UNIVERSITY OF CALIFORNIA,
IRVINE

Evolution of RNA Secondary Structure and Epigenetic Features in Plants

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Galen Thomas Martin

Dissertation Committee:
Professor Brandon S. Gaut, Chair
Professor R. Michael Mulligan
Associate Professor Jose M. Ranz

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Most importantly, I would like to thank my advisor, Dr. Brandon S. Gaut, who supported me intellectually, personally, and financially throughout my time at UCI. I have been deeply lucky to have spent these years learning from you.

For improving my dissertation work, I would like to thank my committee members, Drs. Jose Ranz and Mike Mulligan, as well as Drs. J.J. Emerson, Dan Koenig, Grace Lee, Danelle Seymour, Aline Muyle, and Rebecca Gaut.

Thanks Mom, Dad, and Colin, for everything.

And thank you, in no particular order, to the many others who gave me interesting things to think about: Kevin Norris, Jeff Holmes, Alisa Hove, J.J. Apodaca, Danny Schnell, Stephen Cartier, Langdon Martin, Bibin Paulose, and Bob Youker.

Chapter 1: we thank D. Seymour for assisting with the manuscript. A. Muyle for comments, G. Gaut for statistical advice, and three anonymous reviewers for constructive comments. The work was supported by NSF grants 1542703 and 1655808 to BSG.

Chapter 2: We thank E. Solares, A. Muyle, and A. Bousios, and four anonymous reviewers for comments. The 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.

VITA

Galen Thomas Martin

EDUCATION

2017	B.S. in Chemistry and Biology, Warren Wilson College
2021	M.S. in Biological Sciences, University of California Irvine
2023	Ph.D. in Biological Sciences, university of California Irvine

PUBLICATIONS

Martin GT, Seymour DK, Gaut BS. 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 Biology and Evolution [Internet] 13. Available from: https://doi.org/10.1093/gbe/evab144

Robinson CS, Wyderko JA, Vang Y, **Martin GT**, Youker RT. 2021. Differential Effects of Oleuropein and Hydroxytyrosol on Aggregation and Stability of CFTR NBD1-ΔF508 Domain. Journal of Respiration 1:204–215.

TALKS AND PRESENTATIONS

Society of Molecular Biology and Evolution 2020 *Poster*

Paul Burton Seminar Series at Western Carolina University 2022 *Seminar*

Focus on Evolution Seminar at UC Irvine 2023 *Seminar*

AWARDS

Genome Biology and Evolution Best Graduate Student Paper Award; 2021 Runner-Up

Edward A. Steinhaus Teaching Award, UC Irvine; 2023 Awardee

ABSTRACT OF THE DISSERTATION

Evolution of RNA Secondary Structure and Epigenetic Features in Plants

by

Galen Thomas Martin

Doctor of Biological Sciences

University of California, Irvine, 2023

Professor Brandon S. Gaut, Chair

RNA transcription is the primary route through which genomes determine the phenotype of organisms. However, proper execution of this process on a genome-wide scale requires that (1.) the DNA of transcribed genes lies within portions of the genome that are accessible to RNA polymerase proteins, and (2.) that the mRNAs produced from these genes are stable enough in the cytoplasm to be translated. To complicate matters, many genomes are saturated with transposable elements (TEs) that must remain silent and not be transcribed. My work revolves around understanding the molecular mechanisms and evolutionary processes that govern DNA accessibility and RNA structure/stability.

My first chapter focused on methylated CHH (mCHH) islands, short regions of high methylation near genes that are linked to TE silencing and partitioning the genome between actively transcribed and non-transcribed components. We analyzed the evolutionary conservation of mCHH islands among grass (family Poaceae) genomes, as well as their relationships with gene expression, genic methylation, and proximity to TEs. We found that they were seldom conserved in orthologous genes between species, but they

often corresponded to insertions of certain DNA transposon families. They were also significantly negatively associated with methylated but positively associated with gene expression. Based on these findings, we propose a model wherein mCHH islands are a consequence of aberrant transcription leading to RNA-directed DNA methylation.

An unsolved mystery in genome partitioning is how TEs are initially identified and targeted for silencing. One way that this process has been observed is through hairpin secondary structures that form whenever TEs escape silencing and are transcribed. These hairpins act as a signal that allows structured transcripts to be broken down into small (21–24-nt) RNAs, which then methylate complementary parts of the genome. My second chapter focused on analyzing the genome-wide prevalence of this phenomenon in maize (*Zea mays*), where we found that it is widespread across many types of TEs. We also found that, where they exist in genes, these hairpin-like structures have the same effect. The prevalence of these structures despite their epigenetic effects suggests a conflict between RNA function and stability.

Finally, I studied the evolutionary dynamics of secondary structure in Arabidopsis thaliana using a novel method to identify derived mutations that interrupt ancestral mRNA secondary structures. I found that these mutations, even those at synonymous sites, exist at reduced frequencies relative to putatively neutral mutations in the global Arabidopsis population. Based on population genetic data, I estimated the selective effects of these mutations; they are more deleterious than neutral mutations but not as deleterious as most missense mutations. The population frequencies also varied between Arabidopsis subpopulations on a geospatial scale and were correlated with temperature. I hypothesize that the correlates with temperature reflect the fact that secondary structures vary in part as a function of heat.

INTRODUCTION

Genomic information is carried by DNA and RNA molecules, but the phenotypes produced from this information rely on the interactions between the nucleotide molecules and their cellular environments. Particularly, genomic information affects cellular processes through the cipher of gene expression, and nucleic acid physicality influences both the transcription of DNA and the stability of RNA. In the nucleus, DNA wraps around histone proteins to form nucleosomes that themselves interact to form higher order structures known as chromatin (Luger et al. 1997; Khorasanizadeh 2004; Dong et al. 2017), which can either by more or less accessible to RNA polymerase enzymes. Similarly, RNA is single-stranded and can form intramolecular base-pairing bonds that contort it into a complex secondary structure. This structure informs the function of the RNA molecule, including its splicing, translation, and localization (Vandivier et al. 2016). As a result, natural selection and genome evolution are guided by the myriad processes that control the physical structure of these molecules. Understanding genome evolution—and, consequently, biology as a whole (Dobzhansky 1973)—therefore relies on understanding how these molecular processes influence (and are influenced by) evolutionary processes. My work herein describes variations on this theme.

One reason that the physical structure of nucleic acids is important for genome function lies in the fact that most of the space within genomes is not taken up by functional genes. Instead, transposable elements (TEs)—selfish DNA sequences that can move or copy themselves—make up major portions of these genomes (Wicker et al. 2017; Stitzer et al. 2021). TEs are especially pervasive in plants; in fact, Barbara McClintock discovered them by studying maize (*Zea mays*) lines with large effect mutations (McClintock 1947). Such mutations are a natural consequence of TE mobility, because TEs can insert into gene bodies (disrupting the protein sequence) or promoters (disrupting expression), and they can grossly disrupt chromosomal architecture via ectopic recombination (Langley et al. 1988; Blumenstiel 2011). One famous example of the large-effect phenotypic consequences of these insertions in plants comes from the maize gene *teosinte branched 1* (*tb1*), which is a major quantitative trait locus that determines the apical dominance of maize compared to its progenitor species, teosinte (*Zea mays* ssp. *parviglumis*)(Doebley et al. 1995). The difference in apical dominance phenotypes between these two species comes from a TE insertion, *Hopscotch*, in the regulatory region of *tb1* that increases *tb1* expression in maize compared to teosinte (Studer et al. 2011). Because TEs can have these severe phenotypic consequences, their DNA sequences are often physically differentiated from those of genes to prevent their transcription.

Plant genomes achieve this differentiation through epigenetic means such as DNA methylation, the addition of an extra methyl group to the pyrimidine ring of cytosine bases. DNA replete with methylated Cs (mC) exhibits different chromatin states compared to non-methylated DNA, typically causing methylated DNA to be heterochromatic and inaccessible to polymerase enzymes (Zhang et al. 2018). Genomes target and maintain methylation states through various mechanisms, such as RNA-directed DNA methylation, where small RNAs produced from the destruction of TE transcripts directs methylation to complementary portions of the genome (Matzke and Mosher 2014; Erdmann and Picard 2020), so the post-transcriptional silencing (degradation of transcripts into small RNAs through RNA interference) and transcriptional silencing (methylation) of TEs are interlinked. Since RNA secondary structure partially controls the entry of transcripts into

RNA interference (Slotkin et al. 2003; Bousios et al. 2016), the physical structures of RNA and DNA are also interlinked.

However, studying the evolution of repetitive regions is not simple because TE genomic sequences tend to be degraded over time and are not conserved. As these regions are highly variable, it is difficult to study the specific epialleles involved in TE silencing. Instead, general methylation patterns can be compared between species, such as by Zemach et al. (2010) and Niederhuth et al. (2016). These comparative studies have given rise to two major conclusions: (1.) most plant TEs are heavily methylated and silent, and (2.) overall levels of methylation are divergent between different taxa, depending largely on genomic TE content, life history (e.g., clonal vs sexual propagation), and evolution (Seymour et al. 2014; Niederhuth et al. 2016). The fact that TEs are nearly ubiquitously silenced indicates that this process is important for genome function, but the fact that it differs by taxa indicates that conflicting pressures exist that are largely unknown. Hollister & Gaut (2009) showed that trade-offs between TE silencing and negative consequences for gene function may account for these variable patterns, which has largely been supported by the modern accumulation of methylome sequencing data (Niederhuth et al. 2016).

One such way that TE methylation can negatively affect gene function is through spreading of repressive methylation (Ahmed et al. 2011). Methylated TEs act as a “nucleation sites” of heterochromatin which spreads to surrounding regions (Choi and Lee 2020). This process has been observed to affect loci as far as 20 kb away from TEs (Lee and Karpen). In a genome—like that of maize—that is mostly composed of TEs, most genes are within proximity to TEs (Stitzer et al. 2019). Therefore, mechanisms must exist to protect expression of important genes from heterochromatic spreading (Phillips-Cremins and Corces 2013; Li et al. 2015). Another way that gene function may be hampered by the TE silencing comes not from DNA chromatin structure, but from RNA secondary structure; when genes have strong secondary structures, it can cause them to enter into the RNA interference pathway and be degraded (Li et al. 2012). Both of these complications from TE-silencing mechanisms may affect gene function, so the evolutionary pressures on genomes with differing TE content are sometimes contradictory: a major challenge for plant genomes lies in distinguishing TE sequences from gene sequences and maintaining expression of genes in the face of widespread TE silencing.

However, in plants, DNA methylation is more complex than a simple heterochromatic mark (Harris et al. 2018). It occurs in three possible sequence contexts: CG, CHG, and CHG (where H = any DNA residue other than G, guanine)(Zhang et al. 2018), and methylation in these different contexts is deposited by separate enzymes which confer different chromatin states (Stroud et al. 2014). One example of a non-heterochromatic relationship between methylated DNA and chromatin state can be seen in gene-body methylation (gbM), a phenomenon wherein constitutively expressed gene exons are heavily methylated in the CG context (Takuno and Gaut 2013). Evolutionarily, this type of methylation may be of functional importance as it is conserved in orthologous genes and evolves slowly between diverged species (Takuno and Gaut 2012; Takuno and Gaut 2013; Bewick et al. 2016; Bewick and Schmitz 2017; Seymour and Gaut 2020), but its function and importance remain a subject of debate, as it is not present in at least one species, *Eutrema salsugineum* (Bewick et al. 2016; Zilberman 2017). The way that methylation affects evolution is therefore complicated by the fact that the functional effects of methylation differ depending on the type of methylation involved.

My first chapter investigates a specific type of non-heterochromatic methylation feature, termed methylated CHH islands, which are areas of CHH methylation concentrated upstream of genes. These regions are are typically found at the boundaries between tightly packed chromatin (heterochromatin) and loosely packed chromatin (euchromatin), often near actively expressed genes. To explore the evolutionary dynamics of mCHH islands, compared their presence in eight different grass (family Poaceae) species. We discovered that mCHH islands are quite common and are associated with approximately 39% of genes, on average. We observed that genes near transposable elements (TEs) were more likely to have mCHH islands. While TEs play a role in this association, mCHH islands were not solely determined by TEs. Other factors, such as gene length and gene-body methylation (gbM), also influenced the presence of 5' mCHH islands. In some species, the absence of gbM was a stronger predictor of 5' mCHH islands than TE proximity. Additionally, gene expression level had a weak influence on the presence of mCHH islands. Finally, we assessed the conservation of mCHH islands across evolutionary time and found that they were generally not conserved. Overall, we concluded that mCHH islands are not solely a result of the TE silencing process, and perhaps they emerge from other properties of expressed genes, such as aberrant expression.

The second chapter focuses on the effects of RNA secondary structure on epigenetic response in maize. RNA secondary structure can trigger a response in the host organism controlled by small RNA molecules (smRNAs) in a process similar to RNA interference

(RNAi). To study this, we used computer-based methods to predict folded structures in the maize genome, particularly in regions that folded similarly to pre-miRNA (microRNA precursor) locations that are known to act as RNAi substrates. We discovered that these miRNA-like folded structures are common in genes and in most, but not all, groups of transposable elements (TEs). These miRNA-like regions had more smRNAs and methylation associated with them compared to regions without such structures, but genes with miRNA-like structures tended to be more highly expressed compared to other genes. However, the expression of these genes varied more across different maize lines, and this variability was positively linked to the number of smRNAs associated with them. Together, our results suggest that these hairpin structures serve a functional purpose but could also have negative consequences in the form of smRNA production.

The third chapter investigated the evolution of RNA secondary structure explicitly in *Arabidopsis thaliana*. In addition to their epigenetic effects, RNA structure is essential for proper mRNA processing, but its evolutionary patterns have not been characterized. We studied mutations that may alter RNA structure using computer predictions and empirical pairing data from the 1,001 genomes dataset. We categorized mutations as either conserving ancestral secondary structure or interrupting it. We found that structure-interrupting mutations had lower genetic diversity and were less. Additionally, the impact of these mutations varied depending on where they occurred within genes. We used demographic models to estimate that synonymous structure-changing mutations had selection coefficients about a third those of those for non-synonymous mutations. We conclude that RNA structure experiences subtle but widespread selection.

References:

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

Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, Wang L, Lu Z, Rohr NA, Hartwig B, et al. 2016. On the origin and evolutionary consequences of gene body DNA methylation. *Proc. Natl. Acad. Sci.* 113:9111–9116.

Bewick AJ, Schmitz RJ. 2017. Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.* 36:103–110.

Blumenstiel JP. 2011. Evolutionary dynamics of transposable elements in a small RNA world. *Trends Genet.* 27:23–31.

Bousios A, Diez CM, Takuno S, Bystry V, Darzentas N, Gaut BS. 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.

Choi JY, Lee YCG. 2020. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLOS Genet.* 16:e1008872.

Dobzhansky T. 1973. Nothing in Biology Makes Sense except in the Light of Evolution. *Am. Biol. Teach.* 35:125–129.

Doebley J, Stec A, Gustus C. 1995. teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141:333–346.

Dong P, Tu X, Chu P-Y, Lü P, Zhu N, Grierson D, Du B, Li P, Zhong S. 2017. 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Mol. Plant* 10:1497–1509.

Erdmann RM, Picard CL. 2020. RNA-directed DNA Methylation. *PLOS Genet.* 16:e1009034.

Harris CJ, Scheibe M, Wongpalee SP, Liu W, Cornett EM, Vaughan RM, Li X, Chen W, Xue Y, Zhong Z, et al. 2018. A DNA methylation reader complex that enhances gene transcription. *Science* 362:1182.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* 19:1419–1428.

Khorasanizadeh S. 2004. The Nucleosome: From Genomic Organization to Genomic Regulation. *Cell* 116:259–272.

Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B. 1988. On the role of unequal exchange in the containment of transposable element copy number. *Genet. Res.* 52:223–235.

Lee YCG, Karpen GH. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. *eLife* 6:e25762.

Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. 2012. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell* 24:4346–4359.

Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, Hirsch CN, Dawe RK, Madzima TF, McGinnis KM, 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.

Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389:251–260.

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

McClintock B. 1947. Mutable loci in maize. *Mutable Loci Maize* [Internet]. Available from: https://www.cabdirect.org/cabdirect/abstract/19491603312

Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA, et al. 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* 17:194.

Phillips-Cremins JE, Corces VG. 2013. Chromatin Insulators: Linking Genome Organization to Cellular Function. *Mol. Cell* 50:461–474.

Seymour DK, Gaut BS. 2020. Phylogenetic Shifts in Gene Body Methylation Correlate with Gene Expression and Reflect Trait Conservation. *Mol. Biol. Evol.* 37:31–43.

Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. 2014. Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. *PLOS Genet.* 10:e1004785.

Slotkin RK, Freeling M, Lisch D. 2003. Mu killer Causes the Heritable Inactivation of the Mutator Family of Transposable Elements in *Zea mays*. *Genetics* 165:781–797.

Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. 2019. The Genomic Ecosystem of Transposable Elements in Maize. Evolutionary Biology Available from: http://biorxiv.org/lookup/doi/10.1101/559922

Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. 2021. The genomic ecosystem of transposable elements in maize. *PLOS Genet.* 17:e1009768.

Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. 2014. The roles of non-CG methylation in Arabidopsis. *Nat. Struct. Mol. Biol.* 21:64–72.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* 43:1160–1163.

Takuno S, Gaut BS. 2012. Body-Methylated Genes in Arabidopsis thaliana Are Functionally Important and Evolve Slowly. *Mol. Biol. Evol.* 29:219–227.

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

Vandivier LE, Anderson SJ, Foley SW, Gregory BD. 2016. The Conservation and Function of RNA Secondary Structure in Plants. *Annu. Rev. Plant Biol.* 67:463–488.

Wicker T, Schulman AH, Tanskanen J, Spannagl M, Twardziok S, Mascher M, Springer NM, Li Q, Waugh R, Li C, et al. 2017. The repetitive landscape of the 5100 Mbp barley genome. *Mob. DNA* 8:22.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328:916–919.

Zhang H, Lang Z, Zhu J-K. 2018. Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* 19:489.

Zilberman D. 2017. An evolutionary case for functional gene body methylation in plants and animals. *Genome Biol.* 18:87.

CHAPTER 1

CHH Methylation Islands: A ubiquitous but non-conserved feature of grass genomes

1.1 Abstract

mCHH islands are peaks of CHH methylation that occur primarily upstream to genes. These regions are actively targeted by the methylation machinery, occur at boundaries between heterochromatin and euchromatin, and tend to be near highly expressed genes. Here we took an evolutionary perspective by studying upstream mCHH islands across a sample of eight grass species. Using a statistical approach to define mCHH islands as regions that differ from genome-wide background CHH methylation levels, we demonstrated that mCHH islands are common and associate with 39% of genes, on average. We hypothesized that islands should be more frequent in genomes of large size, because they have more heterochromatin and hence more need for defined boundaries. We found, however, that smaller genomes tended to have a higher proportion of genes associated with 5' mCHH islands. Consistent with previous work suggesting that islands reflect the silencing of the edge of transposable elements (TEs), genes with nearby TEs were more likely to have mCHH islands. However, the presence of mCHH islands was not a function solely of TEs, both because the underlying sequences of islands were often not homologous to TEs and because genic properties also predicted the presence of 5' mCHH islands. These genic properties included length and gene-body methylation (gbM); in fact, in three of eight species the absence of gbM was a stronger predictor of a 5' mCHH island than TE proximity. In contrast, gene expression level was a positive but weak predictor of the presence of an island. Finally, we assessed whether mCHH islands were evolutionarily

conserved by focusing on a set of 2,720 orthologs across the eight species. They were generally not conserved across evolutionary time. Overall, our data establishes additional genic properties that are associated with mCHH islands and suggests that they are not just a consequence of the TE silencing machinery.

1.2 Introduction

Epigenetic marks—such as DNA methylation, histone modifications and nucleosome positioning—affect the function and evolution of plant genomes (Diez, Roessler and Gaut, 2014; Vidalis et al., 2016). Perhaps the best characterized epigenetic effect is transposable element (TE) silencing. Epigenetic silencing within TEs is achieved through a complex series of biochemical reactions that usually result in the methylation of cytosines in three contexts: CG, CHG, and CHH (where H = C, T, or A). Methylation in all three contexts is associated with transcriptional silencing and a heterochromatic state, which effectively renders a TE unable to propagate (Slotkin and Martiennsen, 2007; Fultz et al. 2015). This silencing has evolutionary effects both because it alters the potential trajectory of genome content, which is dominated by TEs in large plant genomes (Lee and Kim 2014), and because TE methylation affects the expression of nearby genes (Lippmann et al., 2004; Choi and Lee 2020).

The processes of DNA methylation and maintenance vary by cytosine context. In Arabidopsis thaliana, CG methylation is deposited and then maintained across generations by the DNA methyltransferase MET1 (see Law and Jacobsen, 2010 for a review). Once it is established, CHG methylation is maintained by a separate methyltransferase (CMT3). In contrast to CG and CHG methylation, CHH methylation is not maintained but must be deposited de novo every generation. This deposition is achieved by one of two pathways. One is RNA-directed DNA methylation (RdDM), which uses homology of small interfering RNAs (siRNAs) to guide methyltransferase machinery to complementary DNA sequences (Law and Jacobsen, 2010). At siRNA target sites, the methyltransferase enzyme deposits methylation in all three contexts (CG, CHG, and CHH), particularly at the edges of targeted TEs (Zemach et al. 2013; Gent et al. 2013). The second pathway includes the plant-specific methylases CHROMOMETHYLASE 2 (CMT2) and CHROMOMETHYLASE 3 (CMT3) (Gouil and Balcombe, 2016), which methylate CHH and CHG cytosines in deep heterochromatin (Bewick et al. 2017). Unlike RdDM, CMT2 tends to methylate TEs across their full length (Zemach et al. 2013), but the regions methylated by RdDM and CMT2 do frequently overlap in Arabidopsis thaliana (Zemach et al., 2013).

These pathways contribute to the epigenetic features known as methylated CHH (mCHH) islands. mCHH islands are short regions of elevated methylation typically found upstream and downstream from genes. mCHH islands were first identified in rice, where they were associated with Miniature Inverted-Repeat Transposable Elements (MITEs) (Zemach et al., 2010b), a group of Terminal Inverted Repeat (TIR) DNA elements that often insert near genes. mCHH islands have also been characterized in maize (*Zea mays* ssp. mays); they were located near ~50% of genes and tend to be nearby genes with high expression levels (Gent et al. 2013). The maize analyses suggest that mCHH islands do not represent typical TE methylation, because maize TEs within 1 kb of genes are more heavily CHH methylated than other TEs and are more heavily methylated on the side of the TE closest to the gene that contained the mCHH island.

Thus far, the function of these mCHH islands is unclear. Given that they occur along boundaries between euchromatin and heterochromatin and also that mCHH island-associated genes tend to be more highly expressed than other genes, Gent et al. (2013) proposed that they partition the genome between different chromatin states, either by preventing the spread of epigenetic modifications into genes or, vice versa, by preventing the spread of euchromatin into TEs, thereby potentially reactivating them. Li et al. (2015) explored this potential function using mop1 maize mutants, which lack mCHH islands. They confirmed that the loss of RdDM leads to an increase of transcribed RNA from some TEs (between 29–179, depending on the tissue examined), suggesting that mCHH islands contribute to TE silencing. Similarly, others have found that the loss of near-gene RdDM in mop1 mutants can lead to unstable TE silencing that may be more susceptible to spontaneous reactivation during heat stress (Guo et al. 2021). Nonetheless, these observations do not fully explain why mCHH islands are concentrated near expressed genes. One potential explanation is that mCHH islands are a result, rather than a cause, of gene expression; this explanation is consistent with the observation mop1 mutants do not display widespread downregulation of mCHH-deficient genes (Li et al., 2015). There is, however, some evidence for a causal relationship between mCHH islands and gene expression, because recent work has shown that two gene A. thaliana products (SUVH1 and SUVH3) form a complex that binds CHH methylated sequences and enhances transcription (Harris et al., 2018). Raju et al. (2019) suggest that this mechanism implicates mCHH islands in protecting and promoting the expression of genes nearby TEs.

Until recently, it has been unclear whether mCHH islands are an idiosyncrasy of rice and maize or instead a general feature of plant epigenomes. To address this question, Niederhuth et al. (2016) surveyed genome-wide methylation patterns across a panel of 34 angiosperms. They defined mCHH islands as 100 bp windows within two kb of genes that were methylated in at least 25% of reads mapped to cytosines in the CHH context. This definition was based on previous work (Li et al., 2015), but it did not account for the widely varying background levels of CHH methylation found across species. Their survey also predominantly contained genomes of relatively small size. Their survey did include the TE rich ~2.3 Gb maize genome, but the remaining species had genomes of < 1.25 Gb in size, which is much smaller than the angiosperm average of 5.7 Gb (Dodsworth et al., 2015). This size distribution makes it difficult to assess whether mCHH islands correlate with genome size, as do other features of plant epigenomes (e.g., Alonso et al., 2015; Takuno et al, 2016; Niederhuth et al., 2016). Nonetheless, the Niederhuth et al. (2016) survey was remarkably informative about many aspects of DNA methylation variation among angiosperms, including mCHH islands. It reported, for example, that species vary markedly in the percentage of genes associated with upstream mCHH islands, from < 1% in Vitis vinifera to ~74% in Beta vulgaris. They also found that several species did not demonstrate an obvious association between mCHH islands and gene expression, making the relationship unclear.

Here we study mCHH islands in members of the grass family (Poaceae). We have chosen to focus on grasses for several reasons, including that they are economically important, that their intermediate evolutionary age makes them a useful comparative system, and that they encompass extensive variation in diploid genome size. They are also an interesting system from the perspective of CHH methylation, because all of the grass species surveyed thus far have low background levels of CHH methylation compared to other angiosperms (Bewick et al. 2017). This is a useful property for studying mCHH islands because they can be easily detected as exceptions to the background pattern of low CHH methylation.

To study mCHH islands, we focus on a set of eight grass taxa that span the breadth of the family, that vary widely in genome size, and that have available data—i.e., whole genome bisulfite sequencing (WGBS) data and RNAseq data (Seymour and Gaut 2020). Importantly, 2,720 1-to-1 orthologs have been identified among these same taxa, so that we can assess the evolutionary conservation of mCHH islands across species for specific genes. Given these data, we identify mCHH islands using methods that recognize that genome-wide mCHH levels vary across species. We then address four questions: First, what is the genome-wide pattern of mCHH islands across species? Is there, for example, a correlation with genome size for mCHH islands, as for other features of DNA methylation? Second, are islands located near genes that have nearby TEs, reinforcing the notion that mCHH islands are associated with TEs? Third, is there a relationship between mCHH islands and gene expression? That is, does gene expression predict the presence of a nearby island, or do other genic features better predict an island's presence? Finally, we take advantage of orthologous genes to investigate whether mCHH islands are evolutionarily conserved across species. Once established, is an island conserved, or is it an evolutionarily short-lived feature of the epigenomic landscape?

1.3 Results

General patterns of CHH methylation near genes

We analyzed the near-gene distributions of cytosine methylation in a dataset of WGBS and RNA-seq data from leaf and shoot tissue of eight grass species (Seymour and Gaut 2020) (Figure S1.1; see Methods). These species represent most of the evolutionary breadth of the Poaceae and span a 15-fold range of genome sizes from 5,428 Mb (*Hordeum vulgare*) to 355 Mb (*Brachypodium distachyon*) (Table 1.1). We first examined methylation in and near genes by measuring the weighted methylation level across all genes with available flanking data for 2 kb both up and down-stream. Following precedent (Schultz et al 2012; see Methods), we defined the weighted methylation level of a region as the proportion of methylated versus unmethylated bases that align to a single site in the appropriate context, and then averaged across all such sites in a defined region or window. We applied this approach to plot CG, CHG and CHH methylation in 200 bp windows and merged the results across genes (Figures 1.1A & S1.2). These analyses revealed well-known patterns—e.g., CG methylation within genes predominated over CHG and CHH methylation, and methylation was relatively low near both transcription start sites (TSS) and downstream of transcription termination sites (TTS) (Zemach et al., 2010b; Feng et al., 2010) (Figure 1.1A & S1.2).

These plots also demonstrated that peaks of CHH methylation within most species are located immediately upstream to the TSS and downstream of the TTS (Figure 1.1A & S1.2). Four features of the mCHH peaks merit further comment. First, the peaks were identifiable despite the fact that these figures average over all genes, not just the genes with mCHH islands. Hence, the peaks likely underestimate the magnitude of methylation levels for the subset of genes that are associated with mCHH islands. Second, the CHH peaks varied in magnitude. They were most prominent in *Z. mays* (Figure S1.2), suggesting either that *Z. mays* had a higher proportion of genes with mCHH islands than other species or that its mCHH islands were more highly methylated. However, mCHH peaks were also notable in *H. vulgare*, which has the largest genome in our sample, in *B. distachyon*, with the smallest genome in our sample, and in *S. bicolor*, with an intermediate size genome (Figure 1.1A). In Oryza sativa, another species with a small genome (Table 1.1), the peak height was also pronounced, reaching >10% of methylated reads across all cytosines in the CHH context (Figure S1.2). Third—as previously found in *Z. mays* (Gent et al. 2013)— mCHH peaks were far more evident in 5' upstream regions compared to 3' downstream regions; accordingly, most of our subsequent analyses focus on 5' islands. Finally, the analyses of *Triticum urartu* and *Phyllostachys heterocycla* yielded the least obvious 5' bumps in mCHH levels (Figure S1.2). Genome-wide analyses of *T. urartu* genes also yielded non-standard patterns of genic methylation (Figure S1.2). In this context, it is worth noting that these two genomes had the lowest contiguity among our sample (Table S1.1). In theory, low contiguity should not affect our results, because we only analyzed genes that had 2.0 kb flanking regions. However, the potential effects of fragmented data and/or poor annotations for these two species must be kept in mind.

One argument about mCHH islands is that they separate euchromatin from heterochromatin. If true, a simple prediction is that mCHH levels should be higher in species with large genomes, because they are more likely to have a high density of TEs interspersed with genic regions. To assess the relationship between mCHH levels and genome size, we measured weighted mCHH methylation in 1.0 kb regions upstream of TSS and downstream of TTS across all genes. We focused on 1.0 kb regions because this distance usually encompassed near-genic CHH peaks (Figure 1.1A & S1.2). Following Gent et al. (2013), we then estimated the fold-enrichment of those 1.0 kb regions by comparing them to an equal number of randomly determined 1.0 kb sites across each genome. All eight species exhibited greater than two-fold enrichments of near-gene mCHH, ranging from 2.34x enrichment near genes in *B. distachyon* to 6.25x enrichment in *H. vulgare*. We tested the relationship between GS and near-gene mCHH enrichment using phylogenetic generalized least squares (PGLS) regression (Symonds and Blomberg, 2014). The relationship was not significant with all eight species (p = 0.157), but *T. urartu* was a clear outlier. When we performed a post hoc analysis without *T. urartu*, the remaining seven species represented a strongly positive correlation between genome size and near-gene mCHH enrichment (p = 4e-4).

To probe this result further, we examined the relationship between genome size separately with levels of CHH methylation in near-gene 1.0 kb windows (Figure S1.3A) versus randomly chosen windows from throughout the genome (Figure S1.3B). Near-gene mCHH levels had a slight but nonsignificant negative relationship with genome size (Figure S1.3A) ($R^2 = -0.04$; p = 0.43). In contrast, random genomic windows had a stronger negative relationship with genome size ($R^2 = -0.1$; p = 0.22), mirroring a previous study that measured genome-wide mCHH levels in this same sample of eight species (Seymour and Gaut, 2020). Putting these results together, they suggest that any relationship between genome size and mCHH enrichment (i.e., the ratio of near-gene to random windows) reflects that background levels of CHH tend to be lower in large genomes. Thus, we find no compelling relationship between genome size and near-gene mCHH levels.

mCHH islands are methylation islands

To characterize mCHH islands more fully, we modified the method of Li et al. (2015) by splitting each genome into non-overlapping 100 bp windows and calling windows with elevated mCHH levels as mCHH islands when they were <2.0 kb from a gene. While Li et al (2015) called mCHH islands using an empirical >25% mCHH cutoff, we performed a binomial test on each window to determine whether there was significantly more CHH methylation than the genome-wide level ($p < 0.01$, after FDR correction) (see Methods). Note that we also applied alternative methods that either focused on the fraction of significantly methylated cytosine sites, rather than weighted methylation levels (Schultz et al. 2012), and also used an empirical cut-off rather than the binomial test (see Materials and Methods). All methods yielded qualitatively identical results with nearly-identical quantitative results. For simplicity, we present the results based on weighted methylation levels, to follow the precedence of previous mCHH island analyses (Li et al., 2015; Niederhuth et al., 2016), and on the binomial test, because it is an inherently statistical approach.

The binomial method yielded information about the mCHH level of statistically identifiable islands. For example, the median mCHH level of islands was highest in *Z. mays* at 53.8%, followed by *S. bicolor* and *O. sativa*, which were between ~40–50%. The remaining species all had median island levels of ~30–40% mCHH (Table 1.1), which is much higher than background levels of 12% or less (Figure S1.3A). Given the identification of islands, we characterized genes as mCHH island-associated (hereafter mCHH island genes) if they had at least one significantly elevated 100 bp region within 2.0 kb upstream. After examining >30,000 annotated genes per genome, we found that the proportion of island genes varied widely between species, from 17.3% in *P. heterocycla* to 71.9% in *O. sativa*, with an average of 38.8% across all eight species (Table 1.1).

To assess methylation levels around mCHH islands, we focused on the center of a single 5' 100 bp mCHH island window and plotted the average upstream and downstream of that center. As expected, we found that methylation distributions were elevated in the CHH context, but the results showed that islands were also elevated in the CG and CHG contexts in all eight species (Figure 1.2). Thus, as noticed previously (Niederhuth et al., 2016), mCHH islands are really "methylation islands," because they contain elevated methylation levels in all three cytosine contexts. This result further reinforces previous conclusions that mCHH islands represent RdDM deposition (Gent et al. 2013; Li et al., 2015), because RdDM is agnostic with respect to cytosine context (Matzke and Mosher, 2014).

Genic attributes of mCHH island-associated genes

mCHH islands are hypothesized to function as a boundary between TE-enriched heterochromatin and gene-rich euchromatin (Gent et al., 2013). This hypothesis predicts that island-genes should be adjacent to TE-rich regions more often than non-island genes. One way to examine this prediction is to investigate genome size, but this approach does not recognize that different genomes may have different organization of TEs and genes that may not be tightly correlated with genome size. Hence, to test this prediction in more detail, we explored the relationship between CHH island genes and TEs. For each species, we first downloaded publicly available annotations of each genome (Table S1.2). Then, for each gene in each species, we identified the annotated repetitive element closest to the TSS and measured the distance from the gene TSS to the nearest end of the repeat. As expected given previous research (Li et al., 2015), we found that mCHH island genes are much closer to repeats, on average, than non-island genes, and this was true for each of the eight species (logistic regression; $p < 0.01$). Although the signal was consistent across each species, note that the quality of repeat annotations likely vary across genomes, as does genome quality.

Another attribute of maize mCHH islands was their association with gene expression (Gent et al., 2013; Li et al., 2015), but the relationship between islands and gene expression did not hold across angiosperms in a more extensive dataset (Niederhuth et al. 2016). We re-investigated this relationship on a smaller scale by first repeating the analyses of previous studies (Gent et al., 2013; Li et al., 2015; Niederhuth et al., 2016). These studies separated genes into quartiles of expression and plotted mCHH levels upstream of genes. Our results were similar to previous work, showing that more highly expressed genes were slightly enriched for mCHH in all species (Figure 1.4A & S1.4). We also contrasted expression differences between mCHH island- vs non-island genes (Figure 1.4B). In all eight species, mCHH island-genes had slightly higher average expression levels than non-island genes, but the difference was significant for only three species (*Z. mays*, *H. vulgare*, and *P. heterocycla*). These results mimic Niederhuth et al. (2016) by suggesting that a relationship between mCHH islands and gene expression is either not universal or that it is so subtle as to be difficult to support statistically in some species.

We investigated additional genic features that may be associated with mCHH islands. For example, Li et al. (2015) reported a small (but non-significant) enrichment of gene-body methylation (gbM) genes among mCHH island-genes. We assessed the relationship between mCHH islands and gbM in two ways, using gbM either as a binary trait or as a quantitative variable (weighted mCG levels within exons) (see Methods). In both cases, we found a negative relationship between mCHH islands and genic methylation, and this negative relationship held for all species (logistic regression with %GC, $p < 3.9e-6$ for each species). We also tested whether island-associated genes were longer than other genes, because gbM genes are typically longer than unmethylated genes (Takuno et al., 2012). mCHH island genes were significantly longer than non-island genes in all species except *H. vulgare* and *P. heterocycla* ($p < 0.05$, logistic regression) (Figure 1.4D), and this relationship held for both total gene length and length of longest transcript (Figure S1.6). As a comparison to gene expression, we plotted genes by length quartiles (Figure 1.4C & S1.5), illustrating that the relationship with gene length is more obvious.

Finally, we incorporated all four predictors (TE distance, gene expression, gbM and total gene length) into a logistic regression model for each species. Gene expression and gene length were positive predictors of island presence. TE distance and gbM were negative predictors and significant in all eight species (Table S1.2). A limitation of logistic regression is that the estimates for predictors are on different scales, so it is difficult to compare their effects directly to one another from the estimates. To circumvent this problem, we applied variable importance analysis (Kuhn, 2008), which scales predictors for direct comparison within a model (Figure 1.5; see Methods). Three notable patterns emerged. First, TE proximity was generally—but not always—the most powerful predictor of the presence of an mCHH island. TE proximity was the most important variable in 5 of 8 species, but gbM was the strongest predictor in the remaining three species. Second, TE proximity was least important in *T. urartu*, which could again reflect features of genome or annotation quality. Third, gene length was also consistently significant, but its importance was always eclipsed by gbM and TE proximity. Finally, gene expression was comparatively unimportant, even in the three species for which it was a significant predictor (Table S1.2).

Assessing evolutionary conservation of mCHH islands

The availability of a set of orthologs from these species facilitates the address of another question: are mCHH islands conserved over evolutionary time? To address this question, we investigated mCHH island conservation among 2,720 orthologs (Seymour and Gaut, 2020). Islands were recorded as a binary trait for each ortholog; that is, each gene was or was not associated with an island in each species (Table 1.1). We then contrasted pairs of species and calculated the enrichment of island conservation. Enrichment was measured as the ratio of the number of orthologs with conserved islands between species to the number expected at random (see Materials and Methods). mCHH islands did not exhibit a signal consistent with a signal of evolutionary conservation (Figure 1.5A). Enrichment between species never exceeded 1.1x (Table 1.1), and the number of orthologs with conserved island-association was not significantly greater than expected by random chance in any pairwise comparison (permutation test, $p > 0.05$). As a contrast, we also investigated gbM conservation, because it is an epigenetic state that is known to be conserved between orthologs from different species (Takuno et al., 2013; Seymour et al., 2014; Takuno et al., 2016; Niederhuth et al., 2016; Seymour and Gaut 2020). In comparison to mCHH island enrichment levels of < 1.1x, gbM conservation ranged from a minimum of 2–fold enrichment to as much as 3.5x enrichment (Figure 1.5A).

We further examined some of the features that may contribute to rare cases of mCHH island conservation. We began by plotting, for each of 2,720 orthologs, the number of islands across eight species. The distribution of mCHH island conservation among orthologs (Figure S1.7) had a median of four species and a mean of 3.57 species, which was statistically indistinguishable both from the expected mean of 3.55 species under a purely random model (simulation, p = 0.272) and from normality (Shapiro-Wilkes test, p > 0.05). To investigate further, we applied linear models to test for correlations between gene-associated variables and maintenance of mCHH island status over evolutionary time. For example, the average exonic %CG across orthologs in all eight species was significantly negatively correlated with the number of species that had a gene island (R^2 = -0.0003, p = 0.003) (Figure 1.5B). Using the same approach, we found that the average expression of an ortholog was not correlated with the number of species that have an mCHH island (R^2 = 9.2-2e-4, p = 0.0567, Figure 1.5C) but that average gene length was positively correlated (avg. gene length, R^2 = 0.0029, p = 0.011, Figure 1.5D). The largest correlation was between conservation and TE distance (R^2 = -0.051, p = 9.6e-21, Figure 1.5E), providing further evidence of the link between mCHH islands and TEs. Although the magnitude of these significant correlations were very low, they largely recapitulated our within-species analyses.

mCHH islands and TE superfamilies in maize, rice and barley

Finally, we brought together data on mCHH islands, TEs and orthologs to further investigate the link among mCHH islands, genes and specific types of TEs. For these analyses, we narrowed our focus to three well-studied species—*Z. mays* (maize) *O. sativa* (rice), and *H. vulgare* (barley)—that had both reasonably contiguous genomes (Table S1.1) and careful TE annotations that distinguished among element superfamilies (Wicker et al. 2007; Table 1.2).

mCHH islands and homology to TEs: If the primary function of mCHH islands is to silence near-gene TEs (Li et al., 2015), their lack of evolutionary conservation is unsurprising because TE content often varies between species. Under this model one expects mCHH islands to be associated with sequences that have homology to TEs and perhaps to specific TE families (Zemach et al. 2010a, Li et al, 2015). Given data from maize, rice and barley, we first counted how often TEs were 2 kb upstream of the TSS of an annotated gene and then assessed whether those genes had a 5' mCHH island. The results varied markedly among species; ~30% of genes had both a TE and an mCHH island in barley and maize, but 74% of genes fell into this category in rice (Table 1.2). The interesting point about these values is that many mCHH islands—about 70% in maize and barley—are not obviously associated with nearby TEs.

One likely possibility for the low overlap with annotated TEs is incomplete annotations, particularly if mCHH island sequences are within fragmented remnants of TEs. To investigate further, we aligned mCHH island DNA sequences to a database of annotated TE sequences from Poaceae genomes, using Blast, and tallied the e-values of mCHH island sequences (see Materials and Methods). As expected, a large proportion of island sequences had high-threshold hits to TEs—e.g., 65.8%, 72.0, and 82.0% of island sequences had homology to TEs at an e-value < 1e-5 in *Z. mays*, *O. sativa*, and *H. vulgare*, respectively. Nonetheless, this implies that from 18.0% to 34.2% of sequences had little homology to TEs. As a genome-wide comparison for context, we sampled the same number of random 100 bp regions from throughout each genome and mapped them to the TE database. In the case of *Z. mays* and *H. vulgare* (Figure 1.6A), a smaller proportion of mCHH island sequences had significant (<1e-5) sequence homology to TEs than the random regions (65.8% vs. 78.9% in *Z. mays*; 82.0% vs 86.2% in *H. vulgare*). Moreover, in both species there was a substantial dearth of mCHH island sequences with exact (e-value < 1e-40) hits to annotated TEs. The situation differed somewhat in *O. sativa*, because it had a greater proportion of mCHH islands (72.0%) with < 1e-5 e-values compared to control regions (55.1%), but it again had a lower proportion of islands with stringent hits (e-value < 1e-40)(Figure 1.3A). Overall, these results indicate: i) that a substantial proportion of mCHH islands were not obviously derived from TEs and, ii) when they did exhibit homology to TEs, they were often diverged such that they did not have especially stringent matches.

