Supplementary Text and Methods

Plant 24-nt reproductive phasiRNAs from intramolecular duplex mRNAs in diverse monocots

Atul Kakrana[1,2], Sandra M. Mathioni[3], Kun Huang[2], Reza Hammond[1,2], Lee Vandivier[4], Parth Patel[1,2], Siwaret Arikit[5], Olga Shevchenko[2], Alex E. Harkess[3], Bruce Kingham[2], Brian D. Gregory[4], James H. Leebens-Mack[6], Blake C. Meyers[3,7*]

[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] Donald Danforth Plant Science Center, St. Louis, MO 63132, USA
[4]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[5]Department of Agronomy, Kasetsart University, Nakhon Pathom 73140, Thailand
[6]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA
[7] Department of Plant Science, University of Missouri – Columbia, MO 65211, USA

List of contents

3. References

1. Long-miRNAs in *Asparagus* and comparative study of their prevalence

Long miRNAs (lmiRNAs) are not common in plants, but are well represented in *Amborella* (Albert et al. 2013), rice (Wu et al. 2010) and eudicot species (Vazquez et al. 2008). *Asparagus* displayed many loci (n=236) generating candidate 24-nt miRNAs resulting in 203 mature 24-nt miRNAs. To assess the validity of the high numbers in *Asparagus*, we ran the same analysis on banana, *Brachypodium* and *Arabidopsis* using public data (Nakano et al. 2006); this identified 132, 32 and nine 24-nt candidate miRNAs, all lineage- or species-specific with no significant conservation, and presence in many libraries (**data not included here**). Grouped by abundances in different *Asparagus* tissues and stages, many of these lmiRNA candidates were found at lower levels in meiotic-stage anthers. Like canonical 21- or 22-nt miRNAs, these lmiRNA precursors were predicted to form a hairpin structure. Next, we tested whether these lmiRNAs might direct cleavage of their targets. We used Parallel Analysis of Read Ends (PARE) libraries for *Asparagus* male and female flowers, spears, leaf and shoot samples (**Table S1**). Only a small proportion (2.9%) of the predicted targets (n=10) demonstrated enrichment in the PARE libraries, perhaps more consistent with a role in directing DNA methylation at target sites (Chellappan et al. 2010; Wu et al. 2010).

2. Methods

2.1 Sample Collection and RNA isolation

Asparagus officinalis samples were collected from a commercial field in the T.S. Smith and Son's Farm (http://www.tssmithandsons.com/), Bridgeville, Delaware. Flowering *Lilium* and daylily plants were purchased from Home Depot (Newark, Delaware). Anther stages were examined on propidium iodide-stained (*Asparagus* and *Lilium*) or cleared tissue (daylily) using confocal microscopy. Samples were collected and anthers were dissected using a 2 mm stage micrometer (Wards Science, cat. #949910) in a stereo microscope, and immediately frozen in liquid nitrogen until total RNA isolation was performed. Total RNA was isolated using the *PureLink Plant RNA Reagent* (ThermoFisher Scientific, cat. #12322012) following the manufacturer's instructions. Total RNA quality and quantity were assessed before proceeding to the next step. Small RNAs (20 to 30 nt) were size selected in a 15% polyacrylamide/urea gel and used for small RNA library preparation. An aliquot of 3 µg of total RNA was used for size selection.

2.2 Anther stages-size correlation microscopy

Anthers from *Asparagus* and *Lilium* were dissected and vacuum fixed using 4% paraformaldehyde, and submitted to histology lab (A.I DuPont Hospital for Children) for paraffin embedding. Then *Lilium* samples were examined using PI-staining (Propidium Iodide). Briefly, the paraffin slides were de-paraffinized with histoclear, and washed with 100% ethanol. Then samples were equilibrated in 2x SSC (pH 7.0) and stained in 500 mM PI (in 2xSSC) for 1-5 min and mounted in slow-fade gold (ThermoFisher Scientific, Inc.). Stages

were assigned based on the morphology of archesporial AR and tapetum cells. For daylily, anthers were dissected and vacuum fixed using 4% paraformaldehyde, then cleared with ScaleP solution for 1 week (Warner et al. 2014). Histology and cell division of the longitudinal images of anther were examined using confocal microscope for stage determination.

2.3 Small RNA, mRNA and PARE library construction, and Illumina sequencing

Small RNA libraries were constructed using the *TruSeq Small RNA Library Preparation Kit* (Illumina, cat # RS-200-0024) as per manufacturer's instructions and as described by Mathioni et al. (2017). RNA-seq libraries were constructed using the *TruSeq Stranded Total RNA Library Preparation Kit* with *Ribo-Zero Plant* (Illumina, cat # RS-122-2401), and RNA was treated with DNase I (NEB, cat # M0303S) and then cleaned using the *RNA Clean & Concentrator™-5* (Zymo Research, cat # R1015). PARE libraries were constructed as previously described (Zhai et al. 2014), with the exception of using 10 ug of total RNA. Small RNA and PARE libraries were single-end sequenced with 51 cycles, and stranded RNA-seq libraries were paired-end sequenced with either 101 or 151 cycles. All libraries were sequenced on an *Illumina HiSeq 2500* instrument at the University of Delaware Sequencing and Genotyping Center in the Delaware Biotechnology Institute.

