Gene capture by transposable elements leads to epigenetic conflict

Aline Muyle[1], Danelle Seymour[1,2], Nikos Darzentas[3], Brandon S. Gaut[1], Alexandros Bousios[4]

[1] Department of Ecology and Evolutionary Biology, UC Irvine, Irvine, CA 92697, USA, [2] Department of Botany and Plant Sciences, UC Riverside, Riverside, CA 92521, USA, [3] Central European Institute of Technology, Masaryk University, Brno, Czech Republic, [4] School of Life Sciences, University of Sussex, Brighton, UK

Author for Correspondence: Alexandros Bousios, alexandros.bousios@gmail.com; Brandon S. Gaut, bgaut@uci.edu

ABSTRACT Plant transposable elements (TEs) regularly capture fragments of host genes. When the host employs siRNAs to silence these TEs, siRNAs homologous to the captured regions may target both the TEs and the genes, potentially leading to their silencing. This epigenetic cross-talk establishes an intragenomic conflict: silencing the TEs comes with the potential cost of silencing the genes. If the genes are important, however, natural selection will act to maintain function by moderating the silencing response. Such moderation may advantage the TEs. Here, we examined the potential for these epigenetic conflicts by focusing on three TE families in maize - Helitrons, Pack-MULEs and Sirevirus LTR retrotransposons. We documented 1,508 TEs with fragments captured from 2,019 donor genes and characterized the epigenetic profiles of both. Consistent with epigenetic conflict, donor genes mapped more siRNAs and were more methylated than 'free' genes that had no evidence of exon capture. However, these patterns differed between syntelog vs. transposed donor genes. Syntelog genes appeared to maintain function, consistent with moderation of the epigenetic response for important genes before reaching a deleterious threshold, while transposed genes bore the signature of silencing and potential pseudogenization. Intriguingly, transposed genes were overrepresented among donor genes, suggesting a link between capture and gene movement. We also investigated the potential for TEs to gain an advantage. TEs with captured fragments were older, mapped fewer siRNAs and had lower levels of methylation than 'free' TEs without gene fragments, but they showed no obvious evidence of increased copy numbers. Altogether, our results demonstrate that TE capture triggers an epigenetic conflict when genes are important, contrasting the loss of function for genes that are not under strong selective constraint. The evidence for an advantage to TEs is currently less obvious.

KEYWORDS transposable elements; intragenomic conflict; gene capture; epigenetic silencing; methylation; siRNAs; maize

Introduction

Transposable elements (TEs) constitute the majority of plant genomes[1], and are major drivers of both genomic and phenotypic evolution[2]. Most TEs are silenced under normal conditions by host epigenetic mechanisms that rely on small interfering RNAs (siRNAs). These siRNAs act against homologous sequences to modify the activity of TEs either before or after transcription. To limit transcription, siRNAs prime the RNA-directed DNA methylation (RdDM) machinery, which in turn guides the deposition of cytosine methylation and heterochromatic histone marks, epigenetic modifications that can be maintained through cell division[3,4]. Silenced TEs are usually heavily methylated in the CG, CHG and CHH contexts (H = A, C, or T) and associated with a closed heterochromatic state[5]. These chromatin characteristics can influence the function and expression of genes, especially when TEs and genes reside in close proximity. For example, methylated and siRNA-targeted TEs are associated with altered expression of neighbouring genes[6-9] and, as a result, are subject to stronger purifying selection compared to unsilenced TEs or TEs far from genes[7,10].

In contrast to the epigenetic effects of TEs near genes, much less is known about epigenetic interactions between TEs and genes over long distances, particularly through the *trans*-activity of siRNAs[11]. For siRNAs to mediate long distance interactions, there must be sequence similarity between genes and TEs, so that siRNAs are homologous to both. The requirement of sequence similarity can be satisfied by varied evolutionary scenarios, such as the exaptation of portions of TEs into coding genes[12], but it is especially relevant in the phenomenon of gene capture by TEs. Gene capture has been investigated widely in both animals and plants[13-16]. Within plant genomes, capture has been best characterized for Helitrons and Pack-MULE DNA transposons, which together have captured thousands of gene fragments[17-19]. Capture is common enough that a single TE often contains fragments of multiple host genes from unlinked genomic locations[20,21]. Although it is clear that gene capture is common, the mechanisms remain uncertain. However, evidence suggests that capture can occur through both DNA and RNA-mediated processes[14,20,22], and several mechanisms have been proposed[23-25].

The evolutionary consequences of gene capture are not well characterized either. One potential consequence is that the shuffling and rejoining of coding information within a TE leads to the emergence of novel genes[21,26]. Consistent with this conjecture, a substantial proportion of TE-captured gene sequences are expressed[22,27-29], a subset of those are translated[29,30], and few exhibit signatures of selective constraint[19,30-32]. Another distinct possibility is that gene capture is a neutral mutational process caused by inexact TE reproduction with few downstream evolutionary ramifications.

Finally, gene capture may establish evolutionary conflicts between TEs and genes. Lisch[33] argues that it is in a TE's evolutionary interest to blur the line between host and parasite "by combining both transposon and host sequences, …, to increase the cost of efficiently silencing those transposons". This argument suggests a model of genomic conflict in which a TE captures a fragment from a gene, and the host mounts an siRNA-mediated response against the TE. Because the siRNAs from the captured fragment within the TE can also target the captured region of the 'donor' gene (i.e., the gene from which the fragment has been captured), the host response to the TE can simultaneously act in *trans* against the donor gene. Under this scenario, transcriptional silencing of the TE may have collateral effects on the donor gene, including targeting by siRNAs that lead to DNA methylation and subsequent silencing (Figure 1a). If the donor gene has an important function, however, then selection is likely to limit potential silencing effects on this gene. This creates an intragenomic conflict, whereby the advantage of silencing the TE is balanced by potential damage to donor gene function. Conversely, selection to moderate the host response potentially advantages the TE with the captured gene fragment. Importantly, this conflict model makes testable predictions: *i*) donor genes bear the signature of *trans*-epigenetic effects, including increased siRNA targeting and consequent methylation, *ii*) selection may limit these *trans*-epigenetic effects for important genes, and *iii*) TEs benefit from capture via decreased host response.

The possibility of epigenetic links between TEs and donor genes has been discussed previously[21], but to our knowledge only one study in 2009 has examined how often siRNAs map to both donor genes and to their captured fragments[30]. This study focused on Pack-MULEs in rice (*Oryza sativa*) and found siRNAs that map to both TEs and donor genes, thus providing the potential for siRNA 'cross-talk' between donor genes and captured gene fragments. The study also found that genes with cross-talk are less expressed compared to genes without any mapped siRNAs. Two recent studies[28,29] of rice Pack-MULEs extended this line of enquiry by investigating whether donor genes are methylated, which could be indicative of epigenetic effects consistent with the conflict model. They found, however, that donor genes have low methylation levels not substantially different from genes with no apparent history of capture by TEs (hereafter termed 'free' genes). These studies provide some, but limited, evidence for epigenetic conflict.

The study of Pack-MULEs in rice suffers from two potential shortcomings with respect to investigating epigenetic interactions. The first is Pack-MULEs themselves. They commonly capture genes and therefore provide a rich template for study, but often have lower methylation levels than other TE families[29,34], possibly because they preferentially insert near the 5' termini of genes[35]. This tendency may lessen the potential for intragenomic conflict with their donor genes. The second shortcoming is the small genome size of rice (390 Mb[36]). Large genomes differ from small genomes in their TE content and, as a result, their DNA methylation patterns. For example, Takuno et al. (2016)[37] showed that only 6% of genes in rice have high levels (>90%) of CG methylation compared to 24% of genes in the much larger (2,300 Mb) genome of maize. The contrast is even stronger for high (>90%) CHG methylation - 12% vs. 1% of maize and rice genes, respectively - reflecting the strong positive correlation between gene CHG methylation and genome size[37-39].

Here, we hypothesize that gene capture by TEs may have epigenetic consequences for endogenous genes in maize. To test this hypothesis, we identify capture events representing all major TE classes, i.e. Helitron rolling circle transposons, Pack-MULE Class II DNA transposons, and, for the first time to our knowledge in plants a representative of Class I retroelements, Sirevirus LTR retrotransposons[40]. Sireviruses are crucial because they compose ~20% of the maize genome[41], are targeted by large numbers of siRNAs, and are highly methylated[42]. Given sets of TEs with gene capture events, we integrate evolutionary analyses with siRNA, methylation and gene expression data to address two sets of predictions. The first set focuses on the genic viewpoint. If the conflict model holds, we predict that donor genes bear the signature of *trans*-epigenetic effects compared to free genes, but also that these epigenetic effects have minimal impact on the function of important genes. In the second set of predictions, we focus on TEs with captured gene fragments. Is there any evidence that they benefit from gene capture via decreased host response?

