Reproductive phasiRNAs in grasses are compositionally distinct from other classes of small RNAs

Parth Patel[1, 2], Sandra Mathioni[2,4], Atul Kakrana[1, 2] , Hagit Shatkay[1, 2, 3], Blake C. Meyers[4, 5]

[1] Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19714, USA.

[2] Delaware Biotechnology Institute, University of Delaware Newark, DE 19714, USA.

[3] Department of Computer and Information Sciences, University of Delaware, Newark, DE 19714, USA.

[4] Donald Danforth Plant Science Center, St. Louis, MO 63132, USA.

[5] University of Missouri – Columbia, Division of Plant Sciences, 52 Agriculture Lab, Columbia, MO 65211, USA.

Short Title: sRNA classification via machine learning.

Author for correspondence:

Blake C. Meyers

Tel: +1 (314) 587-1422

Email: bmeyers@danforthcenter.org

ABSTRACT

Little is known about the characteristics and function of reproductive phased, secondary, small interfering RNAs (phasiRNAs) in the Poaceae, despite the availability of significant genomic resources, experimental data, and a growing number of computational tools. We utilized machine-learning methods to identify sequence-based and structural features that distinguish phasiRNAs in rice and maize from other small RNAs (sRNAs). We developed Random Forest classifiers that can distinguish reproductive phasiRNAs from other sRNAs in complex sets of sequencing data, utilizing sequence-based (k-mers) and structural features. The classification performance attained is >80% in accuracy, sensitivity, specificity, and positive predicted value. Feature selection identified important features in both ends of phasiRNAs. We demonstrated that phasiRNAs have strand specificity and position-specific nucleotide biases potentially influencing AGO sorting; we also predicted targets to infer functions of phasiRNAs, and computationally-assessed their stability and structure characteristics relative to other sRNAs. Our results demonstrate that machine-learning methods effectively identify phasiRNAs despite the lack of characteristic features typically present in precursor loci of other small RNAs, such as sequence conservation or structural motifs. The 5'-end features we identified provide insights into AGO-phasiRNA interactions; we describe a hypothetical model of competition for AGO loading between phasiRNAs of different nucleotide compositions.

INTRODUCTION

Molecular and genomic studies coupled with deep sequencing have identified roles of many endogenous non-coding RNAs (ncRNAs) and small RNAs (sRNAs) at numerous developmental stages in many organisms (Tisseur et al., 2011; Guttman and Rinn, 2012; Kung et al., 2013; Borges and Martienssen, 2015; Axtell, 2013). Flowering plants have three major classes of sRNAs, all derived from ncRNAs: microRNAs (miRNAs), heterochromatic or Pol IV-dependent small interfering RNAs (P4-siRNAs), and phased, secondary, small interfering RNAs (phasiRNAs). This latter class has grown considerably with the discovery of germline-enriched, reproductive phasiRNAs most well described in the Poaceae, namely maize and rice (Johnson et al., 2009; Komiya et al., 2014; Zhai et al., 2015b). Two classes of reproductive phasiRNAs are known: 21-nt pre-meiotic phasiRNAs that peak in abundance during somatic cell specification in maize (one week after anther initiation), and 24-nt meiotic phasiRNAs that peak during meiosis and are

detectable until pollen maturation (one to two weeks after pre-meiotic phasiRNAs peak) (Zhai et al., 2015b). The timing, localization, and narrow developmental time window of accumulation of the 21- and 24-nt phasiRNAs is conserved in rice and maize (Fei et al., 2016). While the biogenesis and spatiotemporal patterns of accumulation of these reproductive phasiRNAs are now well described, our understanding of their function is still limited.

An analogy can be drawn between phasiRNAs of grass anthers and the PIWI-interacting RNAs (piRNAs) of animals, in aspects such as their biogenesis, developmental timing, and enrichment in reproductive organs. piRNAs play crucial roles in transposable element (TE) silencing and germline development from flies to fish to mammals (Meister, 2013). Yet, plants have a highly elaborate RNA-directed DNA methylation pathway (RdDM) that effectively silences most TEs (Matzke and Mosher, 2014), thus their need for yet another TE-silencing pathway is debatable. Emerging evidence implicates plant reproductive phasiRNAs in development; for example, MEL1, a rice Argonaute (AGO), is required for normal anther development (Nonomura et al., 2007), and this AGO binds to 21-nt reproductive phasiRNAs (Komiya et al., 2014). The functions and targets are yet to be determined for both 21- and 24-nt reproductive phasiRNAs, and it is not known whether they function in *cis* or *trans*. In fact, it is possible that they are merely decay products of more functionally relevant long ncRNA precursors. Understanding the role of phasiRNAs requires more detailed molecular and computational analyses that could also serve to direct future experiments. For example, identifying characteristic features or motifs that differentiate reproductive phasiRNAs from other sRNAs (miRNAs, P4-siRNAs, etc.) may provide clues as to their AGO loading or targets.

Work on animal piRNAs has used sequence-based characteristics to demonstrate their unique properties; significant insights have resulted from so-called *alignment-free approaches*. These methods use short nucleotide sequences, k-mers, and other features to distinguish between different types of sRNA sequences, and classify them into distinct groups. For example, Zhang *et al.*, 2011 developed a classifier that can distinguish piRNAs from non-piRNAs (miRNAs, snoRNAs, tRNAs, and lncRNAs) with precision over 90% and a recall over 60%, within a five-fold cross-validation. This work utilized data from five species including mice, humans, rats, fruit flies, and nematodes, effectively discriminating piRNAs. In a test of the validity of their classifier,

Zhang *et al.* (2011) detected >87,000 of ~130,000 piRNAs, in a total set of >600,000 sRNAs. Brayet *et al.* (2014) used a similar approach to identify piRNAs from sequences of several types (miRNAs, tRNAs, and 25-33 nt sequences from protein coding genes) in human and fruit flies with precision over 85% and a recall over 88%. As such, these alignment-free approaches are quite promising for characterizing subsets of sRNAs within large and complex pools of un-sorted sequences.

Our aim was to start with a set of known reproductive phasiRNAs (21- or 24-nt), develop and optimize a classification pipeline, and ultimately use this to sort previously unknown sequences from plants to find reproductive phasiRNAs from among other types of small RNAs. An additional product of this work was the sequence-based characteristics that comprise the output of the classifier, as these might identify novel aspects of reproductive phasiRNAs. In this work, we implemented machine-learning approaches to examine plant 21-nt pre-meiotic and 24-nt meiotic reproductive phasiRNAs, and to build a classifier that can automatically distinguish them from other sRNAs (i.e., miRNAs and P4-siRNAs). Our results provide insights into phasiRNA sequence composition profiles and biases, sequence- and structure-based features, aspects of their biogenesis, features that may influence AGO sorting, predicted targets and possible functions.

RESULTS

Cross validation results distinguishes reproductive phasiRNAs from other sRNAs

We sought to identify unique attributes of rice and maize reproductive phasiRNAs relative to other, better-described small RNA classes. To do this, we developed a machine learning-based workflow focused on sequence-based and structural features of plant small RNAs (Figure 1). To train the classifier, we used as positive examples known reproductive phasiRNAs from rice and maize, including both 21-nt and 24-nt phasiRNAs, while the negative sets consisted of P4-siRNAs, miRNAs, tRNAs, and rRNAs (see Methods). We built and evaluated classifiers by utilizing different negative sets; the performance measurements were achieved via five-fold CV, and this 5-fold CV was completed five times on our datasets (see Methods for a more complete explanation). As noted above, the classification results were in terms of ACC, SE, SP, PPV, and AUC.

The results obtained from our classification pipeline using different negative sets, are shown in Table 1. The results, according to all performance measures, exceed 0.8 (with one exception, see below), for both 21- and 24-nt phasiRNAs. We first examined 21-nt phasiRNAs, and we compared phasiRNAs to a mixture of sRNAs that include selected miRNAs, P4-siRNA, tRNAs, and rRNAs; these latter four cases represent the four major negative sets (i.e. not phasiRNAs) found in a typical plant sRNA dataset (Table 1). In an initial comparison, the negative sets included miRNAs, tRNAs, and rRNAs of different lengths (randomly selected endogenous sequences); all P4-siRNAs were 24-nt. The classifier identified the combined negative set as quite distinct relative to 21-nt phasiRNAs (Table 1). In addition, we computed the area under the ROC curve (AUC) (Supplemental Figure 1C), demonstrating the performance of the above-mentioned classifier with an averaged AUC of 0.97. Next, we combined untrimmed miRNAs and 24-nt P4-siRNAs and still achieved high classification performance (Table 1). This classification result could indicate that length is a primary factor in classification, and thus we used trimmed negative sets to assess this possibility.

We classified 21-nt phasiRNAs relative to 24-nt P4-siRNAs with 3 nt trimmed from the 3' end (see Methods); the classifier performed reasonably well (>0.8 for all measurements, ACC, SP, SE, PPV, and AUC). We also trimmed P4-siRNAs 3 nt from the 5' end or from the internal 11th, 12th, and 13th positions, observing no substantial changes in classification. We concluded that 21-nt reproductive phasiRNAs are compositionally distinct from P4-siRNAs. Finally, we related 21-nt phasiRNAs to 21-nt miRNAs (some trimmed, see Methods), and found similar ACC, higher SE, but slightly lower SP and PPV; the lower SP may be attributed to fewer miRNAs (756 vs 2000 21-nt phasiRNAs in the positive set). This imbalance possibly misclassified some miRNAs, hence low specificity and a high number of false positives (lower PPV). We followed the same procedure in classifying the 24-nt phasiRNAs, first with 24-nt P4-siRNAs and next with the combined negative set. In both cases, the classification of the negative set against 24-nt phasiRNAs, resulted in strong scores for all four performance measurements (Table 1), again indicative that the 24-nt phasiRNAs are also compositionally distinctive. In addition, we observed an averaged AUC of 0.942 when classifying 24-nt phasiRNAs with the combined negative set (Supplemental Figure 1D). We concluded that our classification pipeline successfully classified reproductive phasiRNAs relative to other endogenous plant sRNAs with high values for ACC, SE, SP, PPV, and AUC.

Next, we investigated the predictive sensitivity of our pipeline, asking whether it can correctly classify previously unutilized members of a larger positive set of reproductive phasiRNAs. In other words, these new sequences were different from the 2000 used in the positive set during cross validation study. The classifier was given, first, 27500 21-nt phasiRNA sequences and, next, 7750 24-nt phasiRNA sequences (rice and maize combined, in each case). The classification pipeline based on models that combined each of the negative sets (miRNAs + P4-siRNAs + tRNAs + rRNAs) predicted 26208 21-nt phasiRNAs (SE > 0.95) and 7093 24-nt phasiRNAs (SE > 0.91), achieving high sensitivity (Table 2). We concluded that the pipeline is an effective method for *de novo* classification of plant small RNAs.

Position-specific biases in phasiRNAs relative to other small RNAs

Next, knowing that reproductive phasiRNAs are distinct from other classes of small RNAs, we sought to characterize these differences in greater detail, at the single nucleotide level. We computed single-nucleotide sequence profiles for the most abundant 1000 reproductive phasiRNAs (for 21-nt or 24-nt, rice and maize data combined), miRNAs, and 24-nt P4-siRNAs, determining the frequencies of each nucleotide (A, C, G, and U) at each position (Figure 2). We then compared the position-specific base usage between the reproductive phasiRNAs and either miRNAs or 24-P4-sRNAs by conducting a two-tailed, rank sum test ($P = 1e^{-5}$) to identify positions with statistically significant base usage that would distinguish phasiRNAs from either miRNAs or P4-siRNAs (Figure 2).

