UC Riverside

UC Riverside Previously Published Works

Title

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

Permalink

https://escholarship.org/uc/item/41q71897

Journal

Proceedings of the National Academy of Sciences of the United States of America, 114(49)

ISSN

0027-8424

Authors

Lu, Lu
Chen, Jinfeng
Robb, Sofia MC
et al.

Publication Date

2017-12-05

DOI

10.1073/pnas.1716459114

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-ShareAlike License, availalbe at https://creativecommons.org/licenses/by-nc-sa/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

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

Lu Lu[a,b,1], Jinfeng Chen[a,b,c,1], Sofia M. C. Robb[a,b,c], Yutaka Okumoto[d], Jason E. Stajich[b,c], and Susan R. Wessler[a,b,2]

[a]Department of Botany and Plant Sciences, University of California, Riverside CA 92521; [b]Institute for Integrative Genome Biology, University of California, Riverside, CA 92521; [c]Department of Microbiology and Plant Pathology, University of California, Riverside, CA 92521; and [d]Graduate School of Agriculture, Kyoto University, Kitashirakawa-oiwake Sakyo, Kyoto 606-8502, Japan

Contributed by Susan R. Wessler, October 27, 2017 (sent for review September 19, 2017; reviewed by Damon Lisch and Nathan M. Springer)

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 *Ping* 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. Evasion is possible because host recognition of *mPing* sequences 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. This finding suggests that an important feature of MITE success is that host recognition does not lead to the silencing of the source of transposase.

MITE | genome evolution | transposon silencing | rice | *mPing*

Transposable elements (TEs) comprise the largest proportion of all characterized plant and animal genomes (1). They make up at least half of the human genome (2) and ∼60–85% of some grass genomes (3, 4). Virtually all characterized genomes contain TEs that have attained very high copy numbers. The phenomenon often reflects the ability of a subset of TEs to undergo a "burst," a term that describes a rapid increase in TE copy number to thousands, even tens of thousands. It has been suggested that TE bursts have generated new gene-expression networks through the rapid dispersal of potential TE-encoded regulatory elements into genes throughout the genome (5).

Despite the prevalence of high-copy-number TEs, the strategies that enable TEs to attain high copy numbers without killing their host or triggering their inactivation through epigenetic silencing is not readily apparent from analysis of extant genomes (6). These questions need to be addressed by identifying active TEs in the midst of a burst and characterizing their impact in real time on the host. The first identified active TEs, now called "class 2 elements," were discovered through genetic analysis of mutant alleles and include the *Ac/Ds* and *Spm/dSpm* elements of maize (7), the *Tc1/mariner* elements of *Caenorhabditis elegans* (8), and the *P* element of *Drosophila* (9). Class 2 (DNA) elements transpose through a DNA intermediate and are organized into families containing autonomous and nonautonomous elements (10). Autonomous elements encode the protein(s) necessary for their own transposition and for transposition of nonautonomous family 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 intronic 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

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.

Author contributions: L.L., J.C., S.M.C.R., J.E.S., and S.R.W. designed research; L.L., J.C., and S.M.C.R. performed research; L.L., J.C., S.M.C.R., and Y.O. contributed new reagents/analytic tools; L.L., J.C., S.M.C.R., J.E.S., and S.R.W. analyzed data; and L.L., J.C., J.E.S., and S.R.W. wrote the paper.

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. JSUF00000000 and JSUG00000000) and the BioProject database, https://www.ncbi.nlm.nih.gov/bioproject/ (accession nos. PRJNA198499 and PRJNA264731).

[1]L.L. and J.C. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: susan.wessler@ucr.edu.

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

genes (16). The proximity of MITEs to genes may explain why rice mutants in the RdDM pathway such as *OsMet1* (18), *OsDCL3a* (15), *OsDRM2* (19), and *OsCMT3a* (20) exhibit a more severe spectrum of phenotypes than mutants of orthologous genes in *Arabidopsis*, where TEs and genome methylation are largely restricted to centromeric regions (21).

Study of the dynamics and impact of MITE amplification was enabled by the discovery of an actively transposing rice TE family composed of the autonomous *Ping* element and nonautonomous *mPing*, the first active MITE isolated from any organism (22–24). *Ping* is a member of the *PIF*/*Harbinger* class 2 superfamily that encodes two proteins required for the movement of all family members: a transposase and a second ORF of unknown function (25, 26) (Fig. 1*A*). While transpositions of *Ping* and *mPing* have rarely been detected in the great majority of rice strains analyzed (22, 27), *mPing* movement was initially demonstrated in rice cell and anther cultures (22, 24) and later in mutant backgrounds with reduced DNA methylation (20). Bursts of *mPing* were subsequently detected in five *japonica* rice strains, Gimbozu EG4 (hereafter "EG4"), Gimbozu HEG4 (hereafter "HEG4"), Aikoku A123 (hereafter "A123"), Aikoku A119 (hereafter "A119"), and A157, with up to 40 new insertions per plant per generation (27, 28). Characterization of the thousands of new *mPing* insertion sites in these strains revealed a preference for insertion upstream of the transcription start site and, surprisingly, a significant deficit of exon insertions (27, 28). This latter feature is a powerful success strategy for a TE bursting in copy number, and the avoidance of exons is likely due to the unusually high GC content of rice exons and *mPing*'s AT-rich 9-bp insertion sequence preference (26, 27). This hypothesis is supported by experiments demonstrating that *mPing* is an effective mutagen in transgenic soybean and readily inserts into exons, which typically have a lower GC content than rice (29).