Results

Identifying captured gene fragments and their donor genes

We began this work by retrieving carefully annotated datasets of full-length elements for Helitrons[43], Pack-MULEs[35], and Sireviruses[44]. After further curation (see Methods), our TE dataset consisted of 11,144 full-length elements derived from 1,090 Helitrons, 248 Pack-MULEs and 9,806 Sireviruses. We then performed strict BLASTN comparisons (*E*-value cutoff of 1 x 10^{-40}) between these TEs and the exons of a curated set of 32,551 maize genes (see Methods) to identify both captured gene fragments within TEs and their donor genes. We generated hits between 1,688 TEs and 4,814 candidate donor genes, with the remaining 27,737 genes termed 'free' genes. After further curation of the BLASTN results (see Methods;

Figure S1a), we derived a final set of 2,019 donor genes captured by 1,508 TEs. Most Helitrons (938; 86%) and Pack-MULEs (196; 79%) contained gene fragments in contrast to only a small proportion of Sireviruses (374; 4%). The three families, in turn, captured 1,653, 233 and 242 genes respectively, a total that exceeds 2,019 because 100 genes were captured by more than one family (Figure S1b). Like previous studies[20,21], we found that individual elements often contained multiple independent capture events: 76% of Helitrons harbored ≥2 captured fragments, as did 63% of Pack-MULEs and 49% of Sireviruses.

Donor genes are targets of siRNAs and are highly methylated

Under our conflict model, the first prediction is that gene capture should lead to siRNA cross-talk between genes and TEs, potentially leading to increased methylation of donor genes. Accordingly, we began our study by contrasting donor vs. free genes for siRNA mapping and methylation characteristics. Throughout this study, we relied on published siRNA and bisulfite-sequencing (BS-seq) datasets, focusing on libraries from unfertilized ears[45,46], leaves of maize seedlings[47,48] and tassels[49] (see Methods). We analyzed 21nt, 22nt, and 24nt siRNAs, because these lengths are involved in TE silencing[4]. We combined data from the three siRNA lengths, because genic mapping was strongly correlated across lengths (Figure S2). For each gene, we then calculated the number of distinct siRNA sequences per kb of a locus (see Methods). Results were generally consistent among tissues; hence, we report data from ear in the main text, but provide relevant results from the other two tissues in Supplementary Information.

We first contrasted siRNA mapping profiles of the exons of donor and free genes. Overall, the difference was striking: the 2,019 donor genes mapped 2.5 times more siRNAs per kb on average than the 27,737 free genes (mean 8.97 vs. 2.94 siRNA/kb, Figure 1b, Mann-Whitney-Wilcoxon Test p<2.2e-16). This difference was due in part to the fact that most (62%) free genes did not map siRNAs compared to only 26% of donor genes. However, the difference between the two groups remained even when genes with no siRNAs were removed (Mann-Whitney-Wilcoxon Test p<2.2e-16).

In theory, differences in siRNA mapping should affect methylation patterns. We used BS-seq data to calculate the proportion of methylated cytosines in the CG, CHG and CHH contexts of exons (see Methods). Of the 2,019 donor and 27,737 free genes, 1,807 and 24,641 passed CG methylation filters (≥10 covered CG sites), representing ~89% of the genic dataset, with similar proportions retained for CHG and CHH methylation. We found that donor genes were significantly more methylated than free genes according to Mann-Whitney-Wilcoxon one-sided tests in CG (mean 50.7% vs. 24.2% respectively, p<2.2e-16), CHG (37.1% vs. 14.1%, p<2.2e-16) and CHH context (5.59% vs. 4.17%, p<2.2e-16) (Figure 1c). Overall, the trends were clear and consistent across all tissues (Figure S3): donor genes map more siRNAs and are more highly methylated.

There were also differences across TE families. Donor genes captured by Helitrons had the highest mean CG (54.5%) and CHG (39.6%) methylation level, followed by Sireviruses (52.6% and 40.5%, respectively) and then by Pack-MULEs (22.8% and 14.9%, respectively) (Table S1). This last observation is consistent with the low methylation levels of Pack-MULEs[29,34] that may hamper full understanding of their epigenetic interactions with donor genes.

Captured regions of syntenic ortholog donor genes are enriched for cross-talk siRNAs

Our results support the predictions of the conflict model by showing that donor genes are heavily enriched for both siRNA mapping and methylation levels. But there are alternative explanations, e.g. TEs may often capture genes that are already highly methylated. Moreover, the model specifically proposes that conflict arises for functional genes, but many donor genes contain high levels of CHG methylation (Figure 1c), which is a potential signature of silencing and pseudogenization. To better test the conflict model, we therefore enriched the dataset for functional genes by using synteny as an additional filter. Previous studies have documented that syntenic orthologs (hereafter termed 'syntelogs') tend to be functionally constrained and more often associated with phenotype compared to non-syntelogs[50,51]. After parsing the genic dataset by requiring synteny between maize and *Sorghum bicolor* (sorghum) (see Methods), we retrieved 951 donor and 18,293 free syntelogs and contrasted their epigenetic profiles. Although siRNA and methylation levels were overall lower for syntelogs compared to the complete genic dataset, the differences remained: syntelog donor genes mapped more siRNAs per kb across their exons than syntelog free genes (mean 6.1 vs. 1.2 siRNA/kb, Mann-Whitney-Wilcoxon Test p=2.5e-16) and had higher methylation levels (mean CG 26.7% vs. 15.5%, p<2.2e-16; CHG 9.6% vs. 4.9%, p<2.2e-16; CHH 5.3% vs. 3.6%, p=6.1e-11).

We then focused on an additional prediction of the conflict model: siRNAs should be overrepresented in the region of the gene that was captured by the TE (Figure 1a). To examine this prediction, we retrieved the total number of siRNAs for the 951 syntelog donor genes and then compared mapping between captured vs. non-captured regions of their exons. As predicted, more siRNAs mapped to the captured regions (mean 29.86 vs. 2.11 siRNA/kb, Mann-Whitney-Wilcoxon Test p<2.2e-16; Figure 2a). We also focused on cross-talk siRNAs to test whether they, too, represent an enriched fraction of the total number of siRNAs that mapped to donor genes. To do so, we first used a binomial test to compare the observed proportion of cross-talk siRNAs (cross-talk siRNAs / all siRNAs = 0.43) to the proportion of captured gene length (captured exon length / total exon length = 0.16) across all syntelog genes combined. This revealed a significant enrichment of cross-talk siRNAs in captured regions (p<2.2e-16). We then investigated each gene separately using the same

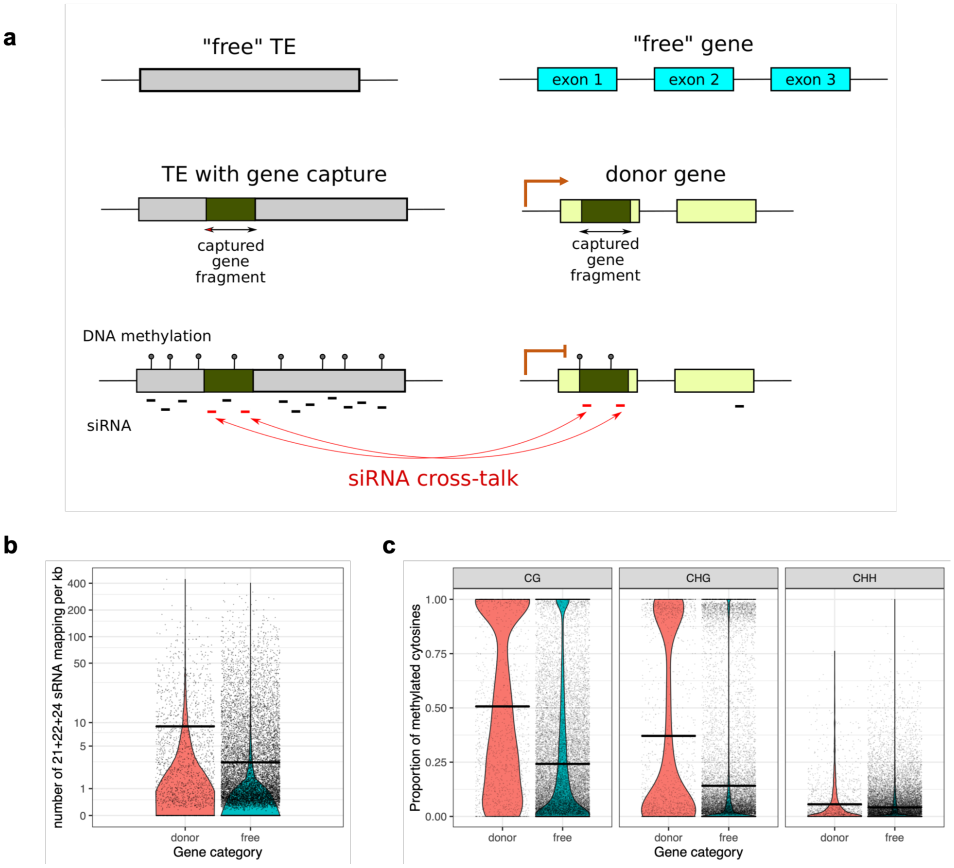

Figure 1. Epigenetic effects of capture on donor genes. (a) Schematic of a capture event by a TE and ensuing epigenetic interactions. Definitions used in the text are shown, including donor and free genes, free TEs and TEs with captured fragments, and cross-talk siRNAs that have the potential to act in *trans*. The orange arrows indicate expression. **(b)** Number of 21+22+24nt distinct siRNA sequences per kb mapping to donor (n=2,019) and free genes (n=27,737). **(c)** Proportion of methylated cytosines in CG, CHG and CHH contexts for donor (n=1,807) and free genes (n=24,641) that passed methylation coverage filters (see Methods). The horizontal black lines in (b) and (c) show the mean, and each dot is a gene. Data are from the ear tissue.

approach. Focusing on a set of 188 syntelog donor genes that had ≥10 siRNAs mapping in their exons and at least one siRNA mapping in the captured region within the TE (to allow cross-talk to occur), we found 100 genes (53.2%) had statistically higher cross-talk than expected given the length of the captured fragment, 77 genes (41%) yielded no significant difference, and, somewhat surprisingly, 11 genes (5.9%) had significantly fewer cross-talk siRNAs in the captured region (Figure 2b). Although there was variation among individual genes, altogether these results document a strong trend toward enhanced numbers of siRNAs in captured regions, which was also consistent across tissues (Figure S4).

