miRNA-like secondary structures in maize (Zea mays) genes and transposable elements correlate with small RNAs, methylation, and expression

Galen Martin, Edwin Solares, Jeanelle Guadardo-Mendez, Aline Muyle, Alexandros Bousios, Brandon Gaut

miRNA-like secondary structures in maize (*Zea mays*) genes and transposable elements correlate with small RNAs, methylation, and expression

Galen Martin[1], Edwin Solares[1,2], Jeanelle Guadardo-Mendez[1], Aline Muyle[1,3], Alexandros Bousios[4], Brandon S. Gaut[1,*]

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine

[2]Department of Ecology and Evolutionary Biology, University of California, Davis

[3]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France.

[4] School of Life Sciences, University of Sussex, Brighton, UK

*Address for Correspondence: bgaut@uci.edu

ABSTRACT

RNA molecules carry information in their primary sequence and also their secondary structure. Secondary structure can confer important functional information, but it is also a signal for an RNAi-like host epigenetic response mediated by small RNAs (smRNAs). In this study, we used two bioinformatic methods to predict local secondary structures across features of the maize genome, focusing on small regions that had similar folding properties to pre-miRNA loci. We found miRNA-like secondary structures to be common in genes and most, but not all, superfamilies of RNA and DNA transposable elements (TEs). The miRNA-like regions mapped a higher diversity of smRNAs than regions without miRNA-like structure, explaining up to 27% of variation in smRNA mapping for some TE superfamilies. This mapping bias was more pronounced among putatively autonomous TEs relative to non-autonomous TEs. Genome-wide, miRNA-like regions were also associated with elevated methylation levels, particularly in the CHH context Among genes, those with miRNA-like secondary structure were 1.5-fold more highly expressed, on average, than other genes. However, these genes were also more variably expressed across the 26 Nested Association Mapping founder lines, and this variability positively correlated with the number of mapping smRNAs. We conclude that local miRNA-like structures are a nearly ubiquitous feature of expressed regions of the maize genome, that they correlate with higher smRNA mapping and methylation, and that they may represent a trade-off between functional need and the potentially negative consequences of smRNA production.

KEYWORDS: small RNA, *Zea mays*, secondary structure, transposable elements, epigenome

INTRODUCTION

In a highly simplified view, plant genomes consist of transposable elements (TEs) and genes. Both of these components use RNA to transmit coding information between one state (DNA) to another (protein). These RNA molecules carry information in their primary sequence of bases but also by their shape. This shape is primarily defined by the secondary structure of the transcript, which is a product of the intramolecular hydrogen bonds between RNA bases. Secondary structure can mediate the relationship between genotype and phenotype, because it affects the localization (Bullock et al., 2010), splicing (Buratti & Baralle, 2004), and translation (Ding et al., 2014) of mRNAs. As a result, secondary structure influences nearly every processing step in the life cycle of transcripts (Vandivier et al., 2016).

Secondary structures can have another effect: they act as a template for small RNA (smRNA) production (Carthew & Sontheimer, 2009; Li et al., 2012; Hung & Slotkin, 2021). This production takes place through the binding of *Dicer-like* proteins (DCL) (Axtell 2013; Fukudome & Fukuhara 2017) that degrade double-stranded RNA (dsRNA). In other words, when single-stranded RNA (ssRNA) forms a hairpin-like secondary structure, DCLs can recognize structured ssRNA as dsRNA and then degrade the dsRNA to produce smRNAs. This mechanism is essential for the biogenesis of microRNAs (miRNAs), a class of smRNAs that are generally ~22-nt in length and that are derived from longer pre-miRNA transcripts with strong hairpin secondary structures (Carthew & Sontheimer 2009). However, this process is not limited to miRNAs, because 21–24-nucleotide RNAs can also originate from the secondary structure of other non-miRNA transcripts (Li et al., 2012, Slotkin et al., 2003). These small RNAs can, in turn, cause transcripts to enter into the RNA interference (*RNAi*) pathway (Baulcombe 2004; Li et al., 2012; Cuerda-Gil & Slotkin, 2016; Hung & Slotkin, 2021). These observations suggest that sufficiently structured mRNAs, like miRNAs, form secondary structures that act as dsRNA substrates for degradation into smRNAs .

Little is known about how host genomes initially distinguish TEs from genes and target them for smRNA production, but some studies suggest that hairpin structures in TE transcripts act as an immune signal for *de novo* silencing of certain TEs (Slotkin et al., 2003; Sijen and Plasterk, 2003; Bousios et al., 2016; Hung & Slotkin 2021). One such example is *Mu-killer,* a locus that generates small RNAs and thereby silences *MuDR* elements (a DNA transposon) in maize (*Zea mays* ssp. *mays*) (Slotkin et al., 2003). *Mu-killer* consists of a truncated, duplicated,

and inverted copy of *MuDR* that, when transcribed, creates a hairpin secondary structure and is subsequently cut into trans-acting small-interfering RNAs (siRNAs) that target active *MuDR* transcripts. Another potential example comes from Sirevirus long terminal repeat (LTR) retrotransposons in maize (Bousios et al., 2016), which occupy 21% of the maize B73 genome (Bousios et al., 2011). In this study, the authors mapped smRNAs to full-length Sirevirus copies, reasoning that loci important for host-plant recognition and silencing should be associated with a larger number of smRNA sequences than other regions of the elements. Indeed, an excess of smRNAs mapped to regions that had strong predicted secondary structure due to clusters of palindromic motifs (Bousios et al., 2016). These studies present evidence that secondary structure helps initiate silencing of some TEs. In fact, one review has argued that the only characterized pathway to *de novo* smRNA production relies on RNA secondary structure (Hung and Slotkin, 2021). [It should be noted, however, that some phased siRNAs are caused by miRNA cleavage events that apparently do not require secondary structure (Creasey et al., 2014).]

If RNA sequences form miRNA-like hairpin structures, leading to the production of smRNAs, two important questions must be addressed. First, how common are miRNA-like secondary structures across the immense diversity of plant TEs? One prominent review of small RNAs argued that there is an urgent need to annotate hairpins that may have the capacity to act as a template for smRNA production (Axtell, 2013), but this need has not yet been met. Thus far, the importance of hairpin structure for *de novo* silencing has been implicated only in a few individual TE families. Second, secondary structure is not unique to TEs and exists within genes too. How often do genes have such structure, and is there evidence that genes form dsRNA substrates in these regions, too? Li et al. (2012) documented a positive relationship between stability of mRNA structure and small RNA abundance for *Arabidopsis thaliana* genes, suggesting that genes do form dsRNA substrates. Yet these genes are still expressed, potentially due to countermeasures that moderate the potential effects of smRNAs on genes, including hypothesized protection against RNAi caused by high GC content (Hung and Slotkin 2021) and active gene demethylation (Gong et al., 2002; Zhang et al., 2022). Although it has long been thought that miRNA loci may be derived from TE sequences (Roberts et al., 2014), there has not yet been, to our knowledge, a genome-wide comparison of miRNA-like secondary structures among genes and TE superfamilies.

In this study, we predict secondary structures in genes and TEs of the maize B73 genome. Secondary structure can be empirically measured through sequencing techniques such as DMS-seq and SHAPE-seq (Yang et al., 2018), which is applied to the transcribed component of whole genomes (Ding et al., 2014; Ferrero-Serrano et al., 2022). However, this approach requires that the sequences of interest are expressed, preventing comprehensive investigation of plant TEs, most of which are silent. These methods are also difficult to perform on large genomes with high repeat content, so that genome-wide 'structurome' sequencing has thus far only been completed on plants with relatively small genomes, like *Arabidopsis* (Ding et al., 2014; Bevilacqua et al., 2016) and rice, *Oryza sativa* (Ritchey et al., 2017). The second approach, which we adopted here, relies on bioinformatic predictions based on genome sequence data. Secondary structure prediction is a subject of active research, and methods vary in their predictions and accuracy. Here we employ two separate methods that rely on distinct algorithms to identify regions with properties similar to miRNA-like hairpins. Briefly, the first uses RNAfold (Lorenz et al., 2011), which estimates the minimum free energy (MFE) of the most likely secondary structure of a given sequence (Nussinov and Jacobson, 1980; Zuker and Stiegler, 1981). Following precedence, we apply RNAfold in a windows-based approach. The second relies on a newer tool, LinearPartition (Zhang et al., 2020), that calculates a partition function for a complete (i.e., not windows-based) RNA sequence. The LinearPartition function includes the sum of equilibrium constants for all possible secondary structures for a sequence (i.e, not just the most likely structure). We focus specifically on detecting regions with miRNA-like secondary structures, because miRNA are known to fold and thereby act as a dsRNA substrate for *Dicer*-like mechanisms. .

After performing computational annotation to predict miRNA-like regions in the genes and TEs of maize, we investigate the relationship between these regions to smRNAs, methylation levels, chromatin accessibility and, where applicable, gene expression (**Fig S1**). With these data, we address four sets of questions. The first focuses on predicted secondary structure: How often do TEs and genes contain regions of miRNA-like regions? And are these regions in specific locations? The second set of questions focuses on the relationship between secondary structure and smRNAs. Do miRNA-like regions consistently map more smRNAs, and, if so, of what size? The question of size is important because it is thought that dsRNA degradation via *Dicer* feeds into post-transcriptional gene silencing (PTGS) pathways, which

tends to rely on 21- and 22-nt smRNAs. In contrast, pathways that lead to transcriptional gene silencing (TGS) tend to rely more often on 24-nt smRNAs, although these size distinctions are neither strict nor universal (Fultz & Slotkin, 2017; Panda et al., 2020). Our third set of questions focuses on the potential genomic implications of hairpins and smRNAs. Do these miRNA-like regions have higher methylation levels or specific chromatin properties? Finally, we assess the effects of miRNA-like secondary structures on gene expression by including data from 26 parents of the maize Nested Association Mapping (NAM) lines (McMullen et al., 2009; Hufford et al., 2021).

RESULTS

Two methods to predict miRNA-like secondary structures and their comparison

We adopted two complementary bioinformatic methods to identify miRNA-like hairpin regions (**Fig 1a**). The details of their implementation are given in the Materials and Methods. Here we provide an overview of the methods and compare their performance. To aid the reader, we also provide terms that are used to characterize analyzed sequences (**Table 1**).

RNAfold: The first method applied RNAfold to sliding windows of 110 nt, following previous work (Wang et al., 2009; Bousios et al., 2016). The 110 nt windows were originally designed by Wang and co-authors to include regions that map 20-25 nt small RNAs, along with ~90 bp of flanking sequence (Wang et al., 2009). This approach established that pre-miRNA windows of this size typically have MFEs <-40 kcal/mol (Wang et al., 2009); we used that empirical cutoff to define windows of secondary structure with miRNA-like stability. By focusing on regions of similar size to pre-miRNA transcripts and by employing their empirical threshold cutoff of -40 kcal/mol, we in effect used miRNA loci as a 'positive control' for ssRNAs that are expected to form secondary structures.

We applied RNAfold across features of the B73 reference maize genome (version 4.0)(Jiao et al., 2017). The features included miRNA precursor loci, TEs and genes. The TEs included all families annotated in Jiao et al. (2017), including Long Terminal Repeat elements (LTRs), Terminal Inverted Repeat elements (TIRs), Helitrons, Long Interspersed Nuclear Elements (LINEs), and Short Interspersed Nuclear Elements (SINEs). Within these TE types, we focused on superfamily categories (Wicker et al., 2007), which distinguished (for example) between *Ty3*/RLG and *Copia*/RLC LTR elements and among TIR elements like *Mutators*/DTM

and *Harbingers*/DTH. [Note that throughout the paper we refer to TE superfamilies by their names and also their three-letter designation from Wicker et al., 2007 (**Table 2**)]. Notably, these annotations do not typically include miniature inverted terminal repeats (MITEs), a class of small non-autonomous TEs that often contain strong secondary structures. For genes, we studied both the annotated gene—which included untranslated regions (UTRs), exons, introns—as well as mature transcripts that lacked introns. Altogether, with this method we examined 373,485 features representing 15 distinct feature categories (**Table 2**). Because we used sliding windows, each nucleotide within a feature corresponded to one sliding window (for all but the final 109 nucleotides of a sequence). This approach was a massive bioinformatic undertaking, requiring an MFE calculation for a total of 3.56 billion windows.

Because each feature consisted of many RNAfold windows, we used summary statistics to characterize local secondary structure in each feature (**Table 1**). These included the minimum MFE (minMFE), which was the MFE of the window with the strongest predicted secondary structure for each feature, and mean MFE (meanMFE), which averaged MFE across windows within a feature. For each feature, we also concatenated overlapping windows with MFE < -40 kcal/mol, designating these as lowMFE regions (**Table 1; Fig 1a,b**).

One concern about using MFE as a quantitative statistic is that it varies by G:C composition (e.g., higher G:C content tends to induce more stable secondary structures) and primary sequence (e.g., whether the order of bases forms palindromes and stem-loop structures). Because we were primarily interested in secondary structure resulting from the latter, we controlled for base composition by randomizing the sequence of each feature five times and then repeating MFE predictions each time, requiring another 17.8 billion (=5 x 3.56 billion) window computations. By randomizing, we identified features that had more stable secondary structures than expected given their nucleotide composition. We then classified a feature as "RF-structured" (RF for RNAfold) when it contained windows with MFEs < -40 kcal/mol and also had a minMFE significantly lower than permutations ($p < 0.05$, one-sided Wilcoxon test, Benjamini and Hochberg corrected) (**Table 1**). Conversely, we labeled features as "unstructured" when their minMFE was not significantly lower than that of randomized sequences. [We report the differences between randomized and observed minMFE values for each feature category in **Fig S2**.] Overall, 76% (286,774 of 373,485) of features were RF-structured - i.e., contained regions of miRNA-like structures by this criteria (**Table 2**).

LinearPartition: The second prediction method was based on LinearPartition (Zhang et al., 2020). This approach did not rely on sliding windows to infer local secondary structure but analyzed the complete sequence of each feature. The advantage of this was that each feature required only one computational analysis, vastly improving computational burden and speed. Accordingly, we applied this method to the same set of 373,485 features as RNAfold but also to a larger, updated version of maize TE annotations (Stitzer et al., 2021), resulting in an expanded dataset of 467,255 features (**Table 2**).