2.4 Pre-processing sRNA, PARE and mRNA-sequencing libraries

Small RNA and PARE libraries were pre-processed using the script "prepro.py" version 0.2 (https://github.com/atulkakrana/preprocess.seq) with default settings as described earlier (Patel et al. 2016; Mathioni et al. 2016). Preprocessing included trimming of 5' and 3' adapters, cropping of reads to 20-nt for PARE libraries, and finally retaining 18- to 36-nt and 20nt reads for sRNA and PARE libraries, respectively. All the reads in processed files were aligned to the *Asparagus* genome (v.1) using Bowtie (v0.12.8) with no allowed mismatches. Mapped reads were finally normalized to empirically derived, 30 million reads base depth. Please refer **Table S1**, for number of sequenced-, mapped-, and distinct-reads, with corresponding GEO IDs for each library. RNA-sequencing libraries were processed using the same script (as above) with default settings. These reads were cropped by 5 nt from 3'-ends to increase the proportion of reads mapped to genome.

2.5 Single-molecule real time (SMRT) Sequencing

The collected plant material was ground in a cold mortar and pestle using liquid nitrogen. Total RNA was isolated using the *PureLink® Plant RNA Reagent* (Life Technologies, cat. # 12322-012), treated with DNAse I (NEB, cat. # M0303S) cleaned and concentrated with *RNA Clean and Concentrator-5* (Zymo Research, cat. # R1015). Then the *MicroPoly(A) Purist™ Kit* (Ambion, cat. # AM1919) was used for isolation of poly(A) RNAs. The poly(A) RNA samples were then converted into cDNA using the SMARTer™ PCR cDNA Synthesis Kit (Clontech, cat. # 634926) and the *SageELF Size Selection System* protocol as described by Pacific Biosciences in protocol # PN100-574-400-02. The cDNA was size selected and fractionated into 12 fractions, which were then pooled into three size ranges: 0.8-2.0 kb, 2.0-5.0 kb, and > 5.0 kb. SMRTbell libraries were prepared for the three cDNA size ranges using the DNA Template Library Preparation kit

(SMRTbell Template Prep Kit 1.0) following the Pacific Biosciences protocol # PN100-574-400-02. A total of 9 SMRT Cells (Pacific Biosciences part # 100-171-800), for each species (*Asparagus* and daylily) and three per library, using the P6C4 polymerase (Pacific Biosciences part #100-372-700) were run on a *PacBio RS II Instrument* at the University of Delaware Sequencing and Genotyping Center (Delaware Biotechnology Institute, Newark). Raw sequencing data was pre-processed using the *pbtranscript-tofu* tool set (v2.3.0) using the default settings. Please note the *pbtranscript-tofu* is now available as Iso-seq3 developer version (https://github.com/PacificBiosciences/IsoSeq_SA3nUP). The pre-processing included classification of reads to full-length and non-full-length categories, followed by clustering of transcripts to consensus isoforms by ICE algorithm and final polishing by Quiver algorithm (min. accuracy = 0.99). For all downstream analysis, "high QV" transcript set generated from Quiver analyses was used. This set was further collapsed based on sequence similarity i.e. without the reference genome, to remove any redundancy in transcripts, especially for transcripts corresponding to same isoforms, by using CD-HIT with recommended parameters https://github.com/PacificBiosciences/cDNA_primer/wiki. In case of *Asparagus*, an additional step was performed to identify novel isoforms and transcriptional-loci. The collapsed "high QV" set was compared with the annotated gene-models using *MatchAnnot* (MA) tool (https://github.com/TomSkelly/MatchAnnot). FL transcripts that matched annotated gene structure with MA score > 2 and on same strand were considered as known, those with MA score <= 2 on same strand were considered as novel isoforms to known genes, and finally those either with MA score <= 2 on opposite strand or no MA assigned score were considered as novel transcription loci. Please see main text for species-specific tallies of known, novel isoforms or transcriptional loci.