Associations with specific TE superfamilies: Both Zemach et al. (2010a) and Li et al (2015) found especially strong signals of association between mCHH islands and terminal inverted repeat (TIR) DNA transposons. We therefore investigated particular classifications of TEs, asking whether their presence within 2 kb of a gene led to mCHH enrichment. We performed this analysis for 12 TE classifications (Table 1.2) that were present in all three species. The enriched TE types varied among species, but there was a clear general trend: DNA transposons tended to be enriched for mCHH islands and retrotransposons were not (Table 1.2). For each 5' mCHH island within an annotated TE, we also measured the distance to the closest 5' or 3' end of the TE and the distance to the TSS of the gene (Figure 1.6B). By definition, the mean distance of within-TE mCHH islands to the edge of the TE was smaller than the distance to the TSS. Surprisingly, however, the coefficient of variation (CV) of distance to the TSS was always smaller than the CV of the distance from the mCHH island to the TE end; this was true for every TE classification and species (Table S1.3 and

Figure 1.6B; P ≅ 0, Feltz and Miller asymptotic test for equality). Assuming the TE annotations were accurate, these results suggest that the location of islands are influenced by their position relative to genes more than their location within TEs.

mCHH islands and TEs between orthologs: If TE movement contributes to low conservation of mCHH islands between orthologs, the presence/absence of a TE should frequently coincide with the presence/absence of an mCHH island between species. We leveraged the set of 2,720 orthologous genes for *Z. mays*, *O. sativa*, and *H. vulgare* to test this idea. For each ortholog, we examined the presence or absence of mCHH islands between two species and then evaluated whether the orthologs had a TE within 2 kb. Focusing on orthologs that had lineage specific mCHH islands (i.e., an island in only one of the two species), we determined whether the mCHH island was 'dissonant' or 'coincident' with the TE, as defined in Figure 1.6C. As expected from our within-genome analyses (Figure 1.4), the presence of an mCHH island often corresponded with the presence of a TE, because coincident events were more frequent than dissonant events for each of the three species contrasts (chi-square; p <0.006). The effect also varied by TE types, because coincident lineage-specific mCHH islands were: i) significantly overrepresented for DTH (Harbinger) transposons and ii) significantly underrepresented for RLC (Copia), RLG (Gypsy) and DHH (Helitrons) (Table S1.4). Overall, the cross-species comparisons support inferences based on within-species data (Table 1.2) by suggesting that TEs—and specific TE superfamilies—are associated with mCHH islands.

1.4 Discussion

We have identified mCHH islands across a sample of eight grass species and documented their patterns relative to genome structure and function. Our study agrees with previous work by showing that mCHH islands have elevated methylation in all three sequence contexts (Niederhuth, 2016), that they vary in prevalence across species (Niederhuth, 2016), and that they tend to be associated with TEs (Zemach et al., 2010b; Li et al. 2015). Our work complements and confirms previous work, but it also provides novel insights into the evolutionary dynamics of mCHH islands as well as associations between mCHH islands and features of nearby genes.

TEs are associated with, but are not sufficient to explain, mCHH islands

Because mCHH islands in maize may act as a boundary between euchromatin and heterochromatin (Gent et al., 2013; Li et al., 2015), we predicted that the prevalence and level of mCHH islands varies with genome size, because larger genomes have more TEs (Tenaillon et al. 2010) and presumably more heterochromatin. We tested the relationship between mCHH islands and genome size in a few ways. We first examined levels of CHH methylation near genes against randomly chosen background windows of similar size. While we could recapitulate a modest negative correlation between background mCHH levels and genome size (Seymour and Gaut, 2020), near-gene mCHH levels were not correlated with genome size (Figure S1.3). The ratio of these two measures—i.e., the enrichment of mCHH levels near genes relative to the background—was negatively correlated with genome size *T. urartu* is not considered. To the extent that these enrichment analyses are accurate, it appears to be driven by the fact that larger genomes have lower genome-wide mCHH levels. We suspect this negative correlation reflects that larger genomes have a higher proportion of deeply-silenced heterochromatin, which is typically not targeted by RdDM for de novo CHH methylation (Zemach et al. 2013).

Separately, we leveraged our mCHH island annotations to measure the median mCHH level of mCHH islands in each species and to identify the proportion of genes across the genome that have an mCHH island within 2.0 kb upstream of their TSS. Neither of these values were obviously positively associated with genome size (Table 1.1); if anything, small genomes tended to have higher (although non-significant; R^2=-0.63; p = 0.09) proportions of genes associated with islands. The higher proportions in smaller (and more densely CHH methylated, Figure S1.3B) genomes are particularly notable given the biases in our statistical approach (see Materials and Methods), which favors identification of islands in larger genomes with lower CHH background methylation levels. Ultimately, the evidence for a relationship between genome size and mCHH islands remains ambiguous: larger genomes have lower background mCHH levels and thus experience somewhat higher near-gene mCHH enrichment, but smaller genomes tend to have a higher proportion of genes with mCHH islands.

Failing to find any compelling relationships with genome size, we turned to genome architecture and particularly to the potential association between mCHH islands and TEs. Consistent with previous work, we find that the presence of a nearby 5' repeat is a significant predictor of the presence of an mCHH island (Niederhuth et al., 2016). We also focused more carefully on three species—maize, rice and barley—that have well-established TE annotations, allowing us to assess whether specific TE classes and superfamilies are particularly associated with mCHH islands. Similar to previous studies of rice and maize (Zemach et al., 2010b, Li et al., 2015), mCHH islands are most consistently associated with Terminal Inverted Repeat (TIR) DNA transposons across species (Table 1.2). The details do vary somewhat because some TIR superfamilies like DTA (hAT elements) are associated with mCHH islands in maize but not significantly so in rice and barley. Nonetheless, TIR elements contrast markedly with retrotransposons, which are usually not enriched for CHH island associations (Table 1.2). It is worth noting that our method to test for enrichment only considers elements within 2 kb of a gene. Thus, these results do not simply reflect that most retrotransposons are located far from genes; when they are close to genes, they are associated with an mCHH island less often than DNA elements.

Previous work has hinted that mCHH islands are evolutionarily labile, because only ~64% of B73 genes had conserved mCHH enrichment (>10% mCHH) across five maize accessions (Li et al. 2015). By examining a set of 2,720 1:1 orthologs identified across all eight species (Seymour and Gaut, 2020), we have shown that 5' conservation of mCHH islands was never greater than expected by random (Figure 1.5A). However, the presence of lineage specific TEs coincides significantly with the presence of a lineage-species mCHH island (Figure 1.6C). TEs turnover rapidly in non-coding regions; this turnover provides at least a partial explanation for the lack of conservation of mCHH islands.

Genic properties associated with mCHH islands

Although TEs (and particularly DNA transposons) are clearly associated with the presence of mCHH islands, TEs are not sufficient to explain the presence of mCHH islands. This was illustrated aptly by Gent et al. (2013), who found that the proximal of near-genic TEs was more highly CHH methylated than the distal half. Gent et al (2013) ultimately concluded that mCHH islands are the product of “an interaction between genes and neighboring sequences” that can be independent of TEs. A subsequent study of maize showed that islands are enriched at the edge of transposons, particularly (TIR) elements, due to RdDM activity (Li et al., 2015). However, they also found that only ~40% of maize mCHH islands are associated with TIR elements, again supporting the view that the TEs may are not fully sufficient to explain mCHH islands. Consistent with previous work, our analyses show that an appreciable proportion of mCHH island sequences do not have strong Blast hits (e-value < 1e-5) to a TE database and that most do not have strong homology to existing TEs. Thus, many mCHH islands may not be derived from active silencing of annotated TEs.

If mCHH island sequences are not specific to a TE, what explains their presence? One possibility is that TEs trigger epigenetic modifications that then spread to adjacent chromosomal regions. If spreading occurs over sufficient distances, it could in theory explain two observations—i.e., that mCHH islands often exist when a TE is not within 2 kb of a gene and that a large proportion of mCHH islands have little homology to TE sequences. Yet, mCHH islands are also clearly a function of genic properties. For example, the maize literature has established that mCHH island genes tend to be highly expressed (Gent et al., 2013, 2014; Li et al., 2015), although it has not been clear if this relationship holds across species (Niederhuth et al., 2016). We have measured gene expression in all eight species and contrasted expression levels between genes that had and did not have nearby 5’ mCHH islands. mCHH island genes are generally more highly expressed than genes without islands, but this relationship is not significant in five of eight species. Intriguingly, the three species that have a significant association have the largest genomes, an observation for which we have no ready explanation (Figure 1.4 & S1.4). There is also an important caveat: we have only examined expression in one tissue, but the tissue(s) under study may be critical, as may be expression breadth (Li et al., 2015). Future studies need to interrogate across more tissue types.

Surprisingly, in all species, gbM is a stronger predictor of mCHH islands than gene expression; in fact, gbM is even a stronger predictor than TE-proximity in three of eight species (Figure 1.4). Our observed negative gbM relationship differs from the positive association documented previously in maize (Li et al., 2015), which examined a subset of syntenic genes. It is difficult to know whether differences between studies reflect the particular subset of genes or specific features of their data. However, we retrieve the same negative relationship when we focus only on the ortholog gene set and on alternative measures of gbM (e.g., presence/absence instead of quantitative measures). Altogether, our results show that 5' mCHH islands are associated with genic properties that include (from stronger to weaker associations): gbM, gene length and gene expression. Intriguingly, mCHH islands are also located at a more consistent distances from the TSS than from the edge of the TE in which they reside (Figure 1.6B), suggesting that spacing relative to the gene is more important than the physical confines of a TE.

Additional questions about mCHH islands

This study has confirmed several features of mCHH islands and discovered more, but it leaves at least two important questions unanswered: how are mCHH islands formed and what is their function? We cannot answer either question, but we can provide a few additional insights. Previous work in maize has shown that the proximal mechanism of formation is RdDM (Gent et al., 2013; Li et al., 2015), which is consistent with the fact that mCHH islands have high methylation across all three methylation contexts (Figure 1.2). Our genome-wide results uphold the view that this is not solely a TE driven phenomenon, suggesting again that mCHH islands represent an interaction between active genes and their neighboring sequences (Gent et al., 2013). A crucial feature of this interaction may be RNA polymerase II (Pol II) (Gent et al., 2013), because it is necessary for both genic transcription and for non-canonical (RDR6) RdDM (Zheng et al., 2009; Cuerda-Gil and Slotkin, 2016).

The specific characteristics of genes or their neighboring sequences that trigger island formation remain unclear. Recent work has shown that maize mCHH island targets are enriched for a specific CG-rich sequence motif (Long et al. 2021), but this motif neither fully explains the existence of islands nor our observations about gbM and gene length. Another possibility is that mCHH islands represent a consequence of erroneous gene transcription (Gent et al., 2013). In this model, genes occasionally experience internal and bidirectional initiation of transcription, leading to transcripts which extend beyond the 5' end of the gene or beyond the polyadenylation site. This transcription of neighboring sequences could engage RdDM and precipitate mCHH islands, especially when those transcripts encompass nearby TEs. Once established, CHH islands may help moderate the effects of neighboring TEs on gene expression by binding the SUVH1 and SUVH3 mediated complex (Harris et al., 2019; Raju et al., 2019).

This proposed mechanism of island formation complements one of our primary observations, which is that mCHH islands and gbM are negatively associated, because one of the presumed functions of gbM is to suppress internal transcription (Zilberman et al.

2008). Although evidence for this gbM effect is admittedly mixed (Neri et al., 2017; Teissandier and Bourc'his 2017, Zilberman 2017; Le et al. 2020), it could drive the observed negative association between gbM and mCHH islands. Under this model, gbM suppresses aberrant transcription but mCHH islands result from aberrant transcription, leading to a negative association. This model is also consistent with our finding that mCHH island genes are generally longer than other genes, because longer genes have a higher probability of containing a cryptic internal promoter. The model also helps explain the relationship between gene expression and TE proximity, because non-expressed genes have no Pol II activity and hence could not develop islands.

Interestingly, a small proportion of genes (ranging from 5.5% in *P. heterocycla* to 26.0% in *O. sativa*) have both gbM and mCHH islands. This is not predicted by our model unless this subset of genes is particularly prone to aberrant transcription. We predict that such genes should be highly expressed and may represent rare cases in which the two epigenetic features are reinforcing and perhaps even synergistic. Consistent with the prediction, genes with both epigenetic features are more highly expressed than genes than with just one of the two features, and this observation holds across all eight species (Figure S1.8). Although intriguing, it is at best preliminary evidence for the model that posits that both gbM and mCHH islands are related to aberrant transcription. Further analyses of aberrant transcription may prove insightful, recognizing that the effect may be subtle, just as the effects of gbM on gene expression are subtle but have become evident with the analysis of larger and more expansive data sets (Muyle et al., 2021). Another important avenue for future research will be analyses of expression breadth and responsiveness as they relate to mCHH islands.

1.5 Materials and Methods

Data and Methylation Calls

These analyses used RNAseq and BSseq data from eight grass species. The *H. vulgare*, *T. urartu*, *S. italica*, and *S. bicolor* data were retrieved from the NCBI Short Read Archive under accession PRJNA340292, all of which were generated from leaf tissue in 6-week-old plants. Data from *B. distachyon* (SRR628921, SRR629088, SRR629207) and *O. sativa* (SRR1035998, SRR1035999, SRR1036000) RNAseq and BSseq were also generated from young leaf tissue. Finally, *Z. mays* (SRR850328) data were generated from seedling tissue. The differing tissues used in this study should have little effect, as methylation typically varies little between tissues (Schmitz et al., 2013; Roessler et al. 2016). These data were chosen to make our mCHH island results comparable to gbM results from the same species, using the same data and reference genomes (Seymour and Gaut, 2020). Genome sizes in Table 1.1 for each of the species were from the Kew C-value database (http://data.kew.org/cvalues/ last accessed September 2, 2019) except for that of *P. heterocycla*, which came from Peng et al. (2013).

For all eight species, we used methylome data provided by Seymour and Gaut (2020). Briefly, they trimmed BSseq reads for quality and adapter sequences using trimmomatic (v0.35) and used Bismark (v0.15.0) with bowtie2 (v 2.2.7) to align trimmed reads to the reference genomes of each species, with seed parameters of -N 0 -L 20. After alignment, Bismark methylation extractor (0.15.0) was used to determine numbers of methylated and unmethylated reads at each cytosine site. The accessions used in this study were the same as those used to generate the reference genomes. The reference genomes were: *S. italica* (Bennetzen et al. 2012; Sitalica_312_v2.fa), *O. sativa* (International Rice Genome Sequencing Project, 2005), *Z. mays* (Schnable et al. 2009), *T. urartu* (Ling et al. 2013), *S. bicolor* (Paterson et al. 2009), *P. heterocycla* (Peng et al. 2013), *B. distachyon* (International Brachypodium Initiative 2010), and *H. vulgare* (Mascher et al. 2017). Coverage information for methylomes can be found in Table S1.6.

Measuring mCHH in windows and defining mCHH islands

We calculated the weighted mCHH level of defined genomic windows using a custom R script (R version 3.5.1). Following Schultz et al. (2012), the weighted methylation of a window was calculated separately for each cytosine context (CG, CHG or CHH) as the number of methylated reads in that window divided by the number of unmethylated reads at cytosines in the same context. We applied this metric to windows of various lengths for different analyses (see text). When this metric was compared to randomly chosen windows (e.g., Figures 1B and 3), we identified those windows using the sample_n() function in the R package dplyr v1.0.2 (Wickham et al. 2019).

mCHH islands were identified using the method of Li et al (2015), altered to be applicable across species with varying genome-wide mCHH levels. Each chromosome was divided into non-overlapping 100 bp windows and weighted methylation levels were calculated for each window. Each window was then assigned a p-value with a one-sided binomial test for mCHH hyper-methylation, similar to the method of Takuno and Gaut (2012) for genes. 100 bp windows were annotated as mCHH islands if they were within 2 kb of gene TSS, contained more than 5 mCHH cytosines, and possessed an Benjamini–Yekutieli FDR-corrected P value < 0.01. Coverage across CHH residues was counted in the near-gene (2 kb 5' of TSS) region for each gene; genes were excluded if they lacked WGBS CHH site coverage >2x in more than half of this region.

For completeness, we performed the same analysis based on methylation levels calculated as the fraction of methylated cytosines (Schultz et al., 2012). In this variation, each cytosine in the CHH context was determined to be either methylated or not based on the binomial test (Lister et al., 2008) when the site had 2 or more reads. Genome-wide and window-wide methylation were then calculated as the percentage of cytosines that were methylated among cytosines with sufficient data. The two methods (weighted vs. per site methylation) yielded nearly identical results; for example, >95% of genes in *Z. mays* had the same designation as island or non-islands genes. All downstream analyses were qualitatively identical for the two analytical methods; we report only weighted methylation levels for simplicity.

Similarly, to ground truth our binomial test, we also explored analyses without using the binomial—i.e, by employing the empirical 25% cutoff (Li et al., 2015). We found similar results between the two methods. For example, we asked what proportion of the 21,760 (=2,720 orthologs x 8 species) genes had mCHH islands based on the binomial method and the 25% cutoff. The two methods agreed for 85.8% of the genes. These results strongly suggest that our overall results are robust to some variation in the mCHH island detection approach. As a further test of robustness, we applied linear regression and variable importance analysis to mCHH islands detected with 25% cut-offs, yielding qualitatively similar results.

Expression analyses

We used the expression information calculated by Seymour and Gaut (2020) from RNAseq data to evaluate expression of mCHH island genes. RNAseq data came from the same tissues and accessions as BSseq data. The raw RNAseq data were filtered for quality and adapter trimming with trimmomatic (v0.35), requiring 45 bp read lengths after trimming. Alignments to reference annotations were performed using bwa (v0.7.12) allowing two mismatches (-n 2). Raw read counts were normalized (TMM) in edgeR (v3.20.9) for each species and reads-per-kilobase-mapped (RPKM) was estimated from the fitted values. Trimmed reads were aligned to annotations available for each genome and reported in supplementary table S4 of Seymour and Gaut (2020).

Expressed genes were divided into quartiles of expression based on log2 RPKM using the quantile() function in R. Genes that were not present in RNA-seq data were marked as being in quartile 0. Metaprofiles showing near-gene mCHH at different expression quartiles were generated by demarcating 100 bp windows across the 2 kb regions 5' to the TSS and separately calculating mCHH means per window for each expression quartile. To compare expression of orthologs between species, expression in RPKM was normalized to zero-mean unit variance using the scale() function in R.

Gene characteristics and regression analyses

We used PGLS regression to query the relationship between GS (log 10 1C) and levels of CHH DNA methylation. PGLS regression corrects for phylogenetic relationships and requires information about branch lengths between species. For the latter, we used a phylogenetic tree inferred by Seymour and Gaut (2020) from 2,982 single-copy orthologs across the eight species of interest (Figure S1.1). The single copy orthologs were identified using orthomcl (v2.0.9) and BLASTP (v.2.2.30), with the "-evalue 1e-5 -outfmt 6" options. The phylogeny was inferred from concatenated nucleotide alignments of orthologs using ape (v5.2) and phangorn (v2.4.0) using a GTR substitution model. PGLS regression using these branch lengths was performed using nlme (v3.1.131) in R. Genome size was retrieved from the Kew C-values database (https://cvalues.science.kew.org/)

We downloaded repeat annotations for all 8 species (Table S1.5). Given annotations, we calculated the TE distance to a gene by taking the absolute value of the difference between the TSS of each gene and the 5' or 3' edge (whichever was closest) of the nearest TE (indiscriminate of strand). Genes without a detectable TE upstream, which were generally the first or last genes on a scaffold, were not included in this analysis. The distance was marked as zero when a TE overlapped with a gene.

We also calculated genic parameters. Gene length included both introns and exons and was calculated by subtracting the minimum from maximum annotated chromosomal position for each gene. As a comparison to gene expression, we divided genes into quartiles of length using the same method as described above for gene expression data. Weighted exonic mCG levels were calculated as before (#mCG reads / # total reads) inside exons of the longest transcript in each gene. Logistic regression models were built in R using the glm() function, using genic variables (expression, distance to a TE, length, and gbM) to predict island association as a qualitative, binary variable. We standardized each variable to a 0–1 scale by subtracting the lowest value of each set from all values in each, then dividing by the highest value in each set. We built this model separately using both gbM as a qualitative and quantitative variable, to make sure that the inclusion of a qualitative variable did not affect the outcome. We evaluated the contribution of each predictor variable to the model using the varImp() function in the caret package (Kuhn, 2008).

We assessed conservation of mCHH island by focusing on the list of orthologs identified by Seymour and Gaut (2020). After filtering for near-gene coverage in our WGBS, we included 2,720 genes (Table S1.7). We calculated fold-enrichment of mCHH island and gbM conservation by comparing observed and expected counts between pairs of species. The observed was the number of orthologs that were mCHH island associated or gbM within both species; the expected was the product of proportions of mCHH island orthologs (Table 1.1) between the two species in each pair. We modeled the relationship between mCHH island frequency (counts from 0 to 8 across species) and each genic variable using the lm() function in R. The Feltz and Miller test (1996) was used to assess the equivalence of CVs between distances to TE-edges and genes.

Blast analyses

To investigate the homology of mCHH island sequences in maize, rice and barley to TEs, we built a reference TE database. The data set consisted of: i) *O. sativa* TE fasta files from the Rice Transposable Element Database (last accessed Feb 10, 2020) (Copetti et al. 2015), ii) *Z. mays* and *H. vulgare* TEs extracted from their reference genomes using samtools, iii) full length TEs from Stitzer et al., 2019, accessed at https://github.com/mcstitzer/maize_TEs/blob/master/B73.structuralTEv2.fulllength.2018-09-19.gff3.gz (Stitzer et al. 2019) and iv) repeat sequences from the Transposable Element Platform (TREP) database (Wicker et al. 2002) (https://botserv2.uzh.ch/kelldata/trep-db/index.html, last accessed Apr 1, 2021). This TE reference database was used as a reference to query mCHH island sequences. The island sequences were run through BLASTn (v2.8.1) (Altschul et al., 1990) using discontiguous megablast (-task dc-megablast) against a custom reference fasta file containing the combined TE sequences in the database. To identify a set of random, “control” sequences, we sampled a number of non-mCHH island 100 bp windows equal to the number of mCHH islands from each genome using the sample_n() function in the R package dplyr v1.0.2 (Wickham et al. 2019). These sequences were BLASTed against the TE reference in the same manner. Sequences with no BLAST hit were assigned an e-value of 1.0.

Figures

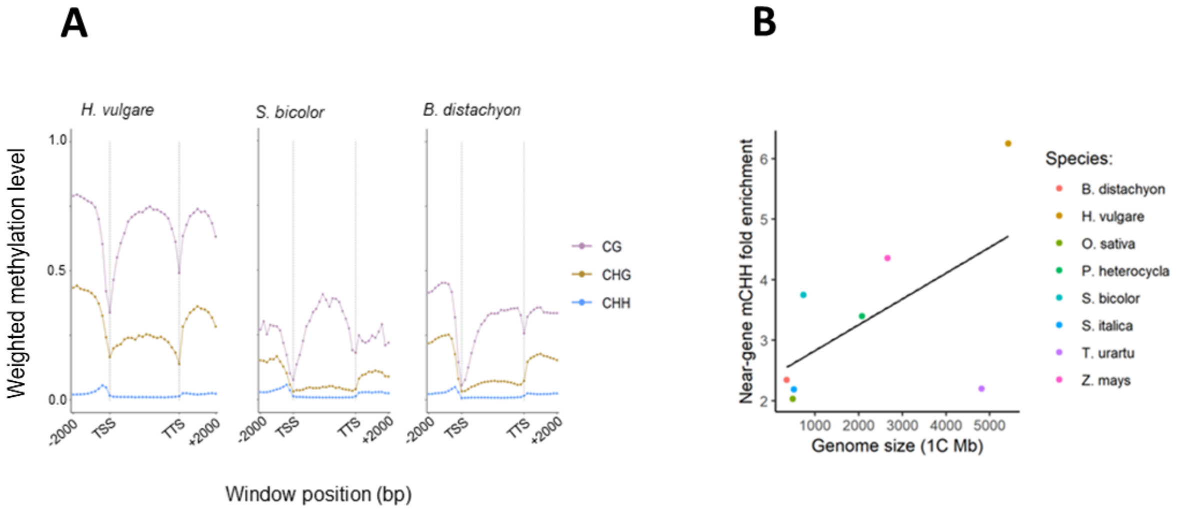


Figure 1.1. Near-gene methylation across Poaceae species. (A) Profiles of methylation across genes and their 2.0 kb 5’ and 3’ flanking regions. Weighted methylation levels are summarized in 10 200 bp windows upstream and downstream of genes, and in 20 equally sized windows within genes that vary in size depending on gene length. These figures summarize across full genes, with exons and introns. Here we show three species that span the range of genome size (Table 1.1), with the remaining species shown in Figure S1.2. TTS refers to the transcription termination site. (B) Near-gene enrichment of mCHH increases with genome size. Near-gene mCHH enrichment represents the mean weighted mCHH levels in 1 kb regions upstream of the transcription start site (TSS) divided by the mean weighted mCHH levels in an equal number of 1 kb regions randomly selected throughout the genome.

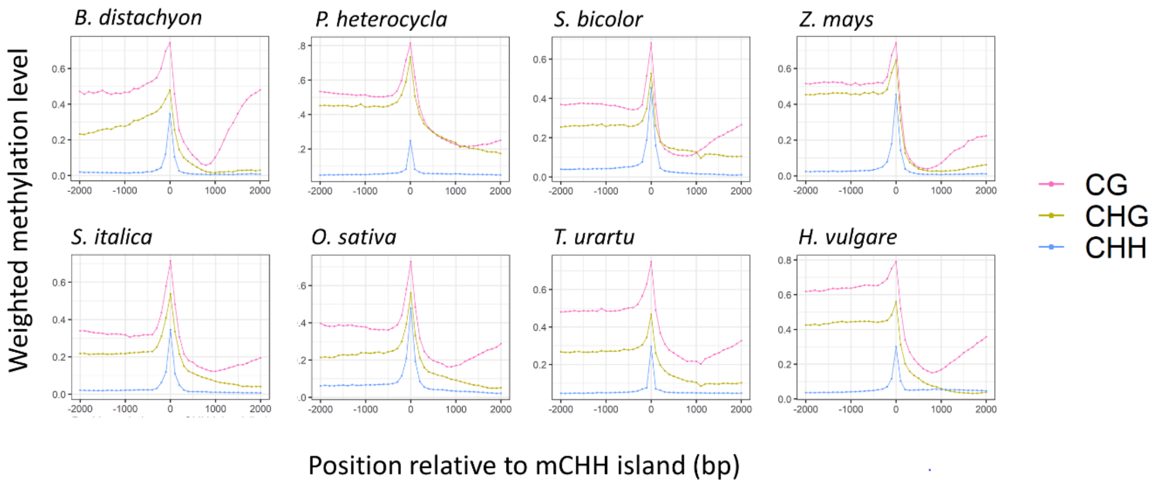


Figure 1.2. Profiles of methylation across mCHH islands in each sequence context. The x-axis provides the distance in base pairs (bp) from a detected island, which is centered at zero. The points on the graph represent weighted methylation levels in 100 bp windows. The islands were not at a fixed distance from genes, because they were determined by significance tests, but they were within the 2 kb 5’ flanking region of genes.

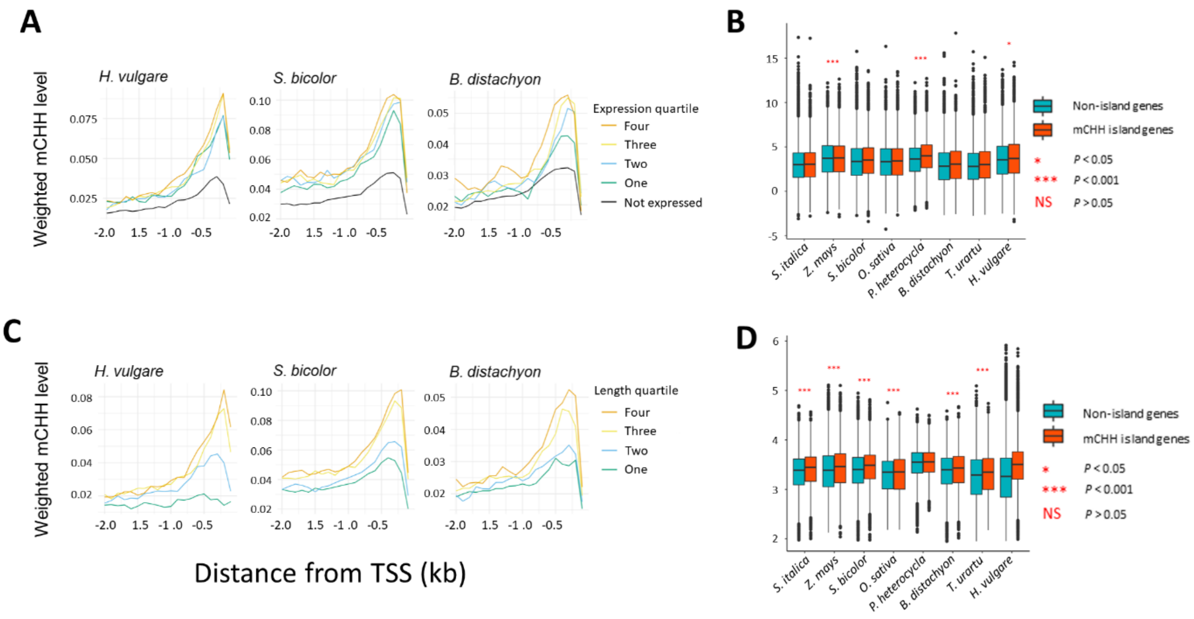


Figure 1.3. mCHH island relative to gene expression and length. (A) Profiles of near-gene methylation in genes separated into four quartiles of expression and into non-expressed genes. The graphs illustrate for some species that genes in the higher quartiles tend to have higher 5' flanking CHH methylation. (B) Expression levels between mCHH island genes and non-island genes. Significance levels between the two categories are shown for each species, with NS = not significant. (C) Profiles of near-gene methylation in genes separated into four quartiles for gene length. (D) The length of island and non-island genes. Significance levels between the two categories are shown for each species, with NS = not significant. For B) and D), the box plots present the median, with the edges representing the upper and lower quartiles. These length measures were based on distances from the TSS to the TTS, but the results hold using the length of exons in the longest transcript (Figure S1.6). For panels A) and C), the species were chosen because they represent a range of genome size, as in Figure 1.1. The remaining species are shown in Figures S1.4 & S1.5.

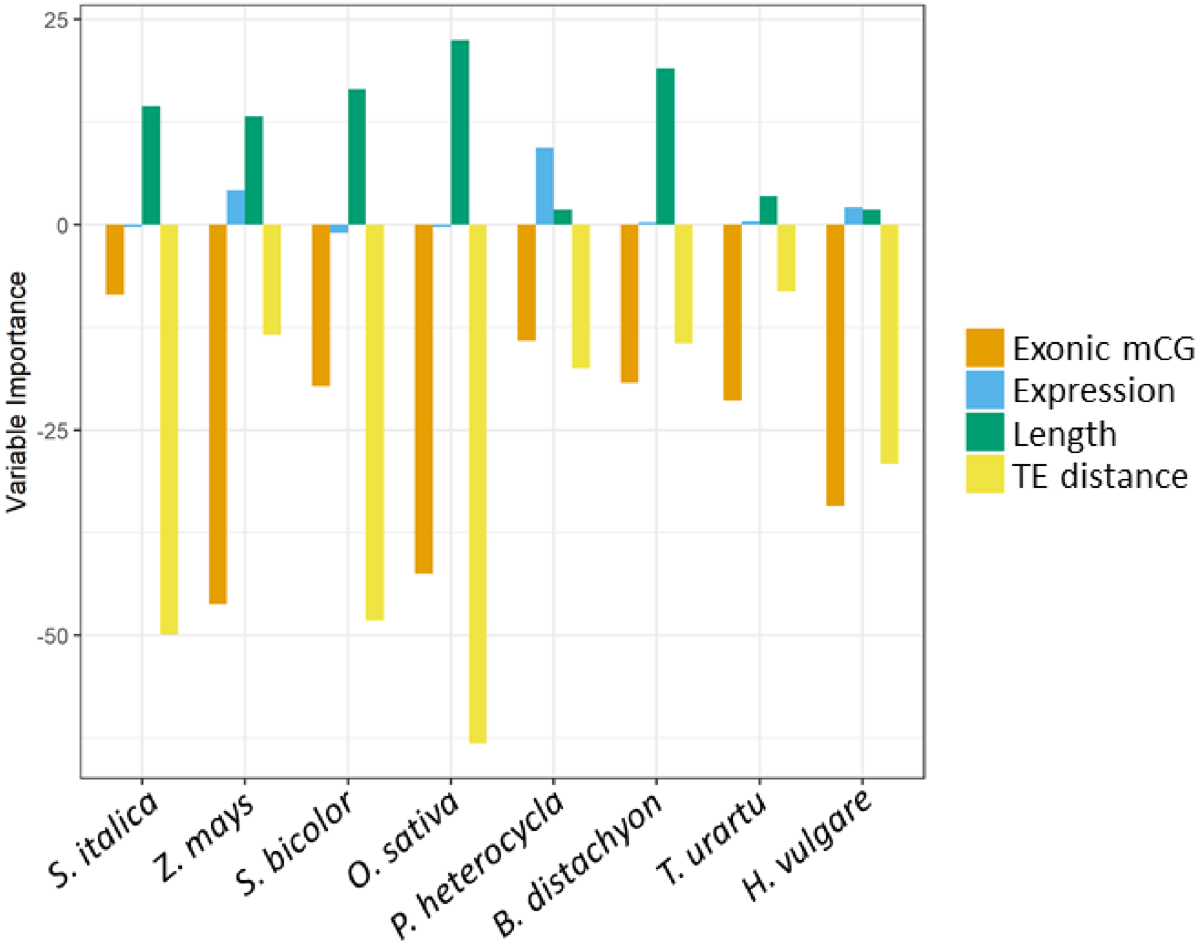


Figure 1.4. Variable importance analysis of the logistic regression model presenting the contribution of each variable to the model on an equivalent scale. Values <0 on the y-axis denote a negative association between the predictor and the presence of a CHH island; values > 0 are positive predictors.

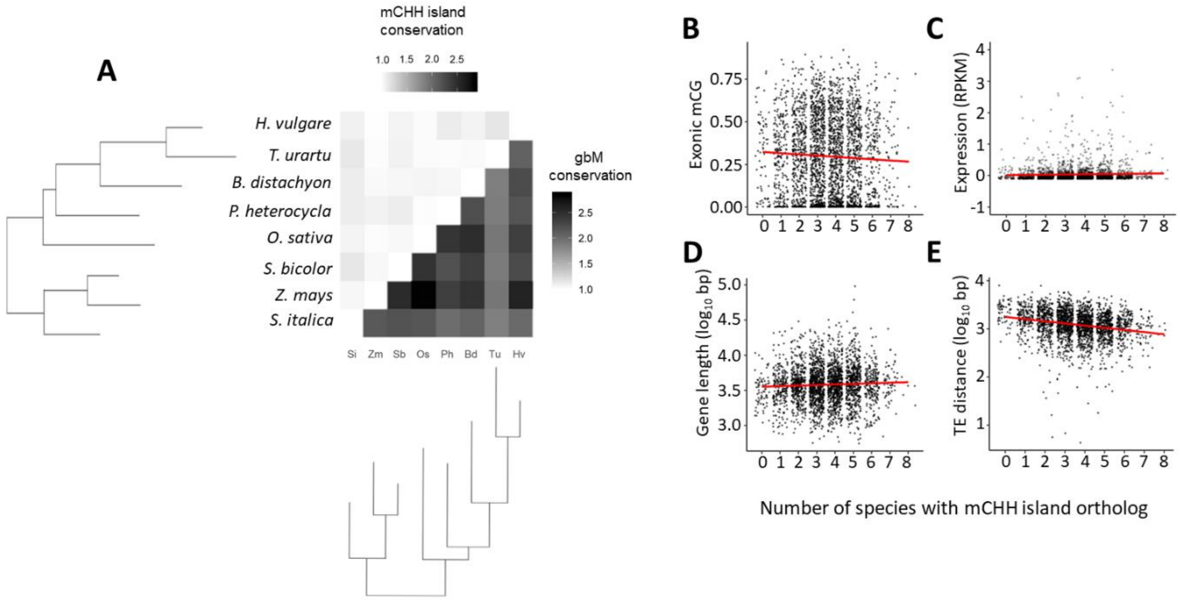


Figure 1.5. Conservation of mCHH islands across orthologs in grass species. (A) A heatmap of the enrichment of features over the random expectation of 1.0. Top half: enrichment of mCHH island conservation between pairs of species based on one-to-one orthologs. Bottom half: enrichment of gene-body methylation between pairs of species based on one-to-one orthologs. (B-E) Graphs of the relationship between mCHH island conservation and each genic predictor variable: exonic mCG level (B), expression (C), length (D), and TE distance (E). For each graph, the x-axis denotes the number of orthologs, of eight total, with a 5' mCHH island, and the *y*-axis denotes the average value of the stated statistics in the ortholog across species.

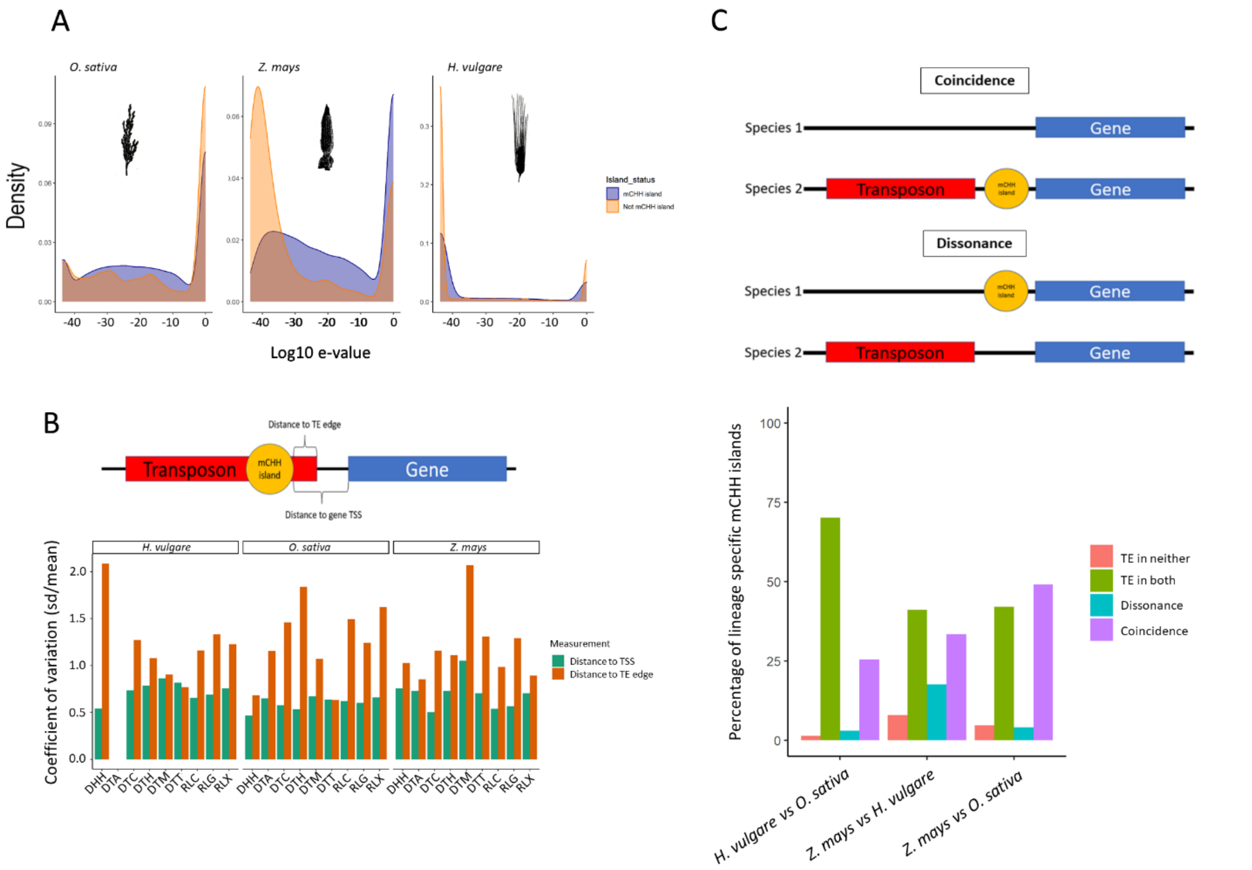


Figure 1.6. mCHH islands in relation to TE presence. (A) The distribution of e-values after blasting sequences to an annotated TE database for *Z. mays* (left) and *O. sativa* (middle) and *H. vulgare* (right). Each graph plots the results for 100 bp mCHH island DNA sequences and an equal number of randomly chosen 100 bp non-island sequences for comparison. (B) The coefficients of variation for mCHH island distances from gene TSS- (orange) and TE-edges (green) for each of the different types of TEs analyzed (Wicker et al., 2007). The schematic above the graphs defines the distances measured. (C) A schematic that defines the use of the terms Coincident and Dissonant. Each term describes a comparison of orthologs between pairs of species, with a lineage specific 5’ mCHH island in only species. Coincidence is when there is a lineage species TE and island in the same species; Dissonance is when the TE and island are in different species. The bar graph shows the frequency with which orthologs possess a lineage-specific TE alongside a lineage-specific mCHH island (coincidence) or the opposite (dissonance) in the three pairwise comparisons between maize, rice, and barley.

Tables

Table 1.1. A list of species examined in this study, with their genome size, the number of genes used in analyses and information about CHH-island characteristics.

Species	Genome Size (Mb)[1]	No. Genes[2]	% mCHH Island Genes[3]	Median island mCHH level[4]	% mCHH Island orthologs[5]	Median ortholog island mCHH level[6]
Brachypodium distachyon	355	34,257	55.16%	31.49%	58.20%	32.76%
Hordeum vulgare	5,428	35,200	28.22%	34.19%	41.34%	34.64%
Oryza sativa	489	41,806	71.85%	41.80%	76.61%	49.19%
Phyllostyachys heterocyla	2,075	30,946	17.27%	29.03%	18.51%	28.57%
Setaria italica	513	34,170	29.80%	38.89%	34.31%	41.00%
Sorghum bicolor	734	33,972	54.01%	46.05%	59.91%	49.06%
Triticum urartu	4,817	33,612	22.94%	34.78%	28.29%	35.71%
Zea mays	2,655	37,534	30.85%	53.85%	38.73%	53.80%

[1] Genome sizes estimated by flow cytometry, primarily from the Kew C-values database (see Materials and Methods)

[2] Number of genes used in genome wide summaries in Figures 1 and 2, including only genes with near-gene BSseq coverage (see Methods).

[3] The percentage of genes associated with mCHH islands within the flanking 5' or 3' 2.0 kb.