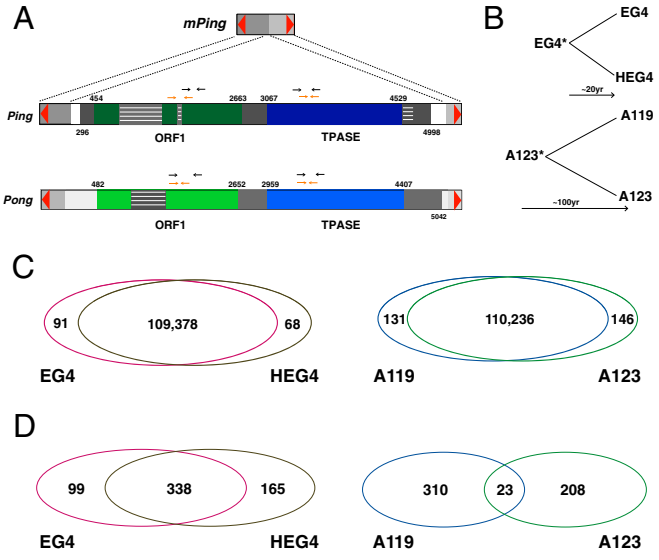

Fig. 1. Proportions of shared *Ping*/*mPing* insertions and SNPs indicate strain pairs have been maintained by self or sibling pollination. (*A*) Schematic structures of *mPing*, *Ping*, and *Pong*. Both autonomous elements encode ORF1 and TPASE. Numbers above the figures show the positions of the coding 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 *Ping* 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 homozygous 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 ancestor and maintained by self or sibling pollination for decades. Since divergence, *Ping* has continued to produce transposase that has catalyzed massive *mPing* transposition while all other potentially active TEs have remained inactive. Analyses of DNA methylation and other epigenetic marks and *Ping* 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 *Ping* 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 ∼20 y (Fig. 1*B*). 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. 1*B*). Because even closely related strains of *O. sativa* var. *japonica* differ by tens to hundreds of thousands 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 sequenced to produce 68× genome coverage for EG4 and 193× coverage for HEG4 (Table S1), which was suitable for identification 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). Identification and analysis of sequence variants was performed using the Nipponbare (NB) genome as a reference (12). SNPs were identified 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. 1*C*). 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. 1*B*).

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× and 197× coverage, respectively. Multiple insert-size libraries were constructed and sequenced for A123 to support the assembly of a draft genome sequence (Tables S1 and S2). A119 and A123 share 110,236 SNP positions that differ from NB and differ from each other by ∼277 SNPs (Fig. 1*C*). These patterns of variation are also consistent with a demographic history in which A119 and A123 share a recent common ancestor (A123*) and have been propagated without outcrossing (Fig. 1*B*).

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 RelocaTE and classified as heterozygous or homozygous with CharacTErizer (Dataset S1) (35). Only homozygous insertions were further classified as shared or private alleles (Fig. 1*D*) 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. 1*D*) 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 burst of the *Ping*/*mPing* 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. 1*A*) (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*s in the four strains, with seven copies in EG4, HEG4, and A119 and 10 copies in A123 (Fig. 2 *A* 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. 2*A*). 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. 2*B*), 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 *Ping*s, 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. 1*A*) (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

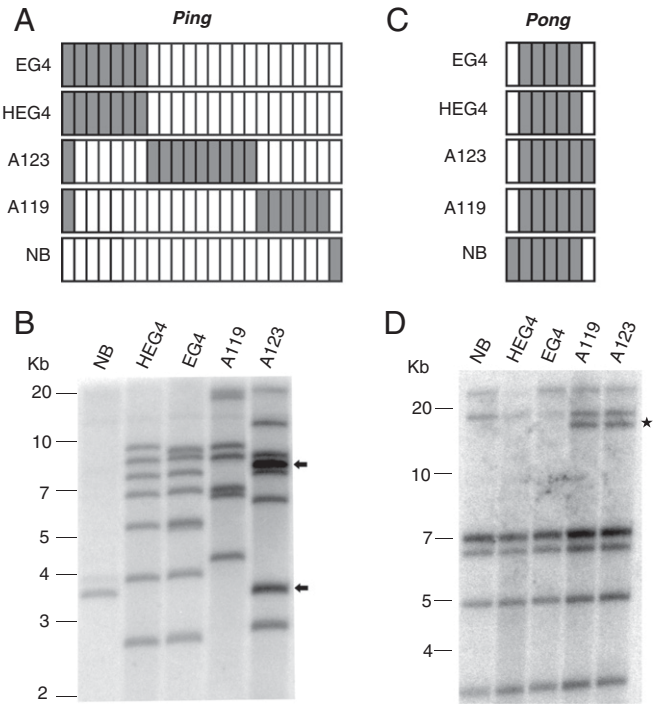

Fig. 2. *Ping* and *Pong* genomic loci in five strains. (*A* and *C*) Schematics of the loci containing full-length *Ping* (*A*) or *Pong* (*C*) elements in the genomes of the indicated strains. Each column is a unique genomic locus, and a filled box represents the presence of *Ping* or *Pong* in that strain at that locus. Precise genome positions are summarized in Tables S3 and S4. (*B* and *D*) Southern blot analysis of *Ping* (*B*) or *Pong* (*D*) is depicted for the five strains. Genomic DNA of each sample was digested with EcoRI and hybridized with *Ping*- or *Pong*-specific probes. Size markers inferred from the ladder are shown on the left. Arrows in *B* indicate doublet unresolved bands in A123, and the star in *D* indicates the private *Pong* locus shared by A119 and A123 (see text for details).