For each sequence, LinearPartition calculated the partition function, summarized by the parameter *Q*. For each nucleotide site within a feature, the method calculated a pairing probability between all nucleotides in the feature. We focused on nucleotide pairs with high probabilities of pairing (> 0.90) and searched within each feature for runs of nucleotides that matched widely-accepted miRNA annotation guidelines for plants (Axtell and Meyers 2018). These guidelines defined hairpins consisting of consecutive stretches of ≥21-nucleotides that were likely to pair (>90% probability) with <5 mismatched nucleotides, including <3 mismatches in putative asymmetric bulges (i.e., places where the gap on one side of a hairpin was > the gap on the other side of the hairpin)(**Fig. 1a**; see **Methods** for details). We called sequences that fit these criteria "LP-hairpins" (**Table 1**).

Comparing the methods: It is worth emphasizing similarities and differences between the two methods. Both focused on identifying regions of strong local secondary structures within features, based on known properties of miRNA-like regions. The MFE method focused on regions of high local structure (MFEs < -40 kcal/mol), without reference to the properties of those structures, like the length of stem loops. In contrast, LinearPartition focused on regions along the complete sequence that matched specific length and size criteria. Because the two methods utilized different miRNA-like properties, we did not expect them to correlate perfectly throughout the genome.

Yet, they did yield significant consistencies and overlaps. For example, we contrasted the two entire-sequence summary statistics—i.e., meanMFE and the partition function normalized for feature length (Q_{norm}). Across structured features, Q_{norm} correlated strongly with meanMFE (**Fig 1c**)($R^2 = 0.73$ across all feature types and $R^2 = 0.97$ across genes; P = 0) and weakly ($R^2 = 0.04$) but still significantly ($P = 3.05 \times 10^{-10}$) with minMFE. The low correlation between Q_{norm} and minMFE was not unexpected, because minMFE focuses on one window within a feature, as

opposed to the property of an entire sequence. However, we also compared the overlap in genomic locations between LP-hairpins and low (<-40) MFE regions (**Fig 1a**). Across all of the 287,744 RF-structured features (**Table 2**), 78.46% of LinearPartition hairpins were within a lowMFE region. Given that lowMFE regions collectively comprised ~22.95% of annotated features, this represented a substantial 12.2-fold enrichment of LP-hairpins within lowMFE regions. By design, lowMFE regions were much larger (median = 348 nt) than LP-hairpins (median = 25 nt), and therefore took up a much larger proportion of the space inside of comparable features. (In total, lowMFE regions constituted 1.9×10^8 nt vs 1.7×10^7 nt for LP-hairpins). These comparisons demonstrate that LP-hairpins are based on a narrower definition, but that the two methods generally agree.

Finally, we compared the performance of the two methods based on a control dataset: annotated pre-miRNA loci from the B73 reference (*n*=107; **Table 2**). Most (71.0%) of this set were RF-structured (**Table 2**), indicating that the MFE threshold defined by Wang et al (2009) generally conformed to existing annotations. Similarly, most (66.36%) of the annotated pre-miRNA loci had LP-hairpins (**Table 2**).

The prevalence and locations of miRNA-like secondary structures

Prevalence of miRNA-like secondary structure across TE superfamilies: Using both methods of prediction, we detected substantial variation in the prevalence of miRNA-like secondary structures among TE categories. Some TE superfamilies contained little evidence of structure. For example, the *LINE* (RIL and RIT) elements had no RF-structured elements and also had no detectable LP-hairpins (**Table 2**). Because the 2017 annotation from Jiao et al. (2017) contained few (*n*=65) RIL and RIT elements, we repeated the LinearPartition analysis with an expanded set of *n*=773 elements from Stitzer et al. (2021), finding again that only a small subset (~3%) contained hairpins (**Table 2**). *SINEs*/RST also had very low incidences of miRNA-like structure, with no RF-structured elements and <2% containing LP-hairpins (**Fig. 1b**). In contrast to LINEs and SINEs, LTR elements generally had abundant miRNA-like structures. For example, 98% of *Copia*/RLC elements had RF-structure and 58.0% had LP-hairpins (**Table 2; Fig 1b**). We note, however, that LTR elements were longer on average than the other TE subfamilies, and also that there was an overall negative relationship between feature length and minMFE across all 15 feature categories ($P < 2.2 \times 10^{-16}$, R2 = 0.20, linear model; **Fig S3**).

Just as the prevalence of miRNA-like regions varied across RNA-based superfamilies, they also varied among DNA-based TE superfamilies. *Mutator*/DTM elements were especially notable for the high percentage of elements with LP-hairpins, at up to 62.82%, while 32.52% of *CACTA*/DTC elements contained LP-hairpins. Fewer than half of the annotated *Tc1-Mariner*/DTT and *PIF-Harbinger*/DTH elements were RF-structured or contained LP-hairpins (**Table 2**), but this corresponded to thousands of elements in these superfamilies that contain miRNA-like regions.

It is worth making two overarching observations from the analyses reported in Table 2. First, the percentage of sequences identified by RNAfold and LinearPartition were correlated across the 15 feature categories ($R=0.65$; $p<0.001$), suggesting again that the two methods identified similar characteristics in most superfamilies. Second, the expanded TE dataset of Stitzer et al. (2021) exhibited similar trends to the Jiao et al. (2017) annotation dataset ($R=0.96$; $p<0.001$). For example, LINEs, SINEs and *hAT*/DTA elements generally had low proportions of elements with LP-hairpins in both annotation sets, while LTR superfamilies had high proportions in both annotation sets.

Biases in the locations of miRNA-like regions: We next examined the locations of miRNA-like secondary structure across the length of each feature type. For these analyses, we focused only on the 286,744 features that were predicted to have RF-structure (**Table 2**). For each feature category, we separately mapped the positions of lowMFE regions and LP-hairpins along their lengths (**Fig 2**). Consistent with previous work (Bousios et al., 2016), both lowMFE and LP-hairpins were concentrated within the LTRs of *Copia*/RLC elements. In contrast, *Ty3*/RLG elements generally lacked an obvious peak for miRNA-like structures. Most DNA transposon superfamilies had relatively uniform distributions of lowMFE regions across their lengths (**Fig S4**), but LP-hairpins were biased heavily towards the terminal inverted repeats for TIR elements like *Mutator*/DTM (**Fig 2**), *hAT*/DTA and *CACTA*/DTC elements (**Fig. S4**). Finally, *Helitrons*/DHH had a distinct 3' bias for both lowMFE regions and LP-hairpins (**Fig 2).** reflecting the ~11 nt stem-loop structure common to *Helitron* 3' ends (Kapitonov & Jurka 2007; Xiong et al., 2014). The take-home messages were that: *i*) some superfamilies – like *Helitron*/DHH, *Mutator*/DTM and *Copia*/RLC – exhibited notable biases in the locations of miRNA-like regions and *ii*) these inferences were similar between the two prediction methods.

Motifs within miRNA-like structures: Distinct sequence motifs could define lowMFE regions. For each TE superfamily, we extracted all the sequences of lowMFE regions and input them into the Multiple EM for Motif Elicitation (MEME) suite motif discovery tool (Bailey and Elkan, 1994), which finds overrepresented sequence motifs within a set of sequences. As expected (Bousios et al., 2016), we recovered the previously identified consensus Sirevirus palindrome, CACCGGACNNNGTCCGGTG (**Fig S5**) as the most abundant motif in *Copia*/RLC elements (MEME e-value = $5.3x10^{-677}$). This motif appeared in 42.9% of RLC structured regions. This same palindrome was also the most abundant motif in *Helitron*/DHH transposons (MEME e-value = 1.0e-165), appearing in 5,231 DHH structured regions (10.7%). This observation could reflect independent emergence of these motifs in the two superfamilies or frequent insertion of one type of element into the other.

miRNA-like secondary structure within genes: A higher percentage (69.0%) of genes were RF-structured than contained LP-hairpins (29.8%) (Table 2). When we examined the distributions of miRNA-like structures across genes and their mature transcripts, we found that the two methods differed in their predictions. In 85% of genes (**Fig 2**), lowMFE regions overlapped the 5' UTRs, where secondary structures are known to participate in ribosome binding and translation (Babendure et al., 2006; Matoulkova et al., 2012). In contrast, LP-hairpins were fairly uniformly distributed across gene lengths (**Fig 2**), with perhaps a slight bias towards the middle of the gene as documented previously in *Arabidopsis* (Li et al. 2012). Most (76.19%) of these LP-hairpins were found in introns, so that far fewer (5.02%) of mature mRNA transcripts had LP-hairpins (**Table 2**). The lowMFE results demonstrate that 5' UTRs commonly have regions of local secondary structure but infrequently contained LP-hairpins.

Comparing miRNA-like secondary regions to smRNA diversity

Correlations between miRNA-like regions and smRNA mapping abundance: Under the dsRNA-substrate model, genomic regions of high secondary structure should have homology to more smRNAs than non-structured regions. To test the hypothesis, we mapped 21, 22, and 24-nt smRNAs from up to 42 published smRNA libraries (see **Methods**; **Table S1**) to the B73 maize genome, and then counted the number of distinct smRNA sequences (also known as 'smRNA species') (Bousios et al., 2017) that mapped with 100% identity to genomic regions. Because of their different functions (Axtell, 2013; Borges and Martienssen, 2015), we examined smRNAs in

the three size classes (21, 22, and 24 nt) separately. Two caveats should be mentioned regarding these small RNAs. First, although we suspect many of these small RNAs to be hairpin-derived RNAs (hpRNAs) (Axtell, 2013), we do not know their origin and refer to them by the more general 'smRNA' term for clarity and concision. Second, we do not know that each smRNAs identified here function as siRNA, merely that they are the correct size to act as a canonical siRNAs.

We first examined the relationship between miRNA-like regions and smRNAs using a linear model across all 373,485 features of the Jiao et al. (2017) annotation set, using correlation statistics. The correlation coefficient was generally small—e.g., R^2 was ~0.1 for models incorporating minMFE—but highly significant (**Table 3**). Moreover, the results were significantly positive for all RNAfold and LinearPartition summary metrics **(Table 3**). Extending this approach separately to the 15 individual feature categories, three smRNA lengths, and three metrics (minMFE, meanMFE and Q_{norm}), 82% of correlations were significant after false discovery rate (FDR) correction (**Table S2**).

Overall, these results indicate a weak but consistent relationship between presence of miRNA-like secondary structure in features and the number of smRNAs that map to those features. We did find some interesting outliers, however. First, the relationship between smRNAs and minMFE statistics were generally not significant for miRNAs (**Table S2**), perhaps reflecting small sample sizes (n=107) or perhaps the fact that miRNA loci generate few distinct smRNAs, despite being highly expressed. Similarly, some LINE comparisons also were typically not significant; LINEs were heavily saturated with for all three smRNA size classes (**Fig S6**) but few had detectable miRNA-like regions. Second, the estimated linear relationships were typically higher for 21 and 22-nt smRNA than for 24-nt smRNA, which is consistent with their role during the initiation of silencing (**Table 3&S2**) and with the observation that DCL-like processing of dsRNA substrates typically yield 21- and 22-nt smRNAs. In genes, for example, correlations between minMFE and 21-22 nt smRNAs were again weak but highly significant ($R^2 = 0.01$, $P < 4.12 \times 10^{-106}$), but the correlation with 24-nt smRNAs was not ($R^2 = 8.35 \times 10^{-05}$, $P = 0.072$)(**Table S2**).

Measuring smRNA abundance with skew: We also examined the relationship between miRNA-like structures and smRNA counts within features by measuring smRNA mapping *skew*, which measures the ratio of smRNA mapping in miRNA-like vs. non-miRNA-like regions

(**Table 1** and Methods). We defined skew to be zero when smRNA mapping was equivalent on a per nucleotide basis between miRNA-like vs. non-miRNA-like regions, and skew ranged from -1.0 to 1.0. When it was positive, smRNA mapping was more abundant in miRNA-like regions.

Generally, TEs in all superfamilies exhibited positive skews, reflecting the tendency for more smRNAs to map to LP-hairpins (**Fig 3a,b)** and the lowMFE regions of RF-structured elements (**Fig S7**). As just one example, *Copia/RLC* elements had positive skews, with slightly higher skews for 22-nt smRNAs as opposed to 21 and 24-nt smRNAs (**Fig 3a)**. These results were confirmed by a linear mixed effects models, because all three smRNA lengths were significantly higher in *Copia*/RLC LP-hairpin regions with all three metrics (i.e., minMFE, meanMFE and Qnorm; all P-values < 1.23×10^{-04}; **Table S2; Fig S8** & **S9**). Overall, LTR elements had more obvious skew than DNA elements, although five of six DNA superfamilies had positive skews for all three smRNA lengths (**Fig. 3a**). These observations were largely supported by mixed effects models (**Table S3** & **S4**), where all TE superfamilies showed significantly higher smRNA mapping to both LP-hairpin and lowMFE regions at all three smRNA lengths (*P*-value range 9.3×10^{-04} in *Rle*/RIT elements to 0.0 in many LTRs, TIRs, and helitrons).

We also examined skew within genes. Genes had homology to far fewer smRNA species than most TE types—nearly 100-times less in most cases (**Fig S6**)—but smRNA species abundance was roughly equivalent between genes and their transcripts. Although genes mapped fewer smRNAs overall, they had stronger skews than any of the TE superfamilies. For example, roughly three-fold more smRNAs (of all size classes) mapped to lowMFE in genes, compared to the 1.5- and 1.3-fold difference in *CACTA*/DTC transposons and *Copia*/RLC retrotransposons. This effect was more pronounced for LP-hairpins. For example, LTR retrotransposons (which includes the RLC, RLG and RLX superfamilies) had a 2.9-fold greater smRNA density in LP-hairpins compared to non-hairpin regions, but genes had a ~89-fold greater density. Consistent with these observations, linear mixed effect models were significant for higher smRNA abundance in lowMFE regions and LP-hairpins of genes for all three smRNA lengths ($P \cong 0$; **Table S3** & **S4**). Comparisons of overall smRNA mapping densities between miRNA-like regions and other regions in genes and TEs can be seen in **Figs S8** (lowMFE) & **S9** (LP-hairpins).