[4] The median level of CHH methylation in islands within 2 kb of genes.

[5] The percent of orthologs, of 2,720 total, associated with a 5' mCHH island in each species.

[6] Median mCHH level for islands associated with orthologs.

Table 1.2: Counts of TEs within 2kb for a common set of TE superfamilies across species and their enrichment status for mCHH islands.

	Barley			**Rice**			**Maize**		
TE Family[1]	**#TEs <2kb**[2]	**% with island**[3]	**Enriched**[4]	**#TEs <2kb**	**% with island**	**Enriched**	**#TEs <2kb**	**% with island**	**Enriched**
DHH	69	0.174	NS	132	0.417	Under	5235	0.283	Under
DTA	21	0.095	NS	517	0.768	NS	653	0.542	**Enriched**
DTC	3480	0.280	NS	1243	0.474	Under	184	0.429	**Enriched**
DTH	478	0.460	**Enriched**	40	0.700	NS	2677	0.536	**Enriched**
DTM	654	0.378	**Enriched**	1106	0.806	**Enriched**	122	0.623	**Enriched**
DTT	474	0.430	**Enriched**	2143	0.898	**Enriched**	2307	0.389	**Enriched**
DTX	332	0.497	**Enriched**	6476	0.882	**Enriched**	299	0.408	**Enriched**
RIX	880	0.227	Under	551	0.611	Under	87	0.253	NS
RLC	4896	0.239	Under	1433	0.651	Under	3148	0.280	Under
RLG	4413	0.213	Under	2230	0.580	Under	3891	0.292	Under
RLX	11356	0.315	**Enriched**	10753	0.704	Under	2632	0.224	Under
RSX	76	0.316	NS	796	0.932	**Enriched**	43	0.302	NS
TOTAL	**27129**	**0.285**		**27420**	**0.747**		**21278**	**0.333**	

[1] TE classification code as described by Wicker et al. (2007). Abbreviations are DHH: *Helitron*; DTA: *hAT*; DTC: *CACTA* ; DTH:*PIF-Harbinger* ; DTM: *Mutator*; DTT: *Tc1-Mariner*; DTX: unknown DNA elements; RIX: unclassified *LINE*; RLC: *Copia*; RLG: *Gypsy*; RLX: unclassified *LTR*; RSX: unclassified *SINE*.

[2] The number of TEs within each class that are within 2kb upstream of an annotated gene, based on counting only the closest TE to a gene.

[3] The proportion of genes that have both an mCHH island and a TE within 2kb upstream.

[4] Based on a binomial test (FDR corrected, p< 0.05), classes of TEs were determined to be enriched for CHH islands or under-enriched, relative to the total proportion estimated across all TE superfamilies. NS = non significant.

Supplemental Figures

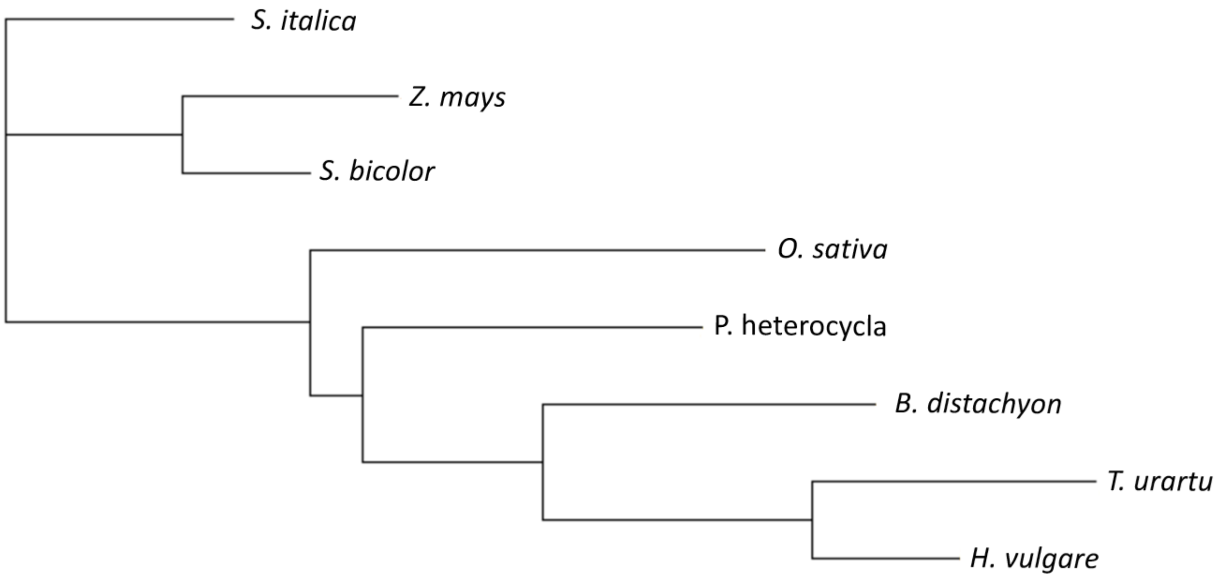
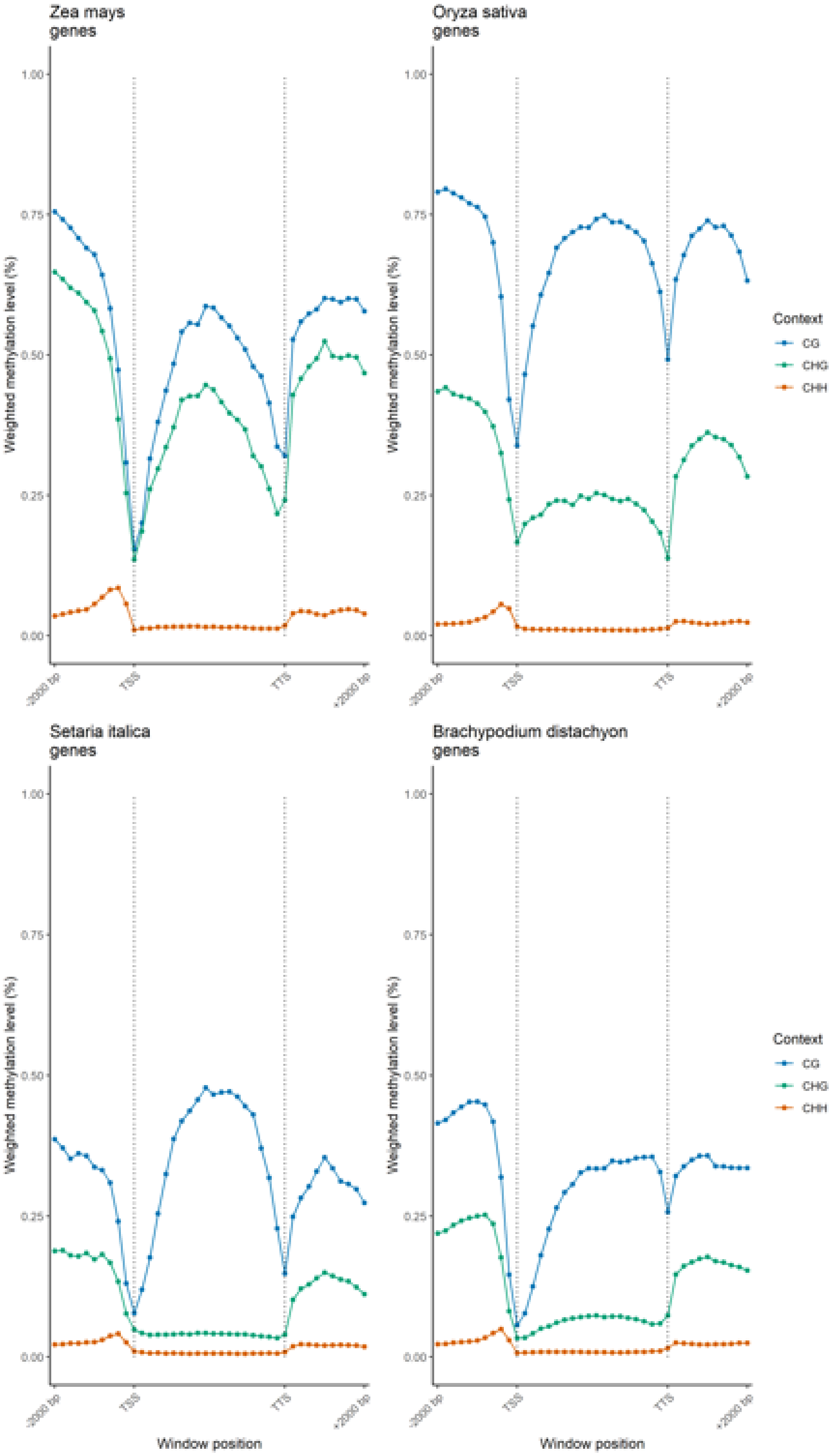


Figure S1.1. Maximum likelihood phylogeny of the eight Poaceae species used in this study, inferred from single-copy orthologous genes.



(cont. next page)

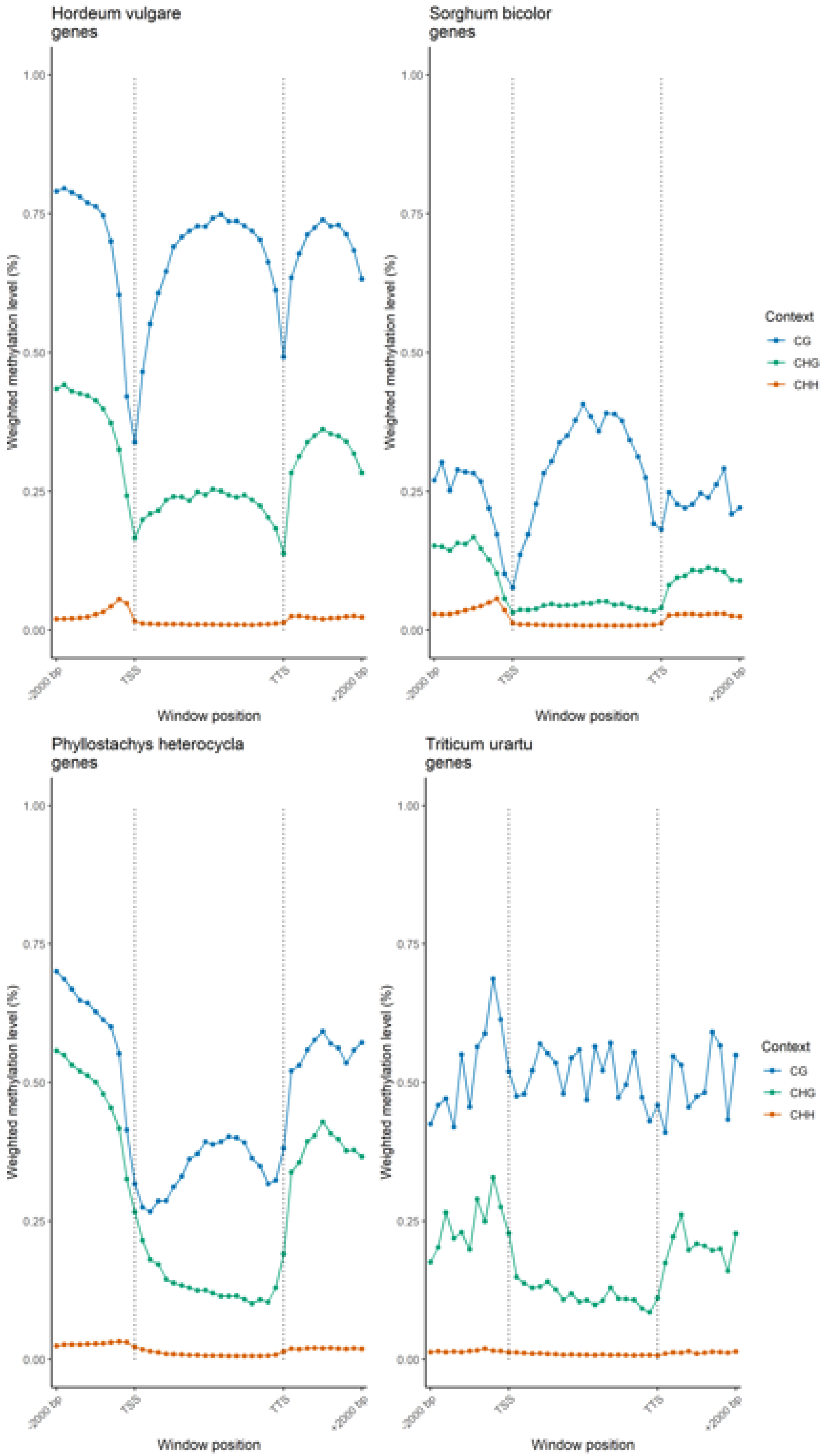


Figure S1.2. Metaprofiles summarizing methylation levels in all three contexts across genes and near-gene regions. Weighted methylation levels were calculated in 100 bp windows 2 kb upstream and downstream of genes, and transcription start sites (TSS) and transcription termination sites (TTS) are marked by dotted lines.

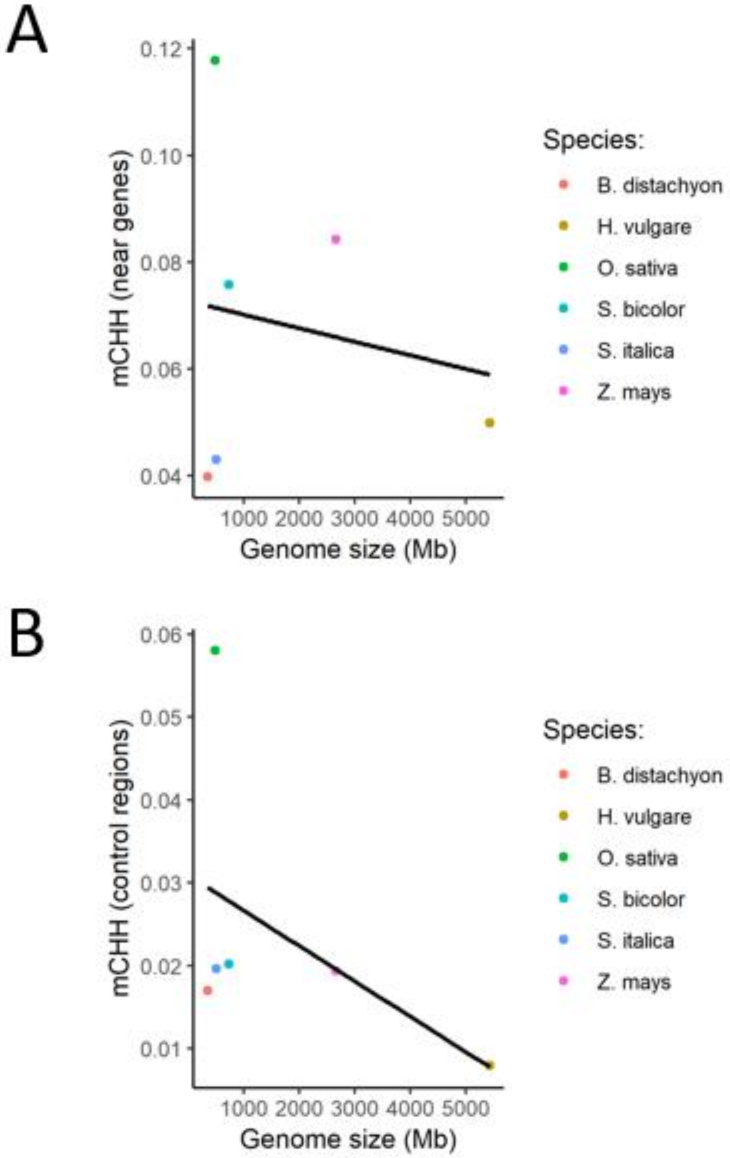


Figure S1.3. Correlations between GS and mCHH levels near genes (S1.3a) and among control sites (S1.3b). These figures represent the numerator and denominator of Figure 1b respectively.

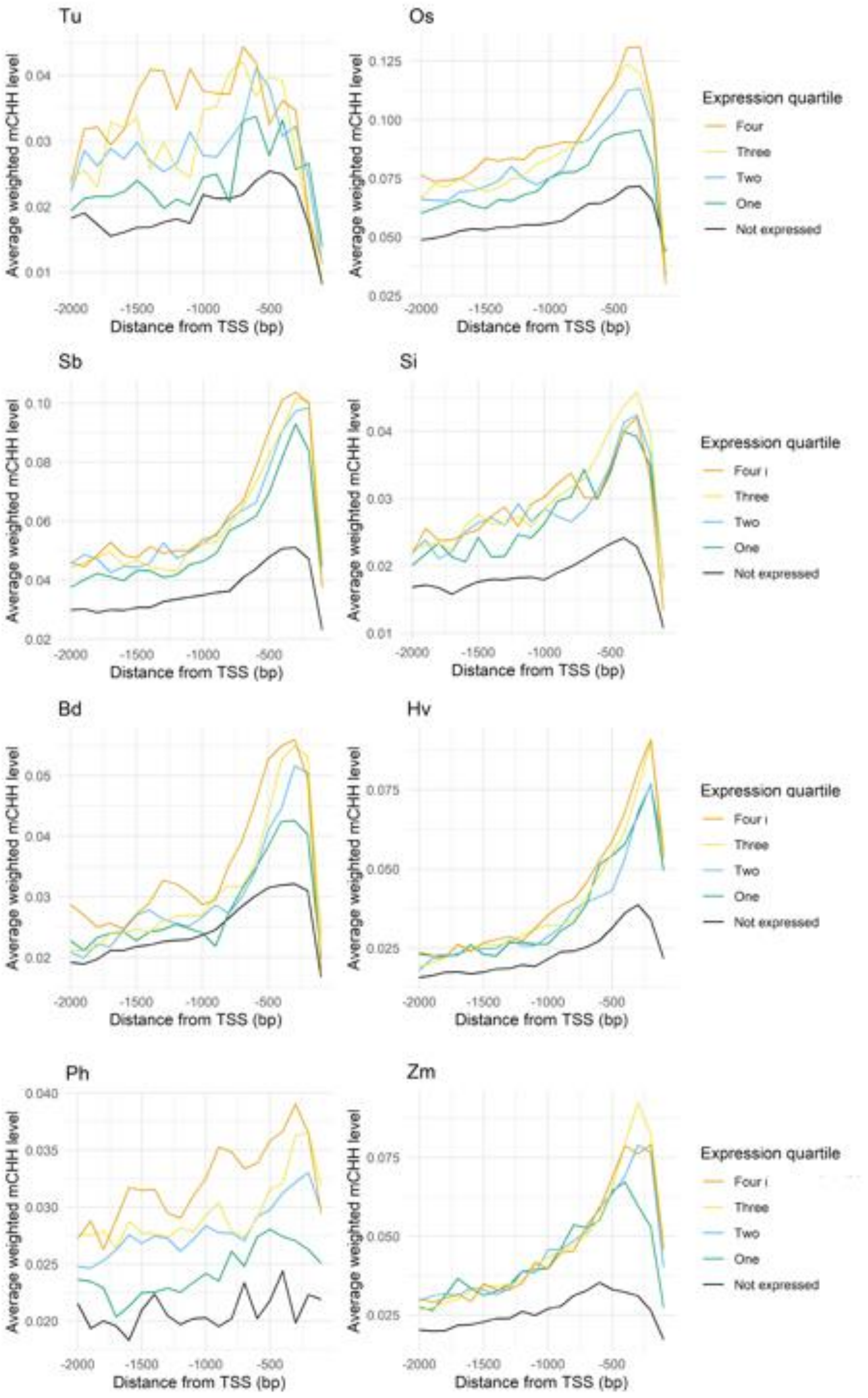


Figure S1.4. Metaprofiles of mCHH levels across 5' near-gene regions in genes separated into non-expressed genes and four quartiles of expression (quartile one being those with the lowest expression, and four with highest).

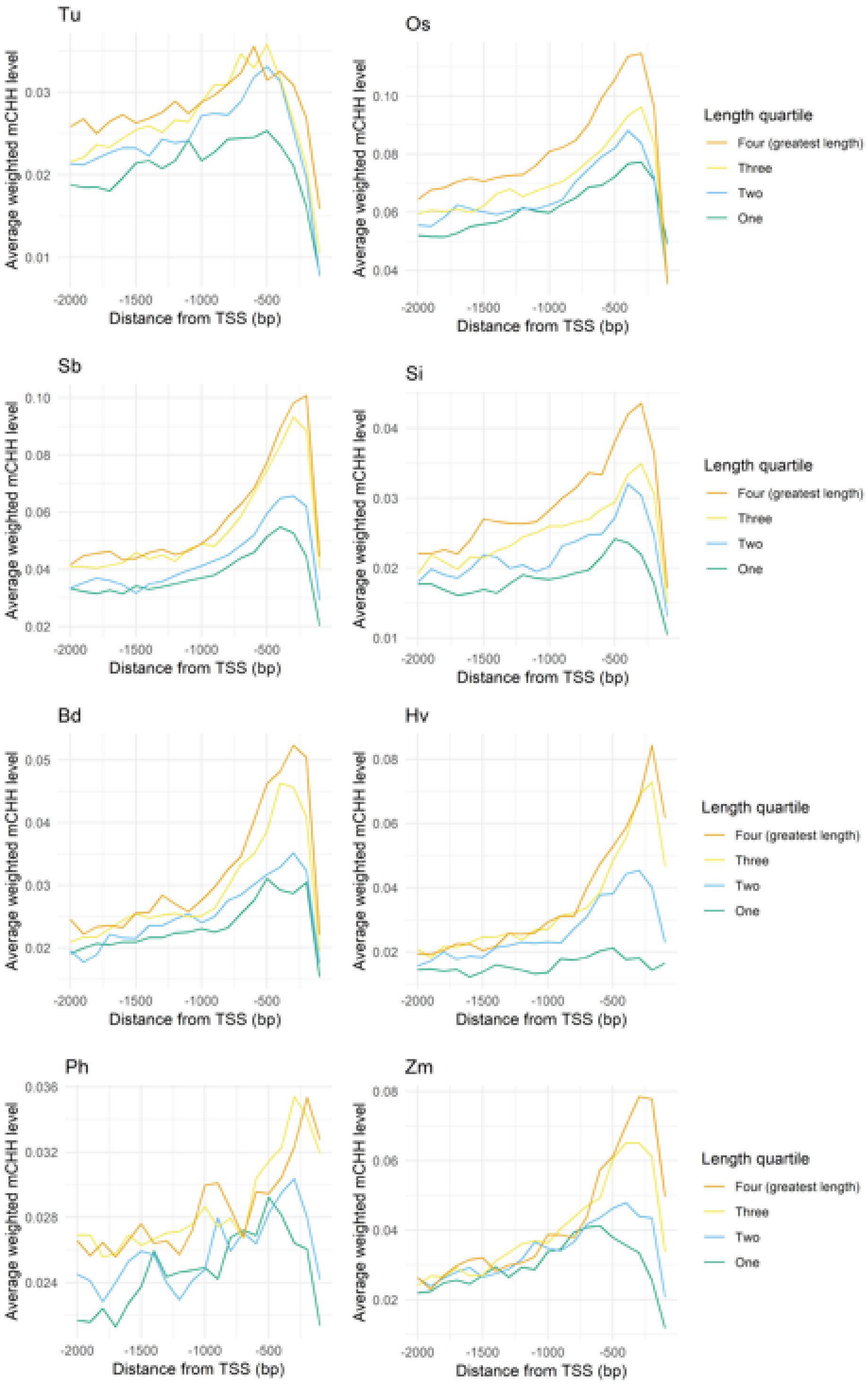


Figure S1.5. Metaprofiles of mCHH levels across 5' near-gene regions in genes separated into four quartiles of length.

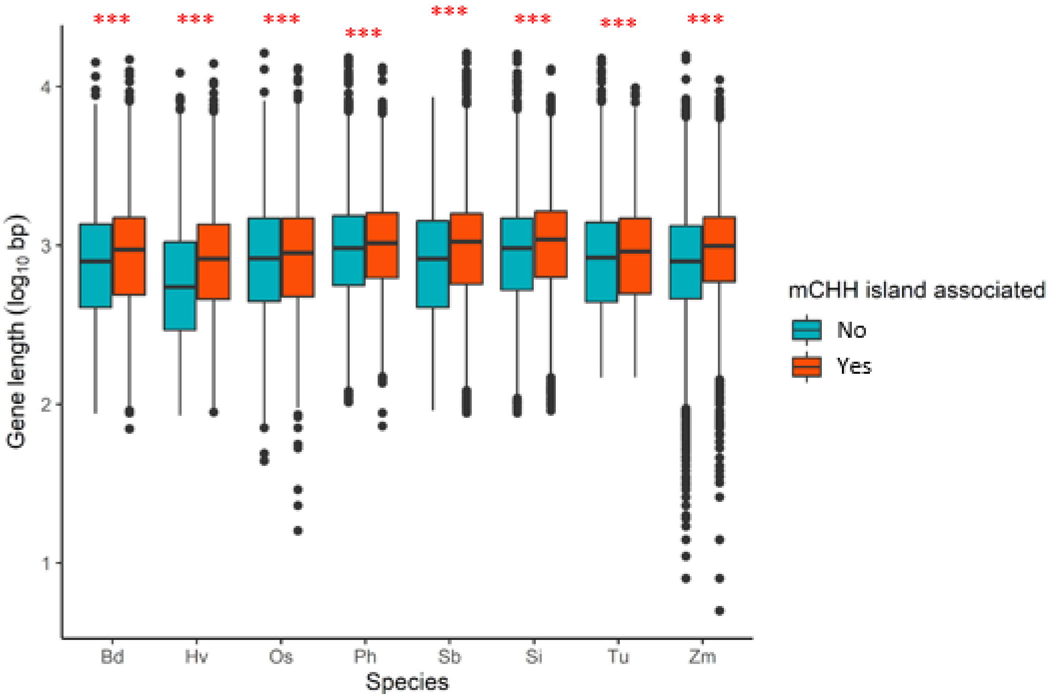


Figure S1.6. Length of longest transcript between mCHH island associated genes and other genes. Boxplots show the four quartiles of the transcript length distributions in each species, and middle lines represent median. Significance was established by replacing full gene length with transcript length in the logistic regression model from the main text.

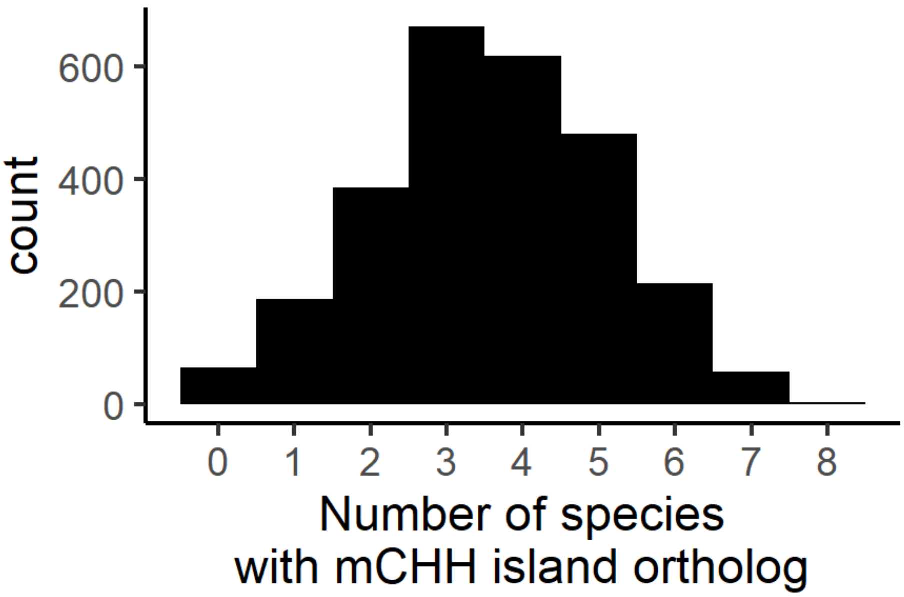


Figure S1.7. Histogram illustrating the distribution of mCHH island conservation among orthologs of the eight species. The numbers on the x-axis represent the number of species in which orthologs were mCHH island associated. The count on the y-axis is the number of orthologs.

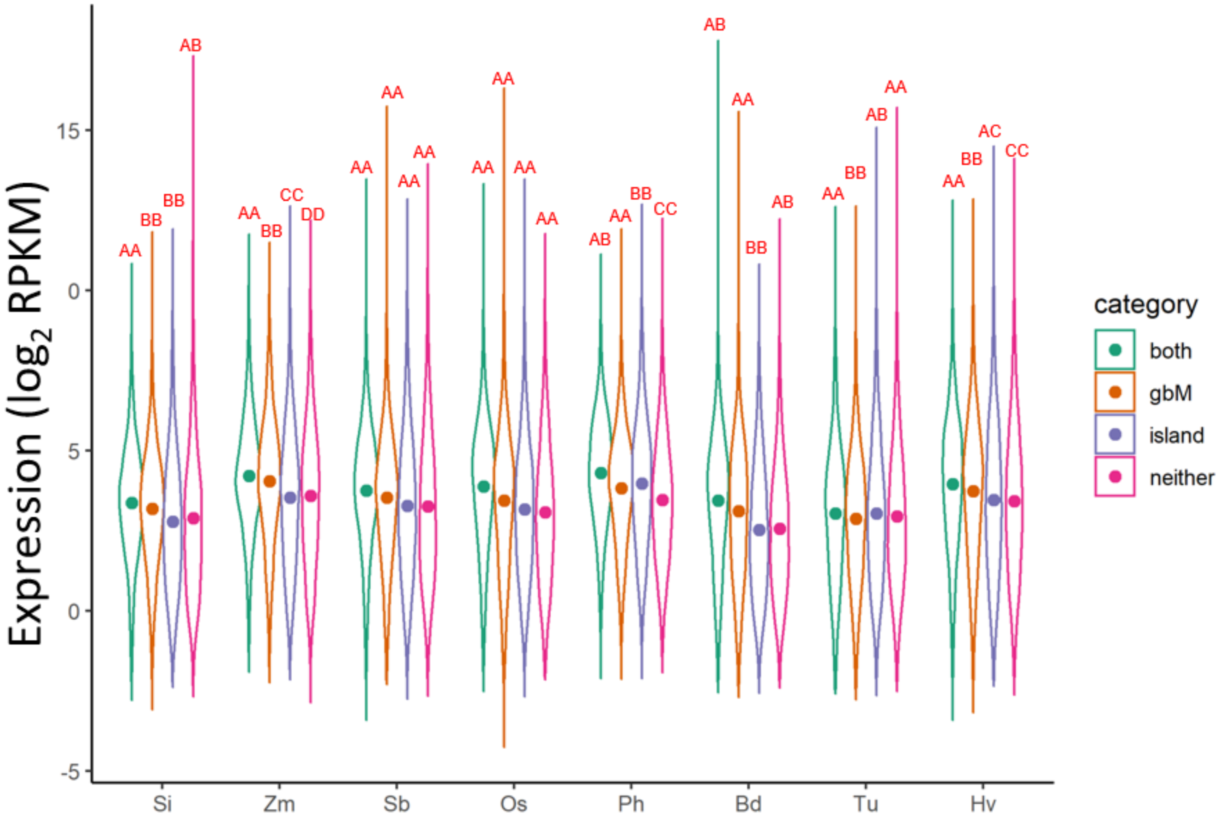


Figure S1.8. Gene expression levels between categories of gbM genes, mCHH island associated genes, genes with both and genes with neither. Genes were separated into four non-overlapping categories: mCHH island only, gbM only, both island and gbM, and neither. Violin plots depict distributions of expression within each category, with mean values marked by dots. Letter codes represent significance (P < 0.05, unpaired t-test), where any distributions that share a letter (e.g., AB and AA) are not significantly different but those that do not share letters (e.g., AA and BB) are significantly different. All comparisons are within, not between, species.

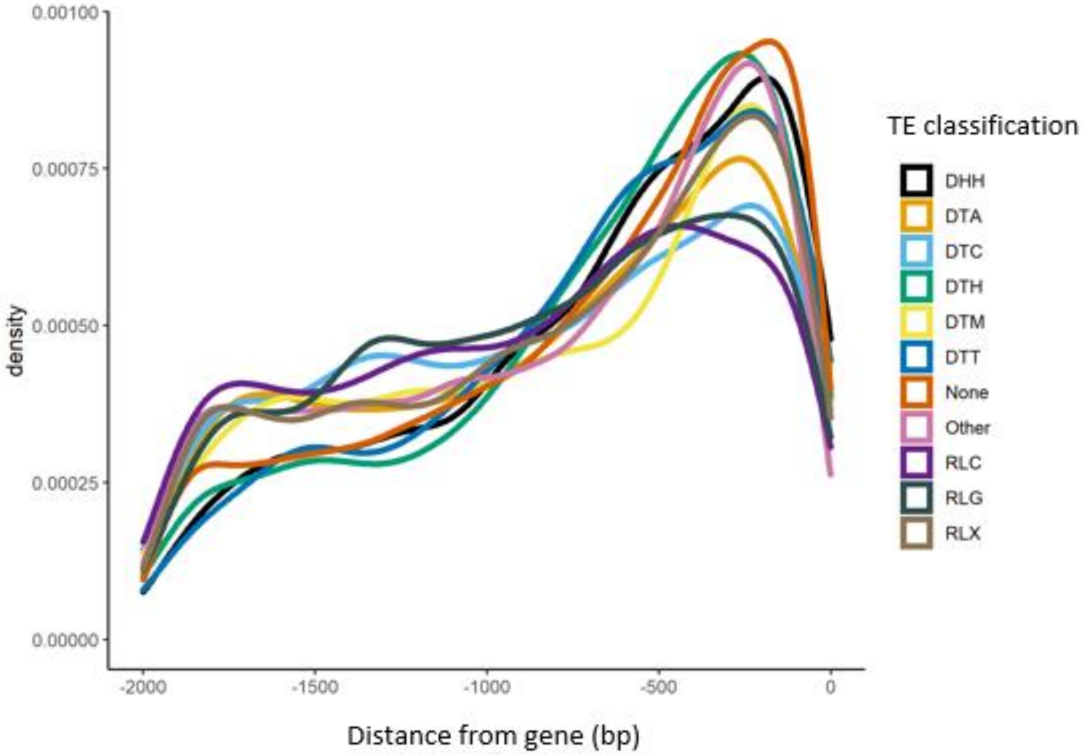


Figure S1.9. Distributions of distances from the gene TSS to mCHH islands located within different types of TEs and not within TEs (“None”). These distributions include data from all three species.

Supplemental Tables

Table S1.1. Reference genome statistics

Species	# contigs	Largest contig size (bp)	Total Length (bp)	N50 (bp)	N50 / Total Length
O. sativa	63	4.33E+07	3.75E+08	3.00E+07	0.07988
H. vulgare	10	7.68E+08	4.83E+09	6.57E+08	0.13595
P. heterocycla	277278	4.87E+06	2.05E+09	3.29E+05	0.00016
S. bicolor	867	8.09E+07	7.09E+08	6.87E+07	0.09687
S. italica	336	5.90E+07	4.06E+08	4.73E+07	0.11646
T. urartu	248855	1.07E+06	3.68E+09	8.76E+04	0.00002
Z. mays	267	3.07E+08	2.14E+09	2.24E+08	0.10487
B. distachyon	10	7.51E+07	2.71E+08	5.91E+07	0.21806

Table S1.2. Logistic regression models of predictors of mCHH island presence

Species	Expression		Distance to nearest TE		Gene length		Exonic mCG	
	Estimate[1]	*p*-value[2]	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value
Oryza sativa	-6.43E-01	7.48E-01	-8.54E-04	0.00E+00	7.09E+00	1.13E-111	-1.36E+00	0.00E+00
Sorghum bicolor	-1.81E+00	3.00E-01	-7.42E-04	0.00E+00	5.62E+00	9.31E-61	-7.35E-01	3.57E-86
Zea mays	2.94E+00	3.12E-05	-3.76E-05	2.52E-41	3.38E+00	1.03E-39	-1.60E+00	0.00E+00
Hordeum vulgare	2.62E+00	3.50E-02	-3.19E-04	9.78E-187	8.93E-01	6.57E-02	-1.08E+00	3.59E-257
Brachypodium distachyon	4.59E-01	8.18E-01	-5.37E-05	1.71E-47	3.74E+00	6.98E-81	-6.44E-01	4.58E-83
Setaria italica	-8.05E-01	7.82E-01	-7.64E-04	0.00E+00	3.53E+00	4.61E-47	-3.69E-01	1.32E-17
Triticum urartu	5.85E-01	6.89E-01	-2.19E-04	4.80E-16	1.38E+00	5.65E-04	-8.38E-01	1.45E-101
Phyllostachys heterocycla	1.03E+01	6.77E-21	-4.03E-04	1.43E-68	4.33E-01	6.45E-02	-7.61E-01	2.03E-45

[1] Estimate of the effect of the predictor

[2] P-value of significance for the estimated effect of the predictor

Table S1.3. Coefficients of variation in mCHH island-gene distance vs mCHH island-TE edge distance

Species[1]	TE classification[2]	Gene distance CV	TE distance CV	P value[3]	Test statistic[3]	# Cases[4]	Mean distance to TE edge	Mean distance to gene TSS
Hv	Other	0.80	1.77	0	4.83E+04	2527	81.26	672.78
Hv	DTM	0.86	0.91	0	8.17E+05	119	77.18	660.48
Hv	RLX	0.76	1.23	0	2.02E+05	3149	127.52	744.76
Hv	RLG	0.69	1.33	0	2.29E+04	696	178.93	813.62
Hv	DTA	NA	NA	NA	NA	1	14.00	93.00
Hv	RLC	0.66	1.16	0	4.39E+04	958	188.97	862.32
Hv	DTC	0.74	1.27	0	3.58E+04	732	110.21	776.24
Hv	DTT	0.82	0.77	0	5.53E+05	124	38.51	643.79
Hv	DTH	0.78	1.08	0	3.26E+04	213	122.81	656.64
Hv	DHH	0.54	2.09	2.74E-07	2.64E+01	6	143.33	885.17
Os	Other	0.67	1.19	0	6.59E+05	14574	72.15	841.81
Os	DTM	0.67	1.07	0	4.21E+04	600	118.71	852.06
Os	RLX	0.66	1.62	0	7.64E+04	4963	191.86	859.51
Os	RLG	0.60	1.24	0	2.26E+04	821	383.82	935.37
Os	DTA	0.65	1.16	0	8.90E+03	200	88.42	884.95
Os	RLC	0.62	1.49	0	1.10E+04	648	259.71	926.34
Os	DTC	0.58	1.46	0	5.10E+03	334	311.80	977.07
Os	DTT	0.77	0.55	3.26E-127	5.76E+02	5	66.60	948.60
Os	DTH	0.53	1.84	2.23E-20	8.56E+01	13	188.69	935.85
Os	DHH	0.47	0.68	4.47E-75	3.36E+02	4	105.00	396.25
Zm	Other	0.63	1.06	0	3.70E+03	64	135.41	816.39

Zm	DTM	1.05		2.07	0	1.56E+03	64	121.58	515.03
Zm	RLX	0.70		0.89	0	8.44E+04	302	706.97	767.25
Zm	RLG	0.56		1.29	0	1.16E+04	561	234.79	865.79
Zm	DTA	0.73		0.85	0	1.97E+05	305	117.78	778.33
Zm	RLC	0.54		0.98	0	2.00E+04	471	372.23	1012.97
Zm	DTC	0.50		1.16	3.59E-231	1.05E+03	51	115.24	964.25
Zm	DTT	0.70		1.31	0	1.59E+04	425	94.87	731.58
Zm	DTH	0.73		1.11	0	9.19E+04	1049	79.44	728.49
Zm	DHH	0.76		1.03	0	1.95E+05	1149	552.83	683.98

[1] Hv = *Hordeum vulgare*, Os = *Oryza sativa,* Zm = *Zea mays*

[2] TE classification codes from Wicker 2007

[3] Fultz and Miller asymptotic test for CV equality

[4] Total number of TE-mCHH islands analyzed in each category

Table S1.4. Coincidence vs dissonance of lineage-specific mCHH islands and TEs

Comparison (TE, species 1, species 2)[1]	**Observed unique island/TE coincidence[2]**	**Expected unique island/TE coincidence[3]**	**Enrichment unique island/TE coincidence[4]**	**Observed unique island/TE dissonance**	**Expected unique island/TE dissonance**	**Enrichment unique island/TE dissonance**	**Chi-square statistic[5]**	**P value[5]**
DHH_HvOs	2	2.63	0.76	0	2.25	0.00	7.70	2.609E-01
DHH_ZmHv	87	80.35	1.08	84	81.80	1.03	6.90	3.302E-01
DHH_ZmOs	24	30.92	0.78	166	156.37	1.06	9.56	1.444E-01
DTA_HvOs	10	17.10	0.58	4	4.03	0.99	10.84	9.334E-02
DTA_ZmHv	15	11.79	1.27	7	11.99	0.58	8.97	1.754E-01
DTA_ZmOs	22	20.65	1.07	18	24.81	0.73	21.00	1.265E-02
DTC_HvOs	20	19.43	1.03	80	96.00	0.83	10.48	1.058E-01
DTC_ZmHv	63	52.26	1.21	51	51.40	0.99	14.77	9.746E-02
DTC_ZmOs	7	2.78	2.51	4	5.43	0.74	13.71	3.302E-02
DTH_HvOs	9	5.58	1.61	17	23.14	0.73	9.01	1.731E-01
DTH_ZmHv	91	64.99	1.40	41	65.78	0.62	62.02	5.467E-10
DTH_ZmOs	29	21.00	1.38	69	107.15	0.64	47.43	1.536E-08
DTM_HvOs	39	31.14	1.25	28	32.26	0.87	8.12	2.294E-01
DTM_ZmHv	21	16.51	1.27	16	16.30	0.98	9.46	3.962E-01

DTM_ZmOs	32	27.31	1.17	4	9.93	0.40	18.56	4.970E-03
DTT_HvOs	8	4.20	1.90	14	18.02	0.78	12.34	5.477E-02
DTT_ZmHv	68	60.47	1.12	53	61.26	0.87	11.91	2.187E-01
DTT_ZmOs	22	18.97	1.16	95	100.96	0.94	7.93	5.416E-01
RLC_HvOs	63	71.72	0.88	201	171.07	1.17	19.10	2.432E-02
RLC_ZmHv	119	127.99	0.93	113	127.36	0.89	26.53	1.670E-03
RLC_ZmOs	72	63.94	1.13	110	109.66	1.00	5.57	7.824E-01
RLG_HvOs	56	67.80	0.83	167	147.35	1.13	11.49	2.434E-01
RLG_ZmHv	121	127.79	0.95	133	127.56	1.04	20.18	1.684E-02
RLG_ZmOs	63	67.80	0.93	116	128.26	0.90	8.70	4.650E-01
RLX_HvOs	304	269.22	1.13	335	359.98	0.93	36.97	2.660E-05
RLX_ZmHv	239	216.58	1.10	180	213.27	0.84	17.65	3.948E-02
RLX_ZmOs	308	288.15	1.07	84	93.56	0.90	17.20	4.562E-02

[1] Hv = *Hordeum vulgare*, Os = *Oryza sativa,* Zm = *Zea mays.* TE classification codes from Wicker 2007

[2] See model in main text for coincidence/dissonance definitions. These numbers represent the observed occurrences in each comparison

[3] Expected values calculated by product of proportions of lineage specific TEs and proportions of lineage specific mCHH islands

[4] Enrichment = observed / expected

[5] Chi-square test of equality between observed and expected proportions of lineage specific mCHH islands and TEs

Table S1.5. Transposable element and repeat annotations

Species	Repeat annotation	Source
Oryza sativa	irgsp1_repeat_unit.gff	https://rapdb.dna.affrc.go.jp/download/irgsp1.html
Sorghum bicolor	Sbicolor_454_v3.1.1.repeatmasked_assembly_v3.0.1.gff3.gz	https://phytozome.jgi.doe.gov/
Zea mays	B73.structuralTEv2.fulllength.2018-09-19.gff3	https://mcstitzer.github.io/maize_TEs/
Hordeum vulgare	Barley_TE_annotation_v2_18Aug16.tsv	https://doi.org/10.5447/IPK/2016/16
Brachypodium distachyon	Bdistachyon_556_v3.2.repeatmasked_assembly_v3.0.gff3.gz	https://phytozome.jgi.doe.gov/
Setaria italica	Sitalica_312_v2.2.repeatmasked_assembly_v2.gff3.gz	https://phytozome.jgi.doe.gov/
Triticum urartu	Triticum_urartu.GCA_000347455.1.30.gff3	https://plants.ensembl.org/
Phyllostachys heterocycla	P_heterocycla_v1.0.repeats.detail	http://server.ncgr.ac.cn/bamboo/down.php

Table S1.6. Genome assembly information

Species	Bisulfite seq mean coverage	Genome assembly	Source
Oryza sativa	60.3	Osativa_323_v7.0.fa	https://phytozome.jgi.doe.gov/
Sorghum bicolor	17.87	Sbicolor_313_v3.0.fa	https://phytozome.jgi.doe.gov/
Zea mays	6.7	Zea_mays.AGPv4.dna.chromosome.all.fa	http://plants.ensembl.org/index.html
Hordeum vulgare	16.99	150831_barley_pseudomolecules.fa	https://webblast.ipk-gatersleben.de/
Brachypodium distachyon	81.07	Bdistachyon_314_v3.0.fa	https://phytozome.jgi.doe.gov/
Setaria italica	17.88	Sitalica_312_v2.fa	https://phytozome.jgi.doe.gov/
Triticum urartu	14.88	Triticum_urartu.ASM34745v1.31.dna.genome.fa	http://plants.ensembl.org/index.html
Phyllostachys heterocycla	18.17	P_heterocycla_v1.0.Scaffolds.fa	http://server.ncgr.ac.cn/bamboo/

References

Alonso, C., Pérez, R., Bazaga, P., and Herrera, C.M. (2015). Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. Front. Genet. 6.