2.6 microRNA prediction

Mapped sRNA reads from all libraries were used as input to two different computational pipelines for discovery of miRNAs – a stringent pipeline for *de novo* identification and a relaxed pipeline for identification of conserved 'known' miRNAs (Jeong et al. 2013). Steps in both pipelines involved processing using *perl* scripts as described earlier (Jeong et al. 2011), with modified version of miREAP (https://sourceforge.net/projects/mireap/) and CENTROIDFOLD (Sato et al. 2009). In 'stringent' criteria pipeline, sRNAs of length between 20 and 24 nt, with abundance >= 50 TP30M in at least one library, and total genome hits <= 20 were assessed for potential pairing of miRNA and miRNA* using modified miREAP optimized for plant miRNA discovery with parameters *–d 400 –f 25*. Strand bias for precursors was computed as ratio of all reads mapped to sense strand against total reads mapped to both strands. In addition to strand bias, abundance bias was computed as ratio of two most abundant reads against all the reads mapped to same precursor. Candidate precursors with strand bias >= 0.9 and abundance bias >= 0.7 were selected, and foldback structure for precursor was predicted using Centroid Fold. Each precursor was manually inspected to match the criteria as described earlier (Jeong et al. 2013). All the miRNAs identified through this stringent pipeline were then annotated by matching mature sequences to miRBASE (version - 21), and those that did not matched to any known miRNA were considered as lineage or species-specific. In 'relaxed' criteria pipeline, which is implemented to maximize identification of 'known' miRNAs, relaxed filters were applied – sRNA between 20 and 24nt, with hits <= 20 and abundance >= 15 TP30M, and precursors with strand bias >= 0.7 and abundance bias >= 0.4. Stem-loop structure of candidate precursors was visually inspected, same as the 'stringent' pipeline. Mature sequences of identified

miRNAs were further matched with miRBASE entries (v21), and those with total 'variance' (mismatches and overhangs) <= 4 were considered conserved miRNAs.

2.7 Computing degree of overlap between two genomic features

The enrichment or depletion of overlap between sRNA generating locations like lmiRNAs and *PHAS* loci, and genome-features like exons, introns, inverted repeats and transposable-elements is computed based on the overlapping nucleotides between sRNA and genome-feature. For a pairwise comparison, an enrichment or depletion ratio was computed as:

Overlap Ratio = log2(O) -log2(E)

Expected Overlap (E) = (x/g) * (y/g) * g

Where, 'E' is the expected number of overlapping nucleotides between sRNA-location (feature-A) and genome-feature (feature-B) under null hypothesis of random chance, 'O' is the observed nucleotides of feature-A overlapping with feature-B, 'x' is total number of non-redundant nucleotides of any feature-A, 'y' is total number of non-redundant nucleotides of any feature-B, 'g' is the total genome size.

2.8 PhasiRNA prediction and trigger identification

Phased siRNA generating (*PHAS*) loci or precursors were identified using the purpose-built tool 'phasdetect' tool from 'PHASIS' suite (Kakrana *et al.* 2017). The *PHAS* loci (or precursors), predicted from different sRNA libraries were collapsed to a non-redundant set by using 'phasmerg*e'* tool from *PHASIS*. Triggers for these *PHAS* loci (or precursors) were further identified using the 'phastrigs' tool, part of PHASIS). As a control for *phastrigs* predictions, we first predicted *PHAS* loci in maize using publically available sRNA libraries (Zhai et al. 2015), and then tested *phastrigs* to identify trigger miRNAs. It identified triggers for 63% and 40% of 21- and 24-nt reproductive *PHAS* loci. For 21-nt reproductive *PHAS* loci members of miR2118 family members were identified as trigger, and for 24-nt reproductive *PHAS* miR2275 family was identified as trigger. The low proportion of *PHAS* for which triggers were identified could be because of splicing in *PHAS* precursors, so those for which miRNA triggers were not identified are actually spliced portion of other *PHAS* loci in vicinity.

2.9 Coding and non-coding assessment

We built a logical classifier that uses Coding Potential Calculator scores (Kong et al. 2007) and Coding Potential Assessment Tool probabilities (Wang et al. 2013), to use – ORF length, ORF integrity, hit score (with known proteins), ORF coverage, Fickett TESTCODE statistics and hexamer usage, for classification of assembled transcripts into 1) coding 2) non-coding and 3) transcript of unknown coding potential (TUCP). CPC determines coding potential based on sequence homology to known proteins, while CPAT assess coding potential purely on transcript sequence using a logistic regression model from ORF coverage, Fickett TESTCODE statistics and hexamer usage bias. CPAT is particularly useful for less conserved proteins from new species, lncRNAs overlapping with protein-coding genes and addresses the issues with quality of sequence alignment in case of homology based coding potential prediction tools. In order to use CPAT, for which no recommended probability cutoff for plants is available, we first determined an optimum probability cutoff by repeatedly randomly sampling 100 each of protein-coding and non-coding transcripts

and optimizing on the balanced accuracy metric (average of specificity and sensitivity metrics). For this we used "reviewed" proteins from Uniprot and putative lncRNAs submitted to Plant Non-coding RNA Database (Yi et al. 2015) and RNA-central database (2015), corresponding to maize which is the closest well annotated monocot to species included in this study. The average area under curve for 1000 iterations was 0.9092, and the average optimal probability cutoff was 0.2212. This cutoff value displayed accurate discrimination of protein-coding and non-coding transcripts (sensitivity = 0.8, specificity = 0.98 and FDR = 0.061). Using the recommended score for CPC and this empirically derived cutoff for CPAT, we classified the transcripts as follows:

1) Coding, if a) CPC score >= 1 (strong coding evidence) or b) CPC score between 0 to 1 (weak coding evidence) and CPAT cutoff > 0.2213 along with ORF >= 100 aa,

2) Non-coding, a) if CPC score <= -1 (strong non-coding evidence) and ORF <= 100 aa or b) CPC score between -1 to 0 (weak non-coding evidence) and CPAT cutoff < 0.2213 along with ORF <= 100 aa, and finally

3) TUCP if none of the above criteria matches.

2.10 Transcriptome assembly, quality assessment and comprehensive transcriptome

Pre-processed RNA-seq libraries and polished full-length transcripts from SMRT-seq experiments were used to generate species-specific transcriptome libraries. For *Asparagus*, an *ab initio* assembly was generated by following Tophat and Cufflinks protocol (Trapnell et al. 2012). This included mapping of all sample-specific RNA-seq libraries, both single- and paired-end, to the *Asparagus* genome using *Tophat* with default settings, followed by generation of sample-specific transcript assemblies through *cufflinks*, which used annotated gene models as reference and finally merging of these assemblies using *cuffmerge* to give a single combined transcriptome assembly. The (*de novo)* hybrid transcriptome assemblies for *Asparagus* and daylily were generated using Trinity platform (Haas et al. 2013). For this, reads from paired-end libraries were first combined into two (FASTQ) files, one corresponding to left reads and other to right reads. Reads from the single-end libraries were then added to the combined left reads (FASTQ) file. These left and right reads files along with full-length reads supplied through '--long-reads' parameter, were used to generate a hybrid assembly with the default settings except for the minimum assembled contig length (set to 250 nt). Similar to *Asparagus* and daylily, for *Lilium*, paired-end libraries from different samples were first combined into two files, one for left reads and other for right reads. These combined files were then used to generate a *de novo* transcriptome assembly using Trinity (v2.1.1) using default settings except the for the minimum assembled contig length (set to 250 nt) and an additional digital normalization step to reduce memory requirements. ExN50 and the quality of assemblies was accessed as recommended in Trinity workflow. Transcripts from hybrid *de novo* assemblies generated for *Asparagus* and daylily and from *de novo* assembly generated for *Lilium* were annotated using Trinotate workflow with the default settings (https://trinotate.github.io/). Candidate protein transcripts generated as part of the Trinotate annotation process were used for further downstream analysis. Expression-level qualification of transcripts from these species-specific (*de novo*) assemblies was done using the RSEM algorithm (Li and Dewey 2011) with default settings, as implemented in the Trinity platform.

Asparagus *de novo* hybrid assembly resulted in 6,623 transcripts matching the annotated asparagus genes and 69,642 novel isoforms. This *de novo* assembly had an Ex90N50 value of 1,396 and captured near full length transcripts (> 80% alignment coverage) for 6,998 unique proteins from Uniprot, indicating a good transcriptome quality. The daylily *de novo* hybrid transcriptome assembly and lily *de novo* transcriptome assembly yielded 157,913 and 182,225 transcripts with normalized expression greater than 1TPM in at least one library, for lily and daylily respectively. Transcript assemblies for both species displayed a significant Ex90N50 statistic (**Supplemental Fig. S8**) and captured near full-length transcripts for at least 6,550 and 7,384 different proteins (≥ 90% alignment coverage, relative to Uniprot) (The UniProt Consortium 2015).

2.11 *dsRNA*-sequencing library preparation and pre-processing

Structure libraries were created as previously described (Li et al., 2012; Vandivier, Li, and Gregory, 2015). For each sample, 100ug of purified total RNA was split into two 50ug aliquots. One aliquot was treated with 1ul single-stranded *RNase ONE®* (Promega), and the other with 5ul double-stranded RNase V1 (Ambion). Both RNase ONE® and RNase V1 were allowed for cut for 1hr at 37C, cutting away ssRNA and dsRNA to completion and yielding dsRNA and ssRNA fragments, respectively. These fragments were then adapter-ligated, PCR amplified, and barcoded using *Illumina TruSeq® smRNA adapters*. Completed dsRNA-seq and ssRNA-seq libraries were sequenced to 51 bp, single-end, on an Illumina HiSeq 2500 instrument. Note that RNase V1 is no longer commercially available, but can be purified from commercially available cobra venom (Mahalakshmi et al. 2000).All downstream analyses were performed using the *Asparagus* genome assembly and transcriptome annotations. Demultiplexed sequencing reads were first trimmed with Cutadapt v1.9.1 to remove 3' sequencing adapters (adaptor sequence: TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACnnnnnnATCTCGTATGCCGTCTTCTGCTTG). Reads with no detectable adaptor were retained in the trimmed read sets. Trimmed reads were mapped to the *Asparagus* genome using Tophat (v2.1.0), allowing up to 10 multi-mappings of each read.

