Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

A role for palindromic structures in the *cis*-region of maize Sirevirus LTRs in transposable element evolution and host epigenetic response

Alexandros Bousios, Concepcion M. Diez, Shohei Takuno, et al.

Genome Res. published online December 2, 2015
Access the most recent version at doi:10.1101/gr.193763.115

P<P Published online December 2, 2015 in advance of the print journal.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

© 2016 Bousios et al.; Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Research

A role for palindromic structures in the *cis*-region of maize Sirevirus LTRs in transposable element evolution and host epigenetic response

Alexandros Bousios,[1,2] Concepcion M. Diez,[3,4] Shohei Takuno,[5] Vojtech Bystry,[6] Nikos Darzentas,[6] and Brandon S. Gaut[4]

[1]*School of Life Sciences, University of Sussex, Brighton BN1 9RH, United Kingdom;* [2]*Institute of Applied Biosciences, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece;* [3]*Department of Agronomy, University of Cordoba, 14014 Cordoba, Spain;* [4]*Department of Ecology and Evolutionary Biology, UC Irvine, Irvine, California 92697, USA;* [5]*SOKENDAI (Graduate University for Advanced Studies), Hayama, Kanagawa 240-0193, Japan;* [6]*Central European Institute of Technology, Masaryk University, 62500 Brno, Czech Republic*

Transposable elements (TEs) proliferate within the genome of their host, which responds by silencing them epigenetically. Much is known about the mechanisms of silencing in plants, particularly the role of siRNAs in guiding DNA methylation. In contrast, little is known about siRNA targeting patterns along the length of TEs, yet this information may provide crucial insights into the dynamics between hosts and TEs. By focusing on 6456 carefully annotated, full-length Sirevirus LTR retrotransposons in maize, we show that their silencing associates with underlying characteristics of the TE sequence and also uncover three features of the host–TE interaction. First, siRNA mapping varies among families and among elements, but particularly along the length of elements. Within the *cis*-regulatory portion of the LTRs, a complex palindrome-rich region acts as a hotspot of both siRNA matching and sequence evolution. These patterns are consistent across leaf, tassel, and immature ear libraries, but particularly emphasized for floral tissues and 21- to 22-nt siRNAs. Second, this region has the ability to form hairpins, making it a potential template for the production of miRNA-like, hairpin-derived small RNAs. Third, Sireviruses are targeted by siRNAs as a decreasing function of their age, but the oldest elements remain highly targeted, partially by siRNAs that cross-map to the youngest elements. We show that the targeting of older Sireviruses reflects their conserved palindromes. Altogether, we hypothesize that the palindromes aid the silencing of active elements and influence transposition potential, siRNA targeting levels, and ultimately the fate of an element within the genome.

[Supplemental material is available for this article.]

Transposable elements (TEs) comprise the largest proportion of plant genomes, but they are typically silenced by host epigenetic mechanisms. These mechanisms suppress the activity of TEs at both post-transcriptional and transcriptional levels. Post-transcriptional silencing is triggered when TEs escape suppression under stress conditions (Ito et al. 2011), in mutants of methylation maintenance (Miura et al. 2001), and in certain cell types (Slotkin et al. 2009) or developmental stages (Li et al. 2010). Initially, the RNA polymerase II (Pol II)-derived mRNA of the reactivated TE is recognized and processed by RNA-dependent RNA polymerase 6 (RDR6) to produce double-stranded RNA (dsRNA) (Matzke and Mosher 2014). The dsRNA is then cleaved by Dicer-like 2 or 4 (DCL2/DCL4) to generate 21–22 nucleotide (nt) small interfering RNAs (siRNAs); these are loaded onto Argonaute 1 or 2 (AGO1/AGO2) and guide the cleavage of TE mRNA through RNA interference (RNAi). This chain of events is also thought to occur when a new TE invades a "naïve" genome (Panda and Slotkin 2013).

During transcriptional silencing, the RNA-directed DNA methylation (RdDM) pathway orchestrates the deposition of DNA methylation and heterochromatic histone marks on TEs. The key steps initiate with the production of single-stranded TE transcripts (ssRNA) by Pol IV (Fultz et al. 2015); hence, this pathway is termed Pol IV-RdDM. The ssRNA is made double-stranded by RDR2, and in turn, the dsRNA is cleaved by DCL3 into 24-nt siRNAs. These siRNAs are then loaded onto AGO4 (or AGO6). Aided by several associated factors, the AGO4/siRNA duplex targets nascent scaffolding transcripts produced by Pol V to initiate chromatin modifications, including de novo cytosine methylation. This de novo methylation occurs in the symmetric CG and CHG (H = A, C, or T) and also the asymmetric CHH contexts in plants (Feng et al. 2010). Symmetric methylation can be maintained and inherited to daughter DNA strands in an RdDM-independent manner, whereas CHH methylation often requires RdDM and de novo targeting (Zemach et al. 2013). Intriguingly, two recent studies have shown how a transition from post-transcriptional to transcriptional silencing might occur either when some 21- to 22-nt siRNAs are loaded onto AGO6 (McCue et al. 2015) or when DCL3 cleaves some TE mRNAs to produce 24-nt siRNAs (Marí-Ordóñez et al. 2013); in both these RDR6-initiated models (hence termed RDR6-RdDM), the first heterochromatic marks are deposited on active TEs, thereby turning them into suitable substrates for Pol IV-RdDM (Law et al. 2013; Johnson et al. 2014).

Corresponding author: alexandros.bousios@gmail.com
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.193763.115.

© 2016 Bousios et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

One aspect of TE silencing pathways that remains largely unexplored is the patterns of siRNA targeting and methylation along the TE sequence. These patterns may provide important clues into the interaction between TEs and the host response. To date, most studies have focused on the investigation of either consensus sequences (Cantu et al. 2010; Creasey et al. 2014; McCue et al. 2015) or by combining exemplars from different TE orders (Wang et al. 2009; Feng et al. 2010; Zemach et al. 2013). Although informative, these approaches often have low resolution and may not assess how siRNA and methylation levels change as a function of the age of element insertion. Higher resolution work requires the careful annotation of large numbers of individual TEs, which can be difficult in large plant genomes where TEs are numerous, diverse, and often nested. Accordingly, a recent review stressed the need for fine-scale characterization of plant TEs to properly assess the epigenetic interplay with their hosts (Ragupathy et al. 2013).

In this study, we assess epigenetic patterns on Sireviruses within the maize genome. Sireviruses are a genus of plant-specific Long Terminal Repeat (LTR) retrotransposons of the *Copia* superfamily (Bousios and Darzentas 2013). They are unique in structure among LTR retrotransposons, because they contain highly conserved motifs in their noncoding regions (Fig. 1A; Bousios et al. 2010). These motifs include junctions between the LTRs and the internal (INT) domain of the elements, as well as characteristic features such as palindromic repeats within the upstream half of the LTRs that represents the *cis*-regulatory center of LTR retrotransposons (Grandbastien 2015). Due to these conserved motifs, full-length Sireviruses can be identified accurately. We previously reported that ~20% of the maize genome is occupied by Sireviruses and identified thousands of full-length elements that belong to five families, including the abundant *Ji* and *Opie* populations (Bousios et al. 2012a). This TE set provides a unique opportunity to study epigenetic patterns on individual elements that extend from recent to ancient insertions. Here, we map multiple siRNA libraries and bisulfite-sequencing (BS-seq) methylation data to Sireviruses, study the resulting epigenetic patterns, and report that silencing associates with underlying sequence characteristics or the age profile of TEs.

Results

The Sirevirus data

The 13,833 full-length maize Sireviruses currently in MASiVEdb (Bousios et al. 2012b) were originally identified by the Sirevirus-specific MASiVE algorithm (Darzentas et al. 2010). After additional filtering for sequence quality, length, and TE contamination (see Methods), the set consisted of 6456 elements, mainly of the *Ji* (3285) and *Opie* (2926), but also of the *Jienv* (99), *Giepum* (102), and *Hopie* (44) families. Their age distribution, as measured by the sequence divergence of each LTR pair, was similar to that of all 13,833 Sireviruses (Supplemental Fig. S1). This suggests both that our strict requirements did not favor the inclusion of young over old Sireviruses, and the filtered set is a representative sample.

General characteristics of siRNA mapping

We mapped the previously published leaf (Diez et al. 2014), tassel (Zhang et al. 2009), and immature ear (Nobuta et al. 2008) siRNA libraries to the reference maize B73 genome and assessed mapping patterns for 21-nt, 22-nt, and 24-nt siRNAs separately. Each distinct siRNA sequence was termed "species," and its number of reads was termed "expression." We considered siRNA species that mapped to either unique (U_siRNAs) or multiple (M_siRNAs) locations in the genome. Next, we counted both the number of siRNA species that mapped to Sireviruses and also their expression levels. The two measurements were highly correlated (average Pearson $r = 0.91$ for all siRNA lengths; $P < 10^{-20}$); hence, we primarily report results based on species for simplicity. Finally, mapping to individual elements was highly correlated between pairs of the three libraries (Supplemental Fig. S2), indicating that the same Sireviruses were consistently targeted across libraries.