Finally, we included organellar genes as negative controls, because they are typically sequestered from the cytosolic complexes like *DCL* and *RdR6* and hence should not exhibit any skew. smRNAs mapped to organellar genes at low levels, but as expected did not exhibit any skew (**Fig. S10).**

Expression matters: putatively autonomous vs. non-autonomous TEs

Non-autonomous DNA transposons are not transcribed (except when they are within expressed UTRs or introns), and therefore RNA secondary structure generally cannot drive the creation of smRNAs for these elements (Panda et al., 2016). We therefore predicted that there could be a difference in skew between autonomous and non-autonomous DNA elements. To investigate, we separated DNA transposons into nonautonomous and autonomous elements using transposase homology data (Stitzer et al., 2021)(see **Methods**), and then repeated our skew and linear model analyses. In most cases, non-autonomous elements had notably less smRNA skew towards miRNA-like regions than autonomous elements (**Fig 3b**), as we had predicted. This pattern was consistent among *Helitron*/DHH (autonomous mean skew among all smRNA lengths = 0.91, non-autonomous mean = 0.37), *CACTA*/DTC (autonomous mean = 0.44, non-autonomous mean = 0.34), *Harbinger*/DTH elements (autonomous mean = 0.37, nonautonomous mean = 0.27), and *Mutator*/DTM (autonomous mean = 0.51, non-autonomous mean = 0.05), but it was particularly notable for 21 and 22-nt smRNAs ($P < 7.5 \times 10^{-31}$) among *Helitrons*/DHH and *Mutator*/DTM, most of which are non-autonomous in maize (Stitzer et al., 2021). Note that all *Mariner*/DTT elements were non-autonomous, which may relate to their overall lack of skew (**Fig 3b**).

Methylation peaks in miRNA-like regions

One function of smRNAs is to recruit methylases, leading to RNA-directed DNA methylation (RdDM). We reasoned that miRNA-like structures should be more highly methylated because they map more smRNAs. We further predicted that this effect should be primarily detected in the CHH context, because mCHH is deposited *de novo* each generation (Law and Jacobsen, 2010).

We employed B73 whole-genome methylation data (Hufford et al., 2021) to measure weighted methylation levels (Schultz et al., 2012) across the genome. We then plotted

methylation levels centered on regions of miRNA-like structure and 2 kb of the upstream and downstream sequences. Both LP-hairpins (**Fig. 4**) and lowMFE regions (**Fig S11**) demonstrated peaks of CHH methylation centered on the region; this peak dissipated rapidly, especially for LP-hairpins. These peaks were found in all feature types with detectable miRNA-like structures, including RNA elements, DNA elements and genes. We also confirmed that miRNA-like regions had significantly higher levels of CHH methylation than other regions by comparing them to randomly chosen unstructured regions of the same length as LP-hairpins (**Fig. 4**). Finally, we found that CHH methylation levels in LP-hairpins were significantly higher than those in the rest of the corresponding sequence (paired *t*-test; *P* values between 3.43×10^{-81} and 1.16×10^{-165} among genes, TIRs, LINEs, LTRs, and helitrons), with enrichments as high as ~10x in genic hairpins. These observations complement the smRNA mapping results and confirm that our miRNA-like regions have detectable epigenetic correlates.

miRNA-like structures and gene expression

Genes possess regions with stable RNA secondary structure (**Figs 1&2**), and this secondary structure coincides with the presence of smRNAs (**Fig 3c & Table S3-S4**) and methylation (**Fig 4 & S11**). Yet, genes are usually expressed, which raises the question as to whether these miRNA-like structures have a quantifiable relationship to gene expression. To address this question, we used previously published RNA-seq data from 23 B73 tissues across developmental stages (Walley et al., 2016). We focused these analyses on structured genes with lowMFE regions (as opposed to LP-hairpins), both because they were common in the UTRs and gene bodies of genes (**Fig. 2**) and because 5' secondary structure is known to be important to gene function. In contrast, LP-hairpins were detected in only ~5% of genic transcripts (**Table 2**); however, the results presented below for lowMFE regions were often recapitulated with LP-hairpin data.

We began by comparing expression in 27,025 structured *versus* 5,060 unstructured genes. Structured genes had significantly higher expression (*t*-test, $P < 2.0 \times 10^{-16}$)(**Fig 5a**), and this was true for all tissues (**Fig S12**) as well as for genes that contained LP-hairpins (**Fig S13**). We suspected, however, that most unstructured genes were either pseudogenes or misannotated. To focus on evolutionarily conserved (and hence presumably *bona fide*) genes, we identified 24,784 B73 genes with syntelogs in *Sorghum bicolor* (Muyle et al., 2021)(see **Methods**). Among the

syntelog set, 16,171 were structured and 460 were unstructured. Structured syntelogs still had a mean expression level that was slightly higher than unstructured syntologs ($P = 3.7 \times 10^{-4}$; **Fig 5a**). More important, however, was the quantifiable relationship between the minMFE and gene expression. Among structured syntelogs, the relationship was significantly positive—i.e, such that gene expression peaked at a minMFE of ~40 kcal/mol (**Fig. 5b**). The opposite was true among unstructured genes, because higher expression occurred with lower MFEs (**Fig. 5b**). This pattern implies both a relationship between gene expression and the properties of secondary structures and also the existence of an optimal minMFE for gene expression. These trends are present for many of the 23 separate B73 tissues separately (**Fig. S14**) and for the complete gene set of genes—i.e., not just genes with syntelogs (**Fig. S15)**.

Among syntelogs, structured genes also mapped significantly more smRNAs than unstructured genes (**Fig. 5c**), which raises an interesting question: Could this phenomenon modulate the expression of genes? To examine this idea, we examined expression data across the 26 nested association mapping (NAM) founder lines (McMullen et al., 2009). For these analyses, we assumed that the secondary structure designations predicted in B73 applied to its syntelog across all 26 NAM parents (Hufford et al., 2021). We then compared gene expression among lines using the coefficient of variation (CV), based on expression values that were normalized across eight tissues in each line (Hufford et al., 2021)(see Methods). Our analyses revealed that structured genes had significantly higher CVs than non-structured genes ($Ps < 0.01$, permutation test)(**Fig 5d**). This was true both for comparisons between all genes in each group and between a downsampled subset of structured genes that was equal in size to the set of unstructured genes. One concern about this analysis is that the CV is standardized by the mean, which could bias results, but this did not drive our observations for three reasons. First, mean expression did not vary substantially between structured and unstructured syntelogs (**Fig. 5a**). Second, we fitted a linear model of expression CV as a function of B73 gene expression, but the correlation was negative (i.e., more highly expressed genes were slightly less variable across lines; $R^2 = 6.1 \times 10^{-4}$, $P = 1.5 \times 10^{-7}$, estimate = -0.01). Third, we examined CV across 23 B73 tissues. There was no difference in CV between structured and unstructured syntelogs across tissues (**Fig. 5c**), illustrating that the CV metric alone does not explain the significant difference across genotypes.

Can the variable expression of structured genes be explained by smRNAs? We predicted that more smRNAs should lead to more expression variation across lines. To investigate this possibility, we fit a linear model of expression CV as a function of smRNA density and found that CV was positively correlated with smRNA abundance ($P = 6.7 \times 10^{-283}$; $R^2 = 0.010$). To see if an effect was discernible between structured genes of variable minMFE values (as suggested by **Fig 4b**), we separated structured genes into four quartiles based on their minMFE and then plotted the number of smRNAs that map to each gene in B73. Consistent with our hypothesis, genes in the lowest minMFE quartile mapped more smRNAs than the other three quartiles for all three smRNA lengths, and minMFE was significantly but weakly correlated with CV in a linear model ($P = 5.8 \times 10^{-79}$; $R^2 = 0.0031$).

This evidence shows that higher CVs for expression are related to the number of smRNAs that map to a gene, but additional factors likely cause (or contribute) to expression variability across NAM genotypes. One factor is chromatin accessibility. We assessed whether accessibility varies more in lowMFE genic regions by using ATAC-seq data (Hufford et al., 2021), which defines accessible chromatin regions (ACRs) among parents (see Methods). For each NAM parent, we identified whether ACRs overlapped with lowMFE regions more than unstructured (MFE > -40kcal/mol) genic regions. We found no difference between the two categories (**Fig 5e**). Genetic effects, like SNPs and structural variants (SVs), contribute to gene expression variation across the NAM lines, particularly given that regions of structure can have altered mutation rates (Hoede et al., 2006). We therefore also examined SNPs and SVs in these regions, based on the data of Hufford et al. (2021). We found that lowMFE regions were less likely to contain SNPs or SVs than unstructured genic regions (**Fig. 5e**), which superficially discounts the idea that higher CVs for expression are caused by genetic effects due to miRNA-like regions having notably high mutation rates.

DISCUSSION

We have profiled miRNA-like secondary structure in annotated features of the maize genome. To our knowledge, this study is the first to comprehensively catalog such structures, and we have done so by applying two bioinformatic prediction methods. The methods rely on different algorithms (RNAfold vs. LinearPartition), different approaches (overlapping windows vs. no windows) and on different characteristics to define miRNA-like regions. By design, the

LinearPartition analyses relied on a narrower definition (**Fig 2**), and so there were fewer observations. Yet, the two methods provide largely concurrent insights about miRNA-like regions, including their relative abundances among TE superfamilies (**Table 2)**; their locational biases in some TE superfamilies **(Fig 2**); their association with elevated smRNA counts in TEs and genes (**Fig 3**); and their genome-wide correspondence to peaks of methylation (**Fig 4**).

Detecting miRNA-like secondary structures

For detecting secondary structure, we have included two positive controls: miRNA precursor loci (Wang et al, 2009) and *Copia*/RLC elements (Bousios et al., 2016). As expected, these two feature categories have extreme statistics. For example, *Copia*/RLC elements have the highest proportion of RF-structured elements (**Table 2**) and also the lowest average minMFE, reflecting previously recognized regions of strong secondary structure (**Fig. 1**). Our other positive control set, miRNA precursor loci, have a high proportion of RF-structure and the highest proportion of LP-hairpins (**Table 2**). However, these positive controls also indicate an appreciable false negative rate, because 29% (RF-structure) and 38% (LP-hairpin) of pre-miRNA loci do not have detectable miRNA-like structures. It is of course possible that misannotations of miRNA precursors contribute to these false negative rates.

The methods have additional limitations. We need to first reiterate that the approach was not designed to identify *all* secondary structures. Our goal was to identify regions similar to miRNA precursors, because they are thought to be involved in forming dsRNA substrates that lead to the production of smRNAs. Second, there are limitations to the TE annotation sets. For example, miniature inverted repeats (MITEs) are not included in either annotation set. MITEs are short non-autonomous elements that are characterized by their tendency to form stem-loop structures and to insert near genes (Bureau & Wessler, 1992, 1994), where they are often incorporated in read-through transcripts. They are an interesting topic for additional work, but we can provide no insights about them here. Third, we know that some summaries are biased—e.g., minMFE is correlated with feature length and lowMFE regions are more likely in sequences with high G:C composition. We have addressed these biases by using multiple summary statistics, by randomizing the primary sequence to test for significant evidence of structure and by using two prediction methods. Finally, we recognize that bioinformatic predictions are approximations that may not correspond to *in vivo* assessments (Ding et al., 2014).

Nonetheless, despite these limitations, the two distinct prediction methods yield several similar trends, including higher smRNA mapping and methylation levels in miRNA-like regions (**Table 2** and **Figs 1,2**). One prosaic explanation for these results is that they are caused by systematic biases in the prediction methods, but this seems highly unlikely because: *i*) error in secondary structure prediction should lead to randomness—i.e., inconsistent correlations, *ii*) the inclusion of false negatives among unstructured elements makes the measured correlations inherently conservative and *iii*) the results, while not identical, are largely consistent between prediction methods. Since both genes and TEs exhibit this relationship, we conclude that the association between miRNA-like structure and smRNA abundance is a general characteristic of the maize epigenome. Our work extends this relationship from a few examples to the genome-wide scale.

miRNA-like regions, epigenetic signals and potential mechanisms

Given known pathways of miRNA and smRNA biogenesis (O'Brien et al., 2018; Hung & Slotkin, 2021), we believe the most likely explanation for the observed association is that miRNA-like secondary structures lead directly to smRNA production via *Dicer-like* mechanisms. This conclusion is bolstered by the fact that smRNA skew is more pronounced for expressed genomic regions—like genes and putatively autonomous elements—for which this mechanism is expected to be most active (**Fig. 3**). There are likely exceptions to this pattern, though. For example, MITEs can be frequently expressed owing to their insertion near genes (Zhang et al., 2000). We predict, then, that "expressed" non-autonomous MITEs will exhibit skews similar to autonomous elements; future work will address that hypothesis.

Based on our bioinformatic analyses, we cannot prove that the structure:smRNA relationships are caused by the formation and processing of dsRNA substrates by DCL-like mechanisms. Arguably the most-straightforward way to do so would be to map smRNA libraries from maize mutants lacking *Dicer-like* functions. Unfortunately, we found no such librarieswe did map the available libraries from maize RdDM mutants: *mediator of paramutation1 (mop1)* and *required to maintain repression2 (rmr2)* (Gent et al., 2014; Barbour et al., 2012). These mutants affect the repression of TEs that have already been silenced (Barbour et al., 2012); they are thus not particularly good candidates to test the dsRNA-substrate model. We nonetheless assessed the effect of mutants on skew by comparing mutant smRNAs to WT individuals from

the same study (**Fig S16**), but we did not observe any clear or consistent patterns across smRNA lengths or TE superfamilies. These comparisons relied on single libraries and are thus more subject to sampling variability than our other observations, which were based on joint consideration of dozens of smRNA libraries.

Since we cannot prove that processing of dsRNA substrates is a causal mechanism, it is worth considering alternative explanations. For example, structure:smRNA correlations could reflect abundance rather than production; one way this could occur is if smRNAs generated from miRNA-like regions degrade less quickly. It is hard to imagine how this might happen, but it is known that smRNAs that are loaded onto AGO have biases (Mi et al., 2008) and thus some may be more stable with longer half-lives. Another possibility is that these structures correlate with degradation through other, non-DCL pathways. Some studies have attempted to correct for degradation and other effects by focusing only on genomic regions where the proportion of 21, 22 and 24 nt smRNAs exceed an arbitrary threshold compared to smRNAs of all lengths (Lundardon et al., 2020). We did not apply such a threshold here, because this approach necessarily assumes that some 21, 22 and 24-nt smRNAs should be ignored as biologically uninformative. We did, however, assess overlaps in genomic positions between the annotated, 21–24-nt siRNA producing loci of Lundardon et al. (2020) and our miRNA-like hairpin structures. Relative to random chance, we found a modest but significant enrichment in overlapping locations in genes and in all TE superfamilies except SINEs and LINEs (Table S5), which generally lack miRNA-like structures (Table 2). We repeated this exercise with a set of annotated small RNA loci that do not produce smRNAS within the canonical 21-24nt length range (Lunardon et al., 2020); these analyses revealed lower enrichment across all features compared to 21-24nt producing loci, no notable enrichment within TEs and a very slight enrichment within mRNAs (Table S5). Altogether, these analyses suggest that a subset of our miRNA-like secondary structures correspond to loci that produce 21–24-nt siRNAs, presumably through DCL-like mechanisms.

