Received: 6 September 2018 | Revised: 14 November 2018 | Accepted: 16 November 2018

DOI: 10.1002/pld3.101

ORIGINAL RESEARCH

Several phased siRNA annotation methods can frequently misidentify 24 nucleotide siRNA-dominated *PHAS* loci

Seth Polydore[1,2] | **Alice Lunardon**[2] | **Michael J. Axtell**[1,2]

[1]Genetics Ph.D. Program, Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania

[2]Department of Biology, The Pennsylvania State University, University Park, Pennsylvania

Correspondence
Michael J. Axtell, Department of Biology and Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA.
Email: mja18@psu.edu

Funding information
NSF, Grant/Award Number: 1339207; NIH Bioinformatics and Statistics CBIOS Predoctoral Training Program, Grant/Award Number: 1T32GM102057-0A1; the Henry W. Popp Graduate Assistantship Fund

Abstract

Small RNAs regulate key physiological functions in land plants. Small RNAs can be divided into two categories: microRNAs (miRNAs) and short interfering RNAs (siRNAs); siRNAs are further subdivided into transposon/repetitive region-localized heterochromatic siRNAs and phased siRNAs (phasiRNAs). PhasiRNAs are produced from the miRNA-mediated cleavage of a Pol II RNA transcript; the miRNA cleavage site provides a defined starting point from which phasiRNAs are produced in a distinctly phased pattern. 21–22 nucleotide (nt)-dominated phasiRNA-producing loci (*PHAS*) are well represented in all land plants to date. In contrast, 24 nt-dominated *PHAS* loci are known to be encoded only in monocots and are generally restricted to male reproductive tissues. Currently, only one miRNA (miR2275) is known to trigger the production of these 24 nt-dominated *PHAS* loci. In this study, we use stringent methodologies in order to examine whether or not 24 nt-dominated *PHAS* loci also exist in *Arabidopsis thaliana*. We find that highly expressed heterochromatic siRNAs were consistently misidentified as 24 nt-dominated *PHAS* loci using multiple *PHAS*-detecting algorithms. We also find that *MIR2275* is not found in *A. thaliana*, and it seems to have been lost in the last common ancestor of Brassicales. Altogether, our research highlights the potential issues with widely used *PHAS*-detecting algorithms which may lead to false positives when trying to annotate new *PHAS*, especially 24 nt-dominated loci.

KEYWORDS
heterochromatic siRNA, *MIR2275*, phased siRNA, small RNA

1 | INTRODUCTION

Small RNAs regulate key physiological functions in land plants, ranging from organogenesis (Boualem et al., 2008; Kutter, Schöb, Stadler, Meins, & Si-Ammour, 2007; Laufs, Peaucelle, Morin, & Traas, 2004; Williams, Grigg, Xie, Christensen, & Fletcher, 2005) to gametogenesis (Grant-Downton, Hafidh, Twell, & Dickinson, 2009). Three major protein families are involved in the biogenesis of small RNAs. The first family is the DICER-LIKE (DCL) protein family. Consisting of four paralogs in the *Arabidopsis thaliana* genome (Baulcombe, 2004; Chapman & Carrington, 2007), DCL proteins hydrolyze RNA precursors into 20–24 nt double-stranded RNA fragments (Millar & Waterhouse, 2005). These double-stranded RNA fragments are then loaded into ARGONAUTE (AGO) proteins, the second protein family, and one strand of the RNA is discarded (Ender & Meister, 2010). Upon Watson-crick binding to other RNA transcripts in the cell, the AGO/single-stranded RNA complex represses other RNA transcripts

Preprint: https://www.biorxiv.org/content/early/2018/11/15/409417
https://doi.org/10.1101/409417

(Baumberger & Baulcombe, 2005; Qi, Denli, & Hannon, 2005). Overall, 10 AGOs are encoded in the *A. thaliana* genome (Tolia & Joshua-Tor, 2007). RNA DEPENDENT RNA polymerases (RDRs) are the third family of proteins involved in the biogenesis of many small RNAs. RDRs convert single-stranded RNAs into double-stranded RNAs by synthesizing the complementary strand of the RNA molecule (Willmann, Endres, Cook, & Gregory, 2011). Six *RDRs* are encoded in the *A. thaliana* genome (Willmann et al., 2011).