In total, a relatively small fraction of each siRNA library mapped to the 6456 elements (Supplemental Table S1). For example, the immature ear library yielded the highest proportions with 3.1% of the 1,118,020 24-nt species, 11.3% of the 177,719 22-nt species, and 8.8% of the 106,514 21-nt species. In contrast, the leaf library yielded the lowest proportions (2.7%–5.0%) for each siRNA length. Although these percentages appear to be low, we note that the 6456 elements encompass a total sequence length of ~65 Mb, or ~2.8% of the 2300-Mb maize genome. Of the siRNAs that mapped to Sireviruses, >97% were multiple mapping M_siRNAs (likely due to a large number of highly similar elements), which contrasts with the much higher U_siRNA to M_siRNA ratio (~1.2:1) across the entire genome (Supplemental Table S1). Nonetheless,

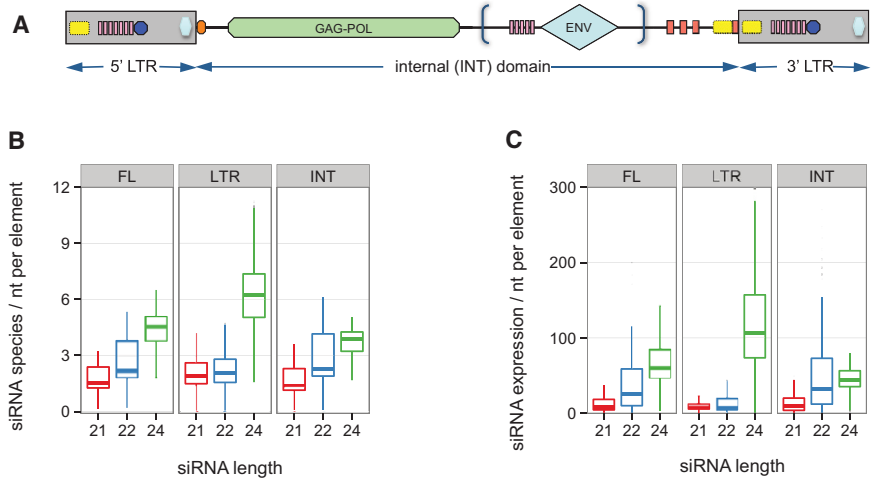

Figure 1. The Sirevirus genome and general siRNA mapping patterns. (*A*) A schematic of a Sirevirus element (based on Bousios et al. 2010). The *gag*/*pol* genes are shown in green, and the *envelope*-like gene (present in some families only) as a light blue diamond. An inverted repeat (IR; yellow) surrounds the junction of the internal (INT) domain with the 3′ LTR. The outmost 5′ side of the junction is occupied by a polypurine-tract (PPT; red). Additional PPTs cluster upstream of the junction. The palindromes (pink) are located in the *cis*-regulatory area of the LTRs preceding a conserved TATA box (blue circle), and near the *envelope*-like gene (when present). The 5′ LTR/INT domain junction harbors a C-rich integrase signal (light blue hexagon) and the primer binding site (orange box). (*B*,*C*) Number of siRNA species (*B*) and expression (*C*) per nucleotide of the full-length (FL), LTR, or INT domain of all 6456 Sireviruses based on the leaf library.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

these M_siRNAs tended to match Sireviruses exclusively, because only 0.1% mapped to both Sireviruses and exons from the Filtered Gene Set, and only 0.6% mapped to both Sireviruses and the non-Sirevirus maize TE exemplars (http://maizetedb.org).

At the family level, *Ji* and *Opie* differed considerably in the number of mapped siRNA species across libraries, despite their similar ages (Supplemental Fig. S1), length distributions (Supplemental Fig. S3), and numbers within the 6456 elements. For example, 54% (78,714) compared to 26% (38,187) of the 144,817 siRNA species of the leaf library mapped to the *Ji* and *Opie* populations, respectively (Supplemental Table S2), of which only 2966 (2.6%) cross-mapped to both families. In comparison to *Ji* and *Opie*, fewer 24-nt siRNAs mapped to the three less abundant families *Jienv*, *Giepum*, and *Hopie*; however, these were often mapped by more 21- to 22-nt siRNAs. These observations were further supported by comparing the average number of siRNA species that mapped to each element of each family, which was first normalized by the family's average genome length to allow cross-family comparisons (Supplemental Table S2). Altogether, these findings corroborate previous evidence that siRNA targeting does not necessarily correlate with TE abundance (Barber et al. 2012; Diez et al. 2014).

siRNA targeting along Sirevirus sequences

We then investigated whether siRNA targeting varied along the length of elements. To do so, we first tagged each Sirevirus nucleotide with the number of siRNA species (or their expression) that mapped to it and then averaged across the length of the locus under investigation, i.e., the full-length element, the LTRs, or the INT domain. Summarizing across families, we found that 24-nt siRNAs targeted Sireviruses more intensely than 21- to 22-nt siRNAs, and this was particularly true for the LTRs (Fig. 1B,C). Conversely, 21-nt siRNAs targeted the LTRs and INT domain similarly, whereas 22-nt siRNAs targeted the INT domain more heavily. These patterns generally held across libraries, although the predominance of 24-nt siRNAs was not as apparent in tassel (Supplemental Fig. S4).

Next, we divided each LTR or INT domain into 100 equally sized windows and calculated the per nucleotide coverage for each library, focusing more on the abundant *Ji* and *Opie* families. Overall, this approach revealed three main patterns. First, mapping to the LTRs was nonuniform, because both *Ji* and *Opie* exhibited distinct siRNA peaks, or mapping "hotspots" (Fig. 2A).

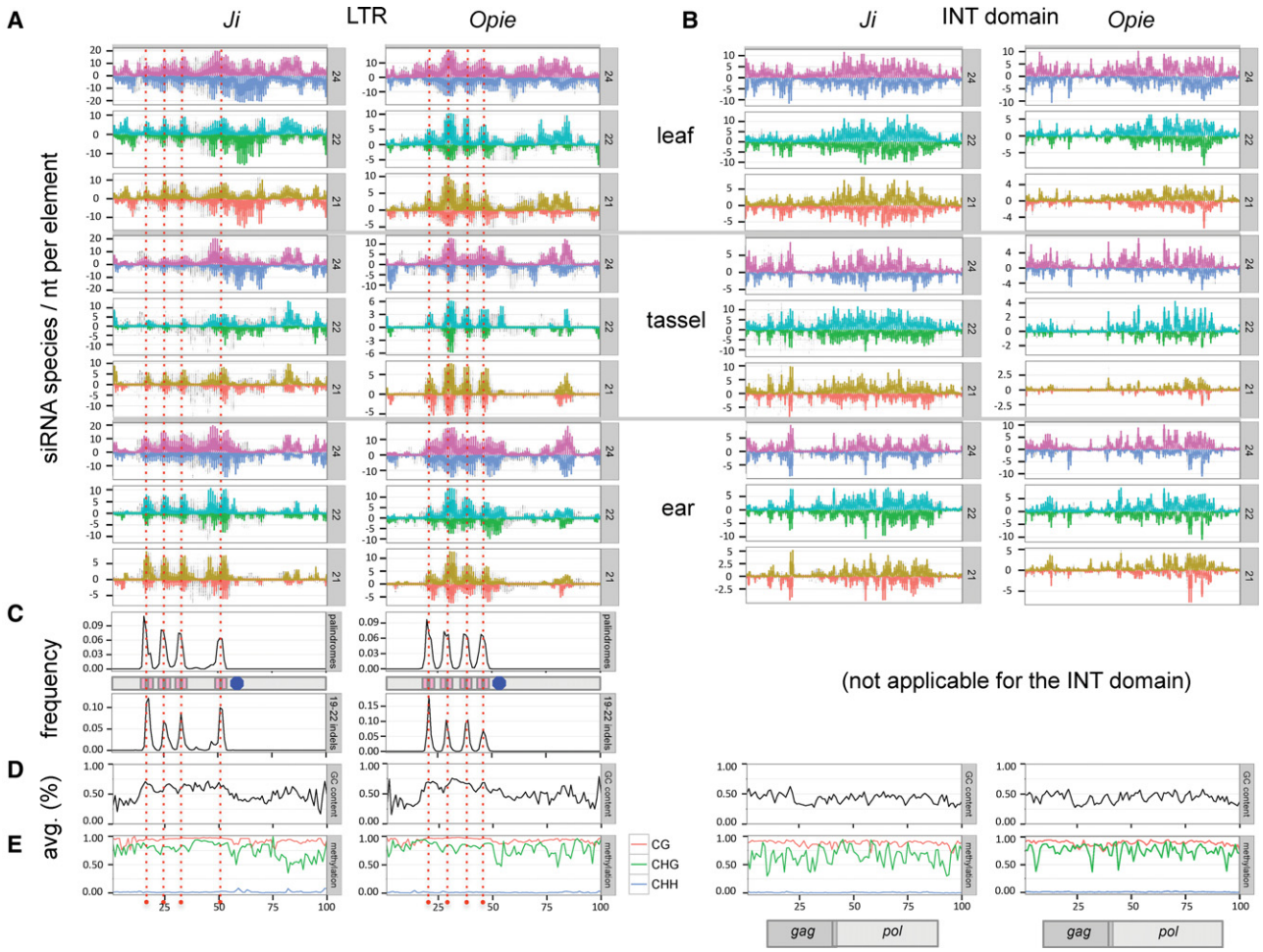

Figure 2. Epigenetic and sequence patterns of *Ji* and *Opie*. Each LTR and INT domain was first split in 100 equally sized windows of ~13-nt and ~67-nt length, respectively. (*A*,*B*) Number of siRNA species per nucleotide along the sense (positive *y*-axis) and antisense (negative *y*-axis) strands of the LTRs (*A*) and INT domain (*B*) of 3285 and 2926 *Ji* and *Opie* elements, respectively (visualized with a box plot for each window). (*C*) Distribution of palindromes (*top*) and 19- to 22-nt-long indels (*bottom*). A schematic shows the approximate positions of the palindromes (pink boxes) and promoter (blue circle). (*D*) Average GC content. (*E*) Average methylation level. For the INT domain, the approximate positions of the *gag* and *pol* genes are shown.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

