplification from one master element (fig. 2*C*). In contrast, MITE families with bimodal or multimodal distributions have several well-supported clades (fig. 2*D*). Interestingly, the pairwise nucleotide diversity within a clade has unimodal distributions (fig. 2*E*). The distribution peak varies for different clades of a family, indicating that the amplification burst for each clade occurred at different time. Therefore, some MITE families were resulted from one amplification burst, whereas other families may have experienced multiple rounds of amplification bursts in their evolutionary history.

MITEs Are Randomly Distributed in Noncoding Regions

To investigate whether MITE inserted randomly in rice genome, we studied the distribution of MITEs on chromosomes. The centromere regions of chromosomes 4, 5, and 8 have been completely sequenced (Feng et al. 2002; Wu et al. 2004). Compared with chromosome arms,

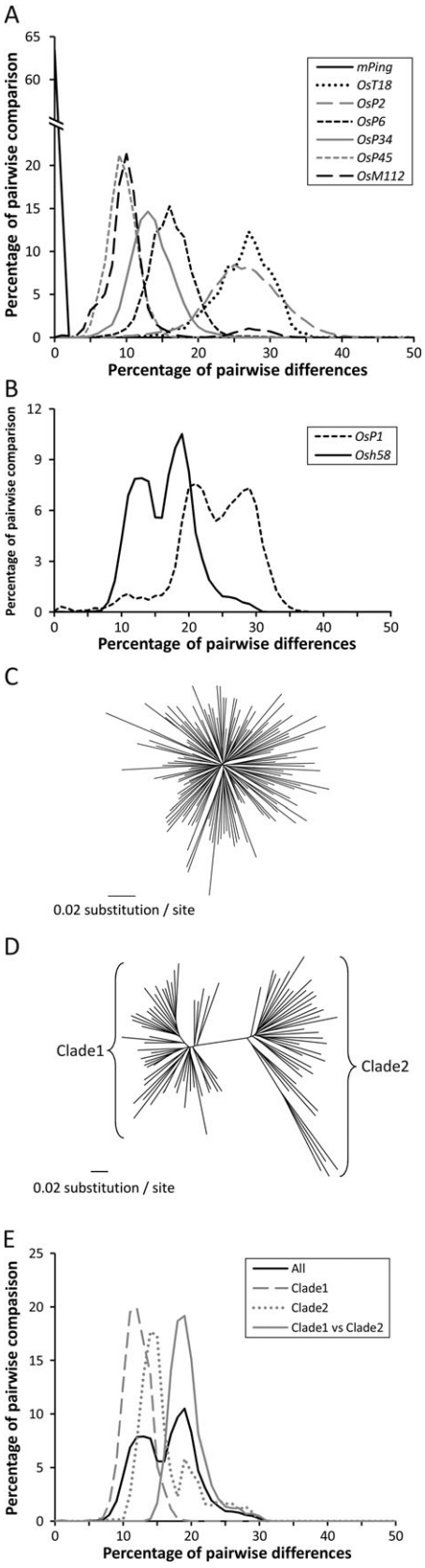

Fig. 2. Rapid amplification of MITE families at different times. (*A*) Unimodal distribution of pairwise nucleotide diversity among full-length elements in some MITE families, suggesting one amplification burst. Only eight families are shown. (*B*) Bimodal distribution, suggesting more than one round of amplification burst. Only two families are shown. (*C*) Phylogenetic tree of MITE family *OsP45* with unimodal distribution of pairwise nucleotide diversity. The star shape tree suggests one round of amplification burst. (*D*) MITE family *Osh58* with bimodal distribution of pairwise nucleotide diversity has two well-supported clades in its phylogenetic tree. (*E*) The distribution of pairwise nucleotide diversity for each of the two clades in family *Osh58*. The peaks of the curves for each clade vary considerably, suggesting different amplification time.

Table 2. The Association of MITE and Rice Genes.

Superfamily	Total Elements	Associated with Genes	Expressed	Expressed with Genes	Small RNAs
Tc1/Mariner	50,207	14,830	2,042	983	33,917
PIF/Harbinger	59,407	14,101	2,298	974	70,257
CACTA	3,859	739	134	58	7,380
hAT	15,299	4,341	737	280	15,395
Mutator	49,126	15,252	2,665	1,162	56,646
Micron	655	138	11	6	242
Total	178,533	49,401	7,887	3,463	183,837

centromeres and their flanking regions contain fewer MITEs, similar to the distribution of rice genes (data not shown). Low density of MITEs near centromeres was also observed on the other nine chromosomes, which have no centromere sequences available (data not shown). Above conclusion is consistent with previous observations when only a subset of MITE families were analyzed (Mizuno et al. 2006).