At a significance level of 10^{-5}, comparing the 21-nt phasiRNAs and miRNAs, we found that the usage of bases at eight positions differed significantly (positions 1, 2, 3, 8, 9, 12, 19, and 21; Figure 2A). Next, we repeated the calculation, comparing 21-nt reproductive phasiRNAs and 24-nt P4-siRNAs (Figure 2B), demonstrating significant differences at positions 1, 3, 14, 19, 20, and 21. Combining these results, we made several observations: (i) in these abundant 21-phasiRNAs, there was a 5' nucleotide preference for C, consistent with a recent report (Komiya et al., 2014), but a strong depletion of G. (ii) We noticed a peak of U at the 14th position in the phasiRNAs (relative to P4-siRNAs), unusual as there were no other biased positions between 2 and 19; the only other internal position showing bias was a G at position 8 in the miRNAs (Figure 2A). (iii) In the 3' end

156 of the 21-nt phasiRNAs, we observed a peak of G at the 19th position (with a depletion of A), and
157 U at the 21st position (G strongly disfavored). This representation of G at the 19th position was
158 investigated in more detail below.
159
160 We conducted a similar analysis comparing the position-specific base usage between the 24-nt
161 reproductive phasiRNAs and P4-siRNAs. We found that positions 1, 3, 7, 10, 11, 14, 18, 20, 21,
162 22, and 23 were statistically different (Figure 2C); in other words, the 24-nt phasiRNAs and P4-
163 siRNAs differed substantially in their base usage over the full length of the molecules. All of the
164 over-represented nucleotides in 24-nt phasiRNAs were either A or U (Figure 2C); again, the 5'
165 and 3' ends showed differences in the two classes of molecules, and internally positions 10/11
166 were overrepresented for U in the 24-phasiRNAs. These correspond to the same two internal
167 positions critical for directing cleavage by AGO proteins in the case of miRNAs (Carrington and
168 Ambros, 2003), so we noted this for subsequent phasiRNA target analysis (see below). The 3' end
169 difference was most striking, as in the P4-siRNAs, there was a high frequency of G from the 20th
170 to 24th positions and a coincident depletion of U (Figure 2C), whereas 24-nt phasiRNAs had an
171 overrepresented A at the 22nd position and U at the 3' end. Therefore, we identified several notable
172 sequence-based features of both classes of reproductive phasiRNAs, observed at both the 5'- and
173 3'-ends and a small number of internal positions; the 24-nt phasiRNAs also displayed an overall
174 nucleotide composition distinct from that of P4-siRNAs. These differences likely have
175 implications for AGO loading and phasiRNA-target interactions, while also potentially explaining
176 the non-stoichiometric abundances of individual phasiRNAs at each *PHAS* locus.
177
178 As observed for animal miRNAs (Chatterjee et al., 2011), it's possible that the non-stoichiometric
179 abundances at a *PHAS* locus results from AGO loading and subsequent stabilization of functional
180 siRNAs. We next computed the sequence profile of 'present' or 'absent' reproductive phasiRNAs
181 in rice and maize; in other words, at a given *PHAS* locus, some phasiRNAs are never observed in
182 the sequenced sRNAs, but we could extract these computationally and assess their sequence
183 composition biases relative to those we detected experimentally. In a comparison to those
184 phasiRNAs detected in the sequencing data, we observed a substantial, overall sequence
185 composition difference for 21-nt phasiRNAs (Figure 2A,C versus Supplemental Figure 2A, left).
186 The differences for present versus absent 24-nt phasiRNAs were less pronounced and mainly

towards the 3' end (Figure 2C, left, versus Supplemental Figure 2A, right). To ensure that the profiles for detected phasiRNAs were not unduly biased by the selected use of only the top 1000 sequences, we also plotted sequence profile of all sequenced 21-nt phasiRNAs (Supplemental Figure 2B, left) and the 24-nt phasiRNAs (Supplemental Figure 2B, right) from the positive set (see Methods). We observed no noticeable changes in the sequence profile relative to the abundance-selected subset (i.e. Figure 2A,C), except a slightly higher representation of 5'U compared to 5'C in the 21-nt phasiRNAs. The comparison of present versus absent reproductive phasiRNAs demonstrated significant differences in nucleotide composition, consistent with relative stabilization of those detected reproductive phasiRNAs after biogenesis; this may reflect AGO loading, target interactions, or other sequence-specific functions of these phasiRNAs.

The duplex nature of phasiRNA biogenesis impacts nucleotide composition

The observed nucleotide biases at the 19th position in the 21-nt phasiRNAs and at the 22nd position in the 24-nt phasiRNAs were the next subject of our investigation. Dicer cleavage of dsRNA typically yields a 2-nt 3′ overhang (Macrae et al., 2006), and these two positions, from a long, dsRNA precursor, complement the 5'-end of the antisense, duplex sRNA. In a schematic integrating position-specific biases (Figure 3A,B), the influence of the most-frequent nucleotides in the top strand on the bottom strand composition is highlighted for the first and last three nucleotide positions; for example, the 19th position G corresponds to a 5' C (1st position) for the duplex phasiRNA. Thus there is a potential co-bias between the 1st and 19th positions, such that if both strands of a 21-nt phasiRNA duplex require a specific 5' nucleotide to ensure proper AGO loading (like a 5' C), the 19th position will co-vary with the 1st position. Alternatively, if only one strand of the duplex is loaded (due to a requisite 5' nucleotide, the primary biogenesis strand, or other reasons) and the duplex partner is dispensable, then the 19th position of the loaded strand is under no selective constraints. For example, in 21-nt phasiRNAs, the 5' position was predominantly C (40.1%) at the 1st position, and the most prevalent 19th nucleotide was G (35.7%) (Figure 4A, upper chart). This is consistent with a co-bias for the paired positions in the duplex, yielding duplexes with 5' C at each end (Figure 3A). We can infer that 21-nt phasiRNAs may have no strand specificity and either strand is likely to be loaded into the AGO protein as long as there is a 5' C. Similarly, among the 21-nt phasiRNAs, the 19A and to a lesser extent 19U classes were underrepresented (Figure 4A, lower), corresponding to bottom-strand 1U and 1A phasiRNAs in a

duplex; since 1U phasiRNAs were common among the sequenced phasiRNAs (Figure 4A, upper), we could infer a bias against 1U phasiRNAs in the complement to phasiRNAs abundant in our libraries.

To assess positional covariance, we analyzed 21- and 24-nt phasiRNAs versus P4-siRNAs, comparing the 5' nucleotide to the position complementary to the bottom-strand 5' position (19 in 21-nt siRNAs, and 22 in 24-nt siRNAs). We used these results to make inferences (see the discussion section) about strand specificity in the biogenesis of plant siRNAs. First, we compared the nucleotide composition at the 19th position of 21-nt phasiRNAs for a given 1st nucleotide and we performed the same analysis for the 1st position composition with the 19th position fixed (Figure 4B). The 1U phasiRNAs (i.e. 5' U) had an almost uniform distribution of nucleotides at the 19th position, which was striking relative to the 1C, 1A, and 1G phasiRNAs, which were depleted for 19A phasiRNAs (and 19U, to a lesser extent). Another noticeable bias was for 1C phasiRNAs, which were predominantly 19C or 19G, yielding a phasiRNA duplex of either 1C/1G or 1C/1C (top strand/bottom strand). 19G was prevalent for 1A, 1U, and 1G phasiRNAs, which in each case would yield a 1C bottom-strand phasiRNA. Next, we analyzed the 5' nucleotide composition for 21-nt phasiRNAs after fixing the 19th position (Figure 4B, lower panel). Among 19G phasiRNAs (the predominant group based on Figure 4A), 1C was most common, corresponding to a 1C/1C duplex. For 19C phasiRNAs (1G on the complement), a strong bias of 1C was observed; since 1C 21-nt phasiRNAs are most commonly loaded to MEL1 (Komiya et al., 2014), this was perhaps an indication of strand specificity (i.e. 1C/1G duplexes, so only the 1C strand loaded). Therefore, among 21-nt phasiRNAs, there is a co-bias of the 1st and 19th positions, perhaps reflective of strand specificity in AGO loading.

Next, we performed similar analyses for 24-nt phasiRNAs, focused on the 1st and 22nd positions (Figure 3B). The 1st position was less biased than 21-nt phasiRNAs, although 1G was also underrepresented (Figure 4C, upper); at the 22nd position, there was less bias than for the 19th position of the 21-mers (Figure 4C, lower), with an increase of A representation, particularly relative to other nucleotide positions (Figure 2C, left). 22A corresponds to 1U in the complement, and since 1U 24-nt phasiRNAs were common in our dataset (Figure 4C, upper), both phasiRNAs in such a duplex are favored in our data, consistent with a lack of strand specificity. Lower levels

of 1st/22nd position covariation were observed in 24-nt than 21-nt phasiRNAs (Figure 4D), and there was an overall A-U enrichment (Figure 2C), demonstrating more relaxed sequence constraints.

For comparison to the 24-nt phasiRNAs, we measured the position-specific nucleotide biases for P4-siRNAs. Their precursors have been described (Figure 3D; summarized from Blevins *et al.*, 2015; Zhai *et al.*, 2015a), although the nature of RDR2-derived bottom strands is as-yet incompletely understood (i.e. how they initiate and terminate relative to the ends of the P4 precursor). Unlike phasiRNAs, however, there is no expectation of P4-siRNA "duplexes" whereby either strand could be loaded, and data from (Zhai et al., 2015a) indicate that the P4 strand is preferably loaded over the RDR2 strand (Figure 3C). Apart from the strong overall 1A bias mentioned above, no notable co-variation biases were observed (Figure 4E,F); i.e. the proportional representation in the 22nd position was essentially invariant, regardless of the 1st position nucleotide, G>C>A>U, consistent with a strong overall bias to the GGGGC motif in the 3' end (Figure 2C).

Combining the compositional analyses described above, we applied these same approaches to an unusual group of siRNAs, a set of 22-nt, putative heterochromatic siRNAs that are RDR2-independent, thus far found only in maize (Nobuta et al., 2008). We were interested to analyze these "22-nt hc-siRNAs" because they are poorly characterized and their relationship to P4-siRNAs is not known (see Methods for extracting 22-nt siRNAs). The most significant difference between 22-nt hc-siRNAs and 24-nt P4-siRNAs was at 5' end positions 1, 3 and 4 (Supplemental Figure 3A,B), but the level of A in 22-nt hc-siRNAs was significantly lower from position 12 to the 3' end, compared to the 24-nt P4-siRNAs. There were apparent 3' differences as well, but this was from the comparison performed by counting nucleotides from the 5' end. We reassessed differences by aligning the 3' ends and measuring positions starting from the 3' end (i.e. comparing up to five positions at the 3' end minus N nucleotides), in case AGO binding occurs in some cases from the 3' end. Measured this way, we observed only one 3' difference, at the 3' end – 1 position, at which the G-U composition varied significantly (Supplemental Figure 3C). We next looked at covariation between the 20th and 1st nucleotides in the 22-nt hc-siRNAs; as with P4-siRNAs, the 20th nucleotide representation was more or less the same for all 5' nucleotides, and even for the

major class of 5' U siRNAs, 20th position G or C nucleotides were equally represented (Supplemental Figure 3D). This lack of bias would yield many bottom strand 5' G sRNAs which are disfavored, consistent with strand specificity for the 22-nt hc-siRNAs (Supplemental Figure 3E). Thus, these RDR2-independent 22-nt siRNAs may be produced by the activity of other RDRs such as RDR1 or RDR6; although the RNA polymerase generating their primary strand precursor remains to be determined, the 5' difference of 22-nt hc-siRNAs compared to P4-siRNAs suggests an alternative production pathway and/or function.

The results of analysis of the nucleotide and co-variation biases across different classes of siRNAs at the 5' and 3'-proximal ends are consistent with evidence of strand specificity for both 21- and 24-nt phasiRNA duplexes. There is stronger support for strand selection of 21-nt reproductive phasiRNAs, perhaps reflective of selection by the AGO protein of one strand over the other.

Predicted targets of reproductive phasiRNAs as a means to infer function

As little is known about the targets and the functions of the reproductive phasiRNAs, we attempted to predict targets for the 500 most abundant premeiotic (21-nt) and meiotic (24-nt) phasiRNAs in rice. Using standard criteria (i.e. modeled on known miRNA-target interactions), prior reports have failed to find targets of reproductive phasiRNAs, while reporting few details of these analyses due to the negative result (Zhai et al., 2015b; Song et al., 2012). To gain insights about phasiRNA targeting, we conducted a comparative analysis, measuring class-by-class how predicted targets of these abundant phasiRNAs compare to those of other known sRNAs, such as miRNAs and P4-siRNAs.

<u>21-nt phasiRNAs</u>

First, we compared in rice the distribution of predicted target scores (TS) of 21-nt phasiRNAs with a selected set of known, conserved miRNAs (Figure 5A). We selected plant miRNAs with numbers lower than miR1000 (i.e. osa-miR162) (n=288), as these are generally abundant, conserved, and better characterized than any more recently-described miRNAs. For each class, miRNAs versus 21-nt phasiRNAs, targets were predicted using sPARTA (Kakrana et al., 2014). We retained two sets of results, either all targets or only the "best" targets (those with a lowest target penalty score, meaning a high degree of complementarity). Each sRNA would also have at least one perfect

match in the genome, a target score of 0, potentially the result of targeting in *cis*. For 21-nt phasiRNAs, the TS distribution showed a peak in the number of best targets at 3.5 (Figure 5A, left), compared to ~1 for miRNAs (Figure 5A, right). The relative paucity of TS matches in the range of 0.5 to 1.5 for 21-nt phasiRNAs was striking, particularly since many miRNAs have predicted targets in this range. We inferred based on this pattern of sequence complementarity that 21-nt phasiRNAs, unlike miRNAs, either may function largely in *cis* via perfect matches or have been selected to avoid closely-matched targets.