We can think of one additional explanation for the association between miRNA-like regions and smRNAs. In *Arabidopsis*, miRNA target sites within mRNAs are significantly less structured than surrounding regions (Li et al., 2012), which may confer accessibility to the endoribonucleases involved in RNAi (Vandivier et al., 2016). This pattern hints that small RNA binding and RNAi could be less effective in structured regions of TEs than in non-structured

regions, as is likely the case in viruses (Gebert et al., 2019). If this is the case, miRNA-like regions of TEs may have evolved to protect those primary sequences from targeting through RNAi-like mechanisms. In this explanation, the regions are first highly targeted by smRNAs and then structure evolves as a component of the evolutionary arms race between TEs and their hosts.

While we cannot document a definitive mechanism, precedence suggests that processing of dsRNA substrates likely contributes to the genome-wide structure:smRNA relationship. If true, then we can add insights about its effects. First, we can estimate the relative amount of smRNAs that are produced via processing of dsRNA substrates compared to other smRNA-generating mechanisms. Across the entire dataset of 373,485 features (Jiao et al., 2017), minMFE explains 10% of the smRNA mapping results for 21-nt smRNAs (**Table 3**), providing a rough estimate for the proportion of smRNAs produced from dsRNA substrates. This value is larger for some metrics within specific feature categories—e.g., Q_{norm} explained 24% of 22-nt smRNA mapping variation in genes and meanMFE explained 21% of 21nt variation for *CACTA*/DTC elements (**Table S2**). On average, across feature categories and smRNA lengths, the summary statistics minMFE, meanMFE and Q_{norm} explained 8% of mapping variation between miRNA-like regions and non-miRNA-like regions (**Table S2**). These low but highly significant values are consistent with the fact that dsRNAs are only one of several routes to smRNA production (Carthew & Sontheimer, 2009).

Second, our data show that miRNA-like regions are associated with peaks of elevated methylation (**Fig 4**). Since siRNAs guide DNA methylation mechanisms (Law and Jacobsen, 2010), these peaks likely reflect causal relationships among structure, smRNAs and methylation. It is especially notable that these peaks are elevated for CHH methylation, which is deposited *de novo* each generation and thus represents active methylation mechanisms (Law and Jacobsen, 2010). Methylation in these peaks is also elevated in other contexts—e.g., the CG context (**Fig. 4**)—such that the peaks resemble mCHH islands. mCHH islands are short (~100 bp) regions of elevated methylation typically found both up- and downstream of genes. They were first identified in rice as associated with MITEs (Zemach et al. 2010). In maize, mCHH islands are associated with several TE types, found near roughly half of genes, and enriched near highly expressed genes (Gent et al. 2013; Li et al, 2015; Martin et al., 2021). It is not yet known if mCHH islands typically correspond to miRNA-like secondary structures, but it is a fitting topic

for future investigations that may shed further insights into this mysterious epigenetic phenomenon.

TE superfamilies vary in the number and pattern miRNA-like regions

Our work was motivated, in part, by a lack of knowledge about the incipient stages of plant host recognition that leads to TE silencing (Bousios and Gaut, 2016). Since processing of dsRNA substrates remains the only recognized pathway to *de novo* smRNA production (Hung and Slotkin, 2021), we had hoped that characterizing miRNA-like regions would provide clues into properties of host recognition across specific TE superfamilies. Our work does not inform this mystery, except to show that *most* annotated TEs have some miRNA-like regions and also to provide a snapshot of variation across TE superfamilies. That snapshot shows that DNA elements generally have less evidence for miRNA-like structures than LTR elements (**Fig. 1**), but non-LTR RNA elements (LINEs and SINEs) contain almost no miRNA-like structures (**Table 2**). There is also marked variation among LTRs, because *Copia*/RLC exhibit a concentration of secondary structures in the LTRs, but *Ty3*/RLG do not show a similar locational bias (**Fig. 2**). Finally, *Helitrons*/DHH warrant separate mention because 84% are RF-structured, with a strong bias of LP-hairpins at the 3' end (**Fig. 2**). The lowMFE regions of *Helitrons*/DHH often contain the same palindrome sequence that defines structured regions of *Copia*/RLC elements (Bousios et al., 2016).

One cannot help but wonder why miRNA-like regions are common within TEs. If secondary structure can lead to the potential for host recognition through smRNAs, there should be selective pressure to lose structure. We suspect that there is a cost to loss related to function. In Sireviruses (the principal representative of the *Copia*/RLC superfamily), there is evidence that palindromic motifs define the *cis*-regulatory region of the LTR (Grandbastien et al., 2015). In fact, studies of different TE families in different organisms have revealed that *cis*-regulatory regions are often arranged as arrays of complex, sometimes palindromic, repeats (Vernhettes et al., 1998; Araujo et al., 2001; Fablet et al., 2007; Ianc et al., 2014; Martinez et al., 2016), implying that secondary structures often assume a *cis*-regulatory function. We hypothesize that *Copia*/RLC elements are engaged in a tug-of-war between the functional necessities of secondary structure and the tendency of these same regions to act as templates for smRNAs . We presume similar dynamics apply to other TE superfamilies, although clearly this conjecture

requires further detailed analyses of structure and function in specific TEs. However, the location differences between *Copia*/RLC and *Ty3/RLG* are interesting in this context (**Fig. 2**), because it superficially suggests that *cis*-regulation modules for *Ty3/RLG* elements have either moved or have modified function. Another potential function for miRNA-like regions relates to the fact that retrotransposons and autonomous DNA transposons need to co-opt the host's translation machinery to extend their life-cyle. miRNA-like structures may be as crucial for translation for TE transcripts as it is for genes (see below).

Genes: evidence for a trade-off

Our analyses have uncovered a few unexpected features of genes. One is that the two methods provide different insights. The RNAfold approach identifies 85% of genes as RF-structured (**Table 2**), with an evident bias toward 5' UTR regions (**Fig. 2**). This result is not unexpected, given that secondary structures in 5' UTRs are tied to crucial functions in ribosome binding and translation (Babendure et al., 2006; Matoulkova et al., 2012). In contrast, LP-hairpins are primarily found in introns. We conclude that 5' UTRs commonly have miRNA-like regions (as defined by MFEs) but apparently lack the stem-loop structures identified by LinearPartition. Nonetheless, both lowMFE regions and LP-hairpins associate positively with smRNAs and demonstrate elevated CHH methylation levels within genes (**Figs. 3,4 & S11**).

This is not the first such observation for plant genes, because Li et al. (2012) discovered that *Arabidopsis* mRNA transcripts with more stable secondary structures had higher smRNA expression and lower genic expression. Our work expands this previous work in two ways. First, we have extended the observations to maize; it is notable that genes in maize and *Arabidopsis* share these trends, because maize has a larger genome with more TEs. Second, we have shown that secondary structure does not universally correlate negatively with gene expression. Rather, the relationship is tiered: there is a qualitative difference in expression between genes with and without RF-structure (**Fig 4A,B**), probably reflecting that secondary structure in 5' UTRs is crucial for some aspects of gene function. Among genes with RF-structure, however, genes with strong structure (as measured by minMFE) tend to be less expressed than genes with moderate RF-structure (**Fig. 5B**). That is, genes with particularly strong secondary structures (i.e., very low MFEs) have lower expression.

This relationship suggests that there can be "too much of a good thing" when it comes to miRNA-like structures. The potential functional consequence of "too much" is illustrated across the NAM parental genotypes, because structured genes with higher coefficients of variation tend to map more smRNAs (**Fig. 5B**) and have more variable expression among genotypes (**Fig. 5C**). We investigated whether this observation could be explained by other features of the miRNA-like regions, such as especially high variability in chromatin accessibility. We also investigated SNPs and SVs, because some work has shown that structured regions can have higher mutation rates (Hoede et al., 2006). Unfortunately, none of these variables have provided insights that explain higher expression variation across genotypes. In fact, the miRNA-like regions tend to have fewer SNPs and SVs than the rest of the gene (**Fig. 5E**), suggesting that the miRNA-like regions are under purifying selection.

Altogether, these results suggest the possibility of an evolutionary tradeoff between selection for stable secondary structure against too much secondary structure. Even so, we are still left by a paradox: if genes have miRNA-like regions that serve as a template for smRNA production, why are they not silenced? We do not have the answer, but we believe it must rely on the bevy of differences between hetero- and euchromatin. It is known, for example, that genic regions have distinct sets of chromatin markers relative to heterochromatin and also that demethylases like *Increased in Bonsai Methylation 1 (IBM1)* and *repressor of silencing 1* (ROS1) (Gong et al., 2002; Penterman et al., 2007) actively demethylate expressed genes (Saze et al. 2008; Miura et al. 2009). Some aspects of genic methylation are under selection (Muyle et al., 2022), and selection will be particularly strong against mechanisms that silence genic regions. We hypothesize that these mechanisms have evolved in part to counter the potentially deleterious effects of the formation of dsRNA structures and subsequent production of smRNAs.

Overall, we have created a catalog of miRNA-like structures across many features of the maize genome. Our catalog shows that miRNA-like secondary structures are common. These regions also correlate weakly, but highly significantly, with smRNA abundance, and they associate visibly with DNA methylation, especially in the CHH context. Finally, we tentatively suggest that the dynamics of gene expression are affected by these structures and their epigenetic associations. We hope this work sparks further exploration of the roles of secondary structure in plant genome evolution, because it raises questions about unstudied TE categories (e.g., MITEs), about the strength of population genetic evidence against mutations in miRNA-like regions

(Ferrero-Serrano et al., 2022), whether secondary structure characteristics are conserved among species, and whether miRNA-like regions contribute to the previously documented relationship between secondary structure and stress response (Zhang et al., 2018).

METHODS

B73 annotation and secondary structure prediction

Version 4 of the B73 maize genome and version 4.39 of the genome annotation were downloaded from Gramene (www.gramene.org). B73 TE annotations were retrieved from https://mcstitzer.github.io/maize_TEs/ (Jiao et al., 2017; Stitzer et al., 2021). TE and gene annotations were cleaned for redundancy (e.g., the same feature annotated by different annotation authorities) using custom scripts, and separated into annotation files for different feature categories. BED files were then generated for each annotation feature, with a standardized naming convention for each feature: Feature Type::Chromosome:Start Position-End Position (e.g., exon::Chr1:47261-47045).

FASTA files for each feature were generated using BEDtools v2.27 (Quinlan & Hall 2010) getFASTA. These FASTA files were divided into 110 nucleotide sliding windows (1-nt step size) for use in the secondary structure prediction program RNAfold v2.4.9 from ViennaRNA (Lorenz et al., 2011). MFE calculations per window were extracted from RNAfold predictions using a Python script, and the MFE summary metrics (minMFE and meanMFE) were calculated for each feature, based on all windows in that feature. As described in the main text, minMFE was calculated as the lowest MFE window in the feature and meanMFE was the mean of all 110 bp window MFE values. The partition function, *Q*, was calculated by LinearPartition. Q_{norm} was calculated by dividing *Q* by the length of each feature in R. BED files representing regions of lowMFE were created by combining all overlapping windows of <-40 kcal/mol MFE. Overlapping MFE windows were converted to BED format using an inhouse Python script. The scripts used for MFE calculations and analyses are available on GitHub (https://github.com/GautLab/maize_te_structure).

To determine whether a feature contained significant structure, the feature sequence was randomized by shuffling the position of nucleotides across the length of the feature. This approach maintained the GC content of the feature but not the primary sequence. Randomized sequences were then subjected to identical MFE calculations—i.e., they were split into 110 bp

windows for RNAfold prediction. This process was repeated five times for each feature, and the minMFE of each randomization was recorded. The significance of observed structure vs the five randomizations was assigned using a Wilcoxon one-sided test with Benjamini-Hochberg correction in R.

For plotting the location of lowMFE regions across features (**Figs 2 & S4**), we split each feature into 100 equally-sized bins across the length of the feature from 5' to 3' end and counted the number of < -40 kcal/mol regions overlapping each bin. To find motifs in lowMFE regions of different feature types, BED files from concatenated low MFE regions were extracted using BEDtools v2.27 getFASTA. These FASTA files were fed into the MEME motif finder (v5.4.0)(Bailey & Elkan 1994) with the DNA alphabet in Classic mode (i.e., enrichment of sequences in a single reference sequence and no control sequence) for each feature category. We selected the top 10 overrepresented sequences.

Separately, we used LinearPartition v1.0 (Zhang et al., 2020) to annotate miRNA-like regions in each feature. We extracted the sequence of each feature using BED tools getFASTA and ran LinearPartition with default arguments on each sequence. The base-pairing probability files generated by LinearPartition contain estimated pairing probabilities for each pair of likely-pairing positions. We used these probabilities to infer the locations of miRNA-like hairpins by searching for consecutive runs of likely pairing bases in R using functions from the IRanges and GenomicRanges (Lawrence et al., 2013), data.table (Dowle & Srinivasan, 2023), and tidyverse (Wickham et al., 2019) packages. We focused on bases with >0.90 pairing probabilities and search for evidence of miRNA-lik hairpin structure based on the criteria of Axtell and Meyers (2018). Specifically, we required LP-hairpins to be ≥21-nt long with <5 mismatched nucleotides (<3 of mismatches in asymmetric bulges). We did not place an upper limit on the length of predicted LP-hairpins, because we sought to find genomic regions with folding potentials equal to or greater than known miRNAs.