The similar distributions of MITEs and genes on chromosomes and frequent association between them have prompted the hypothesis that MITEs are preferentially associated with genes. To test this hypothesis, we compared the number of MITEs located in genes' noncoding regions (GNRs, defined in Materials and Methods) and that in intergenic regions (IRs). The rice genome has 90.7 Mb GNR, which harbors 49,401 MITEs (1 MITE per 1.84 kb; table 2). In comparison, the 241 Mb IR contains 128,748 MITEs (1 MITE per 1.87 kb). The MITE density in the two regions varies only 1.6%. Therefore, MITEs seem randomly distributed on noncoding regions of chromosome arms. The high percentage (58.2%; 23,623 genes) of genes associated with MITEs in rice genome might be due to random distribution of a large number of MITEs. The frequency of MITEs at different distance away from a gene was calculated, and results showed that MITEs are evenly distributed at different distance from genes (fig. 3*A*), further supporting above observation that MITEs are randomly distributed in intergenic regions.

To investigate if different MITE families have similar distribution patterns, each MITE family was analyzed separately, and a regression curve between the number of genes in GNR and that in IR for each family was constructed (fig. 3*B*; $r^2 = 0.95$, $P < 0.05$). Its regression slope (0.37) is almost equal to the ratio of GNR/IR (90.1 Mb/241 Mb = 0.376). Figure 3*B* shows that nine MITE families (only families with more than 1,000 elements were considered) have more elements in the GNR region than expected, whereas eight families have more elements in the IR region than expected (95% confidence interval). However, chi-square tests suggested that the variations were not statistically significant.

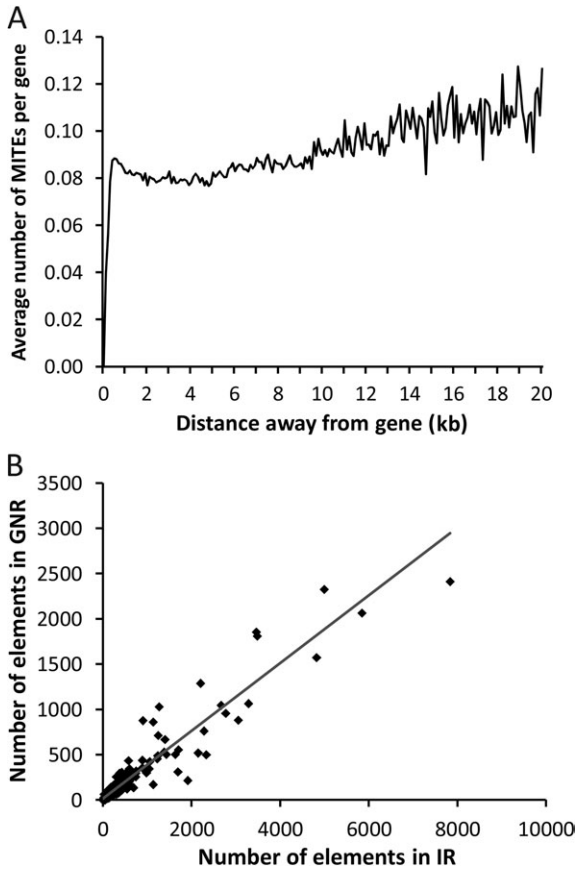

FIG. 3. Distribution of MITEs on noncoding regions. (*A*) Frequency of MITEs at different distances away from genes. (*B*) Regression curve of the number of MITEs located in the GNS region on that in the IS region. Each MITE family was counted separately.

MITEs were frequently found in resistance gene (*R*-gene) loci and may play a role in *R*-gene evolution (Meyers et al. 1998; Song et al. 1998; Kuang et al. 2009). However, our genome-wide analysis showed that MITEs are not preferentially associated with any gene categories (gene ontology), including *R*-genes (data not shown).

Natural Sense/Antisense Transcripts Derived from MITE Sequences

Analysis of 37,961 full-length cDNAs from KOME and NCBI and RNA-Seq data (Lu et al. 2010) showed that at least 7,887 MITEs in Nipponbare are transcribed (table 2). Approximately 8.47% (3,214) of full-length cDNAs and 3.76% (4,447) of 118,064 assembled transcripts contain MITE sequences, and at least 3,463 MITE sequences are present in transcripts with coding sequences. Different superfamilies have similar proportion of elements transcribed with rice genes (table 2). Interestingly, at least 88 MITE families were inserted into genes in both orientations, forming at least 1,130 pairs of potential natural sense/antisense transcripts, if coexpressed (fig. 4*A*). Besides the MITE sequences in mature mRNAs, more than 10,429 MITEs are located in introns and are predicted to be transcribed into RNAs, which will form at least 6,492 additional pairs of potential double-stranded RNAs.