Small RNAs can be divided into two major categories: microRNAs (miRNAs) which are precisely processed from single-stranded RNA with a hairpin-like secondary structure (Millar & Waterhouse, 2005), and short interfering RNAs (siRNAs), which are derived from double-stranded RNA precursors (Axtell, 2013). siRNAs are further divided into several different groups, including phased siRNAs (phasiRNAs) and heterochromatic siRNAs (hc-siRNAs). Predominantly 24 nts in length, hc-siRNAs function to repress transcription of deleterious genomic elements such as transposable elements or repetitive elements (Ahmed, Sarazin, Bowler, Colot, & Quesneville, 2011) and the promoters of certain genes (Baev et al., 2010) by reinforcing the presence of heterochromatin in targeted areas (Baulcombe, 2004; Sugiyama, Cam, Verdel, Moazed, & Grewal, 2005). Biogenesis of hc-siRNAs begins with transcription by the plant-specific, holoenzyme DNA DEPENDENT RNA POLYMERASE IV (Pol IV) (Onodera et al., 2005). The resulting transcript is then converted into double-stranded RNA by RDR2 and this double-stranded transcript is hydrolyzed by DCL3 (Matzke, Kanno, Daxinger, Huettel, & Matzke, 2009). phasiRNAs are derived from DNA-dependent RNA polymerase II (Pol II) transcripts that have been targeted by miRNAs (Fei, Xia, & Meyers, 2013). Upon miRNA-mediated hydrolysis, the RNA transcript is converted into double-stranded RNA by RDR6 (Cuperus et al., 2010). The resulting double-stranded RNA is then cleaved into 21 nt double-stranded RNA fragments by DCL4 (and less frequently DCL2) (Axtell, Jan, Rajagopalan, & Bartel, 2006).

21–22 nt phasiRNA-producing loci (*PHAS*) are clearly represented in all land plants that have been sequenced thus far (Fei et al., 2013; Zheng, Wang, Wu, Ding, & Fei, 2015). However, 24 nt-dominated *PHAS* loci are only currently described in rice (Song et al., 2012), maize (Zhai, Bischof, et al., 2015; Zhai, Zhang, et al., 2015), and other non-grass monocots (Kakrana et al., 2018). Much like 21 nt-dominated *PHAS*, the biogenesis of these 24 nt-dominated *PHAS* loci begins with the Pol II-dependent transcription of a single-stranded RNA precursor which is then targeted by miR2275 and hydrolyzed. To date, miR2275 is the only miRNA known to trigger the production of 24 nt-dominated phasiRNAs (Fei et al., 2013). The resulting RNA transcript is then converted into a double-stranded RNA molecule by RDR6 (Zhai, Zhang, et al., 2015). However, these phasiRNA precursors are then hydrolyzed by DCL5 (a DCL3 homolog sometimes called DCL3b) to produce 24 nt phasiRNAs (Fei et al., 2013).

Aside from the combination of their size and biogenesis patterns, 24 nt-dominated *PHAS* loci are distinct in various ways. These loci as well as their triggering miRNA, miR2275, are very specifically expressed in the tapetum during early meiosis and quickly recede in expression in other stages of male gametogenesis in rice and maize (Tamim et al., 2018). The AGO protein that loads these phasiRNAs is unknown; however, in maize, *AGO18b* expression levels match those of the 24 nt-dominated *PHAS* loci quite closely and is therefore the most likely candidate to load 24 nt phasiRNAs (Komiya et al., 2014; Zhang, Xia, Meyers, & Walbot, 2015). The targets of these 24 nt phasiRNAs are unknown, but they are apparently necessary for proper male gametogenesis (Ono et al., 2018). 24 nt-dominated *PHAS* loci were also described in the non-grass monocots asparagus, lily, and daylily (Kakrana et al., 2018). These phasiRNAs are produced from processing of inverted repeat (IR) RNAs, instead of the double-stranded RNA precursors observed in rice and maize (Kakrana et al., 2018). Although the 24 nt-dominated *PHAS* loci from non-grass monocots are still expressed most greatly in male reproductive tissue, in asparagus they are also expressed in female reproductive tissue (Kakrana et al., 2018).

We set out to search for evidence of 24 nt-*PHAS* loci in plants besides monocots. We searched for 24 nt *PHAS* loci in the *A. thaliana* genome using small RNA-seq data. Currently, several distinct algorithms are available to calculate the "phasing" of a sRNA-producing locus (Dotto et al., 2014; Guo, Qu, & Jin, 2015; Zheng, Wang, & Sunkar, 2014). In general, these algorithms calculate the number of reads that are "in-phase" against those that are "out-of-phase" in order to determine the likelihood that a particular locus truly produces phasiRNAs (Axtell, 2010). However, 24 nt-dominated siRNA loci are very numerous in angiosperms, and therefore are a potential source of false-positives during searches for *PHAS* loci. We therefore carefully examined *A. thaliana* 24 nt-dominated loci that consistently passed *PHAS*-locus detecting algorithms using multiple methods and find that they are likely just heterochromatic siRNAs (hc-siRNAs). We also use two other methods to examine the presence of 24 nt-dominated *PHAS* loci in the *A. thaliana* genome. We searched for *rdr6*-dependent, 24 nt-dominated loci and found 18 such loci. We also examined homology of the miR2275 which triggers 24 nt phasiRNA biogenesis in rice and maize but found that the Brassicales clade contains no potential homologs for this miRNA. Overall, our results suggest that there are no true 24 nt *PHAS* loci in *A. thaliana*. Furthermore, our analysis shows that existing phasing score algorithms to detect novel *PHAS* loci can lead to false positives.