2.12 Secondary-structure score computation

Base-wise structure scores were defined by calculating a normalized ratio of reads from dsRNA-seq to ssRNA-seq. For multi-mapping reads (>5 hits), only one random mapping was considered in calculating coverage. Raw coverage (rds_i and rss_i) for each library was then normalized to the total number of primary aligned mapped bases in each library (N_{ds} and N_{ss}). Structure score (S_i) was calculated as the generalized log ratio (glog) of normalized dsRNA-seq (ds_i) to normalized ssRNA-seq (ss_i)

$$S_i = glog(ds_i) - glog(ss_i) = log_2\left(ds_i + \sqrt{1 + ds_i^2}\right) - log_2\left(ss_i + \sqrt{1 + ss_i^2}\right)$$

$$ds_i = rds_i \cdot \frac{N_{ds}}{N_{ss}} \; ; \; ss_i = rss_i \cdot \frac{N_{ss}}{N_{ds}}$$

Similarly, strand scores were computed as generalized log ratio (glog) of sense versus anti-sense ds-RNA sequencing reads. All structure mapping scripts, including the modified scripts derived from CSAR, are available on https://github.com/GregoryLab/structure

2.13 Probing secondary structure of sRNA associated loci of interest

We used hc-siRNA generating loci as one control for *PHAS* loci for secondary structure studies. For this we first identified sRNA-associated clusters using ShortStack (Axtell 2013). All the sRNA libraries (Table S1) were used as an input to ShortStack. Clusters with phasing p-value >= 0.05, dicer call = 24, showing overlap (>30%) with transposable-elements, and not annotated as miRNA or hpRNA were considered putative hc-siRNAs generating loci. A representative set for comparison with *PHAS* loci was selected by randomly picking 300 loci. *PHAS*-, hc-siRNA loci are computationally defined regions based on sRNAs population, unlike the protein-coding regions that have empirically derived 5' and 3' co-ordinates along and gene-structure information based on mRNA data. Therefore, to ensure that sufficient (per-bp) data is captured for these computationally defined regions in the RNA secondary structure libraries, we computed a locus-specific coverage threshold representing reliable coverage. This 'reliable coverage cutoff' was determined for every locus by randomly sampling regions (n=500) of same length and computing 97.5^{th} percentile of coverage. This process is repeated 1000 times (iterations) and median of 97.5^{th} percentiles from these iterations is considered as coverage cut off for specific locus. *PHAS*-, miRNA- and hc-siRNA loci passing the 'reliable coverage cutoff' were considered for other downstream analyses.

Average structure- and strand-scores for these sRNA-associated loci was computed as described earlier (Li et al. 2012). Empirical FDR thresholds for these scores was calculated by randomly permuting dsRNA- and ssRNA-sequencing reads for structure scores, and shuffling dsRNA-sequencing reads between "Watson" and "Crick" strands for strand-scores, and finally determining the threshold at which 5% of permuted peaks are called as significant (Li et al. 2012). For all analyses involving an average structure- or strand-score, positions with a score of '0' were ignored. Regions with 6-fold or higher structure- and strand-scores were considered as structured and stranded respectively.

To infer the structural pattern within the 24-*PHAS* loci, first, the structured strand (one with high structure scores) was selected for these loci and per-base-pair scores (including replicates) for each *PHAS* were congregated into a set of 100 bins with median scores representing each bin. Mean of these binned scores from 24-*PHAS* loci were used to plot the consensus. Randomly permuted samples (n=5) were used as control to compute statistical significance for secondary structure at the arms by using Wilcoxon signed-rank test (paired) test. The scores for each bin were compared with same bin in shuffled data and the overall differences between the real and shuffled data was tested using Wilcoxon signed-rank test. Such that every test (arm-region) compared with shuffled control-1 (arm-region), then with control-2 and with control-3 and so on. *P-value* from each test, in this case 5 *p-values* for 5 controls are combined using fisher's method.

