Tracking the genome-wide outcomes of a transposable element burst over decades of amplification

Lu Lu,∗,†, Jinfeng Chen,‡,§, Sofia M. C. Robb,∥,⊥, Yutaka Okumoto,∥ Jason E. Stajich,** and Susan R. Wessler∗,†

∗Department of Botany and Plant Sciences, University of California, Riverside CA 92521; †Institute for Integrative Genome Biology, University of California, Riverside, CA 92521; ‡Department of Microbiology and Plant Pathology, University of California, Riverside, CA 92521; and ∥Graduate School of Agriculture, Kyoto University, Kitashirakawa-oiwake Sakyo, Kyoto 606-8502, Japan

To understand the success strategies of transposable elements (TEs) that attain high copy numbers, we analyzed two pairs of rice (Oryza sativa) strains, EG4/HEG4 and A119/A123, undergoing decades of rapid amplification (bursts) of the class 2 autonomous mPing element and the nonautonomous miniature inverted repeat transposable element (MITE) mPing. Comparative analyses of whole-genome sequences of the two strain pairs validated that each pair has been maintained for decades as inbreds since divergence from their respective last common ancestor. Strains EG4 and HEG4 differ by fewer than 160 SNPs and a total of 264 new mPing insertions. Similarly, strains A119 and A123 exhibited about half as many SNPs (277) as new mPing insertions (518). Examination of all other potentially active TEs in these genomes revealed only a single new insertion out of ∼40,000 loci surveyed. The virtual absence of any new TE insertions in these strains outside the mPing bursts demonstrates that the Ping/mPing family gradually attains high copy numbers by maintaining activity and evading host detection for dozens of generations. Episomal elements necessary for their own transposition and for transposition of autonomous and nonautonomous elements (10). Autonomous elements encode the protein(s) necessary for their own transposition and for transposition of nonautonomous element members (10). To our knowledge, none of these actively transposing elements has attained very high copy number, likely because their mutagenic behavior can negatively impact host survival and preclude significant amplification (1).

In plant genomes, two types of TEs have attained high copy numbers: LTR retrotransposons and class 2 miniature inverted repeat transposable elements (MITEs) (1). LTR retrotransposons are class 1 (RNA) elements that move through an RNA intermediate. Most are long (>3 kb) elements that accumulate in intergenic regions where they form clusters of nested insertions (11). LTR retrotransposons are largely responsible for the dramatic differences in genome sizes between related plant species, e.g., the sixfold size difference between the maize and rice genomes (4, 12). In contrast, MITEs are short (<600 bp) nonautonomous elements that are usually deletion derivatives of autonomous DNA transposons (1). MITEs can generate significant allelic diversity, as has been documented in several grasses (e.g., maize, rice, wheat), in which they attain high copy numbers and insert preferentially into or near genes (1, 13).

MITEs are numerically the most abundant TE type in rice (Oryza sativa) with over 23,500 (~58%) MITE-associated genes (14). Given their typical insertion near genes, MITEs are a major source of 24-nt siRNAs, which direct and maintain DNA methylation via the RNA-directed DNA methylation pathway (RdDM) (15–17). For example, transcription of intrinsic MITEs can generate siRNAs that target CHH methylation of identical or nearly identical copies scattered throughout the genome (15). In this way, MITEs are largely responsible for the high levels of CHH methylation in the 5′ and 3′ regulatory regions of rice.


Reviewers: D.L., Purdue University; and N.M.S., University of Minnesota.

The authors declare no conflict of interest.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. JQ500000000 and JQ5000000000 and the BioProject database, https://www.ncbi.nlm.nih.gov/bioproject/ (accession nos. PRJNA184949 and PRJNA264731).

†L.L. and J.C. contributed equally to this work.

∥To whom correspondence should be addressed. Email: susan.wessler@ucr.edu.

This article contains supporting information online at www.pnas.org/cgi/suppl/doi:10.1073/pnas.1716459114/DCSupplemental.

Significance

Rice (Oryza sativa) has a unique combination of attributes that made it an ideal host to track the natural behavior of very active transposable elements (TEs) over generations. In this study, we have exploited its small genome and propagation by self or sibling pollination to identify and characterize two strain pairs, EG4/HEG4 and A119/A123, undergoing bursts of the nonautonomous miniature inverted repeat transposable element mPing. Comparative sequence analyses of these strains have advanced our understanding of (i) factors that contribute to sustaining a TE burst for decades, (ii) features that distinguish a natural TE burst from bursts in cell culture or mutant backgrounds, and (iii) the extent to which TEs can rapidly diversify the genome of an inbred organism.

www.pnas.org/cgi/doi/10.1073/pnas.1716459114

PNAS Early Edition | 1 of 10
Proportions of shared insertion sites | mPing and BHEG4 and A119/A123 movement was initially demonstrated in rice cell is an avoidance of exon OsDRM2 Illumina paired-end sequencing libraries were se-
O. sativa and Insertions in the Two Strain Pairs. class 2 superfamily that gene expression led to the surprising bursts (27). One coverage,
(18), japonica OsDCL3a has no impact on is " Fig. 1. proportions of shared Pong mPing insertions and SNPs indicate strain pairs have been maintained by self or sibling pollination. (A) Schematic structures of mPing, Pong, and Pong. Both autonomous elements encode ORFs. The black horizontal arrows show the positions of the coding regions of both ORFs. Numbers below indicate transcription start sites or termination sites of the ORFs. Red triangles represent TIRs, and white stripes represent introns. The orange and black arrowheads indicate the position of primers for qRT-PCR and bisulfite PCR, respectively. mPing and Pong share 5' ends of 253 bp and 3' ends of 177 bp. (B) Schematics of the presumed lineages of the two strain pairs (see text for details). (C) Venn diagrams of the number of unique and shared SNPs found in strain pairs EG4/HEG4 or A119/A123 compared with the NB reference genome. (D) Venn diagrams of the number of homologous nonreference mPing insertion sites that are shared or private in each strain pair.