2 | MATERIALS AND METHODS

2.1 | Finding potential *MIR2275* homologs in angiosperms

The Phytozome (ver 12.1)-curated angiosperm genome sequences were downloaded. The mature *Oryza sativa* miR2275a sequence was downloaded from miRBase (ver 21) (Griffiths-Jones, Saini, van Dongen, & Enright, 2008) and searched against all the other genomes using Bowtie v1.0 (Langmead, Trapnell, Pop, & Salzberg, 2009) allowing for two mismatches.

In order to determine the predicted secondary structures of the Bowtie results of interest, the sequences corresponding to the Bowtie result, plus 200 nucleotides upstream and downstream, were extracted from the genome. The secondary structures of the sequences were predicted using the mFOLD web server (Zuker, 2003) and visually inspected to determine if the sequence formed a hairpin structure consistent with accepted norms for miRNA biogenesis (Axtell & Meyers, 2018). The sequences were aligned using ClustalX ver. 2 with default parameters (Larkin et al., 2007).

For those species for which publicly available small RNA-seq data existed, we downloaded, trimmed, and merged the small RNA libraries. The merged libraries were collapsed to non-redundant reads and investigated using CD-HIT (ver 4.6.8) using the options -n 4, -d 0, and -g 1. The *O. sativa* miR2275a sequence was used as a query.

2.2 | Determination of potential 24 nt-dominated *PHAS* loci in different wild-type, inflorescence libraries

Wild-type biological triplicate small RNA libraries (GSE105262) (Polydore & Axtell, 2018) were merged and aligned against the *A. thaliana* (TAIR 10) genome using ShortStack (ver 3.8) (Johnson, Yeoh, Coruh, & Axtell, 2016) using 27 known *A. thaliana PHAS* loci as a query file (Supporting information Table S1). Three distinct *PHAS*-detecting algorithms (Dotto et al., 2014; Johnson et al., 2016; Zheng et al., 2014) were used to determine phase scores in the merged run. These scores were used as the basis to determine cutoffs for calling significantly phased loci (Supporting information Figures S3–S4). Phase scores of each known *PHAS* locus for each of the three algorithms in the eight wild-type libraries used is listed in Supporting information Dataset S4. Note that we did not use the multiple testing correction for *PHAS* loci *p*-values as done in Dotto et al., 2014 as we wished to test the three algorithms against each other, and the algorithms that yield phase scores could not be adjusted for multiple testing.

Wild-type and *rdr1-1/2-1/6-15* (*rdr1/2/6*) triple mutant small RNA libraries (GSE105262) (Polydore & Axtell, 2018) were aligned against the *A. thaliana* (TAIR 10) genome using ShortStack (ver 3.2) with option –pad 75 and option –min_cov 0.5 rpm (Supporting information Table S2). With these settings, small RNA loci are found as follows: All distinct genomic intervals containing one or more primary sRNA-seq alignments within 75 nts of each other were obtained, and then filtered to remove loci where the total sRNA-seq abundance with a locus was less than 0.5 reads per million. This produced a final set of distinct, non-overlapping small RNA loci. Differential expression to determine downregulated loci was performed as previously described (Polydore & Axtell, 2018). Downregulated, 24 nt-dominated small RNA loci were catalogued into a list. Eight wild-type, inflorescence small RNA libraries (Supporting information Table S2) were run individually against the *A. thaliana* (TAIR 10) genome utilizing ShortStack (ver 3.8.1) using the results of our wild-type and *rdr1/2/6* libraries run as a query file. The phase scores of loci corresponding to the down-regulated, 24 nt-dominated small RNAs loci were evaluated using the binary sequence alignment (BAM)-formatted alignments from each run. ShortStack and an in-house Python script were used to perform the three phase score calculations. For *Brassica rapa*, *Cucumis sativus*, *Phaseolus vulgaris*, and *Solanum tuberosum*, the previous Shortstack-merged small RNA alignments were analyzed in the same way.