Small RNA Library Analysis

Small RNA-seq libraries were downloaded using NCBI SRA tools and SRAExplorer (https://github.com/ewels/sra-explorer) from the sources indicated in **Table S1**. Adapters, regions with low quality, and low quality reads were trimmed from small RNA RNA-seq libraries using FastQC and cutadapt v0.39 (Bolger et al., 2014). Adapter sequences varied among

libraries, and so were identified and validated in each library using a custom bash script that searched for sets of known maize smRNAs of each length (21–24 nt) in each unprocessed library and confirmed the identity of the adapter sequence connected to each known smRNA sequence. The list of adapters derived for each library is included in **Table S6**. Trimmed reads were then filtered and split based on size matching 21, 22 and 24 nucleotides in length, creating three FASTQ files for each library. We identified the unique smRNA sequences, which we refer to as 'species', following previous methods (Bousios et al., 2016, 2017).

smRNA species were mapped using Bowtie 2 v2.4.2 (Langmead & Salzberg 2012) to the B73 genome, preserving only perfect alignments. SAM tools v1.10 (Danecek et al., 2021) was used to convert and sort the alignment output. BED tools bamtoBED was used to convert the sorted BAM file to BED files. smRNAs from each library were mapped separately for all three lengths, generating a total of 72 (3 sizes × 24 libraries) alignment files. Both uniquely and non-uniquely mapping smRNAs were used to calculate the number of smRNA species corresponding to each genomic locus (Bousios et al., 2017), and strand was not taken into account. Thus, any given position in the genome can be overlapped by several smRNA species, up to two-times the length of the smRNA size class in question (21, 22, or 24).

Bedtools was used to find intersections and coverage counts (per nucleotide) between the smRNA alignment BED files for each library and the MFE region bed files. Subsequently, the smRNA alignment BED files were split into two categories: alignments that intersected low (<-40 kcal/mol) MFE regions and those that did not. Coverage and count files were subsequently generated that contained information of how many smRNA species aligned at each nucleotide, and coverage files contained a normalized count per nucleotide for classification. Normalization was performed by summing the counts and dividing by the length of the region in nucleotides.

For correlations between smRNA species density vs. MFE measurements of features (**Table 3**), linear models of smRNA species per nucleotide as a function of secondary structure metrics (minMFE, meanMFE, etc) were fitted using the base R (v4.1.0) lm() function. To fit these models, smRNA species were summed across all 24 libraries for each feature so that observed smRNA species had an equal weight across libraries. These linear models can be expressed as:

$$log(\text{smRNA counts per kb across feature} + 1) \sim \text{MFE metric}$$

To test the significance of differences in smRNA species density between high and low MFE regions within features, mixed effects models were fit for each smRNA size class using the R package *lme4* (Bates et al., 2015). In these models, smRNA mapping counts from each library were not combined, meaning that each smRNA library:feature pair was counted individually. These mixed effects models can be expressed as:

log(smRNA counts per kb across region + 1) ~ structure designation + (1|feature)

Skew measurements (**Fig 4**) were calculated separately for each TE superfamily and genes as

$$\frac{lowMFE\left(\frac{species}{nt}\right) - highMFE\left(\frac{species}{nt}\right)}{lowMFE + highMFE\left(\frac{species}{nt}\right)}$$

For these calculations, feature-library pairs with zero smRNA species in either non-structured or structured regions were removed from each dataset. We further tested skew differences from zero using Wilcoxon one-sided tests in R.

Autonomous vs non-autonomous designations for TEs were defined differently depending on TE type, but they were determined based on the presence or absence of open reading frames within the TEs, as identified by Stitzer et al. 2021 (downloaded from https://github.com/mcstitzer/maize_genomic_ecosystem). TIRs were considered autonomous if they contained sequence homology to a transposase, and helitrons were considered autonomous if they contained *Rep*/*Hel*, as per Stitzer et al. (2021).

Methylation analyses

Pre-processed B73 genome-wide methylation data from Hufford et al. (2021) were downloaded from https://datacommons.cyverse.org/browse/iplant/home/shared/NAM/NAM_genome_and_annotation_Jan2021_release/DNA_METHYLATION_UMRs/DNA_methylation_coverage_bigwig_files/NAM_methylation_coverage_on_B73v5_coordinates. These data originated from enzymatic methyl-seq (EM-seq) and were mapped against the B73 V5 reference. For this analysis,

coordinates of miRNA-like regions annotated using the B73 V4 reference genome were converted to the V5 reference using the EnsemblPlants CrossMap (v0.6.4) converter.

The methylation data were downloaded as bigW ig files; we converted these data to genome-wide coverage files by multiplying EM-seq coverage at each cytosine position by proportion of methylated and unmethylated reads at each position (yielding, for each cytosine, a number of methylated and unmethylated reads at that position). For each region with miRNA-like structure, we calculated the weighted methylation level for each cytosine sequence context (CG or CHH) by dividing the number of methylation-supporting mapped cytosines by the total number of cytosines in the reference within that region (see Schultz et al., 2012). To find random control regions for comparison, we separated nucleotide positions in each feature into two groups: those that fell within miRNA-like regions and those that did not. For each miRNA-like region in each feature, we randomly assigned a region of equal size to that miRNA-like region but which did not overlap with the miRNA-like region. We did not consider methylation of miRNA-like regions in features where over half of the features fell within miRNA-like regions, because control regions could not be determined by this method.

B73 RNA-seq analyses

B73 gene expression data were downloaded from the ATLAS expression database (www.ebi.ac.uk/gxa/) in transcripts per million (TPM) based on RNA-seq data from 23 maize tissues (E-GEOD-50191)(Walley et al., 2016). The statistical significance of differences between expression of genes in different structure classifications was determined using unpaired *t*-tests between structured and unstructured genes, implemented with t.test() in R. Linear models of expression versus each measurement of secondary structure were separately fit for expression in each tissue type with lm() in R and graphed using ggplot2 (Wickham, 2016). These linear models can be expressed as:

$$Log(Gene\ expression\ +1) \sim MFE\ metric$$

For each of the downstream analyses, we focused on genes with *Sorghum bicolor* syntelogs. We relied on a list of syntelogs in Table S10 of Muyle et al. (2021).

Comparative analyses among NAM founders

Expression, ATAC-seq, SNP data and SV data for each NAM line were downloaded with B73 coordinates from CyVerse at https://datacommons.cyverse.org/browse/iplant/home/shared/NAM/NAM_genome_and_annotation_Jan2021_release (Hufford et al., 2021). Secondary structure predictions were performed in B73 assembly V4, so gene IDs were converted to V5 using the EnsemblPlants ID History Converter web tool (https://plants.ensembl.org/Zea_mays/Tools/IDMapper). Coordinates of TEs and structured regions were converted using the EnsemblPlants CrossMap (v0.6.4) converter with the B73_RefGen_v4 to Zm-B73-REFERENCE-NAM-5.0 parameter. Only genes shared across all lines were included.

Normalized expression data were downloaded in RPKM format from merged RNA-seq libraries from CyVerse at https://datacommons.cyverse.org/browse/iplant/home/shared/NAM/NAM_genome_and_annotation_Jan2021_release/SUPPLEMENTAL_DATA/pangene-files. Only data from genes shared among all lines (as determined by Hufford et al.) were included. These data include RNA-seq normalized across eight tissues in each line: primary root and coleoptile at six days after planting, base of the 10th leaf, middle of the 10th leaf, tip of the 10th leaf at the Vegetative 11 growth stage, meiotic tassel and immature ear at the V18 growth stage, anthers at the Reproductive 1 growth stage. Details for how these data were normalized can be found in Hufford et al., (2021).

The coefficient of variation (CV) of expression was calculated for each gene between the 26 lines using the normalized RPKM expression data from Hufford et al. (2021). For each gene, CV was defined as the standard deviation of its expression across lines divided by its mean normalized across lines. We calculated CV using the sd() and mean() functions in base R. We plotted CVs between categories of structure (RF-structured and RF-unstructured) using ggplot2 (Wickham 2016) and determined statistical significance of differences between categories using unpaired *t*-tests in R. We measured these differences in two different ways: first, using all genes and, second, removing genes with CV = 0 (920 genes, 3.3% of genes). We also built a linear model with lm() in R to correlate the magnitude of gene expression in B73 with the CV of that gene across lines. This linear model can be expressed as:

$$log(B73\ expression + 1) \sim NAM\ line\ CV$$

We also measured epigenetic and genetic features across the NAM lines, and tracked their overlap with miRNA-like regions. For the former, we concatenated ACRs that overlapped positions between lines, producing a set of merged ACRs. We produced these merged sets using the R libraries IRanges and GenomicRanges (Lawrence et al., 2013). We extracted the positions of SNPs from the filtered VCF file from Hufford et al. (2021). The expected overlap was calculated as the proportional of genic space taken up by low MFE regions * the total length of features. We assessed overlap between ACRs/SVs/SNPs and miRNA-like regions using GenomicRanges in R. Custom scripts for these analyses can be found at https://github.com/GautLab/maize_te_structure, and additional supplementary files can be found at https://figshare.com/projects/siRNAs_and_secondary_structure_in_maize_genes_and_TEs/150714.

COMPETING INTERESTS STATEMENT

The authors have no competing interests.

ACKNOWLEDGEMENTS

This work was supported by NSF grant to B.S.G. and by Royal Society awards UF160222, URF\R\221024, RGF/ R1/180006 to A.B. A.M. was supported by the Human Frontier Science Program (HFSP) Fellowship LT000496/2018-L.

Author Contributions: GTM, ES, AM, AB and BSG designed the research questions. RNAfold analyses were performed by ES and J G-M, with ES contributing new computational tools; smRNA mapping was also performed by ES. GTM devised and performed the LinearPartition analyses. GTM, ES, AM and BSG performed statistical analyses of the results. GTM, AB and BSG wrote the paper; BSG supervised the work.

FIGURE LEGENDS

Figure 1: Characteristics of miRNA-like secondary structure across two methods. (**A**) A schematic contrasting the two prediction methods for a genic region on C hromosome 2. The LinearPartition (LP) method focuses on identifying small regions with hairpin characteristics, while the RNAfold method focuses on regions with low Minimum Free Energy (MFE). This example illustrates lowMFE regions in red, with overlapping LP-hairpins in blue. Note that lowMFE regions exceed 110 bp, because they represent the concatenation of overlapping windows with MFE < -40 kcal/mol. (**B**) The correlation between meanMFE and Q_{norm} based on 39,179 genes. (**C**) The distributions of three summary statistics—minMFE, meanMFE and Q_{norm} —across seven feature categories. In the key, helitrons correspond to DHH elements (see **Table 2** for the three letter designations); LTRs consist of RLC, RLG and RLX; LINEs are the RIL and RIT elements; SINEs are RST; and terminal repeat elements consist of DTA, DTC, DTH, DTM, and DTT elements.

Figure 2. Landscapes of miRNA-like regions across feature types. Each row represents a metaprofile that combines data from all members of each feature type, based on structured members. Features were divided into 100 equally sized bins from the 5' end to the 3' end. The left column shows the number of features with lowMFE (<-40 kcal/mol) windows, while the right column shows the number of features with LP hairpins. A peak in the landscape represents a region that commonly contained miRNA-like structures. All panels share the same x-axis, which is represented proportionally across the length of features, from 0.00 (5' end) to 1.00 (3' end). This figure shows these locations for a subset of the 15 categories in Table 2; the remainder of the categories are shown in Figure S4.

Figure 3. The distribution of skew for smRNA mapping in different feature categories. Skew is presented on the x-axis. Height on the y-axis represents the Gaussian estimated kernel density of skew values. Skew measures the relative enrichments of smRNAs in miRNA-like regions compared to non-miRNA regions and ranges from 1.0 (enrichment in miRNA-like regions) to -1.0 (enrichment in non-miRNA-like regions. All panels use the same x-axis. The dotted vertical line represents zero where smRNA density is not skewed to either low or high MFE regions. A. Skew for retrotransposons for 21, 22 and 24-nt smRNAs, separately for Copia (RLC), Ty3

(RLG) and unknown retrotransposons (RLX). B. Skew for DNA transposons, with names for the three letter codes provided in Table 2. The dashed lines represent skew for putatively autonomous elements, while solid lines represent non-autonomous elements. C. Skew measured in genes. These graphs are based on LP-hairpins, but analogous for lowMFE regions and all feature categories are presented in Figure S7.

Figure 4. Methylation at LP-hairpins. The left column shows methylation in the CG context (mCG) and the right shows methylation in the CHH context (mCHH). Each row represents a different feature type. The blue lines summarize the patterns of methylation in the hairpin (variable sizes, median = 25 nt) across all hairpins in a given feature type (e.g., all TIR hairpins, gene hairpins, etc.) and their flanking regions, divided into 40 nonoverlapping 100 bp windows. We assigned a control window to each hairpin in the dataset by choosing a random window of the same size as the hairpin within the same element. The red line corresponds to methylation patterns around these randomized control loci.

Figure 5. Expression between structured and unstructured genes, as defined by RNAfold analysis, in B73. The expression data are based on combined data across 23 tissues. **A.** Difference in the overall magnitude of expression in all structured (n=27,034) vs unstructured (n=5054) genes and in structured vs. unstructured genes with a syntelog in *S. bicolor*. The box plots report the range of the middle quartiles, whiskers report the range, and lines represent the median. **B.** Expression as a function of minMFE for structured (dashed line) and unstructured genes with a *S. bicolor* syntelog (solid line). Both lines report the linear regression; both slopes are highly significant, as indicated by P-values on the figure. **C.** The coefficient of variation (CV) of gene expression across the 26 NAM parents compared between structured vs unstructured genes with a *S. bicolor* syntelog. The two categories differ significantly ($P < 2.22 \times 10^{-16}$). The graph also reports CV among B73 tissues, which does not differ significantly between structured and unstructured genes ($P = 0.32$). **D.** smRNA mapping to structured and unstructured genes and for three smRNA lengths. For all three lengths, the difference is significant ($P < 2.22 \times 10^{-16}$). The violin plots show the distributions of smRNA counts, and the boxplots are formatted the same as in (A.) **E.** Epigenetic and genetic features in lowMFE regions of genes. The plots plot the number of expected and observed features overlapping (or not-

overlapping) the lowMFE region. For example, the number of ACRs (left graph) overlapping lowMFE regions is very similar to the number expected, based on the distributions along genes. In contrast, the numbers of observed SVs (middle) and SNPs (right) are highly underrepresented in lowMFE regions.

REFERENCES

Ahmed, I., Sarazin, A., Bowler, C., Colot, V., and Quesneville, H. 2011. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. Nucleic Acids Res 39, 6919–6931. https://doi.org/10.1093/nar/gkr324.

Araujo, P.G., Casacuberta, J.M., Costa, A.P., Hashimoto, R.Y., Grandbastien, M.A., and Van Sluys, M.A. 2001. Retrolyc1 subfamilies defined by different U3 LTR regulatory regions in the Lycopersicon genus. Mol Genet Genomics 266, 35–41. https://doi.org/10.1007/s004380100514.