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. 2 *C* 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 Ty1-*copia* retrotransposon, is also found in several other *japonica* strains (Omachi, Nongken 58, and Kitaake) (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 at levels roughly proportional to their copy numbers in NB (one *Ping*) and EG4 (seven identical *Ping*s) (Fig. 3*A*, *Left*). The same samples yielded negligible transcript levels from the six *Pong* elements in NB and the five *Pong* elements in EG4 (Fig. 3*A*, *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. 3*B*, *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. 3*B*,

Right), reminiscent of silenced transposons with high levels of symmetric methylation (40).

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 K9 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. 3*C*, *Left*), similar to profiles observed in active rice genes (Fig. 3*D*). In contrast, the entire *Pong* element was modified by the repressive H3K9me2 mark (Fig. 3*C*, *Right*), similar to modification patterns of silenced retrotransposons (Fig. 3*E*). 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. 4*A*). 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. 4*B*), 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. 4*C*). 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. 1*A*). 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*

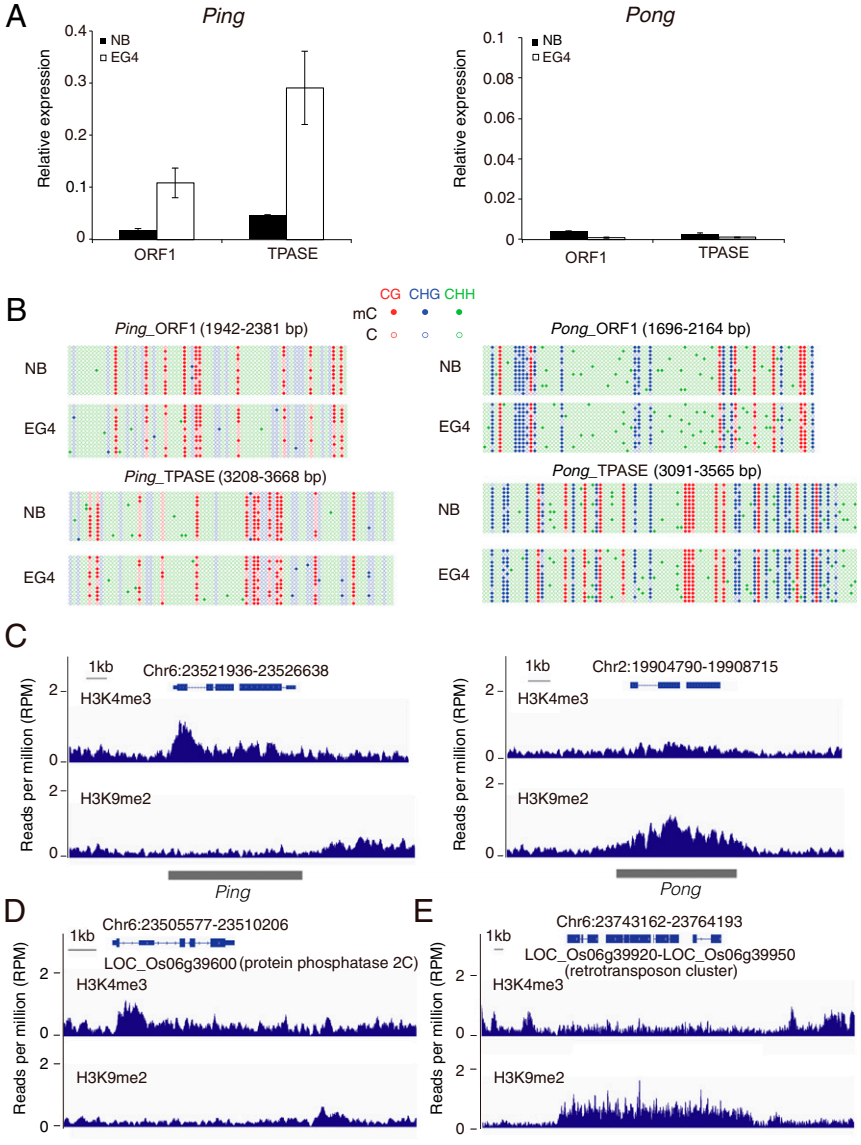

Fig. 3. *Ping* but not *Pong* has actively transcribed genes in both NB and EG4. (*A*) ORF1 and TPASE transcript expression of *Ping* and *Pong* quantified by qRT-PCR using RNA isolated from seedlings of NB and EG4. Transcript levels are shown normalized to actin, differ in magnitude for *Ping* and *Pong*, and depict mean ± SD of three independent biological replicates. (*B*) Dot plots showing DNA methylation of internal regions (start and end positions are shown) of ORF1 and TPASE from *Ping* (*Left*) and *Pong* (*Right*) in NB and EG4. Bisulfite-treated DNAs were amplified and sequenced and 13–20 bisulfite clones are shown. The methylation state of each cytosine is shown as an open circle (not methylated) or closed circle (methylated). Cytosines in CG, CHG, and CHH contexts are shown in red, blue, and green respectively. (*C*) IGV genome browser views of H3K4me3 and H3K9me2 modification patterns in *Ping* (*Left*) and one *Pong* locus (Chr2:19904309–19909474) (*Right*) in NB. Gray bars indicate the positions of full-length *Ping* or *Pong*. (*D* and *E*) Examples of H3K4me3 and H3K9me2 modifications in a typical expressed rice gene (*D*) and typical silenced rice retrotransposons (*E*).