For the four other species besides *A. thaliana*, we downloaded publicly available small RNA libraries (Supporting information Dataset S2) for *B. rapa*, *C. sativus*, *P. vulgaris*, and *S. tuberosum* and merged and aligned them against their respective genomes using ShortStack (ver 3.8.1) with default options. The genomes used were ver 1.0 for *B. rapa* (Wang et al., 2011), ver 1.0 for *P. vulgaris* (Schmutz et al., 2014), ver. 2 for *C. sativus* (Huang et al., 2009), and dm_v404 for *S. tuberosum* (Hardigan et al., 2016).

2.3 | AGO immunoprecipitation, genetic dependency, and properties of loci analyses

Calculating the lengths, proportion of multimapping reads, small RNA expression levels (in Reads Per Million (RPM)), determining the genetic dependencies, and the AGO enrichments of various loci were performed as previously described (Polydore & Axtell, 2018). In order to compare the properties of 24 nt-dominated loci that passed the *PHAS*-detection algorithms to those that did not, 10 subsets of 20 loci were randomly selected from the 24 nt-dominated loci that did not pass the *PHAS*-detection algorithms with replacement.

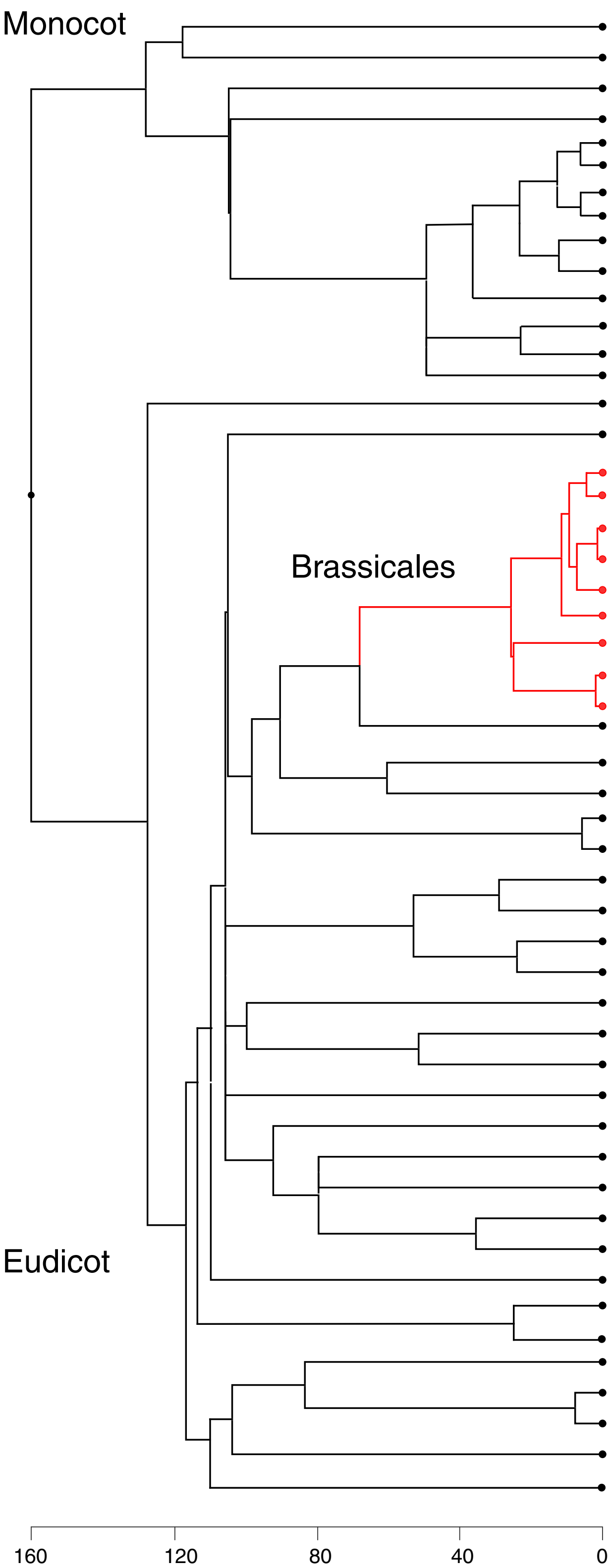
2.4 | Examining *PHAS*-test passing loci for potential miRNA targeting

Sequences corresponding to the *A. thaliana* loci that consistently passed the *PHAS*-detection algorithms plus 200 base pairs upstream and downs-stream were extracted from the *A. thaliana* genome**.** Mature miRNA sequences were downloaded from miRBase (ver 21) and aligned against these sequences using the Generic Small RNA-Transcriptome Aligner. A miRNA was considered to be potentially targeting a sequence if it aligned with an Allen et al. score of 3 or less.