Bennetzen, J.L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A.C., Estep, M., Feng, L., Vaughn, J.N., Grimwood, J., et al. (2012). Reference genome sequence of the model plant Setaria. Nature Biotechnology 30, 555–561.

Bewick, A.J., and Schmitz, R.J. (2017). Gene body DNA methylation in plants. Curr. Opin. Plant Biol. 36, 103–110.

Bewick, A.J., Niederhuth, C.E., Ji, L., Rohr, N.A., Griffin, P.T., Leebens-Mack, J., and Schmitz, R.J. (2017). The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. Genome Biology 18, 65.

Bewick, A.J., Zhang, Y., Wendte, J.M., Zhang, X., and Schmitz, R.J. (2019). Evolutionary and Experimental Loss of Gene Body Methylation and Its Consequence to Gene Expression. G3: Genes, Genomes, Genetics 9, 2441–2445.

Choi, J.Y., and Lee, Y.C.G. (2020). Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. PLOS Genetics 16, e1008872.

Cuerda-Gil, D., and Slotkin, R.K. (2016). Non-canonical RNA-directed DNA methylation. Nature Plants 2, 1–8.

Diez, C.M., Roessler, K., and Gaut, B.S. (2014). Epigenetics and plant genome evolution. Current Opinion in Plant Biology 18, 1–8.

Dodsworth, S., Leitch, A.R., and Leitch, I.J. (2015). Genome size diversity in angiosperms and its influence on gene space. Current Opinion in Genetics & Development 35, 73–78.

Feltz, C.J., and Miller, G.E. (1996). An asymptotic test for the equality of coefficients of variation from k populations. Statistics in Medicine 15, 647–658.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. PNAS 107, 8689–8694.

Forestan, C., Farinati, S., Aiese Cigliano, R., Lunardon, A., Sanseverino, W., and Varotto, S. (2017). Maize RNA PolIV affects the expression of genes with nearby TE insertions and has a genome-wide repressive impact on transcription. BMC Plant Biol 17.

Fultz, D., Choudury, S.G., and Slotkin, R.K. (2015). Silencing of active transposable elements in plants. Current Opinion in Plant Biology 27, 67–76.

Gaut, B.S., and Ross-Ibarra, J. (2008). Selection on Major Components of Angiosperm Genomes. Science 320, 484–486.

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.

Guo, W., Wang, D., and Lisch, D. (2021). RNA-directed DNA methylation prevents rapid and heritable reversal of transposon silencing under heat stress in *Zea mays*. BioRxiv 2021.01.08.425849. (PREPRINT).

Harris, C.J., Scheibe, M., Wongpalee, S.P., Liu, W., Cornett, E.M., Vaughan, R.M., Li, X., Chen, W., Xue, Y., Zhong, Z., et al. (2018). A DNA methylation reader complex that enhances gene transcription. Science 362, 1182.

International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463, 763–768.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. Nature 436, 793–800.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, et al. (2020). caret: Classification and Regression Training.

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.

Le, N.T., Harukawa, Y., Miura, S., Boer, D., Kawabe, A., and Saze, H. (2020). Epigenetic regulation of spurious transcription initiation in Arabidopsis. Nature Communications 11, 3224.

Lee, S.-I., and Kim, N.-S. (2014). Transposable elements and genome size variations in plants. Genomics Inform 12, 87–97.

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.

Ling, H.-Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y., et al. (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. Nature 496, 87–90.

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.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell 133, 523–536.

Lu, Z., Marand, A.P., Ricci, W.A., Ethridge, C.L., Zhang, X., and Schmitz, R.J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat. Plants 5, 1250–1259.

Long, J., Liu, J., Xia, A., Springer, N.M., and He, Y. (2021). Maize Decrease in DNA methylation 1 targets RNA-directed DNA methylation on active chromatin. Plant Cell.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P.E., Russell, J., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature 544, 427–433.

Matzke, M.A., and Mosher, R.A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nature Reviews Genetics 15, 394–408.

Muyle, A., Ross-Ibarra, J., Seymour, D.K., and Gaut, B.S. (2020). Investigation Gene body methylation is under selection in Arabidopsis thaliana. BioRxiv 2020.09.04.283333. (PREPRINT).

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T., and Wessler, S.R. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461, 1130–1134.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. Nature 543, 72–77.

Niederhuth, C.E., Bewick, A.J., Ji, L., Alabady, M.S., Kim, K.D., Li, Q., Rohr, N.A., Rambani, A., Burke, J.M., Udall, J.A., et al. (2016). Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 17, 194.

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. Philosophical Transactions of the Royal Society B: Biological Sciences 375, 20190417.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature 457, 551–556.

Peng, Z., Lu, Y., Li, L., Zhao, Q., Feng, Q., Gao, Z., Lu, H., Hu, T., Yao, N., Liu, K., et al. (2013). The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla). Nature Genetics 45, 456–461.

Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S., and Fischer, R.L. (2007). DNA demethylation in the Arabidopsis genome. PNAS 104, 6752–6757.

Raju, S.K.K., Ritter, E.J., and Niederhuth, C.E. (2019). Establishment, maintenance, and biological roles of non-CG methylation in plants. Essays in Biochemistry 63, 743–755.

Roessler, K., Takuno, S., and Gaut, B.S. (2016). CG Methylation Covaries with Differential Gene Expression between Leaf and Floral Bud Tissues of Brachypodium distachyon. PLOS ONE 11, e0150002.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C., Iwamoto, M., Abe, T., et al. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. 54, e6.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326, 1112–1115.

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.

Seymour, D.K., and Gaut, B.S. (2020). Phylogenetic Shifts in Gene Body Methylation Correlate with Gene Expression and Reflect Trait Conservation. Molecular Biology and Evolution 37, 31–43.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8, 272–285.

Stitzer, M.C., Anderson, S.N., Springer, N.M., and Ross-Ibarra, J. (2019). The Genomic Ecosystem of Transposable Elements in Maize. BioRxiv doi.org/10.1101/559922.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J., and Jacobsen, S.E. (2014). Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat. Struct. Mol. Biol. 21, 64–72.

Symonds, M.R.E., and Blomberg, S.P. (2014). A Primer on Phylogenetic Generalised Least Squares. In Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice, L.Z. Garamszegi, ed. (Berlin, Heidelberg: Springer), pp. 105–130.

Takuno, S., and Gaut, B.S. (2012). Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol. Biol. Evol. 29, 219–227.

Takuno, S., and Gaut, B.S. (2013). Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. PNAS 110, 1797–1802.

Takuno, S., Ran, J.-H., and Gaut, B.S. (2016). Evolutionary patterns of genic DNA methylation vary across land plants. Nat Plants 2, 15222.

Takuno, S., Seymour, D.K., and Gaut, B.S. (2017). The Evolutionary Dynamics of Orthologs That Shift in Gene Body Methylation between Arabidopsis Species. Mol Biol Evol 34, 1479–1491.

Teissandier, A., and Bourc'his, D. (2017). Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. The EMBO Journal 36, 1471–1473.

Vidalis, A., Živković, D., Wardenaar, R., Roquis, D., Tellier, A., and Johannes, F. (2016). Methylome evolution in plants. Genome Biology 17, 264.

Wicker, T., Matthews, D.E., and Keller, B. (2002). TREP: a database for Triticeae repetitive elements. Trends in Plant Science 7, 561–562.

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.

Wicker, T., Schulman, A.H., Tanskanen, J., Spannagl, M., Twardziok, S., Mascher, M., Springer, N.M., Li, Q., Waugh, R., Li, C., et al. (2017). The repetitive landscape of the 5100 Mbp barley genome. Mobile DNA 8, 22.

Zemach, A., Kim, M.Y., Silva, P., Rodrigues, J.A., Dotson, B., Brooks, M.D., and Zilberman, D. (2010). Local DNA hypomethylation activates genes in rice endosperm. PNAS 107, 18729–18734.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328, 916–919.

Zemach, A., Kim, M.Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell 153, 193–205.

Zhang, H., Lang, Z., and Zhu, J.-K. (2018). Dynamics and function of DNA methylation in plants. Nature Reviews Molecular Cell Biology 19, 489.

Zheng, B., Wang, Z., Li, S., Yu, B., Liu, J.-Y., and Chen, X. (2009). Intergenic transcription by RNA Polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. Genes Dev. 23, 2850–2860.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat. Genet. 39, 61–69.

CHAPTER 2

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

2.1 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.

2.2 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 (Figure S2.1). 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).

2.3 Results

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

We adopted two complementary bioinformatic methods to identify miRNA-like hairpin regions (Figure 2.1a). 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 2.1).

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 2.2)]. 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 2.2). 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 2.1). 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 2.1; Figure 2.1a,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 2.1). 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 Figure S2.2.] Overall, 76% (286,774 of 373,485) of features were RF-structured—i.e., contained regions of miRNA-like structures by this criterion (Table 2.2).

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 2.2).

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)(Figure 2.1a; see Methods for details). We called sequences that fit these criteria "LP-hairpins" (Table 2.1).

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 (Figure 2.1c)($R^2 = 0.73$ 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 (Figure 2.1a). Across all of the 287,744 RF-structured features (Table 2.2), 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 x 10^8 nt vs 1.7x 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 2.2). Most (71.0%) of this set were RF-structured (Table 2.2), 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 2.2).

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 2.2). 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 2.2). *SINEs*/RST also had very low incidences of miRNA-like structure, with no RF-structured elements and <2% containing LP-hairpins (Figure 2.1b). 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 2.2; Figure 2.1b). 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 ($P < 2.2 \times 10^{-16}$, $R^2 = 0.20$, linear model; Figure S2.3).

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 2.2), 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 2.2. 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 2.2). For each feature category, we separately mapped the positions of lowMFE regions and LP-hairpins along their lengths (Figure 2.2). 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 (Figure S2.4), but LP-hairpins were biased heavily towards the terminal inverted repeats for TIR elements like *Mutator*/DTM (Figure 2.2), *hAT*/DTA and *CACTA*/DTC elements (Figure S2.4). Finally, *Helitrons*/DHH had a distinct 3' bias for both lowMFE regions and LP-hairpins (Figure 2.2). 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.

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 (Figure S2.5) 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 2.2). 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 (Figure 2.2), 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 (Figure 2.2), 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 2.2). 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 S2.1) 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 2.3). Moreover, the results were significantly positive for all RNAfold and LinearPartition summary metrics (Table

2.3). 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 S2.2).

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 S2.2), 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 (Figure S2.6) 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 2.3&S2) and with the observation that Dicer-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 S2.2).

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 2.1 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 (Figure 2.3a,b) and the lowMFE regions of RF-structured elements (Figure S2.7). 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 (Figure 2.3a). 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 x 10^{-04}; Table S2.2; Figure S2.8 & 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 (Figure 2.3a). These observations were largely supported by mixed effects models (Table S2.3 & 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 x 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 (Figure S2.6)—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 ≅ 0; Table S2.3 & 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 (Figure S2.10).

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 (Figure 2.3b), 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 (Figure 2.3b).

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 (Figure 2.4) and lowMFE regions (Figure S2.11) 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 (Figure 2.4). 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 x 10^{-81} and 1.16 x 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 (Figure 2.3c & Table S2.3-S2.4) and methylation (Figure 2.4 & 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 (Figure 2.2) 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 2.2); 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 x 10^{-16})(Figure 2.5a), and this was true for all tissues (Figure S2.12) as well as for genes that contained LP-hairpins (Figure S2.13). 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 x 10^{-4}; Figure 2.5a). 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 (Figure 2.5b). The opposite was true among unstructured genes because higher expression occurred with lower MFEs (Figure 2.5b). 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 (Figure. S2.14) and for the complete gene set of genes—i.e., not just genes with syntelogs (Figure S2.15).

Among syntelogs, structured genes also mapped significantly more smRNAs than unstructured genes (Figure 2.5c), 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)(Figure 2.5d). 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 (Figure 2.5a). 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 (Figure 2.5c), 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 Figure 2.4b), 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 (Figure 2.5e). 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 (Figure 2.5e), 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.

2.4 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 (Figure 2.2), 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 2.2); their locational biases in some TE superfamilies (Figure 2.2); their association with elevated smRNA counts in TEs and genes (Figure 2.3); and their genome-wide correspondence to peaks of methylation (Figure 2.4).

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 2.2) and also the lowest average minMFE, reflecting previously recognized regions of strong secondary structure (Figure 2.1). Our other positive control set, miRNA precursor loci, have a high proportion of RF-structure and the highest proportion of LP-hairpins (Table 2.2). 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 2.2 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 (Figure 2.3). 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 Dicer-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 libraries, but we 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 (Figure S2.16), 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 particular 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 between siRNA loci and miRNA-like structures in genes and in all TE superfamilies except SINEs and

LINEs (Table S2.5), which generally lack miRNA-like structures (Table 2.2). We repeated this exercise with a set of annotated small RNA loci that do not produce 21-24 nt smRNAS (Lunardon et al., 2020), revealing no notable enrichment within TEs with a very slight enrichment within mRNAs (Table S2.5). 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, t 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 2.3), 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 S2.2). 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 S2.2). 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 (Figure 2.4). 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 (Figure 2.4)—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 (Figure 2.1), but non-LTR RNA elements (LINEs and SINEs) contain almost no miRNA-like structures (Table 2.2). 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 (Figure 2.2). Finally, *Helitrons*/DHH warrant separate mention because 84% are RF-structured, with a strong bias of LP-hairpins at the 3' end (Figure 2.2). 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 assumes 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 (Figure 2.2), 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 2.2), with an evident bias toward 5' UTR regions (Figure 2.2). 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 (Figure 2.4A,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 (Figure 2.5B). 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 (Figure 2.5B) and have more variable expression among genotypes (Figure 2.5C). 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 (Figure 2.5E), 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).

2.5 Materials and 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 BEDtools 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 S2.1. 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 S2.6. 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. SAMtools v1.10 (Danecek et al., 2021) was used to convert and sort the alignment output. BEDtools 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 2.3), 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:

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

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 (Figure 2.4) 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_J an2021_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 bigWig 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.

Figures

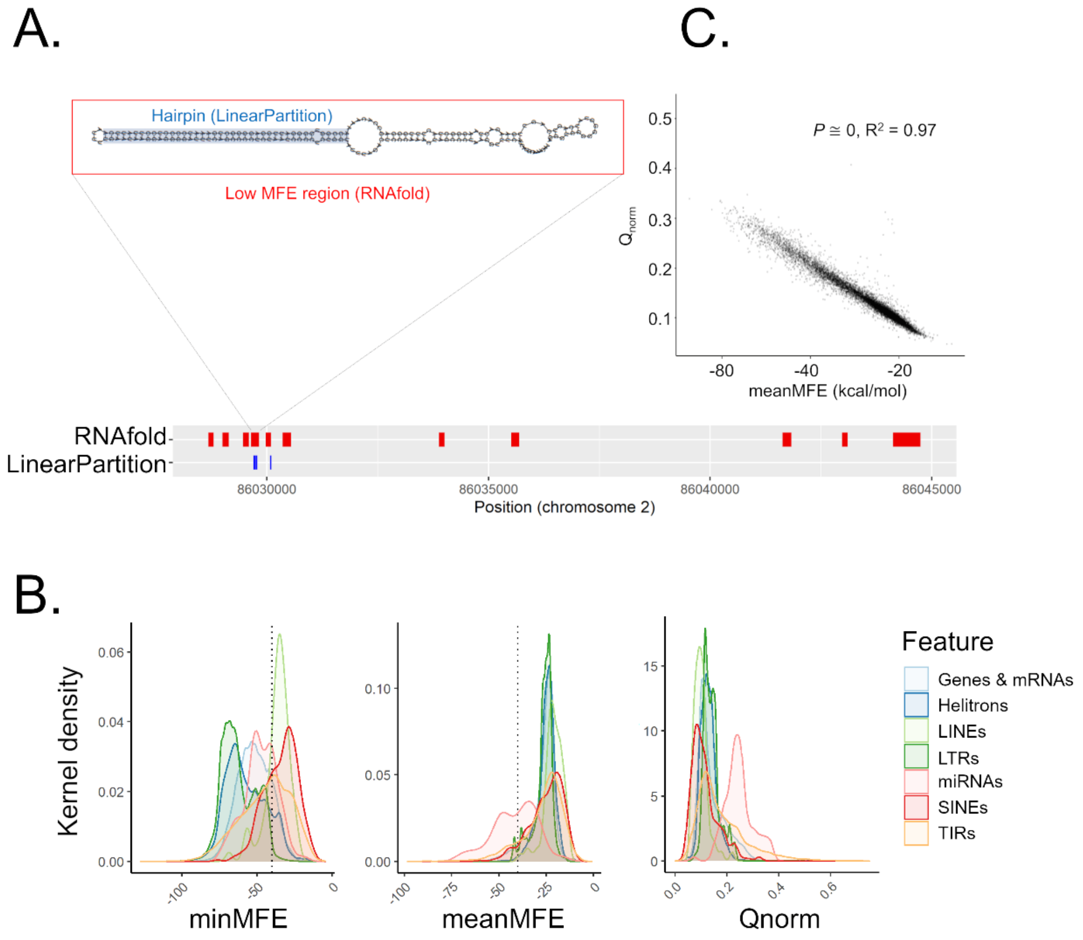


Figure 2.1: Characteristics of miRNA-like secondary structure across two methods. (A) A schematic contrasting the two prediction methods for a genic region on Chromosome 2. The LinearPartition (LP) method focuses on identifying small regions with hairpin characteristics, while the RNAfold method focuses 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 Qnorm based on 39,179 genes. (C) The distributions of three summary statistics—minMFE, meanMFE and Qnorm —across seven feature categories. In the key, helitrons correspond to DHH elements (see Table 2.2 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.

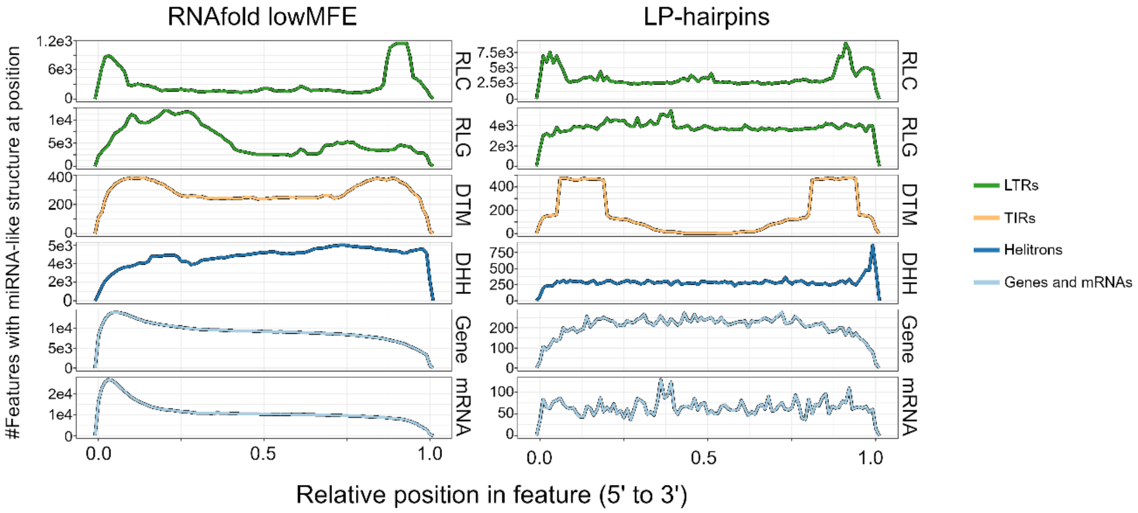


Figure 2.2. Landscapes of miRNA-like regions across feature types. Each row represents a metaprofile that combines data from all members of each feature type, based on structured members. Features were divided into 100 equally sized bins from the 5' end to the 3' end. The left column shows the number 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 2.2; the remainder of the categories are shown in Figure S2.4.

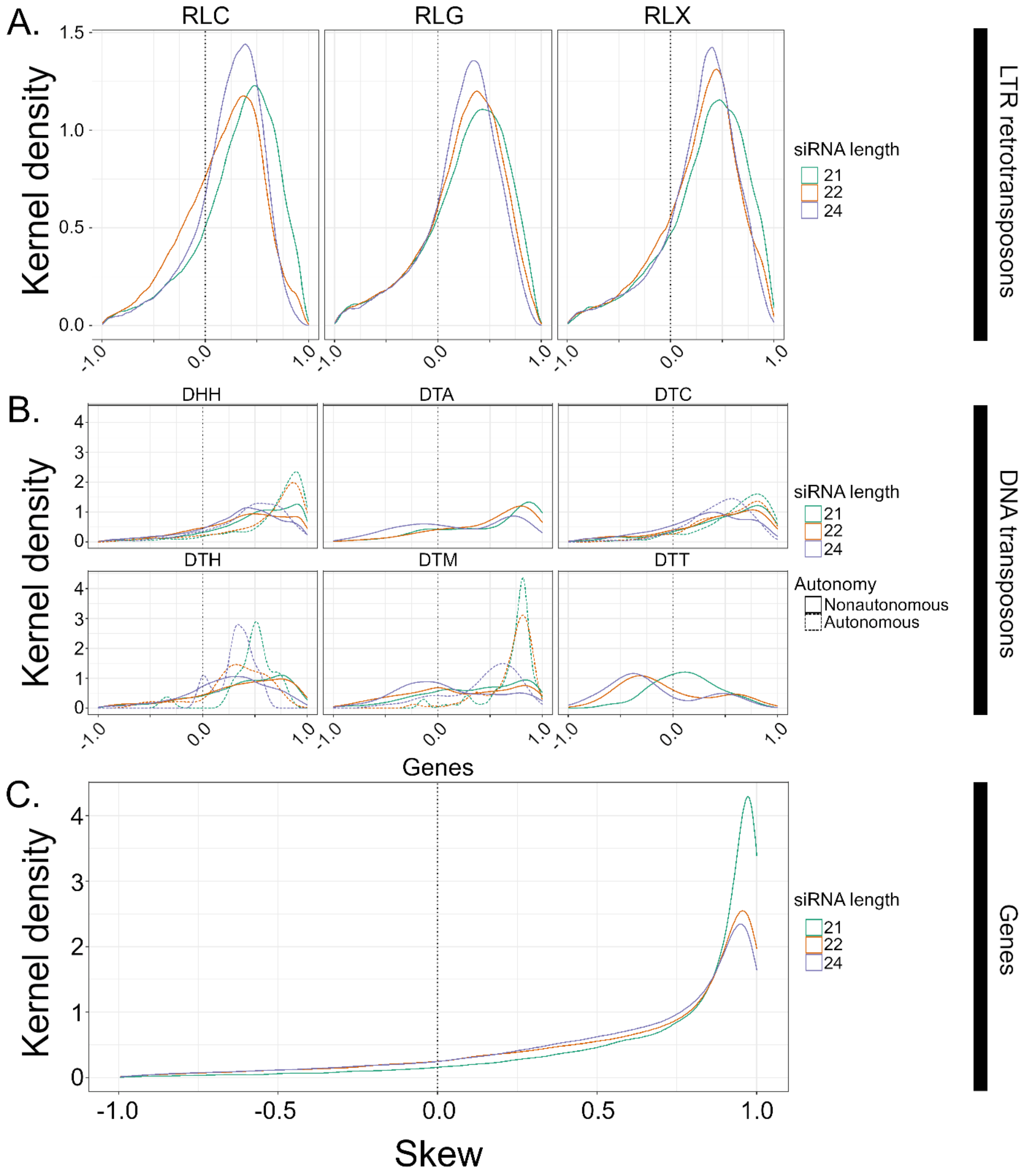


Figure 2.3. The distribution of skew for smRNA mapping in different feature categories. Skew is presented on the x-axis. Height on the y-axis represents the Gaussian estimated kernel density of skew values. Skew measures the relative enrichments of smRNAs in miRNA-like regions compared to non-miRNA regions 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 2.2. 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 S2.7.

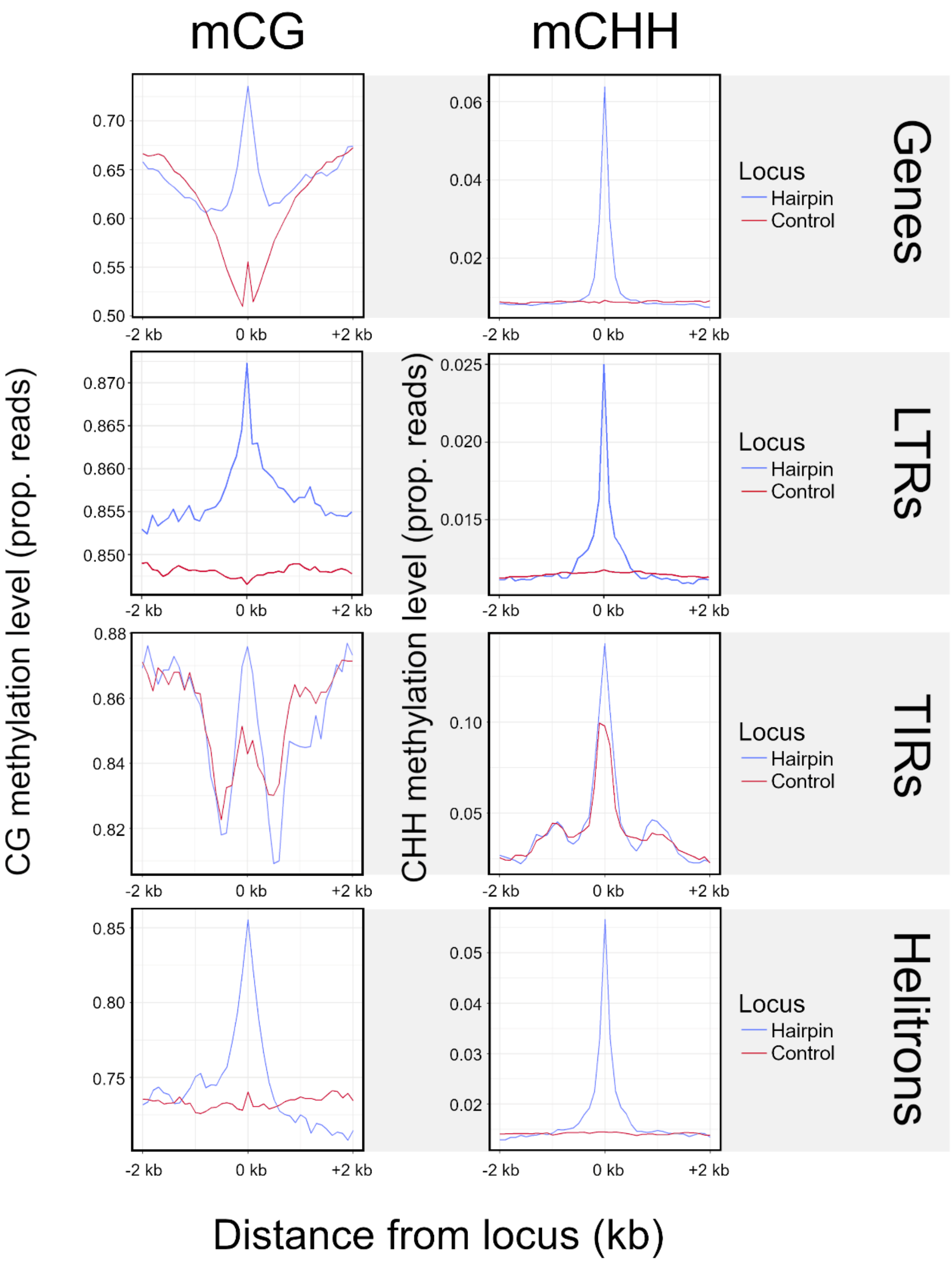


Figure 2.4. Methylation at LP-hairpins. The left column shows methylation in the CG context (mCG) and the right shows methylation in the CHH context (mCHH). Each row represents a different feature type. The blue lines summarize the patterns of methylation in the hairpin (variable sizes, median = 25 nt) 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.

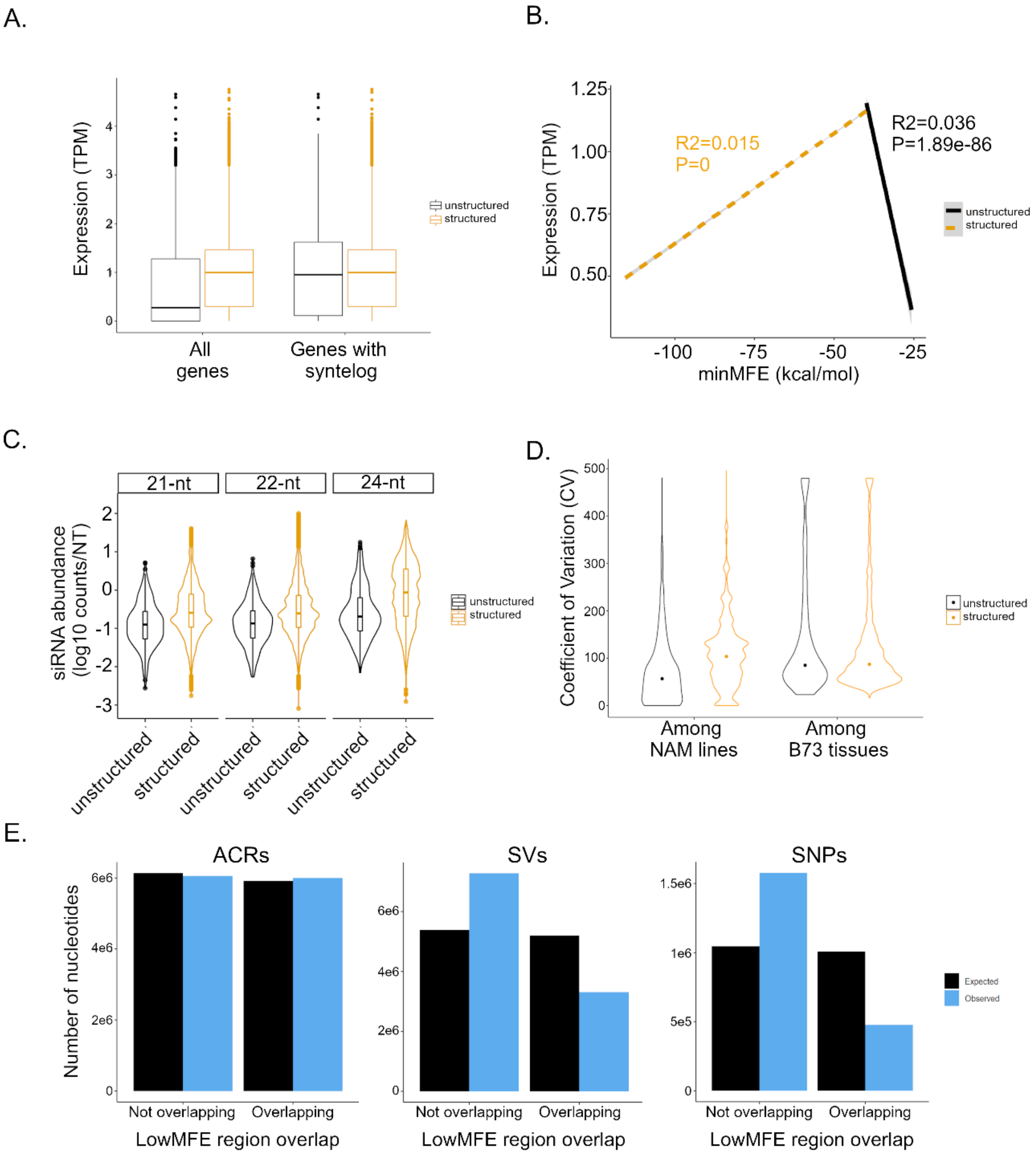


Figure 2.5. Expression between structured and unstructured genes, as defined by RNAfold analysis, in B73. The expression data are based on combined data across 23 tissues. A. Difference in the overall magnitude of expression in all structured (n=27,034) vs unstructured (n=5054) genes and in structured 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 x 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 x 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.

Tables

Table 2.1: 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 2.2: 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	133,81 2	64.80%	5.02%	133,81 2	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]	**373,48 5**	**286,744**	**90,088**	**467,25 5**	**182,7 49**

[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 2.3: 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)

Supplementary Figures

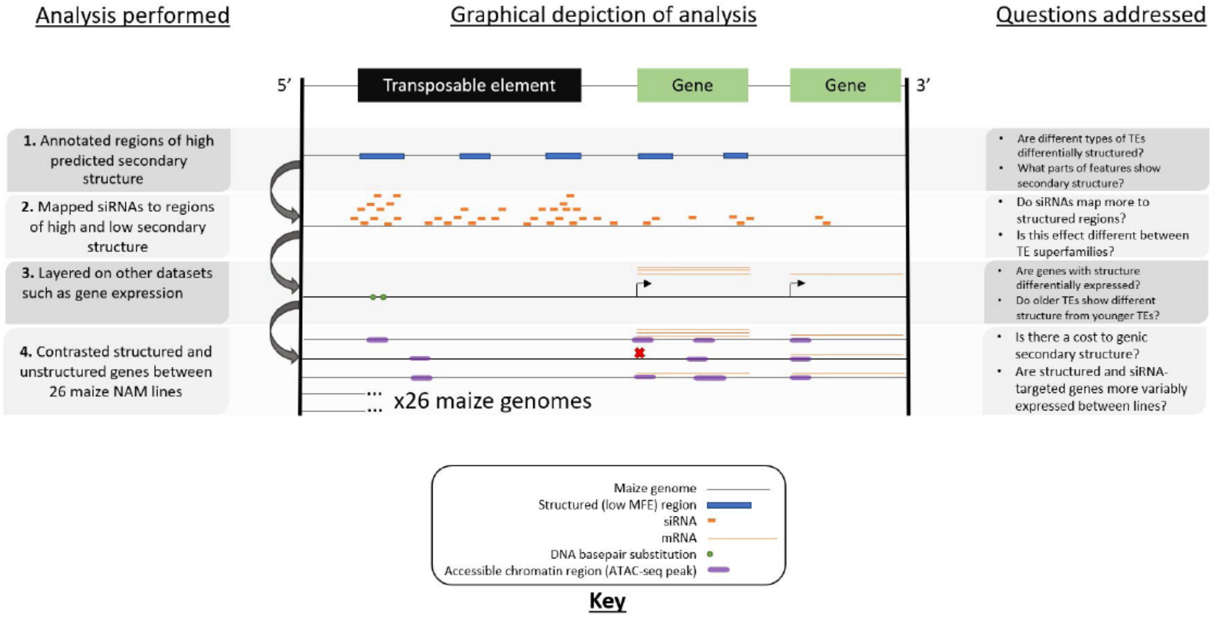

Analysis performed
Graphical depiction of analysis
Questions addressed
5′
Transposable element
Gene
Gene
3′
1. Annotated regions of high predicted secondary structure
2. Mapped siRNAs to regions of high and low secondary structure
3. Layered on other datasets such as gene expression
4. Contrasted structured and unstructured genes between 26 maize NAM lines
x26 maize genomes
Are different types of TEs differentially structured?
What parts of features show secondary structure?
Do siRNAs map more to structured regions?
Is this effect different between TE superfamilies?
Are genes with structure differentially expressed?
Do older TEs show different structure from younger TEs?
Is there a cost to genic secondary structure?
Are structured and siRNA-targeted genes more variably expressed between lines?
Maize genome
Structured (low MFE) region
siRNA
mRNA
DNA basepair substitution
Accessible chromatin region (ATAC-seq peak)
Key

Figure S2.1: Scheme of analyses carried out. Each sequential layer includes an additional analysis performed on annotated features, including genes, ncRNA loci, and TEs. We (1.) annotated regions of predicted miRNA-like secondary structure (Figure 2.1 & 2.2; Table 2.1), then (2.) mapped siRNAs across features of varying secondary structure and between regions with miRNA-like structure and without (Figure 2.4; Table 2.2), then (3.) compared expression (Figure 2.4) , and (4.) examined underlying genetic and epigenetic features of genes with differing secondary structure between 26 inbred NAM lines representing the breadth of global maize diversity (Figure 2.5)

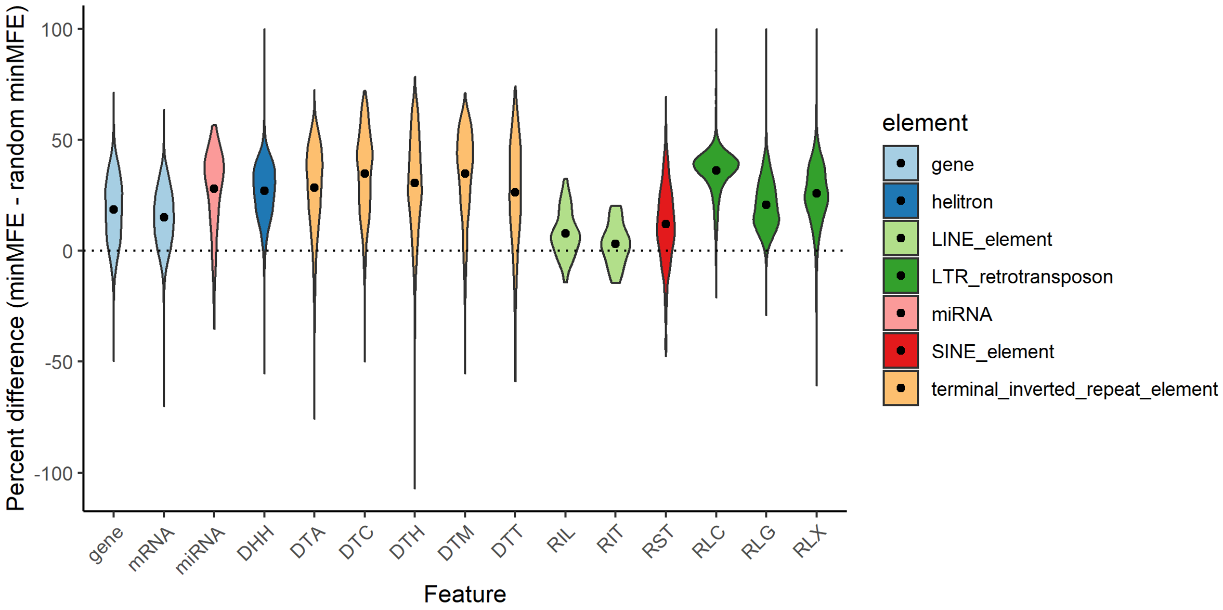


Figure S2.2. Distributions of percent differences between observed and random minMFEs in each feature type. Differences represent how much more negative (and therefore more stabily structured) observed minMFEs were compared to mean minMFE across five randomizations. To find percent differences, these differences were divided by the observed minMFE and multiplied by 100 [e.g., if the observed minMFE was -100 and the mean randomized minMFE was -50, percent difference would be ((-100 + -50) / -100) * 100 = 100%]. Superfamilies are colored by their broader TE category (LTR, TIR, etc.) and dots represent the mean of each distribution. The dotted line represents 0%, or zero difference from random minMFE.