While one successful strategy of mPing is an avoidance of exon insertions, this study addresses another important but poorly understood aspect of a TE burst, that is, how the burst is sustained for generations without triggering host silencing of the transposition machinery. We exploited the availability of two pairs of rice strains, EG4/ HEG4 and A123/A119, which previously have been shown to have an actively bursting mPing (27, 30). Comparative sequence analysis demonstrates that each strain pair was derived from a common an-
cessor and maintained by self or sibling pollination for decades. Since divergence, Pong has continued to produce transposase that has cat-
yzed massive mPing transposition while all other potentially active TE's have remained inactive. Analyses of DNA methylation and other epigenetic marks and Pong gene expression led to the surprising finding that mPing continues to transpose and increase in copy number despite being highly methylated before and during the burst. Importantly, recognition of mPing has no impact on Pong activity, because these elements do not share any coding sequences.

Results

Sequencing and Variant Analysis of Select Rice Strains. This study was possible due to the availability of two pairs of rice strains that had previously been shown to be in the midst of mPing bursts (27). One pair, HEG4 and EG4, is a direct descendant of Gimbozu accession EG4+ arising from two single seeds and reported to have been maintained as separate lines by self or sibling pollination for ~30 y (Fig. 1B). The relationship between the second pair, A123 and A119, had not been documented before this study. According to incomplete breeding records, their last common ancestor (LCA, called “A123*” in this study) was a local variety (also called a “landrace”) that was widely cultivated in northern Japan during the early part of the 1900s (31). Like EG4*, A123* is no longer available. In 1912, A119 was initiated as a pure line from A123*, much like HEG4 from EG4*. Therefore, whether the two existing strains (called “A119” and “A123”) had been maintained as pure lines since their divergence from A123* ~100 y ago needed to be validated (Fig. 1B). Because even closely related strains of O. sativa var. japonica differ by tens to hundreds of polymorphisms (32), comparative analysis of the two strain pairs would reveal any accidental outcross during their propagation.

EG4 vs. HEG4. Illumina paired-end sequencing libraries were se-
sequenced to produce 68× genome coverage for EG4 and 193× coverage for HEG4 (Table S1), which was suitable for identifi-
cation of high-confidence polymorphisms that could distinguish the two strains. A draft genome assembly of the HEG4 strain was de novo assembled from these short reads (Table S2). Identi-
fication and analysis of sequence variants was performed using the Nipponbare (NB) genome as a reference (12). SNPs were iden-
tified with the Genome Analysis Tool Kit (GATK) following best practices to reduce false positives and low-quality variant calls (33). A high-confidence polymorphism set for strain analyses was produced by further removing ambiguous or heterozygous SNP sites in any individual. Based on this dataset, the strains EG4 and HEG4 differ from each other by ~159 SNPs and share 109,378 SNPs compared with NB (Fig. 1C). These observed polymorphism differences are consistent with a history of a recent shared common ancestor (called “EG4+”) and subsequent propagation of each line by self or sibling pollination (without outcrossing) (Fig. 1B).

A123 vs. A119. To determine whether strains A119 and A123 have been maintained as pure lines since 1912, as stated in breeding 
records, the strains were sequenced to 62× coverage and differ from each other by ~159 SNPs and share 109,378 SNPs compared with NB (Fig. 1C). These observed polymorphism differences are consistent with a history of a recent shared common ancestor (called “EG4+”) and subsequent propagation of each line by self or sibling pollination (without outcrossing) (Fig. 1B).

Comparative Analysis of mPing Insertions in the Two Strain Pairs. The majority of characterized O. sativa strains have ~1–50 mPing copies (22, 27, 34). To study the dramatic increase in mPing copy
number in each of the strain pairs, insertion sites were identified from paired-end Illumina sequence reads with RelocuTE and classified as heterozygous or homozygous with CharacTErizer (Dataset S1) (35). Only homozygous insertions were further classified as shared or private alleles (Fig. 1D) because it could not be determined whether heterozygous insertions resulted from germinal (heritable) or somatic events. As mentioned above, EG4 and HEG4 arose from the common ancestor EG4* about 20 y ago. The number of shared (338) vs. unshared (264) insertion sites in EG4 vs. HEG4 is consistent with a scenario in which mPing amplification occurred in the EG4* ancestor followed by divergence of the strains. Unshared sites are additional insertions that accumulated as the mPing burst continued independently in EG4 and HEG4. These private mPing insertions are classified as new insertions, as no evidence of an excision footprint is found at the orthologous position in the strain lacking an mPing site.

In contrast, the proportion of shared (23) to unshared (518) mPing insertion sites in A119 vs. A123 (Fig. 1D) was quite different, given their recent divergence inferred from SNPs. With 23 shared mPing elements, the A123* ancestor would have had a similar copy number of mPing elements as the great majority of extant japonica strains, in which mPing abundance is low because transposition is extremely rare (22, 27). These data suggest that the LCA of the Pong family began independently in A119 and A123 after divergence from their LCA. Furthermore, comparison of two strain pairs offers dramatic illustrations of the extent of TE-mediated genome diversity that is possible over a few decades in a self-pollinating species. For both pairs, the number of new homozygous mPing insertions exceeded the number of new homozygous point mutations by almost 2 to 1 (1.66 for EG4/HEG4 and 1.87 for A119/A123).