3 | RESULTS

3.1 | miR2275 is not found in the *Brassicales* clade

Currently known 24 nt-dominated phasiRNA precursors are known to be targeted only by a single miRNA family, miR2275 (Song et al., 2008; Zhai, Zhang, et al., 2015). We examined all available angiosperm genomes on Phytozome (ver 12.1) for potential homologs of *MIR2275*. In monocots, all but *Brachypodium stacei*, *Spirodela polyrhiza*, and *Zostera marina* had potential miR2275 homologs (Figure 1; Supporting information Figure S1). Twelve eudicots had potential miR2275 homologs based on sequence similarity (Figure 1). We therefore interrogated small RNA libraries for these species to see if a mature miR2275 homolog was expressed. Five eudicots had evidence of mature miR2275 accumulation (Figure 1). All of the monocots for which we obtained small RNA-seq data expressed mature miR2275, except *Musa acuminata* (Figure 1).



Organism Name	Homolog(s)?	Mature miRNA Expressed?
Spirodela polyrhiza	No	–
Zostera marina	No	–
Musa acuminata	No	No
Ananas comosus	Yes	–
Panicum hallii	Yes	–
Panicum virgatum	Yes	–
Setaria viridis	Yes	–
Setaria italica	Yes	Yes
Zea mays	Yes	Yes
Sorghum bicolor	Yes	Yes
Oropetium thomaeum	Yes	–
B. distachyon	Yes	Yes
Brachypodium stacei	No	–
Oryza sativa	Yes	Yes
Aquilegia coerulea	Yes	–
Eucalyptus grandis	No	–
Capsella grandiflora	No	–
Capsella rubella	No	–
Arabidopsis halleri	No	–
Arabidopsis lyrata	No	–
Arabidopsis thaliana	No	No
Boechera stricta	No	–
Eutrema salsugineum	No	–
Brassica oleracea	No	No
Brassica rapa	No	No
Carica papaya	No	No
Gossypium raimondii	No	–
Theobroma cacao	No	–
Citrus clementina	Yes	Yes
Citrus sinensis	Yes	–
Trifolium pratense	No	–
Phaseolus vulgaris	No	No
Medicago truncatula	No	No
Glycine max	No	No
Fragaria vesca	Yes	Yes
Prunus persica	Yes	Yes
Malus domestica	Yes	Yes
Cucumis sativus	No	No
Linum usitatissimum	Yes	No
Ricinus communis	No	–
Manihot esculenta	Yes	No
Salix purpurea	Yes	–
Populus trichocarpa	Yes	Yes
Vitis vinifera	Yes	–
Kalanchoe laxiflora	No	–
Kalanchoe fedtschenkoi	No	–
Erythranthe guttata	No	–
Solanum tuberosum	No	No
Solanum lycopersicum	No	No
Daucus carota	Yes	No
A. hypochondriacus	No	–

🟦 No homolog found.
🟨 Either homolog found, but no mature miRNA verified or vice-versa.
🟩 Homolog found but no small RNA libraries are available.
■ Homolog and mature miRNA are both verified.

FIGURE 1 Conservation of *MIR2275*. Evidence for existence of *MIR2275* homologs in angiosperms. Phylogeny depicts estimated divergence times per TimeTree of Life (Kumar, Stecher, Suleski, & Hedges, 2017)

Alignment of the *MIR2275* loci in species for which potential homologs could be identified shows strong conservation of the mature miRNA and miRNA* sequences (Supporting information Figure S2), suggesting that these loci evolved from a common ancestor. Importantly, because of the high specificity of miR2275 expression in developing anthers (Tamim et al., 2018), it is possible that our analysis includes false negatives, especially in situations where no reproductive tissue small RNA libraries were available. Altogether, our data suggest that miR2275 is not found in *A. thaliana* and that this loss apparently occurred before the last common ancestor for Brassicales.

3.2 | Three *A. thaliana* hc-siRNA loci consistently pass *PHAS*-detecting algorithms

Although *A. thaliana* lacks miR2275, it is possible that 24 nt *PHAS* loci exist in *A. thaliana* and are triggered by a different small RNA. We reasoned that true 24 nt *PHAS* loci would be dependent on one or more of the well-described *A. thaliana RDR* genes: *RDR1*, *RDR2*, or *RDR6*. We thus examined a previously described differential expression analysis that identified *A. thaliana* small RNA loci that were downregulated in an *rdr1/rdr2/rdr6* triple mutant (Polydore & Axtell, 2018). The phase scores of *rdr1/rdr2/rdr6*-dependent, 24 nt-dominated loci were calculated in eight independent small RNA libraries using three different algorithms (Figure 2a). Reasonable cutoffs for *PHAS* loci detection were determined by examining the phase score distributions in the three merged wild-type libraries when well-known 21 nt *PHAS* loci were analyzed (Supporting information Figure S3). These cutoffs were consistent when a larger number of sRNA-seq libraries were examined (Supporting information Figure S4). Of the 31,750 loci examined, only three (Supporting information Figure S5; Dataset S1) passed the *PHAS* loci algorithms consistently in all 8 libraries examined (Figure 2a).