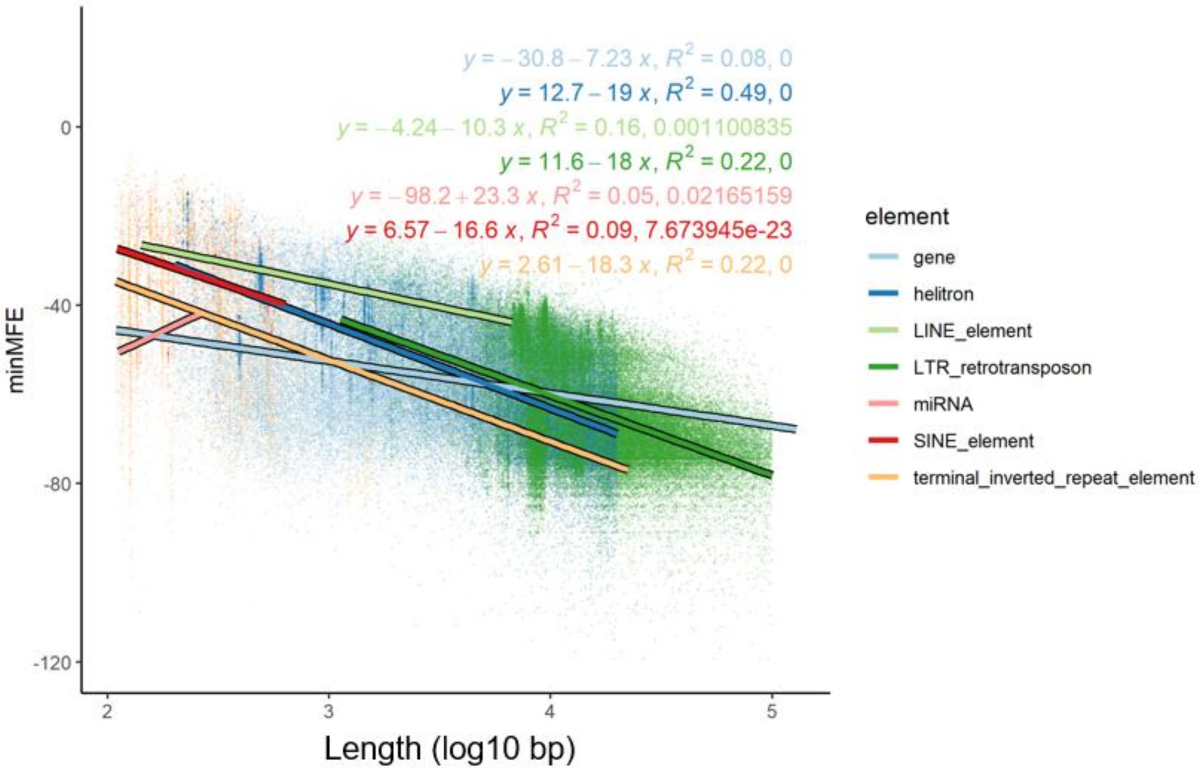


Figure S2.3: Linear models of minMFE as a function of length. Each type of TE was modeled separately using its length from 5’ to 3’ end and observed minMFE value. Plots represent simple linear models from the lm() function in R, and colored text represents the formulae, R2 value, and p-value of each regression.

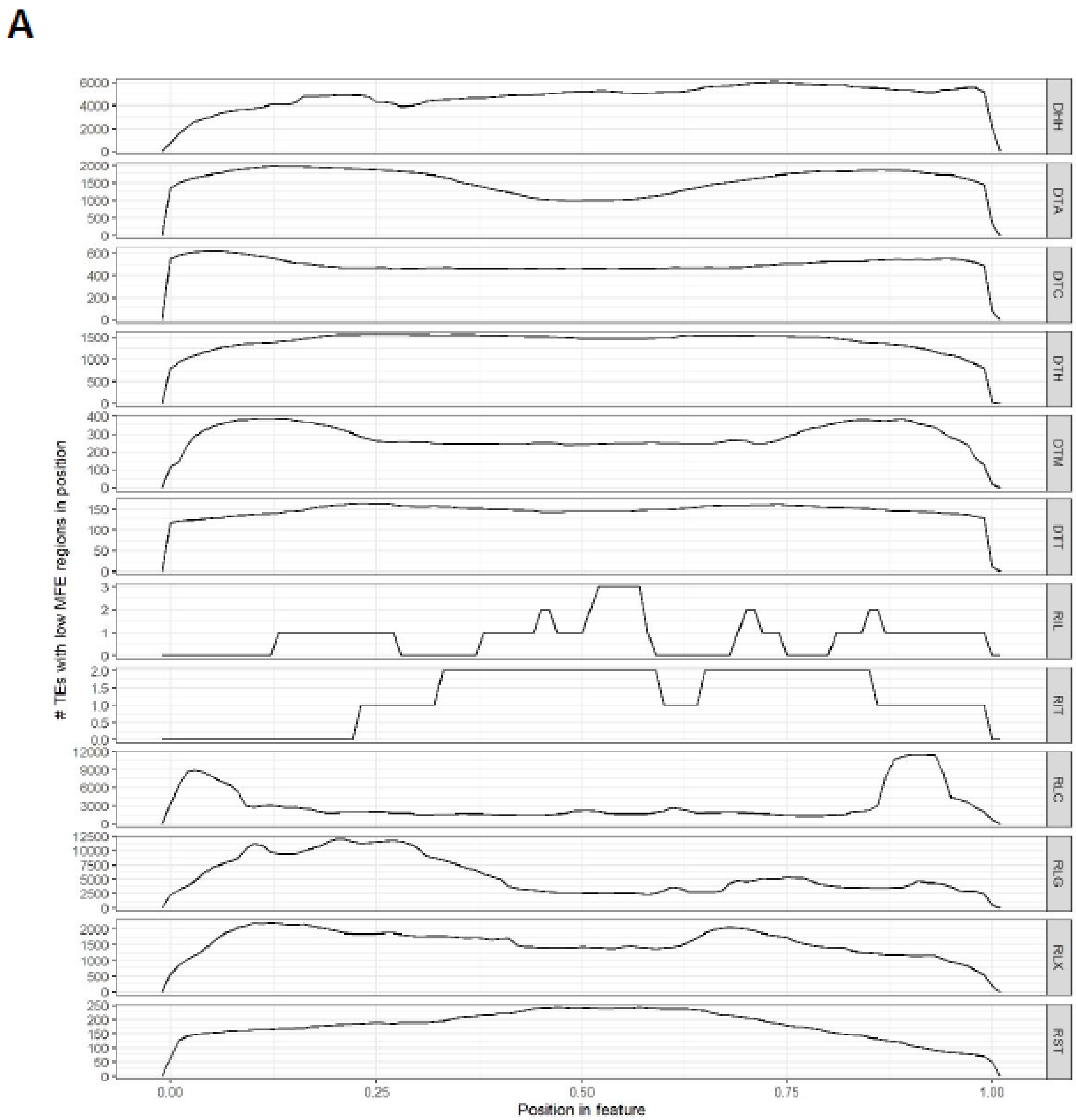


Figure S2.4. Landscapes of miRNA-like regions across feature types. (A) Metaprofiles of lowMFE regions across TE superfamilies, and (C) metaprofiles of LP-hairpins across TEs. Each row represents a metaprofile combining data from all members of each feature type. Features were divided into 100 equally sized bins from the 5' end to the 3' end, and the number of features with miRNA-like regions overlapping each of these bins was counted. A peak in the landscape therefore represents a region of the feature type which often shows very stable secondary structure. All rows share the same X axis, which is represented proportionally across the length of the feature from 0.00 (5' end) to 1.00 (3' end).

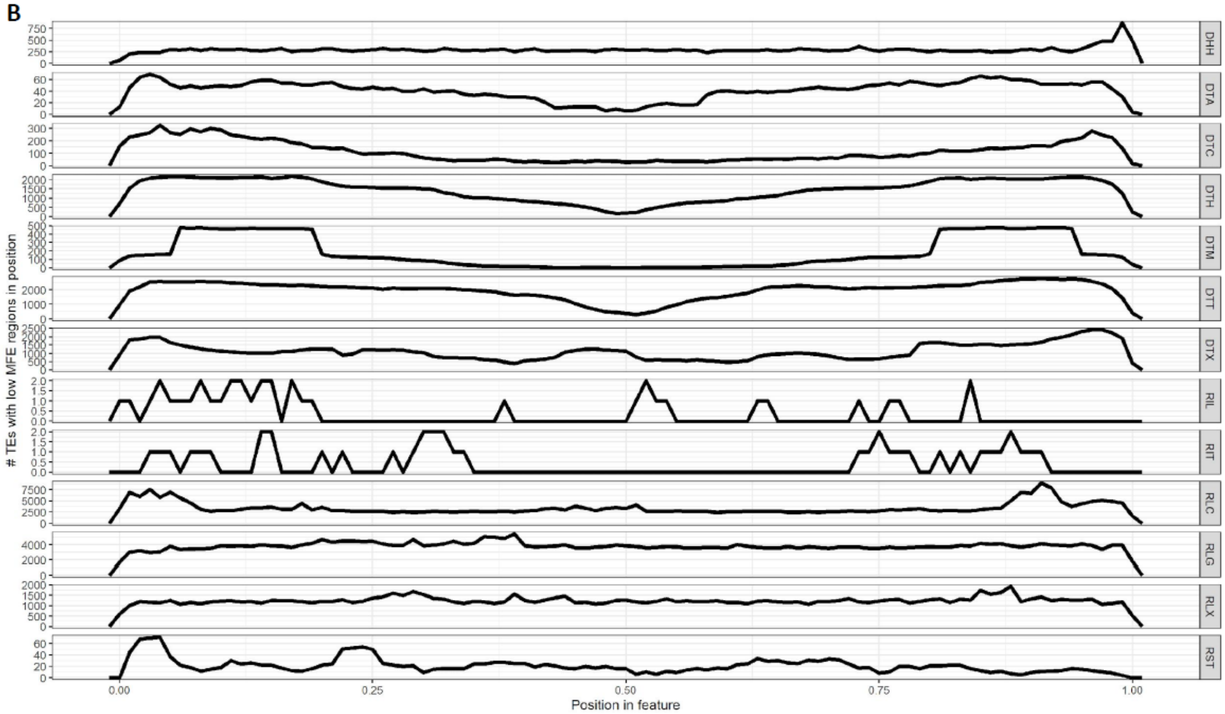


Figure S2.4 (cont). Landscapes of miRNA-like regions across feature types. (A) Metaprofiles of lowMFE regions across TE superfamilies, and (C) metaprofiles of LP-hairpins across TEs. Each row represents a metaprofile combining data from all members of each feature type. Features were divided into 100 equally sized bins from the 5' end to the 3' end, and the number of features with miRNA-like regions overlapping each of these bins was counted. A peak in the landscape therefore represents a region of the feature type which often shows very stable secondary structure. All rows share the same X axis, which is represented proportionally across the length of the feature from 0.00 (5' end) to 1.00 (3' end).

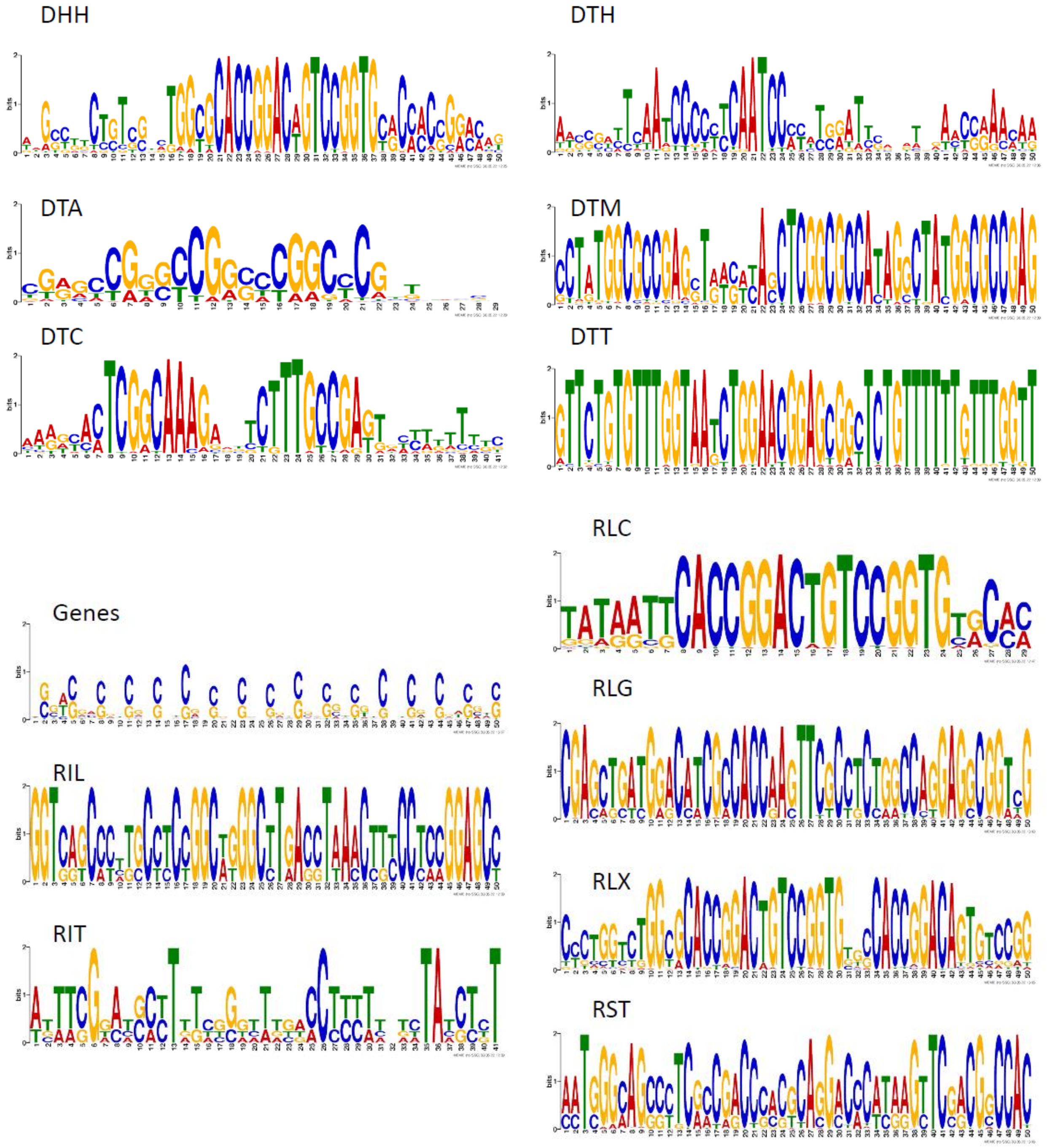


Figure S2.5. Overrepresented motifs in structured/low MFE regions (<-40 kcal/mol) of structured features. Structured regions of each superfamily were entered into MEME motif finder (See Methods), and logos represent the most highly overrespresented motif found in each superfamily.

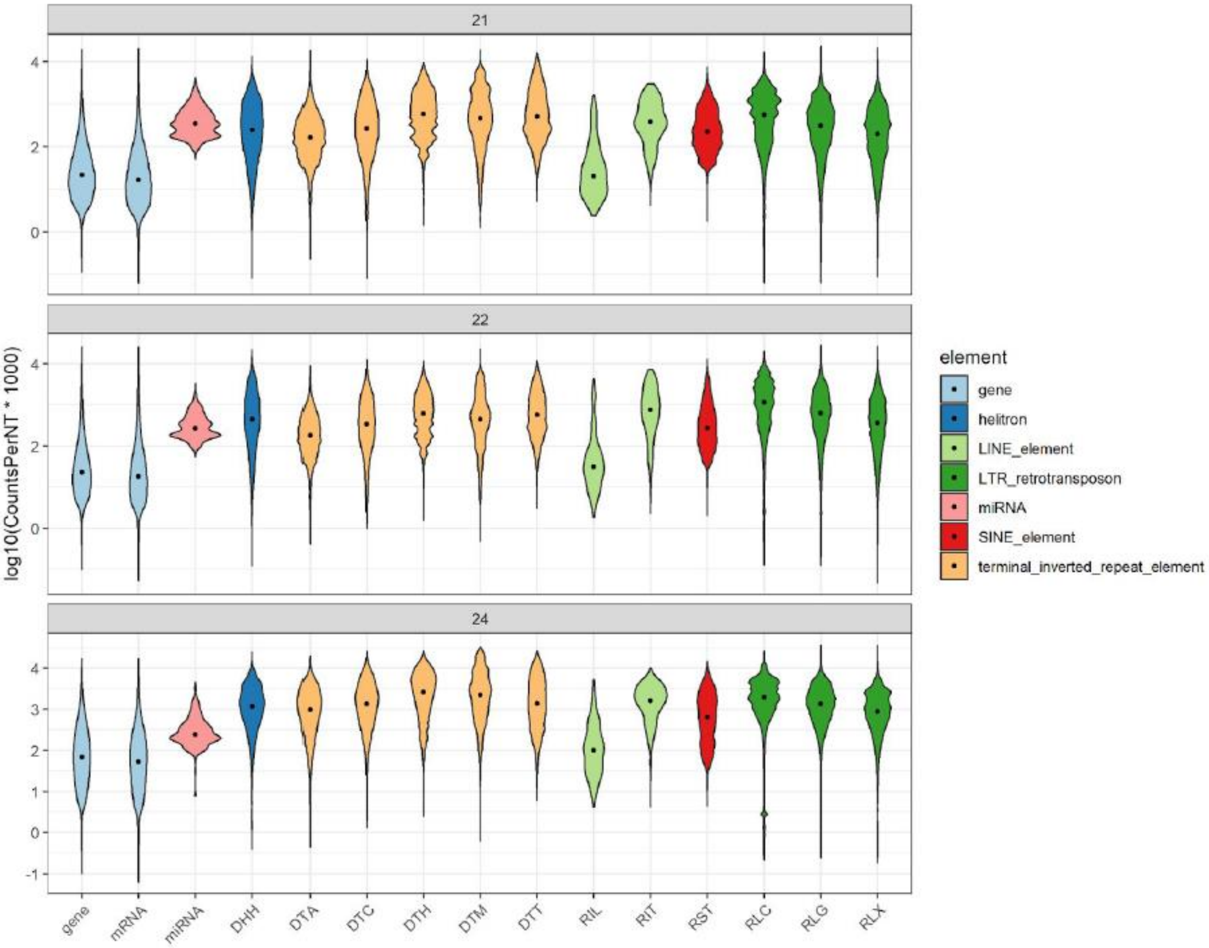


Figure S2.6. Variation in siRNA mapping between feature types. Violin plots show the distributions of siRNA mapping densities in log10(siRNA species counts per kilobase) for each superfamily/genomic feature, and black dots show the mean of the distribution. Panels represent siRNA size classes (21-nt, 22-nt, 24-nt).

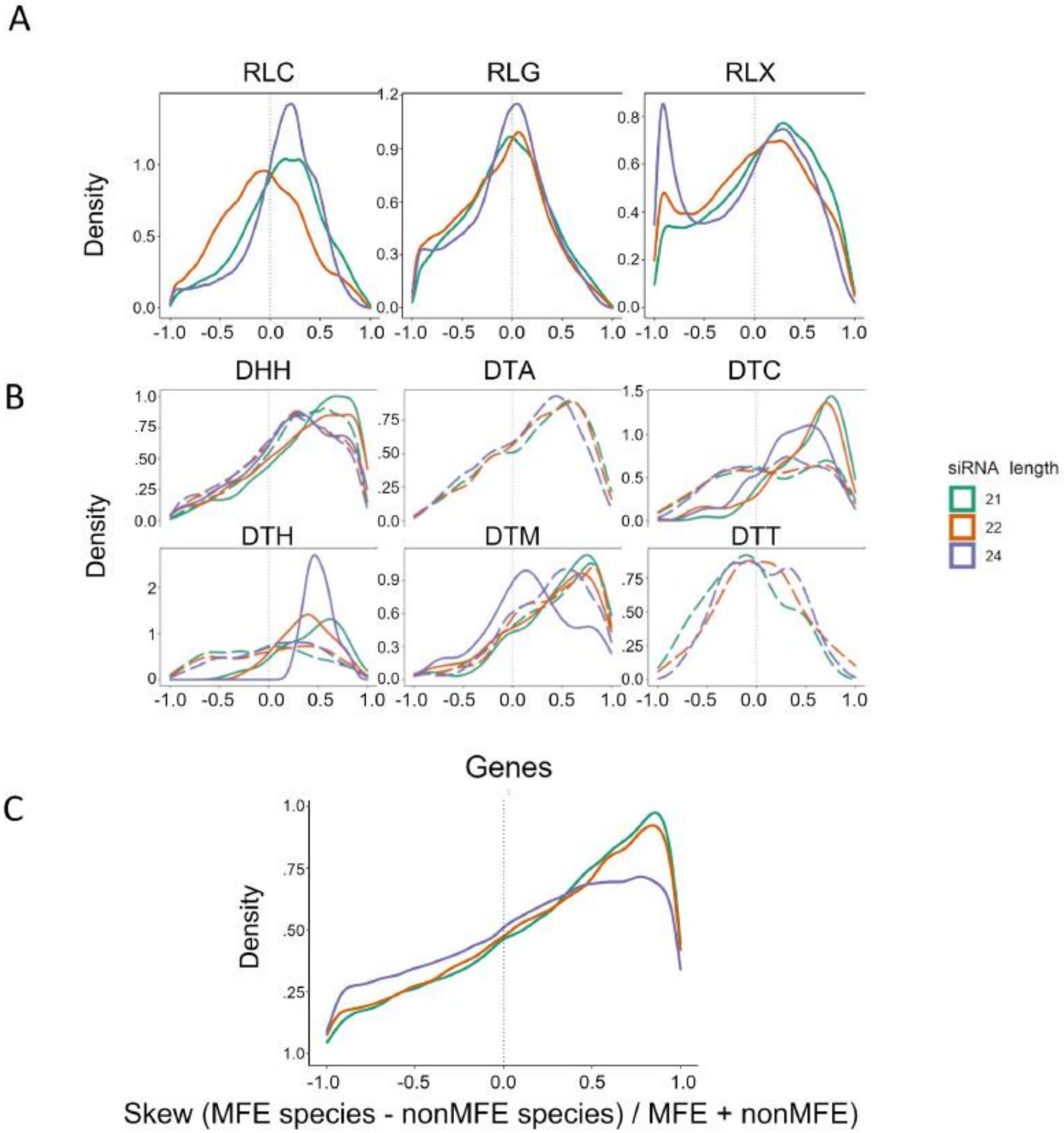


Figure S2.7. siRNA mapping skew towards lowMFE regions. All panels use the same x-axis, which is a measure of skew, and the dotted vertical line represents zero where smRNA density is not skewed to either low or high MFE regions. A. Retrotransposons and their skew for 21, 22 and 24-nt siRNAs, representing Copia (RLC), Ty3 (RLG) and unknown retrotransposons (RLX). B. DNA transposons, with names for the three letter codes provided in Table 2.2. The solid lines represent autonomous elements, while dashed lines represent non-autonomous elements. C. Skew measured in genes.

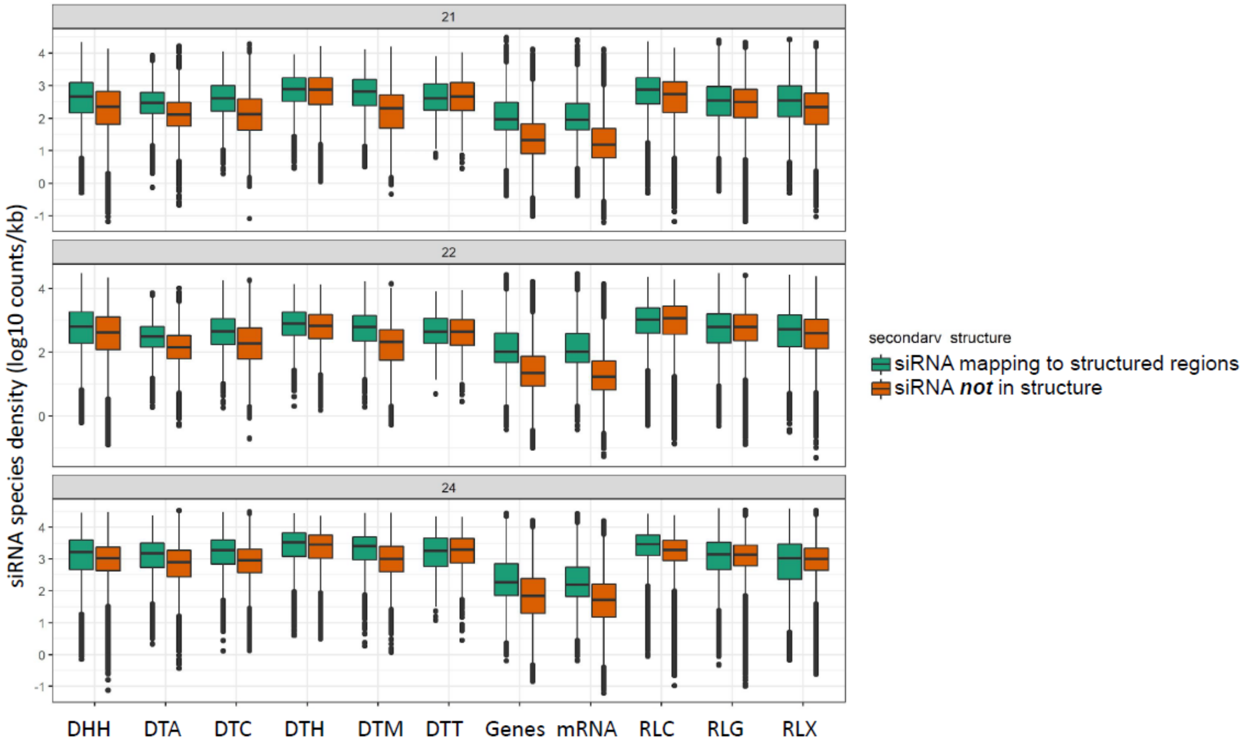


Figure S2.8. siRNA species mapping density in lowMFE regions vs unstructured regions. For each structured feature (minMFE significantly lower than mean randomized minMFE (see Methods), siRNAs mapping to the feature were divided into those mapping to lowMFE regions (<-40 kcal/mol) and those outside of structured regions. Boxplot central lines show the median, and boxes show the 25% and 75% quartiles. Statistical significance in these comparisons can be seen in Table S2.3.

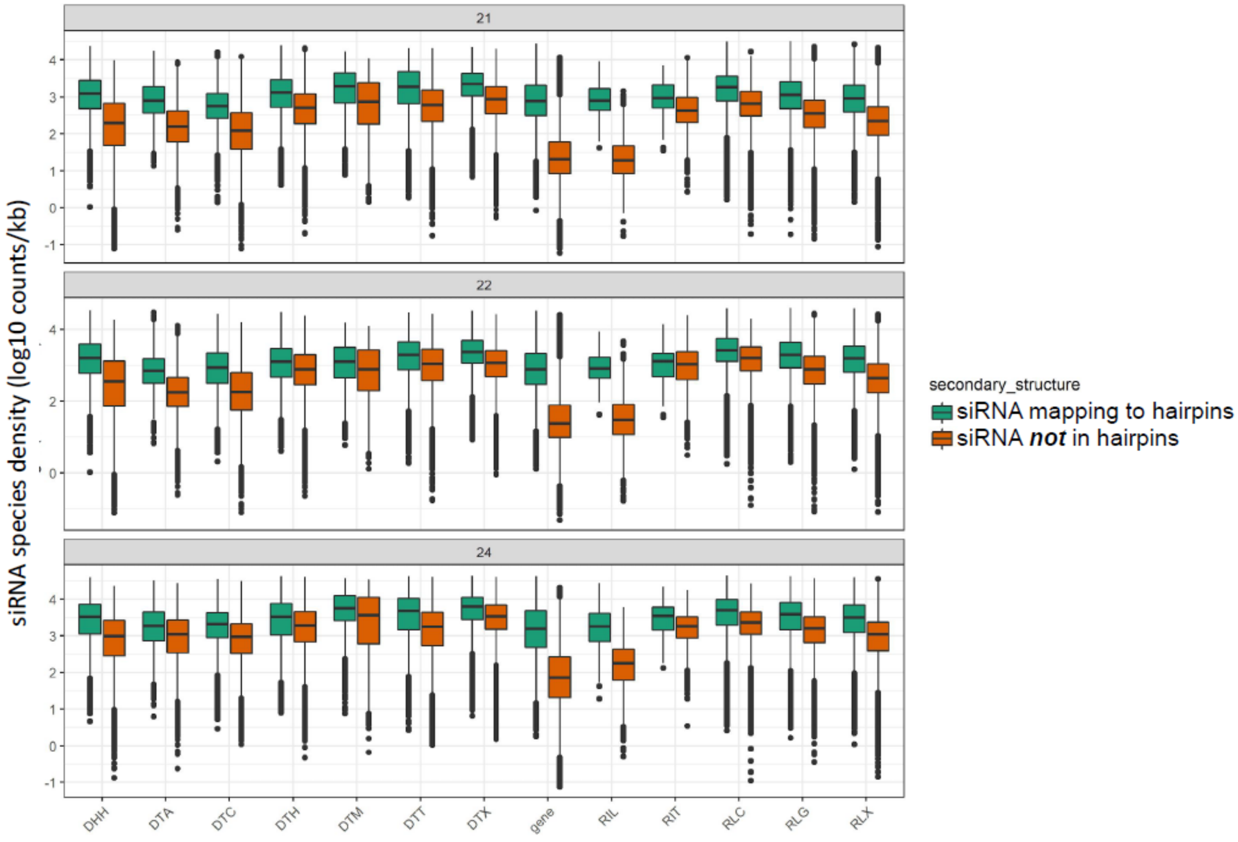


Figure S2.9. siRNA species mapping density in LP-hairpins vs other regions. For each feature, siRNAs mapping to the feature were divided into those mapping to LP-hairpins (<-40 kcal/mol) and those outside of structured regions. Boxplot central lines show the median, and boxes show the 25% and 75% quartiles. Statistical significance in these comparisons can be seen in Table S2.4.

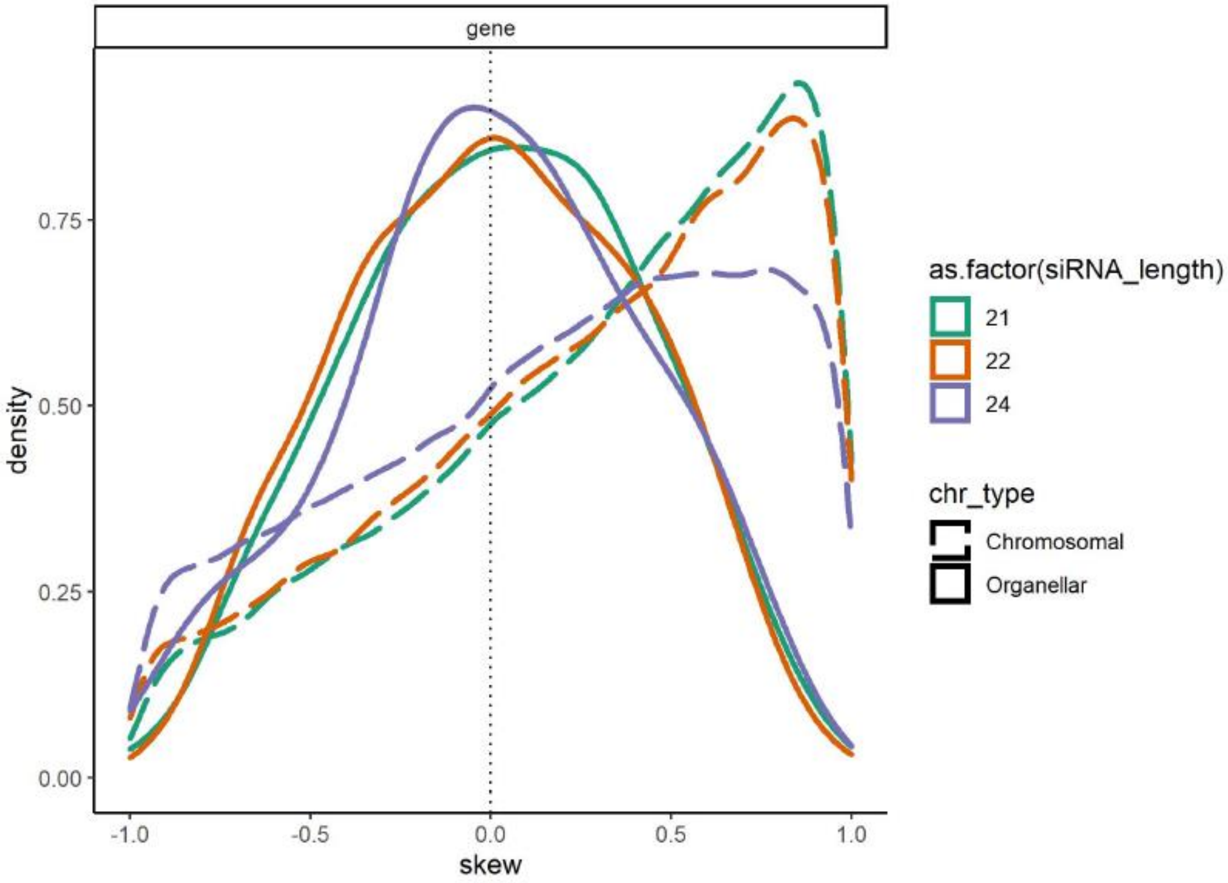


Figure S2.10. siRNA mapping skew towards lowMFE regions in organellar vs nuclear genes. Genes were separated based on position, with mitochrondrial and plastid genes assigned to "organellar" and all other genes assigned to "chromosomal." The dotted line represents chromosomal genes, and the solid line represents organellar genes.

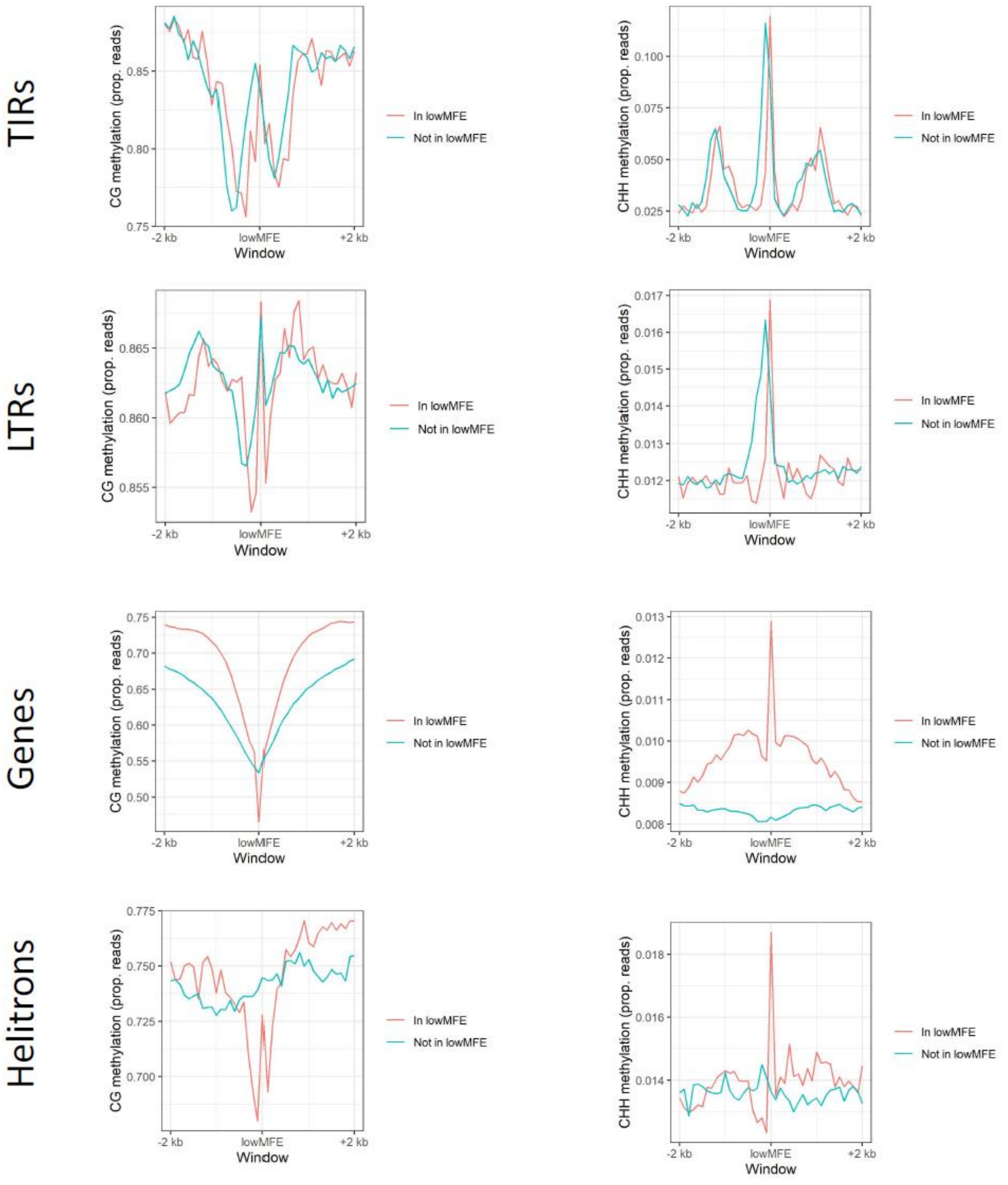


Figure S2.11. Methylation at lowMFE regions. The left column shows methylation in the CG context (mCG) and the right shows methylation in the CHH context (mCHH). Each row represents a different feature type. The red lines summarize the patterns of methylation in the lowMFE regions in a given feature 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 blue line corresponds to methylation patterns around these randomized control loci.

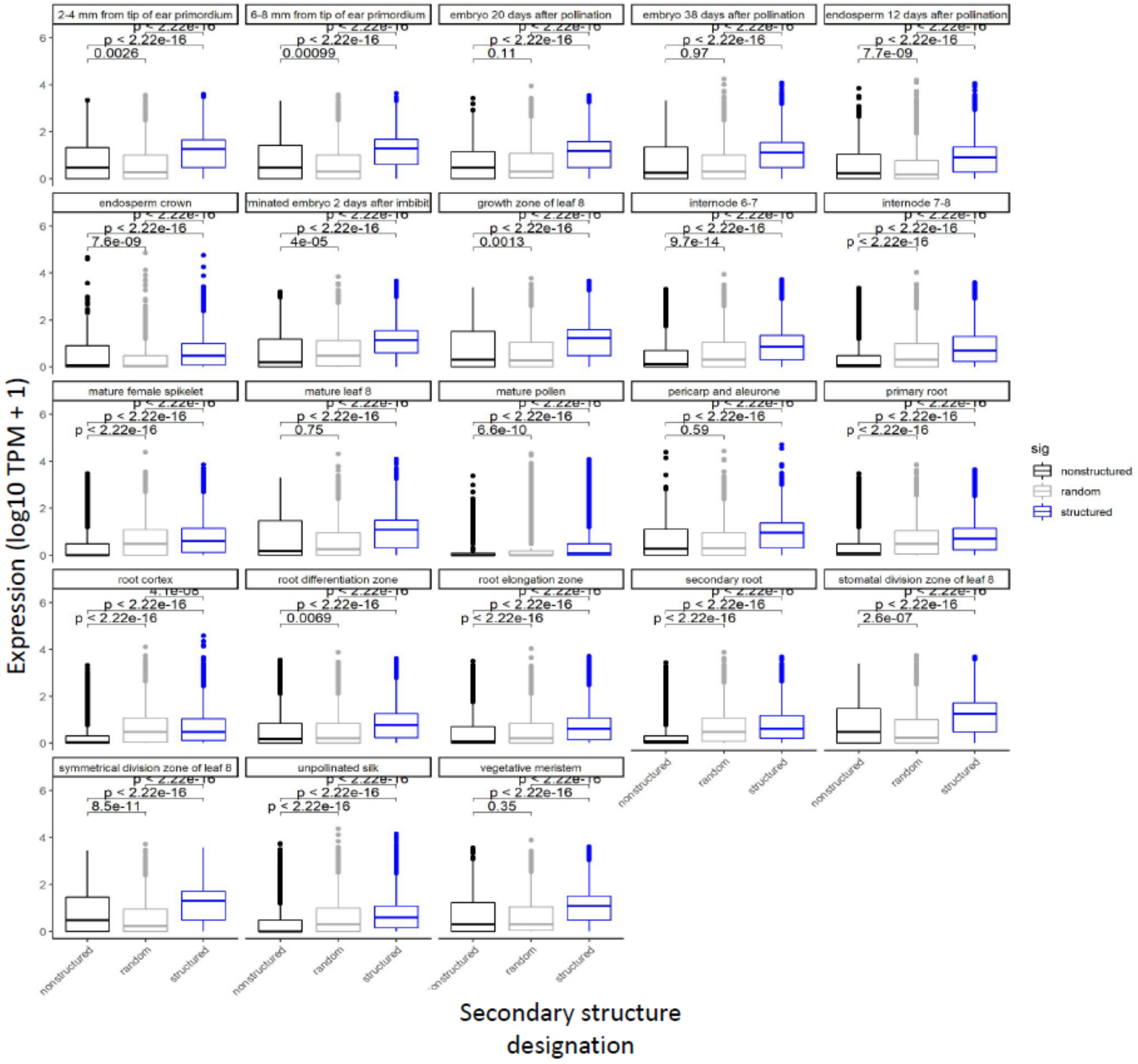


Figure S2.12. Expression between structured, random, and unstructured genes in 23 B73 tissues. “Structured” represent RF-structured genes, “random” represent genes with minMFE < -40 kcal/mol but which are not significantly different from the minMFE of 5 randomizations, and “unstructured” have minMFE > -40 kcal/mol. Expression data are from Walley et al., 2016 and were downloaded from the ATLAS expression database (E-GEOD-50191). Boxplot central lines represent the median, and boxes represent the 25% and 75% quartiles.

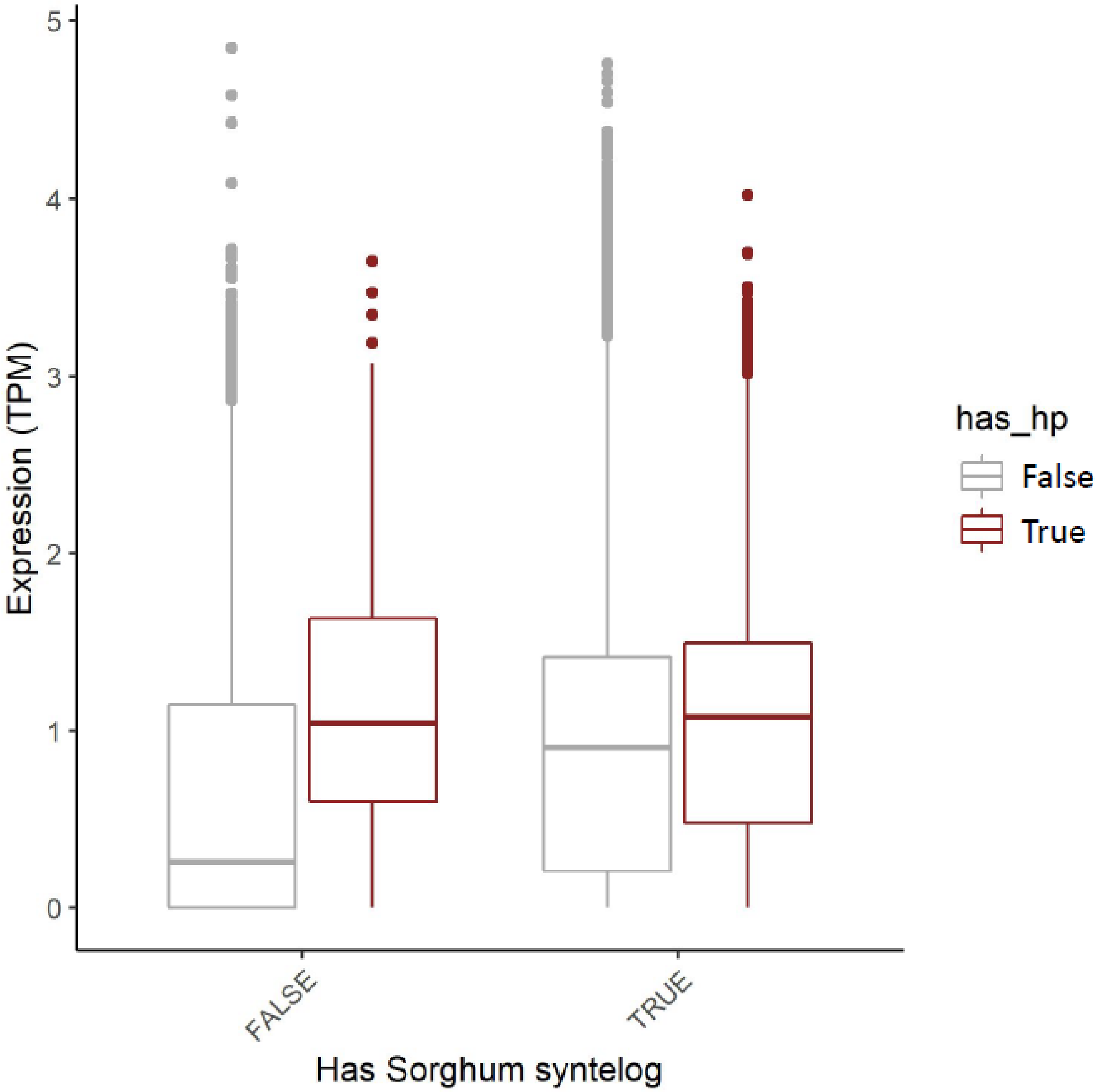


Figure S2.13. Expression between genes with and without LP-hairpins across 23 B73 tissues. Genes are divided into those with and without detectable Sorghum bicolor syntelogs (Muyle et al., 2021). Expression data are from Walley et al., 2016 and were downloaded from the ATLAS expression database (E-GEOD-50191). Boxplot central lines represent the median, and boxes represent the 25% and 75% quartiles.