Finally, we explored the relationship between siRNA cross-talk and time. This is probably a complex relationship, for two reasons. The first is that the initiation of an epigenetic response by the host against a new capture event may not be immediate, so that very recent capture events may not generate enough siRNAs for us to detect cross-talk. The second is that opportunities for cross-talk are finite, because the sequences of the donor gene and the captured fragment within the TE diverge over time. As they diverge, cross-talk can no longer occur because siRNAs no longer match both entities. We examined the relationship between siRNA cross-talk and time since capture, by estimating synonymous divergence (d_s) between the donor gene and the TE-captured exon as a proxy of the age of capture (see Methods). Altogether, we found a positive relationship between the number of cross-talk siRNAs and capture age, whereby syntelog donor genes with older capture events had more cross-talk siRNAs despite the increased divergence of their captured sequences over time (linear model with mixed effects across all tissues z-value=16.95, p<2e-16, marginal R-squared 0.037, see Methods) (Figure S5). When the captured fragment of an exon was part of the 5' or 3' untranslated region of a gene, we estimated non-coding divergence and found a similar effect (linear model with mixed effects across all tissues z-value=45.14, p<2e-16, marginal R-squared 0.15). Overall, we interpret these results to imply that it takes time for cross-talk to evolve after the capture event.

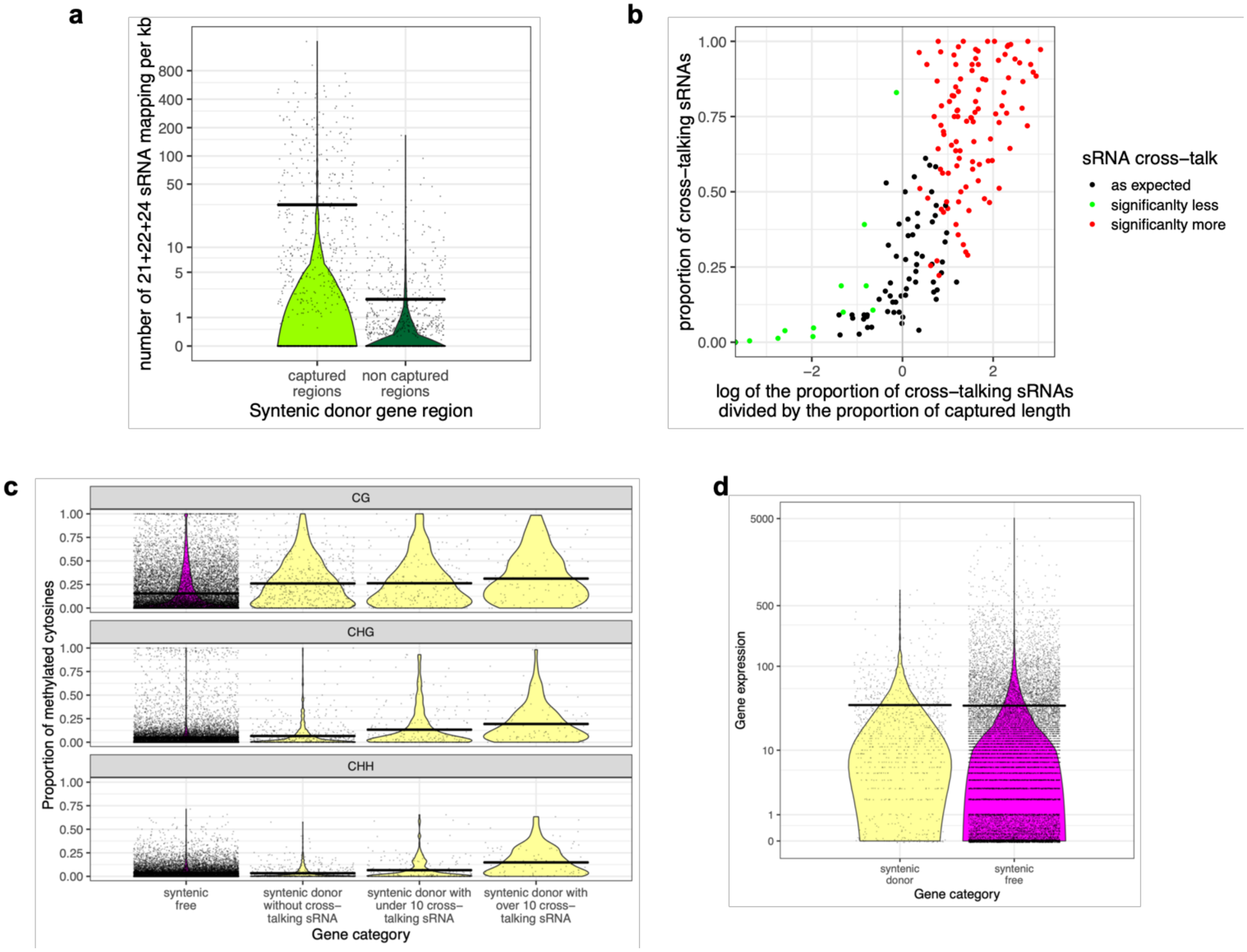

Figure 2. Epigenetic and expression profiles of syntelog donor and free genes. (a) Number of 21+22+24nt distinct siRNA sequences per kb mapping to the captured and non-captured regions of syntelog donor genes (n=951). **(b)** Proportion of cross-talk siRNAs as a function of the log of the proportion of cross-talk siRNAs divided by the proportion of capture exon length for each syntelog donor gene with sufficient siRNA numbers (n=188, see text). The color code represents the binomial test for whether cross-talk siRNAs are enriched compared to expectation (see text). **(c)** Proportion of methylated cytosines in CG, CHG and CHH contexts as a function of the number of cross-talk siRNAs in syntelog donor genes (no cross-talk siRNAs, n=620; <=10 cross-talk siRNAs, n=235; >10 cross-talk siRNAs, n=96). Syntelog free genes are used as control (n=18,293). **(d)** Gene expression of syntelog donor (n=951) and free genes (n=18,293) measured in TPM. The horizontal black lines in (a), (c) and (d) show the mean, and each dot is a gene. Data are from the ear tissue.

siRNA cross-talk affects gene methylation in trans

We have defined cross-talk siRNAs as those that map to both a gene and a TE, but it is not clear whether these siRNAs can act biologically in *trans*. However, a key prediction of the conflict model, i.e. that gene capture has the capacity to modify the epigenetic state of the donor gene, presupposes that siRNAs are *trans*-acting. Hence, we tested for potential epigenetic effects by examining the relationship between the number of cross-talk siRNAs and methylation levels of syntelog donor genes.

We first separated the 951 syntelog donor genes into three categories: those with no cross-talk siRNAs (620), those with ≤10 cross-talk siRNAs (235), and those with >10 cross-talk siRNAs (96). As a control, we also included the 18,293 syntelog free genes that passed the appropriate filters. We then compared these categories for methylation levels in ear (Figure 2c) and leaf (Figure S6) using a generalized linear model, with tissue as a random effect and the length of captured fragments as a fixed effect (see Methods). The relationship between the number of cross-talk siRNAs and methylation was significant for each cytosine context (Table

S2, *p*<2e-16), strongly suggesting that cross-talk siRNAs drive increased methylation levels of donor genes in *trans*. Although significant, however, this relationship explained only a small proportion of the total variance: altogether, capture length and the number of cross-talk siRNA explained 0.22% of variance in the CG context, 1.32% in the CHG context and 3.85% in the CHH context (Tables S2). The fact that most variation was explained for CHH methylation makes biological sense, because methylation in this context is maintained *de novo* by RdDM[4].

Expression of syntelog donor genes

Our analyses are consistent with the interpretation that cross-talk siRNAs drive, to some extent, increased methylation of donor genes in *trans*, hence setting the stage for the conflict model. The model predicts, however, that these epigenetic modifications will have minimal effects on important genes, because natural selection acts against changes that affect function. A proper test requires the ability to compare the expression of genes before and after they have been captured by TEs, but this contrast is not available. As a proxy, we instead contrasted expression of donor vs. free syntelogs, using data retrieved from the Atlas Expression database (see Methods). Consistent with the prediction of the conflict model, we did not find significantly lower levels of expression (in Transcripts per Million, TPM) in donor genes in ear (Figure 2d), leaf and ten different cell types of the maize kernel (Figure S7). In fact, generally across all tissues, we found that donor genes were expressed at significantly higher levels than free genes (for example, in ear Mann-Whitney-Wilcoxon Test p<2.2e-16) and that a lower proportion had zero expression (for example, in ear 4.3% donor vs. 13.2% free genes were not expressed).