The location of hotspots varied between the two families, but for both they were ~40–60 nt in length and more easily discerned in floral libraries and for 21- to 22-nt siRNAs. Second, mapping to the INT domains was also nonuniform, including a ~1-kb siRNA-poor region followed by a ~3.5-kb region of increased targeting (Fig. 2B). The ~3.5-kb region corresponded to the location of the *pol* gene, whereas the siRNA-poor region overlapped with the 3′ end of the *gag* gene. Finally, siRNA mapping to *Jienv*, *Giepum*, and *Hopie* generated similar siRNA hotspots in the LTRs (Supplemental Fig. S5A); these families contain an *envelope*-like gene in their INT domain (Bousios et al. 2012a), which also tended to have many mapped siRNAs (Supplemental Fig. S5A).

siRNA hotspots correspond to the palindrome region

To further investigate the siRNA hotspots, we examined additional sequence features of Sireviruses. Notably, most hotspots were situated within the upstream half of the LTRs, which corresponds to the *cis*-regulatory palindrome-rich region of Sireviruses (Fig. 1A). The consensus sequence of the palindrome for *Ji* and *Opie* is CACCGGACtGTCCGGTG, and it was originally found in multiple proximal pairs in the archetypical *Ji* and *Opie* sequences in GenBank (Bousios et al. 2010). Allowing for one mismatch, we identified the palindrome 86,055 times, for an average of 7.3 copies within each *Ji* and *Opie* LTR (Fig. 3A). Sequence analysis of the 86,055 copies revealed that the symmetrical arms were more conserved than the central nucleotide (Fig. 3B).

It is known that direct and inverted repeats mediate the formation of indels and copy number variation both within TEs and genome-wide (Ma et al. 2004; Wicker et al. 2010). During the process of aligning LTR pairs, we observed that the lengths of indels across all elements were distinctly nonrandom, with an overabundance of 19- to 22-nt indels (Fig. 3C). In total, 5060 19- to 22-nt indels were identified in the *Ji* and *Opie* elements, representing 25% of all their indels. We retrieved whole or fragments of palindromes from >75% of the 19- to 22-nt indel sequences, indicating their overlap. Indeed, the locations of both the palindromes and the long indels formed four narrow loci in each family that also overlapped in large part with the siRNA hotspots (Fig. 2C) and with regions of high (50%–75%) GC content (Fig. 2D). In contrast, for siRNAs mapping to the INT domain, we could not detect any obvious overlaps between siRNA hotspots and underlying sequence characteristics other than the distinct preference for coding over noncoding areas.

We tested whether the four palindrome loci had significantly higher numbers of siRNA species per nucleotide than the rest of the LTR regions, by using the Mann-Whitney *U* test (*P*-values adjusted by Bonferroni correction) and combining information from both strands. Each locus was first allocated four consecutive windows (see previous section) based on the palindrome's locations (*Ji* windows: 16–19; 24–27; 31–34; 50–53; *Opie* windows: 19–22; 28–31; 36–39; 44–47), and the four loci were then combined for this analysis. The test showed that the siRNA/palindromes overlap was statistically significant in most cases, especially in floral tissues and for smaller siRNA lengths (e.g., $P = 3.4 \times 10^{-5}$ and 5.9×10^{-9} for 21-nt siRNAs in tassel for *Ji* and *Opie*, respectively) (Supplemental Fig. S6A). Nonetheless, these results could be caused by mapping artifacts, especially given

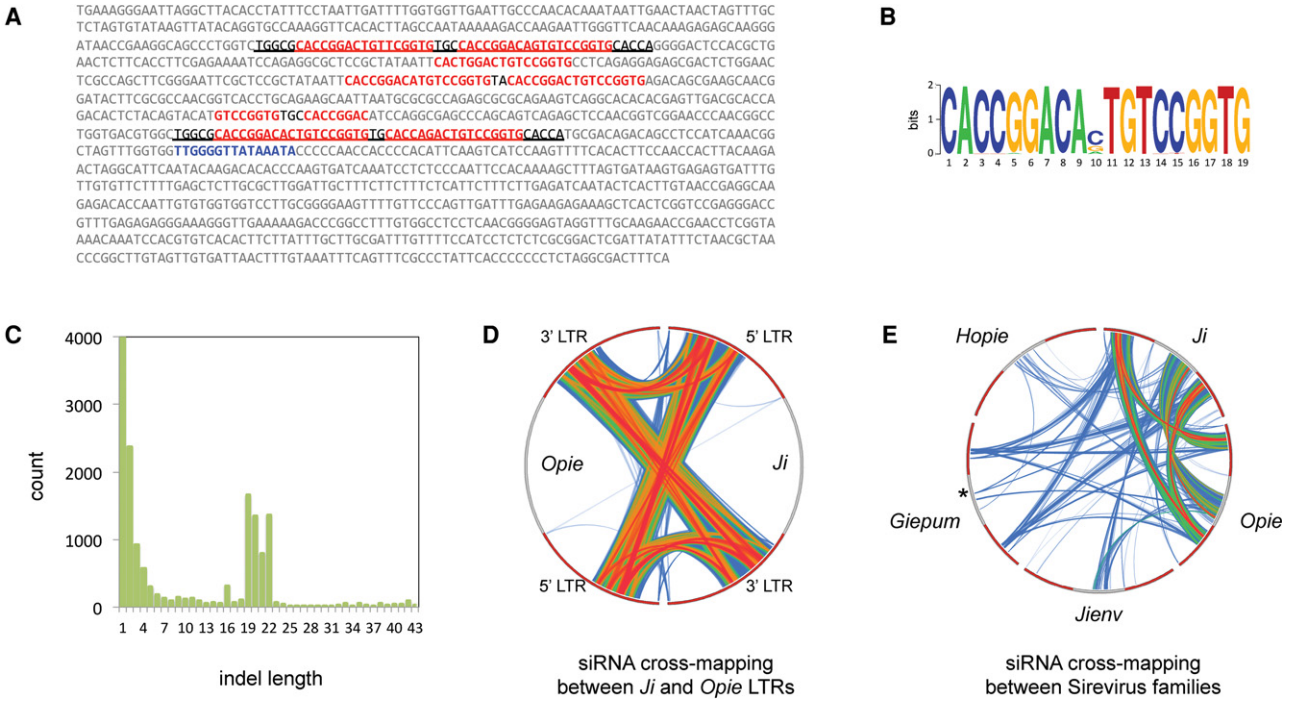

Figure 3. Sequence and cross-mapping characteristics of the palindrome region. (*A*) An example of a *Ji* LTR showing the palindromes (red) and the conserved promoter (blue). Underlined are the fragments that form the long stem indicated by the arrows in Figure 4A. (*B*) Sequence logo of the 86,055 palindromes. (*C*) Length of indels identified during LTR pair alignment for *Ji* and *Opie* elements. The *y*-axis is cut off at 4000 so that the scale is appropriate for indels >1 nt. (*D*,*E*) Location of cross-mapping siRNAs between *Ji* and *Opie* LTRs (*D*) and all Sirevirus families (*E*) using Circos (Krzywinski et al. 2009). siRNAs that did not map to two or more elements within each family were removed. Ribbons connect windows (described in Fig. 2) to which each siRNA species mapped; red indicates an increased number of overlapping ribbons. Moving clockwise, the outer circle denotes the 5′ LTR, INT domain, and 3′ LTR. The black asterisk in *E* indicates the area near the *envelope*-like gene of *Giepum* that "cross-talks" with the palindrome region of *Ji* LTRs.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

the repetitive nature of palindromes, the multiple locations of M_siRNAs, and the fact that we examined only one mapping statistic, i.e., the average number of siRNA species per nucleotide. Accordingly, we investigated additional strategies, such as measurements of siRNA expression and the use of both metrics normalized by their number of genomic locations, hence, controlling for the effect of M_siRNAs. Across all approaches, the palindromes remained statistically significant siRNA hotspots for most combinations of tissue and siRNA length (Supplemental Fig. S6B–D).

Overall, a large proportion of all LTR-targeting siRNA species mapped to the four palindrome loci. For example, 62% and 71% of the 21-nt siRNAs in the immature ear library targeted the *Ji* and *Opie* loci, respectively (Supplemental Table S3), although they collectively comprised only 16% of the ~1.3-kb LTR length. Furthermore, the whole palindrome-rich region, which contained all four loci and the GC-rich part of the LTR (*Ji* windows: 16–53; *Opie* windows: 19–48), mapped up to 89% of LTR-targeting siRNAs (Supplemental Table S3). This was clearly demonstrated when we plotted the total number of siRNA species that mapped to each window of all *Ji* or *Opie* LTRs combined (Supplemental Fig. S7). Within this larger region, however, the four loci were the areas of highest siRNA mapping (Mann-Whitney *U* test; *P* values adjusted by Bonferroni correction; $P < 3.4 \times 10^{-2}$).