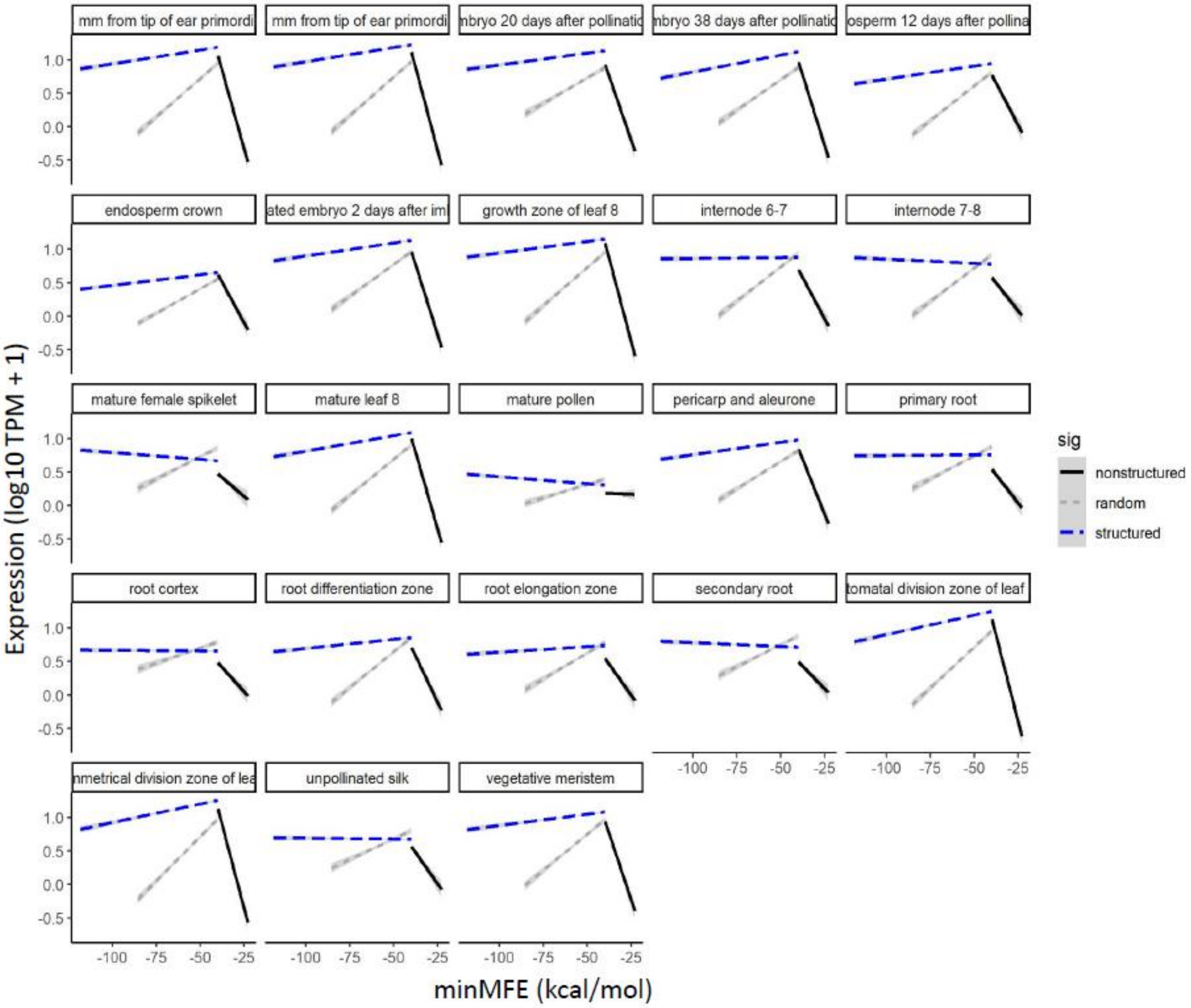


Figure S2.14. Expression as a function of minMFE in 23 B73 tissues. Expression is represented in log10 TPM+1, and structure designations are from the primary sequences in B73 (see Figure S2.13). Expression data are from Walley et al., 2016 and were downloaded from the ATLAS expression database (E-GEOD-50191).

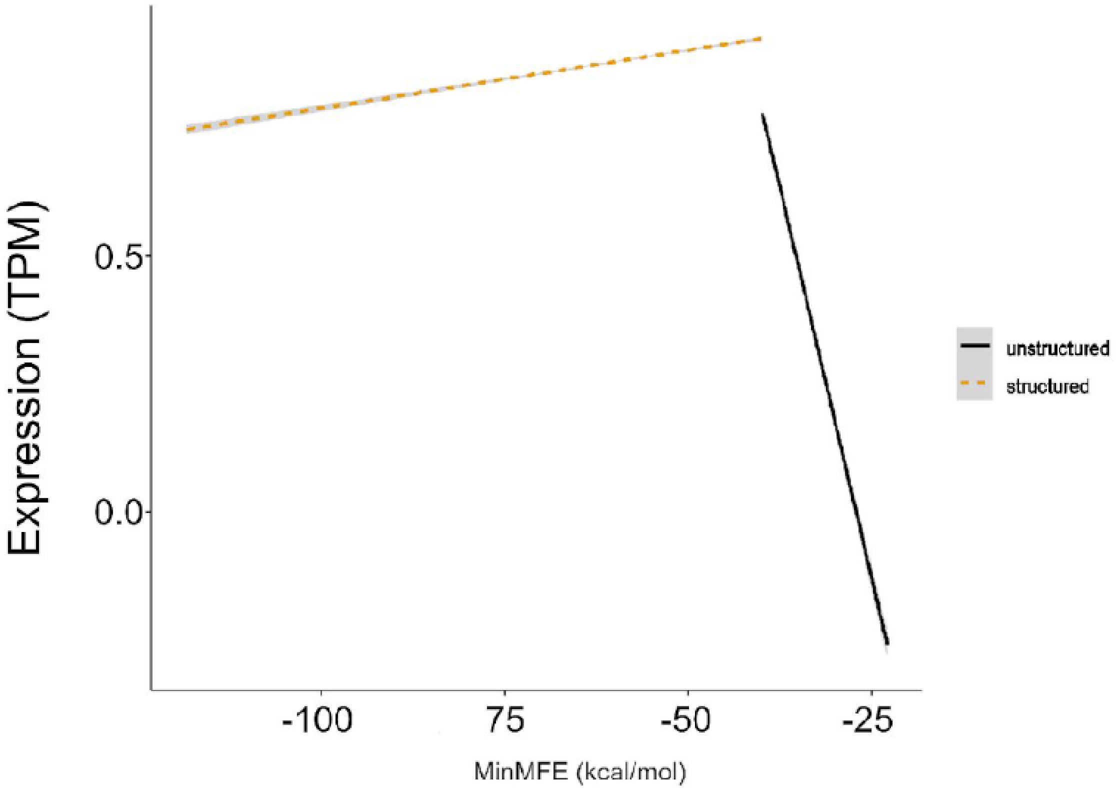


Figure S2.15. Expression as a function of minMFE for RF-structured and unstructured genes. In contrast to Figure 2.5, all genes (with and without Sorghum syntelogs) are included.

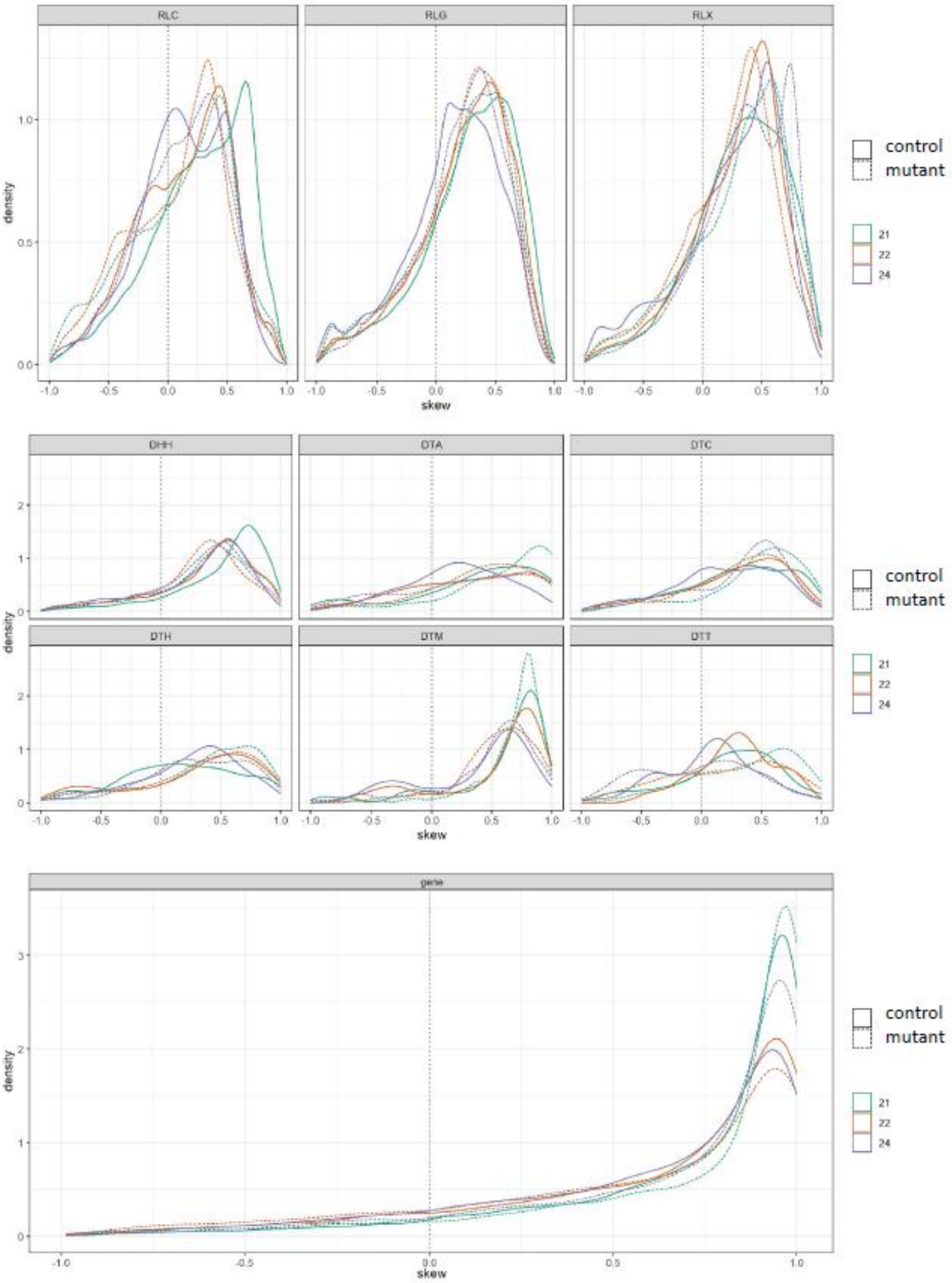


Figure S2.16. siRNA mapping skew for mutant vs wildtype control libraries in LP-hairpins. Solid lines represent control (WT) libraries, while dotted lines represent mop1 libraries. Skew was calculated as in Figure 2.3.

Supplementary Tables

Table S2.1. Small RNA libraries used

Accession #	Reference	Used in RNAfold analysis	Used in LinearPartition analysis	Notes
GSM1342517	Diez et al., 2014	X	X	3rd and 4th leaves
SRR032087	Zhang et al., 2009		X	Ear
SRR032088	Zhang et al., 2009		X	Ear
SRR032089	Zhang et al., 2009		X	Ear
SRR032090	Zhang et al., 2009		X	Ear
SRR032091	Zhang et al., 2009	X	X	Ear
SRR1186264	Diez et al., 2014	X	X	3rd and 4th leaves
SRR1917157	Petsch et al., 2015	X	X	Embryo
SRR1917158	Petsch et al., 2015	X	X	Embryo
SRR2086100	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2086104	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2086116	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2086120	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2086132	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2086136	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2106180	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2106184	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2106196	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR2106200	Lunardon et al., 2016	X	X	Leaf and shoot apical meristem
SRR3684242	Huang et al., 2016	X	X	Endosperm
SRR3684389	Huang et al., 2016	X	X	Endosperm
SRR895785	Liu et al., 2014	X	X	Ear
SRX483603	Diez et al., 2014	X	X	3rd and 4th leaves
SRX708789	Gent et al., 2014		X	mop1 WT control
SRX708787	Gent et al., 2014		X	mop1 mutant
SRX708790	Gent et al., 2014		X	mop1 mutant
SRX708788	Gent et al., 2014		X	mop1 mutant
GSM1178886	Zhai et al., 2013		X	hen1 mutant
GSM913701	Barbour et al., 2012		X	rmr2 mutant

GSM913702	Barbour et al., 2012		X	rmr2 mutant (heterozygote)
GSM448853	Zhang et al., 2009		X	Ear
GSM448854	Zhang et al., 2009		X	Ear
GSM448855	Zhang et al., 2009		X	Ear
GSM448856	Zhang et al., 2009		X	Ear
GSM381738	Wang et al., 2009		X	Seedling roots
GSM381716	Wang et al., 2009		X	Seedling roots
GSM448857	Zhang et al., 2009		X	Tassel
SRX120259	Gent et al., 2013		X	Unfertilized outer ear
GSM306488	Nobuta et al., 2008		X	mop1 mutant
GSM306487	Nobuta et al., 2008		X	mop1 WT control
GSM1178887	Zhai et al., 2013		X	hen1 WT control
GSM433620	Nobuta et al., 2008		X	Leaves
GSM433621	Nobuta et al., 2008		X	Leaves
GSM433622	Nobuta et al., 2008		X	Leaves

Table S2.2. Correlations between metrics of predicted secondary structure and siRNA mapping density within feature types

Feature	Structure measurement	21-nt siRNA	22-nt siRNA	24-nt siRNA

DHH	minMFE	P = 3.070e-78
		R2 = 7.100e-03
			P = 2.180e-117
			R2 = 1.070e-02
				P = 3.950e-103
				R2 = 9.400e-03

	meanMFE	P = 0.000e+00
		R2 = 7.970e-02
			P = 0.000e+00
			R2 = 6.990e-02
				P = 0.000e+00
				R2 = 9.520e-02

	Qnorm	P = 6.250e-101
		R2 = 9.200e-03
			P = 3.050e-198
			R2 = 1.820e-02
				P = 5.080e-68
				R2 = 6.150e-03

DTA	minMFE	P = 4.770e-198
		R2 = 1.490e-01
			P = 8.580e-194
			R2 = 1.460e-01
				P = 3.610e-109
				R2 = 8.420e-02

	meanMFE	P = 1.610e-150
		R2 = 1.150e-01
			P = 4.770e-133
			R2 = 1.020e-01
				P = 2.280e-48
				R2 = 3.740e-02

	Qnorm	P = 5.210e-04
		R2 = 2.020e-03
			P = 8.510e-02
			R2 = 4.980e-04
				P = 3.360e-07
				R2 = 4.360e-03

DTC	minMFE	P = 1.360e-12
		R2 = 3.900e-02
			P = 9.750e-18
			R2 = 5.660e-02
				P = 4.440e-28
				R2 = 9.120e-02

	meanMFE	P = 6.710e-68
		R2 = 2.140e-01
			P = 2.940e-48
			R2 = 1.550e-01
				P = 1.920e-41
				R2 = 1.340e-01

	Qnorm	P = 1.100e-08
		R2 = 2.320e-02
			P = 2.900e-02
			R2 = 3.420e-03
				P = 2.800e-02
				R2 = 3.460e-03

DTH	minMFE	P = 2.990e-76
		R2 = 6.640e-02
			P = 2.740e-48
			R2 = 4.200e-02
				P = 1.460e-32
				R2 = 2.800e-02

	meanMFE	P = 1.800e-69
		R2 = 6.060e-02
			P = 5.940e-33
			R2 = 2.840e-02
				P = 2.800e-13
				R2 = 1.070e-02

	Qnorm	**P = 5.610e-01** R2 = 6.430e-05	**P = 6.270e-01** R2 = 4.500e-05	**P = 1.850e-01** R2 = 3.340e-04

DTM	minMFE	P = 6.100e-04
		R2 = 8.880e-03
			P = 9.640e-01
			R2 = 1.550e-06
				P = 4.210e-04
				R2 = 9.400e-03

	meanMFE	P = 1.220e-68
		R2 = 2.080e-01
			P = 1.830e-76
			R2 = 2.290e-01
				P = 2.080e-48
				R2 = 1.500e-01

	Qnorm	P = 5.380e-65
		R2 = 1.910e-01
			P = 3.490e-38
			R2 = 1.150e-01
				P = 4.980e-39
				R2 = 1.170e-01

DTT	minMFE	P = 1.480e-12
		R2 = 1.040e-01
			P = 3.250e-12
			R2 = 1.010e-01
				P = 9.380e-21
				R2 = 1.740e-01

	meanMFE	P = 2.260e-05
		R2 = 3.870e-02
			P = 3.680e-04
			R2 = 2.750e-02
				P = 1.750e-09
				R2 = 7.650e-02

	Qnorm	P = 2.470e-14
		R2 = 6.620e-02
			P = 1.430e-13
			R2 = 6.240e-02
				P = 1.160e-25
				R2 = 1.210e-01

gene	minMFE	P = 5.560e-129
		R2 = 1.480e-02
			P = 4.120e-106
			R2 = 1.210e-02
				P = 7.050e-02
				R2 = 8.350e-05

	meanMFE	P = 2.850e-124
		R2 = 1.420e-02
			P = 2.520e-158
			R2 = 1.820e-02
				P = 0.000e+00
				R2 = 6.820e-02

	Qnorm	P = 0.000e+00
		R2 = 1.910e-01
			P = 0.000e+00
			R2 = 2.400e-01
				P = 0.000e+00
				R2 = 1.440e-01

miRNA	minMFE	***P = 2.290e-01*** *R2 = 1.380e-02*	***P = 9.360e-02*** *R2 = 2.650e-02*	***P = 4.180e-01*** *R2 = 6.270e-03*
	meanMFE	***P = 3.190e-01*** *R2 = 9.470e-03*	***P = 8.550e-02*** *R2 = 2.790e-02*	***P = 1.360e-01*** *R2 = 2.100e-02*

	Qnorm	***P = 2.050e-01***
		R2 = 1.050e-02
			P = 4.430e-02
			R2 = 2.640e-02
				P = 6.540e-01
				R2 = 1.330e-03

RIL	minMFE	P = 4.860e-02
		R2 = 1.100e-01
			P = 6.800e-02
			R2 = 9.460e-02
				P = 1.240e-03
				R2 = 2.670e-01

	meanMFE	P = 2.780e-02
		R2 = 1.340e-01
			P = 2.960e-02
			R2 = 1.320e-01
				P = 3.180e-04
				R2 = 3.210e-01

	Qnorm	P = 3.390e-02
		R2 = 1.260e-01
			P = 7.750e-02
			R2 = 8.880e-02
				P = 4.470e-03
				R2 = 2.140e-01

RIT	minMFE	***P = 3.090e-01*** *R2 = 3.820e-02*	***P = 2.380e-01*** *R2 = 5.120e-02*	***P = 4.010e-01*** *R2 = 2.620e-02*
	meanMFE	***P = 1.260e-01*** *R2 = 8.460e-02*	***P = 8.900e-02*** *R2 = 1.030e-01*	***P = 2.660e-01*** *R2 = 4.560e-02*
	Qnorm	***P = 3.200e-01*** *R2 = 3.670e-02*	***P = 5.350e-01*** *R2 = 1.440e-02*	***P = 1.830e-01*** *R2 = 6.480e-02*

RLC	minMFE	P = 0.000e+00
		R2 = 6.980e-02
			P = 0.000e+00
			R2 = 6.990e-02
				P = 0.000e+00
				R2 = 1.140e-01

	meanMFE	P = 0.000e+00
		R2 = 7.130e-02
			P = 0.000e+00
			R2 = 6.270e-02
				P = 0.000e+00
				R2 = 9.500e-02

	Qnorm	P = 3.880e-06
		R2 = 1.990e-03
			P = 1.230e-04
			R2 = 1.380e-03
				P = 4.250e-10
				R2 = 3.640e-03

RLG	minMFE	P = 0.000e+00
		R2 = 7.040e-02
			P = 0.000e+00
			R2 = 5.370e-02
				P = 0.000e+00
				R2 = 1.550e-01

	meanMFE	P = 0.000e+00
		R2 = 1.600e-01
			P = 0.000e+00
			R2 = 1.080e-01
				P = 0.000e+00
				R2 = 2.410e-01

	Qnorm	P = 1.370e-173 R2 = 4.860e-02	P = 2.020e-156
			R2 = 4.380e-02
				P = 1.610e-209
				R2 = 5.850e-02

RLX	minMFE	P = 0.000e+00
		R2 = 1.100e-01
			P = 0.000e+00
			R2 = 1.280e-01
				P = 0.000e+00
				R2 = 1.480e-01

	meanMFE	P = 1.050e-275
		R2 = 7.480e-02
			P = 1.690e-254
			R2 = 6.920e-02
				P = 0.000e+00
				R2 = 9.460e-02

	Qnorm	P = 1.090e-70
		R2 = 8.720e-02
			P = 1.930e-90
			R2 = 1.110e-01
				P = 8.230e-71
				R2 = 8.740e-02

RST	minMFE	P = 2.240e-41
		R2 = 1.750e-01
			P = 4.840e-30
			R2 = 1.280e-01
				P = 2.780e-46
				R2 = 1.940e-01

	meanMFE	P = 7.040e-45
		R2 = 1.890e-01
			P = 1.280e-30
			R2 = 1.310e-01
				P = 8.970e-51
				R2 = 2.120e-01

	Qnorm	P = 4.870e-50
		R2 = 1.930e-01
			P = 7.780e-48
			R2 = 1.850e-01
				P = 2.080e-47
				R2 = 1.840e-01

All features	minMFE	P = 0.000e+00
		R2 = 9.060e-02
			P = 0.000e+00
			R2 = 1.030e-01
				P = 0.000e+00
				R2 = 7.380e-02

	meanMFE	P = 0.000e+00
		R2 = 1.660e-02
			P = 0.000e+00
			R2 = 8.610e-03
				P = 5.010e-227
				R2 = 4.310e-03

	Qnorm	P = 0.000e+00
		R2 = 1.010e-01
			P = 0.000e+00
			R2 = 1.330e-01
				P = 0.000e+00
				R2 = 8.930e-02

Table S2.3. Statistics from mixed-effect models comparing siRNA species mapping density between lowMFE and other regions (i.e., regions less than and greater than -40 kcal/mol) of RF-structured features (i.e., features with significantly lower minMFE than five randomizations; see **Methods**)

Feature	**21-nt siRNA**	**22-nt siRNA**	**24-nt siRNA**
RLC	Estimate = -2.400e-01 | P = 0.000e+00 | std. err. = 1.860e-03	Estimate = 1.860e-01 | P = 0.000e+00 | std. err. = 1.750e-03	Estimate = -2.190e-01 | P = 0.000e+00 | std. err. = 1.320e-03
RLG	Estimate = 1.580e-01 | P = 0.000e+00 | std. err. = 1.650e-03	Estimate = 3.000e-01 | P = 0.000e+00 | std. err. = 1.520e-03	Estimate = 2.070e-01 | P = 0.000e+00 | std. err. = 1.230e-03
RLX	Estimate = -1.780e-01 | P = 0.000e+00 | std. err. = 4.140e-03	Estimate = 5.720e-02 | P = 5.350e-47 | std. err. = 3.970e-03	Estimate = 2.670e-01 | P = 0.000e+00 | std. err. = 3.140e-03
DHH	Estimate = -2.830e-01 | P = 0.000e+00 | std. err. = 2.280e-03	Estimate = 1.230e-02 | P = 1.930e-08 | std. err. = 2.190e-03	Estimate = -7.890e-02 | P = 0.000e+00 | std. err. = 1.680e-03
DTC	Estimate = -5.640e-01 | P = 9.980e-279 | std. err. = 1.550e-02	Estimate = -3.370e-01 | P = 7.810e-101 | std. err. = 1.570e-02	Estimate = -3.630e-01 | P = 8.310e-208 | std. err. = 1.170e-02
DTA	Estimate = -6.380e-01 | P = 0.000e+00 | std. err. = 8.910e-03	Estimate = -5.760e-01 | P = 0.000e+00 | std. err. = 8.500e-03	Estimate = -4.450e-01 | P = 0.000e+00 | std. err. = 6.610e-03
DTH	Estimate = -1.270e-01 | P = 6.240e-28 | std. err. = 1.160e-02	Estimate = -2.100e-01 | P = 1.610e-75 | std. err. = 1.140e-02	Estimate = -2.630e-01 | P = 5.490e-209 | std. err. = 8.470e-03
DTM	Estimate = -9.740e-01 | P = 0.000e+00 | std. err. = 1.630e-02	Estimate = -8.830e-01 | P = 0.000e+00 | std. err. = 1.520e-02	Estimate = -7.690e-01 | P = 0.000e+00 | std. err. = 1.230e-02
DTT	Estimate = -3.180e-01 | P = 7.850e-15 | std. err. = 4.060e-02	Estimate = -4.600e-01 | P = 6.030e-31 | std. err. = 3.920e-02	Estimate = -4.430e-01 | P = 8.540e-50 | std. err. = 2.940e-02
gene	Estimate = -8.780e-01 | P = 0.000e+00 | std. err.	Estimate = -7.910e-01 | P = 0.000e+00 |	Estimate = -5.400e-01 | P = 0.000e+00 | std. err. = 4.340e-03

	= 5.980e-03	std. err. = 5.800e-03	
mRNA	Estimate = -1.080e+00 | P = 0.000e+00 | std. err. = 3.830e-03	Estimate = -1.020e+00 | P = 0.000e+00 | std. err. = 3.720e-03	Estimate = -6.840e-01 | P = 0.000e+00 | std. err. = 2.870e-03

Table S2.4. Statistics from mixed-effect models comparing siRNA species mapping density between LP-hairpins and non-hairpin regions.

Feature	21-nt siRNA	22-nt siRNA	24-nt siRNA
gene	Estimate = 1.820e-01 | P = 0.000e+00 | std. err. = 1.200e-03	Estimate = 2.960e-01 | P = 0.000e+00 | std. err. = 1.220e-03	Estimate = 5.420e-01 | P = 0.000e+00 | std. err. = 1.620e-03
mRNA	Estimate = 2.560e+00 | P = 0.000e+00 | std. err. = 1.670e-02	Estimate = 1.980e+00 | P = 0.000e+00 | std. err. = 1.780e-02	Estimate = 1.760e+00 | P = 0.000e+00 | std. err. = 1.580e-02
DHH	Estimate = 1.070e+00 | P = 0.000e+00 | std. err. = 1.840e-03	Estimate = 1.230e+00 | P = 0.000e+00 | std. err. = 1.970e-03	Estimate = 1.520e+00 | P = 0.000e+00 | std. err. = 2.230e-03
DTX	Estimate = 3.650e+00 | P = 0.000e+00 | std. err. = 2.590e-03	Estimate = 3.920e+00 | P = 0.000e+00 | std. err. = 2.650e-03	Estimate = 4.660e+00 | P = 0.000e+00 | std. err. = 2.910e-03
DTT	Estimate = 2.470e+00 | P = 0.000e+00 | std. err. = 1.560e-03	Estimate = 2.830e+00 | P = 0.000e+00 | std. err. = 1.670e-03	Estimate = 3.350e+00 | P = 0.000e+00 | std. err. = 1.790e-03
DTH	Estimate = 2.940e+00 | P = 0.000e+00 | std. err. = 1.670e-03	Estimate = 3.340e+00 | P = 0.000e+00 | std. err. = 1.740e-03	Estimate = 4.220e+00 | P = 0.000e+00 | std. err. = 1.910e-03
DTA	Estimate = 2.510e+00 | P = 0.000e+00 | std. err. = 4.970e-03	Estimate = 2.730e+00 | P = 0.000e+00 | std. err. = 5.020e-03	Estimate = 4.400e+00 | P = 0.000e+00 | std. err. = 6.160e-03
DTM	Estimate = 1.150e+00 | P = 0.000e+00 | std. err. = 1.840e-02	Estimate = 1.560e+00 | P = 0.000e+00 | std. err. = 1.790e-02	Estimate = 2.060e+00 | P = 0.000e+00 | std. err. = 2.020e-02
DTC	Estimate = 1.640e+00 | P = 0.000e+00 | std. err. = 8.270e-03	Estimate = 1.860e+00 | P = 0.000e+00 | std. err. = 8.750e-03	Estimate = 2.640e+00 | P = 0.000e+00 | std. err. = 1.010e-02
RLC	Estimate = 6.910e-01 | P = 0.000e+00 | std. err. = 2.840e-03	Estimate = 8.010e-01 | P = 0.000e+00 | std. err. = 3.110e-03	Estimate = 7.980e-01 | P = 0.000e+00 | std. err. = 3.290e-03

RLG	Estimate = 1.310e+00 | P = 0.000e+00 | std. err. = 1.940e-03	Estimate = 1.410e+00 | P = 0.000e+00 | std. err. = 2.150e-03	Estimate = 1.510e+00 | P = 0.000e+00 | std. err. = 2.370e-03
RLX	Estimate = 1.050e+00 | P = 0.000e+00 | std. err. = 3.750e-03	Estimate = 1.170e+00 | P = 0.000e+00 | std. err. = 4.140e-03	Estimate = 1.350e+00 | P = 0.000e+00 | std. err. = 4.680e-03
RIT	Estimate = 2.990e-01 | P = 9.250e-04 | std. err. = 8.970e-02	Estimate = 3.140e-01 | P = 3.170e-04 | std. err. = 8.670e-02	Estimate = 6.040e-01 | P = 1.530e-14 | std. err. = 7.670e-02
RIL	Estimate = 2.550e+00 | P = 5.970e-75 | std. err. = 1.120e-01	Estimate = 2.080e+00 | P = 6.640e-63 | std. err. = 1.050e-01	Estimate = 1.470e+00 | P = 1.010e-54 | std. err. = 8.570e-02

Table S2.5. Overlap with siRNA and vs other small RNA loci from Lunardon et al., 2020.

Small RNA type[1]	Type	Real overlap[2]	Expected overlap[3]	Fold-enrichment[4]	P-value[5]	Corrected P-value	Significant enrichment
21-24 nt	helitrons	6.04%	3.67%	1.64	0.00	0.00	**TRUE**
21-24 nt	TIRs	31.01%	28.87%	1.07	0.00	0.00	**TRUE**
21-24 nt	LTRs	1.01%	0.81%	1.25	0.00	0.00	**TRUE**
21-24 nt	LINE	8.93%	11.65%	0.77	0.71	1.00	FALSE
21-24 nt	SINE	6.45%	6.45%	1.00	1.00	1.00	FALSE
21-24 nt	gene	18.61%	6.90%	2.70	0.00	0.00	**TRUE**
21-24 nt	mRNA	18.61%	7.63%	2.44	0.00	0.00	**TRUE**
Other sizes	helitrons	0.20%	0.25%	0.80	0.94	1.00	FALSE
Other sizes	TIR element	0.96%	0.85%	1.14	0.03	0.28	FALSE
Other sizes	LTR	0.05%	0.06%	0.90	0.94	1.00	FALSE
Other sizes	LINE	0.00%	0.00%	NA	NA	NA	NA
Other sizes	SINE	0.00%	0.00%	NA	NA	NA	NA
Other sizes	gene	0.95%	1.10%	0.86	1.00	1.00	FALSE
Other sizes	mRNA	2.46%	2.10%	1.17	0.00	0.02	**TRUE**

[1] 21-24-nt smRNA loci are putatively produce siRNAs through the DCL pathway. Other sizes (< 21-nt, 23-nt, and >24-nt) putatively represent products of degradation or production not dependent on the DCL pathway.

[2] Percentage of miRNA-like loci that overlapped genomic positions with smRNA-producing loci

[3] Expected overlap produced from the average overlap in 500 permutations of miRNA-like loci position within features.

[4] Real overlap divided by expected overlap

[5] Significance from permutation test

Annotated small RNA loci were downloaded from https://plantsmallrnagenes.science.psu.edu/ on August 30, 2023

Table S2.6. Illumina adapter sequences used to trim each small RNA library with CutAdapt (see **Methods**)

Library	Adapter
GSM306487	CTGTAGG
GSM306488	CTGTAGG
SRR032087	CTGTAGG
SRR032088	CTGTAGG
SRR032089	CTGTAGG
SRR032090	CTGTAGG
SRR032091	CTGTAGG
SRR1186264	ATCTCGT
SRR1917157	AGATCGG
SRR1917158	AGATCGG
SRR2086100	TGGAATT
SRR2086104	TGGAATT
SRR2086116	TGGAATT
SRR2086120	TGGAATT
SRR2086132	TGGAATT
SRR2086136	TGGAATT

SRR2106180	TGCAGCA
SRR2106184	TGCAGCA
SRR2106196	TGCAGCA
SRR2106200	TGCAGCA
SRR3684242	TCGTATG
SRR3684389	TCGTATG
SRR895785	TCGTATG
SRX483603	ATCTCGT

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.

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.

CHAPTER 3

Quantifying evolution in RNA secondary structure among Arabidopsis thaliana genes

3.1 Abstract

RNA secondary structure serves important functional roles in plants, but its evolutionary dynamics are not well-studied. We characterized evolutionary dynamics of putative secondary structure-altering mutations in *Arabidopsis thaliana* using computational predictions and empirical pairing data combined with population genomic data from the 1,001 genomes dataset. We classified mutations as structure conserving (unpaired mutations: upM) or structure interrupting (pair-changing mutations: pcM) based on effects on inferred ancestral secondary structure, which we estimated based on two separate methods, one that was computational, and one that was empirical. pcM mutations showed reduced nucleotide diversity and reduced allele frequencies compared to pcM mutations, and their allele spectra were shifted toward low frequency alleles compared to mutations unlikely to affect ancestral secondary structure. Additionally, pcM mutations showed varying fitness effects according to their position within genes. We used demographic models comparing the allele frequency spectra of pcM vs upM synonymous alleles to estimate that synonymous pcM mutations had average scaled selection coefficients between 20.3% and 30.3% those of nonsynonymous changes. Our results demonstrate weak but significant and pervasive selection on secondary structure.

3.2 Introduction

RNA molecules are single stranded (ssRNA), which gives them the ability to form Watson-Crick bonds, as well as less stable bonds (Varani and McClain 2000), between bases on the same molecule. This intramolecular base pairing, termed secondary structure, largely determines the three-dimensional shape of the molecule. As a result, the capacity for an RNA sequence to form secondary structures affects the function of transcribed regions of plant genomes in many ways (Vandivier et al. 2016). For example, secondary structures influence function by modulating translation (Kozak 1988; Svitkin et al. 2001), mRNA splicing (Buratti and Baralle 2004), ribozyme activity (Steitz and Moore 2003), localization (Bullock et al. 2010), and protein-RNA interactions (Williams and Marzluff 1995). Additionally, they affect the epigenetic fate of genes, including their stability (Li et al., 2012), small interfering RNA (siRNA) complement, and DNA methylation (Martin et al., 2022). The ultimate impact of a transcribed genomic region on phenotype (Duan et al., 2003) and fitness (Innan & Stephen, 2001) is therefore dependent on its capacity to form secondary structures, and in humans, mutations that affect mRNA structure have been implicated in disease (Halvorsen et al. 2010). Yet, the evolutionary dynamics of mutations affecting secondary structures in mRNAs have received little attention in the evolutionary biology literature, with most such studies focusing on non-coding RNAs (Nowick et al., 2019).

One interesting and unexplored aspect of selection on secondary structure is its potential to contribute to adaptation. In protein coding genes, positive selection could, in theory, act on both the amino acid sequence and mRNA secondary structure of genes. This process is usually viewed from at the protein level through measured by d_N/d_S, the ratio of nonsynonymous to synonymous mutations in a coding sequence, which quantitates the surplus of amino acid-changing mutations in a gene, or through linked-selection (e.g., detection of selective sweeps) where the causative mutation is unknown (Booker et al., 2017). In a new environment, it is likely that the fitness optima of secondary structures also change. For example, mRNAs in very cold environments will fold differently from mRNAs in hot environments, meaning that the selection pressures for stronger or weaker base pairing are likely different between the old and new environments. Ferrero-Serrano et al. (2022) recently demonstrated that this was the case with two experimentally-validated structure-changing SNPs–termed “riboSNitches” (Halvorsen et al. 2010). Nevertheless, the extent to which this phenomenon occurs across the entire transcribed portion of the genome, especially concerning structural mutations that alter the stability of structures, has not been explored.

Selection on secondary structure could also have important methodological consequences for our ability to use molecular data to measure selection. This is because the interpretation of dn/ds ratios assume that synonymous mutations are neutral with regard to fitness (Kimura 1968). Since the 1980s, evolutionary biologists have known that this is not entirely correct because codon usage is non-random (Ikemura 1981), and more recent studies have demonstrated strong non-neutral fitness effects from synonymous mutations (Lawrie et al. 2013; Lebeuf-Taylor et al. 2019). With regard to the fitness effects of secondary structure, we know that (*i*) secondary structures within mRNA coding regions are more stable than expected under randomized codon usage (Seffens and Digby 1999), (*ii*) the location of synonymous substitutions is not random with respect to secondary structure stability (Chamary and Hurst 2005), (*iii*) codon usage is constrained towards weaker structure around miRNA-binding sites (Gu et al., 2012), and (*iv*) synonymous variants disrupting computationally-predicted secondary structure exist at reduced frequencies in human populations (Gaither et al. 2021), implying that purifying selection acts at these sites. Tools such as d_N/d_S (the ratio of nonsynonymous to synonymous mutations in a coding sequence), and the McDonald-Kreitman test for direct selection (McDonald and Kreitman 1991), rely on the assumption that synonymous changes at degenerate sites are selectively neutral. These measures have been shown to be sensitive to even weak selection on synonymous substitutions (Rahman et al. 2021). Depending on the strength and prevalence of RNA-level selection, accounting for secondary structure could therefore be a prerequisite for distinguishing neutral synonymous variants from less neutral variants.

Another reason such mutations could be evolutionarily interesting is through the possibility of pleiotropic effects between the RNA and protein "life stages" of gene expression. Nonsynonymous mutations can alter both amino acid sequence and secondary structure stability, potentially leading to a conflict between selection for protein function (protein-level selection) and mRNA stability (RNA-level selection). For instance, a derived missense substitution may enhance the effectiveness of a protein but compromise the fitness of its mRNA due an effect on secondary structure [through improper splicing, translation, or reduced stability and so on (Vandivier et al. 2016)]. The frequency and importance of this conflict depend on the relative strength of selection acting on mutations affecting secondary structure, which remains unknown. If such conflict arises, it may constrain the efficacy of positive selection by reducing the overall realized fitness coefficient of more effective proteins, as in the case of pleiotropic effects (Fraïsse et al., 2019).

Finally, while secondary structures serve important functions, particularly strong secondary structures have unique properties that may have negative effects on mRNA stability. To date, several studies have suggested that stable genic hairpins can cause genes to behave like pre-microRNA (miRNA) transcripts (Li et al., 2012), which form hairpin structures that are targeted by Dicer-like enzymes and are subsequently degraded into small RNAs. Like in miRNAs, these structured genes map large numbers of small interfering RNAs (Li et al., 2012; Martin et al., 2022), likely because their hairpins are bound by Dicer-like enzymes. In turn, regions of miRNA-like secondary structure in genes correspond to high densities of small RNA mapping as well as high levels of small RNA-associated methylation, which often repress gene expression and function (Li et al., 2012). Given that small RNA mapping and repressive methylation levels are typically associated with silenced sequences, such as transposable elements, it is interesting that many genes (between ~30-70% of *Zea mays* genes, depending on how they are defined) contain these regions (Martin et al., 2022). It is possible that evolutionary conflicts arise within these regions between the crucial functions of hairpins and a concomitant decrease in stability. Alternatively, these structures might be unavoidable if they are constrained by amino acid sequence.

In this paper, we examine the evolutionary dynamics of mutations that affect secondary structure within genes in the *Arabidopsis thaliana* 1,001 genomes population dataset. We establish a two-pronged identification method that includes both computational prediction and empirical data to find a subset of variants that likely change RNA structure. We then examine the frequencies of these variants in global Arabidopsis accessions to determine (1.) whether these mutations are under selection, (2.) how the strength of this selection compares to nonsynonymous variants, and (3.) if RNA-level selection can meaningfully conflict with protein-level selection. Finally, we question whether structural mutations contribute to local adaptation in differing environments using climatic data associated geospatially with Arabidopsis accessions by integrating our dataset of structure altering mutations with climate and landscape data.

3.3 Results

Identifying unpaired mutations (upM) and pair changing mutations (pcM) mutations

In many of our analyses, we make comparisons between mutations likely to change structure and those unlikely to change structure. Identifying causative mutations that change the overall conformation of the RNA molecule (termed "riboSNitches") is a complex and unsolved problem (Ferrero-Serrano et al. 2022). We therefore developed a method to identify derived mutations at bases that have a very high likelihood of being ancestrally paired in *Arabidopsis thaliana* (i.e., those that likely contribute to secondary structure) in order to perform population genetic analyses. We refer to such mutations as "**p**air **c**hanging **m**utations (pcM)" and those that are unlikely to change secondary structure as "**u**n**p**aired **m**utations (upM mutations)."