**Ping Copy Number and Insertion Site Comparison in the Four Strains.** Ping harbors two genes, ORF1 and TPASE, that are both necessary for Ping and mPing transposition (Fig. 1A) (25, 26). Under normal growth conditions, transposition of mPing is rarely seen in the genome of the reference strain NB, which contains only one Ping element (22). Ping copy number in the four rice strains was determined using two independent methods: DNA sequencing and DNA blot analysis. Both techniques identified multiple Ping insertions in the four strains, with seven copies in EG4, HEG4, and A119 and 10 copies in A123 (Fig. 2A and B and Table S3), as is consistent with a previous report (30). Whereas EG4 and HEG4 share the same seven Ping loci, A119 and A123 have only one Ping locus in common, a locus that is also shared with EG4 and HEG4 (Fig. 2A). Although a DNA blot of genomic DNA from EG4 and HEG4 probed with internal Ping sequences revealed that one of seven bands is polymorphic (Fig. 2B), sequence analysis demonstrated that the size difference is due to an insertion of mPing adjacent to one of the Ping elements in HEG4 (Fig. S1). Taken together, the Ping copy number and locations identified in the strains suggest that the LCA of EG4 and HEG4, EG4*, had seven Pongs, consistent with the divergence of EG4 and HEG4 in the midst of the Ping/mPing burst. In contrast, the LCA of A123 and A119, A123*, had only one Ping that subsequently amplified in copy number after their divergence to enable a genomic environment in which mPing could begin to burst.

**Comparative Analysis of Pong Insertion Sites in the Four Strains.** Pong is the closest element by sequence similarity to Ping in the rice genome, and at least five nearly identical copies of Pong can be identified in all strains examined (Fig. 1A) (22, 36). Previous research has demonstrated that Pong is a very active element; its encoded proteins catalyze the transposition of mPing in yeast and Arabidopsis transposition assays at even higher frequencies than Ping proteins (25, 26). However, Pong insertion sites in several japonica strains are nearly invariant, suggesting that Pong is epigenetically silenced in these genomes (22). Further, the observed high-frequency transposition of Pong in rice cell culture, in which Pong-encoded proteins also catalyzed the transposition of mPing (22), supports the contention that Pong proteins are not normally expressed but can be expressed if epigenetic regulation is relaxed. Examination of the Pong sites in EG4, HEG4, A123, and A119 found no evidence of new insertions since divergence of the strain pairs. Specifically, all strains share five Pong loci (Fig. 2C and D and Table S4), while an additional Pong locus is shared between A123 and A119. This Pong locus, containing an element inserted in a Tyl1-copia retrotransposon, is also found in several other japonica strains (Omachi, Nongken S8, and Kitake) (Table S4), indicating that it originated by outcrossing before their divergence from A123* and not by recent transposition (37, 38).

**Ping and Pong Transcripts, Methylation, and Chromatin Modifications.** Quantification of transcript levels and chromatin modifications indicate that Ping loci contain two actively transcribed genes while Pong is silenced in strains NB and EG4 (Fig. 3). Ping ORF1 and TPASE transcripts were detected by qRT-PCR and TPASE transcripts were detected by qRT-PCR at levels roughly proportional to their copy numbers in NB (one Ping) and EG4 (seven identical Pongs) (Fig. 3A, Left). The same samples yielded negligible transcript levels from the six Pong elements in NB and the five Pong elements in EG4 (Fig. 3A, Right).

The patterns of three epigenetic marks including gene body methylation and the chromatin modifications H3K4me3 and H3K9me2 were consistent with the transcript levels. Bisulfite sequencing (BS-seq) of Ping ORF1 and TPASE from both NB and EG4 revealed CG methylation but little CHG or CHH methylation (Fig. 3B, Left). This pattern resembles the gene-body methylation pattern of many protein-coding genes (39). Using the same DNA samples, BS-seq of Pong coding regions revealed hypermethylation in the CHG context of both ORFs (Fig. 3B, Right).
The chromatin status of Ping and Pong elements in NB was elucidated by performing ChIP-sequencing (ChIP-seq) using anti-H3K4me3 and anti-H3K9me2 antibodies. In plants, histone H3 trimethylated at K4 is associated with active transcription (41), whereas H3 dimethylation at K4 is correlated with transcriptional silencing of transposons and other repetitive sequences (42, 43). Enrichment of the H3K4me3 mark was found in the first exon and intron of Ping ORF1 (Fig. 3C, Left), similar to profiles observed in active rice genes (Fig. 3D). In contrast, the entire Pong element was modified by the repressive H3K9me2 mark (Fig. 3C, Right), similar to modification patterns of silenced retrotransposons (Fig. 3E). These results indicate that Ping is transcriptionally active in the high-copy-number strain and in NB, whereas Pong is epigenetically silenced in all strains analyzed.

Methylation of mPing and Shared mPing/Ping Sequences. Heavy methylation of mPing elements in both NB and EG4 was visualized by dot plots of the results of PCR/BS-seq of several independent mPing clones (Fig. 4A). The mPing element was almost completely CG methylated, and lower but significant levels of cytosine methylation in CHG and CHH contexts were also recorded. A more global picture of methylation was derived from analysis of individual mPings using paired-end reads of whole-genome BS-seq of DNAs from NB, EG4 (Fig. 4B), and A119 (Fig. S2). Methylation profiles of individual mPings showed that most mPing elements are heavily CG methylated (80% of the locus on average) and moderately CHG methylated (40–80%). Asymmetric CHH methylation varied among mPing copies from less than 5% to over 70%.