Dramatic differences between syntelogs and recently transposed donor genes

We have thus far focused on syntelogs, because they are expected to be enriched for genes that are functional, subject to natural selection and thus susceptible to intragenomic conflict. In the absence of natural selection, however, the conflict model should not hold. In pseudogenes, for example, capture by TEs should lead to siRNA cross-talk that, in turn, should lead to high levels of methylation without the moderating effects of natural selection.

To examine this idea, we focused on a class of genes that may be facing less selection pressure than syntelogs, specifically 2,732 genes that have moved from their syntenic location in maize in relation to sorghum (hereafter termed 'transposed' genes, see Methods). We made two striking observations. First, a higher proportion of transposed genes were captured by our TEs compared to syntelogs, i.e. 442 of 2,732 (16.2%) vs. 951 of 19,244 (4.9%, Chi-squared=512.37, p<2.2e-16). Second, transposed genes were, as a group, mapped by more siRNAs, methylated at higher levels in the CG, CHG and CHH contexts, and with correspondingly lower levels of expression compared to syntelogs (Figure S8). This profile is in agreement with previous studies that showed that transposed genes have pseudogene-like characteristics[50,52,53]. But, more importantly, by repeating the analysis separately for the 442 donor transposed vs. the 2,290 free transposed genes, it became clear that genes captured by TEs really drive the differences. Donor transposed genes mapped more siRNAs than free transposed genes (mean 7.88 vs. 5.62 siRNAs/kb, Mann-Whitney-Wilcoxon Test p<2.2e-16, Figure 3a), were more methylated in CG (mean 84.8% vs. 46.9%, Mann-Whitney-Wilcoxon Test p<2.2e-16), CHG (mean 77.9% vs. 41.4%, p<2.2e-16), and CHH contexts (mean 3.2% vs. 2.8%, p=2.918e-14) (Figure 3b), and were also less expressed (mean TPM of 8.24 vs. 22.09, Mann-Whitney-Wilcoxon Test p=3.269e-05, Figure 3c). Hence, donor transposed genes are clearly distinguished from free transposed genes and exhibit a signal consistent with run-away epigenetic interactions with TEs that is not moderated by functional constraints and, hence, their expression is dramatically reduced. These patterns were consistent across all tissues examined (Figure S9).

Potential advantages for TEs to capture gene fragments

Besides the impact on genes, the conflict model also predicts that TEs with captured gene fragments gain an advantage, due to a moderation of the host response. To explore this possibility, we focused on 860 TEs that captured fragments from syntelogs and contrasted them to 9,456 'free' TEs that had no BLASTN hit to the gene dataset. Given these two groups, we considered four potential measures of advantage for TEs with captured fragments: *i*) they may be retained within the genome for longer lengths of time, *ii*) they may be targeted by fewer siRNAs, *iii*) they may have lower levels of methylation, and *iv*) they may proliferate more often, leading to higher copy numbers.

To test the first idea, we used age estimates from terminal branch lengths of TE phylogenetic trees generated by Stitzer at al. (2019)[34]. We found that TEs with captured fragments are older than free TEs (mean of 0.134 vs. 0.066 million years, one-sided Mann-Whitney-Wilcoxon Test p<2.2e-16; Figure 4a), suggesting that they have remained intact within the genome for longer periods. We next examined siRNA mapping using a linear model across all tissues (as a random factor) and after removing the captured regions from TEs with captured fragments. The analysis revealed that TEs with captured fragments had significantly less siRNA mapping compared to free TEs (for example ear mean 77.89 vs. 208.90 siRNA/kb, contrast z-ratio=-103.82, p<0.0001, marginal R-squared=31.03%, see Methods) (Figure 4b, Figure S10a); this result remained significant after also including TE age in the model (Table S3).

We then asked whether siRNA differences translated to methylation differences. We found that TEs with captured fragments were less methylated compared to free TEs in both the CG (ear mean 95.9% vs. 98.5%) and CHG (90.3% vs. 91.4%) contexts (Figure 4c, Figure S10b). These differences were small but significant in a linear model across tissues

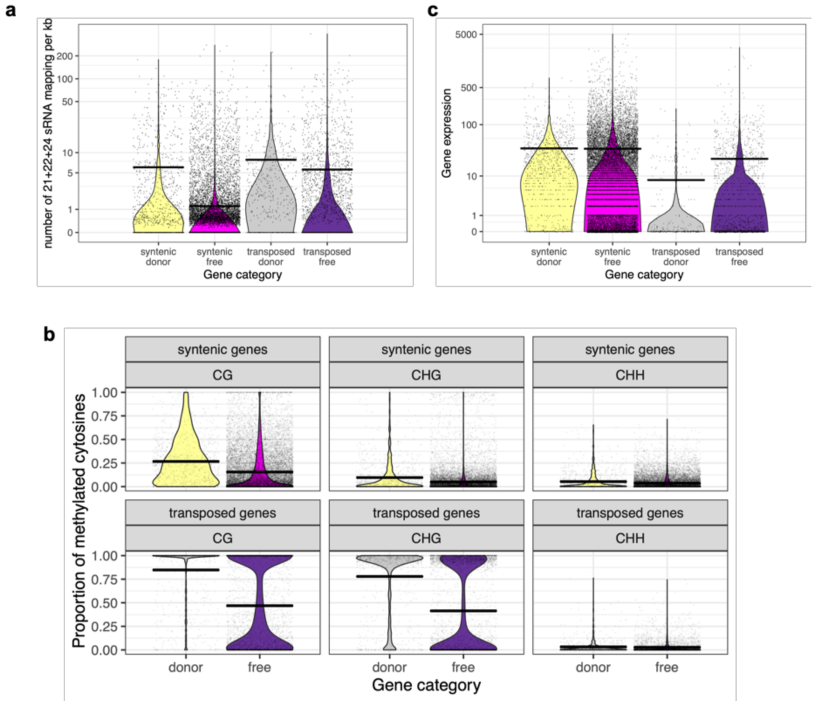

Figure 3. Epigenetic and expression profiles of donor and free genes based on their syntelog or transposed status. (a) Number of 21+22+24nt distinct siRNA sequences per kb. **(b)** Proportion of methylated cytosines in CG, CHG and CHH contexts. **(c)** Gene expression measured in TPM. The four categories of genes in all plots are: syntelog donor (n=951), syntelog free (n=18,293), transposed donor (442), and transposed free (n=2,290) genes. The horizontal black lines show the mean, and each dot is a gene. Data are from the ear tissue.

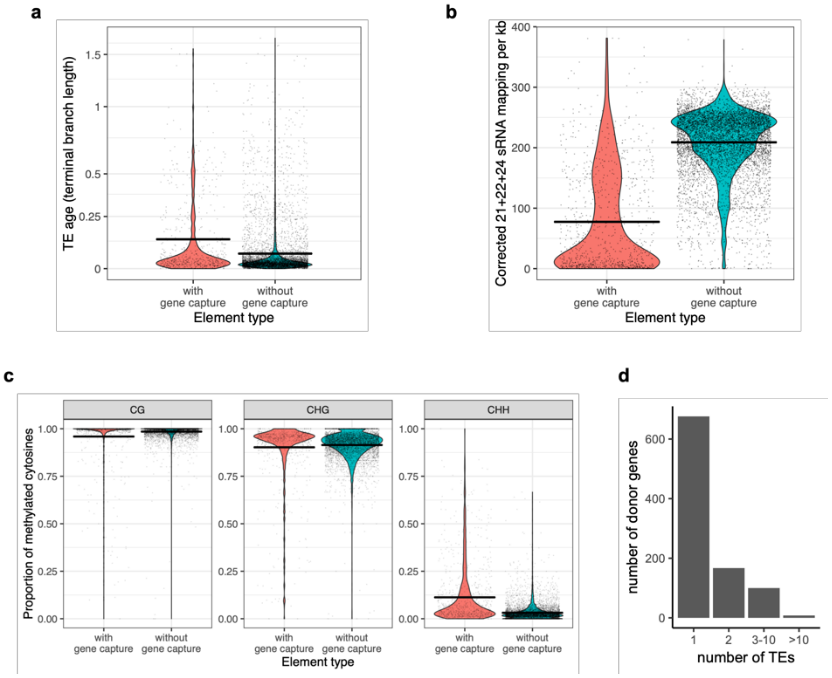

Figure 4. Characteristics of TEs with and without gene capture. (a) TE insertion age (in million years). **(b)** Number of 21+22+24nt distinct siRNA sequences per kb mapping to TEs. This was computed after removing captured regions from TEs, but results were qualitatively identical when they were included. **(c)** Proportion of methylated cytosines in CG, CHG and CHH contexts of TEs. **(d)** Histogram of the number of times that a donor syntelog gene was found within a TE, showing that most capture events were detected in only one TE. In all plots, TEs with syntelog gene capture (n=860) are compared to TEs without gene capture (n=9,456). The horizontal black lines in (a), (b) and (c) show the mean, and each dot is a TE. Data are from the ear tissue.

(as a random factor) and also held after controlling for TE age (CG: contrast t-value=74.21, p<2e-16; CHG contrast t-value=18.59, p<2e-16; Table S4). However, TEs with captured fragments had significantly more CHH methylation compared to free TEs (11.2% vs. 3.1%, contrast t-value=-62.65, p<2e-16, Figure 4c, Figure S10b, Table S4). Finally, to address the issue of copy number, we assessed how many times a syntelog donor gene was found within multiple TEs and found that the majority (71%) had been captured by a single element (Figure 4d). This suggests that true capture events were numerous, but TEs do not subsequently amplify in large numbers.