Nearly a Quarter of Small RNAs in Rice Are Derived from MITEs

Besides the natural sense/antisense transcripts, some transcribed MITE sequences themselves may form double strand RNAs. For example, some elements from the *Tc1/Mariner* superfamily are mainly composed of TIR sequences and can form hairpin structures when transcribed (fig. 4*C*). The double strand RNAs formed by sense/antisense or hairpin structures may be recognized by Dicer and be processed into small RNAs (fig. 4*B*). Of the 781,885 published rice small RNAs, 183,837 (23.5%) were derived from MITE-related sequences (table 2). A total of 104,079 (58.2%) MITE sequences have perfect match with at least one small RNA sequence, and they have representatives from all but two MITE families, which have only 5 and 30 elements, respectively. The number of distinct small RNAs from a MITE family is significantly correlated with the total length of sequences in a family ($r^2 = 0.8415$, $P < 0.001$). Surprisingly, 51.8% of MITE-derived small RNAs

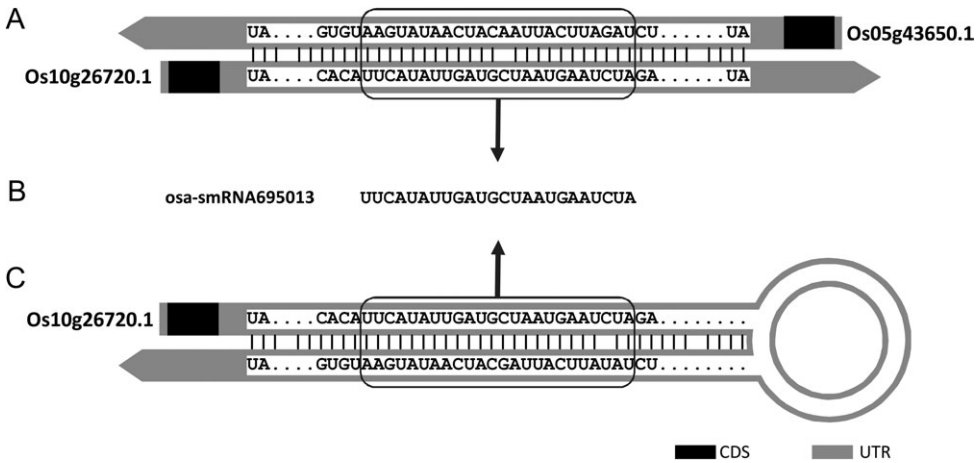

FIG. 4. Double-stranded RNAs generated by MITE sequences. (*A*) *Trans*-sense/antisense transcripts formed between two MITE elements inserted in opposite orientations in UTRs of rice genes *Os10g26720* and *Os05g43650*. (*B*) A small RNA has perfect match with the MITE sequence inserted in gene *Os10g26720*. (*C*) The MITE element inserted in the UTR of rice gene *Os10g26720* can form a stem-loop structure by itself. The *OsT*20 sequences are delimited by its TSD sequences "UA."

were generated exclusively by MITEs located in intergenic regions.

MITE-Derived Small RNAs Were Predominantly 24 nt in Length and Their Positions Vary Dramatically in Different MITE Families

We further investigated which parts of MITE sequences generated small RNAs. Since MITE elements (even in the same family) vary in length, relative positions of small RNAs in MITEs were used to locate them. With all verified full-length MITE sequences from all MITE families investigated, MITE-derived small RNAs were found to be evenly distributed over the sequence, with two modest peaks in the terminals and one in the middle (fig. 5*A*). To investigate the variations between different MITE families, elements with identical or nearly identical length from 19 MITE families were analyzed independently. Strikingly, dramatic variations were found on the positions of small RNAs in different MITE families. Eleven of the 19 MITE families generated small RNAs mainly from the terminals (fig. 5*B*), while 6 other families generated small RNAs predominantly from the central region (fig. 5*C*).

The MITE-derived small RNAs are mostly (128,927, 70.1%) 24 nt in length, similar to the MITE-derived small RNAs in Solanaceae species (Kuang et al. 2009). Like other superfamilies, the *Tc1/Mariner* superfamily also generate predominantly 24-nt small RNAs, though many elements from this superfamily have long TIR sequences and are predicted to form stem-loop structures.