To dissect these predicted sRNA-target interactions in rice, we recorded position-specific matches for both 21-nt phasiRNA-target interactions and 21-nt miRNA-target interactions (Figure 5B). This represented the putative binding pattern as a percentage of each position of predicted matches, gaps, wobbles, and mismatches. We selected only predicted targets (for both phasiRNAs and miRNAs) with a TS between 0.5 and 3.5, omitting self-targeting interactions. Overall, consistent with higher scores, 21-nt phasiRNAs showed lower match rates across all positions than miRNAs (Figure 5B); a few substantial position-specific differences were observed, including higher match rates for phasiRNAs at the 1st and 21st positions, and a higher (yet unexplainable) rate of gaps at the 15th position (Figure 5B). We concluded that unless 21-nt reproductive phasiRNAs target primarily in *cis,* they must have lower levels of complementarity to their targets than miRNAs.

<u>24-nt phasiRNAs</u>

Next, we extended our analysis to attempt to find the targets of the reproductive 24-nt phasiRNAs, again focusing on rice. We performed similar analyses as above and compared the TS distribution of 24-nt phasiRNAs (Figure 5C, left) with the top 500 most abundant 24-nt P4-siRNAs (Figure 5C, right). For the 24-mers, we omitted the higher penalty for a mismatch at the 10th and 11th positions in the target alignment; that penalty is relevant for 21/22-nt sRNAs that direct cleavage at those positions, whereas pairing requirements for individual 24-nt siRNAs have not been described or tested. For the 24-nt phasiRNAs, we observed a peak in the number of best targets at 4.5 (Figure 5C, left); while score of 4.5 to 5 was also the peak for P4-siRNAs (excluding perfect, or 'cis' matches at 0), P4-siRNAs had a much more even distribution of scores. There was a striking gap in the distribution of target scores from 0 to ~2 for the 24-nt phasiRNAs, indicating that these lack highly homologous *trans* targets (Figure 5C, left). In other words, the 24-nt

phasiRNAs are largely quite distinct from most other genome sequences, relative to P4-siRNAs, which find many highly homologous potential target sites.

Again, as for the 21-nt phasiRNAs, we predicted and recorded position-specific matches for both 24-nt phasiRNA-target interactions and P4-siRNA-target interactions (Figure 5D). This represented the putative binding pattern as a percentage of each position of predicted matches, gaps, wobbles, and mismatches. In this case, given the different score distribution relative to 21-mers, we selected only predicted targets (for both phasiRNAs and P4-siRNAs) with a TS between 0.5 and 5, omitting self-targeting interactions. Overall, consistent with higher TS scores, 24-nt phasiRNAs showed much lower match rates across all positions than P4-siRNAs (Figure 5D left versus right), i.e. an average of 15 to 20% mismatches compared to fewer than 15% mismatches for 24-nt P4-siRNAs.

Classes of predicted reproductive phasiRNA targets

As a final step in analyzing the possible targets of reproductive phasiRNAs in rice, we classified the predicted target loci. This analysis used all predicted targets described in the sections above, including both *cis* and *trans* targets. In rice, the top 500 21-nt phasiRNAs were predicted to target 7766 loci (Supplemental Data Set 2). These putative targets included 1400 (18.02 percent) loci classified by RepeatMasker as related to the transposable elements (TEs). The top 500 24-nt phasiRNAs were predicted to target 5631 loci, of which 836 (14.84 percent) are related to TEs (Supplemental Data Set 3). To assess whether these predicted matches to TEs represent an enrichment or depletion compared to random chance, we randomly selected 7800 and 5600 genes from the 35,000+ annotated genes in rice; among these, ~30 to 31% are TE-like. Therefore, the predicted targets of reproductive phasiRNAs are relatively depleted for TE-like targets. Overall, our more detailed results are consistent with earlier statements that classes of potential targets are not evident for reproductive phasiRNAs, and thus the characterization of their functions will require molecular and biochemical investigation.

Structural feature comparison illustrates that phasiRNAs are less stable and structured

Since the structural features are an integral part of the classification pipeline that we devised, we revisited the computationally-inferred stability of the reproductive phasiRNAs. These calculations

373 assessed their MFE and the GC content. We compared the MFE and the GC content of the
374 reproductive phasiRNAs, miRNAs, and P4-siRNAs in rice and maize (Figure 6). In a comparison
375 of the MFE between the two sequences, that with the lower negative value is more stable and
376 structured. In our analysis, P4-siRNAs and miRNAs (Figure 6A) had lower negative values
377 (comparing the mean of the distributions); we concluded that they are more stable or structured
378 than the reproductive phasiRNAs. A lower GC content was also observed for the reproductive
379 phasiRNAs compared to P4-siRNAs and miRNAs (Figure 6B). We also observed that 24-nt
380 phasiRNAs have a lower GC content compared to the 21-nt phasiRNAs (47.64% versus 42.36%,
381 mean values), consistent with lower stability and suggesting that their potential targets may differ,
382 as a class, compared to the 21-nt phasiRNAs. We further assessed the statistical correlation
383 between the GC content and the MFE in the reproductive phasiRNAs, miRNAs, and P4-siRNAs
384 by computing the Pearson correlation coefficient ('r') (Figure 6C). There was a moderate negative
385 correlation (r = -0.56) between the MFE and the GC content, consistent with prior reports (Chan
386 et al., 2009). Thus, we concluded that reproductive phasiRNAs are distinct from miRNAs and P4-
387 siRNAs in terms their computationally-predicted stability and structure characteristics.
388

389 **DISCUSSION**

390 Our machine learning-based workflow focused on sequenced-based and structural features of plant
391 sRNAs, with an emphasis on the poorly characterized set of reproductive phasiRNAs. We
392 demonstrate that this approach can successfully classify reproductive phasiRNAs relative to other
393 endogenous plant sRNAs, with high values for ACC, SE, SP, PPV, and AUC. Feature selection
394 demonstrated the importance of the 5'- and 3'- ends, k-mer features, GC content, and structural
395 features including the MFE. We observed characteristics that may reflect specificity in AGO
396 loading of reproductive phasiRNAs, the key to the function of all sRNAs. Examination of
397 spatiotemporal expression data for AGOs in rice and maize shows a high correlation between
398 peaks of abundance of reproductive phasiRNAs and *AGO* genes, suggesting that there might be a
399 functional connection. From rice and maize data, this includes OsAGO1d, ZmAGO18b,
400 OsAGO18, OsAGO2b (Zhai et al., 2015b; Fei et al., 2016), and OsAGO5c (MEL1) which loads
401 21-nt phasiRNAs in rice. In Arabidopsis, AGO3, close to OsAGO2b (Zhang et al., 2015), recruits
402 24-nt sRNAs with 5'A and effects epigenetic silencing, consistent with the hypothesis that 5'A
403 24-nt phasiRNAs might be loaded into AGO2b in grasses. Moreover, ZmAGO18b, a grass specific

404 AGO, binds both 21-nt phasiRNAs with 5′U and 24-nt phasiRNAs with 5′A to function in
405 inflorescence meristem and tassel development (Sun et al., 2017). Our classification data lay the
406 groundwork for better definition of AGO-phasiRNA interactions.
407
408 One unique aspect of working with the reproductive phasiRNAs is that their production from long,
409 double-stranded RNA precursors from hundreds or thousands of loci yields a rich dataset for which
410 comparable analyses of tasiRNAs or miRNAs are not possible due to their more limited
411 representation. This allowed the large-scale assessment of biases in representation in the libraries,
412 from which we observed significant biases in the representation of specific nucleotides at the 1st
413 and 19th positions among the 21-mers. One possible interpretation of these biases is a model of
414 competition for loading between the two strands of a duplex, whereby one strand is preferentially
415 loaded over the other, typically understood to be driven by the 5' nucleotide (Schwarz et al., 2003),
416 which is a preferred C in the case of 21-nt reproductive phasiRNAs (Komiya et al., 2014). Yet, 1U
417 phasiRNAs are quite abundant, begging the question of whether these are competing with 1C
418 phasiRNAs for loading into MEL1; among sequenced MEL1-associated phasiRNAs, 1U
419 phasiRNAs were less than 10% of the total (Komiya et al., 2014). Perhaps the higher proportion
420 in the sequenced phasiRNAs reflects (1) stability in the absence of loading, or (2) perhaps 1U
421 phasiRNAs are loaded into a different AGO than the 1C phasiRNAs – maybe AGO1, known to
422 have an affinity for 1U 21-nt sRNAs (Zhao et al., 2016). Assuming the latter, for the sake of
423 argument, the difference in the 19th position for a given 1st position nucleotide for the 21-nt
424 reproductive phasiRNAs could be explained by AGO affinity: 1U phasiRNAs may be loaded as
425 well or better than 1C phasiRNAs, but into this different AGO.
426
427 Based on the observation of abundant 1U and 1C 21-nt phasiRNAs, we hypothesized an AGO
428 competition model (Figure 7). We inferred/hypothesized this because of the data in Figure 4B
429 (upper panel) that the sequenced 1V (V = A or C or G, using the IUPAC code) phasiRNAs are
430 depleted for 19A phasiRNAs, which would be 1U on the bottom strand; perhaps this is because in
431 a duplex with a 1U phasiRNA, the 1U phasiRNA is loaded. But sequenced 1U phasiRNAs showed
432 no bias in the 19th position, because they are preferred over the opposite strand, and thus are the
433 "winners" in the competition (Figure 7A). In contrast, the 1R/19G (R = A or G) phasiRNAs are
434 paired with 1C phasiRNAs, which is AGO loaded (Figure 7B). The 1V/19C phasiRNAs are

abundant because these are paired with 1G phasiRNAs, which are not AGO loaded and thus are always "losers" in the competition with their duplex pairs. The 1C phasiRNAs are an interesting case because based on frequency, 1C/19C > 1C/19G > 1C/19W (W = A or U) (Figure 4B, upper panel). The 1C/19G phasiRNAs are paired with 1C phasiRNAs, which compete well, and thus either strand may be loaded and stabilized (Figure 7C). The 1C/19U phasiRNAs are less frequent because they are paired with 1A phasiRNAs that are not particularly stabilized or loaded. In other cases (Figure 7D), the frequency of 1D/19G phasiRNA is higher than 1D/19C (D = A or G or U) (Figure 4B, upper panel); one interpretation of the high frequency of 1D/19G is that since the 1D/19G phasiRNAs are paired with a 1C/19H phasiRNA (H = A or C or T), thus 1C/19H phasiRNA is preferentially loaded and stabilized. Thus, phasiRNAs from a 1D/19G duplex are more abundant than those from a 1D/19C duplex because in the latter, the 1G/19H phasiRNA on the bottom strand is likely not loaded or stabilized. With as rich a dataset as reproductive phasiRNAs provide, we can start to resolve the sequence-based characteristics that influence representation in sequencing data, and infer the mechanistic basis for these differences. For example, we identified novel position-specific biases, like the 14th position in the 21-nt phasiRNAs (Figure 2A, left, and Supplemental Figure 2B, left). These internal positions may be important for AGO loading, or phasing function/targeting, and thus future functional or structural studies should investigate these in greater detail.

METHODS

Classification via machine learning

We use the Random Forest ("RF")(Breiman, 2001) classification method, which is based on building an ensemble of decision trees. This method has proven effective for addressing a variety of classification problems in bioinformatics (Yang et al., 2010; Lertampaiporn et al., 2014). We employed the WEKA implementation of RF (Frank et al., 2016) to build the model for distinguishing phasiRNAs (to which we refer as the *positive* set) from non-phasiRNAs (the *negative* set). As we study two sets of reproductive phasiRNAs, characterized by two distinct lengths, namely 21- and 24-nt, for each set we have trained two distinct classifiers, one for each length. When training each of these classifiers, we have varied the composition of the negative

sets of non-phasiRNAs to which the phasiRNAs were compared (see Dataset used for cross validation study, Methods).

To train and test the classifiers we developed, we have used the commonly used stratified five-fold cross-validation (CV) framework (Kohavi, 1995). Under this framework, the dataset is partitioned into five subsets, where each subset has the same ratio of positive instances to negative instances as the whole dataset. Once the data is partitioned, five iterations of training and testing are performed, where in each iteration four parts of the data (80%) are used for training and the remaining part (20%) is used for testing. To ensure stability and reproducibility of the results, the whole five-fold CV experiment was repeated five times, each using a different five-way split (partition) of the dataset.