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. 4*C* 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*A*). Similarly, few siRNAs mapped to *Pong* loci (Fig. S4*B*). 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 arrows). 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

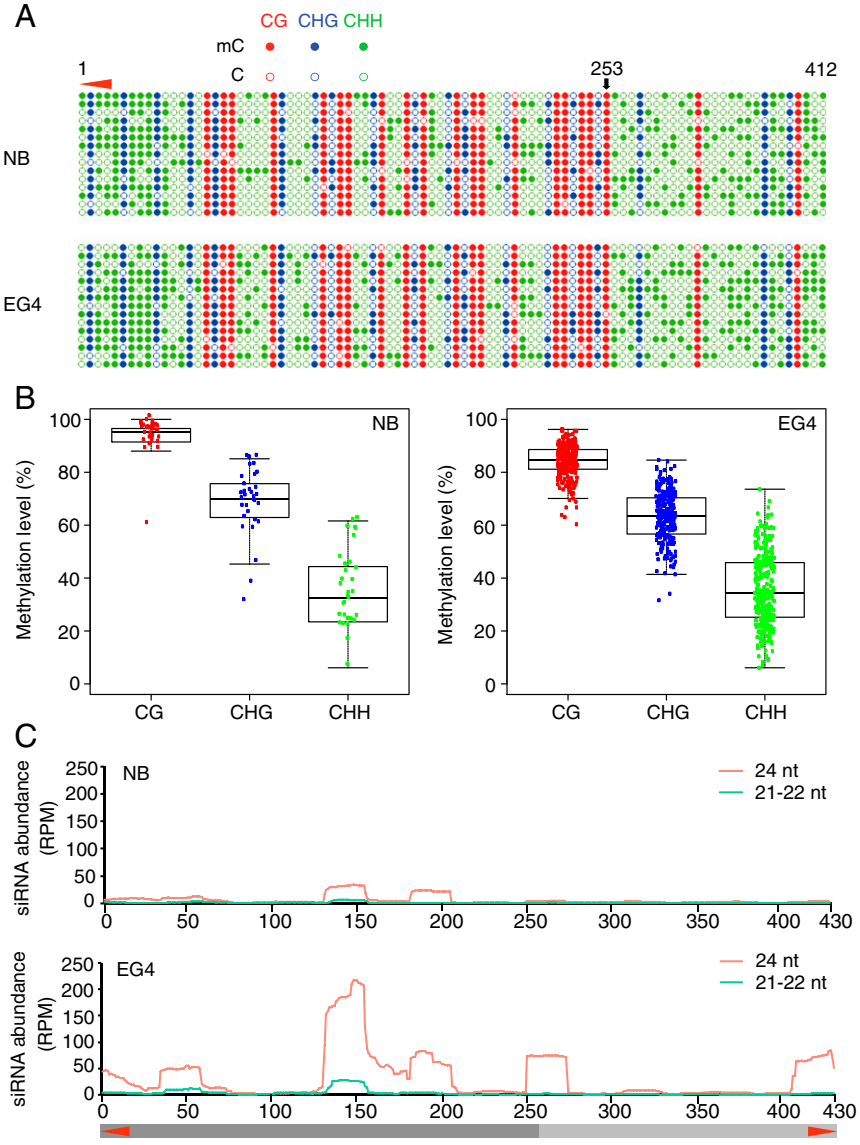

Fig. 4. *mPing* elements are methylated. (*A*) Dot plots showing DNA methylation of *mPing* in NB and EG4. Sequences 1–253 are virtually identical to the 5′ end of *Ping*; sequences 254–430 are identical to the 3′ end of *Ping* (Fig. 1*A*). The red triangle shows the location of the 5′ TIR. The last 18 bp of *mPing* (positions 413–430), including the 3′ TIR, were not calculated due to the limitation of the primers. (*B*) Boxplots and dots plotted for individual *mPing* elements showing percentage of target cytosine sequences methylated in individual *mPing* elements detected by whole-genome BS-seq data. Data shown for NB are from 32 of 51 *mPings* in which 70% of the cytosines in all three sequence contexts (CG, CHG, and CHH) were covered by BS-seq reads. Data for EG4 are from 271 of 437 *mPing* copies in which 80% of the cytosines were covered by sequencing reads. (*C*) Abundance in reads per million (RPM) of 24-nt (red) and 21- or 22-nt (green) siRNAs is plotted along the *mPing* locus (*x* axis) from small RNA sequencing libraries from NB and EG4. The regions shared by *mPing* and *Ping* (1–253 and 254–430 bp) are shown as dark gray and light gray bars, respectively. Reads were directly mapped to the *mPing* sequence, not to the whole-genome sequence. Red triangles at the bottom show the location of the 5′ and 3′ TIRs.