Some other considerations should be addressed: First, defining mutations that change secondary structure is not as straightforward as defining those that change amino acids; while the genetic code is universal and each codon always codes for a specific amino acid, a particular RNA transcript may have multiple possible conformations of secondary structure with varying stabilities (Mathews, 2004). By adopting our approach to define pcM/upM mutations, we are simplifying the true biological complexity of the scenario, shifting it from a quantitative problem to a classification problem. In addition to the physical complexity of secondary structures, finding RNA bases involved in secondary structure ("paired" bases) is itself nontrivial. X-ray crystallography can be used to accurately determine the structure of a transcript at one moment in time (Zhang and Ferré-D'Amaré 2014), but it is prohibitively expensive and cumbersome to perform on a genome-wide level. Computational prediction of secondary structure is widely used but does not always recapitulate known structures from X-ray crystallography and can be computationally intensive (Zhang et al. 2020). Finally, sequencing approaches such as double-stranded RNA (dsRNA) sequencing (Zheng et al. 2010; Li et al. 2012) and SHAPE-seq (Kwok et al. 2013; Liu et al. 2021) have also been successfully used in Arabidopsis, but they are also error prone, dependent on coverage, and also only capture a single possible secondary structure.

Because of these complications, we defined "paired" bases as those implicated in secondary structure by both a computational (LinearPartition) and a sequencing method (dsRNA sequencing) and unpaired bases as those captured by neither (Figure 3.1). To identify these sites, we first polarized ancestral single nucleotide polymorphisms (SNPs) from the Arabidopsis 1,001 genomes project (Weigel and Mott 2009; Alonso-Blanco et al. 2016) and used these ancestral SNPs to create an ancestral pseudo-transcriptome from the TAIR10 assembly (Berardini et al. 2015) by replacing derived alleles present in the Col-0 reference genome with the ancestral SNP. We then extracted mRNA sequences from the pseudo-ancestral reference and inferred base-pairing within these sequences with LinearPartition (Zhang et al., 2020). LinearPartition calculates a partition function for a complete RNA sequence, and it sums equilibrium constants for all possible secondary structures for a sequence (i.e, not just the most likely structure). It outputs a base-pairing matrix that conveys the estimated probability that two bases pair. We focused only on base pairs with high (>0.90) probability of pairing.

We overlapped LinearPartition analysis with empirical data, namely dsRNA data generated by Zheng et al. (2010). These data were generated from 6-week-old Col-0 (flower bud clusters, leaves, and all aerial portions) with the intention of distinguishing true paired bases from less-likely paired bases. We found that dsRNA overlapping regions were, on average, 26.9 nt long, and spanned a total of 187e6 nt, representing X% of the total mRNA database. Given both LinearPartition and dsRNA data, e defined derived pcM mutations as the subset of SNPs: i) that had a LinearPartition probability >0.9, ii) that were detected as paired in dsRNA data, and iii) whose presumed paired base did not also contain a complementary SNP. For example, if the identified base contained an A->G SNP at position 1 and was found to be ancestrally paired with a T at position 20, the SNP at position 1 was not counted if position 20 contained a T->C substitution).

In total, we identified a subset of 201,965 paired bases and 8,469 pcM mutations in Arabidopsis genes. We recognize that these likely do not reflect *all* of the bases involved in secondary structure, because we applied strict criteria. We suspect that many of the bases in the “unpaired” and “unclear” groups are likely more weakly paired. To address this concern, we also compared results from less conservative pcM/upM definitions (e.g., changing the pairing-probability cutoff and relying only on LinearPartition by not considering dsRNA overlap). These less conservative datasets yielded quantitatively similar results (Figs S3,4), but for simplicity we focus on the more conservative subset described above.

Prevalence and distribution of upM and pcM mutations

We categorized SNPs based on their predicted impact on amino acid and secondary structures. This characterization allowed us to partition their fitness effects between distinct "life stages" of gene expression. Because our criteria were strict, genes contained far more upM mutations than pcM mutations: about ~200-times more among both synonymous and nonsynonymous SNPs (Table 3.1). Therefore, our downstream analyses represent a small minority of the total genetic diversity in the 1,001 genomes dataset. We also compared the fractions of SNP effects among upM and pcM mutations using SnpEff annotations (Cingolani et al. 2012). Intriguingly, upM mutations were more likely to occur in untranslated regions (UTRs), and splice sites than pcM, which were more likely to occur in coding regions (Table 3.2). This effect could be caused by two factors: i)paired bases could be less frequent in these non-coding regions, making identification of pcM mutations less likely due to the distribution of structure, or ii) purifying selection maintaining ancestral secondary structure is stronger in these regions, so pcM mutations are purged. Paired bases were about half as likely to to be found in the UTRs as expected given random distributions weighted by the length of these features within genes (i.e., UTRs take up about 15% of the genic space, but only ~8% of paired bases were found within them). However, the percentage of 5' UTR pcM mutations (4.98%) is slightly lower than expected given the percentage of paired bases within 5' UTRs (6.86%). We conducted 10,000 random samples of subsets (n=8,469) from our complete paired site dataset and observed a significant difference (P<0.01) from the anticipated divergence between paired site distribution and pcM distribution.

Reduced nucleotide diversity at paired versus unpaired sites

Nucleotide diversity at synonymous sites (π_S) tends to be higher than at nonsynonymous sites (π_N), which is generally interpreted as the result of purifying selection (Ingvarsson 2010; Osada 2015). A central goal of this paper is to ask whether selection on pcM mutations is observable and comparable to selection on missense mutations. To do so, we focused only on synonymous sites to avoid the confounding effects of selection on nonsynonymous substitutions. We therefore compared nucleotide diversity at segregating synonymous sites between pcM (n=3,214) and upM (n=631,838). We hypothesized that, if pcM mutations are non-neutral, pcM diversity (π_{pcM}) should not be equal to upM diversity (π_{upM}). If selection on secondary structure is strong and similar to that of protein-level selection, we predicted that π_{pcM} should be similar to nonsynonymous diversity (π_N) (n=864,006). As anticipated, π_{pcM} was significantly lower than $\pi_{upM, suggesting that there may be purifying selection on pcM sites}$ However, π_{pcM} was higher than π_N sites, putting it at an intermediate level compared to the other two types of sites (medians: π_{upM} = 8.2e-3, π_{pM} = 7.1e-3, π_N = 3.7e-3)(Figure 3.2a; t-test P < 0.001).

We also wished to ask whether it is likely for selection on secondary structure to interfere with selection for amino acid sequence. One factor which may influence this likelihood is the variation in selection across the length of genes: if different types of selection operate in distinct gene regions, potential interference between the two may be inconsequential, irrespective of the relative strength of each selection type. For example, secondary structure is known to be particularly important at start codons and intron splice sites (Li et al. 2012; Vandivier et al. 2016), while amino acid sequence is more important towards the middle of the protein (Bricout et al. 2023). To understand these spatial differences, we measured π at the three different types of sites across the length of gene coding sequences (Figure 3.2b), finding that the distributions differed between site types: notably, π_N is lowest at the middle of the coding sequence, while π_{upM} is lowest towards the edges. The signal for π_{pM} is noisy, perhaps owing to the low *n* of this category, but it shows a dip towards the 3' end of the coding sequence. We were curious about whether these differences in distribution of π are related to specific feature types, such as translation start sites (start codons), stop codons, and splice junctions. However, we measured π in each category as a function of distance to each of these feature types and found that the correlation was nonsignificant in each case (simple linear model $P>0.05$; Figure S3.1). Overall, selection at the protein and RNA levels seems to target different regions of genes, but it is not clear which features drive those distributions.

Intermediate levels of selection on structural mutations

We next examined allele frequencies of pcM mutations to search for signatures of purifying selection. Based on observed levels of diversity, we hypothesized that synonymous pcM mutations exist at lower frequencies than synonymous upM mutations but at higher frequencies than nonsynonymous mutations; in other words, we predicted that changes to secondary structure should generally be more deleterious than synonymous changes that do not affect secondary structure, but less deleterious than changes to amino acid sequence. Similarly, we predicted that nonsynonymous pcM mutations (i.e., mutations that have an effect on both secondary structure and amino acid sequence) should exist at lower frequencies than nonsynonymous upM mutations, due to effects at both the RNA and protein level. As expected, we found that the site frequency spectrum (SFS) for pcM sites was skewed towards low frequency alleles from that of upM sites, with an abundance of singletons and fewer intermediate and fixed mutations, while frequencies of mutations with unclear structural effects (i.e., those which were identified by either the computational or sequencing approach, but not both) fell at middle frequencies between pcM and upM mutations (Figure 3.3a-b). The set of pcM mutations is a small subset of total SNPs, so we checked for statistical significance between SFSs in two ways: first, using a Kolgorov-Smirnov test, and second using a permutation test (Figure 3.3c-d; Figure S3.2). Both methods found that the spectra for pcM mutations at both synonymous and nonsynonymous sites were highly significantly different from upM mutations. Interestingly, the effect was stronger among nonsynonymous mutations: that is, nonsynonymous mutations that likely interrupt secondary structure appear to be more deleterious than synonymous mutations that interrupt secondary structure (permutation nonsyn. $P \simeq 0$ vs syn. $P = 0.01$).

We evaluated the robustness of these results by investigating datasets based on alternative definitions of pcM and upM. First, we considered sites identified as likely to be paired by LinearPartition, without filtering by dsRNA overlap. For both synonymous (n = 19,252) and nonsynonymous (26,021) pcM, allele frequencies remained significantly different between pcM and upM (Figure S3.3). Second, we used less strict cutoffs for bases likely to be paired by the computational method (filtering by dsRNA cutoff as in the original method). With a base-pairing probability threshold of 0.50 (opposed to the original 0.90), both synonymous (n = 20,596) and nonsynonymous (57,349) pcM allele frequencies remained significantly different between pcM and upM (Figure S3.4).

We next returned to the question of distribution along the length of coding sequences. For each polymorphic site, we measured its distance to the nearest start codon, the nearest stop codon, and the nearest splice site. We categorized sites as "close" to each of these features in order to examine the allele frequencies of pcM mutations that might interrupt important secondary structure at these vital locations along the length of mRNAs. We assigned each site a binary descriptor describing their distance from such features ("proximal" or "distal") using various threshold values of nucleotide (nt) distance (50 nt, 100 nt, 200 nt, and 500 nt) and compared SFS between close and far sites for each feature type (Figure 3.3e-f). In accordance with our results regarding diminished nucleotide diversity at the start and stop sites of coding regions (Figure 3.2), we found that pcM allele frequencies tended to be lower near the start and stop codons. We did not find, however, a clear signal for intron splice sites (Figure S3.5).

Finally, we used these frequency spectra to infer the strength of selection on pcM mutations using ∂a∂I (Gutenkunst et al. 2009; Kim et al. 2017). We first inferred the demographic history of European Arabidopsis populations using the synonymous SFS at unpaired sites (rather than the unfiltered synonymous SFS, as would typically be used). We found that demographic models with a recent bottleneck and a two-epoch model were similarly robust to the data, with modern/ancestral population size ratio estimates between 21.5x and 43.1x and population size expansion 0.3-0.6*$2N_{\text{Ancestral}}$ generations ago. These inferences of recent population size-change fits well with previous demographic inference of Arabidopsis, which is thought to have experienced several bottlenecks and expanded from refugia after the last glacial maximum ~20 KYA (François et al. 2008).

We used both demographic models to separately estimate distributions of fitness effects (DFE) from the SFS of pcM SNPs, and we compared these DFE to distributions estimated from SFS of upM missense SNPs. By doing so, we aimed to partition the fitness effects from changes in secondary structure (synonymous pcM) and compare them to the portion of *non*-structure related fitness effects of mutations which are known to have real fitness effects (nonsynonymous upM). Given our *a priori* expectations about amino acid changes having greater effects than mRNA secondary structure changes, we expected the DFE to be weaker among (synonymous) pcM than among nonsynonymous mutations, and we found that this expectation held true (**Figure 3.3g**); from the gamma distributions, we estimated that synonymous pcM sites had a mean scaled selection coefficient (γ, $2N_{\text{Ancestral}}S$) = 15.81 for the two-epoch model, and γ = 318.48 for the bottleneck model. These mean effect sizes are smaller than the effect sizes for nonsynonymous sites: two-epoch $\gamma = 52.21$ and bottleneck $\gamma = 1570.28$, but they are appreciably different from neutrality ($\gamma < 1$).

Structure-changing SNPs in the geographic landscape

While our previous results hint that secondary structure in many genes may be under directional selection, they provide no functional explanation for why pcM mutations increase in frequency and become fixed. One potential pathway to this outcome is through temperature; secondary structure is guided by intramolecular bonds, and these bonds are subject to the same physical properties as any other chemical bond. Therefore, at lower temperatures, certain secondary structures that are disfavored at high temperatures may occur more frequently, while the opposite occurs at higher temperatures. For the most part, plants are sessile and possess no ability to regulate temperature, so they are subject to the environment in which they germinate. It is therefore possible that evolution may be quite rapid at sites which affect the stability of secondary structures. This phenomenon has been observed in *Arabidopsis* via the computational and experimental validation of two "riboSNitches" (SNPs that change the secondary structure of a transcript)(Halvorsen et al. 2010), which segregate at different frequencies based on geospatial location (Ferrero-Serrano et al. 2022).

The two genes identified by Ferrero-Serrano et al. (2022) represent strong, experimentally validated examples of this phenomenon, where the effect of the mutations on the overall conformation of the molecule has been determined. Given that we were able to detect selection on our set of pcM mutations, we tested whether their findings represent a general trend: in other words, can divergence in frequencies of pcM alleles could be detected in geographically distinct subpopulations of the 1,001 genomes dataset, and does this variation have a climatic component? To define subpopulations, we used two separate methods: first, we used admixture groups from (Alonso-Blanco et al., 2016). Second, we used equally-sized geospatial quadrats, defining subpopulations by their coordinates as accessions which were located together within those transects.

For each of the subpopulations, we found mean frequencies of derived synonymous pcM alleles on a subpopulation-wide level. We then mapped these frequencies geographically (Figure 3.5a; Figure S3.6), finding a slight but clear signal of higher pcM allele frequencies in more northern and eastern subpopulations. We then extracted climatic variables for each individual in each subpopulation based on its geographical coordinates. For each subpopulation, we found the mean for each climatic variable in order to provide each population with a summarized mean for each variable. For subpopulations defined by admixture groups, we found that climatic variables related to temperature were negatively correlated with the frequency of pcM alleles within those populations (Figure 3.5b-c). This effect was strong in BIO1 (mean annual temperature; linear model $P = 0.03$, $R^2 = 0.51$) and BIO6 (minimum temperature of the coldest month; $R^2 = 0.54$, $P = 0.024$). Other variables, such as isothermality (BIO3) and maximum temperature (BIO5) were uncorrelated ($P > 0.05$)(Figure S3.7) For subpopulations defined by spatial quadrats, the results were similar–however, a simple linear regression provided a poor fit for the data, but still showed a weak negative correlation ($R^2 = 0.02$). Instead, we found that a polynomial regression was a better fit, showing a pattern of higher frequencies at both extremes away from the middle (i.e., colder and hotter climates)(generalized linear model BIO1: P=2.69e-3, R^2=0.24; BIO6: P=2.67e-3 R^2=0.15).

3.4 Discussion

Our results suggest that secondary structure is a weak but prevalent factor shaping the evolution of Arabidopsis genes. By classifying mutations at putative ancestrally unpaired sites (upM) versus paired sites (pcM), we found signatures of purifying selection maintaining ancestral base pairing, including reduced diversity at paired sites and skewed allele frequencies. This paper also represents the first (to our knowledge) attempt to compare the strength of this type of selection to selection on mutations that change protein products (nonsynonymous mutations). As expected, since synonymous mutations are conventionally viewed as selectively neutral, we found that selection on secondary structure was weaker than selection on amino acid sequence; depending on the demographic model used, we found that γ for these alleles was, on average, about 20.3% to 30.3% that of nonsynonymous substitutions. Nonetheless, selection on these sites is strong enough to alter the frequencies of alleles, a finding that is in agreement with empirical work showing the existence of strongly deleterious synonymous sites in Drosophila (Lawrie et al. 2013), as well as population genetic estimates in humans that showed reduced frequencies of alleles predicted to change secondary structure (Gaither et al. 2021).

We should address several important caveats regarding our methodology. First, it is important to note that we do not claim to have precise or accurate knowledge about the individual effects of each allele on the broader secondary structure of mRNA. To mitigate structural effects from other mutations, we have excluded alleles with potentially compensatory mutations within the same gene. However, the identification of genuine riboSNitches remains a complex and unresolved challenge, as highlighted by the extensive empirical work by Ferrero-Serrano et al. (2022). Our primary goal was therefore to identify a set of segregating SNPs that with very high confidence to disrupt ancestral secondary structures.

Our process for identifying ancestral secondary structures is based on a number of assumptions. Specifically, we have assumed that the double-stranded RNA (dsRNA) structures present in the Col-0 reference also existed in the ancestral state. While this assumption almost certainly does not hold universally true, given that Col-0 genes have accumulated mutations altering their conformations compared to the ancestral forms, its impact on result accuracy should be limited given that we exclude bases with some evidence for pairing from the "unpaired" category. Moreover, we have shown that we find similar trends with alternative datasets based on lower LinearPartition pairing-likelihoods and/or that disregard the dsRNA data. Nonetheless, our conservative approach, which required confirmation from both computational and empirical methods, likely underestimates the actual count of ancestrally paired bases. Additionally, our methodology assumes minimal effects arising from structural variants (we ignore insertions, deletions, and inversions). It is evident that these types of mutations can have large effects on secondary structure (for example, transposable element insertions are known to be highly structured)(Bousios et al., 2016); as further investigations into the evolution of secondary structures are conducted, addressing the influence of structural mutations should receive special attention, particularly due to recent findings indicating that secondary structure affects the epigenetic response and genome function (Martin et al., 2022).

While it has been long established that synonymous mutations are not entirely neutral (Ikemura 1981), our findings present compelling evidence that mutations influencing structure exert a substantial impact on fitness. Our results also appear to diverge somewhat from recent experimental investigations into the non-neutrality of synonymous alleles (Lawrie et al., 2013), which indicated only a minor influence of secondary structure on selection at synonymous sites in Drosophila. It is difficult to directly compare our results and those of Lawrie et al., because they first identified signals of selection at synonymous sites and then subsequently asked whether structure could account for part of the signal. Instead, we first focused on structure and then investigated differential selection at structured sites. There are several other potential explanations for the discrepancy between studies. One possibility is Drosophila and Arabidopsis differ with respect to the significance of secondary structure. Testing this hypothesis is probably not feasible, given the inherent ambiguity in defining structure within species. Nonetheless, considering the intertwined relevance of secondary structure and small RNAs (Li et al., 2012), coupled with the distinct small RNA production dynamics between plant and animal genomes (), this prospect should not be entirely dismissed. Alternatively, a more probable scenario could be that structure contributes only minimally to the overall fitness effects of synonymous alleles, and these fitness effects are entangled with other collinear variables (such as proximity to translation start/termination sites).

Finally, pcM mutations have quantifiable fitness effects, but are they strong enough to interfere with selection at the protein level? To get a sense for how frequently these types of pleiotropic effects may occur, we built a simple simulation of the coding portion of the *Arabidopsis* genome. We used ther observed CDS lengths (mean = 1275.39), the observed fraction of pcM mutations at nonsynonymous sites, 0.431% (**Table 3.1**), and inferred DFE to estimate the proportion of nonsynonymous sites and genes where this phenomenon may occur. We assumed that the average CDS contains ~66% nonsynonymous sites, and that sites with $\gamma < 1$ are essentially neutral. From these observed/inferred values and assumptions, we estimate that there are roughly 5.06e4 sites across 2.00e4 genes (73% of genes) where the selection coefficient for amino acid sequence dominates selection for non-neutral pcM mutations. On the other hand, there are about 2.59e4 sites where the opposite is true (selection for secondary structure is larger selection for protein) across 1.47e4 genes (54%). While these are rough estimates subject to numerous caveats, they do suggest that (1.) even the small fitness effects observed among pcM mutations in this dataset may affect many sites, and (2.) although the average γ for pcM mutations is an order of magnitude smaller than that of nonsynonymous mutations, pleiotropic interference from this selection is not proportional to this average difference (i.e., RNA-level selection overwhelmed protein-level selection only half as frequently as the inverse, rather than 1/10th as frequently). Of course, the evolutionary importance of such pleiotropic effects depends on the efficacy of selection, and these effects may only infer meaningfully at extremely large population sizes.

Overall, our work sets a framework for studying the evolutionary interplay between selection at the RNA versus protein life stages of gene expression. Our findings also suggest that mRNA structures merit increased attention in plant molecular evolution and perhaps gene function. Further disentangling the numerous (perhaps contradictory) pressures shaping genome evolution will require integrating structural dynamics into molecular and population genetic analysis.

3.5 Materials and Methods

Segregating sites data

We downloaded the segregating sites VCF and SnpEff files from the 1,001 Genomes data center (https://1001genomes.org/data/GMI-MPI/releases/v3.1/)(Weigel and Mott 2009; Alonso-Blanco et al. 2016). We downloaded the TAIR10 *A. thaliana* assembly, *A. lyrata* (v1.0)(Hu et al. 2011) assembly, associated gene/exon/UTR annotations, and CDS sequences from EnsemblPlants. dsRNA coverage data was retrieved in from the NCBI Gene Expression OmniBus Accession # GSE23439 (Zheng et al. 2010) and converted from the TAIR9 to TAIR10 assembly using CrossMap (0.6.4). Assembly conversion was validated by checking overlap with TAIR10 genes using Bedtools (2.27.1) intersect.

Identification of derived upM and pcM mutations

We polarized the ancestral state of biallelic *A. thaliana* sites in the 1,001 genomes dataset by whole genome alignment with *A. lyrata* using AnchorWave (1.0)(Song et al. 2022), and anchoring to CDS sequences from *A. thaliana*. We checked the polarization using a separately polarized VCF produced using minimap2, and we found 95% agreement between the two methods (i.e., 95% of alleles had the same ancestral state call).

To identify ancestrally paired and unpaired sites, we first constructed a pseudo-ancestral genome from the polarized 1,001 genomes VCF using GATK FastaAlternateReferenceMaker. We then extracted pseudo-ancestral longest mRNA sequences using bedtools getfasta. We ran each sequence through LinearPartition (v1.0)(Zhang et al. 2020) and extracted positions from the LinearPartition base pairing probability matrices where probabilities were > 0.9 (we separately analyzed sets with less strict criteria), and we associated these positions along the length of genes with their chromosomal positions. We used samtools tabix to extract VCF positions which overlapped computationally inferred pairing bases. We then looked for overlap between these positions and the dsRNA coverage data in R (4.2.1) using IRanges (2.30.1) and GenomicRanges (1.48.0)(Lawrence et al. 2013). Positions where both criteria were met (high base pairing probability and dsRNA coverage) were considered pcM sites. We further filtered these sites by finding overlap with potential compensating mutations using the base pairing probability files from LinearPartition. We estimated overlap with UTRs and introns using GenomicRanges in R.

Nucleotide diversity, allele frequency, and DFE analysis

We analyzed nucleotide diversity using VCFtools (Danecek et al. 2011). First, we extracted subsets of the full 1,001 genomes VCF file for each category (syn. pcM, nonsyn. pcM, etc.) using samtools tabix and the annotations from our paired/unpaired site identification above and the 1,001 genomes SnpEff file. We calculated π with the per-site method in each gene using VCFtools. We measured distance between sites and various genic features (starts, stops, and intron junctions) using GenomicRanges in R.

Allele frequencies and frequency spectra were calculated in R using custom code integrating vcfR (Knaus and Grünwald 2017), data.table (Dowle & Srinivasan, 2023), and tidyverse (Wickham et al., 2021) functions. Permutation tests for differences between spectra were calculated by looping the sample_n function on the upM mutation VCF file 10,000 times, building a SFS for each subsample, and taking the difference between that permutation spectrum and the true upM spectrum. Sites "distal" and "proximal" from genic features were identified using GenomicRanges, and frequency spectra for these categories were constructed by subsetting the full VCF in R.

To estimate the distributions of fitness effects for different mutation types, we used dadi (Gutenkunst et al. 2009) with functions from fitdadi (Kim et al. 2017) in Python3. We first inferred demography using the synonymous upM SFS, which we read into dadi from the VCF file (subset using samtools). We analyzed various demographic models with various starting parameters to optimize the most robust demographic models. Our final demographic models for the DFE in Figure 3.3 used the following parameters. Two-epoch: nu = 26.9, T = 0.113; and bottlegrowth-1d: nuB = 0.00634688, nuF = 2.60504245, and T=0.0762755. We estimated the DFE using the synonymous pcM SFS and the nonsynonymous pcM+upM SFS separately, modeling DFE as a simple gamma distribution. We plotted DFEs in Python using matplotlib (Hunter 2007). We estimated the mean of each gamma distribution by multiplying the shape*scale parameter of each.

Geospatial and climatic correlations

We downloaded geographical coordinates for each accession in the 1,001 genomes dataset using the AraPheno database (https://arapheno.1001genomes.org/studies/). We defined subpopulations by admixture groups defined in (Alonso-Blanco et al. 2016), and defined quadrats by nonoverlapping square areas defined by fixed increments of latitude and longitude, and we only analyzed quadrats with >3 accessions. The quadrats used in Figure 5 are incremented by 3.5 degrees each. We extracted climatic data from WorldClim using raster (Hijmans et al. 2023) in R. For each admixture subpopulation, we found the mean of each BIO climatic variable by extracting the climatic data for each accession and taking the mean of all accessions in the population. For quadrats, we found the mean within the quadrat based on its location. We found frequencies of pcM alleles in each subpopulation by subsetting the pcM VCF by accession ID and taking the mean frequency of all alleles present in each subpopulation in R. Models of climate variable association were determined using the lm() and glm() functions for admixture and quadrat subpopulations respectively.

Pleiotropy simulation

We multiplied the CDS lengths of each 1:1 *A. thaliana*:*A. lyrata* ortholog by 0.66 to approximate the number of nonsynonymous sites across the genome. For each gene, we assigned each site as either paired or unpaired based on the probability data from **Table 3.1**. We then assigned each site a “protein-level” fitness effect (γ) by pseudo-randomly drawing a value from the nonsynonymous DFE gamma distribution in R. This selection coefficient assignment was pseudo-random, because the maximum value of assignments was capped at 1,000 ($2N_AS$). We then assigned paired sites a “RNA-level” fitness effect by the same method, but this time sampling from the synonymous pcM DFE. We evaluated the accuracy of our DFE simulations by comparing the means of these sampled DFE to the “true” means estimated from the gamma distributions (shape*scale). We repeated the simulation several times, finding that the derived solutions changed minimally.

Figures

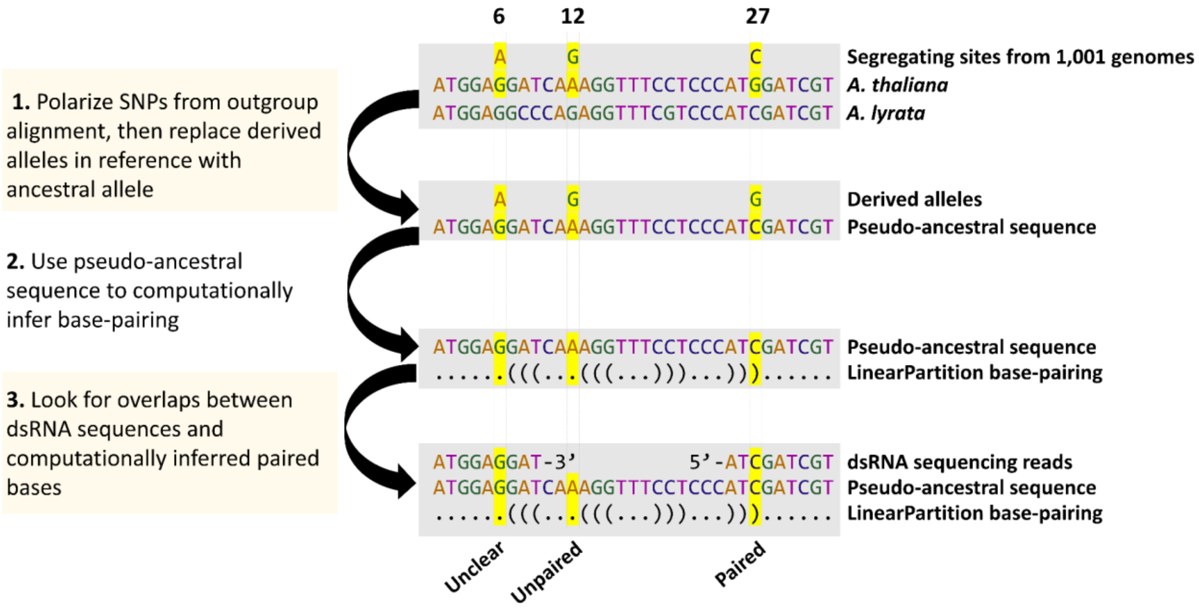

6
12
27
A G C
ATGGAGGATCAAAGGTTTCCTCCCATGGATCGT
ATGGAGGCCCAGAGGTTTCGTCCCATCGATCGT
Segregating sites from 1,001 genomes
A. thaliana
A. lyrata
1. Polarize SNPs from outgroup alignment, then replace derived alleles in reference with ancestral allele
A G G
ATGGAGGATCAAAGGTTTCCTCCCATCGATCGT
Derived alleles
Pseudo-ancestral sequence
2. Use pseudo-ancestral sequence to computationally infer base-pairing
ATGGAGGATCAAAGGTTTCCTCCCATCGATCGT
......(((...(((...)))...)))......
Pseudo-ancestral sequence
LinearPartition base-pairing
3. Look for overlaps between dsRNA sequences and computationally inferred paired bases
ATGGAGGAT-3’ 5’-ATCGATCGT
ATGGAGGATCAAAGGTTTCCTCCCATCGATCGT
......(((...(((...)))...)))......
dsRNA sequencing reads
Pseudo-ancestral sequence
LinearPartition base-pairing
Unclear
Unpaired
Paired

Figure 3.1. Schematic representation of upM/pcM identification method. Three 1,001 genomes SNPs are shown at positions 6, 12, and 27 within a hypothetical gene: one that is unclear (6, *left*), one that is an upM (12, *middle*) and one that is a pcM (27, *right*). Derived alleles must have both high LinearPartition base-pairing probability (shown as parentheses above) and dsRNA coverage to be considered pcM.

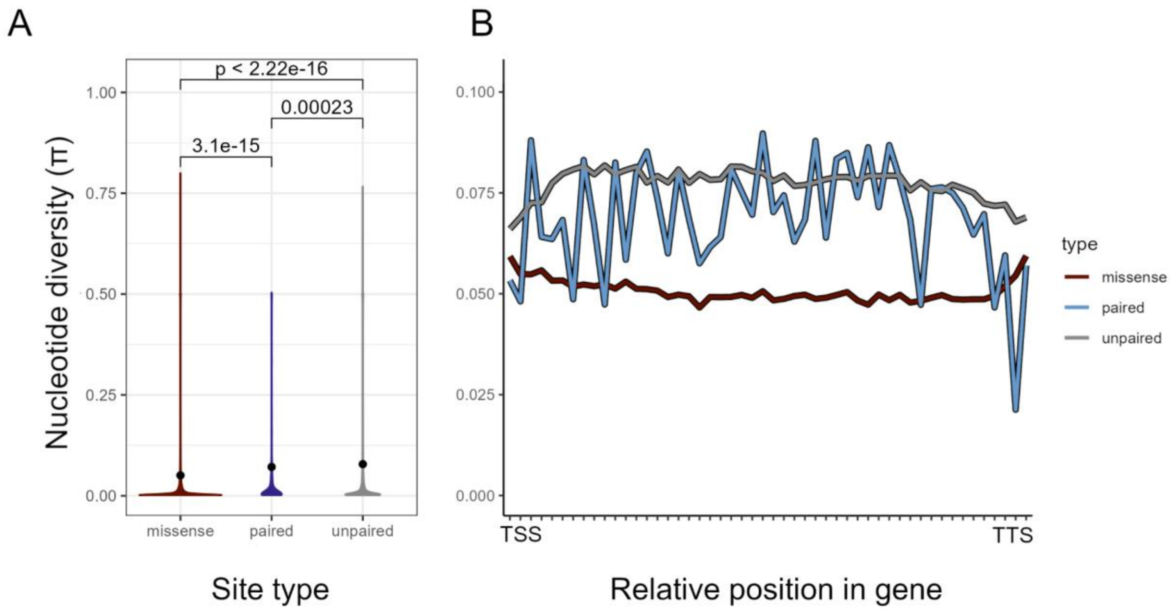


Figure 3.2: Nucleotide diversity from 1,001 genomes dataset at synonymous paired, synonymous unpaired, and nonsynonymous sites. (A) Distributions of nucleotide diversity between site types. (B) π summarized over the length of all analyzed CDS. The x-axis represents length-standardized windows across the span of all analyzed genes from the 5' end (transcription start site) to the 3' end (transcription termination site).

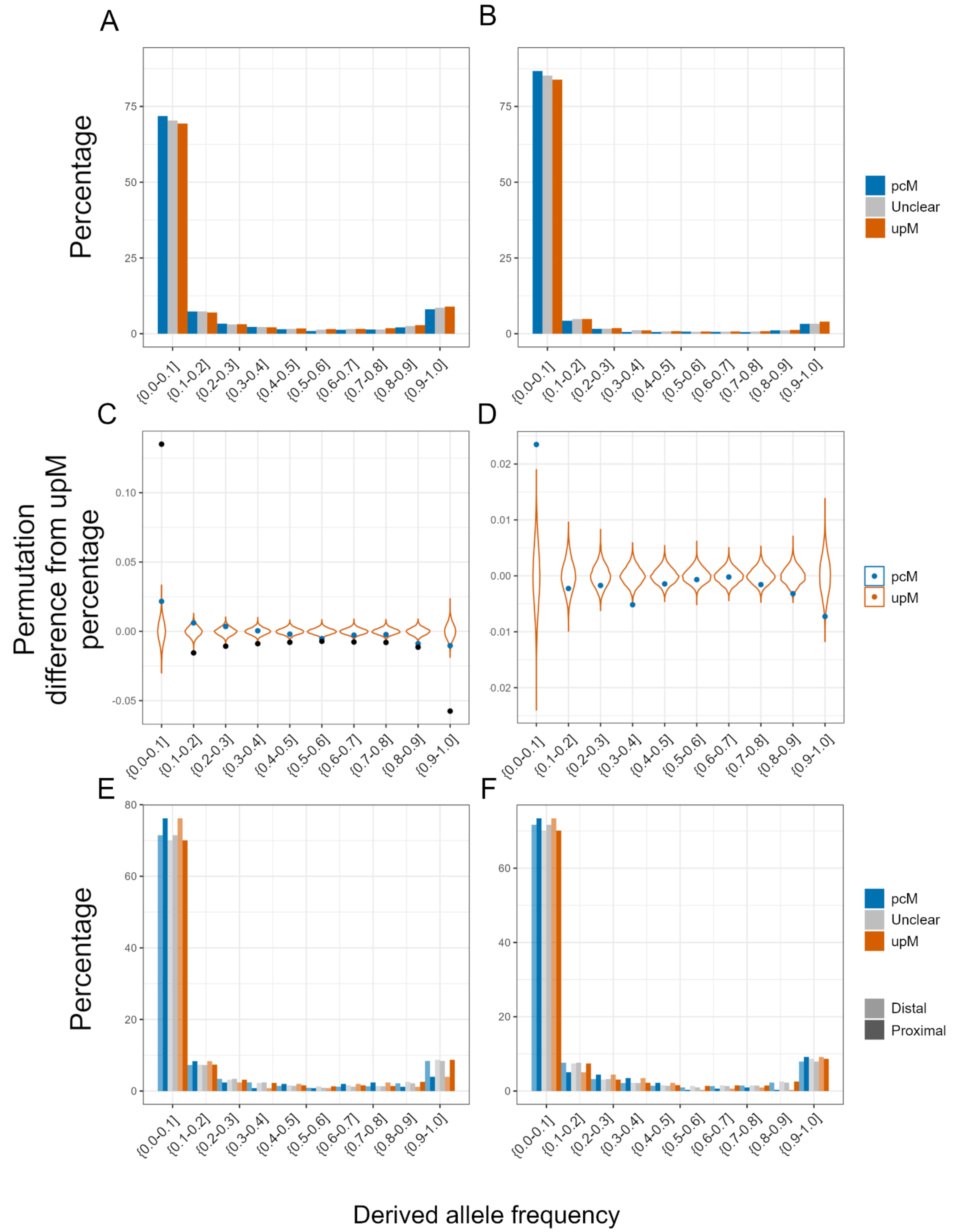


Figure 3.3. (A) Unfolded site frequency spectra (SFS) showing derived allele frequencies of synonymous alleles categorized by their inferred effect on ancestral secondary structure. (B) SFS of nonsynonymous alleles in each category. Permutation distributions for

differences between paired and unpaired SFS at synonymous sites (C)(black dot shows missense for scale) and nonsynonymous (D). Violins represent the distribution of differences in random samples (same *n* as paired sites) from the unpaired data, while points show the true differences. Differences (y-axis) were calculated as the percentage of alleles in each pcM SFS bin subtracted from the percentage in the same upM SFS bin (e.g., {0-0.1] in pCM minus {0-0.1] in upM etc.)(E) SFS of distal vs proximal sites (100 nt cutoff) to 5' UTR (start codon). (F) the same as (E)(100 nt cutoff) but for stop codon/3' UTR proximity. Solid bars represent alleles frequencies at sites near intragenic features.

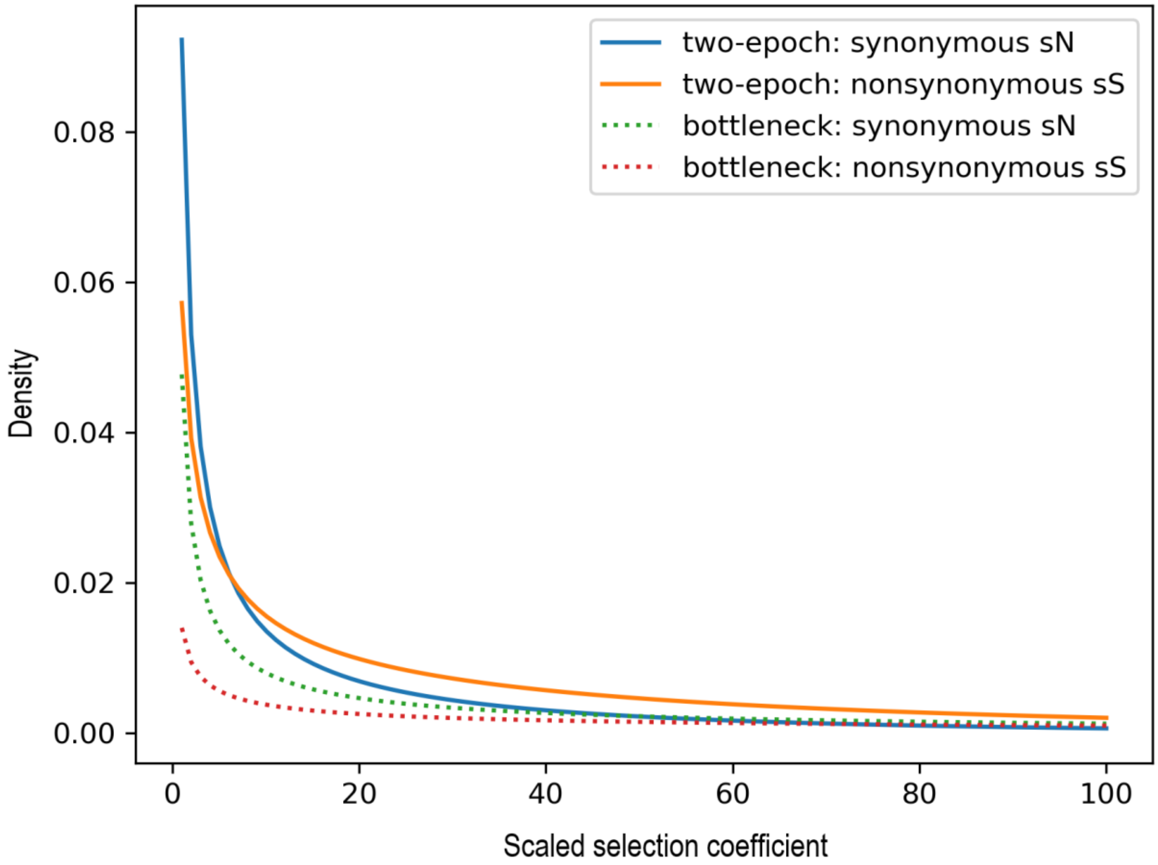


Figure 3.4. Distribution of fitness effects (DFE) for mutation types under two demographic models. Distributions of the scaled selection coefficient, γ (2*$N_{ancestral}$S) are governed by shape and scale parameters estimated from dadi under a simple gamma distribution and were inferred from the SFS of synonymous upM (i.e., the subset of mutations unlikely to have an effect at either the protein or RNA level).

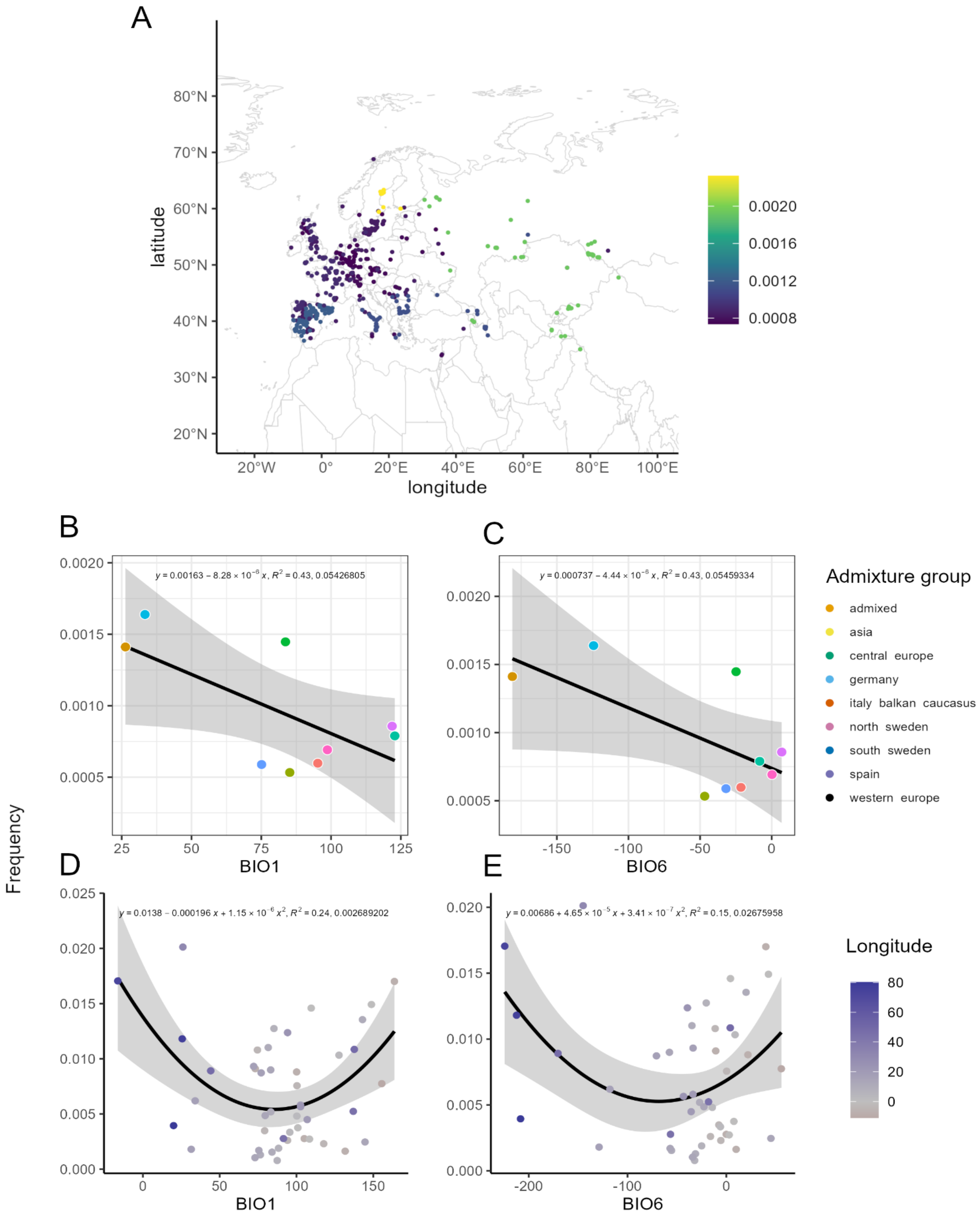


Figure 3.5: Geospatial variation in structural mutations. (A) Mean pcM frequencies within admixture groups across geographic space. (B-C) Mean pcM frequencies within admixture

groups as a function of mean climate variable within group. (D-E) Mean pcM frequencies within quadrat-defined subpopulations as a function of mean climate variable within quadrat. Points are colored by the longitude of the middle of the quadrat.