Genes Associated with MITE Have Significant Lower Expression Than Those Away from MITEs

The prevalence of MITE-derived small RNAs makes all genes associated with MITEs their potential targets. To illustrate the effects of MITEs, the expression of genes associated with MITEs was compared with that of genes away from MITEs. The gene expression data (RKPM value) were generated through RNA sequencing (Lu et al. 2010). A total of 32,587 genes have an RPKM threshold value of 0.15, which corresponds to a false discovery rate (FDR) and false negative rate of 8%. Among them, 11,059 genes have MITE insertions within 500 bp 5′ to their start codon (upstream), 8,514 genes have MITE insertions within 500 bp 3′ to their stop codon (downstream), and 9,112 genes have MITE insertions in their introns, whereas 13,079 genes are away from MITE insertion. MWU test and *t*-test showed that genes with MITE insertions in upstream, intron, and downstream have significantly lower expression than genes away from MITEs in Nipponbare, respectively ($P < 0.001$). Comparison between the distribution density curves of expression level showed that all genes with MITE insertions have a larger proportion of weakly expressed genes than the genes with no MITE insertions (fig. 6*A*).

To investigate the effects of MITEs on gene expression in another rice cultivar, 93-11 (*O. sativa* ssp. *indica*), MITE sequences were retrieved from its partially sequenced genome. A total of 197,316 MITE elements were discovered

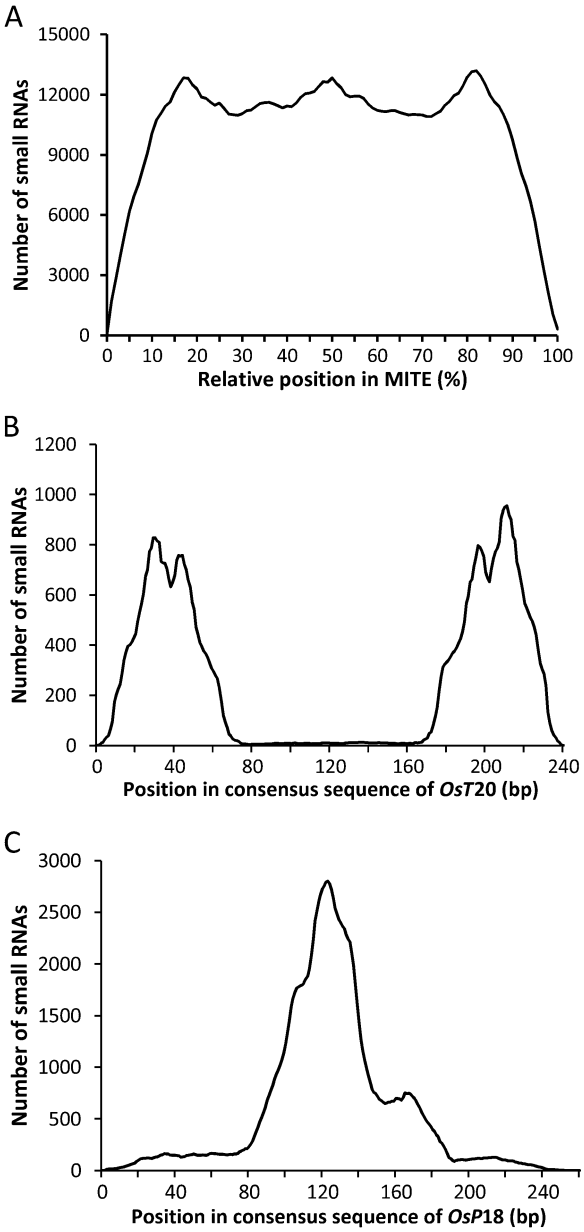

Fig. 5. The positions of small RNAs in MITE sequences. (*A*) The relative positions of small RNAs in all full-length MITEs. (*B*) Small RNAs are mainly generated from the terminals of MITE elements in 11 of the 19 MITE families investigated. Only family *OsT20* is shown. (*C*) Small RNAs are mainly generated from the middle of MITE elements in 6 of the 19 MITE families investigated. Only family *OsP18* is shown.

from the partially sequenced genome. Of the 40,745 genes annotated using GLEAN method, 22,976 (56.4%) genes are associated with MITEs, whereas 17,077 genes are away from MITEs. Among them, 20,164 genes with MITE insertions and 12,265 genes with no MITE insertions have RPKM > 0.15, and they were analyzed further. Like in Nipponbare, the genes with MITE insertion in 93-11 have a larger proportion of weakly expressed genes than the genes without MITE (fig. 6*B*). MWU test and *t*-test both showed that genes with MITE insertions have significantly lower expression than genes away from MITEs in 93-11 ($P < 0.001$).