Performance evaluation

To assess classification performance we use the standard measures of *accuracy* (ACC), *specificity* (SP), *sensitivity* (SE), *positive predictive value* (PPV), and *area under the receiver operating characteristic curve* (AUC), whose formulae and descriptions are as follows:

- Sensitivity $SE = \frac{|TP|}{|TP|+|FN|}$;
- Specificity $SP = \frac{|TN|}{|TN|+|FP|}$;
- Accuracy $ACC = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$;
- Positive Predictive Value $PPV = \frac{|TP|}{|TP|+|FP|}$;

  where *True Positives* (TP) denotes the set of correctly classified phasiRNAs, *True Negatives* (TN) denotes the set of correctly classified non-phasiRNAs, *False Positives* (FP) denotes the set of non-phasiRNA sequences that were classified as phasiRNAs, and *False Negatives* (FN) denotes the set of phasiRNA sequences that were not classified as such by our classifier. The number of items in the sets TP, TN, FP, and FN is denoted by |TP|, |TN|, |FP|, and |FN|, respectively.
- The Area Under the ROC Curve (AUC) is an effective and joint measure of sensitivity and specificity, which is calculated by the Receiver Operating Characteristic curve (ROC). AUC determines the relative performance of classifiers for correctly classifying

phasiRNAs and non-phasiRNAs. Values of AUC is between 0 (worst performance) and 1 (best performance). ROC illustrates the true positive rate (sensitivity) against the false-positive rate (1 - specificity).

Development of a machine learning classifier for plant small RNAs

The classification pipeline we developed takes as input a set of plant small RNA sequences to assess for each sequence whether it has attributes or not of a reproductive phasiRNA, based on a training/test set, returning a "yes" or "no" response. Thus, for this decision, feature characterization is crucial. The pipeline used several sequence- and structural-based features. One known feature of reproductive phasiRNAs is a 5′-terminal cytosine, described for 21-nt phasiRNAs bound by MEL1, a rice Argonaute (Komiya et al., 2014). Another known characteristic of both 21- and 24-nt reproductive phasiRNAs is their origin from unique or low copy regions in the genome (Johnson et al., 2009; Zhai et al., 2015b). Beyond these features, little was known about their sequence composition, true even for other classes of plant small RNAs.

Thus, to build a classifier, we utilized an alignment-free approach based on k-mers. These k-mer motifs (see Methods), together with the GC content of the small RNA, comprised the sequence-based features of the classifier. The other major component of the classifier was structural features, calculated for each sequence using RNAfold (Hofacker, 2003) to assess the RNA secondary structure and minimum free energy (MFE) of folding (for intramolecular interactions of each small RNA). MFE is considered a vital feature to assess the stability of the secondary structure of non-coding RNAs; it has been widely used, for example in several piRNA and miRNA prediction tools (Batuwita and Palade, 2009; Meng et al., 2014; Seyeddokht et al., 2016). These two sets of attributes for each sequence comprised 1372 features, most of which were short k-mers or words that we could use to classify plant small RNAs. Before each classification, feature selection of the top 250 most informative features (of the 1372) was performed as a step to better understand which features play key roles in classifying phasiRNAs; this allowed us to reduce the feature dimensionality comprising classification without compromising or negatively impacting the classifier's performance (see Methods). We have also experimented with different number of trees and of features used. Consequently, to estimate the performance of the classifier, RF was applied

using 100 trees, five out of 250 features assessed (five randomly sampled features selected as candidates at each split) at each split, and five complete runs of the 5-fold CV.

Dataset used for cross validation study

We used several representative positive and negative datasets to perform a five-fold, cross validation. The use of these sets in the cross-validation steps allowed us to compare the performance of the classifiers with these different datasets. The positive sets consisted of non-redundant sequences of reproductive phasiRNAs, including sets of phasiRNAs whose length is either 21- or 24-nt. The sets characterized by either sequence length comprise phasiRNAs from previously-identified *PHAS* loci published for rice and maize. Specifically for rice, we utilized small RNA libraries published by Fan et al., 2016 and analyzed these libraries using our previously published python-based FASTQ processing script to produce the 'tag count' formatted files (Patel et al., 2015). Next, we identified 2024 21-nt *PHAS* loci and 59 24-nt *PHAS* loci from rice using a tool called - PHASIS (Kakrana et al., 2017), and we used 463 21-nt *PHAS* loci and 176 24-nt *PHAS* loci previously described for maize (Zhai et al., 2015b) to obtain 21- and 24-nt phasiRNAs. From these loci, we identified non-redundant sets of 20240 (rice) and 9260 (maize) premeiotic 21-nt phasiRNAs, and 2950 (rice) and 8800 (maize) meiotic 24-nt phasiRNAs; these comprise the small RNAs from these *PHAS* loci that were represented in the sequencing data from the reproductive tissues. To ensure high quality of the data, we used in our cross validation studies for training/testing the classifiers only the top 1000 most-abundant phasiRNA sequences from each species represented within the sets of sequences of either length. The negative training sets were built by gathering several types of non-redundant sRNA sequences. For simplicity, we refer to all of them as *non-phasiRNAs*, as each sequences of each type were selected based on the criterion of not containing reproductive phasiRNAs

- Mature miRNA sequences were downloaded from the miRBase (Kozomara and Griffiths-Jones, 2014), version 21. We combined miRNAs from rice (n=553) and maize (n=203), totaling 756 miRNAs. We trimmed miRNAs longer than 21-nt to 21-nt, trimming from the 3'- end to obviate any length differences in comparison to the 21-nt phasiRNAs. Thus, we used this set (n=756) only while classifying 21-nt phasiRNAs.
- P4-siRNAs sequences were selected from either rice or maize small RNA libraries (Supplemental Data Set 1). These were selected using a length filter of 24 nucleotides,

and the criterion that the sequence should match the genome more than 10 times and overlaps with repetitive regions, consistent with repetitive origins. As an additional filter to avoid reproductive phasiRNAs, the P4-siRNAs were from vegetative tissues in which reproductive phasiRNAs are absent (Johnson et al., 2009). In the libraries we analyzed, using these criteria, we identified 14008 non-redundant P4-siRNAs from maize and 12530 from rice as our data set. Of these, we selected the top 1000 most abundant sequences from rice and separately from maize. As with the miRNAs, for comparisons to 21-nt phasiRNAs, we trimmed P4-siRNAs to 21-nt from the 3'- end. For comparisons to 24-nt phasiRNAs, no trimming was required since the P4-siRNAs were preselected as 24-nt small RNAs.

- Small RNAs from tRNA and rRNA loci in either rice or maize were extracted from the vegetative libraries described above. The tRNA loci in both genomes were identified using tRNAscan-SE (Lowe and Eddy, 1997) with default parameters. The rRNA loci were identified using RepeatMasker (Smit et al., 2013) with a default cut-off score of 225. The small RNAs derived from these loci were combined for the two genomes (since tRNAs and rRNAs are highly conserved between rice and maize), and we randomly selected 500 tRNAs and 500 rRNAs.

Finally, to assess the predictive performance of the algorithm on new sequences, we selected the remaining 8260 and 19240 21-nt phasiRNAs and 7800 and 1950 24-nt phasiRNAs from maize and rice, distinct from those that comprised the positive set used for cross validation study.

Features included in the machine learning algorithm, and their selection

The machine learning classification pipeline integrated both sequence-based and structural features, as follows:

Sequence Based Features:

1. The frequency of k-mer motifs: We computed normalized frequencies of all 1 to 5 nt k-mer patterns (see patterns below) using a sliding-window of length k (k = 1, 2, 3, 4, 5). A particular string of length k could slide along the length of the sequence, denoted as L, by a step of 1 nt. If the string in the sequence matched with some pattern *i* within the window, the count of that pattern in the sequence, denoted as C_i, was increased by one, using the following formula to calculate frequencies:

586 $F_i = C_i / S_k$, for i = 1 to 1364 , where $S_k = L - k+1$, and S_k is the total number of times the
587 sliding-window of length k could slide along the sequence.
588 The frequencies of all 1 to 5 nt k-mer patterns (4+16+64+256+1024=1364 patterns) were
589 as follows:
590 1. 4 one-mer patterns (A,C,G,U):
591 2. 16 two-mer patterns (AA,AC,AG,AU,…,UU)
592 3. 64 three-mer patterns (AAA,AAC,AAG,AAU,…,UUU)
593 4. 256 four-mer patterns (AAAA,…,UUUU)
594 5. 1024 five-mer patterns (AAAAA,…,UUUUU)
595
596 Thus, phasiRNAs and non-phasiRNAs were characterized as a vector consisting normalized
597 frequencies of 1364 k-mer motifs. We used the repDNA package (Liu et al., 2015) in Python to
598 generate these features.
599 2. GC_content = (|C|+|G|) / L*100, where |C| and |G| denote the number of nucleotides C or
600 G in the sequence, respectively.
601
602 **Structural Features:**
603 We used the RNAfold program (Hofacker, 2003) with default parameters to obtain structural
604 features based on the RNA secondary structures using minimum free energy (MFE) of folding.
605 Since MFE is an important index for measuring the stability, we computed several MFE-related
606 features, such as dG and MFE indices (MFEI1, MFEI2, MFEI3, and MFEI4). MFEI1 is a ratio
607 between the MFE and the GC content; MFEI2 is a ratio between the MFE and the number of
608 stems; MFEI3 is a ratio between the MFE and the number of loops; and MFEI4 denotes the ratio
609 between the MFE and the total number of base pairs in the secondary structure. Table 3 lists all
610 the structural features computed in this study. Thus, for each sequence in the positive and the
611 negative sets, we calculated 1372 features (1365 sequence-based + seven structure-based).
612
613 To eliminate redundancy from among the 1372 features, the attribute selection tool of WEKA
614 was used prior to performing classification, separately for each positive vs negative set
615 comparison. Information gain (also known as entropy) was employed to evaluate the value of
616 each feature, for either the phasiRNA or non-phasiRNA classes. Values of information gain (i.e.

the relative contribution) for each feature varied from zero (no information) to one (maximum information); features conveying more information thus had a higher information gain value. Features with lower values could be removed, with the aim to reduce the dimensionality of the feature set by using a minimum number of maximally informative features. Thus the features were ranked by information gain, and the top 250 were selected (Supplemental Figure 1A, B), as this subset included 95 to 97% of informative features (Supplemental Figure 1E), reducing dimensionality without compromising or negatively impacting the classifier's performance. As examples of our feature selection process, we listed the top 25 out of 250 selected features in Table 4.

Extraction of a set of maize 22-nt hc-siRNAs

22-nt, putative heterochromatic siRNAs that are RDR2-independent, thus far found only in maize (Nobuta et al., 2008) were selected maize small RNA libraries (Supplemental Data Set 1). These were selected using a length filter of 22 nucleotides, and the criterion that the sequence should overlap with highly repetitive regions (i.e. hits > 50 in the genome), consistent with repetitive origins.

Generation and computational analysis of sequencing data from maize

Four biological replicate libraries were constructed from the *rdr2/mop1* maize mutant (Alleman et al., 2006), using young ears (3 to 5 cm in length); the materials were generously donated by Drs. Mario Arteaga-Vazquez and Vicki Chandler. Small RNA library construction was as previously described (Lu et al., 2007). Small RNA libraries were processed using a custom python-based FASTQ pipeline (Patel et al., 2015). Briefly, the processing script removes adaptors and low-quality reads, and retain sequences with lengths between 18 and 31 nt. Then, these sequences were mapped to the maize genome (version AGPv2) using Bowtie (Langmead et al., 2009). Reads that perfectly matched (no mismatch) to the maize genome, excluding those matching to structural RNAs (tRNAs or rRNAs) were used for further study.

Accession numbers

The data utilized here were either obtained from public data available in Gene Expression Omnibus (GEO) database, or several additional libraries have been submitted to GEO, with accession numbers not yet available, to be added in a subsequent version (see Supplemental Data Set 1).

SUPPLEMENTAL DATA

The following materials are available in the online version of this article.

Supplemental Figure 1. Information gain (IG) was used to reduce feature dimensionality together with ROC curve to select top most 250 most informative features.

Supplemental Figure 2. Position-specific nucleotide biases in either the not-detected positions from reproductive *PHAS* loci phasiRNAs or in all detected phasiRNAs.