We were interested in determining why these three loci consistently pass the *PHAS*-detection algorithms. As phasiRNA precursors are known to be targeted by miRNAs, we predicted whether or not any known *A. thaliana* miRNAs could target these three loci. We were unable to find any obvious miRNA target sites at these loci. Although miRNA target sites were not apparent at these loci, it is possible that other siRNAs might target them and initiate siRNA phasing. We therefore attempted to determine the most common phase register at each of these loci in the eight different wild-type libraries (Supporting information Figure S6). If these loci are truly phased, then we would expect the same phase register to predominate in each library examined. While this was the case for *TAS2* (a positive control), none of the three 24 nt loci passing the *PHAS*-detection algorithms had consistent phase registers (Supporting information Figure S6). We also note that these three loci did not have the majority of their mapped reads falling into a single phase register in any of the libraries examined, as one would expect for a true *PHAS* locus (Supporting information Figure S6).

We then determined where in the genome these loci were encoded. Similar to hc-siRNA loci, these three loci primarily overlap with repeat- and transposable elements (Figure 2b). In comparison, many known *A. thaliana PHAS* loci are primarily derived from genes (Figure 2b). Rice and maize annotated 24 nt *PHAS* loci are derived from long noncoding RNAs that are encoded in regions of the genome that are devoid of protein-coding genes, transposable elements, or repeats (Song et al., 2008; Zhai, Bischof, et al., 2015; Zhai, Zhang, et al., 2015). We also note that sRNA accumulation from these three loci is downregulated in *nrpd1-3* (NRPD is the largest subunit of Pol IV), *nrpe* (NRPE is the largest subunit of Pol V), and *rdr2* (Figure 2c). hc-siRNAs have the same genetic requirements (Matzke et al., 2009), which strongly suggests that these three loci produce hc-siRNAs. Notably, these loci are not down-regulated in *dcl3* backgrounds. This is in line with past data that show that DCL4 and DCL2 can partially complement production of hc-siRNAs in *dcl3* backgrounds (Gasciolli, Mallory, Bartel, & Vaucheret, 2005).

The Pol IV transcripts from which hc-siRNAs are derived are around 26–50 nts in length, and accumulate to detectable levels in the *dcl2-1/3-1/4-2t* (*dcl234*) triple mutant (Ye et al., 2016; Zhai, Bischof, et al., 2015). We therefore analyzed the lengths of reads mapping to our three putative 24 nt *PHAS* loci in *dcl234* triple mutants. We found most reads were less than 40 nts long, indicating that these putative 24 nt *PHAS* loci are associated with short precursors similar to hc-siRNA loci (Figure 2d). In comparison, reads mapping to the known *TAS* loci had a wider range of sizes, ranging from 24 to 80 nts (Figure 2e). Overall, the size profile of the precursor RNAs further delineate these three loci from known *PHAS* loci.

Finally, we examined the AGO enrichments of small RNAs from these loci. Canonical hc-siRNAs are loaded into AGO4 in order to repress other loci transcriptionally (Mi et al., 2008). While sRNAs from our three putative 24 nt *PHAS* loci are not particularly enriched in AGO4-immunoprecipitation libraries, known 21 nt *PHAS* loci are quite depleted in the same dataset (Supporting information Figure S7). We also note that sRNAs from the three putative 24 nt *PHAS* loci are depleted in AGO1 immunoprecipitation libraries (Supporting information Figure S7), probably owing to the lack of 21 nt sRNAs, which AGO1 primarily loads, produced at these loci (Mi et al., 2008).

3.3 | The three *PHAS*-Test passing loci have distinct characteristics

hc-siRNAs are known to produce small RNAs in a very imprecise manner, unlike the largely precise processing of phasiRNAs (Axtell, 2013). We were interested in how three hc-siRNAs could consistently pass three different *PHAS*-detecting algorithms. One simple explanation is possible erroneous placement of ambiguously mapped reads. These multimapping reads could cause the *PHAS*-detecting algorithms to overestimate the number of “in phase” reads. We therefore determined the proportion of multimapping reads in each locus, but these loci have low proportion of multimapping reads compared to other 24 nt-dominated loci (Figure 3a). We note that these three loci are more highly expressed