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. 6*A*). Validation of this inversion by PCR is shown in Fig. 6*B*. Comparison of this region in NB and EG4 reveals no inversion and near structural identity except for two *mPing* insertions in EG4 (Fig. 6*A*, 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. 6*A*, 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

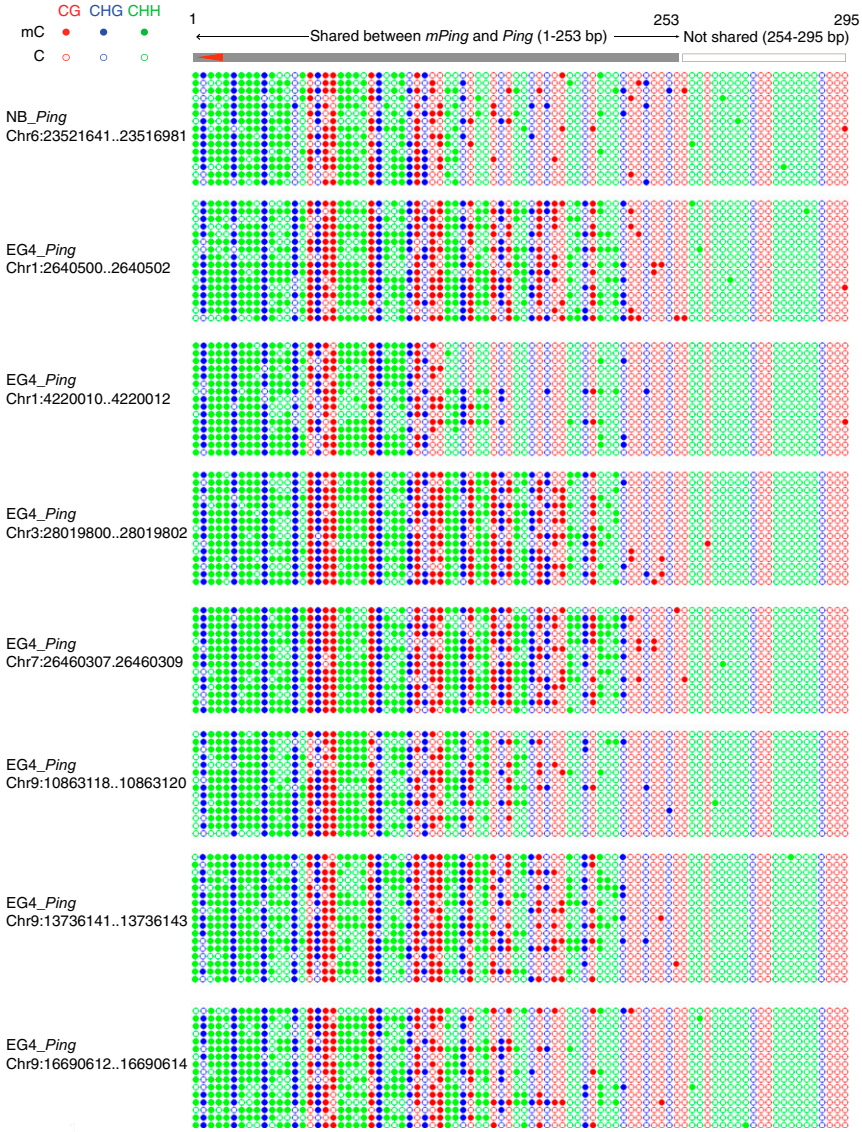

Fig. 5. DNA methylation of the 5′ shared sequences between *mPing* and *Ping* in NB and EG4. Dot plots display DNA methylation of the region from 1 to 295 bp of *Ping*. Bisulfite-treated DNAs from NB and EG4 were amplified using a forward primer in unique flanking sequences and a reverse primer inside *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 A119 from A123 (Fig. 1*C*). 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 A119/A123 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 A119 vs. A123 found that all TE insertions were shared except for a single *Dasheng* locus in A123 but not in A119. 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. 1*B*, EG4*, A123*).

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 A123/A119, A123* 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 A123 and A119 from A123*, and the potential to burst was inherited in both lineages. Of possible significance is that the single *Ping* locus Chr1:2640500–2640502 shared by A123/A119 strains is also the only *Ping* locus shared with EG4/HEG4 (Fig. 2*A*). 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 A123* (and in the progenitor of EG4). Experiments designed to test this scenario for activation are currently underway.

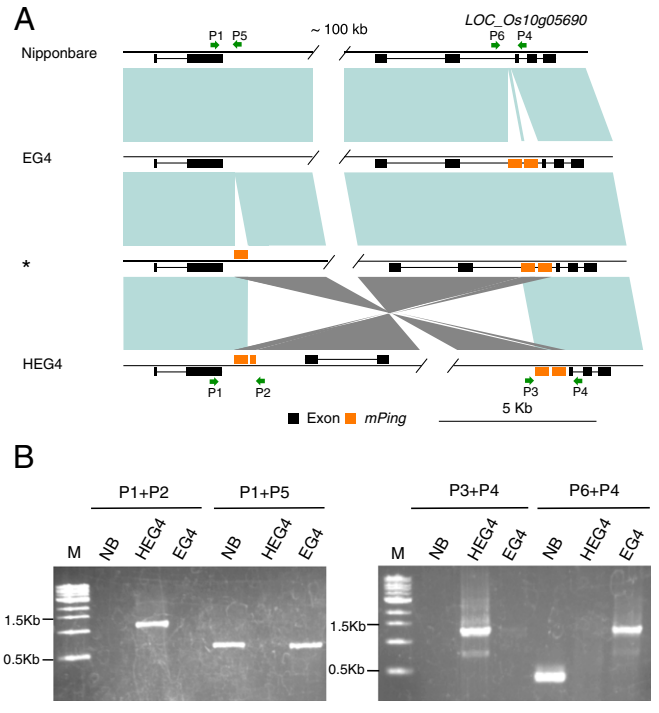

Fig. 6. A large inversion, unique to HEG4, is flanked by *mPing* elements. (*A*) Schematic representation of the inversion breakpoints. Genomic sequences are horizontal black lines with the positions of exons (black rectangles) and *mPing* elements (orange) shown. Homologous regions are denoted by light green shaded regions. Inverted fragments are denoted by gray shaded regions. The asterisk represents a hypothesized intermediate state in HEG4 before the inversion event. Positions of PCR primers P1–P6 are indicated by green arrows. (*B*) PCR validation of the structures of NB, EG4, and HEG4 predicted by genomic sequences.