2.14 Identification of isomiRs and putative miRNA loci in sequenced genomes

The isomiRs were identified by matching the small RNA reads from vegetative and reproductive libraries, against the mature miRNA sequences in miRBASE (v.21). Those matching the miRBASE entries with 5 nt or less variance i.e. sum of mismatches, single nucleotide insertion or deletion (only one instance allowed), and single nucleotide 5' or 3' offset (only two nucleotide offset allowed) is five nucleotides or less, are considered as valid isomiRs. Candidate loci for miR2118 and miR2275 family members in *Amborella* and *Zostera* were identified by employing a reverse approach. At first, mature sequences for both these families, from maize and rice along with the species-specific isomiRs were mapped to genome by allowing five or less edits i.e. mismatches, single nucleotide insertion or deletion. A 300 nt (+150/-150) region flanking these mapped sites was then investigated for foldback structure. The loci capable of forming a foldback, and with miRNA map site located in 5'- or 3'-arm i.e. not located in terminal loop or unstructured region, were considered as valid candidates. Finally, these loci were manually investigated to exclude those that display characteristics of hc-siRNA associated regions.

2.15 Identification of candidate precursors processed from loop-to-base direction

The first phased-position from 5'-end of double-stranded region in foldback precursors was considered as start-site for phased-siRNA production. The following phased positions for which no phasiRNA was detected, their abundance was set to zero and an abundance ratio was computed for phasiRNAs emanating from the 5'-start (base-side) against those emanating from the 3'-end (loop-side) of fold-back structure by dividing double-stranded region into two parts. Foldback precursors that displayed 8-fold (log2 ratio >= 3) bias in phasiRNAs abundance towards the 3' end of foldback were considered as candidate precursors that are likely processed from loop-to-base direction. These precursors were then manually checked for absence of phased-positions towards 5'-end and to exclude those candidates that showed bias due to one or two highly abundant phasiRNAs. The final representative set (n=9 precursors) was used for comparison with those triggered by miR2275 and displaying raggedness at first phase-cycle

2.16 Identification of Dicer and AGO families

Species-specific transcriptome annotations from the Trinotate workflow were manually curated to identify Dicer and Argonaute family members in *Lilium* and daylily. In *Asparagus,* protein- and nucleotide-BLAST was used to identify protein transcripts from annotated gene models and genomic copies of AGO and DCL members. Orthologs from monocots (rice, maize) and dicot (*Arabidopsis* and soybean) species were used as query sequences in both scans, and their results were manually curated. Computationally predicted protein transcripts for these candidates were aligned to orthologs from rice, maize, *Arabidopsis* and soybean using T-COFFEE multiple sequence alignment tool (v.3.8) (Notredame et al. 2000) in 'accurate' mode. Finally, a phylogenetic tree was generated using PhyML (Guindon et al. 2010) with default parameters and Non parametric bootstraps (n= 1000) replicates along with the BEST approach used to optimize tree topology. The latter combines both nearest neighbor interchanges (NNI), and subtree pruning and regrafting (SPR) approach and returns the best solution among two.

2.17 Fluorescent *in situ* hybridizations for *PHAS* precursors

Small RNAs were detected using LNA probes by *Exiqon* (Woburn, MA). Samples were vacuum fixed using 4% paraformaldehyde, and submitted to histology lab (A.I DuPont Hospital for Children) for paraffin embedding. We followed the protocols for the pre-hybridization, hybridization, post-hybridization and detection steps as previously described (Javelle and Timmermans 2012). For fluorescent *in situ* hybridization of DCL3b mRNA, paraffin slides were de-paraffinized with 'histoclear' and then washed in ethanol series (100%, 95%, 80% 70%, 50%, 30% 10% and water). Protease treatment for 20 min (final concentration 65 µg/ml) followed by 0.2% glycine treatment in 1xPBS 2 min. Then wash in 1x PBS for 2 min, 95% ethanol 1 min, 100% 1 min. Samples were then hybridized overnight at 55°C in 100 µl of a mixture containing 10% dextran sulfate, 2 mM vanadyl-ribonucleoside complex, 0.02% RNAse-free BSA, 40 µg *E. coli* tRNA, 2x SSC, 50% formamide, 30 ng of probe. After hybridization, samples are washed twice for 45 min at the appropriate stringency: 0.2x SSC, 55 °C, and rinsed twice in TBS. Digoxigenin-labeled probes were detected with sheep anti-digoxigenin antibodies (1/500) from Sigma-Aldrich (cat# 11214667001), and then with donkey anti-sheep antibodies conjugated to AlexaFluor647 (1/1000) from Thermo-Fisher Scientific (cat# A-21448). Slides are incubated overnight at 4°C with primary antibody, and then washed in washing buffer three times for 20 min at room temperature. Slides were incubated overnight at 4°C with secondary antibody, and then washed in washing buffer three times for 20 min at room temperature. For final mounting, samples were washed in 1X TBS, and mounted in slow-fade gold with DAPI (ThermoFisher Scientific, Inc.).