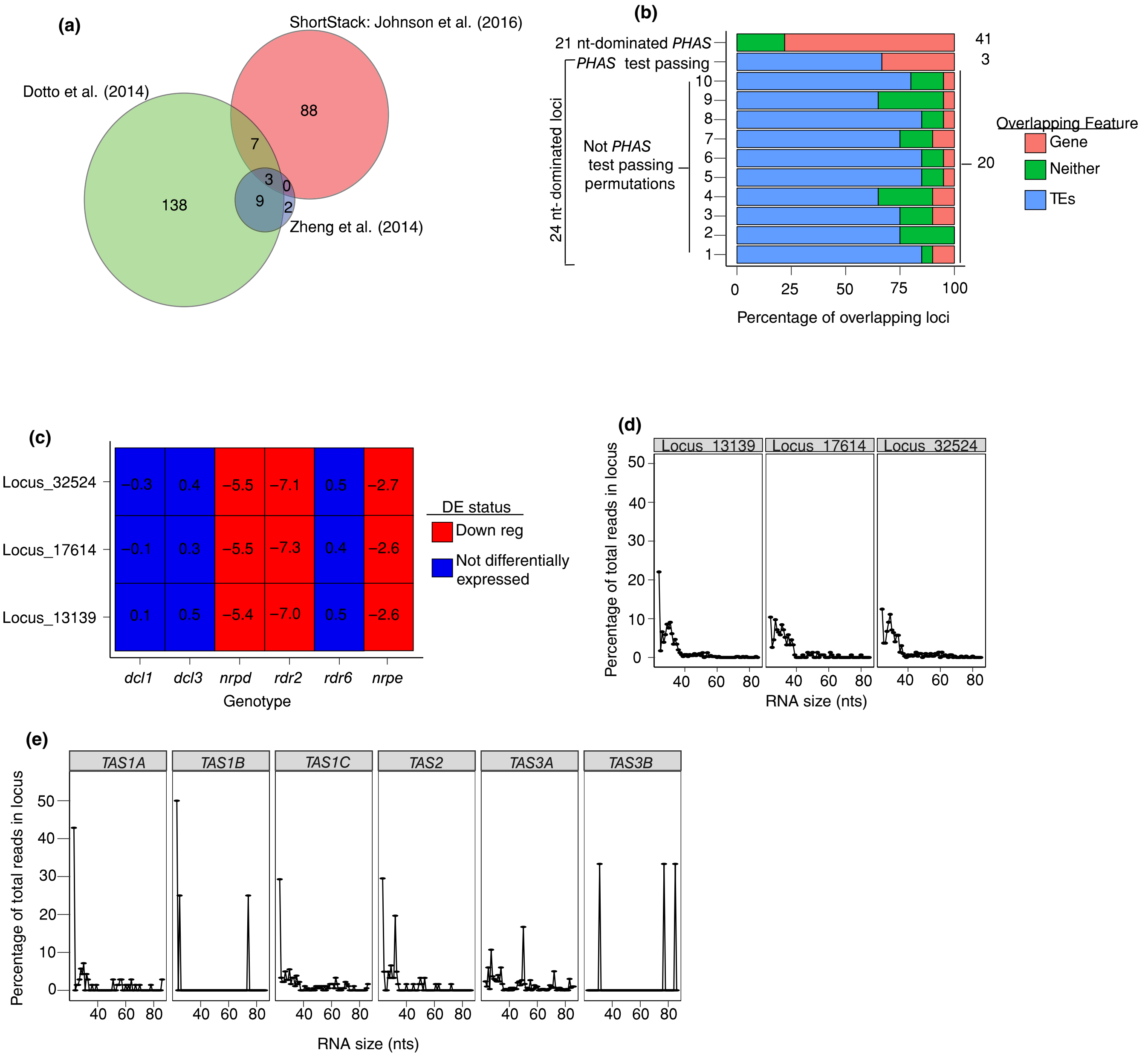


FIGURE 2 Properties of three *Arabidopsis thaliana* 24 nt-dominated small RNA loci that were called 'phased' by three different methods. (a) Venn diagram shows numbers of 24 nt-dominated loci that were called 'phased' by the indicated algorithms. (b) Percentage of the three *PHAS*-test passing loci overlapping either genes or transposable elements. The percentage is calculated as: (number of loci intersecting a feature/total number of loci in category)*100. The total number of loci in each category is given on the right. For 24 nt not-*PHAS* loci, 10 randomly selected cohorts of 20 loci each are shown. (c) Accumulation of the three *PHAS*-test passing loci in different genetic backgrounds. Numbers represent the ratio of small RNA accumulation in the indicated genotypes over that in corresponding wild-type library as computed by DESeq2. The differential expression status was determined via DESeq2 at an FDR of 0.1. (d) Percentage of short RNAs from *dcl2/dcl3/dcl4* triple mutant libraries by read length for the three *PHAS*-test passing loci. (e) Same as in Panel d, except for five known *A. thaliana TAS* loci

than most other 24 nt-dominated loci and even known *PHAS* loci in *A. thaliana* (Figure 3b). These three loci accumulate to nearly 1,000 RPM, while most other 24 nt-dominated loci and known *PHAS* loci only produce around 1 RPM (Figure 3b). Another interesting feature of these loci is their length. All three of these loci are around 10,000 nts in length, far greater than the 200–800 nt length of canonical *PHAS* loci and most other 24 nt-dominated loci (Figure 3c). These results indicate that these loci are simply highly expressed and particularly long hc-siRNA producing loci (Supporting information Figure S3).