We also repeated these analyses at the family level, and only Sireviruses generated significant trends for age and siRNA mapping (Figure S11, Table S5 and S6). The lack of significance for Helitrons and Pack-MULEs may reflect the fact that very few of these elements lack captured gene fragments (Figure S1a).

Discussion

Intragenomic conflicts are a common feature of genome evolution[54]. TE conflicts arise from the fact that their proliferation often has a deleterious effect on host fitness. Here, we have studied a unique feature of intragenomic conflict that arises from the capture of genes by TEs. We began our study by formalizing a model that was initially suggested by Lisch in 2009[33]. This model argues that gene capture can have a beneficial effect on TEs because they become 'camouflaged' and, hence, less apt to be silenced by the host epigenetic machinery. They are less apt to be silenced because of an epigenetic conflict: by silencing these TEs, there is a chance of also mistakenly silencing the host gene.

The case for conflict: donor genes

The model makes several concrete predictions about donor genes and the TEs that capture them. For donor genes, it predicts higher siRNA mapping relative to genes with no history of capture, and that this should be especially true for cross-talk siRNAs. If cross-talk siRNAs act in *trans*, the model also predicts altered methylation dynamics for donor genes. Finally, the epigenetic modification of donor genes eventually reaches some threshold that is likely to affect gene function. It is the existence of this threshold that drives conflict. That is, when the silencing response becomes deleterious, then natural selection favors an amelioration of the silencing response.

What is the evidence to support this model? Based on our dataset of syntelogs, which are enriched for functional genes[50,51] to which the conflict model applies, we found that donor genes map more siRNAs (Figure 3a) and are more methylated (Figure 3b) than free genes, and that these siRNAs are enriched within captured genic regions (Figure 2a,b). There is also a clear and significant relationship between cross-talk siRNAs and gene methylation (Figure 2c). We emphasize that these results are consistent across multiple tissues. However, we also recognize that the magnitude of effects is small; for example, the number of cross-talk siRNAs explains <4% of methylation variation among syntelogs. The low amount of variance undoubtedly reflects that many other features associate with gene methylation, including gene length, exon number, gene expression and nucleosome occupancy[55]. Nonetheless, the clear positive relationship between cross-talk siRNAs and methylation levels (Figure 2c, Figure S6) indicates directionality, as does the relationship between siRNAs and the time since capture (Figure S5).

The conflict model predicts that the epigenetic interactions should not proceed to the extent that gene function is altered, because natural selection will conserve the function of important genes. We used a coarse approach to assess function: we compared gene expression between donor and free syntelogs, expecting to find no evidence of reduction in expression of donor genes. This was indeed the case across all tissues examined (Figure 2d, Figure S7); however, the comparison also revealed that donor genes are more highly expressed than free genes. We propose that this difference likely reflects biases in capture events. This hypothesis presupposes that TEs are better able to capture highly expressed genes in open chromatin, and it conforms to the integration preferences of several TE families across plants and animals for genic regions[56]. Intriguingly, and unlike other abundant LTR retrotransposon families in maize[57], Sireviruses also favor integration in gene-rich regions[41].

Although we interpret the evidence for epigenetic effects of gene capture on donor genes to be relatively strong, we recognize caveats to our analyses. For example, our set of donor genes does not represent all capture events throughout the history of the maize genome, for two reasons. The first is that we did not examine all TE families but instead relied on highly-curated sets of Helitrons and Pack-MULEs - the two best studied TE families for gene capture - and Sirevirus LTR retrotransposons that comprise a fifth of the maize genome with a significant impact on its evolution[41]. The second is that we used criteria to identify captured events that were stricter than previous studies[19,20,30,35,58-60], a conservative approach that favors specificity over sensitivity (see Methods). That is, we know that the set of free genes likely contains several undetected capture events, leading to a systematic underestimation of the differences between donor and free genes. Yet, epigenetic differences between these groups remain detectable.

The case for conflict: TEs

The conflict model also predicts that TEs with captured gene fragments gain an advantage. It is an open question as to how to measure such an advantage, and so we investigated several potential measures. We asked, for example, whether TEs with gene fragments have a tendency for camouflage, as measured by siRNAs mapping. Consistent with the conflict model, TEs with captured fragments map fewer siRNAs than free TEs, even when the captured region was masked or when TE age was taken into consideration (Figure 4b). One caveat to this result is that we likely underestimated the size of the captured region; this could bias analyses if captured regions tend to map fewer siRNAs than TE-specific regions.

Fewer siRNAs could lead to lower methylation levels, which is another possible indicator of camouflage. We found that TEs with capture events tend to have lower CG and CHG methylation, even after accounting for TE age (Figure 4c). However, this is a nuanced result, for two reasons. First, we find that differences are not large in magnitude: for CG and CHG levels, there are ~1% to 2% differences between the two TE sets, and both have >90% methylation on average. At these high levels of methylation, any TE is probably effectively silenced. Second, TEs with capture events have ~3-fold higher levels of CHH methylation (Figure 4c), which is hard to reconcile with the lower number of matching siRNAs. The cause of this CHH difference remains elusive.

If gene fragments provide camouflage for TEs, one reasonable prediction is that they will exist within the genome for longer periods of time than free TEs. Using the inferred age of elements from a previous study[34], we found that this is indeed the case (Figure 4a). The above differences were principally caused by Sirevirus elements, perhaps in part reflecting their higher proportion of free TEs. Altogether, summing across information on siRNAs, methylation and age, we consider the case for TE advantage to be tantalizing and perhaps correct, but not yet fully convincing especially considering that we also did not find evidence for increased amplification rates for these TEs (Figure 4d).

Finally, there is another interesting scenario to consider, which is an alternative to the conflict model. Given our analyses of syntelog donor genes (see above), there seems to be little doubt that gene capture by TEs leads to epigenetic interactions that affect genes. Moreover, given the extensive evolutionary literature on the conservation of relative levels of gene body methylation between species[37,38,61,62], it is reasonable to assert that methylation levels of functional genes affect some aspects of function[63] and are thus directly or indirectly visible to natural selection. We have proposed that natural selection ameliorates the host epigenetic response, leading to conflict between TEs and donor genes. Another possibility, however, is that the epigenetic response against TEs continues unabated - leading to no advantage for TEs - but the epigenetic effects on donor genes are moderated by other mechanisms, such as active CHG demethylation[64].

Exceptions that prove the rule: transposed genes

The conflict model, if true, is bound to vary substantially among genes. Some genes are under strong selection for function, leading to the potential for strong conflict, while others are functionally redundant and may be silenced without substantial costs to host fitness. Our focus on transposed donor genes may provide insights into the latter scenario. These genes are less likely to be annotated with a specific function; 64% of donor transposed genes have been assigned a specific function in v4 compared to 91.6% of donor syntelogs. These genes also follow the hallmarks of run-away epigenetic interactions with TEs, including high levels of both siRNA mapping and methylation (Figure 3a,b). Nearly 80% of transposed donor genes have >90% CG and CHG methylation, a pattern consistent with complete gene silencing that is supported by their low average expression levels (Figure 3c). These genes may be the exceptions that prove the rule - i.e., they illustrate the run-away effects of epigenetic interactions in the absence of strong selection for function. It is important to emphasize that these epigenetic patterns are not a necessary feature of transposed genes, because many of the free transposed genes have low CG and CHG methylation levels (Figure 3b) and high expression levels (Figure 3c). It is tempting to suggest that some of the transposed free genes with epigenetic characteristics similar to transposed donor genes may in reality represent false negatives for which we failed to identify their capture by a TE.

Another interesting facet of transposed genes is that they are captured by TEs more frequently than syntelogs; we detected 16.2% of transposed genes to be donors, compared to 4.9% of syntelogs. Previous work has shown that TEs contribute to modifications of synteny[65], suggesting that TE capture can trigger gene movement. It is therefore possible that the categories of ‘donor’ and ‘transposed’ may be linked mechanistically.

Concluding remarks

To summarize, our study provides evidence of enhanced siRNA targeting and methylation of donor genes, particularly in their captured regions, with tantalizing hints of advantages to the TEs that capture them. We propose that these epigenetic interactions trigger conflict between TEs and the host genome when genes are important – as is often the case for syntenic orthologs – and, hence, their effects are moderated by functional constraints to avoid gene silencing (Figure 5a). In contrast, less important genes, such as those that have moved from their syntenic loci to new positions, may be the exceptions that prove the rule. We posit that they demonstrate the outcome of epigenetic interactions without the moderating force of natural selection, leading to high levels of methylation and silencing, hence potentially representing a route towards pseudogenization (Figure 5b).

We propose that these conflicts apply generally to plant genomes, for which gene capture by TEs appears to be a common occurrence[14,17,20,30,35]. We suspect, however, that conflict is more pervasive for species with larger (e.g. maize) than smaller (e.g. rice) genomes, because methylation levels and TE load are generally higher in large genome species[37,38]. Finally, we note that the strength of conflict may vary by the type of TE that performs the capture, as suggested by our analysis of Pack-MULEs.