miRNA	Probe sequence	Probe T_m (°C)	Hybridization temperature (°C)	Probe concentration
Asp_miR2118	AAGGATTAGGTGGCATCGGGA/3Dig_N/	85	55	250 nM
Asp_miR2275	TGAGATGTTGGAGGAAACCGA/3Dig_N/	85	55	250 nM
Asp_24-nt PhasiRNA	TCCTATGTCGGTTCACAGTT/3Dig_N/	84	55	250 nM
Asp_IR_based 21nt-phasiRNA	TCTGAGTCCAACCAAGTGT/3Dig_N/	84	55	250 nM
Asp_nonIR_based 21nt-phasiRNA	GGCGTTCAAGTTGTTTAATGA/3Dig_N/	85	55	250 nM
Asp_24-nt phasiRNA precursor	TGGGACAATGAAACAACTCTA/3Dig_N/	82	55	250 nM
Lilium_miR2275	AGATATCAGAGGAAATTGA/3Dig_N/	79	55	250 nM
Lilium_inferred IR-based 24-nt phasiRNA	AGTCATGCTCAGAGAGTTAACA/3Dig_N/	84	55	250 nM
Lilium_inferred IR-based 24-nt phasiRNA precursor	TCACTAATTTTTACGCATGA/3Dig_N/	83	55	250 nM
Lilium_direct IR-based 24-nt phasiRNA	AGGCCGGAGGGAGTTATGTT/3Dig_N/	84	55	250 nM
Lilium_direct IR-based 24-nt phasiRNA precursor	AGTTTACTAGGATGACTCCTTCA/3Dig_N/	84	55	250 nM
Scrambled control	/5DigN/GTGTAACACGTCTATACGCCCA	87	55	250 nM

T_m, melting temperature
5DigN, 5’ Digoxigenin NHS Ester

Primers for amplifying DCLs,
LA-DCL3b For GAAGGAACTTCATGGGATGGT Rev GGATGCTGGAGCGTGATATT
And amplified with T7 promoter in vitro using T7 in vitro transcriptase (Roche).

2.18 Confocal microscopy

Confocal images were taken with *Zeiss LSM880 using a C-Apochromat 40X* (NA=1.3) oil immersion objective lens. For NBT-stained slides, blocks were excited at 458 nm and auto-fluorescence was detected using a 578 nm – 674 nm band pass detector. We also used transmitted light for generating DIC images. For Fluorescent *in situ* hybridization, images were taken under 633 nm excitation and emission 649-758 nm wavelength.

2.19 Real-Time qRT-PCR

Total RNA was extracted as described above, treated with DNase I (NEB, cat # M0303S), and then cleaned using the *RNA Clean and Concentrator-5* (Zymo Research, cat # R1015) columns. An aliquot containing 800 ng of clean total RNA was used for reverse transcription using the *SuperScript IV First-Strand Synthesis System* (Thermo Fisher Scientific, cat # 18091050). Then, the first-stranded RNA was 3x diluted and 1 µL was used in the qPCR reaction, for which was used the *SsoAdvanced Universal SYBR Green Supermix* (Bio-Rad, cat # 172-5271) for a 20 µL reaction. The qPCR runs were performed in the *CFX96 Real-Time PCR Detection System* (Bio-Rad) and the run condition was as follow: 95.0°C – 30 sec; 40 cycle of 95.0°C – 5 sec, 61.0°C - 30 sec; Melt curve 65.0°C to 95 with 0.5 increment, for 5 sec. The sequence of primers tested is listed below. Actin (AoAct-2, primer ID-1 29 and 30) was used as endogenous control.

Samples

Name	Description
Asparagus BM14-72	leaf
Asparagus BM14-181	<0.5 mm anthers (whole buds)
Asparagus BM14-182	0.5 - 1.0 mm anthers (whole buds)
Asparagus BM14-183	1.0 - 1.5 mm anthers (whole buds)
Asparagus BM14-184	0.5 - 1.0 mm anthers
Asparagus BM14-185	1.0 - 1.5 mm anthers

Primers

Primer ID-1	Primer ID-2	Sequence
19	AoDCL5-1F	TGA CTC TGC TCA TGT AAA CTA CG
20	AoDCL5-1R	ATT AGC CCA GGT CCC AGA TA
21	AoDCL5-2F	TAT CTG GGA CCT GGG CTA AT
22	AoDCL5-2R	GTT GCC TCT ATC AAG AGA ACA AAT C

23	AoDCL5-3F	ACA TCA TAC TGC GAA CCA TCT AC
24	AoDCL5-3R	GGC CAC CTT TCT CCA TCT TAA T
25	AoDCL5-4F	CTT CGA CCT CTG TCG AAT ACT T
26	AoDCL5-4R	GTT GAA ACC CAT CAC TCC ATT C
27	AoAct-1F	CCA AGG CCA ACA GAG AGA AA
28	AoAct-1R	GTA CGA CCA CTA GCG TAA AGA G
29	AoAct-2F	CTG GTA TTG CTG ACC GTA TGA G
30	AoAct-2R	CCA ATC CAG ACA CTG TAC TTC C
31	AoGAPDH-1F	CGA CAT TCT GTC AGG AGT ACA A
32	AoGAPDH-1R	CCT CCC AAG CAA TCC TCA TAT C
33	AoUBC2-1F	TGT GAC CCA AAT CCC AAC TC
34	AoUBC2-1R	CTC TGC TCC ACT ATC TCT CTC A

3. Supplemental References

Albert V a., Barbazuk WB, DePamphilis CW, Der JP, Leebens-Mack J, Ma H, Palmer JD, Rounsley S, Sankoff D, Schuster SC, et al. 2013. The Amborella Genome and the Evolution of Flowering Plants. *Science* **342**: 1241089.