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. 4 *A* and *B* and Fig. S2), likely by RdDM (44) guided by abundant 24-nt siRNAs that target *mPing* sequences for methylation (Fig. 4*C*). 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 A123, A119, 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 A123/A119 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 A123. 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

the 5′ ends of the seven *Ping* elements (253 bp shared, 42 bp unshared with *mPing*) in EG4 and the one *Ping* in NB was resolved. For all eight *Pings*, 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, TEs 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 multicopy 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.affrc.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 or sibling pollination for ~20 generations. A123* was a popular cultivar grown in northern Japan in the early 1900s. A119* arose as a pure line from A123* ~100 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 (68×) for EG4, 23.2 Gb (62×) for A119, and 73 Gb (197×) 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 MseI + T for *mPing*, *nDart*, *Osr10*, and *Osr37*; MseI + A for *SPMLIKE*, *Tami2*, and *RIRE2*; MseI + *G* for *Bajie*, *Dasheng*, and *RIRE3*; BfaI + G for *Tourist6*; BfaI +T for *TYPEU*; and BfaI + 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 ^{32}P-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 RNeasy 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 first-strand synthesis supermix (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 rice 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 TOPO 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.plantbiology.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. Bismarck'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 CG, 24 in 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

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 buffers, resuspended in nuclei lysis buffer (7), and sheared using a Covaris instrument. Chromatin was incubated with anti-H3K4me3 (Millipore catalog no. 07-473), anti-H3K9me2 (Cell Signaling catalog no. 4658S), and anti-IgG antibodies (Cell Signaling catalog no. 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 (BioScientific) 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 randomly selected from multiple equally best hits. Sequence depth peaks of ChIP 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 TruSeq 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) SRX1396995] were used. After the removal of adapter sequences using Cutadapt (66), sequence reads of 18–30 nt were mapped directly to *mPing*, *Ping*, or *Pong* 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(r963:234) to produce sorted BAM files (72). SNP and indel identification were performed with GATK v1.2-64-gf62af02 (33) following recommended best practices from the GATK team (https://software.broadinstitute.org/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://broadinstitute.github.io/picard/). To prevent false-positive variant calls due to alignment artifacts, sequence reads containing any indel were realigned using GATK RealignerTargetCreator to create an updated BAM file. Final SNP identification was made using all these BAM files as input to the GATK UnifiedGenotyper. The resulting Variant Call Format (VCF) file of variants was filtered to remove those falling 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, MQ0 ≥4 && [(MQ0/(1.0 * DP)] >0.1), FS >60.0, HaplotypeScore >13.0, MQRankSum less than −12.5, ReadPosRankSum less than −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 compute a table used to construct the Venn diagrams 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 RelocaTEv1-0-2 (35) to classify sites as shared between the HEG4/EG4 pair or the A119/A123 pair or as private. TE positions in MSU7 were also computed using RelocaTE (option–reference_ins 1). Because all *mPing*s are virtually identical, RelocaTE was run 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 *Pong* insertions followed by processing with ConstrucTEr, a RelocaTE companion tool, to differentiate between similar elements. Detection of *Pong* loci in other public rice strains was performed using reads of Omachi, Nongken 58, and Kitaake [DNA Data Bank of Japan (DDBJ) accession nos. DRA000307, ERA009071, and SRA054074, respectively].

To assess whether other TEs were polymorphic in each of the strain pairs, 812 TE families (Dataset S2) were analyzed with RelocaTE allowing up to 10% base mismatches in the alignment of the reads to account for sequence divergence of older TEs. Private and shared insertions were classified using a two-step process. First, TE insertions with reads that align to both sides of the insertion site when mapped to MSU7 were used to establish an initial set of locations. A site found in only one individual was classified as private; TE sites found in multiple individuals were classified as shared. Second, to reduce falsely classified private insertions, private sites were reclassified as shared when at least one read from the other strain pair aligned to the site. All private TE sites identified were confirmed by manual inspection of the read mapping in IGV v 2.3.74 (70) and were validated by PCR.