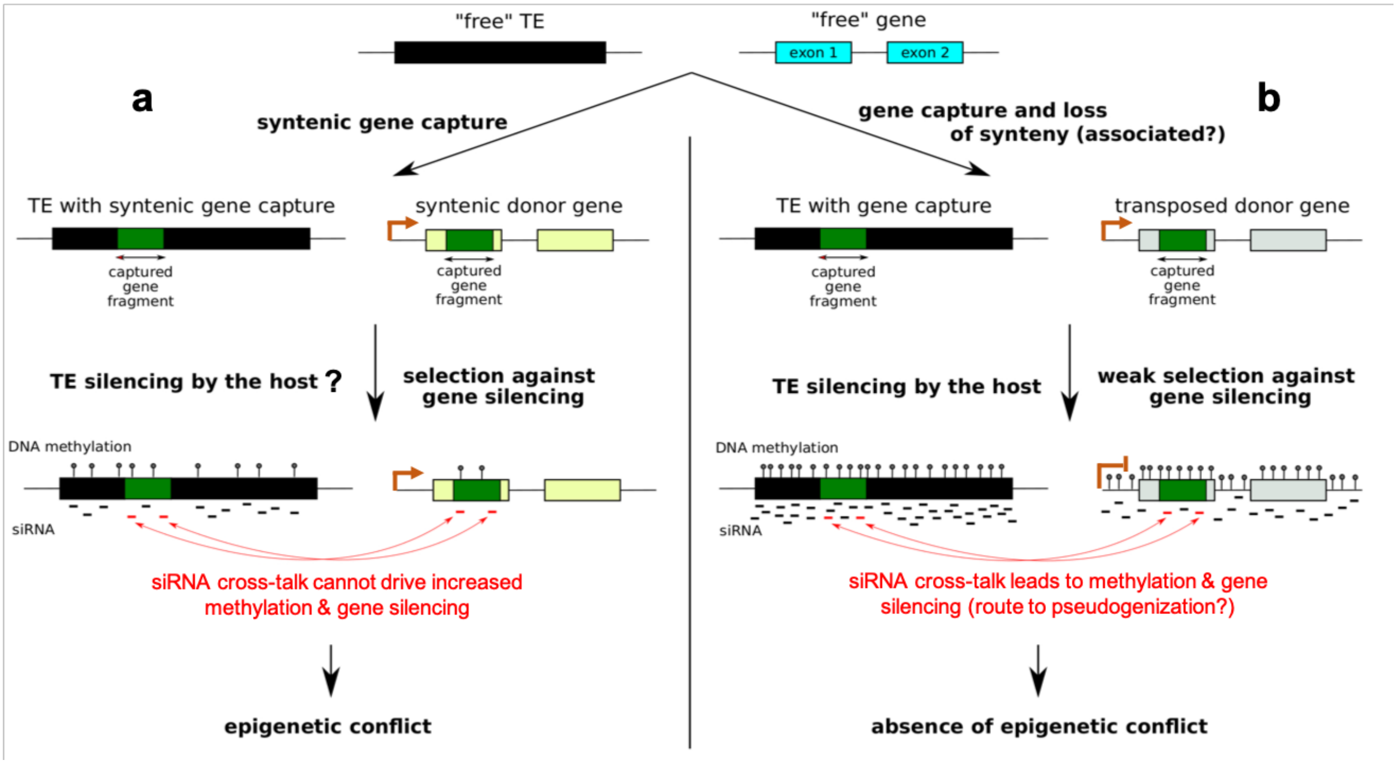
"free" TE
"free" gene
exon 1
exon 2
a
syntenic gene capture
gene capture and loss of synteny (associated?)
b
TE with syntenic gene capture
syntenic donor gene
TE with gene capture
transposed donor gene
captured gene fragment
TE silencing by the host ?
selection against gene silencing
TE silencing by the host
weak selection against gene silencing
DNA methylation
siRNA
siRNA cross-talk cannot drive increased methylation & gene silencing
siRNA cross-talk leads to methylation & gene silencing (route to pseudogenization?)
epigenetic conflict
absence of epigenetic conflict

Figure 5. The conflict model of gene capture. Under the conflict model, when TEs capture fragment of donor genes, siRNAs derived by the TEs may act *in trans* through sequence homology to accidentally mediate an epigenetic response against the donor genes, which, in turn, may lead to increased methylation and reduced expression. (**a**) The conflict comes from evolutionary pressure to silence TEs without simultaneously silencing functionally important genes – syntelogs in our example. As a result, epigenetic effects on these genes are moderated by natural selection, siRNA and methylation levels remain low (but higher than syntelog free genes, an observation that indicates the existence of a threshold), and expression is not affected; meanwhile, TEs may advantage from this moderation, although this is only partially supported based on our data (but see text for an alternative explanation). **(b)** In contrast, for genes that are not under strong selective constraint, siRNA mapping and methylation can increase in the absence of conflict, leading to loss of expression and function. This profile is characteristic of genes that have moved from their syntenic loci. In fact, these genes are overrepresented among donor genes, suggesting that capture may trigger movement. This process may therefore represent a route towards pseudogenization.

Materials & Methods

TE and gene datasets

Identifying true gene capture events by TEs is a challenging task. It is therefore important to use high-quality gene and TE datasets, for two reasons. First, some genes may be TEs that were misannotated by gene prediction algorithms. Second, some TEs may be misannotated as full-length, while in reality being fragmented elements or a mosaic of different TEs, leading to false positive capture events and erroneous sequence analysis.

In this work, we opted to favor specificity and apply strict criteria for the generation of the two datasets. For TEs, we utilized three published datasets that we know to be carefully curated, representing full-length Helitrons, Pack-MULEs and Sireviruses. For Helitrons, we downloaded the coordinates on the B73 RefGen_v2 genome of 1,351 elements from Xiong et al. (2014)[43]. These represent a high-quality subset of ~30,000 Helitrons identified by the HelitronScanner algorithm that were additionally validated with *in silico* comparisons with the genome of the Mo17 inbred line[43]. For Pack-MULEs, the coordinates of 275 full-length elements from Jiang et al. (2011)[35] were based on the RefGen_v1 genome; hence, we aligned their sequences (BLASTN, *E*-value 1 x 10^{-180}) on the RefGen_v2 genome requiring 100% identity on the complete length of each element. This approach yielded 251 Pack-MULEs. For Sireviruses, we downloaded from MASiVEdb[44] the sequences of 13,833 elements identified in RefGen_v2 using the MASiVE algorithm[66]. We further filtered out elements from all families with >5 consecutive 'N' nucleotides, based on evidence that BLASTN hits between genes and TEs often mapped precisely at the border of these stretches, indicating potential errors during scaffold assembly. Finally, we found cases where elements of the same or different families overlapped with each other. We removed all partially overlapping pairs and the outer and, therefore, fragmented TE from each full insertion. Our final TE population consisted of 1,090 Helitrons, 248 Pack_MULEs and 9,806 Sireviruses. We converted the chromosomal coordinates of TEs from v2 to the most recent v4 genome version using the Assembly Converter tool available in http://www.gramene.org/ (accepted TEs had >80% of length converted), and then overlapped the v4 coordinates with the recent maize TE annotation available at https://mcstitzer.github.io/maize_TEs/ (accepted TEs had >80% overlap). In this way, we retrieved information on insertion age based on terminal branch lengths generated by Stitzer et al. (2019)[34].

For genes, we produced a dataset of 32,551 out of 39,423 genes of the maize B73 RefGen_v2 Filtered Gene Set (FGS) downloaded from http://ftp.gramene.org/maizesequence.org/ alongside other useful files (see below). The FGS genes were filtered to include only evidence-based and not *ab initio* predictions (3,045 genes). The gene set was also i) free of >5 'N' nucleotides like TEs, ii) filtered for the presence of keywords such as 'TE', 'transposable', 'pseudogene', 'copia' and 'gypsy' in various annotation files (ZmB73_5b_FGS_info.txt, ZmB73_5b_FGS.gff, ZmB73_5b_WGS_to_FGS.txt, ZmB73_5a_gene_descriptors.txt, ZmB73_5a_xref.txt), and iii) filtered for similarity (BLASTN, *E*-value 1 x 10^{-20}) of their exons to the conserved domains of the reverse transcriptase and integrase genes of LTR retrotransposons that were identified using Hidden Markov Models (PF07727 and PF00665 respectively) from Pfam[67]. Finally, we linked the v2 gene IDs to the v4 genome version by using files 'updated_models' in https://download.maizegdb.org/B73_RefGen_v3/ and 'maize.v3TOv4.geneIDhistory.txt' in http://ftp.gramene.org. This allowed us to access information on the function of each gene ('Zea_mays.B73_RefGen_v4.43.chr.gff3' in http://ftp.gramene.org) and the syntenic relationships with *Sorghum bicolor* that were generated by Springer et al. (2018)[68] and kindly provided to us by Dr. Margaret Woodhouse of MaizeGDB.

Identification of capture events

The sequence comparison between TE and gene datasets is also critical. High sensitivity (e.g. BLASTN *E*-value of 1 x 10^{-5}), which was a common choice in previous studies[19,20,30,35,58-60], will certainly yield more results, but at the expense of specificity. Here, maintaining our intention to minimize false positive capture events, we opted for a strict BLASTN *E*-value cutoff of 1 x 10^{-40} between the exons of the 32,551 genes and the 11,144 TEs. We only kept BLASTN results when exons belonged to the longest transcript of a gene. The average capture length was 280nt, with a minimum of 90nt and a maximum of 1,932nt. To avoid potential biases in the epigenetic analysis, we removed cases of physical overlaps between genes and TEs, even if the TE contained fragments of genes other than the overlapping one. When exons from multiple genes overlapped partially or fully with a TE, we selected the highest BLASTN bit score to define the true donor gene[30,31,35]. If exons from multiple genes had the same bit score, they were all regarded as true donors and kept for downstream analyses; this was not a common incidence however, as most (88%) had only one true donor.