Supplemental Figure 3. 22-nt siRNAs from maize are distinct from P4-siRNAs, but have similar 3' end sequence composition.

Supplemental Data Set 1. sRNA libraries from maize and rice used in this study.

Supplemental Data Set 2. Predicted targets of 21-nt phasiRNAs in rice.

Supplemental Data Set 3. Predicted targets of 24-nt phasiRNAs in rice.

ACKNOWLEDGMENTS

We are grateful to members of the Meyers lab for useful discussions. We also acknowledge the contribution of maize *mop1 (rdr2)* mutant tissue from the lab of Vicki Chandler; and Stacey Simon for the construction of those libraries.

FUNDING

This research was supported by NSF IOS Plant Genome Research program award #1339229 (to B.C.M.), and a University of Delaware Graduate Fellow Award (to P.P.).

AUTHOR CONTRIBUTIONS

Experiments were designed by P.P., H.S., S.M., and B.C.M. P.P. implemented methods and conducted the analyses. A.K. contributed methods and algorithmic refinements. H.S. contributed

conceptual ideas. S.M. performed data generation. P.P. and B.C.M. wrote the paper with input from all authors; all authors read and approved the manuscript.

REFERENCES

Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J.E., White, J., Sikkink, K., and Chandler, V.L. (2006). An RNA-dependent RNA polymerase is required for paramutation in maize. Nature **442**: 295–298.

Axtell, M.J. (2013). Classification and comparison of small RNAs from plants. Annu. Rev. Plant Biol. **64**: 137–59.

Batuwita, R. and Palade, V. (2009). microPred: Effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics **25**: 989–995.

Blevins, T., Podicheti, R., Mishra, V., Marasco, M., Wang, J., Rusch, D., Tang, H., and Pikaard, C.S. (2015). Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding *de novo* DNA methylation in Arabidopsis. eLife **4**.

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

Brayet, J., Zehraoui, F., Jeanson-Leh, L., Israeli, D., and Tahi, F. (2014). Towards a piRNA prediction using multiple kernel fusion and support vector machine. Bioinformatics **30**: i364-70.

Breiman, L. (2001). Random forests. Mach. Learn. **45**: 5–32.

Carrington, J.C. and Ambros, V. (2003). Role of microRNAs in plant and animal development. Science **301**: 336–338.

Chan, C.Y., Carmack, C.S., Long, D.D., Maliyekkel, A., Shao, Y., Roninson, I.B., and Ding, Y. (2009). A structural interpretation of the effect of GC-content on efficiency of RNA interference. BMC Bioinformatics **10 Suppl 1**: S33.

Chatterjee, S., Fasler, M., Büssing, I., and Großhans, H. (2011). Target-mediated protection of endogenous microRNAs in *C. elegans*. Dev. Cell **20**: 388–396.

Fan, Y. et al. (2016). *PMS1T*, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. Proc. Natl. Acad. Sci. **113**: 15144–15149.

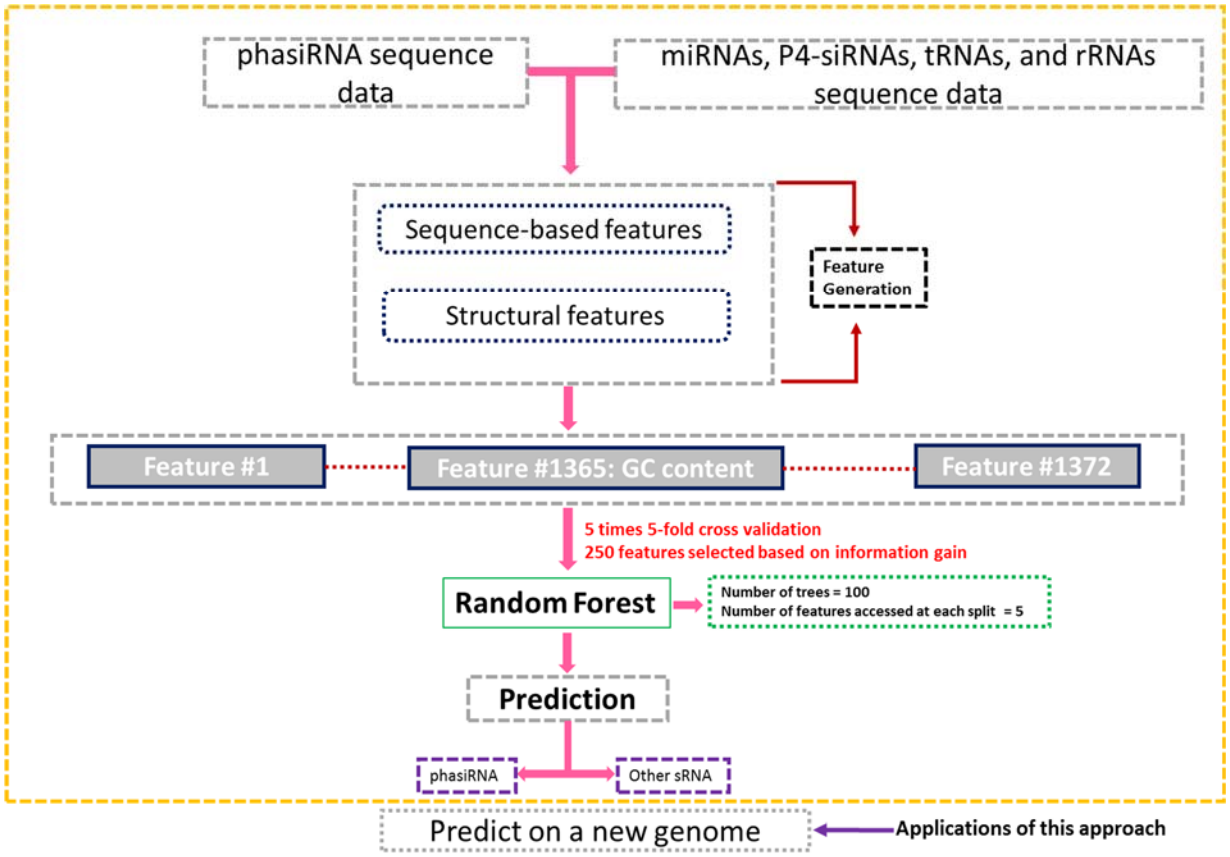
Fei, Q., Yang, L., Liang, W., Zhang, D., and Meyers, B.C. (2016). Dynamic changes of small RNAs in rice spikelet development reveal specialized reproductive phasiRNA pathways. J.

705 Exp. Bot. **67**: 6037–6049.
706 **Frank, E., Hall, M.A., and Witten, I.H.** (2016). The WEKA Workbench.
707 **Guttman, M. and Rinn, J.L.** (2012). Modular regulatory principles of large non-coding RNAs.
708 Nature **482**: 339–346.
709 **Hofacker, I.L.** (2003). Vienna RNA secondary structure server. Nucleic Acids Res. **31**: 3429–
710 3431.
711 **Johnson, C., Kasprzewska, A., Tennessen, K., Fernandes, J., Nan, G.L., Walbot, V.,**
712 **Sundaresan, V., Vance, V., and Bowman, L.H.** (2009). Clusters and superclusters of phased
713 small RNAs in the developing inflorescence of rice. Genome Res. **19**: 1429–1440.
714 **Kakrana, A., Hammond, R., Patel, P., Nakano, M., and Meyers, B.C.** (2014). sPARTA: a
715 parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets,
716 including new miRNA target-identification software. Nucleic Acids Res. **42**: e139.
717 **Kakrana, A., Li, P., Patel, P., Hammond, R., Anand, D., Mathioni, S., and Meyers, B.** (2017).
718 PHASIS: A computational suite for de novo discovery and characterization of phased,
719 siRNA-generating loci and their miRNA triggers. bioRxiv158832; doi:
720 https://doi.org/10.1101/158832
721 **Kohavi, R.** (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and
722 Model Selection. In the International Joint Conference on Artificial Intelligence (IJCAI), pp.
723 1–7.
724 **Komiya, R., Ohyanagi, H., Niihama, M., Watanabe, T., Nakano, M., Kurata, N., and**
725 **Nonomura, K.-I.** (2014). Rice germline-specific Argonaute MEL1 protein binds to
726 phasiRNAs generated from more than 700 lincRNAs. Plant J. **78**: 385–397.
727 **Kozomara, A. and Griffiths-Jones, S.** (2014). miRBase: annotating high confidence microRNAs
728 using deep sequencing data. Nucleic Acids Res. **42**: D68-73.
729 **Kung, J.T.Y., Colognori, D., and Lee, J.T.** (2013). Long Noncoding RNAs: Past, Present, and
730 Future. Genetics **193**: 651–669.
731 **Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L.** (2009). Bowtie: An ultrafast memory-
732 efficient short read aligner. [http://bowtie.cbcb.umd.edu/]. Genome Biol.: R25.
733 **Lertampaiporn, S., Thammarongtham, C., Nukoolkit, C., Kaewkamnerdpong, B., and**
734 **Ruengjitchatchawalya, M.** (2014). Identification of non-coding RNAs with a new
735 composite feature in the Hybrid Random Forest Ensemble algorithm. Nucleic Acids Res. **42**.

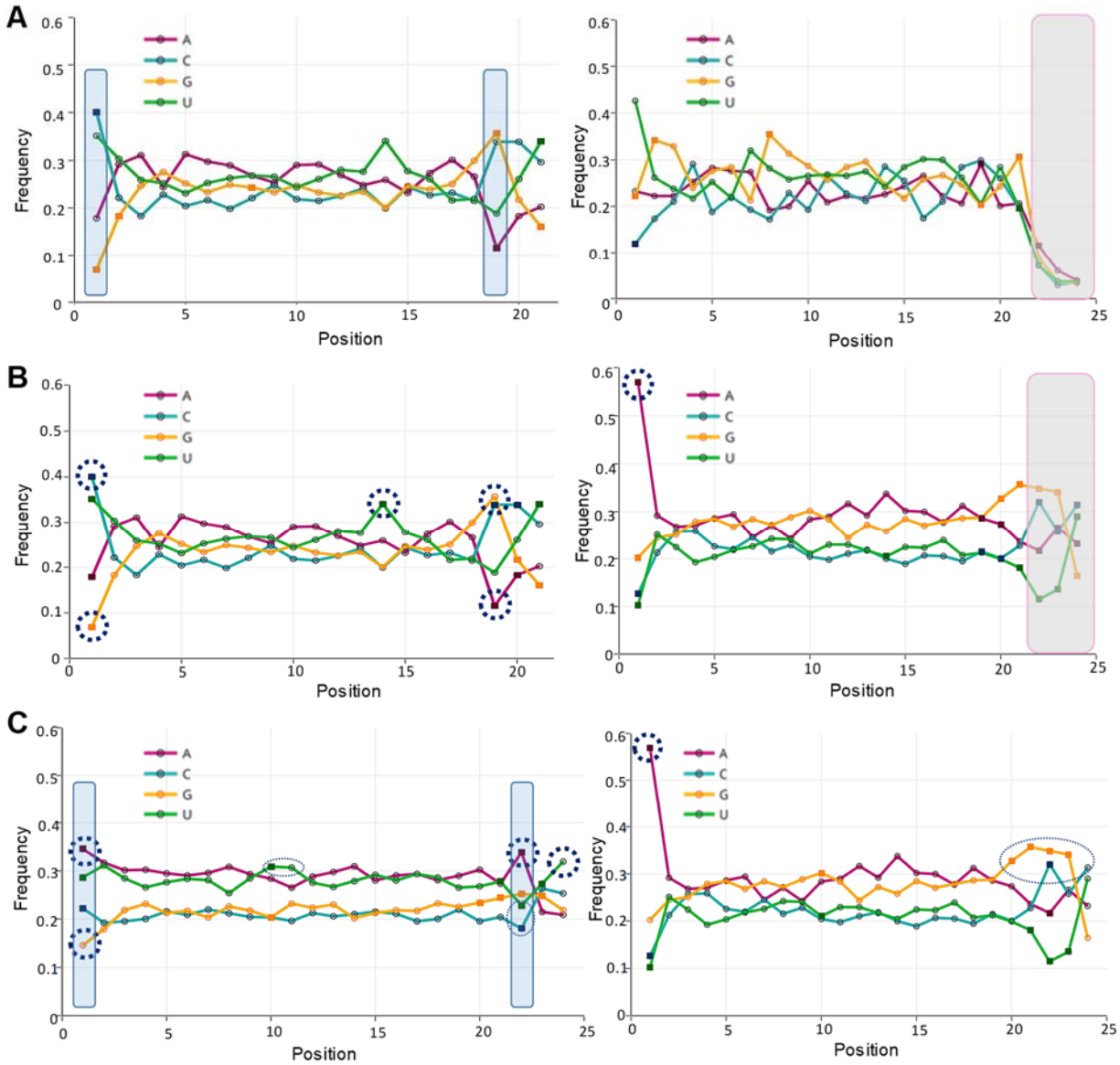
736 **Liu, B., Liu, F., Fang, L., Wang, X., and Chou, K.C.** (2015). RepDNA: A Python package to
737 generate various modes of feature vectors for DNA sequences by incorporating user-defined
738 physicochemical properties and sequence-order effects. Bioinformatics **31**: 1307–1309.
739 **Lowe, T.M. and Eddy, S.R.** (1997). tRNAscan-SE: a program for improved detection of
740 transferRNA genes in genomic sequence. Nucleic Acids Res. **25**: 955–964.
741 **Lu, C., Meyers, B.C., and Green, P.J.** (2007). Construction of small RNA cDNA libraries for
742 deep sequencing. Methods **43**: 110–117.
743 **Macrae, I.J., Zhou, K., Li, F., Repic, A., Brooks, A.N., Cande, W.Z., Adams, P.D., and**
744 **Doudna, J.A.** (2006). Structural basis for double-stranded RNA processing by Dicer. Science
745 **311**: 195–8.
746 **Matzke, M.A. and Mosher, R.A.** (2014). RNA-directed DNA methylation: an epigenetic
747 pathway of increasing complexity. Nat. Rev. Genet. **15**: 394–408.
748 **Meister, G.** (2013). Argonaute proteins: functional insights and emerging roles. Nat. Rev. Genet.
749 **14**: 447–59.
750 **Meng, J., Liu, D., Sun, C., and Luan, Y.** (2014). Prediction of plant pre-microRNAs and their
751 microRNAs in genome-scale sequences using structure-sequence features and support vector
752 machine. BMC Bioinformatics **15**: 423.
753 **Nobuta, K. et al.** (2008). Distinct size distribution of endogenous siRNAs in maize: Evidence
754 from deep sequencing in the *mop1-1* mutant. Proc. Natl. Acad. Sci. **105**: 14958–14963.
755 **Nonomura, K.-I., Morohoshi, A., Nakano, M., Eiguchi, M., Miyao, A., Hirochika, H., and**
756 **Kurata, N.** (2007). A germ cell-specific gene of the ARGONAUTE family is essential for
757 the progression of premeiotic mitosis and meiosis during sporogenesis in rice. Plant Cell **19**:
758 2583–2594.
759 **Patel, P., Ramachandruni, S.D., Kakrana, A., Nakano, M., and Meyers, B.C.** (2015).
760 miTRATA: A web-based tool for microRNA truncation and tailing analysis. Bioinformatics
761 **32**: 450–452.
762 **Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D.** (2003).
763 Asymmetry in the assembly of the RNAi enzyme complex. Cell **115**: 199–208.
764 **Seyeddokht, A., Aslaminejad, A.A., Masoudi-Nejad, A., Nassiri, M., Zahiri, J., and Sadeghi,**
765 **B.** (2016). Computational detection of piRNA in human using support vector machine.
766 Avicenna J. Med. Biotechnol. **8**: 36–41.