Methylation of asymmetric CHH sites in dividing cells is orchestrated by 24-nt siRNAs (44). Consistent with high CHH methylation of mPing is the accumulation of siRNAs, predominantly 24 nt, which map directly to mPing sequences in both NB and EG4 (Fig. 4C). The almost 10-fold difference in mPing copies between EG4 and NB may explain the higher levels of siRNAs in EG4. These data indicate that mPing is detected by host surveillance in both low- and high-copy-number mPing strains.

The first 253 bp of Ping (containing sequences upstream of ORF1) are nearly identical with mPing (Fig. 4D). Because ORF1 is an actively transcribed gene, it was of interest to determine the extent of methylation in this region in Ping loci. Recall that NB has a single Ping element, and EG4 has seven identical Pings. To determine the methylation state of sequences shared between Ping and mPing, the first 295 bp of each of the eight Ping

![Graph](https://www.pnas.org/cgi/doi/10.1073/pnas.1716459114)
elements, which contain all sequences upstream of ORF1 transcripts, were amplified from bisulfite-treated NB and EG4 genomic DNA with primers from unique flanking DNA. Dot plot visualization of 14–20 clones of each Ping compared with the overlapping mPing regions revealed uniformly high levels of methylation over the first ~125 bp, diminishing to different extents for individual Pings (Fig. 5). Interestingly, sequences downstream of the breakpoint at position 253 are almost completely unmethylated in all Pings, as are unshared upstream sequences adjacent to the start of ORF1 transcripts at position 296.

The 5′ end of Ping is characterized by almost total CHH methylation (Fig. 5), likely due to the abundance of 24-nt siRNAs derived from the nearly identical mPing sequences (Fig. 4C and Fig. S3). Additional support for the involvement of mPing siRNAs in the methylation of Ping sequences comes from the absence of siRNAs in regions of Ping not shared with mPing (Figs. S3 and S4). Similarly, few siRNAs mapped to Pong loci (Fig. S4B). However, low levels of Pong siRNAs indicate that the RdDM pathway, which requires siRNAs, is not required to silence Pong.

**Comparative Analysis of Other TE Insertion Sites in the Two Strain Pairs.** Several rice TEs have been reported to be active in rice cell culture, in certain mutant backgrounds, or in progeny of interspecific hybrids. Elements active in these situations include the class 2 (DNA) TEs dTok (45), nDart (46), nDaiz (47), and mGing (48) and the class 1 retrotransposons Tos17 (49), lullaby (50), Osr7, Osr17, and Osr23 (51), and Karma (52). To assess whether these and other potentially active rice TEs have transposed since the divergence of the two strain pairs, we performed both experimental (transposon display) and computational genome-wide analyses. Transposon display of mPing insertion sites provides dramatic visual evidence of recent transposition in the four strains (Fig. S5). Consistent with the comparative sequence analysis presented above, HEG4 (lane 2) and EG4 (lane 3) have both shared and unshared mPing insertions, whereas most of the mPing insertions in A119 (lane 4) and A123 (lane 5) are unshared. In contrast, for 12 other rice TEs (Fig. S5), insertions were readily identified that are shared by HEG4 and EG4 (red arrows), shared by A123 and A119 (yellow arrows), shared by the four strains but not NB (blue arrow), or present only in NB (green arrow). For all TEs tested, no polymorphisms were detected between EG4 and HEG4 or between A123 and A119.

To determine whether any other TE elements had transposed since the divergence of the two strain pairs, a comprehensive comparative sequence analysis using RelocaTE (35) was performed on over 40,000 TE insertion sites for over 800 rice TE
families (Dataset S2). This analysis revealed that all sites were shared in EG4 and HEG4, while A123 had one unique insertion of the retrotransposon Dasheng that was not present in A119 (Dataset S2).

Comparative Analysis of Large Structural Variations in the Two Strain Pairs. TEs have been associated with other classes of genomic rearrangements including deletions, duplications, inversions, and translocations (53, 54). The very first TE discovered by McClintock, Ds, was initially identified through its ability to promote chromosome breakage (55). To investigate whether the mPing burst contributed in any significant way to genome rearrangements, genome-wide comparisons of structural variations (SVs) larger than 100 bp (other than mPing or Ping insertions) were performed for each strain pair using the newly assembled genomes HEG4 and A123 as references. In total, there were four SVs in HEG4-EG4 and three in A123-A119. Five of the seven are deletions in HEG4 (2), EG4 (1), A123 (1), and A119 (1) (Table S5). None of these deletions occurred in the vicinity of any mPing insertions, and only one of the seven, the previously noted Dasheng in A123 (vs. A119), was due to a TE insertion. Validation of these SVs by PCR is shown in Fig. S6. Several SVs affected coding sequences, resulting in full-gene or exon disruptions (Table S5).

In contrast, the most extensive SV, an ∼120-kb inversion found exclusively in HEG4, correlates with the presence of multiple mPing elements (Fig. 6A). Validation of this inversion by PCR is shown in Fig. 6B. Comparison of this region in NB and EG4 reveals no inversion and near structural identity except for two mPing insertions in EG4 (Fig. 6A, orange boxes). The inversion in HEG4 is flanked by two full-length mPings at one end and by one full-length and one truncated mPing at the other end. A hypothesized intermediate structure, based on signatures of breakpoint sequences, is also shown (Fig. 6A, asterisk). A scenario for the origin of the inversion is included in Fig. S7. Based on this scenario, the inversion in HEG4 is likely generated through template switching during DNA replication, facilitated by the three clustered mPing insertions.