Often, a TE contained multiple independent capture events, defined as non-overlapping areas within the TE. In total, we identified 6,838 such areas across all our TEs. We tested how this number changed after merging areas located in close proximity to each other, with the assumption that they may in reality represent a single capture event that BLASTN failed to identify in its entirety. By allowing a window of 10nt or 50nt, the number only slightly reduced to 6,724 (98.3%) and 6,379 (93.3%) respectively, suggesting that the majority represent truly independent capture events.

siRNA, methylation and expression data

For siRNA mapping, we retrieved short read libraries for ear (GSM306487), leaf (GSM1342517) and tassel (GSM448857). We used Trimmomatic[69] to trim adaptor sequences, and FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to remove low quality nucleotides until reads had ≥3 consecutive nucleotides with a phred Q score >20 at the 3' end. Reads of 21nt, 22nt and 24nt in length were kept because they represent the vast majority of siRNAs that map to TEs. They were filtered for tRNAs (http://gtrnadb.ucsc.edu/), miRNAs (http://www.mirbase.org/), and rRNAs and snoRNAs (http://rfam.xfam.org/), and then mapped to the RefGen_v2 genome using BWA[70] with default settings and no mismatches. We retrieved with a custom Perl script the ID and number of times each distinct siRNA sequence mapped to a locus (e.g. full-length TE, captured region within the TE, exon) to calculate mapping of distinct siRNA sequences per kb as suggested in Bousios et al. (2017)[71] and to identify siRNAs that crosstalk between the captured fragment within the TE and the exons of donor genes.

For DNA methylation analysis we used previously published BS-seq data from ear (SRA050144) and leaf (SRR850328). Reads were trimmed for quality and adapter sequences with Trimmomatic[69] using default parameters and a minimum read length of 30nt. Trimmed reads were then mapped to the RefGen_v2 genome using bowtie2[72] (v2.2.7, parameters: -N 0 -L 20 -p 2) within the bismark (v0.15.0) software suite[73]. The number of methylated and unmethylated reads at each cytosine in the genome was calculated using bismark_methylation_extractor. Positions with >2 reads were retained for further analysis. Bisulfite conversion error rates, or false methylation rates (FMR), were estimated from reads that mapped to the chloroplast genome. as chloroplast DNA is not expected to be methylated. For the ear sample, FMRs were 0.016, 0.014, and 0.008 for CG, CHG, and CHH sites, respectively. Similarly, FMRs for the leaf sample were 0.008, 0.007, and 0.006. A binomial test incorporating the estimated rates of bisulfite conversion errors (*P*<0.05 after Benjamini-Yekutieli FDR correction) was then used to identify methylated cytosines[74].

For each locus we retrieved the number of total, covered and methylated CG, CHG and CHH sites. Methylation levels across genes and TEs were inferred for each context by dividing the number of methylated to covered cytosines[37]. Only genes with ≥10 covered cytosines in their exons were kept for each context separately. We additionally tried various coverage cutoffs, including previously published ones for maize[37], and obtained highly similar results in relation to the methylation profiles of TEs, free and donor genes.

Finally, we downloaded gene expression data from the ATLAS Expression database (www.ebi.ac.uk/gxa/) for ear (6-8mm from tip of ear primordium; E-GEOD-50191), leaf (tip of transition leaf; E-MTAB-4342) and various tissues of the maize kernel (E-GEOD-62778). Only genes with >0.1 TPM are included in the ATLAS database, hence we classified all other genes as having no expression.

Statistical analyses of donor genes for siRNA mapping, expression and methylation

siRNAs were defined as 'cross-talk' if they mapped to both the exons of the donor gene and the captured fragment within the TE. In order to test whether siRNAs that map to donor genes cross-talk more often than expected by chance, we used a one-sided binomial test. The number of successes is the number of siRNAs that crosstalk, the number of trials is the number of siRNAs mapping to the donor gene, and the probability of success is the proportion of the donor gene length that has been captured by all TEs. If we assume a random distribution of siRNA across the donor gene, the expected probability of mapping of any siRNA onto the captured area is the length of the captured area divided by total gene length. Binomial exact test p-values were corrected for multiple testing using Benjamini & Hochberg (1995).

In order to study the link between methylation levels of donor genes, the number of cross-talk siRNAs and capture length, the glmer function of the R package lme4[75] was used to write a generalized linear model with mixed effects. The r.squaredGLMM function of the R package MuMIn[76] was used to compute the marginal R-squared (the variance explained by the fixed effects, here the number of cross-talk siRNAs and capture length). The binomial family was used, and tissue was set as a random factor. The analysis was repeated separately for the three methylation contexts:

proportion of methylated cytosines ~ number of cross-talk siRNA + capture length + (1|tissue)

Age of capture events

In order to estimate the age of gene capture events, we estimated synonymous and/or non-coding divergence between donor genes and the captured fragments within TEs. The v2 genome GFF file was used to split sequences into coding and non-coding (since in v2 UTRs are included in the first/last exons). The coding parts of donor genes and captured fragments were aligned using MACSE v2[77]. In cases where stop codons were found in the captured gene fragment, they were replaced by 'NNN' in order to compute synonymous divergence (dS) using the yn00 program in the paml package[78]. The non-coding parts of donor genes and captured fragments were aligned using MAFFT v7[79]. The number of substitutions and the number of gap openings were computed using the R package ape. The non-coding divergence was defined as the sum of the number of substitutions plus the number of gap openings divided by the alignment length. To obtain capture age, non-coding divergence and dS were divided by 2 x (1.3 x10-8) as in Ma & Bennetzen (2004)[80].

The glmer function of the R package lme4[75] was used to write a generalized linear model with mixed effects to study the link between capture age and the number of cross-talk siRNAs. The r.squaredGLMM function of the R package MuMIn[76] was used to compute the marginal R-squared (the variance explained by the fixed effects, here capture age). The poisson family was used and tissue was set as a random factor:

cross-talk siRNA number ~ capture age + (1|tissue)

Capture age was either coding (dS) or non-coding divergence between the donor gene and the captured fragment within the TE.

Statistical analyses of TEs for siRNA mapping, expression and methylation

In order to study siRNA mapping to TEs, the glmer function of the R package lme4[75] was used to write an exponential model with mixed effects to study the effect of TE type (with or without gene capture) and TE age. The r.squaredGLMM function of the R package MuMIn[76] was used to compute the marginal R-squared (the variance explained by the fixed effects, here TE type and age). The lsmeans function from the R package lsmeans[81] was used to compute the contrast between TEs with and without gene capture. Tissue was set as a random factor and the number of siRNAs per kb was log transformed:

log(siRNA per kb+1) ~ TE type + TE age + (1|tissue)

A simpler model was also used:

log(siRNA per kb+1) ~ TE type + (1|tissue)

Similarly, a generalized linear model with mixed effects was used to study the effects of TE type and TE age on TE methylation. The binomial family was used, and tissue was set as a random factor. The analysis was repeated separately for the three methylation contexts:

proportion of methylated cytosines ~ TE type + TE age + (1|tissue)

Acknowledgements

A.M. is supported by an EMBO Postdoctoral Fellowship ALTF 775-2017 and by HFSPO Fellowship LT000496/2018-L. D.K.S. is supported by a Postdoctoral Fellowship from the National Science Foundation (NSF) Plant Genome Research Program (1609024). B.S.G. is supported by NSF grant 1655808. A.B. is supported by The Royal Society (award numbers UF160222 and RGF/R1/180006).