Discovery of SVs in Strain Pairs Using Paired-End Reads. Raw reads were processed by Trimmomatic v0.3 (74) for adapter trimming (LEADING:0 TRAILING:0 ILLUMINACLIP:adaptor.fa:2:40:15 MINLEN:50). A de novo genome was constructed for HEG4 and A123 by assembling adapter-trimmed reads with ALLPATHS-LG v 41554 (75) with default parameters, except that MIN_CONTIG was set to 300. Sequence gaps in the assembly of ALLPATHS-LG were filled by GapCloser using one cycle of closure (76). Paired-end reads and synteny in *Oryza* genomes, including MSU7, *Oryza glaberrima* (77), and *Oryza punctata* (GenBank accession no. GCA_000573905.1) were used to construct pseudochromosome sequences of HEG4 and A123 using Reference-Assisted Chromosome Assembly (RACA v 0.9.1) (78). SVs in strain pairs (EG4 vs. HEG4 and A119 vs. A123) were detected using read-mapping approaches in which the genome of HEG4 was used as a reference to compare EG4 and likewise A123 assembly was used to discover variants in A119. Read-mapping approaches involved the use of BreakDancer v 1.1.2_2013_03_08 (79), Pindel v 0.2.4o (80), and Meerkat v 0.175 (81), which utilize paired-read information, detection of splits within individual read alignments, and a combination of read pair and split read methods, respectively. RelocaTE was also used to assist in the identification of TE-specific events (35). The variants identified by three tools were merged using BEDtools (73). The merged variations were verified by local sequence assembly of all available short reads from the 1-kb flanking region using Velvet v 1.2.09 (82). All variations except for those smaller than 100 bp or known *Ping*/*mPing* insertions were further confirmed by manual inspection of the read mapping in the IGV v 2.3.74 (70). Seven confirmed SVs as shown in Table S5 were verified by PCR (Fig. S6).

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 Aikoku landraces (A119 and A123). This work was supported by National Science Foundation Grant IOS-1027542 (to S.R.W. and J.E.S.).

1. Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: Where genetics meets genomics. *Nat Rev Genet* 3:329–341.
2. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 7:e1002384.
3. Paterson AH, et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457:551–556.
4. Schnable PS, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326:1112–1115.
5. Sundaram V, et al. (2014) Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* 24:1963–1976.
6. Lisch D, Slotkin RK (2011) Strategies for silencing and escape: The ancient struggle between transposable elements and their hosts. *Int Rev Cell Mol Biol* 292: 119–152.
7. McClintock B (1954) Mutations in maize and chromosomal aberrations in *Neurospora*. *Carnegie Institution of Washington Year Book* (Carnegie Institution of Washington, Washington, DC), pp 254–260.
8. Emmons SW, Yesner L, Ruan KS, Katzenberg D (1983) Evidence for a transposon in *Caenorhabditis elegans*. *Cell* 32:55–65.
9. Bingham PM, Kidwell MG, Rubin GM (1982) The molecular basis of *P-M* hybrid dysgenesis: The role of the *P* element, a *P*-strain-specific transposon family. *Cell* 29:995–1004.
10. Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. *Proc Natl Acad Sci USA* 103:17600–17601.
11. SanMiguel P, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274:765–768.
12. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800.