Tables

Table 3.1. SNPs categorized by effect on amino acid sequence and secondary structure.

	upM	pcM	Structurally unclear[1]
Nonsynonymous	3,790	864,006	22,231
Synonymous	3,214	631,838	16,038

[1]SNPs detected by the computational method (LinearPartition) or dsRNA coverage, but not both.

Table 3.2. SnpEff annotations for upM vs pcM SNPs.

SNP effect	Number upM	Percentage upM in SnpEff category[1]	Number pcM	Percentage pcM in SnpEff category[1]	Percentage difference (upM - pcM)
synonymous variant	631838	27.23%	3214	37.95%	-10.72%
missense variant	864006	37.23%	3790	44.75%	-7.52%
disruptive inframe deletion	2825	0.12%	19	0.22%	-0.10%
inframe insertion	2437	0.11%	11	0.13%	-0.02%
inframe deletion	3449	0.15%	14	0.17%	-0.02%
frameshift variant+start lost	450	0.02%	3	0.04%	-0.02%
frameshift variant+stop gained	222	0.01%	2	0.02%	-0.01%
frameshift variant+stop lost	377	0.02%	2	0.02%	-0.01%
initiator codon variant	231	0.01%	1	0.01%	0.00%
stop retained variant	1315	0.06%	4	0.05%	0.01%
disruptive inframe insertion	808	0.03%	2	0.02%	0.01%
start lost	1232	0.05%	2	0.02%	0.03%
stop lost	1532	0.07%	2	0.02%	0.04%
splice acceptor variant	3389	0.15%	8	0.09%	0.05%
splice donor variant	3537	0.15%	3	0.04%	0.12%
stop gained	15577	0.67%	41	0.48%	0.19%
frameshift variant	24822	1.07%	64	0.76%	0.31%
5' UTR premature start codon gain variant	18977	0.82%	42	0.50%	0.32%
5' UTR variant	161299	6.95%	422	4.98%	1.97%
splice region variant	94490	4.07%	55	0.65%	3.42%
3' UTR variant	276907	11.93%	681	8.04%	3.89%
intron variant	210835	9.09%	87	1.03%	8.06%

Total	**2320555**	**8469**

[1] Percentage of SNPs in category (upM or pcM) that had a particular SNP effect.

Supplementary Figures

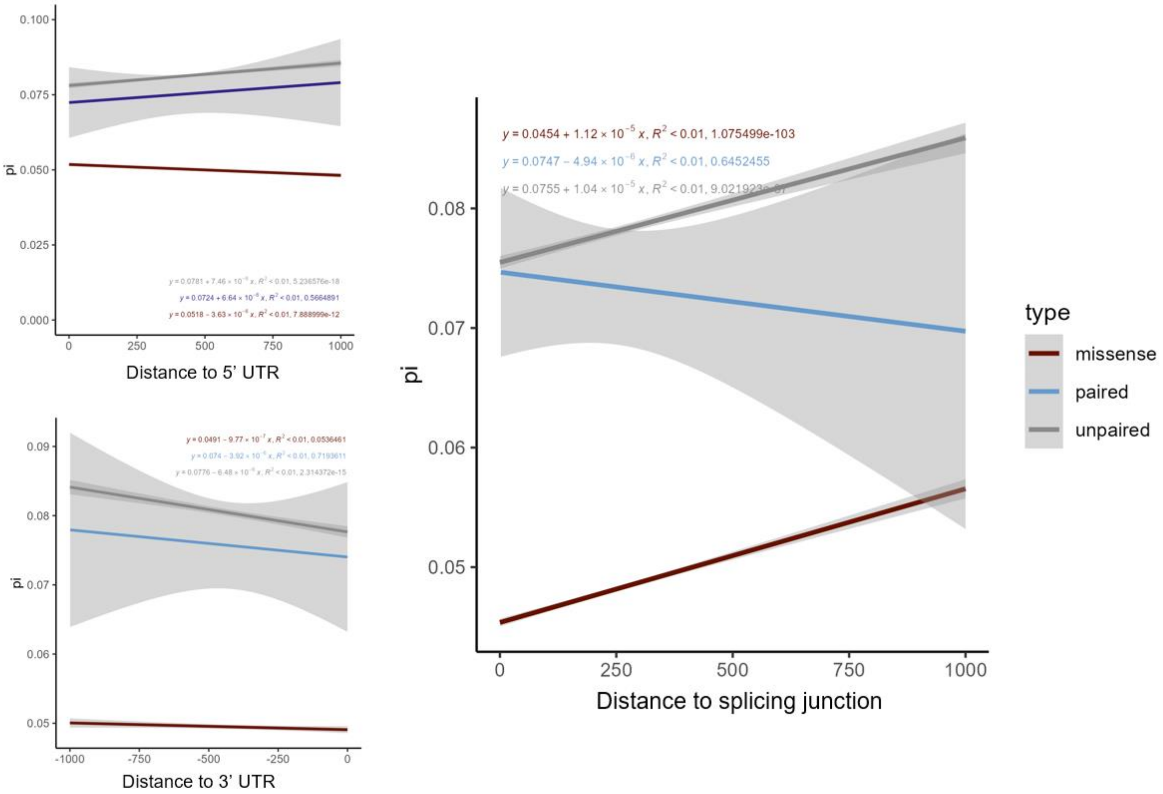


Figure S3.1. Correlations between sitewise π (nucleotide diversity) at different site types and distance to intragenic features. None of the correlations at paired sites were significant, the correlations were weak (R2 < 0.01) but significant for unpaired sites. This is probably a reflection of the greater sample size in these site types (Table 3.2) rather than a reflection of biologically important differences. Distance to 3' UTR is given as a negative value (more negative = farther from UTR)

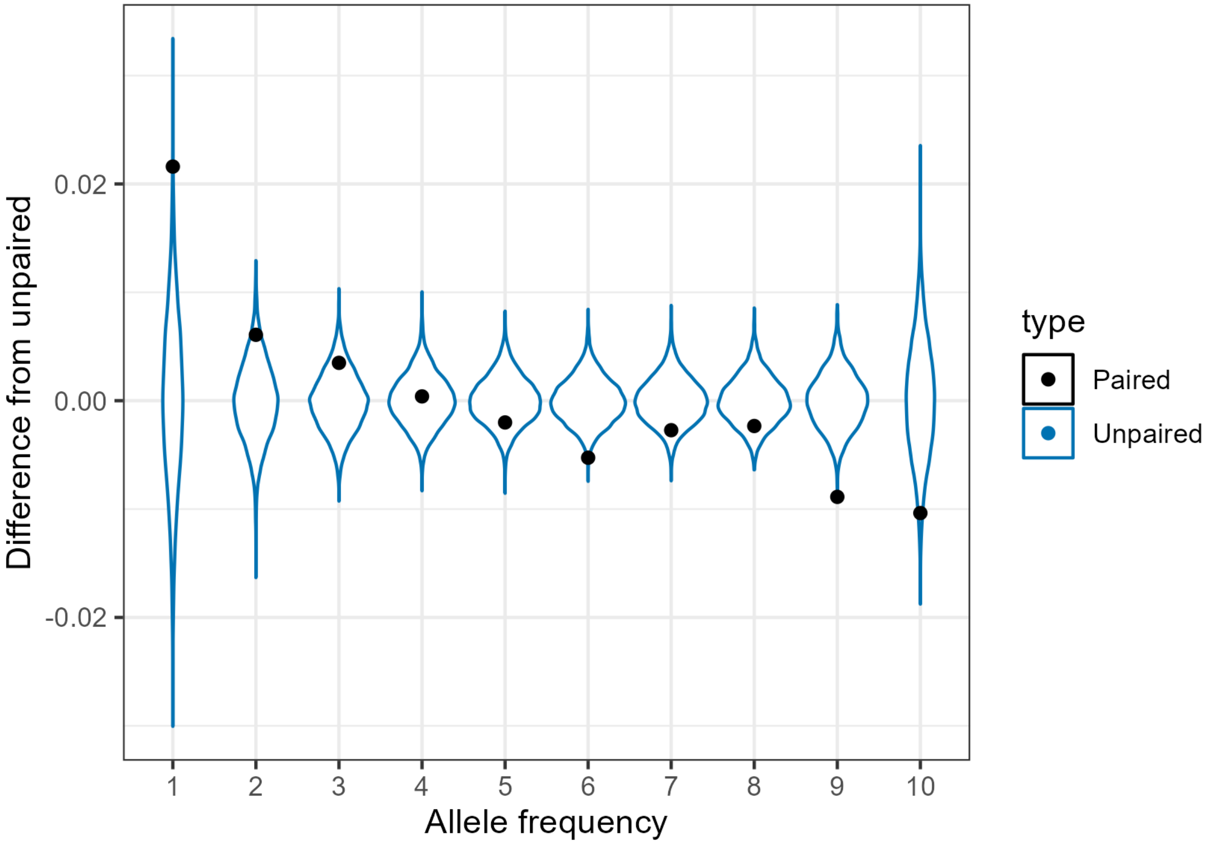


Figure S3.2. Permutation tests for frequencies of synonymous pcM and upM. These sites required a pairing probability > 0.90 (50%) and dsRNA coverage. Violins represent the distribution of differences in random samples (same n as paired sites) from the unpaired data, while points show the true differences. Differences (y-axis) were calculated as the percentage of alleles in each pcM SFS bin subtracted from the percentage in the same upM SFS bin (e.g., {0-0.1] in pCM minus {0-0.1] in upM etc.). Differences in bins 1, 9, and 10 were significant.

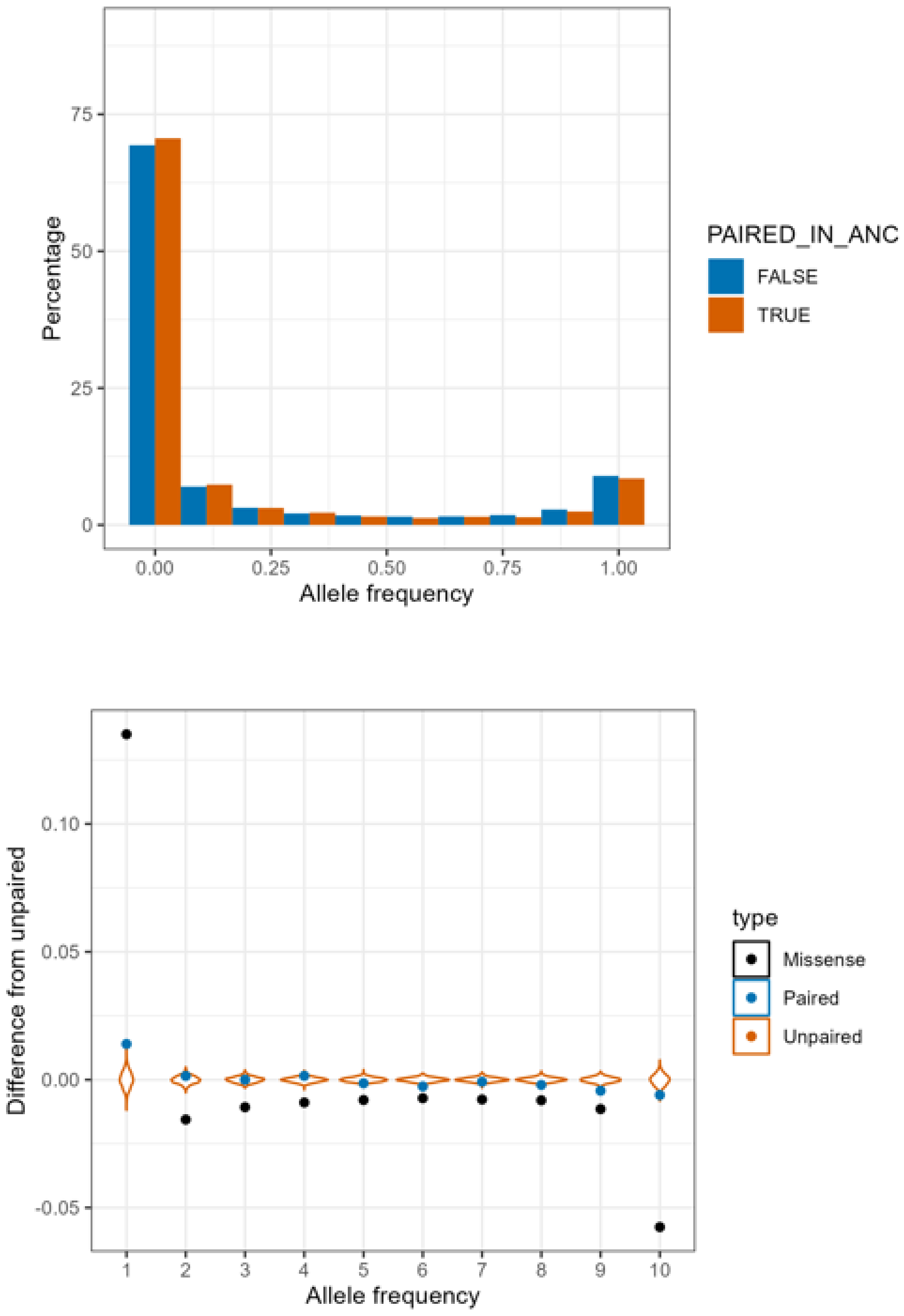


Figure S3.3. Site frequency spectrum and permutation tests for frequencies of synonymous pcM, unclear, and upM defined without the requirement of dsRNA coverage (all sites with base pairing probability > 0.90 were considered paired sites). No unclear category is included because only one metric of pairing was used (LinearPartition).

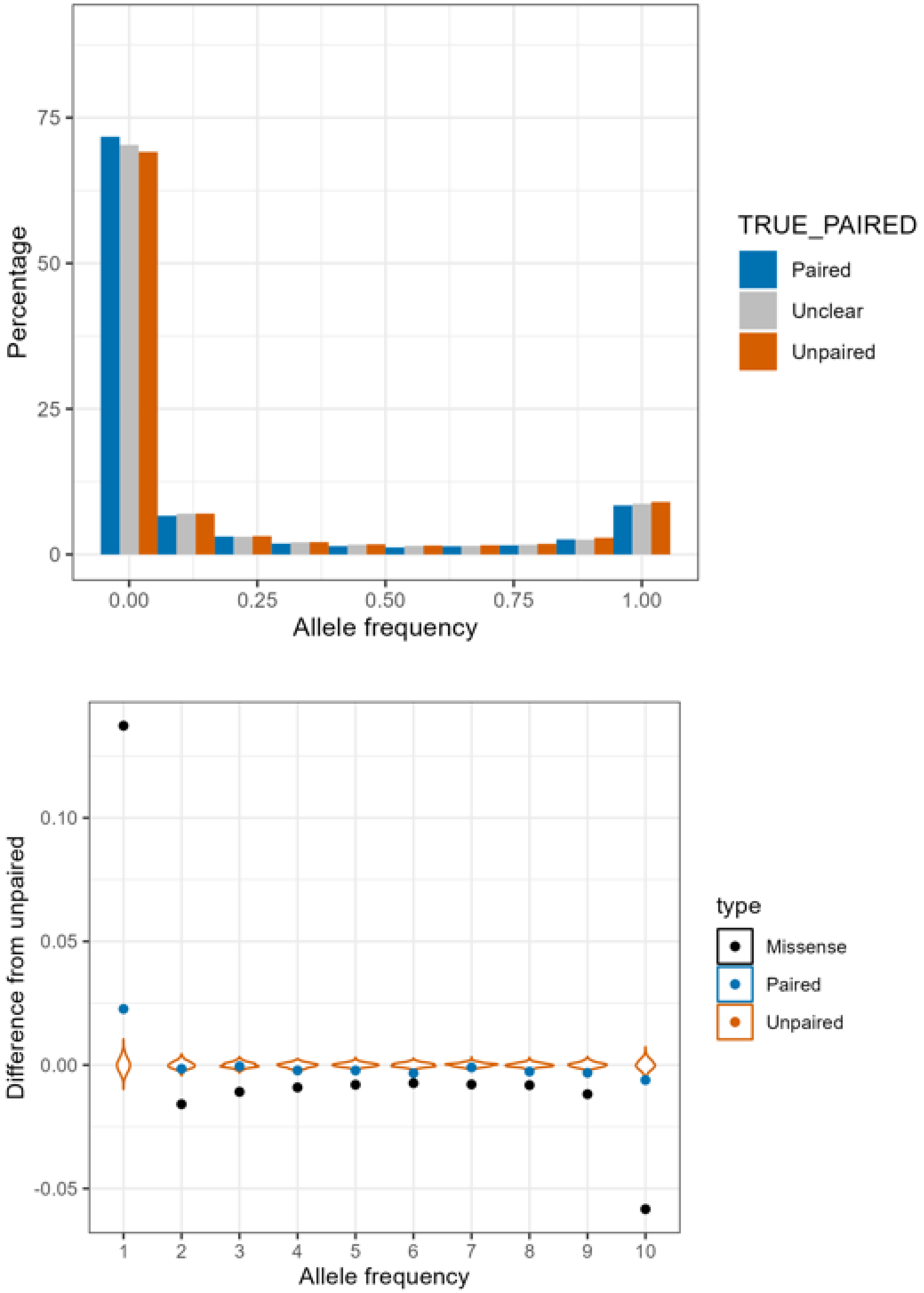


Figure S3.4. Site frequency spectrum and permutation tests for frequencies of synonymous pcM, unclear, and upM defined with a lower base pairing probability requirement from LinearPartition. These sites required a pairing probability > 0.50 (50%) and dsRNA coverage.

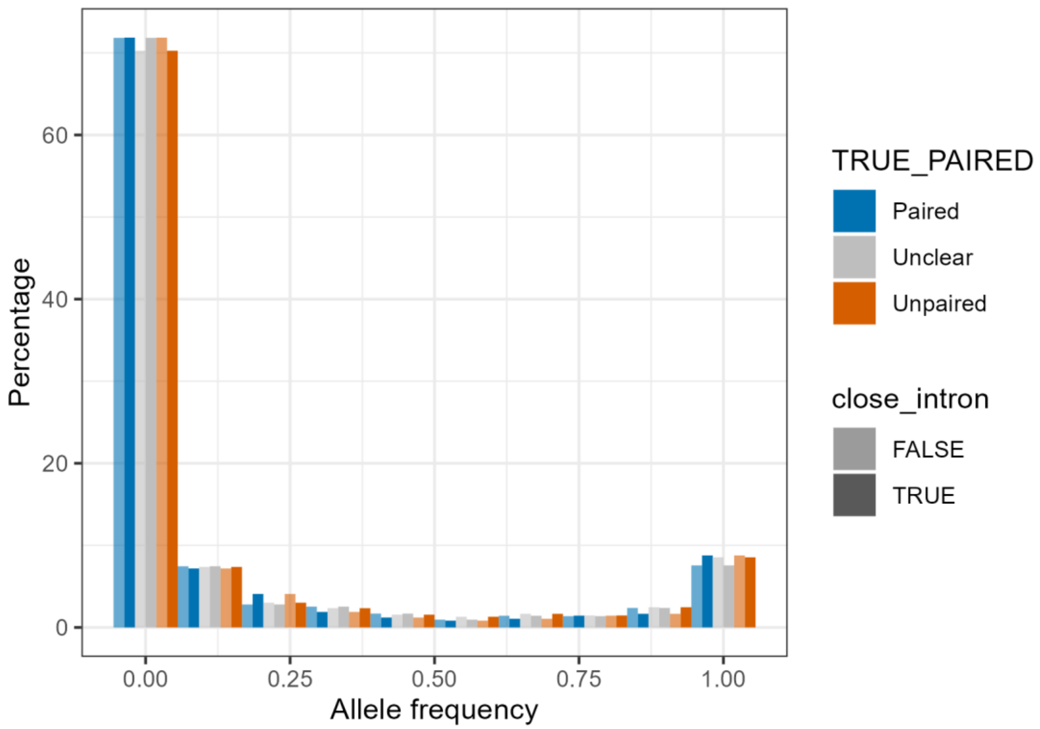


Figure S3.5. SFS of distal vs proximal sites (100 nt cutoff) to intron splice sites. Solid bars represent alleles frequencies at sites near intragenic features. pcM alleles do not show differences depending on distance from splice sites.

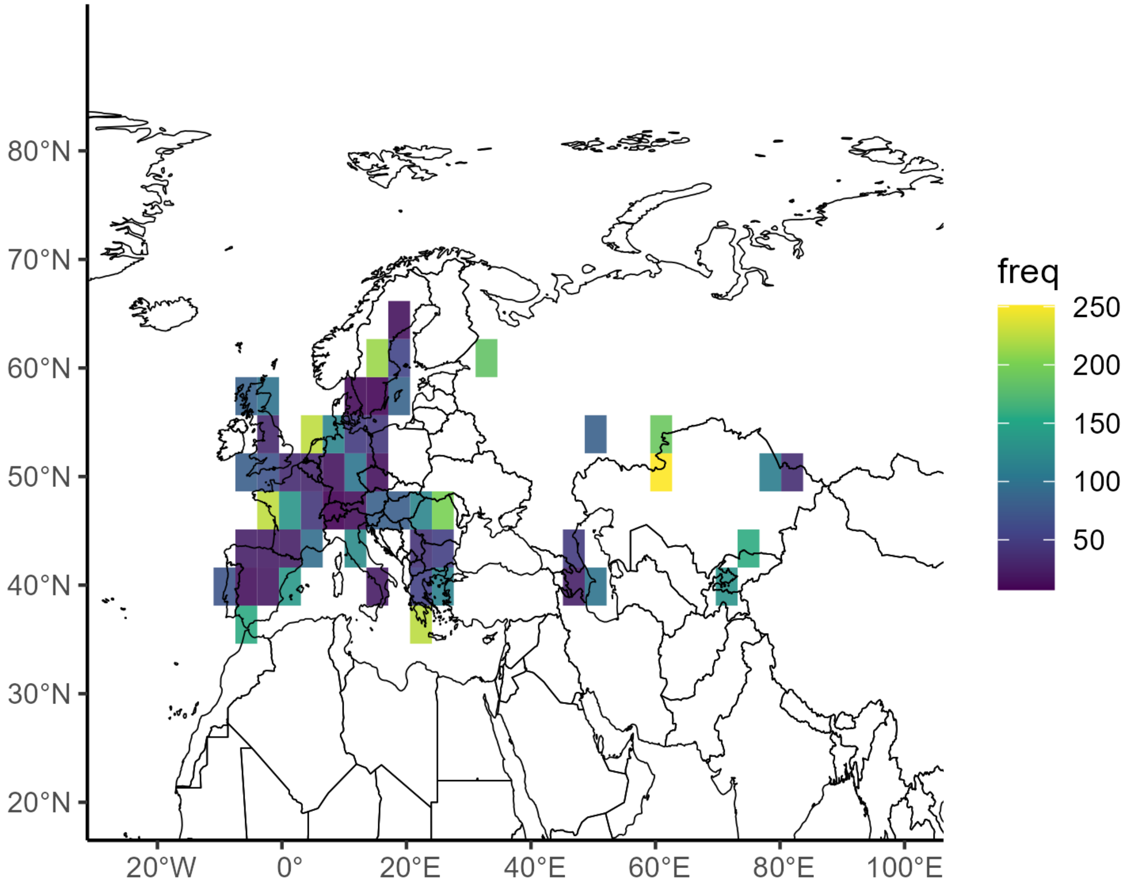


Figure S3.6. Geospatial variation in structural mutations. Heatmap shows sums of mean pcM frequencies within quadrate-defined subpopulations across geographic space. Quadrats were defined arbitrarily by distance. They are non-overlapping, and each accession belongs to a single quadrat-population.

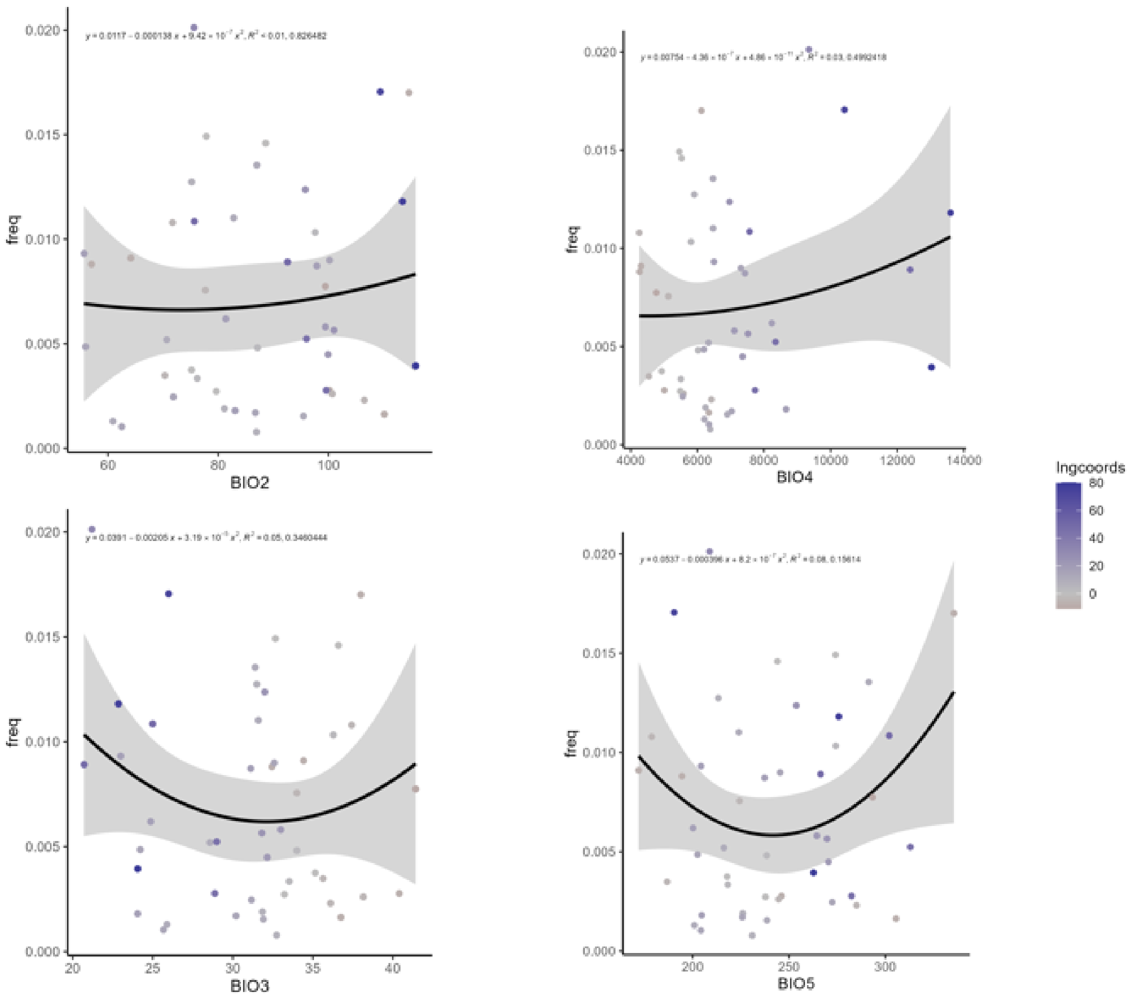


Figure S3.7. Mean pcM frequencies within quadrat-defined subpopulations as a function of mean climate variable within quadrat. Points are colored by the longitude of the middle of the quadrat. Lines are defined by a log regression. The correlations for these models were not qualitatively different from other models used.

References

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166:481–491.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. genesis 53:474–485.

Booker TR, Jackson BC, Keightley PD. 2017. Detecting positive selection in the genome. BMC Biology 15:98.

Bousios A, Diez CM, Takuno S, Bystry V, Darzentas N, Gaut BS. 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.

Bricout R, Weil D, Stroebel D, Genovesio A, Roest Crollius H. 2023. Evolution is not Uniform Along Coding Sequences. Mol. Biol. Evol. 40:msad042.

Bullock SL, Ringel I, Ish-Horowicz D, Lukavsky PJ. 2010. A’-form RNA helices are required for cytoplasmic mRNA transport in Drosophila. Nat. Struct. Mol. Biol. 17:703–709.

Buratti E, Baralle FE. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol. Cell. Biol. 24:10505–10514.

Chamary J, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 6:R75.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6:80–92.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinforma. Oxf. Engl. 27:2156–2158.

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.

Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Human Molecular Genetics 12:205–216.

Ferrero-Serrano Á, Sylvia MM, Forstmeier PC, Olson AJ, Ware D, Bevilacqua PC, Assmann SM. 2022. Experimental demonstration and pan-structurome prediction of climate-associated riboSNitches in Arabidopsis. Genome Biol. 23:101.

Fraïsse C, Puixeu Sala G, Vicoso B. 2019. Pleiotropy Modulates the Efficacy of Selection in Drosophila melanogaster. Mol Biol Evol 36:500–515.

François O, Blum MGB, Jakobsson M, Rosenberg NA. 2008. Demographic History of European Populations of Arabidopsis thaliana. PLOS Genet. 4:e1000075.

Gaither JBS, Lammi GE, Li JL, Gordon DM, Kuck HC, Kelly BJ, Fitch JR, White P. 2021. Synonymous variants that disrupt messenger RNA structure are significantly constrained in the human population. GigaScience 10:giab023.

Gu W, Wang X, Zhai C, Xie X, Zhou T. 2012. Selection on Synonymous Sites for Increased Accessibility around miRNA Binding Sites in Plants. Molecular Biology and Evolution 29:3037–3044.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genet. 5:e1000695.

Halvorsen M, Martin JS, Broadaway S, Laederach A. 2010. Disease-Associated Mutations That Alter the RNA Structural Ensemble. PLOS Genet. 6:e1001074.

Hijmans RJ, Etten J van, Sumner M, Cheng J, Baston D, Bevan A, Bivand R, Busetto L, Canty M, Fasoli B, et al. 2023. raster: Geographic Data Analysis and Modeling. Available from: https://cran.r-project.org/web/packages/raster/index.html

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. 43:476–481.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9:90–95.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151:389–409.

Innan H, Stephan W. 2001. Selection Intensity Against Deleterious Mutations in RNA Secondary Structures and Rate of Compensatory Nucleotide Substitutions. Genetics 159:389–399.

Ingvarsson PK. 2010. Natural Selection on Synonymous and Nonsynonymous Mutations Shapes Patterns of Polymorphism in Populus tremula. Mol. Biol. Evol. 27:650–660.

Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. Genetics 206:345–361.

Kimura M. 1968. Evolutionary Rate at the Molecular Level. Nature 217:624–626.

Knaus BJ, Grünwald NJ. 2017. vcfr: a package to manipulate and visualize variant call format data in R. Mol. Ecol. Resour. 17:44–53.

Kozak M. 1988. Leader Length and Secondary Structure Modulate mRNA Function under Conditions of Stress. Mol. Cell. Biol. 8:2737–2744.

Kwok CK, Ding Y, Tang Y, Assmann SM, Bevilacqua PC. 2013. Determination of in vivo RNA structure in low-abundance transcripts. Nat. Commun. 4:2971.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. PLOS Comput. Biol. 9:e1003118.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong Purifying Selection at Synonymous Sites in D. melanogaster. PLOS Genet. 9:e1003527.

Lebeuf-Taylor E, McCloskey N, Bailey SF, Hinz A, Kassen R. 2019. The distribution of fitness effects among synonymous mutations in a gene under directional selection.Landry CR, Wittkopp PJ, Venkataram S, editors. eLife 8:e45952.

Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. 2012. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. Plant Cell 24:4346–4359.

Liu J, Zhang Y, Lei X, Zhang Z. 2008. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biol. 9:R69.

Liu Z, Liu Q, Yang X, Zhang Y, Norris M, Chen X, Cheema J, Zhang H, Ding Y. 2021. In vivo nuclear RNA structurome reveals RNA-structure regulation of mRNA processing in plants. Genome Biol. 22:11.

Mathews DH. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA 10:1178–1190.

Martin G, Solares E, Muyle A, Bousios A, Gaut BS. 2022. Diverse patterns of secondary structure across genes and transposable elements are associated with siRNA production and epigenetic fate. :2022.10.17.512609. Available from: https://www.biorxiv.org/content/10.1101/2022.10.17.512609v1

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

Monroe JG, 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.

Osada N. 2015. Genetic diversity in humans and non-human primates and its evolutionary consequences. Genes Genet. Syst. 90:133–145.

Rahman S, Kosakovsky Pond SL, Webb A, Hey J. 2021. Weak selection on synonymous codons substantially inflates dN/dS estimates in bacteria. Proc. Natl. Acad. Sci. 118:e2023575118.

Seffens W, Digby D. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res. 27:1578–1584.

Song B, Marco-Sola S, Moreto M, Johnson L, Buckler ES, Stitzer MC. 2022. AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. Proc. Natl. Acad. Sci. 119:e2113075119.

Steitz TA, Moore PB. 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem. Sci. 28:411–418.

Svitkin YV, Pause A, Haghighat A, Pyronnet S, Witherell G, Belsham GJ, Sonenberg N. 2001. The requirement for eukaryotic initiation factor 4A (elF4A) in translation is in direct proportion to the degree of mRNA 5' secondary structure. RNA N. Y. N 7:382–394.

Vandivier LE, Anderson SJ, Foley SW, Gregory BD. 2016. The Conservation and Function of RNA Secondary Structure in Plants. Annu. Rev. Plant Biol. 67:463–488.

Varani G, McClain WH. 2000. The G·U wobble base pair. EMBO Rep. 1:18–23.

Weigel D, Mott R. 2009. The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol. 10:107.

Wickham H, François R, Henry L, Müller K, RStudio. 2021. dplyr: A Grammar of Data Manipulation. Available from: https://CRAN.R-project.org/package=dplyr

Williams AS, Marzluff WF. 1995. The sequence of the stem and flanking sequences at the 3' end of histone mRNA are critical determinants for the binding of the stem-loop binding protein. Nucleic Acids Res. 23:654–662.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15:496–503.

Zhang H, Zhang L, Mathews DH, Huang L. 2020. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. Bioinformatics 36:i258–i267.

Zhang J, Ferré-D'Amaré AR. 2014. New molecular engineering approaches for crystallographic studies of large RNAs. Curr. Opin. Struct. Biol. 26:9–15.

Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, Yang J, Cao K, Wang L-S, Gregory BD. 2010. Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in Arabidopsis. PLOS Genet. 6:e1001141.

CONCLUSIONS

The physical structures of DNA and RNA molecules are crucial for proper genome function and evolution. DNA wraps around histone proteins to form chromatin, which controls accessibility and transcriptional activity. Meanwhile, RNA folds into structures that determine splicing, translation, and localization. Because these three-dimensional aspects of nucleic acid structure influence phenotype, natural selection acts on the many processes governing these structures. Moreover, TEs compose the majority of genomic space in many species, and they have large effects on both chromatin as well as secondary structure.

In my dissertation, I used epigenetic and genomic information in plants to study the mechanisms and effects of epigenetic modifications, small RNA expression, and RNA secondary structure. I hoped to bridge the gap between two disparate fields: the bioinformaticians and molecular biologists who study epigenetics and RNA structure tend to do so with an eye towards the molecular processes that establish and maintain these patterns. Meanwhile—partially because understanding the evolutionary effects of these structures is somewhat reliant on understanding their molecular biology—evolutionary biologists tend to focus on genes and mutations that change the amino acid composition of their products (or, perhaps more frequently, they do not characterize the mechanical effects of the mutations involved in evolution).

In my first chapter, I studied evolutionary dynamics of methylated CHH (mCHH) islands—which are peaks of CHH methylation upstream of genes (Gent et al. 2013; Li et al. 2015)—across eight grass species spanning the breadth of the Poaceae family with a wide range of genome sizes and structures. Using whole genome bisulfite sequencing and expression data, I examined interspecies patterns of CHG, CHG and CHH methylation near genes. I found expected patterns, like CG methylation predominating over CHG and CHH within genes and methylation being relatively low near transcription start and stop sites (Niederhuth et al. 2016). I also found that peaks of CHH methylation occurred immediately upstream and downstream of genes in most species.

Based on work by Hollister & Gaut (2009), I hypothesized that mCHH islands should be more frequent in large genomes with more heterochromatin and transposable elements (TEs). However, I found no relationship between genome size and mCHH island prevalence or levels, and smaller genomes even tended to have a higher proportion of mCHH island genes. Consistent with islands reflecting TE silencing, mCHH island genes were closer to TEs than non-island genes in all species. Focusing on maize, rice, and barley, I confirmed mCHH islands were associated with DNA transposons, especially terminal inverted repeats (TIRs), more than retrotransposons. However, examining mCHH island sequences, many islands had little homology to known TEs, indicating that additional factors beyond TE silencing contribute to islands.

I found mCHH island genes were slightly more expressed than non-island genes, but significantly so in only a few species. Surprisingly, absence of gene-body methylation (gbM) was a stronger island predictor than TE proximity. Analyzing one-to-one orthologs across species, I showed mCHH islands were generally not evolutionarily conserved, unlike gbM. Presence of lineage-specific TEs coincided with lineage-specific mCHH islands, partially explaining lack of conservation. Additional factors like gbM and gene length also correlated with mCHH island conservation.

My work in Chapter 1 established previously-unknown genic properties correlated with mCHH islands, beyond just TE silencing effects, but cast doubt on their evolutionary conservation. Important questions remain about how islands form and function. My results implicate aberrant gene transcription, which could engage RNA-directed DNA methylation, especially when encompassing nearby TEs. mCHH islands may then help moderate TE effects on gene expression, or simply be products of this process. The negative association between mCHH islands and gbM, which suppresses aberrant transcription (Teissandier and Bourc’his 2017), fits this model.

In my second chapter, I studied microRNA (miRNA)-like secondary structures across the maize genome, including in TEs and genes. I used two computational methods to identify regions with properties resembling miRNA precursor hairpins (Lorenz et al. 2011; Zhang et al. 2020). These miRNA-like regions were common, present in the majority of annotated features. Different TE types varied in prevalence of structures, with DNA transposons and LTR retrotransposons containing more miRNA-like regions than non-LTR retrotransposons. The locations of structures also differed among TEs; for example, Copia element hairpin structures tend to lie within their LTRs.

I hypothesized miRNA-like regions may act as substrates for small RNA production through Dicer-like processing of double-stranded RNA (Li et al. 2012). As predicted, miRNA-like regions in both TEs and genes consistently mapped a higher diversity of small RNAs. In support of our hypothesis, this was especially true of small RNAs 21-22 nucleotides in length, which are the lengths usually produced from RNA interference/Dicer-like degradation. Additionally in support of this model, the small RNA enrichment was reduced in non-autonomous TEs unlikely to be transcribed. I also found miRNA-like regions exhibited peaks of DNA methylation, particularly in the CHH context deposited by small RNA-directed pathways. In many ways, secondary structures correlated with downstream repressive epigenetic signals.

In genes, I propose that tradeoffs occur between important secondary structure and deleterious effects of small RNA production. Structured genes were more highly expressed than unstructured genes, suggesting functional importance of structure, but genes with extremely stable structures had reduced expression. Across diverse maize lines, structured genes also exhibited greater expression variability related to their small RNA levels. This implies secondary structure in genes may spawn small RNAs that destabilize expression.

These genome-wide analysis revealed widespread miRNA-like secondary structures with traceable molecular and phenotypic associations. I propose structured regions in genes are experience conflict between maintenance of beneficial RNA folding while avoiding excess double-stranded RNA that risks small RNA-mediated silencing. My work expands our understanding of the ubiquity of strong endogenous structures and their diverse effects on genome regulation and evolution.

In Chapter 3, I studied mutations affecting mRNA secondary structure in *Arabidopsis thaliana* genes using population genomic data (Weigel and Mott 2009; Alonso-Blanco et al. 2016). I developed a novel method to identify derived alleles that likely disrupt ancestral base pairing. I termed these mutations "pair changing" (pcM) and compared them to mutations at unpaired sites (upM). I found pcM mutations exhibited signatures of purifying selection relative to upM mutations, including reduced nucleotide diversity and skewed allele frequencies, indicating selection maintains ancestral mRNA structures. The strength of selection on synonymous pcM mutations was weaker than on nonsynonymous mutations but sufficient to alter allele frequencies.

My results suggest RNA-level selection is prevalent but weaker than protein-level selection, as expected since synonymous mutations are conventionally viewed as neutral. I found spatial differences in diversity levels along genes, with purifying selection targeting distinct regions for amino acid sequence versus secondary structure. This spatial separation may limit potential conflicts between protein- and RNA-level selection.

Nonetheless, based on inferred selection coefficients and the prevalence of inferred ancestral secondary structure, I estimated that over half of Arabidopsis genes contain sites where secondary structure selection could exceed protein-level selection. Hence, pleiotropic effects on both RNA structure and amino acids may conflict at some loci, reducing the realized fitness gains of adaptive amino acid substitutions. Overall, my work reveals widespread selection maintaining mRNA structures and suggests that RNA-level selection should be considered in molecular evolution and population genetics.

References

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166:481–491.

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

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19:1419–1428.

Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. 2012. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. Plant Cell 24:4346–4359.

Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, Hirsch CN, Dawe RK, Madzima TF, McGinnis KM, 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.

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms Mol. Biol. 6:26.

Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA, et al. 2016. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 17:194.

Teissandier A, Bourc'his D. 2017. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. EMBO J. 36:1471–1473.

Weigel D, Mott R. 2009. The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol. 10:107.

Zhang H, Zhang L, Mathews DH, Huang L. 2020. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. Bioinformatics 36:i258–i267.