Discussion
Rice has a unique combination of attributes that made it an ideal host to track the natural behavior of very active TEs over generations. In this study, we have exploited its small genome and propagation by self or sibling pollination to identify and characterize two strain pairs, EG4/HEG4 and A119/A123, undergoing massive amplification (burst) of the mPing element. Availability of these four genome sequences facilitated inference of the TE content of their last common ancestors. Comparative analyses of

![Fig. 5. DNA methylation of the 5' shared sequences between mPing and Ping in NB and EG4. Dot plots display DNA methylation of the region from 1 to 295 bp of Ping. Bisulfite-treated DNAs from NB and EG4 were amplified using a forward primer in unique flanking sequences and a reverse primer inside Ping. The PCR fragments were sequenced, and 14–20 bisulfite clones were compared for each. The labels refer to the genomic locus of the seven Ping copies in EG4 and one copy in NB. The gray bar indicates sequences shared by mPing and Ping; the open box indicates unshared adjacent sequences only in Ping, and the red arrowhead represents the 5' TIR.](#)
these strains have advanced our understanding of (i) factors that contribute to sustaining a TE burst for decades, (ii) features that distinguish a natural TE burst from bursts in cell culture or mutant backgrounds, and (iii) additional features that allow MITEs to attain high copy numbers.

Each Strain Pair Has Been Maintained as Inbreds Since Divergence from Their Last Common Ancestor. Two lines of evidence, the paucity of private SNPs and the high proportion of shared TE insertion sites, confirmed that members of each strain pair are nearly identical. Only 159 private SNPs distinguish EG4 from HEG4, and 277 SNPs separate AI19 from AI23 (Fig. 1C). In contrast, most members of the same rice subspecies, japonica, usually differ by well over 80,000 SNPs (32). The EG4/HEG4 lineage differs from AI19/AI23 by ~60,000 SNPs (Fig. S8). Analysis of the tens of thousands of TE loci (other than Ping and mPing loci) common to EG4 vs. HEG4 or AI19 vs. AI23 found that all TE insertions were shared except for a single Dasheng locus in AI23 but not in AI19. In contrast, there are almost 200 polymorphic TE insertion sites when the strain pairs are compared with each other, including 23 polymorphic Dasheng loci (Dataset S2). We interpret these data as indicating that members of each strain pair have been maintained as inbreds since divergence from their LCA because even a single outcross would have substantially increased the number of private SNPs and polymorphic TE insertion sites (Fig. 1B, EG4*, AI23*).

Analysis of the timing of the respective mPing bursts in the two strain pairs shows a different pattern. EG4 and HEG4 share 338 mPing loci and all seven Ping loci, indicating that the burst was well underway in their LCA, EG4*. In contrast, the LCA of AI123/AI19, AI23* had a single Ping locus and ~23 mPing loci, resembling many extant japonica strains that have between 1 and 50 mPings (22, 27, 34). Thus, the increase in copy numbers of both Ping and mPing elements occurred after the divergence of AI23 and AI19 from AI23*, and the potential to burst was inherited in both lineages. Of possible significance is that the single Ping locus Chr1:2640500–2640502 shared by AI23/ AI119 strains is also the only Ping locus shared with EG4/HEG4 (Fig. 2A). This suggests that possession of the shared Ping locus may be responsible for conferring the capacity to catalyze a massive increase in Ping and mPing elements in the descendants of AI23* (and in the progenitor of EG4). Experiments designed to test this scenario for activation are currently underway.

Epigenetic Regulation Has Been Maintained Since Divergence. Only one new heritable insertion (Table S5 and Dataset S2) was detected from all other potentially active rice TEs, indicating the maintenance of normal genome surveillance during decades of the Ping/mPing bursts. This is best illustrated by our inability to detect movement of the class 2 Pong element and the class 1 Tos17 element, both shown previously to be simultaneously activated in rice cell culture (along with mPing) (22, 49), where epigenetic regulation is known to be relaxed (22, 49). Activation of multiple rice TEs was reported in rice DNA methyltransferase and chromomethylase mutants, providing an explanation for the severity of the observed mutant phenotypes (20, 56, 57). In contrast, our data suggest that natural bursts, like those characterized in this study, may be sustained for decades because they are less harmful; only a single TE family is transposing, and its members avoid inserting into exons (27, 28).

Methylation of mPing Sequences Does Not Prevent mPing Transposition or Ping Activity. After establishing that the strain pairs were propagated as inbreds and that epigenetic regulation was maintained during each burst, we chose to investigate how the bursts were sustained for decades. Our initial hypothesis that both mPing and Ping elements avoided silencing led to analyses of their epigenetic marks in one member of each strain pair and in NB. Surprisingly, the majority of mPing elements were highly methylated in all strains examined, indicating host recognition of mPing before and during the bursts (Fig. 4A and B and Fig. S2), likely by RdDM (44) guided by abundant 24-nt siRNAs that target mPing sequences for methylation (Fig. 4C). These siRNAs probably derive from transcripts of host genes with mPing insertions in introns or 3′ flanking sequences. Of the 51 mPing insertions in NB, eight are located in introns of rice genes, while 40 EG4 mPings are in introns.