References

1. Tenaillon, M.I., Hollister, J.D. & Gaut, B.S. A triptych of the evolution of plant transposable elements. *Trends in Plant Science* **15**, 471-478 (2010).
2. Lisch, D. How important are transposons for plant evolution? *Nature Reviews Genetics* **14**, 49-61 (2013).
3. Cuerda-Gil, D. & Slotkin, R.K. Non-canonical RNA-directed DNA methylation. *Nature Plants* **2**(2016).
4. Matzke, M.A. & Mosher, R.A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature Reviews Genetics* **15**, 394-408 (2014).
5. Sigman, M.J. & Slotkin, R.K. The First Rule of Plant Transposable Element Silencing: Location, Location, Location. *Plant Cell* **28**, 304-313 (2016).
6. Hollister, J.D. *et al.* Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 2322-2327 (2011).
7. Lee, Y.C.G. & Karpen, G.H. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. *Elife* **6**(2017).
8. Maumus, F. & Quesneville, H. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. *Nature Communications* **5**(2014).
9. Quadrana, L. *et al.* The Arabidopsis thaliana mobilome and its impact at the species level. *Elife* **5**(2016).
10. Hollister, J.D. & Gaut, B.S. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research* **19**, 1419-1428 (2009).
11. Cho, J. Transposon-Derived Non-coding RNAs and Their Function in Plants. *Frontiers in Plant Science* **9**(2018).
12. Lockton, S. & Gaut, B.S. The Contribution of Transposable Elements to Expressed Coding Sequence in Arabidopsis thaliana. *Journal of Molecular Evolution* **68**, 80-89 (2009).
13. Kapitonov, V.V. & Jurka, J. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics* **23**, 521-529 (2007).
14. Morgante, M. *et al.* Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics* **37**, 997-1002 (2005).
15. Pritham, E.J. & Feschotte, C. Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1895-1900 (2007).
16. Thomas, J., Phillips, C.D., Baker, R.J. & Pritham, E.J. Rolling-Circle Transposons Catalyze Genomic Innovation in a Mammalian Lineage. *Genome Biology and Evolution* **6**, 2595-2610 (2014).
17. Dong, Y.B. *et al.* Structural characterization of helitrons and their stepwise capturing of gene fragments in the maize genome. *Bmc Genomics* **12**(2011).
18. Ferguson, A.A., Zhao, D.Y. & Jiang, N. Selective Acquisition and Retention of Genomic Sequences by Pack-Mutator-Like Elements Based on Guanine-Cytosine Content and the Breadth of Expression. *Plant Physiology* **163**, 1419-1432 (2013).
19. Yang, L.X. & Bennetzen, J.L. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 19922-19927 (2009).
20. Jiang, N., Bao, Z.R., Zhang, X.Y., Eddy, S.R. & Wessler, S.R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**, 569-573 (2004).
21. Thomas, J. & Pritham, E.J. Helitrons, the Eukaryotic Rolling-circle Transposable Elements. *Microbiology Spectrum* **3**(2015).
22. Barbaglia, A.M. *et al.* Gene Capture by Helitron Transposons Reshuffles the Transcriptome of Maize. *Genetics* **190**, 965-975 (2012).
23. Catoni, M., Jonesman, T., Cerruti, E. & Paszkowski, J. Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling. *Nucleic Acids Research* **47**, 1311-1320 (2019).
24. Grabundzija, I. *et al.* A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature Communications* **7**(2016).
25. Lal, S., Oetjens, M. & Hannah, L.C. Helitrons: Enigmatic abductors and mobilizers of host genome sequences. *Plant Science* **176**, 181-186 (2009).
26. Cerbin, S. & Jiang, N. Duplication of host genes by transposable elements. *Current Opinion in Genetics & Development* **49**, 63-69 (2018).
27. Lynch, B.T. *et al.* Differential pre-mRNA Splicing Alters the Transcript Diversity of Helitrons Between the Maize Inbred Lines. *G3-Genes Genomes Genetics* **5**, 1703-1711 (2015).
28. Wang, J. *et al.* DNA methylation changes facilitated evolution of genes derived from Mutator-like transposable elements. *Genome Biology* **17**(2016).
29. Zhao, D.Y. *et al.* The unique epigenetic features of Pack-MULEs and their impact on chromosomal base composition and expression spectrum. *Nucleic Acids Research* **46**, 2380-2397 (2018).
30. Hanada, K. *et al.* The Functional Role of Pack-MULEs in Rice Inferred from Purifying Selection and Expression Profile. *Plant Cell* **21**, 25-38 (2009).
31. Hoen, D.R. *et al.* Transposon-mediated expansion and diversification of a family of ULP-like genes. *Molecular Biology and Evolution* **23**, 1254-1268 (2006).
32. Juretic, N., Hoen, D.R., Huynh, M.L., Harrison, P.M. & Bureau, T.E. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Research* **15**, 1292-1297 (2005).
33. Schnable, P.S. *et al.* The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* **326**, 1112-1115 (2009).
34. Stitzer, M.C., Anderson, S.N., Springer, N.M. & Ross-Ibarra, J. The Genomic Ecosystem of Transposable Elements in Maize. *BioRxiv* **559922**(2019).

35. Jiang, N., Ferguson, A.A., Slotkin, R.K. & Lisch, D. Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 1537-1542 (2011).
36. Matsumoto, T. *et al.* The map-based sequence of the rice genome. *Nature* **436**, 793-800 (2005).
37. Takuno, S., Ran, J.H. & Gaut, B.S. Evolutionary patterns of genic DNA methylation vary across land plants. *Nature Plants* **2**(2016).
38. Niederhuth, C.E. *et al.* Widespread natural variation of DNA methylation within angiosperms. *Genome Biology* **17**(2016).
39. Seymour, D.K. & Gaut, B.S. Phylogenetic shifts in gene body methylation correlate with gene expression and reflect trait conservation. *BioRxiv* **687186**(2019).
40. Bousios, A. & Darzentas, N. Sirevirus LTR retrotransposons: phylogenetic misconceptions in the plant world. *Mobile DNA* **4**(2013).
41. Bousios, A. *et al.* The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story. *Plant Journal* **69**, 475-488 (2012).
42. Bousios, A. *et al.* A role for palindromic structures in the cis-region of maize Sirevirus LTRs in transposable element evolution and host epigenetic response. *Genome Research* **26**, 226-237 (2016).
43. Xiong, W.W., He, L.M., Lai, J.S., Dooner, H.K. & Du, C.G. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 10263-10268 (2014).
44. Bousios, A. *et al.* MASiVEdb: the Sirevirus Plant Retrotransposon Database. *Bmc Genomics* **13**(2012).
45. Gent, J.I. *et al.* CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Research* **23**, 628-637 (2013).
46. Nobuta, K. *et al.* Distinct size distribution of endogenous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 14958-14963 (2008).
47. Diez, C.M., Meca, E., Tenaillon, M.I. & Gaut, B.S. Three Groups of Transposable Elements with Contrasting Copy Number Dynamics and Host Responses in the Maize (Zea mays ssp mays) Genome. *Plos Genetics* **10**(2014).
48. Li, Q. *et al.* Genetic Perturbation of the Maize Methylome. *Plant Cell* **26**, 4602-4616 (2014).
49. Zhang, L. *et al.* A Genome-Wide Characterization of MicroRNA Genes in Maize. *Plos Genetics* **5**(2009).
50. Schnable, J.C. Genome Evolution in Maize: From Genomes Back to Genes. in *Annual Review of Plant Biology, Vol 66*, Vol. 66 (ed. Merchant, S.S.) 329-343 (2015).
51. Schnable, J.C. & Freeling, M. Genes Identified by Visible Mutant Phenotypes Show Increased Bias toward One of Two Subgenomes of Maize. *Plos One* **6**(2011).
52. Eichten, S.R. *et al.* Heritable Epigenetic Variation among Maize Inbreds. *Plos Genetics* **7**(2011).
53. El Baidouri, M. *et al.* Genic C-Methylation in Soybean Is Associated with Gene Paralogs Relocated to Transposable Element-Rich Pericentromeres. *Molecular Plant* **11**, 485-495 (2018).
54. Gardner, A. & Ubeda, F. The meaning of intragenomic conflict. *Nature Ecology & Evolution* **1**, 1807-1815 (2017).
55. Takuno, S. & Gaut, B.S. Body-Methylated Genes in Arabidopsis thaliana Are Functionally Important and Evolve Slowly. *Molecular Biology and Evolution* **29**, 219-227 (2012).
56. Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics* **18**, 292-308 (2017).
57. Baucom, R.S. *et al.* Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. *Plos Genetics* **5**(2009).
58. Du, C., Fefelova, N., Caronna, J., He, L. & Dooner, H.K. The polychromatic Helitron landscape of the maize genome. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 19916-19921 (2009).
59. Holligan, D., Zhang, X.Y., Jiang, N., Pritham, E.J. & Wessler, S.R. The transposable element landscape of the model legume Lotus japonicus. *Genetics* **174**, 2215-2228 (2006).
60. Sweredoski, M., DeRose-Wilson, L. & Gaut, B.S. A comparative computational analysis of nonautonomous Helitron elements between maize and rice. *Bmc Genomics* **9**(2008).
61. Seymour, D.K., Koenig, D., Hagmann, J., Becker, C. & Weigel, D. Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. *Plos Genetics* **10**(2014).
62. Takuno, S. & Gaut, B.S. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 1797-1802 (2013).
63. Zilberman, D. An evolutionary case for functional gene body methylation in plants and animals. *Genome Biology* **18**(2017).
64. Wendte, J.M. *et al.* Epimutations are associated with CHROMOMETHYLASE 3-induced de novo DNA methylation. *Elife* **8**(2019).
65. Wicker, T., Buchmann, J.P. & Keller, B. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Research* **20**, 1229-1237 (2010).
66. Darzentas, N., Bousios, A., Apostolidou, V. & Tsaftaris, A.S. MASiVE: Mapping and Analysis of SireVirus Elements in plant genome sequences. *Bioinformatics* **26**, 2452-2454 (2010).

67. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Research* **47**, D427-D432 (2019).
68. Springer, N.M. *et al.* The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nature Genetics* **50**, 1282-+ (2018).
69. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
70. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
71. Bousios, A., Gaut, B.S. & Darzentas, N. Considerations and complications of mapping small RNA high-throughput data to transposable elements. *Mobile DNA* **8**(2017).
72. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U54 (2012).
73. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
74. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536 (2008).
75. Bates, D., Machler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **https://www.jstatsoft.org/article/view/v067i01**(2015).
76. Barton, K. MuMIn : multi-model inference. *http://r-forge.r-project.org/projects/mumin/* (2009).
77. Ranwez, V., Douzery, E.J.P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Molecular Biology and Evolution* **35**, 2582-2584 (2018).
78. Yang, Z.H. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
79. Katoh, K. & Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).
80. Ma, J.X. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12404-12410 (2004).
81. Length, R.V. Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software* **69:1–33**(2016).