Axtell, M.J. 2013. Classification and comparison of small RNAs from plants. Annu. Rev. Plant Biol. *64*, 137–159.

Axtell, M.J., and Meyers, B.C. 2018. Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. The Plant Cell 30, 272–284. 10.1105/tpc.17.00851.

Babendure, J.R., Babendure, J.L., Ding, J.-H., and Tsien, R.Y. 2006. Control of mammalian translation by mRNA structure near caps. RNA *12*, 851–861.

Bailey, T.L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2, 28–36.

Barbour, J. E. R., Liao, I. T., Stonaker, J. L., Lim, J. P., Lee, C. C., Parkinson, S. E., ... & Hollick, J. B. 2012. required to maintain repression2 is a novel protein that facilitates locus-specific paramutation in maize. *The Plant Cell*, *24*(5), 1761-1775.

Bates, D., Mächler, M., Bolker, B., and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67, 1–48. https://doi.org/10.18637/jss.v067.i01.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.-M., Westerman, R.P., SanMiguel, P.J., and Bennetzen, J.L. 2009. Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. PLOS Genet. *5*, e1000732.

Baulcombe, D. 2004. RNA silencing in plants. Nature 431, 356–363. https://doi.org/10.1038/nature02874.

Bureau, T. E., & Wessler, S. R. 1992. Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *The Plant Cell*, *4*(10), 1283-1294.

Bureau, T. E., & Wessler, S. R. 1994. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *The Plant Cell*, *6*(6), 907-916.

Bevilacqua, P.C., Ritchey, L.E., Su, Z., and Assmann, S.M. 2016. Genome-Wide Analysis of RNA Secondary Structure. Annu Rev Genet 50, 235–266. 10.1146/annurev-genet-120215-035034.

Borges, F., and Martienssen, R.A. 2015. The expanding world of small RNAs in plants. Nat. Rev. Mol. Cell Biol. *16*, 727–741.

Bousios, A., and Gaut, B.S. 2016. Mechanistic and evolutionary questions about epigenetic conflicts between transposable elements and their plant hosts. Curr. Opin. Plant Biol. *30*, 123–133.

Bousios, A., Gaut, B.S., and Darzentas, N. 2017. Considerations and complications of mapping small RNA high-throughput data to transposable elements. Mobile DNA 8, 3. https://doi.org/10.1186/s13100-017-0086-z.

Bousios, A., Diez, C.M., Takuno, S., Bystry, V., Darzentas, N., and Gaut, B.S. 2016. A role for palindromic structures in the cis-region of maize Sirevirus LTRs in transposable element

evolution and host epigenetic response. Genome Res. *26*, 226–237.
Bousios, A., Kourmpetis, Y.A.I., Pavlidis, P., Minga, E., Tsaftaris, A., and Darzentas, N. 2012. The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story. The Plant Journal 69, 475–488. https://doi.org/10.1111/j.1365-313X.2011.04806.x.
Bullock, S.L., Ringel, I., Ish-Horowicz, D., and Lukavsky, P.J. 2010. A′-form RNA helices are required for cytoplasmic mRNA transport in Drosophila. Nat Struct Mol Biol 17, 703–709. https://doi.org/10.1038/nsmb.1813.
Bureau, T.E., and Wessler, S.R. 1992. Tourist: a large family of small inverted repeat elements frequently associated with maize genes. The Plant Cell 4, 1283–1294. 10.1105/tpc.4.10.1283.
Bureau, T.E., and Wessler, S.R. 1994. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6, 907–916. 10.1105/tpc.6.6.907.
Buratti, E., and Baralle, F.E. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol 24, 10505–10514. 10.1128/MCB.24.24.10505-10514.2004.
Carthew, R.W., and Sontheimer, E.J. 2009. Origins and Mechanisms of miRNAs and siRNAs. Cell 136, 642–655. https://doi.org/10.1016/j.cell.2009.01.035.
Choi, J.Y., and Lee, Y.C.G. 2020. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. PLOS Genet. *16*, e1008872.
Creasey, K.M., J. Zhai, F. Borges, F. Van Ex, M. Regulski, B.C. Meyers and R.A. Martienssen. 2014. miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. Nature 508:411-415.
Cruz, C., and Houseley, J. Endogenous RNA interference is driven by copy number. ELife *3*, e01581.
Cuerda-Gil, D., and Slotkin, R.K. 2016. Non-canonical RNA-directed DNA methylation. Nat Plants 2, 16163. https://doi.org/10.1038/nplants.2016.163.
Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008.
Devert, A., Fabre, N., Floris, M., Canard, B., Robaglia, C., and Crété, P. 2015. Primer-dependent and primer-independent initiation of double stranded RNA synthesis by purified Arabidopsis RNA-dependent RNA polymerases RDR2 and RDR6. PloS One *10*, e0120100.
Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505, 696–700. https://doi.org/10.1038/nature12756.
Dowle M, Srinivasan A. 2023. data.table: Extension of 'data.frame'. https://r-datatable.com, https://Rdatatable.gitlab.io/data.table, https://github.com/Rdatatable/data.table.
Eichten, S.R., Ellis, N.A., Makarevitch, I., Yeh, C.-T., Gent, J.I., Guo, L., McGinnis, K.M., Zhang, X., Schnable, P.S., Vaughn, M.W., et al. 2012. Spreading of Heterochromatin Is Limited to Specific Families of Maize Retrotransposons. PLOS Genet. *8*, e1003127.
Fablet, M., Rebollo, R., Biémont, C., and Vieira, C. 2007. The evolution of retrotransposon regulatory regions and its consequences on the Drosophila melanogaster and Homo sapiens host genomes. Gene 390, 84–91. https://doi.org/10.1016/j.gene.2006.08.005.

Ferrero-Serrano, Á., Sylvia, M.M., Forstmeier, P.C., Olson, A.J., Ware, D., Bevilacqua, P.C., and Assmann, S.M. 2022. Experimental demonstration and pan-structurome prediction of climate-associated riboSNitches in Arabidopsis. Genome Biology 23, 101. https://doi.org/10.1186/s13059-022-02656-4.

Fultz, D., and Slotkin, R.K. 2017. Exogenous Transposable Elements Circumvent Identity-Based Silencing, Permitting the Dissection of Expression-Dependent Silencing. The Plant Cell 29, 360–376. https://doi.org/10.1105/tpc.16.00718.

Fukudome, A., and Fukuhara, T. 2017. Plant dicer-like proteins: double-stranded RNA-cleaving enzymes for small RNA biogenesis. J Plant Res 130, 33–44. https://doi.org/10.1007/s10265-016-0877-1.

Gebert, D., Jehn, J., and Rosenkranz, D. 2019. Widespread selection for extremely high and low levels of secondary structure in coding sequences across all domains of life. Open Biology 9, 190020. 10.1098/rsob.190020.

Gent, J.I., Ellis, N.A., Guo, L., Harkess, A.E., Yao, Y., Zhang, X., and Dawe, R.K. 2013. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Res. *23*, 628–637.

Gent, J.I., Madzima, T.F., Bader, R., Kent, M.R., Zhang, X., Stam, M., McGinnis, K.M., and Dawe, R.K. 2014. Accessible DNA and Relative Depletion of H3K9me2 at Maize Loci Undergoing RNA-Directed DNA Methylation. The Plant Cell 26, 4903–4917. 10.1105/tpc.114.130427.

Gong, Z., Morales-Ruiz, T., Ariza, R.R., Roldán-Arjona, T., David, L., and Zhu, J.K. 2002. ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase. Cell 111, 803–814. https://doi.org/10.1016/s0092-86740201133-9.

Grandbastien, M.-A. 2015. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim Biophys Acta 1849, 403–416. https://doi.org/10.1016/j.bbagrm.2014.07.017.

Hoede, C., Denamur, E., and Tenaillon, O. 2006. Selection Acts on DNA Secondary Structures to Decrease Transcriptional Mutagenesis. PLOS Genetics 2, e176. https://doi.org/10.1371/journal.pgen.0020176.

Hollister, J.D., Smith, L.M., Guo, Y.-L., Ott, F., Weigel, D., and Gaut, B.S. 2011. Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proc. Natl. Acad. Sci. U. S. A. *108*, 2322–2327.

Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A., Guo, T., Olson, A., Qiu, Y., et al. 2021. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science 373, 655–662. https://doi.org/10.1126/science.abg5289.

Hung, Y.-H., and Slotkin, R.K. 2021. The initiation of RNA interference (RNAi) in plants. Current Opinion in Plant Biology 61, 102014. 10.1016/j.pbi.2021.102014.

Ianc, B., Ochis, C., Persch, R., Popescu, O., and Damert, A. 2014. Hominoid composite non-LTR retrotransposons-variety, assembly, evolution, and structural determinants of mobilization. Mol Biol Evol 31, 2847–2864. https://doi.org/10.1093/molbev/mst256.

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

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.-S., et al. 2017. Improved maize reference genome with single-molecule

technologies. Nature 546, 524–527. https://doi.org/10.1038/nature22971.
Kapitonov, V.V., and Jurka, J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 23, 521–529. https://doi.org/10.1016/j.tig.2007.08.004.
Langmead, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923.
Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M.S., Strahl, B.D., Patel, D.J., and Jacobsen, S.E. 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. Nature *498*, 385–389.
Law, J.A., and Jacobsen, S.E. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11, 204–220. 10.1038/nrg2719.
Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. 2013. Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology 9, e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
Li, F., Zheng, Q., Vandivier, L.E., Willmann, M.R., Chen, Y., and Gregory, B.D. 2012. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. Plant Cell 24, 4346–4359. https://doi.org/10.1105/tpc.112.104232.
Li, Q., Gent, J.I., Zynda, G., Song, J., Makarevitch, I., Hirsch, C.D., Hirsch, C.N., Dawe, R.K., Madzima, T.F., McGinnis, K.M., et al. 2015. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc. Natl. Acad. Sci. U. S. A. *112*, 14728–14733.
Lippman, Z., Gendrel, A.-V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. Nature *430*, 471–476.
Lisch, D. 2009. Epigenetic Regulation of Transposable Elements in Plants. Annu. Rev. Plant Biol. *60*, 43–66.
Lisch, D., and Slotkin, R.K. 2011. Chapter Three - Strategies for Silencing and Escape: The Ancient Struggle Between Transposable Elements and Their Hosts. In International Review of Cell and Molecular Biology, K.W. Jeon, ed. Academic Press, pp. 119–152.
Liu, J., He, Y., Amasino, R., and Chen, X. 2004. siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. Genes Dev. *18*, 2873–2878.
Liu, P., Cuerda-Gil, D., Shahid, S., & Slotkin, R. K. 2022. The Epigenetic Control of the Transposable Element Life Cycle in Plant Genomes and Beyond. *Annual Review of Genetics*, *56*, 63-87.
Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. 2011. ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.
Lunardon, A. , Johnson , N.R., Hagerott, E., Phifer, T. , Polydore , S., Coruh, C, and Axtell, M.J. 2020. Integrated annotations and analyses of small RNA-producing loci from 47 diverse plants. *Genome Research* 30:497-513.
Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., and Voinnet, O. 2013. Reconstructing de novo silencing of an active plant retrotransposon. Nat. Genet. *45*, 1029–1039.
Martin, G.T., Seymour, D.K., and Gaut, B.S. 2021. CHH Methylation Islands: A Nonconserved Feature of Grass Genomes That Is Positively Associated with Transposable Elements but Negatively Associated with Gene-Body Methylation. Genome Biol. Evol. *13*, evab144.
Martínez, G., Panda, K., Köhler, C., and Slotkin, R.K. 2016. Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. Nat Plants 2, 16030.

https://doi.org/10.1038/nplants.2016.30.
Matoulkova, E., Michalova, E., Vojtesek, B., and Hrstka, R. 2012. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. RNA Biol. *9*, 563–576.
Matzke, M.A., and Mosher, R.A. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat. Rev. Genet. *15*, 394–408.
McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., et al. 2009. Genetic properties of the maize nested association mapping population. Science 325, 737–740. https://doi.org/10.1126/science.1174320.
Miura, A., Nakamura, M., Inagaki, S., Kobayashi, A., Saze, H., and Kakutani, T. 2009. An Arabidopsis jmjC domain protein protects transcribed genes from DNA methylation at CHG sites. The EMBO Journal 28, 1078–1086. 10.1038/emboj.2009.59.
Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., et al. 2022. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature 602, 101–105. https://doi.org/10.1038/s41586-021-04269-6.
Muyle, A., Seymour, D., Darzentas, N., Primetis, E., Gaut, B.S., and Bousios, A. 2021. Gene capture by transposable elements leads to epigenetic conflict in maize. Molecular Plant 14, 237–252. 10.1016/j.molp.2020.11.003.
Muyle, A.M., Seymour, D.K., Lv, Y., Huettel, B., and Gaut, B.S. 2022. Gene Body Methylation in Plants: Mechanisms, Functions, and Important Implications for Understanding Evolutionary Processes. Genome Biology and Evolution 14, evac038. https://doi.org/10.1093/gbe/evac038.
Nuthikattu, S., McCue, A.D., Panda, K., Fultz, D., DeFraia, C., Thomas, E.N., and Slotkin, R.K. 2013. The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. Plant Physiol. *162*, 116–131.
Nussinov, R., and Jacobson, A.B. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A 77, 6309–6313. 10.1073/pnas.77.11.6309.
O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. 2018. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front. Endocrinol. *9*.
Panda, K., McCue, A.D., and Slotkin, R.K. 2020. Arabidopsis RNA Polymerase IV generates 21–22 nucleotide small RNAs that can participate in RNA-directed DNA methylation and may regulate genes. Philos. Trans. R. Soc. B Biol. Sci. *375*, 20190417.
Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S., and Fischer, R.L. 2007. DNA demethylation in the Arabidopsis genome. Proc. Natl. Acad. Sci. *104*, 6752–6757.
R Core Team. 2022. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org.
Ritchey, L.E., Su, Z., Tang, Y., Tack, D.C., Assmann, S.M., and Bevilacqua, P.C. (2017). Structure-seq2: sensitive and accurate genome-wide profiling of RNA structure in vivo. Nucleic Acids Research 45, e135. 10.1093/nar/gkx533.
Saze, H., Sasaki, T., and Kakutani, T. 2008. Negative regulation of DNA methylation in plants. Epigenetics 3, 122–124. 10.4161/epi.3.3.6355.
Schultz, M.D., Schmitz, R.J., and Ecker, J.R. 2012. "Leveling" the playing field for analyses of single-base resolution DNA methylomes. Trends Genet 28, 583–585.