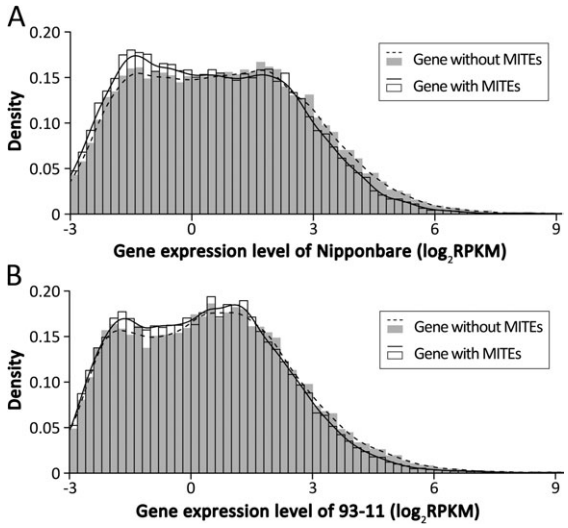

FIG. 6. Probability density curve of gene expression. The unshaded area is for genes associated with MITE and the shaded area is for genes away from MITEs. Genes associated with MITEs have higher proportion with low expression.

We also compared the expressions of alleles between 93-11 and Nipponbare. A total of 13,453 genes have verified alleles in the two rice cultivars and have RPKM > 0.15 from each cultivar. Approximately 16.9% (2,279 genes) of them are differentially expressed between the two cultivars ($P < 0.001$, FDR = 10%): 8.0% have higher expression in Nipponbare and 8.9% have higher expression in 93-11. To study the effects of MITE insertion on expression of individual genes, we identified 372 genes that have MITE insertions in Nipponbare but no MITE in 93-11. No significant differences of expressions of the 372 genes were found between the two cultivars (*t*-test, $P > 0.1$). Nevertheless, 91 of the 372 genes are differentially expressed in the two cultivars but with both upregulation and downregulation in Nipponbare (table 3). On the other hand, there are 623 genes that have MITE insertions in 93-11 but no MITE in Nipponbare. Similarly, no significant differences were found between their expressions in the two cultivars, and both upregulation and downregulation of gene expressions were found in 93-11 (table 3). Note that the upregulation or downregulation is not necessarily caused by MITE insertions since 15.2% of genes are differentially expression regardless of MITE insertions (see above). Therefore, no obvious effects of MITE insertions on expression of individual genes were detected.

MITE-Derived Diversity in Rice Cultivars

Genome-wide polymorphisms generated by MITEs were analyzed in rice cultivars Nipponbare and 93-11. Due to practical reasons, only verified full-length elements were compared between the two cultivars. The 113,840 full-length MITEs (56,390 from Nipponbare and 57,450 from 93-11) were assigned into 70,482 loci. Of them, 43,361 loci have full-length MITEs in both cultivars and 7,485 loci have MITE insertions in one cultivar but not in the other. The remaining 19,636 loci were only found in one cultivar,

Table 3. The Effects of MITE Insertion on Individual Genes.

	Total Number	Differentially Expressed	Upregulated in Nipponbare	Upregulated in 93-11
All genes[a]	13,453	2,279	1,075	1,204
With MITE in Nipponbare[b]	372	91	51	40
With MITE in 93-11[c]	623	113	37	76

[a] All genes that have verified alleles in both cultivars and have at least 0.15 RPKM in RNA sequencing data.
[b] All genes with MITE insertion in the Nipponbare but no MITE insertion in corresponding alleles in 93-11.
[c] All genes with MITE insertion in the 93-11 but no MITE insertion in corresponding alleles in Nipponbare.

while their corresponding loci in the other cultivar could not be confirmed using their flanking sequences, either due to variations of the flanking sequences or sequencing gaps in 93-11 genomic sequences. Therefore, at least 14.8% (7,485/(43,361 + 7,485)) of the full-length MITE elements exhibit presence/absence polymorphism between the two rice cultivars.

Discussion

Efficiency of MITE Identification Using the BLASTN-Based Approach

Many sequences (including gene families) in a genome have multiple copies due to polyploidization or duplication (including segmental duplication). Most duplicated sequences are large and/or have uneven terminals. In contrast, full-length transposable elements have precise terminals. In this study, we developed a method (RSPB) for genome-wide de novo identification of MITEs through identification of groups of short repetitive sequences with precise boundaries. Of the 1,588 groups of repetitive sequences identified from rice genome using this method, the main false positives are solo-LTRs and long transposons, which also have precise boundaries. The solo-LTRs, all with 5-bp TSD sequences, are prone to be misannotated, such as the *MiS5* family in Solanaceae (Kuang et al. 2009).

MITE-hunter, a recently developed program, has excellent success in MITE identification (Han and Wessler 2010). Like all other programs, MITE-hunter identifies repetitive sequences with TIR and TSD structures. We compared the MITE groups identified by MITE-hunter (kindly provided by Dr Y. Han) and those identified using RSPB. There are 37 MITE families detected by RSPB but missed by MITE-hunter. In other words, 37 novel MITE families were discovered in this study. Compared with other methods, RSPB is less likely to miss MITE families with unusual TIR and TSD sequences (such as short TIR sequences). Twenty-one MITE families identified by MITE-hunter have fewer than five copies and were not regarded as MITE by our definition. Besides above low copy MITEs, 26 additional MITE families identified by MITE-hunter were missed by RSPB.