The palindrome loci differ considerably at the sequence level

Because the palindrome loci of *Ji* and *Opie* are composed of the same building block, it is possible that they represent identical sequences mapped by the same set of siRNAs. Three observations, however, suggest this not to be the case. First, when we counted the total number of siRNA species that mapped to each locus, we recorded up to a threefold difference among loci within each family (Table 1). Second, only a small proportion (2.3%–19.9%) of siRNAs mapped to any pair of the four *Ji* or *Opie* loci (Table 1). Finally, only a few siRNAs cross-mapped between the LTRs of the two families, i.e., 1265 (2.1%), 613 (3.7%), and 1298 (5.3%) of all LTR-targeting siRNAs for the leaf, tassel, and immature ear libraries, respectively. These results suggest that each locus represents a distinct sequence that is targeted by distinct sets of siRNAs. That said, the few cross-mapping siRNAs specifically clustered to the palindrome loci (Fig. 3D). It appears, therefore, that the palindromes form a backbone for siRNA targeting but also a template for sequence evolution.

Supporting this conjecture, we found that the LTRs of the three lower copy families exhibited similar correspondences between indel size and location, GC content, and siRNA targeting (Supplemental Fig. S5A–C). Moreover, analysis of the cross-mapping siRNAs among all families revealed that *Giepum* shared part of the *Ji* and *Opie* backbone in its LTRs but also in the vicinity of the *envelope*-like gene (Fig. 3E), where additional palindromes have been previously reported (Bousios et al. 2010). In contrast, the lack of cross-mapping siRNAs for *Jienv* and *Hopie* implies that they may harbor distinct sets of palindromes.

Table 1. Comparison of total numbers of siRNA species that mapped to the four palindrome loci of all *Ji* or *Opie* LTRs combined

		Number of siRNA species from all lengths combined			Pairs of loci	siRNA species that mapped to both loci (%)		
Family	Locus	Leaf	Tassel	Ear		Leaf	Tassel	Ear
Ji	1st	2186	813	1898	1_2	7.4	14.7	13.0
	2nd	3577	1154	2557	1_3	2.8	5.9	4.5
	3rd	4065	1234	2853	1_4	3.3	5.6	7.3
	4th	5223	2389	3008	2_3	16.0	19.9	18.0
					2_4	2.8	5.7	6.9
					3_4	2.3	4.2	5.2
Opie	1st	1316	537	1592	1_2	8.7	15.6	10.8
	2nd	2268	795	2532	1_3	5.0	10.3	7.1
	3rd	1922	635	2025	1_4	6.1	10.0	8.7
	4th	1899	781	1973	2_3	7.3	15.9	10.6
					2_4	5.2	7.6	7.2
					3_4	7.5	11.7	10.5

The palindrome region forms hairpins

It seems likely that the abundance and proximity of palindromes in the LTRs (Fig. 3A) could trigger the formation of stem–loop secondary structures. To test this idea, we calculated the folding potential of *Ji* and *Opie* LTRs using RNAfold (Gruber et al. 2008) and found that the majority could form complex structures in their upstream halves similar to the example shown in Fig. 4A. We then speculated whether this region could also form appropriate precursors for the production of a recently reported class of plant small RNAs termed hairpin RNAs (hpRNAs) (Axtell 2013a). To examine the folding strength of such potential precursors, we followed the methodology of Wang et al. (2009), who reported that loci for known maize miRNAs typically had a minimum free energy (MFE) below −40. Extending the mapping location of each siRNA by 20 nt upstream and 70 nt downstream and calculating the MFE of these 111- to 114-nt fragments, we found that those that mapped to the palindrome region had such an MFE (Fig. 4B). Moreover, when we split siRNAs into three mapping locations, i.e., the palindrome-rich LTR region as previously defined, the rest of the LTR, and the INT domain, the MFE was consistently below −40 only for the first across all *Ji* and *Opie* elements (Fig. 4C). This was also true for *Jienv*, *Giepum*, and *Hopie* elements (Supplemental Fig. S5D). In addition, these families also produced putative precursor structures with MFE below −40 in narrow regions upstream of their *envelope*-like gene, which further supports the existence of palindromes in the specific area. Altogether, these results suggest that the palindrome region may be a suitable template for the production of hpRNAs.

siRNA targeting and the age of Sireviruses

Many studies on TE silencing indicate that the patterns of 21-, 22-, and 24-nt siRNA targeting could vary as a function of TE age (Teixeira et al. 2009; Ito et al. 2011; Marí-Ordóñez et al. 2013); however, these models tend to focus on the initial stages of TE infection or reactivation. In contrast, little is known about the dynamics between siRNA targeting and TE age on a longer time scale. Given that the age distribution of Sireviruses spans millions of years (my) (Supplemental Table S4), our data set is suitable for studying these evolutionary dynamics.

Using the per nucleotide coverage metric, our analyses produced a strong negative correlation between age and siRNA mapping for all siRNA lengths and families except *Jienv* (Fig. 5A). Nevertheless, we noticed an unexpected phenomenon: *Ji* and *Opie* elements older than ~2.0–2.5 my deviated from the general pattern by having a similar number of matching siRNAs as their younger counterparts. As a result, when we excluded elements older than 2 my from the statistical analysis, the strength of the correlation increased substantially (average Pearson $r = -0.84$

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

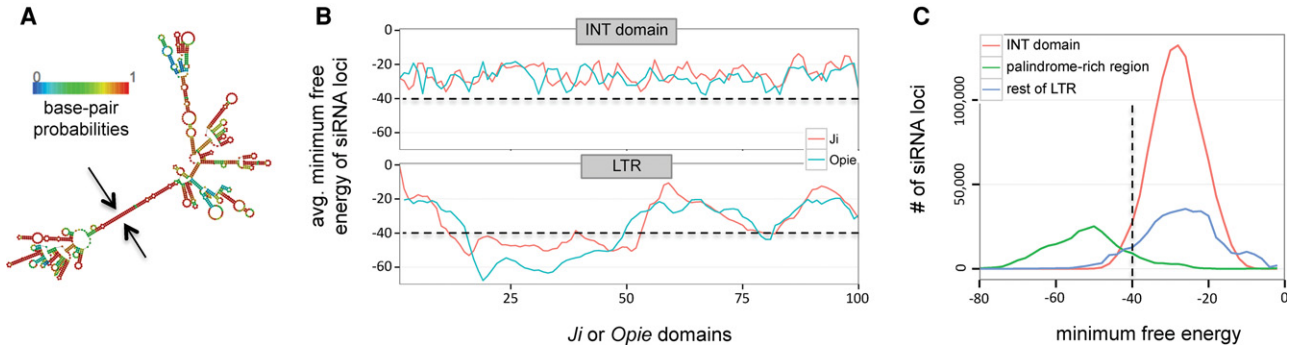

Figure 4. Secondary structure of Sirevirus LTRs. (*A*) Predicted hairpin formation of the *Ji* LTR shown in Figure 3A using RNAfold (Gruber et al. 2008). Arrows represent the underlined fragments in Figure 3A. (*B*) Average minimum free energy (MFE) of the siRNA loci, 20 nt upstream + siRNA length + 70 nt downstream (as in Wang et al. 2009) along *Ji* and *Opie* LTRs and INT domains, which were first split in 100 equally sized windows of ~13-nt and ~67-nt length, respectively. The lines represent averages for the windows across all elements. (*C*) Distribution of MFE for siRNA loci mapping to the INT domain, the palindrome-rich region of the LTR (38 windows for *Ji*; 30 windows for *Opie*), and the rest of the LTR of all *Ji* and *Opie* elements combined. The miRNA cutoff of −40 MFE is highlighted in *B* and *C* by dotted lines.

compared to −0.45 for all siRNA lengths; $P < 10^{-20}$). This pattern was evident in all libraries and even when we analyzed other subsets of elements from MASiVEdb (Supplemental Fig. S8). We were unable to determine whether the same pattern occurred in *Jienv*, *Giepum*, and *Hopie* because of their small populations and lack of elements older than 2.5 my (Supplemental Table S4).

We further contrasted these dynamics among three distinct age groups: "very young" (VY) elements less than 0.5 my; "middle-aged" (MA) elements between 1.5 and 2.0 my; and "very old" (VO) elements older than 3.0 my. Counting the number of mapped siRNA species revealed that VO elements matched more siRNAs than MA elements (Fig. 5B). For example, on average, 435 24-nt siRNA species mapped to the 5′ LTR of VY *Ji* elements compared to approximately 262 and 358 for MA and VO elements, respectively (Supplemental Table S5). The same pattern was observed in the INT domain; this suggests that the high mapping of VO elements was not solely fueled by the expected age-dependent increase of LTR pair divergence (Supplemental Fig. S9) that could naturally generate higher sequence variability and therefore allow targeting by more siRNA species.

siRNA 'cross-talk' between young and old elements

Because siRNAs map to multiple Sireviruses, we asked whether the high targeting of old elements might be due to increased cross-mapping with younger elements. Hence, we examined mapping patterns among *Ji* or *Opie* age groups. For example, we took each *Opie* VY element and calculated the number of 24-nt siRNA species that mapped both to (1) either of its LTRs, and (2) at least one LTR of any MA or VO element. On average, 97.6% of the siRNA species that mapped to the LTRs of any VY *Opie* element also cross-mapped to the LTRs of one or more VO elements; in contrast, only 68.1% mapped to one or more MA elements (Supplemental Table S6).