10.1016/j.tig.2012.10.012.
Seligmann, H., and Raoult, D. 2016. Unifying view of stem–loop hairpin RNA as origin of current and ancient parasitic and non-parasitic RNAs, including in giant viruses. Current Opinion in Microbiology 31, 1–8. https://doi.org/10.1016/j.mib.2015.11.004.
Sigman, M.J., and Slotkin, R.K. 2016. The First Rule of Plant Transposable Element Silencing: Location, Location, Location. Plant Cell *28*, 304–313.
Sijen, T., and Plasterk, R.H.A. 2003. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. Nature *426*, 310–314.
Slotkin, R.K., Freeling, M., and Lisch, D. 2003. Mu killer causes the heritable inactivation of the Mutator family of transposable elements in Zea mays. Genetics 165, 781–797. https://doi.org/10.1093/genetics/165.2.781.
Stitzer, M.C., Anderson, S.N., Springer, N.M., and Ross-Ibarra, J. 2021. The genomic ecosystem of transposable elements in maize. PLOS Genetics 17, e1009768. 10.1371/journal.pgen.1009768.
Su, Z., Tang, Y., Ritchey, L.E., Tack, D.C., Zhu, M., Bevilacqua, P.C., and Assmann, S.M. 2018. Genome-wide RNA structurome reprogramming by acute heat shock globally regulates mRNA abundance. Proceedings of the National Academy of Sciences 115, 12170–12175. https://doi.org/10.1073/pnas.1807988115.
Sun, F.-J., Fleurdépine, S., Bousquet-Antonelli, C., Caetano-Anollés, G., and Deragon, J.-M. 2007. Common evolutionary trends for SINE RNA structures. Trends Genet. TIG *23*, 26–33.
Tenaillon, M.I., Hollister, J.D., and Gaut, B.S. 2010. A triptych of the evolution of plant transposable elements. Trends Plant Sci. *15*, 471–478.
Vandivier, L.E., Anderson, S.J., Foley, S.W., and Gregory, B.D. 2016. The conservation and function of RNA secondary structure in plants. Annu Rev Plant Biol 67, 463–488. https://doi.org/10.1146/annurev-arplant-043015-111754.
Vernhettes, S., Grandbastien, M.A., and Casacuberta, J.M. 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. https://doi.org/10.1093/oxfordjournals.molbev.a025988.
Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., et al. 2016. Integration of omic networks in a developmental atlas of maize. Science 353, 814–818. https://doi.org/10.1126/science.aag1125.
Wang, X., Elling, A.A., Li, X., Li, N., Peng, Z., He, G., Sun, H., Qi, Y., Liu, X.S., and Deng, X.W. 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.
Wang, X., Weigel, D., and Smith, L.M. 2013. Transposon Variants and Their Effects on Gene Expression in Arabidopsis. PLOS Genet. *9*, e1003255.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982.
Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes

A, Henry L, Hester J,et al. 2019. “Welcome to the tidyverse.” Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686.
Xiong, W., He, L., Lai, J., Dooner, H.K., and Du, C. 2014. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proceedings of the National Academy of Sciences 111, 10263–10268. https://doi.org/10.1073/pnas.1410068111.
Yang, X., Yang, M., Deng, H., and Ding, Y. 2018. New Era of Studying RNA Secondary Structure and Its Influence on Gene Regulation in Plants. Frontiers in Plant Science 9.
Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328(5980):916–919.
Zhang, H., Gong, Z., and Zhu, J.-K. 2022. Active DNA demethylation in plants: 20 years of discovery and beyond. Journal of Integrative Plant Biology 64, 2217–2239. 10.1111/jipb.13423.
Zhang, H., Zhang, L., Mathews, D.H., and Huang, L. 2020. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. Bioinformatics 36, i258–i267. 10.1093/bioinformatics/btaa460.
Zhang, Q., Arbuckle, J., & Wessler, S. R. 2000. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize. *Proceedings of the National Academy of Sciences*, *97*(3), 1160-1165.
Zhang, X., and Qi, Y. 2019. The Landscape of Copia and Gypsy Retrotransposon During Maize Domestication and Improvement. Front. Plant Sci. *10*.
Zhang, Y., Burkhardt, D.H., Rouskin, S., Li, G.-W., Weissman, J.S., and Gross, C.A. 2018. A Stress Response that Monitors and Regulates mRNA Structure Is Central to Cold Shock Adaptation. Molecular Cell 70, 274-286.e7. 10.1016/j.molcel.2018.02.035.
Zuker, M., and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9, 133–148. 10.1093/nar/9.1.133.

Table 1: Terms defined in the text and that are used to describe and characterize miRNA-like regions.

Term	**Method**	**Explanation**
minMFE	RNAfold	The Minimum Free Energy (MFE) of the 110 bp window with the lowest MFE score within an individual TE or gene sequence
meanMFE	RNAfold	The average estimated MFE across all 110 bp windows in any TE or gene sequence
lowMFE	RNAfold	A region or regions of a TE or gene that is defined by concatenating overlapping windows of MFE< -40/kcal/mol
RF-structured	RNAfold	Designates any TE or gene that has a significantly lower minMFE value than randomized sequences
LP-hairpin	LinearPartition	Putative hairpin structure identified by combing base-pairing probabilities from LinearPartition with miRNA hairpin criteria
Q_{norm}	LinearPartition	The LinearPartition function reports *Q*, a summary of secondary structure across an entire sequence. Q_{norm} adjusts *Q* by the length of the sequence
skew	Both	Measures the relative proportion of distinct smRNAs that map to miRNA-like regions of a sequence compared to the remainder of that sequence. Ranges from -1.0 to 1.0, where 1.0 denotes that smRNAs map only to miRNA-like regions.

Table 2: Fifteen feature categories and accompanying statistics. The statistics include the

number of individual features in each category, based on two annotation versions for TEs, and the percentage of features that have miRNA like structure (structured) based on RNAfold or detectable LP-hairpins.

Feature type	**No**[1]	**RF**[2]	**LP**[3]	**No**[4]	**LP**
Genes	39,179	69.00%	29.82%	39,179	29.82%
mRNA	133,812	64.80%	5.02%	133,812	5.02%
miRNA precursor	107	71.00%	66.36%	107	66.36%
Helitrons/DHH	49,235	84.00%	13.00%	22,339	6.43%
hAT/DTA	5,602	59.60%	4.15%	5,096	4.28%
CACTA/DTC	1,264	79.00%	32.52%	2,768	41.76%
PIF-Harbinger/DTH	4,971	38.80%	17.57%	63,216	6.22%
Mutator/DTM	1,319	60.30%	62.82%	928	57.54%
Tc1-Mariner/DTT	458	43.90%	16.69%	67,533	6.75%
L1 LINE/RIL	36	0.00%	0.00%	477	2.73%
Rte LINE/RIT	29	0.00%	0.00%	296	3.04%
Copia/RLC	45,009	98.20%	58.04%	44,242	55.88%
Ty3/RLG	72,976	88.00%	40.57%	70,165	38.47%
Unclassified-LTR /RLX	18,457	85.90%	38.18%	16,205	32.98%
SINEs/RST	1,031	0.00%	1.74%	892	1.46%

TOTAL[5]	**373,485**	**286,744**	**90,088**	**467,255**	**182,749**

[1] The number of features in each category in the Jiao et al. (2017) annotation

[2] The percentage of RF-structured features in each category, as determined by RNAfold analyses and permutations.

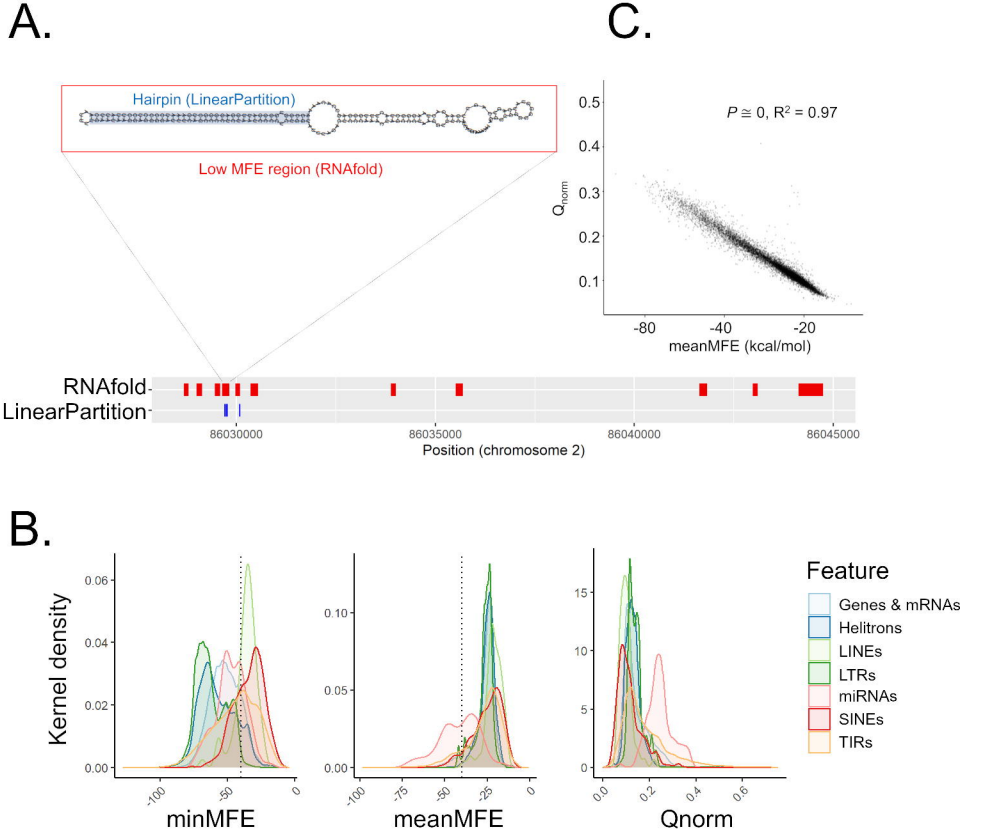
[3] Percentage of features in each category that contained at least one LP-hairpin as inferred from LinearPartition base pairing probabilities and analyses.

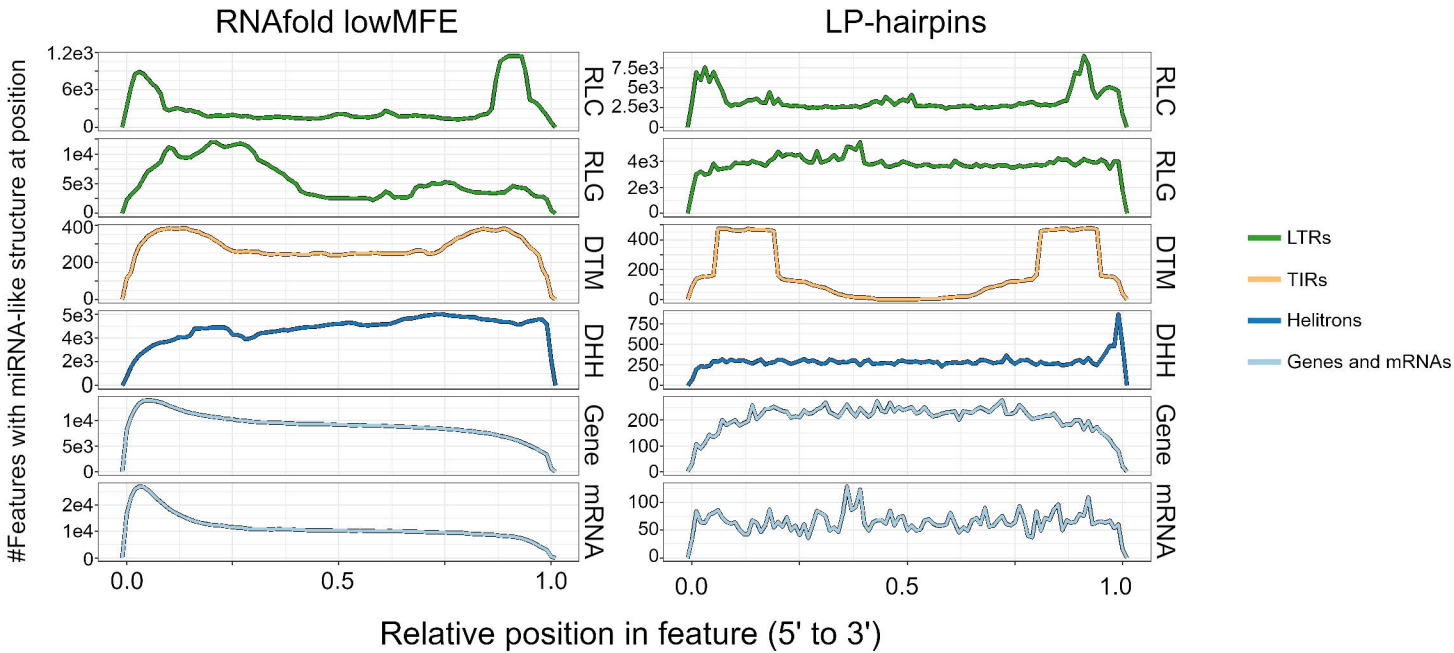
[4] The number of features in each TE superfamily based on the updated annotation by Stitzer et al. (2021).