In summary, the RSPB is a powerful program for MITE identification, with some minor disadvantages: 1) it will miss MITE families with only one or two full-length copies

in a database and 2) it takes a long running time (9 days for RSPB running on a 4-core PC vs. 55 h for MITE-hunter on a cluster to finish the 380-Mb rice genome). The combination of MITE-hunter and RSPB is predicted to identify a vast majority, if not all, MITE families in a genome, with no prior information required.

MITEs in Rice Genome

After combination of MITEs detected by RSPB and those identified previously, 178,533 MITE-related sequences were detected in the genome of rice cultivar Nipponbare. MITE sequences were recommended to be classified into the same family using the 80–80–80 rule: at least 80% (>80 bp) of the entire sequence has more than 80% nucleotide identity (Wicker et al. 2007; Han and Wessler 2010). Such classification was for practical purposes but has no evolutionary meaning. In this study, we focused on evolutionary genomics of MITEs and used a broad sense of definition of MITE family. All sequences with significant similarity (BLASTN e value $< 10^{-10}$) were grouped into a family, and the 178,533 MITE-related sequences were grouped into 338 families. Nevertheless, we cannot rule out the possibility that some elements within a MITE family were originated independently through deletions from different but related long transposons.

MITEs Were Generated Sporadically

Groups of equally related elements were frequently encountered in a MITE family. Such groups of elements were either generated through massive amplifications in a short period or reticulate evolution resulted from sequence exchanges among elements. Four-gamete (Librado and Rozas 2009) and Geneconv analyses (Sawyer 1989) showed that sequence exchanges between different MITE elements were not prominent and should not account for equally related elements (data not shown). Therefore, the equally related elements in a family or a clade were the results of amplification burst. This was confirmed by the wave-like histograms of pairwise nucleotide diversity among elements (fig. 2). Sequence analysis of different clades in a MITE family suggested that there might be several rounds of amplification bursts within a family. Such bursts could not be explained by genome-wide duplications. First, the burst time varied dramatically for different MITE families. Second, the burst results in hundreds or thousands of copies for some MITE families, which should not be generated by a few times of genome-wide duplications. We hypothesized that the MITE amplifications in rice genome were sporadic: they remained inactive most time in evolutionary history and their activation was intermittent. The temporal activation of MITEs could be caused by "genome shock" or temporal activation of cognate transposase by unknown mechanisms. Genome shock may activate only one or a few transposons. For example, *mPing* has been the only MITE family in rice genome to be activated by irradiation, cell culture, or recent domestication (Jiang et al. 2003; Nakazaki et al. 2003; Naito et al. 2006). It remains unknown why and how only the cognate transposase of *mPing* is currently activated.

The Distribution and Selection of MITEs in Rice Genome

MITEs may upregulate the expression of genes if they contain regulatory motifs (Yang et al. 2005; Naito et al. 2009) or downregulate gene expression via MITE-derived small RNAs (Kuang et al. 2009). The insertion of MITEs near genes will be under selection as long as they impose any effects on gene expression or function.

At least 23,623 (58.2%) genes of rice cultivar Nipponbare have MITE insertions in their introns or flanking (500 bp) regions. Though genome-wide analysis showed MITEs are evenly distributed in genic region and intergenic region, we cannot rule out the preference or avoidance of genic region for some MITE families. A previous study showed that the *mPing* family preferentially transposed into the upstream (within 500 bp) of genes (Naito et al. 2009). Selection was ruled out as the cause, since the plants after MITE transposition had not been under either natural or artificial selections. Similarly, gametophyte or dominant lethality cannot explain such preference of insertions (Naito et al. 2009). Unlike the currently active *mPing* family, the vast majority of MITEs in the Nipponbare genome are relatively old and are the results of selection or genetic drift after a long evolutionary history. PCR screening suggested that a large proportion of MITEs are fixed (present in all genotypes) in *O. sativa* (data not shown). The fixed loci may be due to selections or genetic drift. On the other hand, MITE insertions at some loci might be selected against. Therefore, the current distribution of MITEs is the results of both types of selections as well as genetic drift.

MITE-Derived Small RNAs

MITE-related sequences generate 23.5% of all small RNAs in rice. Theoretically, MITEs near genes should be more likely to be transcribed (e.g., as untranslated region [UTR] sequences of plant genes) than those in intergenic regions and thus generate more small RNAs. Surprisingly, 51.8% of MITE-derived small RNAs were generated exclusively by MITEs located away from genes. They may be derived from the large number of transcriptional active regions with no protein-coding sequences (Lu et al. 2010).