High Ping copy number appears to be necessary for maintaining the mPing burst. A previous study reported that strains AI23, AI19, EG4, and HEG4 had multiple copies of Ping (30), but their genomic locations remained unknown. Here we report that EG4/HEG4 share all seven Ping loci, which is consistent with the timing of the strains’ origin having occurred in the midst of the mPing burst. The strain pair AI23/AI19 shares only a single Ping locus, indicating that increase in Ping copy number occurred independently in each lineage after the strains diverged from their LCA. The same prior study (30) also reported expression of transcripts from both Ping genes in EG4 and AI23. We detected Ping transcripts in all strains and in NB at levels consistent with lower copy number. Furthermore, we found that Ping transcription reflects its epigenetic marks: Both genes, ORF1 and TPASE, have active gene body histone modifications in NB and EG4 (Fig. 3). These data support the hypothesis that Ping is expressed in NB and likely in all other strains that contain a single copy, but Ping activity is low and rarely promotes transposition of itself or of mPing. Thus, a likely scenario is that an increase in Ping copy number preceded and continues to drive increases in mPing copy numbers in both strain pairs. In this regard, Ping differs from other TEs such as the maize Ac element, which displays a negative dosage effect (58): One Ac copy catalyzes significantly more transposition than two copies, which catalyze more transposition than three copies (59).

Using oligo primers designed against the unique flanking sequences of each Ping locus, the methylation status of 295 bp from

![Image](image-url)
the 5’ ends of the seven mPing elements (253 bp shared, 42 bp unshared with mPing) in EG4 and the one mPing in NB was resolved. For all eight mPing, most of the DNA sequences shared with mPing are highly methylated (Fig. 5 and Fig. S3). This is likely caused by trans-acting siRNAs targeting mPing for methylation and also recognizing the identical Ping sequences. However, there is no apparent spreading of methylation into Ping ORFs. In fact, the opposite appears to be the case: Shared sequences closer to the Ping ORF1 promoter (region 218–295 bp) are less methylated than regions adjacent to the terminal inverted repeats (TIRs) (Fig. 5). Although the extent of reduced methylation is not uniform, all Pings show this reduction, even the single Ping in NB. The mechanism underlying reduced methylation of shared sequences is under investigation.

In summary, host genome defense recognition of mPing appears to have no impact on initiation or maintenance of the burst. Ping is actively transcribed, and both Ping and mPing can transpose despite methylation of terminal sequences. Furthermore, host recognition of mPing in low-copy strains like NB suggests that mPing was also recognized in the burst progenitor strains (EG4* and A123*), but recognition did not prevent activation of this TE family. This finding suggests that another feature of the Ping/mPing family’s success (in addition to avoiding insertion into exons) is that the mPing MITE does not share any sequences with its autonomous partner that would repress its activity.

In addition to instability caused by the massive TE amplification, there have been documented to underlie other chromosomal changes such as deletions, inversion, and duplication of chromosomal segments. The availability of the two sets of strain pairs and the overall stability of the genomes of an inbred species such as *O. sativa* permitted detection of these large-scale rearrangements and changes through comparison with NB as an outgroup to polarize changes and determine which structural arrangement is ancestral. An extensive comparative analysis identified a single large inversion flanked by mPing elements, suggesting that even over a fairly short period of time this active MITE may be responsible for a major structural rearrangement. As the burst continues and mPing copy number increases, additional sites are created for possible recombination events, which can lead to greater genome instability or increase the frequency of mutations due to TE insertions.

Alternatively, in some individuals, Ping may transpose into a region of heterochromatin or a region where antisense is generated, leading to its silencing and the end of the burst in all inbred and half of any outcrossed progeny.

In conclusion, the remarkable observation about the mPing bursts ongoing in these inbred strains is that they are occurring despite an intact epigenetic regulation system that likely evolved to suppress transposon accumulation. The MITE and autonomous elements escape silencing because the trans-acting siRNAs targeted to the monocopy MITE sequences do not share sequence similarity with the two coding regions of Ping. As shown previously, the mPing MITEs also preferentially insert into regions of the genome that do not interrupt exons (28, 35), thus reducing the chances of generating a lesion that is lethal or reduces strain fitness. The “natural” MITE burst observed in the two strain pairs is fundamentally different from the movement of TEs reported in rice cell culture systems or in rice mutants in which DNA methylation is reduced. In these situations, perturbation of DNA methylation results in the simultaneous activation of several distinct TE families, resulting in severe phenotypes including embryo inviability (17, 20, 56).

**Materials and Methods**

**Rice Strains.** Seeds of Gimbozu HEG4, Gimbozu EG4 (called “HEG4” and “EG4,” respectively in this study), Aikoku A123 (GeneBank accession no. 6730, called “A123*” in this study), and Aikoku A119 (GeneBank accession no. 6158, called “A119*” in this study) were obtained from the GeneBank project of the National Institute of Agrobiological Science, Ibaraki, Japan (www.gene.bioaffr.go.jp/databases-plant_search_en.php). HEG4 originated—25 y ago from a single seed of a strain designated here as EG4*. EG4 and HEG4 were propagated by self and transplanted pollination for ~20 generations. A123* was a popular cultivar grown in northern Japan in the early 1990s. A119* arose as a pure line from A123*—~10 y ago (31).

The precise number of generations from A119* to A119 and from A123* to A123 is not known. Seeds were sterilized in 1% (vol/vol) sodium hypochlorite for 1 h and rinsed with water. Sterilized seeds were placed on wet filter paper for 4 d at 25 °C, and germinated seeds were transplanted into plastic pots and grown in a greenhouse (30 °C daytime, 20 °C night) for 3 wk under natural light.