13. Han Y, Qin S, Wessler SR (2013) Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC Genomics* 14:71.
14. Lu C, et al. (2012) 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*. *Mol Biol Evol* 29:1005–1017.
15. Wei L, et al. (2014) Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. *Proc Natl Acad Sci USA* 111:3877–3882.
16. Zemach A, et al. (2010) Local DNA hypomethylation activates genes in rice endosperm. *Proc Natl Acad Sci USA* 107:18729–18734.
17. Song X, Cao X (2017) Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. *Curr Opin Plant Biol* 36: 111–118.
18. Hu L, et al. (2014) Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. *Proc Natl Acad Sci USA* 111:10642–10647.
19. Moritoh S, et al. (2012) Targeted disruption of an orthologue of *DOMAINS REARRANGED METHYLASE 2, OsDRM2*, impairs the growth of rice plants by abnormal DNA methylation. *Plant J* 71:85–98.
20. Cheng C, et al. (2015) Loss of function mutations in the rice chromomethylase *OsCMT3a* cause a burst of transposition. *Plant J* 83:1069–1081.
21. Li X, et al. (2012) Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* 13: 300.
22. Jiang N, et al. (2003) An active DNA transposon family in rice. *Nature* 421:163–167.
23. Nakazaki T, et al. (2003) Mobilization of a transposon in the rice genome. *Nature* 421: 170–172.
24. Kikuchi K, Terauchi K, Wada M, Hirano HY (2003) The plant MITE *mPing* is mobilized in anther culture. *Nature* 421:167–170.
25. Hancock CN, Zhang F, Wessler SR (2010) Transposition of the *Tourist-MITE mPing* in yeast: An assay that retains key features of catalysis by the class 2 *PIF/Harbinger* superfamily. *Mob DNA* 1:5.
26. Yang G, Zhang F, Hancock CN, Wessler SR (2007) Transposition of the rice miniature inverted repeat transposable element *mPing* in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 104:10962–10967.
27. Naito K, et al. (2006) Dramatic amplification of a rice transposable element during recent domestication. *Proc Natl Acad Sci USA* 103:17620–17625.
28. Naito K, et al. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461:1130–1134.
29. Hancock CN, et al. (2011) The rice miniature inverted repeat transposable element mPing is an effective insertional mutagen in soybean. *Plant Physiol* 157:552–562.
30. Teramoto S, Tsukiyama T, Okumoto Y, Tanisaka T (2014) Early embryogenesis-specific expression of the rice transposon *Ping* enhances amplification of the MITE *mPing*. *PLoS Genet* 10:e1004396.
31. Morinaga T (1957) *Rice of Japan* (Yokendo, Tokyo), pp 324. Japanese.
32. Arai-Kichise Y, et al. (2014) Genome-wide DNA polymorphisms in seven rice cultivars of *temperate* and *tropical japonica* groups. *PLoS One* 9:e86312.
33. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.
34. Baruch O, Kashkush K (2012) Analysis of copy-number variation, insertional polymorphism, and methylation status of the tiniest class I (TRIM) and class II (MITE) transposable element families in various rice strains. *Plant Cell Rep* 31:885–893.
35. Robb SM, et al. (2013) The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. *G3 (Bethesda)* 3:949–957.
36. Zhang X, Jiang N, Feschotte C, Wessler SR (2004) *PIF*- and *Pong*-like transposable elements: Distribution, evolution and relationship with *Tourist*-like miniature inverted-repeat transposable elements. *Genetics* 166:971–986.
37. Good AG, Meister GA, Brock HW, Grigliatti TA, Hickey DA (1989) Rapid spread of transposable *P* elements in experimental populations of *Drosophila melanogaster*. *Genetics* 122:387–396.
38. Catania F, et al. (2004) World-wide survey of an *Accord* insertion and its association with DDT resistance in *Drosophila melanogaster*. *Mol Ecol* 13:2491–2504.
39. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328:916–919.
40. Feng S, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci USA* 107:8689–8694.
41. Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biol* 10:R62.
42. Zhang X (2008) The epigenetic landscape of plants. *Science* 320:489–492.
43. Roudier F, et al. (2011) Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. *EMBO J* 30:1928–1938.
44. Matzke MA, Mosher RA (2014) RNA-directed DNA methylation: An epigenetic pathway of increasing complexity. *Nat Rev Genet* 15:394–408.
45. Moon S, et al. (2006) Identification of active transposon *dTok*, a member of the *hAT* family, in rice. *Plant Cell Physiol* 47:1473–1483.
46. Tsugane K, et al. (2006) An active DNA transposon *nDart* causing leaf variegation and mutable dwarfism and its related elements in rice. *Plant J* 45:46–57.
47. Huang J, et al. (2009) Identification of a high frequency transposon induced by tissue culture, *nDaiZ*, a member of the *hAT* family in rice. *Genomics* 93:274–281.
48. Dong HT, et al. (2012) A *Gaijin*-like miniature inverted repeat transposable element is mobilized in rice during cell differentiation. *BMC Genomics* 13:135.
49. Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Natl Acad Sci USA* 93: 7783–7788.
50. Picault N, et al. (2009) Identification of an active LTR retrotransposon in rice. *Plant J* 58:754–765.
51. Wang H, et al. (2009) Molecular characterization of a rice mutator-phenotype derived from an incompatible cross-pollination reveals transgenerational mobilization of multiple transposable elements and extensive epigenetic instability. *BMC Plant Biol* 9:63.
52. Cui X, et al. (2013) Control of transposon activity by a histone H3K4 demethylase in rice. *Proc Natl Acad Sci USA* 110:1953–1958.
53. Gray YH (2000) It takes two transposons to tango: Transposable-element-mediated chromosomal rearrangements. *Trends Genet* 16:461–468.
54. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41:331–368.
55. McClintock B (1946) Maize genetics. *Carnegie Institution of Washington Year Book* (Carnegie Institution of Washington, Washington, DC), pp 176–186.
56. Tsukahara S, et al. (2009) Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* 461:423–426.
57. Kato M, Miura A, Bender J, Jacobsen SE, Kakutani T (2003) Role of CG and non-CG methylation in immobilization of transposons in *Arabidopsis*. *Curr Biol* 13:421–426.
58. McCLINTOCK B (1951) Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol* 16:13–47.
59. Heinlein M (1996) Excision patterns of *Activator* (*Ac*) and *Dissociation* (*Ds*) elements in *Zea mays* L.: Implications for the regulation of transposition. *Genetics* 144:1851–1869.
60. Doyle JJ (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* 19:11–15.
61. Casa AM, et al. (2000) The MITE family *heartbreaker* (*Hbr*): Molecular markers in maize. *Proc Natl Acad Sci USA* 97:10083–10089.
62. Yang G, Weil CF, Wessler SR (2006) A rice *Tc1/mariner*-like element transposes in yeast. *Plant Cell* 18:2469–2478.
63. Gruntman E, et al. (2008) Kismeth: Analyzer of plant methylation states through bisulfite sequencing. *BMC Bioinformatics* 9:371.
64. Krueger F, Andrews SR (2011) Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* 27:1571–1572.
65. Zhu JY, Sun Y, Wang ZY (2012) Genome-wide identification of transcription factor-binding sites in plants using chromatin immunoprecipitation followed by microarray (ChIP-chip) or sequencing (ChIP-seq). *Methods Mol Biol* 876:173–188.
66. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17:10–12.
67. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359.
68. Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9:R137.
69. Xu S, Grullon S, Ge K, Peng W (2014) Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol Biol* 1150:97–111.
70. Robinson JT, et al. (2011) Integrative genomics viewer. *Nat Biotechnol* 29:24–26.
71. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
72. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.
73. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
74. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.
75. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108:1513–1518.
76. Luo R, et al. (2012) SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1:18.
77. Wang M, et al. (2014) The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat Genet* 46:982–988.
78. Kim J, et al. (2013) Reference-assisted chromosome assembly. *Proc Natl Acad Sci USA* 110:1785–1790.
79. Chen K, et al. (2009) BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 6:677–681.
80. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25:2865–2871.
81. Yang L, et al. (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* 153:919–929.
82. Zerbino DR, Birney E (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 18:821–829.