When all MITEs were considered, MITE-derived small RNAs seem evenly distributed along MITE sequences (fig. 5). However, the positions of small RNAs from different MITE families vary dramatically. Some MITE families produce MITE from their terminals, with no or very few small RNAs from the central region. In contrast, some MITE families produce small RNAs mainly from their central region, with none or very few from their terminals. Such variations cannot be attributed to type of superfamilies, MITE length or TIR length (data not shown). It will be interesting to investigate the mechanisms for such variations in future studies.

The MITE-derived small RNAs from rice genome are predominantly (70.1%) 24 nt in length, similar to those from Solanaceae species but unlike those from wheat (Kuang et al. 2009; Cantu et al. 2010). The small RNAs derived from these MITE sequences were often predicted as miRNAs by computer programs since some MITEs (such as the

Tc1/Mariner superfamily) are mainly composed of TIR sequences, which are predicted to form stem-loop structures similar to that of miRNA genes. In literature and MiRbase, at least 807 MITE-derived small RNAs were annotated as miRNAs. However, the small RNAs derived from these MITEs are predominantly 24 nt rather than 21 nt in length. MITE-derived small RNAs are produced from different positions of predicted stem-loop structures. Furthermore, small RNAs derived from MITEs were shown to be generated by the RNAi biogenesis pathway (Kuang et al. 2009). The gene transcript (as UTRs of plant genes), siRNA biogenesis pathway, poor conservation in different species, and nonspecific targets indicate that these MITE-derived small RNAs are most likely siRNAs rather than miRNAs (Meyers et al. 2008).

The Effects of MITEs on Gene Expression

Our genome-wide analysis showed that MITE insertions are associated with reduced expression of nearby genes. The general trend of downregulation does not rule out the possibility that some MITEs can increase the expression of nearby genes (Yang et al. 2005; Naito et al. 2009). Nevertheless, downregulation of nearby genes by MITEs is consistent with the presence of a large number of MITE-derived small RNAs. Genes with MITE insertions may become the potential targets of MITE-derived small RNAs. Genome-wide analysis of transposons also suggested that methylated transposons decrease the expression of nearby genes through siRNA pathways (Hollister and Gaut 2009; Hollister et al. 2011). The effects of transposable elements (including MITEs) on individual genes and phenotypic diversity of a species will be an interesting field for future research (Liu et al. 2004).

However, comparison between genes with MITE insertions in one cultivar but without MITEs in the other showed no effects of MITE insertion on gene expression. Analysis on individual genes, on the other hand, showed no consensus results. Some alleles with MITE insertions showed increased expression, some showed decreased expression, whereas the majority showed no significant change. The increase or decrease of the expression of some alleles with MITE insertions is not necessarily caused by MITE insertion per se, since 16.9% of genes are differentially expressed regardless of MITE insertions. Divergent genetic backgrounds (different subspecies) and small sample size (i.e., limited number of genes with presence/absence polymorphisms) may obscure the results of such analysis.

Diversity Generated by MITEs

Comparison between the two sequenced rice genomes showed that MITE generated a large amount of diversity for *O. sativa*. The diversity generated by MITEs was also demonstrated using transposon-display technique (Casa et al. 2000, 2004). Compared with old MITEs, which are frequently fixed due to selection or genetic drift, new insertions of MITEs are more likely to produce polymorphisms in different genotypes. When these MITEs are inserted close to genes, they may either upregulate or downregulate genes' expression (see above). Therefore, distinct sets of genes in different genotypes are under the regulation of MITEs, which may contribute considerable phenotypic diversity to a species.

Supplemental Material

Supplementary table 1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Drs Ning Jiang, Feng Li, and Junhua Peng for their critical review of this manuscript and Dr Y. Han for sharing the results of MITE-hunter. This work was supported by the "973" National Key Basic Research Program grant no. 2009CB119000, China Transgenic grant no. 2009ZX08009-045B, and the National Natural Science Foundation of China grant no. 30921002.

References

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

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol.* 11:R106.

Cantu D, Vanzetti LS, Sumner A, Dubcovsky M, Matvienko M, Distelfeld A, Michelmore RW, Dubcovsky J. 2010. Small RNAs, DNA methylation and transposable elements in wheat. *BMC Genomics* 11:408.

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

Casa AM, Nagel A, Wessler SR. 2004. MITE display. *Methods Mol Biol.* 260:175–188.

Chen Y, Zhou F, Li G, Xu Y. 2009. MUST: a system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*. *Gene* 436:1–7.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14:1394–1403.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

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

Feng Q, Zhang Y, Hao P, et al. (74 co-authors). 2002. Sequence and analysis of rice chromosome 4. *Nature* 420:316–320.