**DNA Extraction, Illumina Library Preparation, Sequencing.** Genomic DNA was extracted from one plant of each strain using the cetyl trimethylammonium bromide (CTAB) extraction method (60). Libraries for paired-end sequencing were prepared using the Illumina TruSeq DNA Kit (Illumina Inc.) following the manufacturer’s instructions. Large insert libraries (3–10 kb) were prepared using the standard protocol from Illumina by the Arizona Genomics Institute. Each library (10 nm) was paired-end sequenced on the Illumina HiSeq 2000 platform generating 2 × 100-bp reads at the UC Riverside Institute for Integrative Genome Biology. In total, 72 Gb (193× coverage) were sequenced for HEG4, 25.4 Gb (68x) for EG4, 23.2 Gb (62x) for A119, and 73 Gb (197x) for A123. The mean insert size of each library is shown in Table S1.

**Transposon Display.** Transposon display was performed as described (61). The adapter primers were Msel + T for mPing, N Dart, Osr10, and Osr37; Msel + A for SPMILike, Tam1, and RIRE2; Msel + G for Bajie, Dasheng, and RIRE3; Bafal + G for Tourist6; Bafal + T for TYPYE; and Bafal + C for Copia2. Primer sequences for each TE are given in Table S6.

**DNA Blot Analysis.** Genomic DNAs (10 μg) were digested with EcoRI, resolved by gel electrophoresis, transferred to Hybond-N nylon membranes (GE Healthcare), and hybridized with [32P]-labeled probes as described (62). Primer sequences used to synthesize labeled probes are given in Table S6.

**RNA Extraction and qRT-PCR Analysis.** Total RNA was extracted from 3-wk-old EG4 and NB seedlings with the NReasy Plant Mini Kit (Qiagen). After removal of contaminating DNA by digestion with amplification-grade RNase-free DNase I (Qiagen), RNAs were reverse transcribed by SuperScript III (Invitrogen). Resultant cDNAs served as templates for qPCR using iQ SYBR Green Supermix (Bio-Rad) with the CFX96 system (Bio-Rad). Samples were normalized to the actin gene. Primers used for qRT-PCR are given in Table S6.

**Bisulfite PCR and Sequencing.** Genomic DNA was extracted using the DNeasy Plant Mini kit (Qiagen), and bisulfite conversion was performed using the EpiTect Bisulfite Kit (Qiagen). PCR primer sets were designed for mPing and for ORF1 and TPASE from both Ping and Pong (Table S6). Individual Ping loci were distinguished for BS-seq by using a forward primer in unique flanking DNA and a reverse primer inside Ping (Table S6). PCR products were purified by the QiAquick gel extraction kit (Qiagen) and were cloned with the TOP10 TA cloning kit (Invitrogen). For each sample, 10–20 independent colonies were selected and sequenced, and sequences were analyzed using Kismeth software (63).

**Whole-Genome BS-Seq and Data Analysis.** Genomic DNAs (1–3 μg) isolated from NB, EG4, and A119 seedlings using the DNeasy Plant Mini kit (Qiagen) were sheared with a Covaris instrument to mean size of 300 bp. Fragments were purified, and ends were repaired, A-tailed, and ligated with methylated adapters (Bioo Scientific) following the manufacturer’s instructions for the KAPA LTP library preparation kit (Kapa Biosystems). Bisulfite-treated DNAs (EpiTect Bisulfite Kit; Qiagen) were amplified for 12–16 cycles, and resultant DNAs were multiplexed and applied to paired-end sequencing with read lengths of 100 or 75 nt on the Illumina HiSeq 2500 or NextSeq 500 platform. Raw reads were quality trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and aligned to the NB reference (MSU7) (rice.pbiology.msu.edu) using Bismark (64) allowing two mismatches. EG4 and A119 reads were also mapped to the mPing pseudogenome (all mPing sites with 1-kb flanking sequences). Reads mapping to unique flanking sequences from a read-pair were used to distinguish individual mPing insertion sites. Bismark’s methylation extractor script was used to calculate the methylation level for each cytosine. The bisulfite conversion rate for each library was calculated based on the methylation level of cytosines from reads mapping to the unmethylated chloroplast genome. The mPing sequence contains 42 cytosines in the CG context, 32 in CHG, and 129 in CHH. For EG4, only mPing elements with at least 80% read coverage of cytosines in all three contexts were selected (at least 32 Cs in strand CHG, and 100 in CHH). Because NB was sequenced at a lower depth, mPing elements with 70% cytosine coverage were selected.

**ChIP-Seq and Data Analysis.** ChIP was performed following previously described methods (65). Leaf tissue (5 g) from 2-wk-old seedlings was fixed in