767 **Smit, A., Hubley, R., and Green, P.** (2013). RepeatMasker Open-4.0. 2013-2015 .
768 http://www.repeatmasker.org.
769 **Song, X. et al.** (2012). Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J.
770 **69**: 462–474.
771 **Sun, W., Xiang, X., Zhai, L., Zhang, D., Cao, Z., Liu, L., and Zhang, Z.** (2017). AGO18b
772 negatively regulates determinacy of spikelet meristems on the tassel central spike in maize.
773 J. Integr. Plant Biol. doi: 10.1111/jipb.12596
774 **Tisseur, M., Kwapisz, M., and Morillon, A.** (2011). Pervasive transcription - Lessons from yeast.
775 Biochimie **93**: 1889–1896.
776 **Yang, P., Hwa Yang, Y., B. Zhou, B., and Y. Zomaya, A.** (2010). A Review of Ensemble
777 Methods in Bioinformatics. Curr. Bioinform. **5**: 296–308.
778 **Zhai, J. et al.** (2015a). A One Precursor One siRNA Model for Pol IV-Dependent siRNA
779 Biogenesis. Cell **163**: 445–455.
780 **Zhai, J., Zhang, H., Arikit, S., Huang, K., Nan, G.-L., Walbot, V., and Meyers, B.C.** (2015b).
781 Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize
782 anthers. Proc. Natl. Acad. Sci. U. S. A. **112**: 3146–3151.
783 **Zhang, H., Xia, R., Meyers, B.C., and Walbot, V.** (2015). Evolution, functions, and mysteries
784 of plant ARGONAUTE proteins. Curr. Opin. Plant Biol. **27**: 84–90.
785 **Zhang, Y., Wang, X., and Kang, L.** (2011). A k-mer scheme to predict piRNAs and characterize
786 locust piRNAs. Bioinformatics **27**: 771–776.
787 **Zhao, J.-H., Fang, Y.-Y., Duan, C.-G., Fang, R.-X., Ding, S.-W., and Guo, H.-S.** (2016).
788 Genome-wide identification of endogenous RNA-directed DNA methylation loci associated
789 with abundant 21-nucleotide siRNAs in Arabidopsis. Sci. Rep. **6**: 36247.
790

791 **FIGURES**

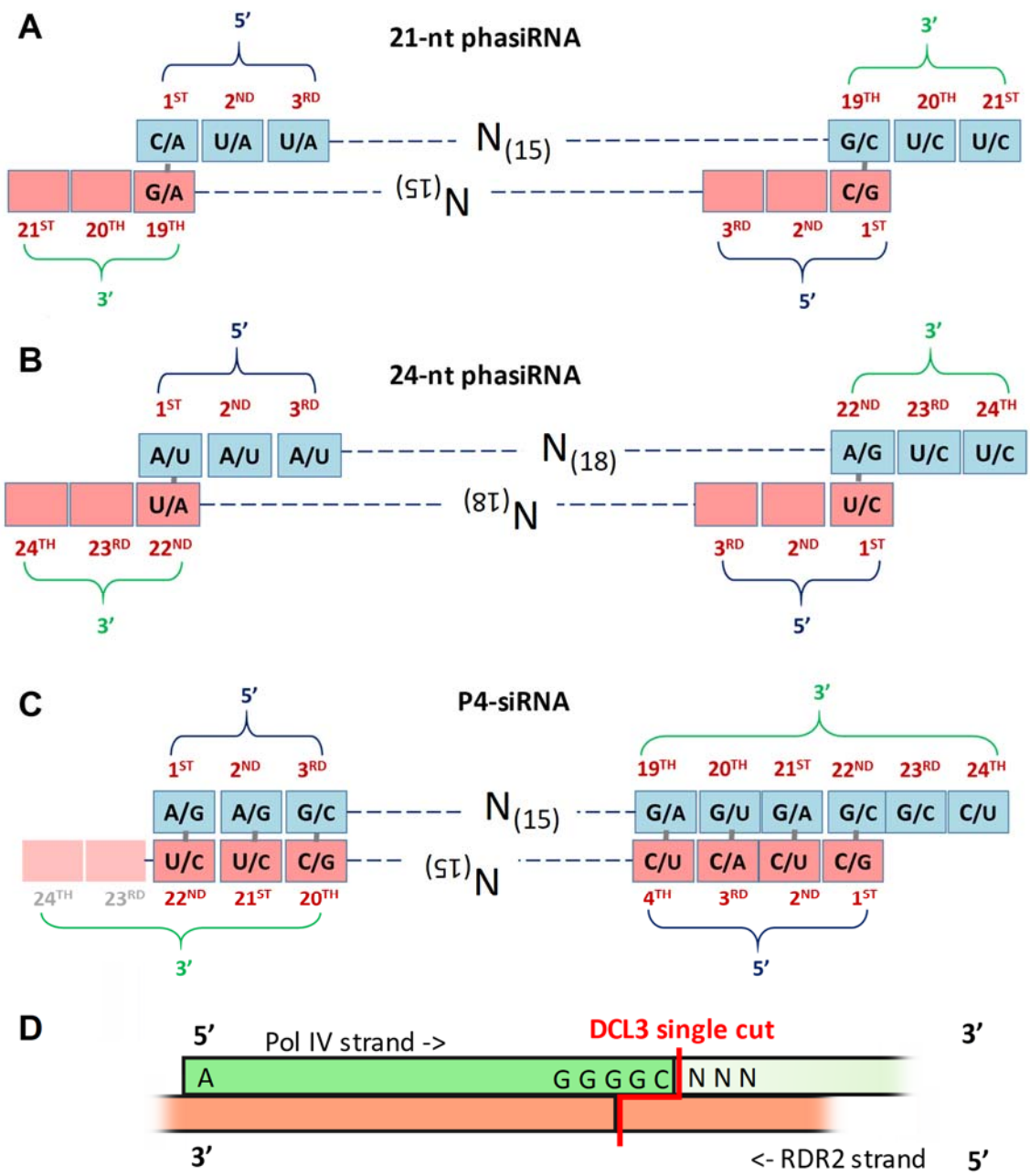


792 **Figure 1. General workflow of our pipeline.**



793 **Figure 2. Reproductive phasiRNAs have characteristic position-specific nucleotide biases.**
794 Single-nucleotide sequence profiles of position specific base usage comparing 21-nt phasiRNAs
795 (left) and either miRNAs (at right in panel A), or 24-nt P4-siRNAs (at right in panel B). For all
796 phasiRNA analyses in this figure, the top most abundant 1000 phasiRNAs from the rice and maize
797 data were combined; in panel A, 553 rice and 203 maize miRBase-annotated miRNAs were used
798 (see methods). The frequencies of each of the four bases (A, C, G, and U) at each position are
799 indicated as an open circle. Markers denoted as small square boxes represent positions at which a
800 statistically significant (p = 1e-5) base usage distinguishes phasiRNAs and either miRNAs (panel
801 A) or P4-siRNAs (panel B), determined by comparison of the data in the two plots. Dotted circles
802 highlight positions in the sequences selected for further discussion in the main text. The gray boxes

803 at right covering the 22nd, 23rd, and 24th positions to retain fair comparison with 21-nt phasiRNAs
804 and the longer sequences, including that those additional position could be disregarded. (C) Single-
805 nucleotide sequence profiles of position specific base usage comparing 24-nt phasiRNAs (at left)
806 and 24-nt P4-siRNAs (at right). In panels A and C, the blue boxes highlight positions that were
807 analyzed in greater detail in Figure 4 (positions #1 & 19 for 21-nt phasiRNAs, and positions #1 &
808 22 for 24-nt phasiRNAs).



809 **Figure 3. Nucleotide biases indicate one of the two siRNA precursor duplex small RNAs is**
810 **preferentially retained.**
811 Schematic duplex structures of different types of plant small RNAs; the 5'- and 3'- ends are
812 annotated and highlighted to emphasize the influence that a nucleotide bias on one strand has on
813 the other due to pairing. The first three and the last three nucleotide positions are indicated from
814 the 5'- and 3'-end positions, respectively, as the analyses focused on sequence composition
815 biases at these positions; red numbering indicates the base position within the small RNA.

816 Within each position, the top two most frequent nucleotides are indicated, with the first
817 representing the most common occurring nucleotide; the sequences analyzed are the same as
818 Figure 2. (A) Position-specific nucleotide biases for abundant 21-nt reproductive phasiRNAs in
819 rice and maize. (B) Position-specific nucleotide biases for 24-nt reproductive phasiRNAs from
820 rice and maize. (C) Position-specific nucleotide biases for P4-siRNAs from rice and maize; for
821 P4-siRNAs, the RDR2-derived bottom strand may terminate at the 22nd position, corresponding
822 to the 5' end of the 'top', Pol IV-derived strand, although this is as-yet poorly characterized
823 (indicated by lighter shading of the 23rd and 24th positions). (D) For comparison to panel C, prior
824 work by Zhai *et al*. (2015) and Blevins *et al*. (2015) described the P4R2 (Pol IV and RDR2-
825 derived) precursors of 24-nt P4-siRNAs as ~26 to 42 nt RNAs; mapped onto the green Pol IV
826 RNA are the biases observed here for P4-siRNAs. The 5' and 3' ends of the RDR2-derived
827 strands are blurred because these ends have not yet been characterized.