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

Chellappan P, Xia J, Zhou X, Gao S, Zhang X, Coutino G, Vazquez F, Zhang W, Jin H. 2010. siRNAs from miRNA sites mediate DNA methylation of target genes. *Nucleic Acids Res* **38**: 6883–6894.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–1512.

Javelle M, Timmermans MCP. 2012. In situ localization of small RNAs in plants by using LNA probes. *Nat Protoc* **7**: 533–541.

Jeong D-H, Park S, Zhai J, Gurazada SGR, De Paoli E, Meyers BC, Green PJ. 2011. Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. *Plant Cell* **23**: 4185–4207.

Jeong D-H, Thatcher SR, Brown RSH, Zhai J, Park S, Rymarquis L a, Meyers BC, Green PJ. 2013. Comprehensive investigation of microRNAs enhanced by analysis of sequence variants, expression patterns, ARGONAUTE loading, and target cleavage. *Plant Physiol* **162**: 1225–45.

Kakrana A, Li P, Patel P, Hammond R, Anand D, Mathioni SM & Meyers BC. PHASIS: a computational suite for *de novo* discovery and characterization of phased, siRNA-generating loci and their miRNA triggers. *bioRxiv,* doi: 10.1101/158832.

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. 2007. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**: 345–349.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, Aiyer S, Valladares O, Yang J, Bambina S, Sabin LR, et al. 2012. Global analysis of RNA secondary structure in two metazoans. *Cell Rep* **1**: 69–82.

Mahalakshmi YV, Jagannadham MV, Pandit MW. 2000. Ribonuclease from cobra snake venom: purification by affinity chromatography and further characterization. *IUBMB Life* **49**: 309–316.

Mathioni SM, Kakrana A, Meyers BC. 2016. Characterization of plant small RNAs by next generation sequencing. In *Current Protocols in Plant Biology*, John Wiley & Sons, Inc. http://onlinelibrary.wiley.com/doi/10.1002/cppb.20043/abstract (Accessed March 26, 2017).

Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res* **34**: D731–D735.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205–217.

Patel P, Ramachandruni SD, Kakrana A, Nakano M, Meyers BC. 2016. miTRATA: a web-based tool for microRNA truncation and tailing analysis. *Bioinforma Oxf Engl* **32**: 450–452.

RNAcentral Consortium. 2015. RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res* **43**: D123–D129.

Sato K, Hamada M, Asai K, Mituyama T. 2009. CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res* **37**: W277-280.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562–78.

Vazquez F, Blevins T, Ailhas J, Boller T, Meins F. 2008a. Evolution of Arabidopsis MIR genes generates novel microRNA classes. *Nucleic Acids Res* **36**: 6429–6438.

Vazquez F, Blevins T, Ailhas J, Boller T, Meins F. 2008b. Evolution of Arabidopsis MIR genes generates novel microRNA classes. *Nucleic Acids Res* **36**: 6429–6438.

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**: e74.

Warner CA, Biedrzycki ML, Jacobs SS, Wisser RJ, Caplan JL, Sherrier DJ. 2014. An optical clearing technique for plant tissues allowing deep imaging and compatible with fluorescence microscopy. *Plant Physiol* **166**: 1684–1687.

Wu G. 2013. Plant microRNAs and development. *J Genet Genomics Yi Chuan Xue Bao* **40**: 217–30.

Wu L, Zhou H, Zhang Q, Zhang J, Ni F, Liu C, Qi Y. 2010. DNA methylation mediated by a microRNA pathway. *Mol Cell* **38**: 465–475.

Yi X, Zhang Z, Ling Y, Xu W, Su Z. 2015. PNRD: a plant non-coding RNA database. *Nucleic Acids Res* **43**: D982-989.

Zhai J, Arikit S, Simon SA, Kingham BF, Meyers BC. 2014. Rapid construction of parallel analysis of RNA end (PARE) libraries for Illumina sequencing. *Methods San Diego Calif* **67**: 84–90.

Zhai J, Zhang H, Arikit S, Huang K, Nan G-L, Walbot V, Meyers BC. 2015. Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proc Natl Acad Sci U S A* **112**: 3146–51.