This observation prompted us to investigate cross-mapping in more detail. To control for potential biases due to the differing sizes of age groups, we randomly sampled 30 elements from each age group for *Ji* or *Opie* and formed "triplets" of one VY, one MA, and one VO element; 27,000 triplets for each family. We next allocated each siRNA species into one of seven possible categories based on the element(s) of the triplet that it mapped to. The process was repeated three times for each siRNA length and tissue. In both families, a large proportion (73% on average) of the siRNAs was unique to one age group (Fig. 5C; Supplemental S10; Supplemental Table S7), unsurprisingly considering that this was an analysis of only three elements at a time. However, a significant fraction of the remaining siRNAs was shared only by the VY and VO elements (ranging between 7.0% and 16.2% for *Ji* and 14.3% and 24.9% for *Opie*) as opposed to only 0.8%–6.0% for VY-MA and MA-VO. This difference was statistically significant for all tissues and siRNA lengths (Wilcoxon signed-rank test; $P < 10^{-10}$), and again more evident for floral tissues and smaller siRNA lengths (Supplemental Table S7). We also checked whether phylogenetic histories within each family might artifactually drive this pattern (e.g., if VY and VO elements represent a distinct clade from MA elements), but this was not the case (see Supplemental Material; Supplemental Fig. S11). Taken together, these results suggest that the VY-VO siRNA "cross-talk" is a true property of Sirevirus evolution.

Degeneration and conservation of the palindrome loci

Thus far, the reasons for the increased mapping of siRNAs to VO elements and their "cross-talk" with VY elements are unclear. Inspired by the nonuniform siRNA targeting pattern along Sireviruses, we examined the distribution of siRNA species along *Ji* and *Opie* elements for the three age groups separately. Although the mapping pattern was complex, there were evident differences in the palindrome region, with fewer siRNAs matching to MA compared to VY and VO elements (Fig. 6A,B; Supplemental S12A,B). In contrast, other regions such as the first ~150 nt or the 3′ end of *Ji* LTRs had similar mapping dynamics across all three age groups.

We further investigated the occurrence of palindromes and 19- to 22-nt indels in each age group. As expected, the average number of all indels in each LTR pair increased with age (Table 2); however, MA elements had the highest number of 19- to 22-nt indels, and accordingly, the lowest number of palindromes. In addition, although the average number of substitutions in each LTR pair also increased with age as expected (Table 2), VO was the only age group with a distinct and statistically significant lack of substitutions in the four palindrome loci (Mann-Whitney *U* test; *P*-values adjusted by Bonferroni correction; 4×10^{-5} for *Ji* VO using windows 5–80 and 2×10^{-5} for *Opie* VO using all windows) (Fig. 6C; Supplemental Fig. S12C). These data indicate that the palindrome regions are conserved in VO elements and contribute both to their higher-than-expected siRNA targeting and to the cross-mapping with VY elements. That said, siRNA mapping to

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

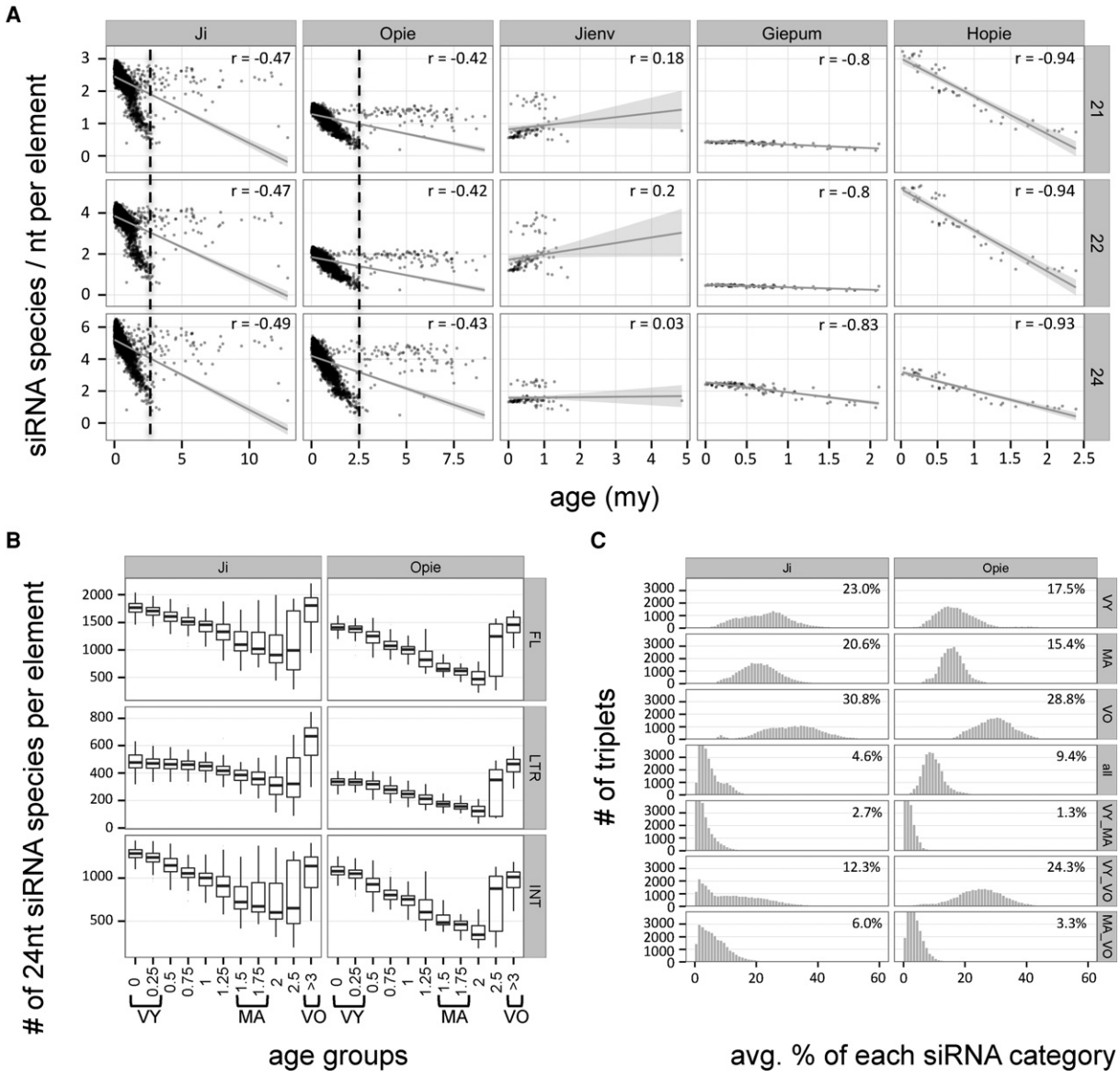

Figure 5. Sirevirus age and siRNA targeting. (*A*) Relationship between age and number of siRNA species of the leaf library calculated per nucleotide of the full-length sequence for each siRNA length and family separately. The dotted lines indicate the cutoff point around 2.0–2.5 my, after which elements did not exhibit decreased targeting. The shading around each regression line represents the 0.95 confidence interval, while the Pearson *r* is shown for each plot (*P* value $<10^{-20}$ in all cases, except *Jienv*). (*B*) Number of 24-nt siRNA species of the leaf library that mapped to the LTRs, INT domain, or full-length sequence of *Ji* or *Opie* elements across age groups (see Supplemental Table S4 for their population sizes). Very young (VY), middle-aged (MA), and very old (VO) elements are indicated. (*C*) Twenty-one nucleotide siRNA species of the immature ear library cross-mapping to the LTRs of 27,000 *Ji* or *Opie* triplets. Each siRNA was classified across seven types based on the element(s) of the triplet it mapped to. The average percentage of each type is shown. The *y*-axis is cut off at 4000 my.

the INT domain was also less intense in MA elements, but it was difficult to pinpoint specific regions of differential targeting (Supplemental Fig. S12D). Thus, although the conservation of the palindrome region may be essential, it does not alone explain the phenomenon.