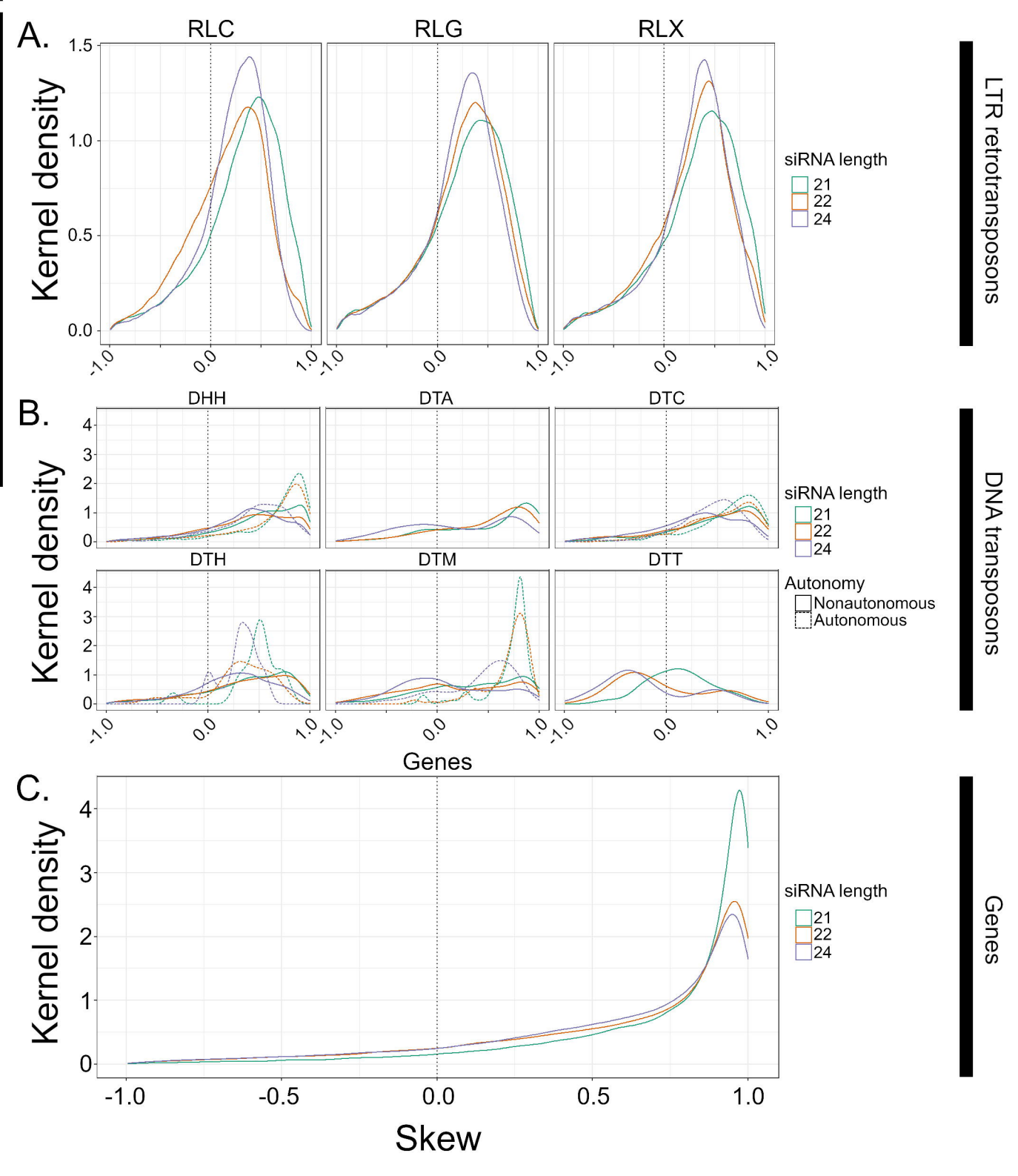
[5] Total refers to the total number (No.) of sequences in each annotation set or it refers to the number of sequences that contain miRNA-like regions based on the RF-structured or LP-hairpin criteria

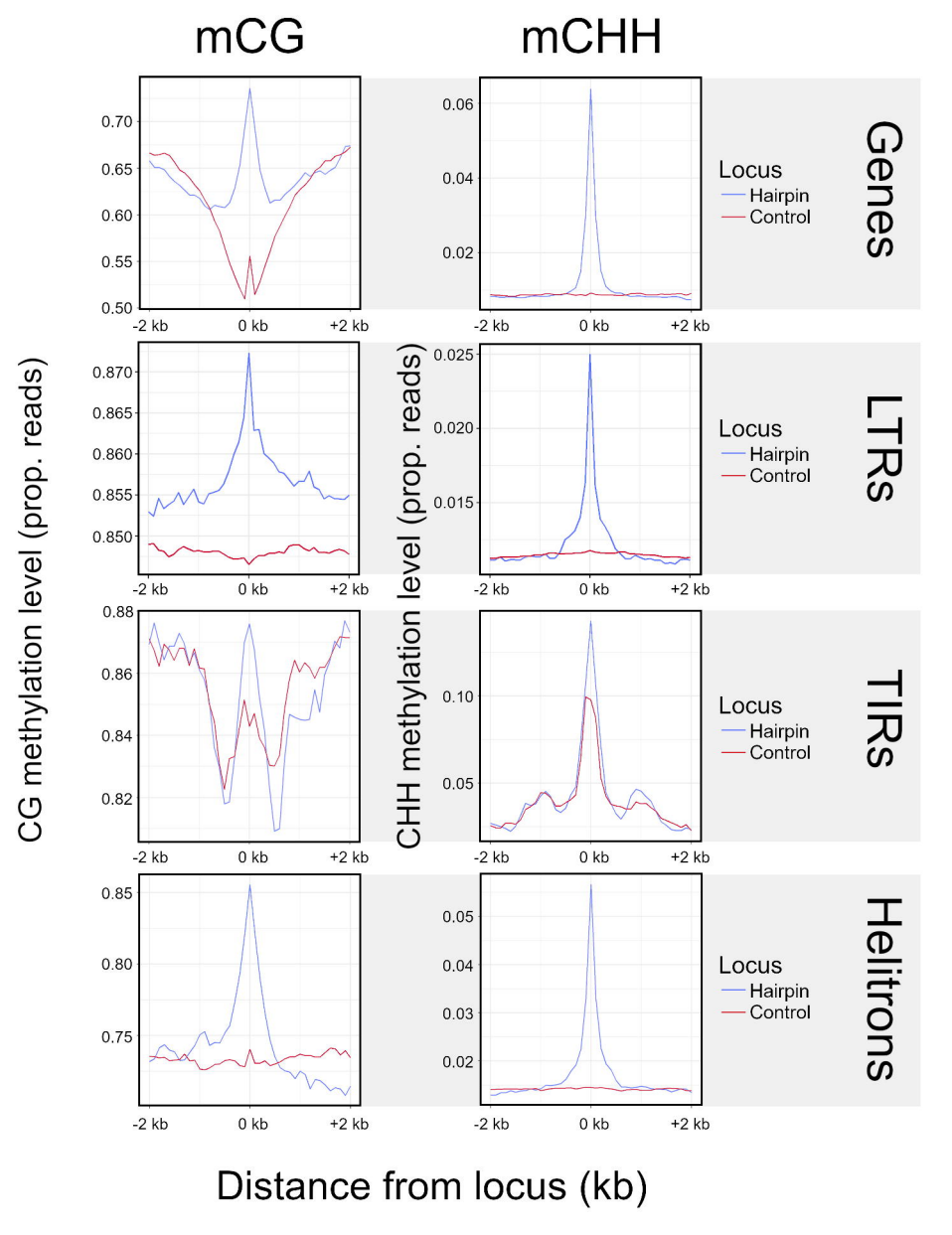
Table 3: Correlation value (with FDR corrected p-value in parentheses) between secondary structure summary statistics and numbers of smRNA species across all 373,485 features.

Summary Metric	**21-nt smRNA**	**22-nt smRNA**	**24-nt smRNA**
minMFE	0.091 (0.00)	0.103 (0.00)	0.074 (0.00)
meanMFE	0.017 (0.00)	8.6×10^{-3} (0.00)	0.004 (5.01×10^{-227})
Q_{norm}	0.101 (0.00)	0.133 (0.00)	0.089 (0.00)

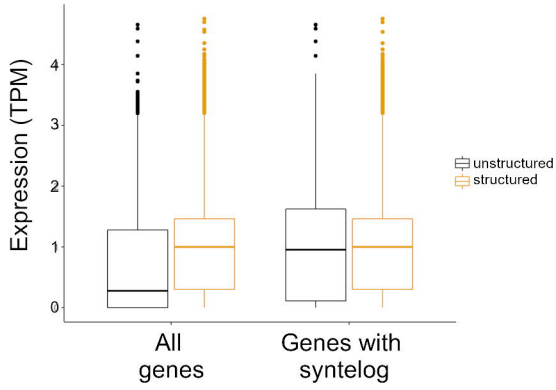

A.
Hairpin (LinearPartition)
Low MFE region (RNAfold)
RNAfold
LinearPartition
86030000
86035000
86040000
86045000
Position (chromosome 2)
C.
P ≅ 0, R² = 0.97
Qnorm
meanMFE (kcal/mol)
B.
Kernel density
minMFE
meanMFE
Qnorm
Feature
Genes & mRNAs
Helitrons
LINEs
LTRs
miRNAs
SINEs
TIRs


RNAfold lowMFE
LP-hairpins
#Features with miRNA-like structure at position
RLC
RLG
DTM
DHH
Gene
mRNA
1.2e3
6e3
0
1e4
5e3
400
200
5e3
4e3
2e3
2e4
7.5e3
2.5e3
750
500
250
100
50
0.0
0.5
1.0
Relative position in feature (5' to 3')
LTRs
TIRs
Helitrons
Genes and mRNAs

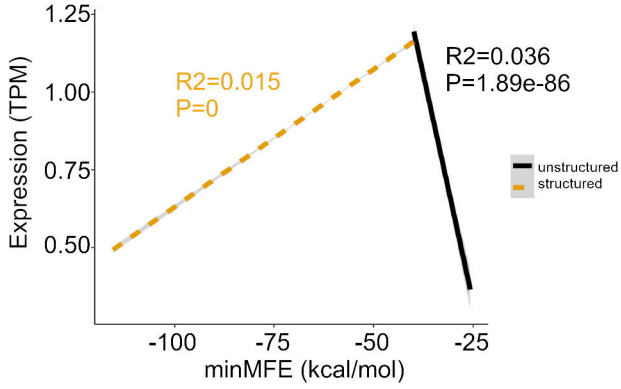

A.
RLC
RLG
RLX
Kernel density
1.5
1.0
0.5
0.0
-1.0
0.0
1.0
siRNA length
21
22
24
LTR retrotransposons
B.
DHH
DTA
DTC
DTH
DTM
DTT
Kernel density
4
3
2
1
0
Autonomy
Nonautonomous
Autonomous
DNA transposons
Genes
C.
Kernel density
4
3
2
1
0
-1.0
-0.5
0.0
0.5
1.0
Skew
siRNA length
21
22
24
Genes


mCG
mCHH
Genes
LTRs
TIRs
Helitrons
Locus
Hairpin
Control
CG methylation level (prop. reads)
CHH methylation level (prop. reads)
-2 kb
0 kb
+2 kb
Distance from locus (kb)

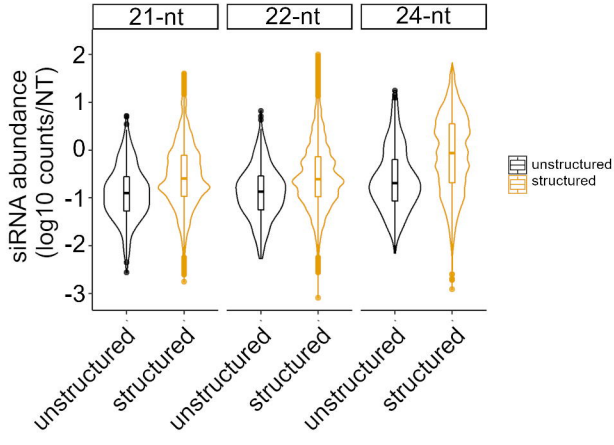
A.


Expression (TPM)
0
1
2
3
4
All genes
Genes with syntelog
unstructured
structured

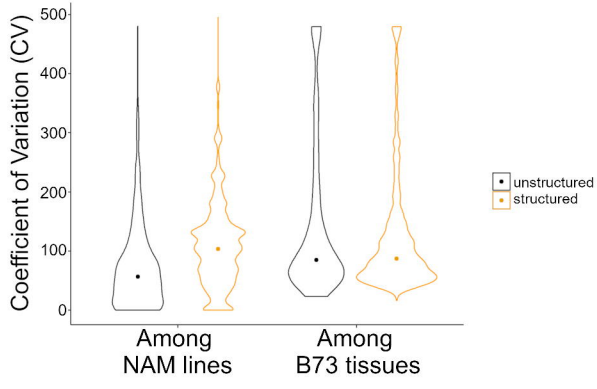
B.


Expression (TPM)
0.50
0.75
1.00
1.25
R2=0.015
P=0
R2=0.036
P=1.89e-86
-100
-75
-50
-25
minMFE (kcal/mol)
unstructured
structured

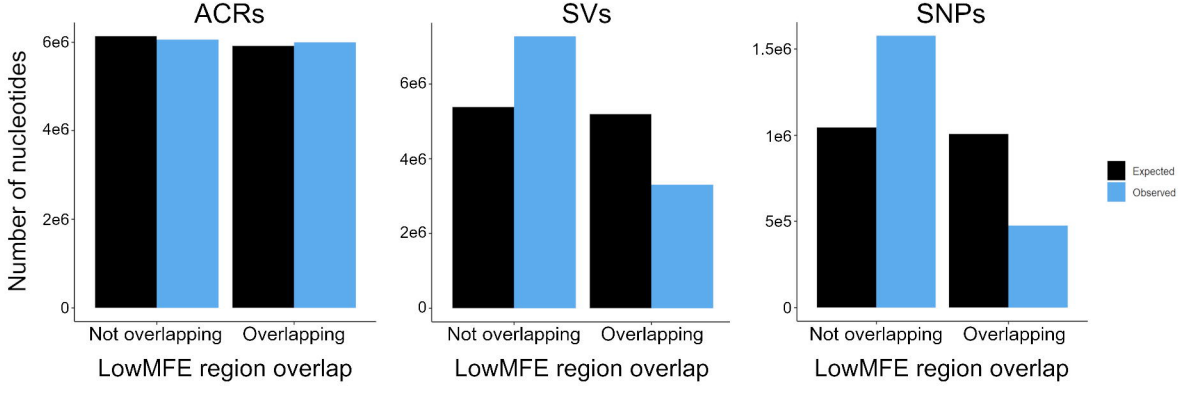
C.


21-nt
22-nt
24-nt
siRNA abundance (log10 counts/NT)
-3
-2
-1
0
1
2
unstructured
structured
unstructured
structured
unstructured
structured
unstructured
structured

D.


Coefficient of Variation (CV)
0
100
200
300
400
500
Among NAM lines
Among B73 tissues
unstructured
structured

E.


ACRs
SVs
SNPs
Number of nucleotides
0
2e6
4e6
6e6
5e5
1e6
1.5e6
Not overlapping
Overlapping
LowMFE region overlap
Expected
Observed