![Image](https://example.com/image.png)
1% formaldehyde and vacuum infiltrated for 25 min. Cross-linking was quenched with 0.125 M glycine, and the tissue was rinsed three times with water and frozen in liquid nitrogen. Chromatin was isolated using extraction reagents, resuspended in nuclei lysis buffer (7), and sheared using a Covaris instrument. Chromatin was incubated with anti-H3K4me3 (Millipore catalog number 07-473, anti-H3K9me2 (Cell Signaling catalog number 4658S)), and anti-IGG antibodies (Cell Signaling catalog number 7074S) as a control. After reverse cross-linking and proteinase K and RNase treatments, the immunoprecipitated DNA was purified by phenol/chloroform extraction. Eluted ChIPed DNA (100 ng) was used for library preparation using the NEXTflex Chip-Seq kit (Bio Scientific) per the manufacturer’s instructions, and libraries were sequenced on the Illumina HiSeq 2000 platform. Yields of single-end reads of 100 bp for H3K4me3, H3K9me2, and IGG libraries were 35.1 million, 44.8 million, and 2.3 million respectively. Quality filtered reads using Cutadapt (66) were aligned to MSU7 with Bowtie2 (67) using default parameters which process nonunique reads (e.g., to a TE) to be assigned to a region a randomly selected from multiple equally best hits. Sequence depth peaks of H3K4me3 reads were identified by MACS software with default parameters (68). The tool SICER (69) was used to detect broad H3K9me2 signals. To view the data, wig files were generated from BAM files using MACS or SICER programs and were visualized using the Integrated Genomics Viewer (IGV v 2.3.74) (70).

Small RNA Sequencing and Analysis. Total RNA (30 μg) was isolated from EG4 seedlings and resolved on 15% denaturing polyacrylamide gels. RNA molecules ranging from 18 to 30 nt were excised from the gel and were recovered by overnight soaking in 0.3 M NaCl followed by ethanol precipitation. Small RNA libraries were constructed using the TrueSeq Small RNA Sample Prep Kit (Illumina, Inc.) and were sequenced on the Illumina HiSeq 2500 platform with read lengths of 50 nt. Previously published small RNA sequences from NB [National Center for Biotechnology Information Sequence Read Archive (SRA) SRX1308895] were used. After the removal of adapter sequences using Cutadapt (66), sequence reads of 18–30 nt were mapped directly to mPing, Ping, or Pång with the flanking 100-bp sequences and to the whole genome of MSU7 using bowtie with perfect match (13). Small RNAs read counts for the TE loci were normalized to the mapped small RNA library size to calculate the number of reads per million.

Sequence Variation and Transposon Identification. Paired-end reads from A119, A123, HEG4, and EG4 were aligned to MSU7 using Burrows–Wheeler Aligner (BWA) v 0.5.9-r16 (71) and were processed with SAMtools v 0.1.16 (96:234) to produce sorted BAM files (72). SNP and indel identification were performed with GATK v1.2-64-gf2a6f02 (33) following recommended best practices from the GATK team (https://software.broadinstitute.org/gatk/gatk-best-practices; v3). PCR artifacts were removed to avoid overconfidence in SNP calls by processing BAM files to mark duplicate reads using Picard-tools MarkDuplicates (https://software.broadinstitute.org/gatk/picard/). To prevent false-positive variant calls due to alignment artifacts, sequence reads containing any indel were realigned using GATK RealignmentTargetCreator to create an updated BAM file. Final SNP identification was made using all these BAM files as input to the GATK Uni-itedGenotyper. The resulting Variant Call Format (VCF) file of variants was filtered to remove those failing in repeats of MSU7 generated by RepeatMasker v 3.3.0 (www.repeatmasker.org) and removed with subtractBed implemented in BEDTools v 2.15.0 (73). Only high-quality variants were retained using GATK VariantFiltration by applying stringent cutoffs for quality and coverage (QD < 5.0, Hrun < 4, QUAL < 60, MQ < 24 & (MQ0/(1.0* DP)) > 0.1), FS > 600, HaplotypeScore > 13.0, MQRankSum > -12.5, ReadPosRankSum < -8.0). Accuracy of called heterozygous sites was determined by Sanger sequencing of 20 loci with high coverage of two alleles; only one of the 20 was found to be an actual heterozygous polymorphism. To avoid overcalling these potentially low-confidence sites, all SNP loci found to be heterozygous in any of the four strains were filtered out (13,353 in HEG4, 13,419 in EG4, 12,149 in A119, and 12,027 in A123). SNP positions and genotypes of resulting high-quality variants for each strain reflect differences between the strain and NB. To calculate whether SNPs represented private alleles in each strain, the VCF was processed to compute the private alleles with a custom Perl script, to reduce false positive and private insertions, the private SNPs were identified in Fig. 1 (compare_strains_in_vcf.pl, available at https://github.com/stajichlab/general-genomics-tools). Sanger sequencing validated 25 of 26 randomly selected sites from the private SNP list.

TE insertion sites were identified using paired-end reads and SologenaTEv1-0-2 (35) to classify sites as shared between the HEG4/HEG4 pair or the A119/A123 pair or as private. TE positions in MSU7 were also computed using RelocaTE (option -r all) (77). Because we were only interested in TE positions with the requirement of a minimum of 10-bp perfect alignment of the read to the TE. Due to high sequence identity of the TIRs, RelocaTE was first used to identify all potential Ping and Pång insertions followed by processing with ConstrucTEr, a RelocaTE companion tool, to differentiate between similar elements. Detection of Ping loci in other public rice strains was performed using reads of Oamochi, Nokken 58, and Kitaake [DNA Data Bank of Japan (DDBJ) accession: DRA000017, ERAD_123/1, and 3.5 S-Retroviral-aided-Com-}

ACKNOWLEDGMENTS. We thank Drs. Matthew Collin and Jinghua Shi for the preparation of some libraries for sequencing; Patrick Schreiner for software development; Drs. James Burnette and Dawn Nagel for critical reading of the manuscript; and Genebank Project of the National Institute of Agrobiological Science in Japan for providing seeds of Alkoku landraces (A119 and A123). This work was supported by National Science Foundation Grant IOS-1027542 (to S.R.W. and J.E.S.).