Methylation levels of Sireviruses

The purpose of the siRNA response is ultimately to deposit heterochromatic marks on TEs. We therefore investigated Sirevirus methylation levels based on a subset of the 6456 elements that exceeded BS-seq coverage cutoffs (see Supplemental Material). Consistent with previous studies of maize TEs (Regulski et al. 2013; West et al. 2014), Sireviruses were highly methylated in the CG (~95%) and CHG (~84%) contexts but lacked extensive CHH (~1.2%) methylation (Supplemental Fig. S13A). Methylation was higher in the LTRs than in the INT domains, whereas CG methylation appeared to plateau across the palindrome region in the LTRs (Fig. 2E). We also investigated the evolutionary relationship between methylation and age. The analysis produced weak, yet significant positive correlations between symmetric methylation and age, but only for the *Ji* and *Opie* LTRs (average Pearson $r = 0.1$;

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

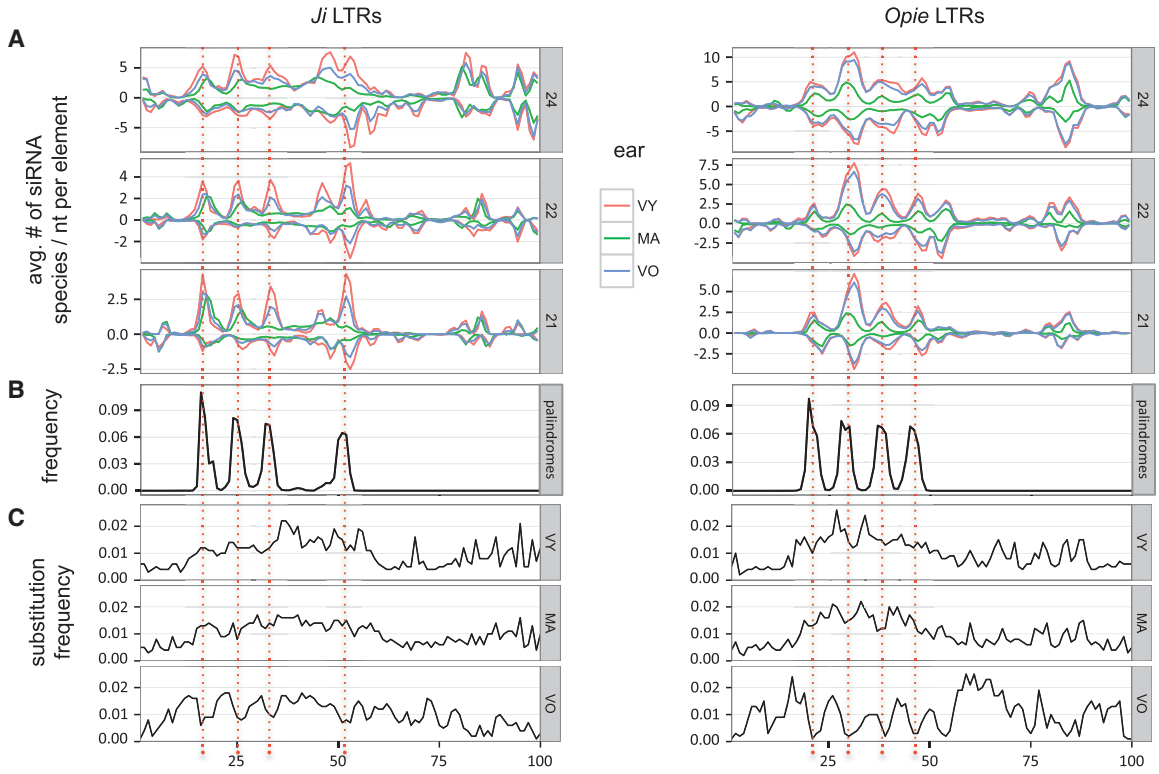

Figure 6. siRNA targeting and conservation of the palindrome-rich region as a function of age. Each *Ji* or *Opie* LTR was first split in 100 equally sized windows of ~13 nt. (*A*) Number of siRNA species of the immature ear library per nucleotide of the sense (positive *y*-axis) and antisense (negative *y*-axis) strands of very young (VY), middle-aged (MA), and very old (VO) elements (see Supplemental Table S4 for their population sizes). The average of each window is depicted instead of a box plot as in Figure 2A. (*B*) Distribution of palindromes. (*C*) Distribution of nucleotide substitutions identified during LTR pair alignment.

$P < 10^{-20}$) (Supplemental Fig. S13B). This correlation coefficient increased to $r = 0.25$ when we excluded elements older than 2 my, which again deviated from the general pattern (Supplemental Fig. S13C). Conversely, CHH methylation in the LTRs correlated negatively with age, which also became stronger by excluding old elements (average Pearson $r = -0.22$ compared to -0.42; $P < 10^{-20}$).

Discussion

We have analyzed epigenetic features of LTR retrotransposons with the goal of gathering insight into the interaction between hosts and TEs. With our carefully annotated Sirevirus data set, we have shown that silencing associates with sequence characteristics of the TE itself, a theme that remains unexplored partially due to difficulties in accurately identifying large populations of TEs. We have uncovered variation in siRNA targeting among families, along the length of elements and as a function of element age. We discuss these findings in more detail below.

siRNA targeting at the family level

Ji and *Opie* are similar in their genomic abundance, age, and length distributions (Bousios et al. 2012a) and are also represented by similar numbers in our data set. However, *Ji* elements are targeted by considerably higher numbers of siRNA species, especially of 21- to 22-nt lengths (Supplemental Table S2). Higher level of targeting to *Ji* elements has been noted previously; Barber et al. (2012) found *Ji* to be the most targeted LTR retrotransposon family in maize. When they categorized TEs by their predominant siRNA length, *Ji* and *Opie* were defined as "22-nt" and "24-nt" families, respectively, which is in agreement with our findings. Among the three

Table 2. Number of indels, palindromes, and nucleotide substitutions per element for each age group

Family	Age group[a]	Number of SVs	All indels[b]	19- to 22-nt indels[b]	Palindromes	Substitutions[b]
Ji	VY	1611	1.9	0.5	7.7	6.9
	MA	161	5.3	1.1	5.7	50.6
	VO	56	15.7	0.9	6.6	188.1
Opie	VY	1503	2.2	0.6	8	6.2
	MA	115	4.8	1.2	5.3	51.4
	VO	80	14.7	0.7	7.5	164

[a](VY) very young; (MA) middle-aged; (VO) very old; (SVs) Sireviruses.
[b]Indels and substitutions were scored during the LTR pair alignment of each element.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

remaining families, *Jienv* and *Hopie* have high 21- to 22-nt targeting, whereas *Giepum* is nearly exclusively targeted by 24-nt siRNAs in leaf and tassel tissues. At present, the biological causes for variable targeting are unclear, but it could reflect differences in the number of elements that escape repression and produce Pol II transcripts in vivo, which are then processed by RDR6-RdDM into 21- to 22-nt siRNAs. Another possible cause may be genomic location, since there are peaks of RdDM activity near genes (Gent et al. 2013; Zemach et al. 2013). However, *Ji* and *Opie* have similar distributions in relationship to gene proximity (Bousios et al. 2012a) and the same weak negative correlation between siRNA mapping and distance to genes (average Pearson $r = -0.08$; $P < 10^{-10}$), suggesting that gene proximity explains at best a minor facet of siRNA targeting variation.

The function, complexity, and evolution of the palindrome region

The most profound similarity between *Ji* and *Opie* is the overlap in the LTRs between siRNA hotspots and the palindromes (Fig. 2A; Supplemental Figs. S6, S7; Supplemental Table S3). The intensity of this overlap, however, varies among tissues and among siRNA lengths: There is a consistent increase from leaf to floral tissues and from 24-nt to 21- to 22-nt siRNAs. These findings prompted us to examine an additional siRNA library from the immature ear of a *mop1* (*rdr2*) mutant (Nobuta et al. 2008). As expected, the *mop1* mutant had lower numbers of 24-nt siRNAs compared to the other libraries (Supplemental Table S1), but it retained the same siRNA hotspots in the palindromes and produced similar patterns to the wild-type ear library (Supplemental Fig. S14).

Several lines of evidence suggest that *Jienv*, *Giepum*, and *Hopie* also contain palindromes: They generate long indels, have high GC content, and are predicted to form secondary structures in the same region. Furthermore, cross-mapping of siRNAs between the *Giepum*, *Ji*, and *Opie* LTRs occurs only in the palindrome region (Fig. 3E). It appears, therefore, that the palindromes are the building blocks of a conserved backbone among families. This conservation possibly extends across the plant kingdom based on our previous finding that most Sireviruses contain similar symmetrical motifs (Bousios et al. 2010). Importantly, the palindromes are located in the highly conserved *cis*-regulatory region of LTR retrotransposons (Grandbastien 2015) and retroviruses (Mergia et al. 1992). In some families, this area is arranged similar to Sireviruses as arrays of few repeats or may display symmetrical features (Araujo et al. 2001; Grandbastien 2015). Variation in the repeats among subfamilies has been suggested to confer new regulatory properties for optimizing coevolution within their hosts or for colonizing new species (Vernhettes et al. 1998; Araujo et al. 2001). Thus, there is substantial evidence that this region is important for the *cis*-regulation of Sireviruses.

Nevertheless, although the palindrome structure may be necessary for the function of Sireviruses, its sequence content is apparently not conserved. It appears to differ (1) among families, (2) among family members, and (3) among the four discrete palindrome loci. The first point is supported by the small proportion of cross-mapping siRNAs between the palindrome-rich regions of the *Ji* and *Opie* LTRs (< 5.3%), which nonetheless represent the only point of "contact" between their LTRs (Fig. 3D). The second point follows from our observation that the number of palindromes changes with age (Table 2). The third point is based on the small proportion (<20%) of siRNAs that match any pair of the four loci (Table 1). Consequently, the makeup of this region is complex and has evolved differently for each locus. The mechanisms that generate such sequence variation have yet to be elucidated. However, reverse transcription has very low fidelity, which could mediate the loss and gain of palindrome variants. Furthermore, elements might occasionally swap regions of their LTRs through recombination, for example, when their transcripts co-package within the same virus-like particle during reverse transcription (Sharma et al. 2008; Du et al. 2010).