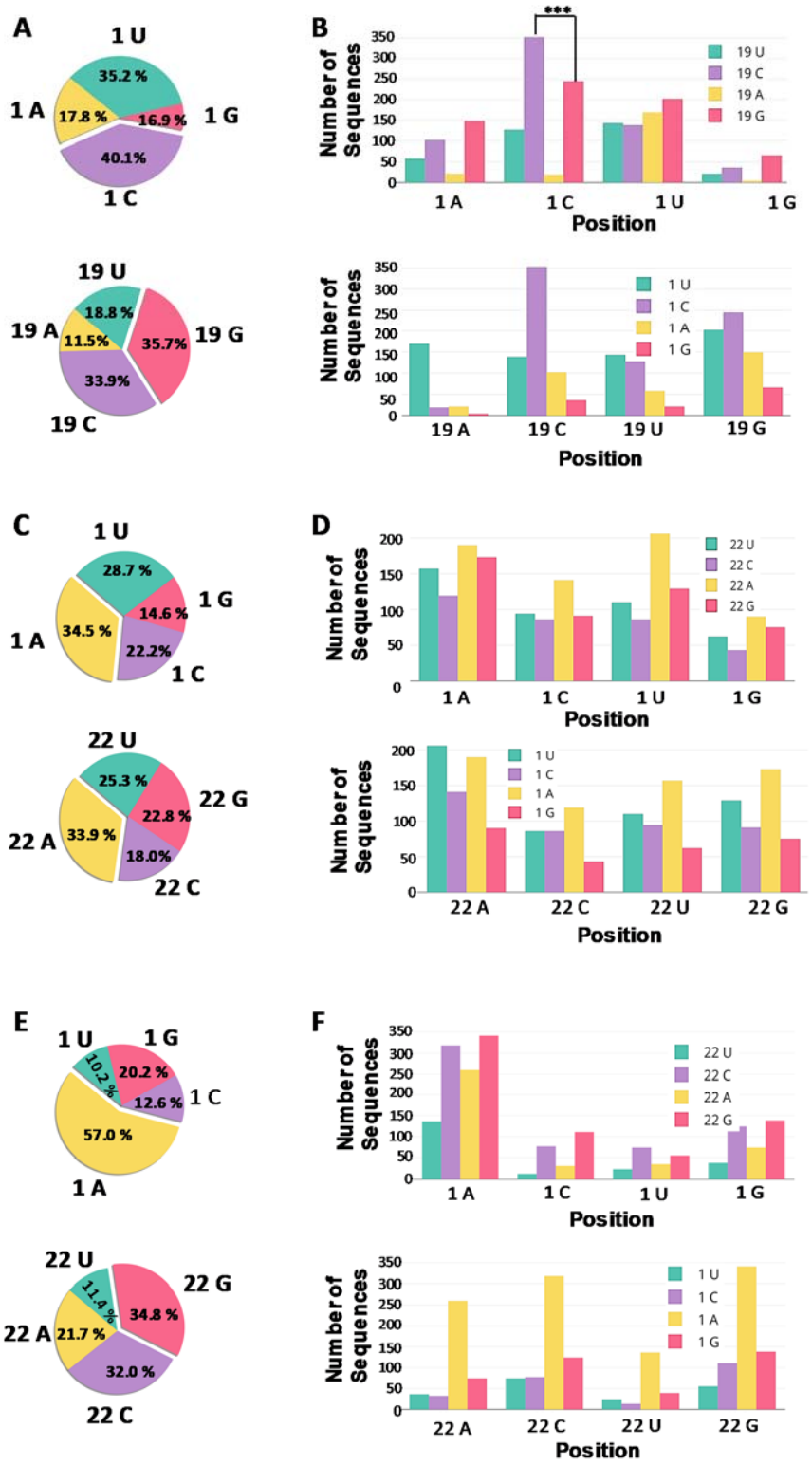


Figure 4. 5' ends in phasiRNA duplexes influence the composition of 3'-proximal nucleotides.

830 (A) The pie charts show the composition as a percent of all four nucleotides at the 1st (above)
831 and at the 19th (below) positions in 21-nt reproductive phasiRNAs, combined from maize and
832 rice. The predominant nucleotide is highlighted by separation from the other three. These data
833 are the same as Figure 2A (blue boxes in that figure), redrawn here for clarity. (B) Above,
834 nucleotide composition at the 19th position of the 21-nt phasiRNAs shown in panel A when the
835 1st position is selected or fixed, as indicated on the X-axis. Below, the same analysis for the 1st
836 position composition when the 19th position is selected or fixed. Significant differences are
837 indicated (Student's t-test): ***, $P \leq 0.001$. (C) Pie charts shows the composition as a percent of
838 all four nucleotides at the 1st and at the 22nd positions in 24-nt phasiRNAs, combined from
839 maize and rice. These data are the same as Figure 2C, left panel (blue boxes in that figure),
840 redrawn here for clarity. (D) Above, nucleotide composition at the 22nd position of the 24-nt
841 phasiRNAs shown in panel C when the 1st position is selected or fixed, as indicated on the X-
842 axis. Below, the same analysis for the 1st position composition when the 22nd position is
843 selected or fixed. (E) Pie charts as above, for P4-siRNAs, combined from maize and rice. These
844 data are the same as Figure 2C, right panel, redrawn here for clarity. (F) Above, nucleotide
845 composition at the 22nd position of the 24-nt P4-siRNAs shown in panel E when the 1st position
846 is selected or fixed, as indicated on the X-axis. Below, the same analysis for the 1st position
847 composition when the 22nd position is selected or fixed.

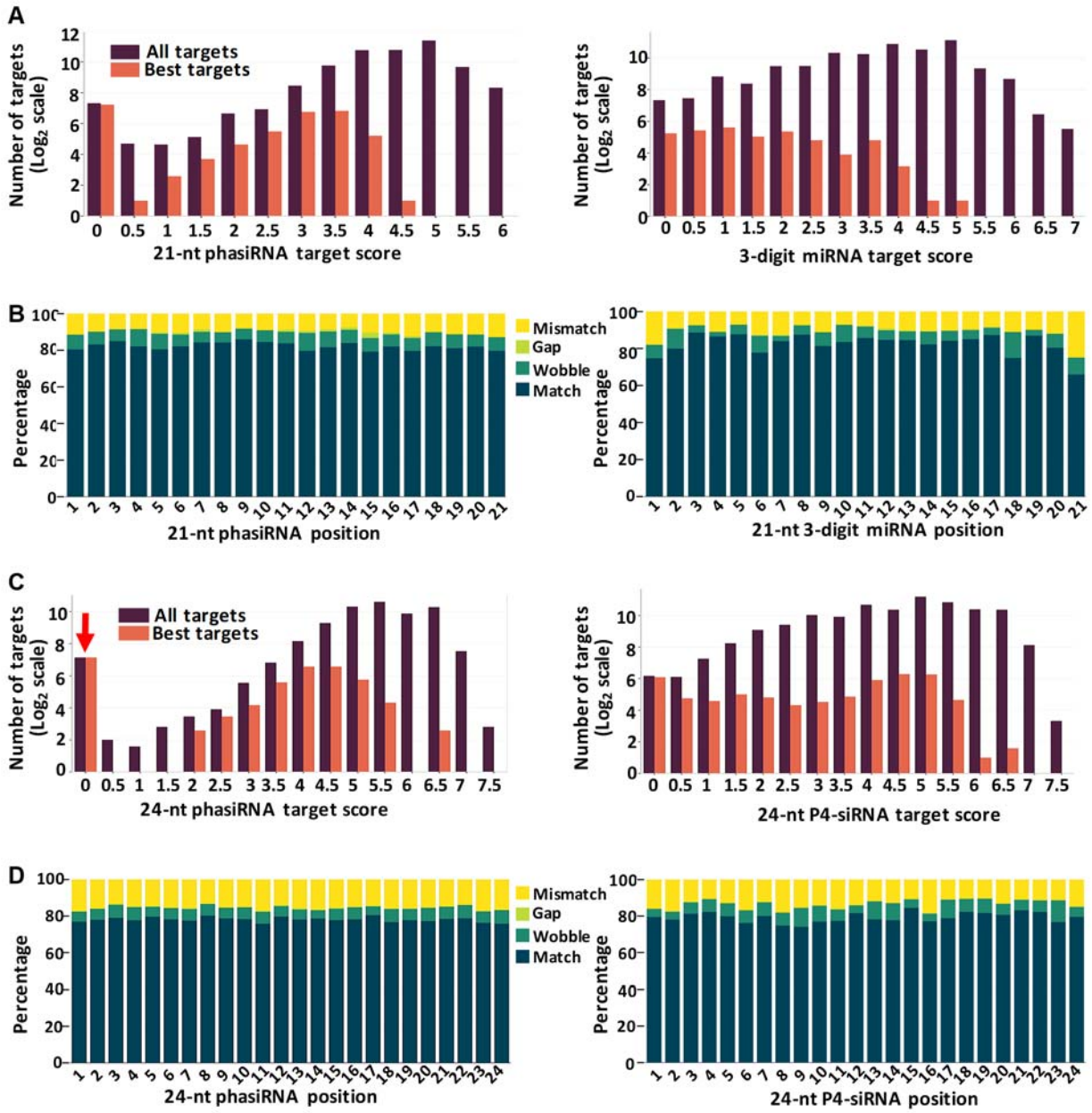


Figure 5. phasiRNA target prediction illustrates low binding affinity compared to other sRNAs to their targets due to sequence diversity.

Target prediction for top 500 most abundant 21- and 24-nt phasiRNAs in rice, rice 3-digit miRNAs (n=288), and top 500 most abundant P4-siRNAs in rice was performed using sPARTA. (A)The bar plots show target score distribution (as indicated on X-axis) for 21-nt phasiRNAs (at left) and 3-digit miRNAs (at right). Dark purple bars depict target score distribution of all targets of 21-nt phasiRNAs and 3-digit miRNAs. Orange bars depict target score distribution of only best targets (targets with a lowest target penalty score, meaning high degree of complementarity

856 between phasiRNAs or miRNAs and their targets) of 21-nt phasiRNAs and 3-digit miRNAs. As
857 indicated, Y-axis (number of targets) is transformed into log2 scale and red arrow indicates
858 potential self-targeting or *cis* interactions (with target score of 0, meaning perfect match). (B)
859 The bar charts record the 21-nt phasiRNA-target interaction (at left) and 3-digit miRNA-target
860 interaction (at right) for all targets with target score between 0.5 and 3.5, capturing binding
861 pattern as a percent (Y-axis) of match, gap, wobble, and mismatch. (C) Bar plots showing target
862 score distribution as above panel A, for 24-nt phasiRNAs (at left) and 24-nt P4-siRNAs (at
863 right). (D) As above panel B, the bar charts indicating the binding pattern as a percent (Y-axis)
864 of match, gap, wobble, and mismatch for 24-nt phasiRNA-target interaction (at left) and 24-nt
865 P4-siRNAs-target interaction (at right) for all targets with target score between 0.5 and 5.