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

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, Swamy L, Wessler SR. 2003. Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* 163:747–758.

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

Han Y, Wessler SR. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 38:e199.

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

Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc Natl Acad Sci U S A.* 108:2322–2327.

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.

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

Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* 16:418–420.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12:656–664.

Kikuchi K, Terauchi K, Wada M, Hirano HY. 2003. The plant MITE *mPing* is mobilized in anther culture. *Nature* 421:167–170.

Kuang H, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Ouyang S, Jiang J, Buell CR, Baker B. 2009. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res.* 19:42–56.

Le QH, Wright S, Yu Z, Bureau T. 2000. Transposon diversity in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 97:7376–7381.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

Liu J, He Y, Amasino R, Chen X. 2004. siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes Dev.* 18:2873–2878.

Lu T, Lu G, Fan D, et al. (12 co-authors). 2010. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* 20:1238–1249.

Lyons M, Cardle L, Rostoks N, Waugh R, Flavell AJ. 2008. Isolation, analysis and marker utility of novel miniature inverted repeat transposable elements from the barley genome. *Mol Genet Genomics.* 280:275–285.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A.* 101:12404–12410.

Mao L, Wood TC, Yu Y, et al. (12 co-authors). 2000. Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. *Genome Res.* 10:982–990.

Meyers BC, Axtell MJ, Bartel B, et al. (21 co-authors). 2008. Criteria for annotation of plant MicroRNAs. *Plant Cell* 20:3186–3190.

Meyers BC, Chin DB, Shen KA, Sivaramakrishnan S, Lavelle DO, Zhang Z, Michelmore RW. 1998. The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. *Plant Cell* 10:1817–1832.

Mizuno H, Ito K, Wu J, Tanaka T, Kanamori H, Katayose Y, Sasaki T, Matsumoto T. 2006. Identification and mapping of expressed genes, simple sequence repeats and transposable elements in centromeric regions of rice chromosomes. *DNA Res.* 13:267–274.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 5:621–628.

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 U S A.* 103:17620–17625.

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

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.

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.

Piriyapongsa J, Marino-Ramirez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* 176:1323–1337.

Ramsköld D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol.* 5:e1000598.

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

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.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol.* 6:526–538.

Shan X, Liu Z, Dong Z, Wang Y, Chen Y, Lin X, Long L, Han F, Dong Y, Liu B. 2005. Mobilization of the active MITE transposons *mPing* and *Pong* in rice by introgression from wild rice (*Zizania latifolia* Griseb.). *Mol Biol Evol.* 22:976–990.

Song WY, Pi LY, Bureau TE, Ronald PC. 1998. Identification and characterization of 14 transposon-like elements in the non-coding regions of members of the *Xa21* family of disease resistance genes in rice. *Mol Gen Genet.* 258:449–456.

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.

Toedling J, Ciaudo C, Voinnet O, Heard E, Barillot E. 2010. Girafe—an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads. *Bioinformatics* 26:2902–2903.

Tu Z. 2001. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proc Natl Acad Sci U S A.* 98:1699–1704.

Wang H, Chua NH, Wang XJ. 2006. Prediction of trans-antisense transcripts in *Arabidopsis thaliana*. *Genome Biol.* 7:R92.

Wessler SR. 1998. Transposable elements and the evolution of gene expression. *Symp Soc Exp Biol.* 51:115–122.

Wicker T, Sabot F, Hua-Van A, et al. (13 co-authors). 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 8:973–982.

Wright SI, Agrawal N, Bureau TE. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Res.* 13:1897–1903.

Wu J, Yamagata H, Hayashi-Tsugane M, et al. (21 co-authors). 2004. Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell* 16:967–976.

Yang G, Dong J, Chandrasekharan MB, Hall TC. 2001. *Kiddo*, a new transposable element family closely associated with rice genes. *Mol Genet Genomics*. 266:417–424.

Yang G, Hall TC. 2003a. *MDM-1* and *MDM-2*: two mutator-derived MITE families in rice. *J Mol Evol*. 56:255–264.

Yang G, Hall TC. 2003b. MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Res*. 31:3659–3665.

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, Nagel DH, Feschotte C, Hancock CN, Wessler SR. 2009. Tuned for transposition: molecular determinants underlying the hyperactivity of a *Stowaway* MITE. *Science* 325:1391–1394.

Zerjal T, Joets J, Alix K, Grandbastien MA, Tenaillon MI. 2009. Contrasting evolutionary patterns and target specificities among three *Tourist*-like MITE families in the maize genome. *Plant Mol Biol*. 71:99–114.

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 U S A*. 97:1160–1165.