Palindrome hpRNAs may trigger silencing of active Sireviruses

Another prominent feature of the palindromes is that they are an obvious source of stem–loop structures and a potential template for hpRNA production (Fig. 4). hpRNAs are an understudied class of small RNAs that resemble miRNAs by being derived by ssRNA precursors (Axtell 2013a). One study has focused on a large 6-kb hairpin in *Arabidopsis* and found that DCL1 was used for hpRNA biogenesis (Henderson et al. 2006). Such long hairpins do not occur within LTR retrotransposons, but Axtell (2013a) hypothesized that numerous shorter hairpins might also produce hpRNAs. In fact, Wang et al. (2009) analyzed the sequence composition of hpRNA loci in maize to show that they are exceptionally GC-rich, an observation consistent with the high GC content of the Sirevirus palindrome region (Fig. 2D).

The capacity to form hairpins may offer insight into a poorly understood step of the silencing pathway. Previous studies have established how RDR6-RdDM initiates silencing of active TEs (Marí-Ordóñez et al. 2013; McCue et al. 2015), but it remains unclear how RDR6 physically recognizes single-stranded TE mRNA to produce dsRNA. It is believed that "primary" siRNAs are required to trigger dsRNA synthesis, either directly as primers or indirectly by mediating AGO-directed cleavage of the ssRNA (Bologna and Voinnet 2014). Recently, miRNAs have been implicated for this role in *Arabidopsis* (Creasey et al. 2014), but this occurs in a subset of TE families, suggesting that additional miRNA-independent mechanisms are needed. One of these additional mechanisms may entail the existence of hairpins within the TE sequence (Lisch and Slotkin 2011), because hairpins may be recognized directly by DCL1/DCL2 and processed into "primary" siRNAs. Secondary RNA structures have been searched for, but not found, during silencing of the *Evade* element in *Arabidopsis* (Marí-Ordóñez et al. 2013), but an inverted repeat in a maize *MuDR* element has been shown to trigger RNAi and de novo transcriptional silencing of the whole family (Slotkin et al. 2005). Under this model, we predict that palindrome-derived hpRNAs may play a role in silencing active Sireviruses by functioning as "primary" siRNAs. Of course, this conjecture requires further experimental confirmation.

Age and evolutionary fate of Sireviruses

Our analysis suggests that age is a major factor in the dynamic between the host response and Sireviruses. Generally, age is negatively correlated with the targeting level of all siRNAs (Fig. 5A,B). This pattern seems reasonable if one assumes that older elements are deeply silenced due to heavy cytosine methylation. The positive correlation between symmetric methylation and age supports this conjecture (Supplemental Fig. S13B,C). However, there is a striking turning point at 2–2.5 my at which older elements remain heavily mapped by siRNAs. We are confident that this is not an artifact of incorrect age estimation, because our methodology generates precise LTR borders (Darzentas et al. 2010). We nevertheless manually examined the LTR alignments of old elements and

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

found no obvious misalignments. Furthermore, this phenomenon is not obviously driven by phylogenetic history. Remarkably, the best explanation seems to be the fact that old Sireviruses contain many intact palindromes (Fig. 6C; Table 2). That said, the INT domain also contributes to these targeting effects, although specific areas of interest could not be identified.

These observations allow us to explore Sirevirus aging under an evolutionary perspective, where the palindromes represent a crucial parameter. The basic assumption is that they are important *cis*-regulators and, hence, major targets of silencing. Under this model, young elements have not yet accumulated many mutations, contain intact palindromes, have a large potential for transposition, and are therefore intensively targeted by siRNAs. Conversely, middle-aged elements have undergone degeneration by mutations, tend to be less fit due to loss of palindromes, and are therefore less targeted by siRNAs. Finally, old elements have escaped degradation of their palindrome region by mechanisms that remain obscure. As a result, these elements may be capable of *trans* activation and are heavily targeted by siRNAs. In addition, evidence from preliminary analysis indicates that some may also retain intact *gag* and *pol* open reading frames and, therefore, may be competent for autonomous *cis* activation despite their advanced age.

But why are the palindromes retained in old Sireviruses? Perhaps, they have simply escaped mutation by chance. Although reasonable, this scenario cannot easily explain how some elements remain highly targeted at estimated ages up to 13 my (Fig. 5A). Another hypothesis entails the existence of "zombie" TEs (Lisch 2009), that is, domesticated elements that act as memory to produce 21- to 22-nt siRNAs and guard against epigenetic loss. Zombies are capable of Pol II transcription and contain appropriate triggers for the initiation of silencing. Through sequence homology with their relatives, the host uses zombies to spread silencing across family members. Several properties of the old Sireviruses are compatible with this hypothesis. First, they contain the presumed epigenetic trigger, i.e., the hairpin-prone palindrome region. Second, they are full-length and presumably competent for *cis* or *trans* activation by Pol II. Third, most of their siRNAs "cross-talk" with young elements (Supplemental Table S6), suggesting that a small population of zombies may suffice for silencing thousands of "untamed" elements. Nevertheless, we note at least two major discrepancies with this hypothesis. First, zombies are a predicted feature of plant genomes with no empirical support. Second, young elements with intact palindromes may also contain the hpRNA triggers and could perhaps act as zombies themselves. Hence, the preservation of old, intact Sireviruses within the maize genome, as documented here, remains an intriguing mystery.

Methods

Sirevirus data

The population of 13,833 full-length maize Sireviruses was downloaded from MASiVEdb (http://databases.bat.infspire.org/masivedb/). This set was further curated to minimize the inclusion of low quality elements, while still retaining thousands of bona fide full-length Sireviruses. We first filtered out 4027 elements with more than five consecutive "N" nucleotides in their sequence, based on evidence that BLASTN hits between Sireviruses and genes often mapped precisely at the border of stretches of Ns, suggesting errors during scaffold assembly. We then filtered out elements, whose full or LTR lengths were outside typical ranges for the family (Supplemental Fig. S3). Furthermore, we used the maize TE exemplars (http://maizetedb.org), excluding the Sirevirus representatives, to screen for TE insertions within the Sirevirus sequences (BLASTN, *E*-value 1×10^{-20}) and filter out potentially hybrid elements. The final set consisted of 6456 elements (see Supplemental Material). To estimate the age of each element, we aligned its LTR pair using MAFFT with default parameters (Katoh and Standley 2013), which also simultaneously produced data on indels and substitutions. We then applied the LTR retrotransposon age formula with a substitution rate of 1.3×10^{-8} mutations per site per year (Ma and Bennetzen 2004).

Mapping of siRNA data

We used published short read data from leaf (GSM1342517), tassel (GSM448857), and immature ear wild-type (GSM306487) and *mop1* (GSM306488) libraries. Adapters were trimmed using Trimmomatic, and low quality nucleotides were removed using the FASTX toolkit until every read had three or more consecutive nucleotides with a quality score of 20 or more at the 3′-end. Reads of length 21, 22, and 24 nt were filtered to eliminate tRNAs (http://gtrnadb.ucsc.edu/), rRNAs and snoRNAs (http://rfam.sanger.ac.uk/), and miRNAs (http://www.mirbase.org/). The remaining siRNAs were mapped to the maize B73 genome (RefGen_V2) using BWA with default settings and no mismatches (Li and Durbin 2010) and classified as uniquely (U_siRNAs) or multiply mapped (M_siRNAs).

Normalization and measurement of siRNA data

The extent to which siRNAs can direct methylation to multiple locations is unclear, but evidence suggests that M_siRNAs account for a considerable subset of RdDM (Lister et al. 2008). Previous studies have followed various approaches for addressing the effect of M_siRNAs, which can be more numerous in TE-rich genomes like maize, compared to species like *Arabidopsis* in which most siRNAs map uniquely to TEs (Hollister et al. 2011). For example, some studies used only U_siRNAs, but mapping was conducted on TE exemplars (Gent et al. 2013; Regulski et al. 2013). In contrast, other studies randomly allocated M_siRNAs to a single genomic locus (Wang et al. 2009; He et al. 2013) or weighted each M_siRNA by its number of mapping locations (Barber et al. 2012; Diez et al. 2014). Recently, tools have been developed that appoint M_siRNAs to single loci based on the densities of U_siRNAs among all possible options (Axtell 2013b). For Sireviruses, however, the dearth of U_siRNAs (Supplemental Table S1) and the high number of locations for M_siRNAs hinder the implementation of such tools.

We therefore analyzed both U_siRNAs and M_siRNAs and used three measures. The first counted the number of distinct siRNA sequences (siRNA "species") or the number of siRNA reads (siRNA "expression") that mapped to an individual nucleotide of a TE. Mapping events were not corrected for M_siRNAs, so that each event had a value of 1.0. In practice, each nucleotide was put into a bin to allow calculation of averages and variances by location across individual TEs. The second counted siRNA species or expression in the same way but corrected for the number of mapping locations, i.e., if a siRNA species or a siRNA read mapped to x locations, each specific mapping event was weighted by x and counted as $1/x$. Finally, we also counted the total number of siRNA species that mapped to a locus. This metric is unaffected by normalization, because it simply shows how many different siRNA species (i.e., the diversity of siRNA species) map to a given locus. This approach was also used to compare between different

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

sets of loci or elements (e.g., cross-mapping between palindromes or age groups).

Mapping of methylation data

Methylation BS-seq data from unfertilized ears (SRA050144) were mapped to the maize B73 genome (RefGen_V2) as previously described (Takuno and Gaut 2013). In brief, BS Seeker (Chen et al. 2010) was used with default settings, allowing no mismatches and retaining only uniquely mapped reads. We applied a binomial method at $P < 0.01$ to classify each cytosine as methylated or unmethylated (Lister et al. 2008) and then measured methylation levels separately in the CG, CHG, and CHH contexts. The methodology and reasoning for the methylation coverage cutoffs used in this study are included as Supplemental Material.