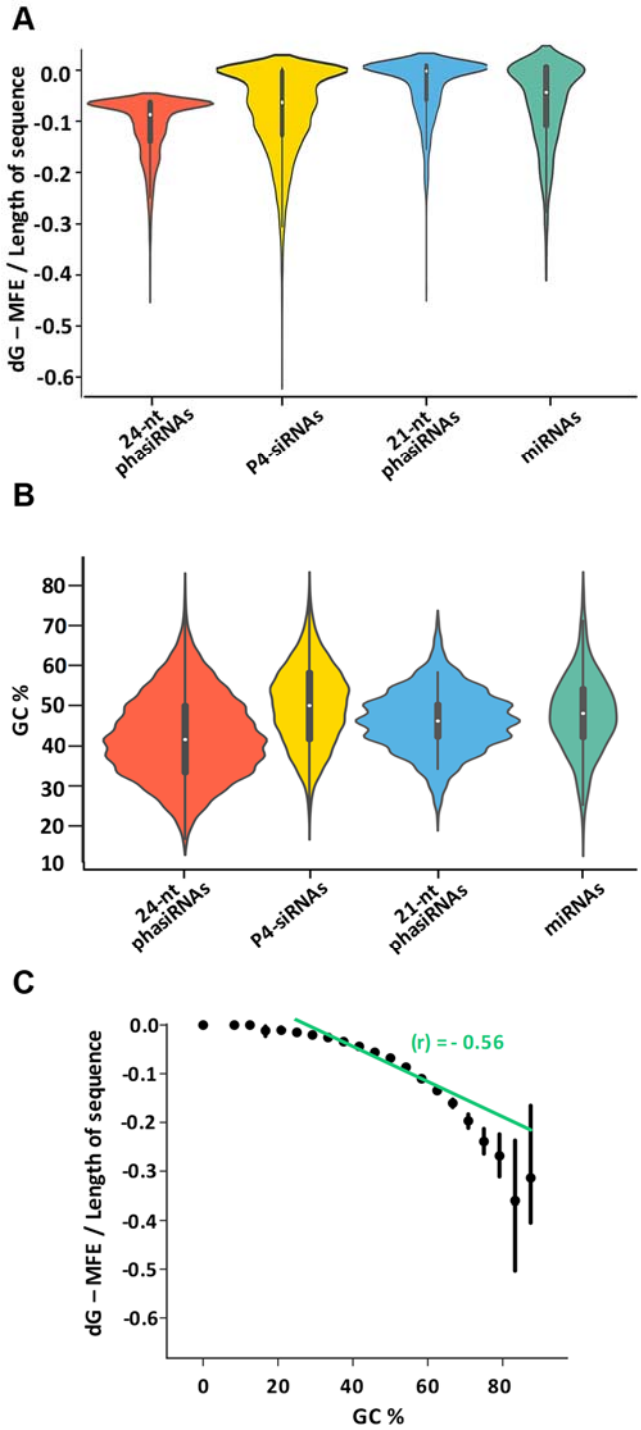
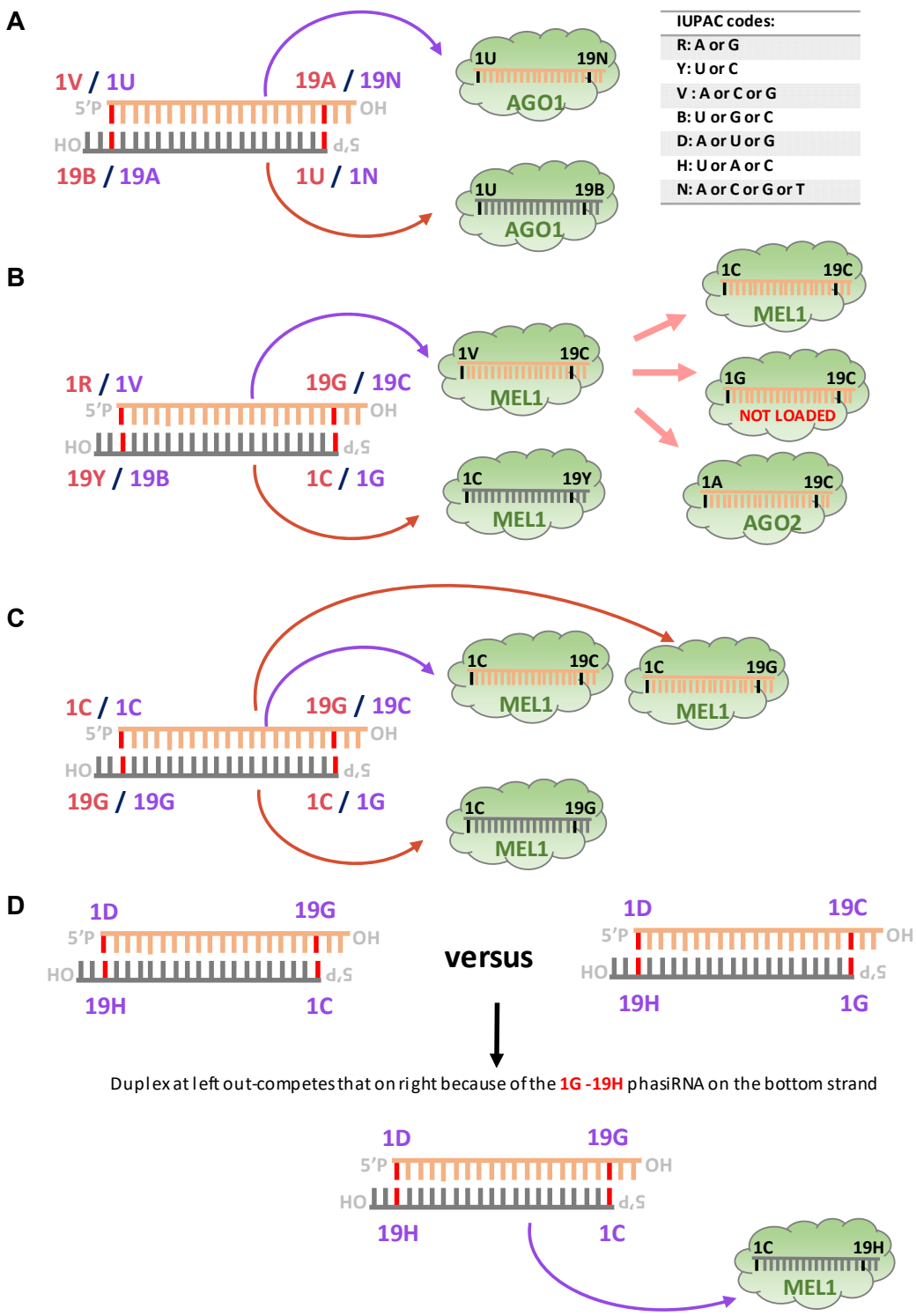


Figure 6. The GC content and the minimum free energy (MFE) in sRNAs indicate stability of RNA secondary structure.

(A) Violin plots compare the normalized MFE between 24-nt phasiRNAs (red), P4-siRNAs (yellow), 21-nt phasiRNAs (blue), and miRNAs (green). The center line in each plot indicates the

870 one standard deviation (thick bar) or two (thin bar) past the mean (white dot) of the distribution.
871 siRNAs with the most negative value of the MFE are considered to be more stable and structured.
872 (B) Violin plots compare the GC content for the same classes shown in panel A. (C) A statistical
873 correlation was observed between the MFE and the GC content in the reproductive phasiRNAs,
874 miRNAs, and P4-siRNAs (data combined); 'r' indicates the Pearson correlation coefficient (green
875 line). The observations in each discrete bin (black dots) were collapsed to plot an estimate of
876 central tendency along with a confidence interval (ci=99).



877 **Figure 7. An AGO competition model may explain abundance differences for 1U, 1C or**
878 **other 21-nt phasiRNAs.**
879 We propose a model for AGO proteins competing for phasiRNAs, distinguished by 1st and 19th
880 position nucleotides, to explain the substantial differences in abundances observed among

881 sequenced or absent phasiRNAs, and when compared to MEL1-associated 21-nt phasiRNAs.
882 IUPAC nucleotide codes (shown in the table) are used for simplifying the representation of
883 nucleotides at the 1st and the 19th positions, highlighted in 'red', in an orange top strand and a
884 gray bottom or complementary strand. Different combinations of the nucleotides at the 1st and
885 the 19th positions on a top and a bottom strand, comprising a single duplex are denoted in "pink"
886 or "purple". Loading of a particular strand from a duplex into a specific AGO is indicated by the
887 arrow of same color as the nucleotides on that strand. (A) In a 1V/19A - 1U/19B duplex, 1U/19B
888 (bottom) phasiRNA is more likely to be loaded into AGO1. In a 1U/19N – 1N/19A duplex,
889 1U/19N (top) is more likely to be loaded into AGO1. (B) As in panel (A), in a 1R/19G – 1C/19Y
890 duplex, 1C/19Y is preferentially loaded into MEL1. In a 1V/19C - 1G/19B duplex, the 1V/19G
891 phasiRNA is either loaded into MEL1 or another AGO (AGO2?). The 1G/19C phasiRNA is
892 likely not loaded into AGO. (C) In a 1C/19G - 1C/19G duplex, phasiRNAs from both strands are
893 loaded into MEL1. In a 1C/19C - 1G/19G duplex, only the top strand is loaded into MEL1. (D)
894 Comparing two duplexes, 1D/19G - 1C/19H outcompetes 1D/19C - 1G/19H, yielding a 1C/19H
895 MEL1-bound phasiRNA (from the bottom strand).

896 **TABLES**

897

898 **Table 1.** Results of classification to distinguishing phasiRNAs of lengths 21-nt (top) and 24-nt
899 (bottom) from other small RNA types.

Classification		**Performance Evaluation Measure**				
Positive set	**Negative set**	ACC (±SD)	SP (±SD)	SE (±SD)	PPV (±SD)	AUC (±SD)
21-nt phasiRNA	miRNAs* + P4-siRNA + tRNA* + rRNA*	0.92 (±0.01)	0.90 (±0.00)	0.93 (± 0.01)	0.93 (± 0.01)	0.97 (± 0.00)
	miRNAs* + P4-siRNA	0.93 (±0.01)	0.89 (±0.00)	0.97 (±0.01)	0.87 (±0.02)	0.96 (± 0.00)
	P4-siRNAs, 3' trimmed	0.81 (±0.01)	0.80 (±0.01)	0.81 (±0.01)	0.81 (±0.01)	0.89 (± 0.00)
	miRNAs, 3' trimmed	0.82 (±0.01)	0.74 (±0.01)	0.90 (± 0.01)	0.78 (± 0.01)	0.88 (± 0.02)
24-nt phasiRNA	P4-siRNA	0.84 (±0.02)	0.83 (±0.00)	0.84 (± 0.02)	0.83 (±0.02)	0.91 (± 0.01)
	miRNAs* + P4-siRNA + tRNA* + rRNA*	0.87 (±0.01)	0.83 (±0.00)	0.91 (± 0.01)	0.85 (±0.01)	0.94 (± 0.00)

900 Note: ACC, accuracy; SP, specificity; SE, sensitivity; PPV, positive predictive value. See methods
901 for further detail. An asterisk (*) next to a negative subset indicates no size selection or trimming
902 of the sequences. Results are averaged over the five-fold cross-validation. The size of positive and
903 negative sets are as follows: 21-nt phasiRNAs (n=2000), 24-nt phasiRNAs (n=2000), miRNA
904 (n=756), P4-siRNAs (n=2000), tRNAs (n=500), and rRNAs (n=500).

905 **Table 2.** Predictive performance of classification models of 21-and 24-nt phasiRNAs.

906

Predictive sensitivity	Classification Model (positive set vs negative set)	Performance Evaluation Measure	
		TP	SE
21-nt phasiRNAs	21-nt phasiRNA vs. miRNA + P4-siRNA + tRNA + rRNA	26208/27500	0.953
24-nt phasiRNAs	24-nt phasiRNA vs. miRNA + P4-siRNA + tRNA + rRNA	7093/7750	0.915

907 Note: TP, true positive prediction; SE, sensitivity.

908 **Table 3.** Structural features used in the machine learning classifier.

Structural Features	Description
dG =MFE/L	Length-normalized minimum free energy of folding (dG) where L is the length of the sequence.
dP = tot_bases/L	Length-normalized base-paring propensity (dP) where tot_bases is the total number of base pairs in the secondary structure; L is the length of the sequence.
NEFE=EFE/L	Normalized Ensemble Free Energy (NEFE).
MFEI1= dG/GC_content	Index 1 based on MFE.
MFEI2 = dG/n_stems	Index 2 based on MFE where n_stems is the number of stems in the secondary structure.
MFEI3 = dG/n _loops	Index 3 based on MFE where n _loops is the number of loops in the secondary structure.
MFEI4 = MFE/tot_bases	Index 4 based on MFE.

909

910 **Table 4.** Top 25 features, from example comparisons, obtained using information gain.

Classification	Feature Group	No. of features	Feature Symbol
21-nt phasiRNA vs. P4-siRNA (3' trimmed)	Sequence-based features	18	CGG, CT, CG, T, ACG, G, GG, ACGG, CGGA, GGA, GC, A, CAT, AAACG, CGGAC, AC, GACG, TGC
	Structural Features	7	NEFE, dG, MFEI1, MFEI2, MFEI3, MFEI4, dP
21-nt phasiRNA vs. miRNA (3' trimmed)	Sequence-based features	18	G, AC, GCCAA, GG, GCC, TGCC, GCCA, GGG, TGCCA, TA, A, AGA, CTGA, AAG, GC, GGG, CTA, GTGTT
	Structural Features	7	NEFE, dG, MFEI1, MFEI2, MFEI3, MFEI4, dP
24-nt phasiRNA vs. P4-siRNA	Sequence-based features	18	GC_content, T, CGG, G, GG, CG, AT, ACGG, ACG, CGGA, GGA, GAC, GGAC, TT, GACG, CGGAC, CAT, TAT
	Structural Features	7	NEFE, dG, MFEI1, MFEI2, MFEI3, MFEI4, dP

911 Note: Sequence-based features denote the frequency of the k-mer motifs as they appear along the
912 length of the sequence.