Bioinformatics and statistics

Most bioinformatics analyses were performed using custom scripts based on the Perl programming language (available as Supplemental Material). Their main functionality involved parsing Sirevirus information from MASiVEdb, parsing siRNA and methylation mapping files, and reporting outputs for individual TEs or regions for use in downstream analyses. All statistical tests were performed in R (R Core Team 2015).

Acknowledgments

We thank Adam Eyre-Walker for helpful discussions on the manuscript, Kamila Reblova for advice on RNA folding, and Pavlos Pavlidis for help with phylogenetic analyses. A.B. is supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number PIEF-GA-2012-329033; N.D. and V.B. by CEITEC MU (CZ.1.05/1.1.00/02.0068), SuPReMMe (CZ.1.07/2.3.00/20.0045), SYLICA (FP7-REGPOT-2011-1), ESLHO::EuroClonality (http://www.euroclonality.org), 7th Framework Programme (NGS-PTL/2012-2015/no.306242), and Czech Ministry of Education, Youth and Sports (2013-2015, number 7E13008). B.S.G. is supported by National Science Foundation (NSF) grant IOS-1542703. Computational resources were provided by the MetaCentrum under the program LM2010005 and the CERIT-SC under the program Centre CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, Registration number CZ.1.05/3.2.00/08.0144.

References

Araujo PG, Casacuberta JM, Costa APP, Hashimoto RY, Grandbastien MA, Van Sluys MA. 2001. Retrolyc1 subfamilies defined by different U3 LTR regulatory regions in the *Lycopersicon* genus. *Mol Genet Genomics* **266:** 35–41.

Axtell MJ. 2013a. Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol* **64:** 137–159.

Axtell MJ. 2013b. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* **19:** 740–751.

Barber WT, Zhang W, Win H, Varala KK, Dorweiler JE, Hudson ME, Moose SP. 2012. Repeat associated small RNAs vary among parents and following hybridization in maize. *Proc Natl Acad Sci* **109:** 10444–10449.

Bologna NG, Voinnet O. 2014. The diversity, biogenesis, and activities of endogenous silencing small RNAs in *Arabidopsis*. *Annu Rev Plant Biol* **65:** 473–503.

Bousios A, Darzentas N. 2013. Sirevirus LTR retrotransposons: phylogenetic misconceptions in the plant world. *Mob DNA* **4:** 9.

Bousios A, Darzentas N, Tsaftaris A, Pearce SR. 2010. Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants? *BMC Genomics* **11:** 89.

Bousios A, Kourmpetis YA, Pavlidis P, Minga E, Tsaftaris A, Darzentas N. 2012a. The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story. *Plant J* **69:** 475–488.

Bousios A, Minga E, Kalitsou N, Pantermali M, Tsaballa A, Darzentas N. 2012b. MASiVEdb: the Sirevirus Plant Retrotransposon Database. *BMC Genomics* **13:** 158.

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

Chen PY, Cokus SJ, Pellegrini M. 2010. BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* **11:** 203.

Creasey KM, Zhai J, Borges F, Van Ex F, Regulski M, Meyers BC, Martienssen RA. 2014. miRNAs trigger widespread epigenetically activated siRNAs from transposons in *Arabidopsis*. *Nature* **508:** 411–415.

Darzentas N, Bousios A, Apostolidou V, Tsaftaris AS. 2010. MASiVE: Mapping and Analysis of SireVirus Elements in plant genome sequences. *Bioinformatics* **26:** 2452–2454.

Diez CM, Meca E, Tenaillon MI, Gaut BS. 2014. Three groups of transposable elements with contrasting copy number dynamics and host responses in the maize (*Zea mays* ssp. *mays*) genome. *PLoS Genet* **10:** e1004298.

Du J, Tian Z, Bowen NJ, Schmutz J, Shoemaker RC, Ma J. 2010. Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through LTR swapping in soybean. *Plant Cell* **22:** 48–61.

Feng SH, Cokus SJ, Zhang XY, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci* **107:** 8689–8694.

Fultz D, Choudury SG, Slotkin RK. 2015. Silencing of active transposable elements in plants. *Curr Opin Plant Biol* **27:** 67–76.

Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, Dawe RK. 2013. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res* **23:** 628–637.

Grandbastien MA. 2015. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim Biophys Acta* **1849:** 403–416.

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA Websuite. *Nucleic Acids Res* **36:** W70–W74.

He G, Chen B, Wang X, Li X, Li J, He H, Yang M, Lu L, Qi Y, Wang X, et al. 2013. Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. *Genome Biol* **14:** R57.

Henderson IR, Zhang X, Lu C, Johnson L, Meyers BC, Green PJ, Jacobsen SE. 2006. Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nat Genet* **38:** 721–725.

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

Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* **472:** 115–159.

Johnson LM, Du JM, Hale CJ, Bischof S, Feng SH, Chodavarapu RK, Zhong XH, Marson G, Pellegrini M, Segal DJ, et al. 2014. SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature* **507:** 124–128.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30:** 772–780.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645.

Law JA, Du JM, Hale CJ, Feng SH, Krajewski K, Palanca AMS, Strahl BD, Patel DJ, Jacobsen SE. 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498:** 385–389.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26:** 589–595.

Li H, Freeling M, Lisch D. 2010. Epigenetic reprogramming during vegetative phase change in maize. *Proc Natl Acad Sci* **107:** 22184–22189.

Lisch D. 2009. Epigenetic regulation of transposable elements in plants. *Annu Rev Plant Biol*, **66:** 43–66.

Lisch D, Slotkin RK. 2011. Strategies for silencing and escape: the ancient struggle between transposable elements and their hosts. In *International review of cell and molecular biology* (ed. Jeon KW), Vol. 292, pp. 119–152. Elsevier Academic Press, San Diego, CA.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133:** 523–536.

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

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

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O. 2013. Reconstructing *de novo* silencing of an active plant retrotransposon. *Nat Genet* **45:** 1029–1039.

Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet* **15:** 394–408.

McCue AD, Panda K, Nuthikattu S, Choudury SG, Thomas EN, Slotkin RK. 2015. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. *EMBO J* **34:** 20–35.

Mergia A, Prattlowe E, Shaw KE, Renshaw-Gegg LW, Luciw PA. 1992. *cis*-acting regulatory regions in the long terminal repeat of simian foamy virus type-1. *J Virol* **66:** 251–257.

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature* **411:** 212–214.

Nobuta K, Lu C, Shrivastava R, Pillay M, De Paoli E, Accerbi M, Arteaga-Vazquez M, Sidorenko L, Jeong DH, Yen Y, et al. 2008. Distinct size distribution of endogenous siRNAs in maize: evidence from deep sequencing in the *mop1-1* mutant. *Proc Natl Acad Sci* **105:** 14958–14963.

Panda K, Slotkin RK. 2013. Proposed mechanism for the initiation of transposable element silencing by the RDR6-directed DNA methylation pathway. *Plant Signal Behav* **8:** pii: e25206.

R Core Team. 2015. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ragupathy R, You FM, Cloutier S. 2013. Arguments for standardizing transposable element annotation in plant genomes. *Trends Plant Sci* **18:** 367–376.

Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, Llaca V, Deschamps S, Smith A, Levy D, McCombie WR, et al. 2013. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res* **23:** 1651–1662.

Sharma A, Schneider KL, Presting GG. 2008. Sustained retrotransposition is mediated by nucleotide deletions and interelement recombinations. *Proc Natl Acad Sci* **105:** 15470–15474.

Slotkin RK, Freeling M, Lisch D. 2005. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat Genet* **37:** 641–644.

Slotkin RK, Vaughn M, Borges F, Tanurdžić M, Becker JD, Feijó JA, Martienssen RA. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136:** 461–472.

Takuno S, Gaut BS. 2013. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci* **110:** 1797–1802.

Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, Ciaudo C, Cruaud C, Poulain J, Berdasco M, Fraga MF, et al. 2009. A role for RNAi in the selective correction of DNA methylation defects. *Science* **323:** 1600–1604.

Vernhettes S, Grandbastien MA, Casacuberta JM. 1998. The evolutionary analysis of the Tnt1 retrotransposon in *Nicotiana* species reveals the high variability of its regulatory sequences. *Mol Biol Evol* **15:** 827–836.

Wang X, Elling AA, Li X, Li N, Peng Z, He G, Sun H, Qi Y, Liu XS, Deng XW. 2009. Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell* **21:** 1053–1069.

West PT, Li Q, Ji L, Eichten SR, Song J, Vaughn MW, Schmitz RJ, Springer NM. 2014. Genomic distribution of H3K9me2 and DNA methylation in a maize genome. *PLoS One* **9:** e105267.

Wicker T, Buchmann JP, Keller B. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res* **20:** 1229–1237.

Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. 2013. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153:** 193–205.

Zhang L, Chia JM, Kumari S, Stein JC, Liu Z, Narechania A, Maher CA, Guill K, McMullen MD, Ware D. 2009. A genome-wide characterization of microRNA genes in maize. *PLoS Genet* **5:** e1000716.

Received April 29, 2015; accepted in revised form December 1, 2015.