3.4 | A small number of *A. thaliana rdr6*-dependent, 24 nt-dominated siRNA loci exist, but are not phased

Canonical monocot 24 nt-dominated *PHAS* loci are produced in part through the biochemical activity of RDR6 (Zhai, Zhang, et al., 2015).

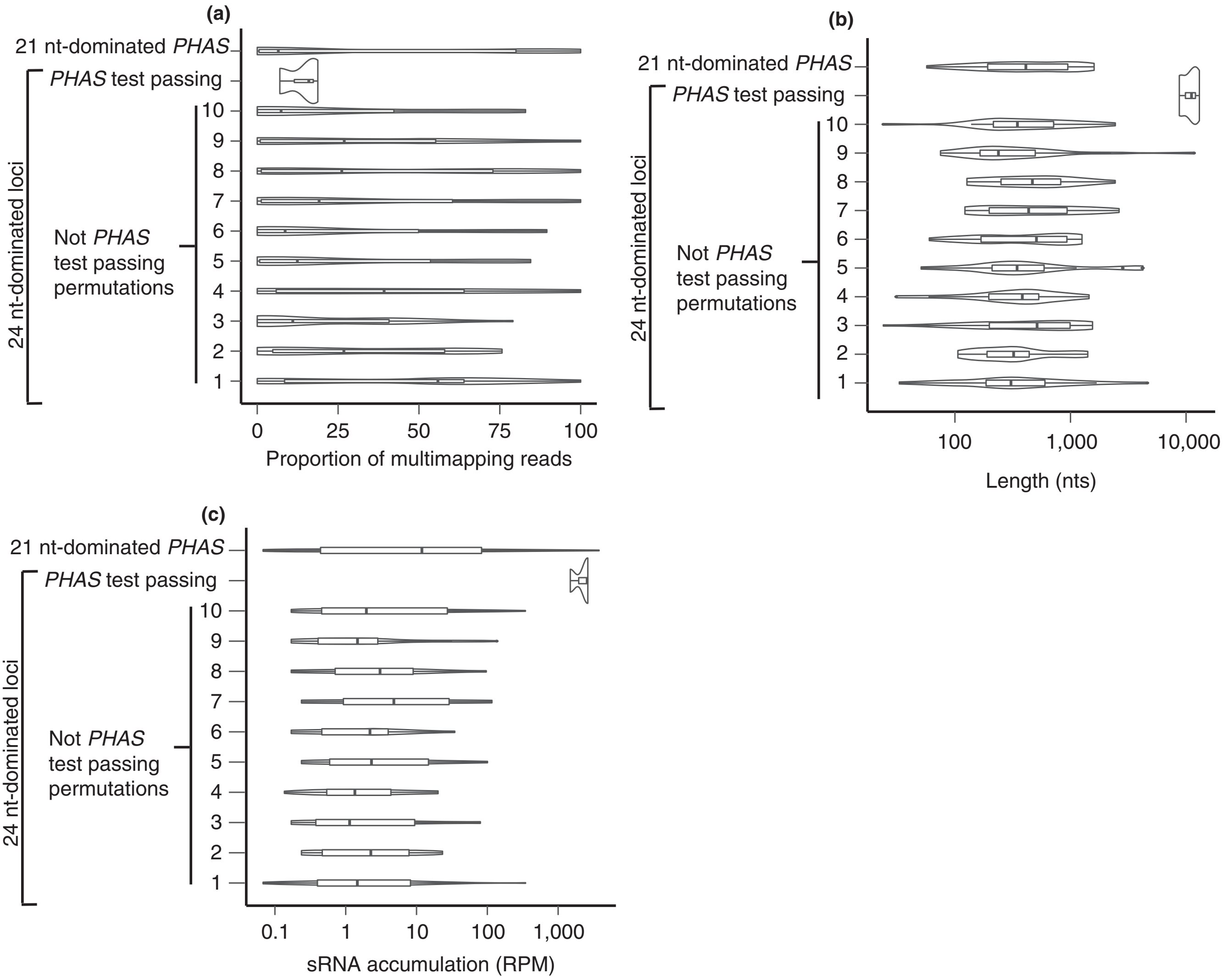
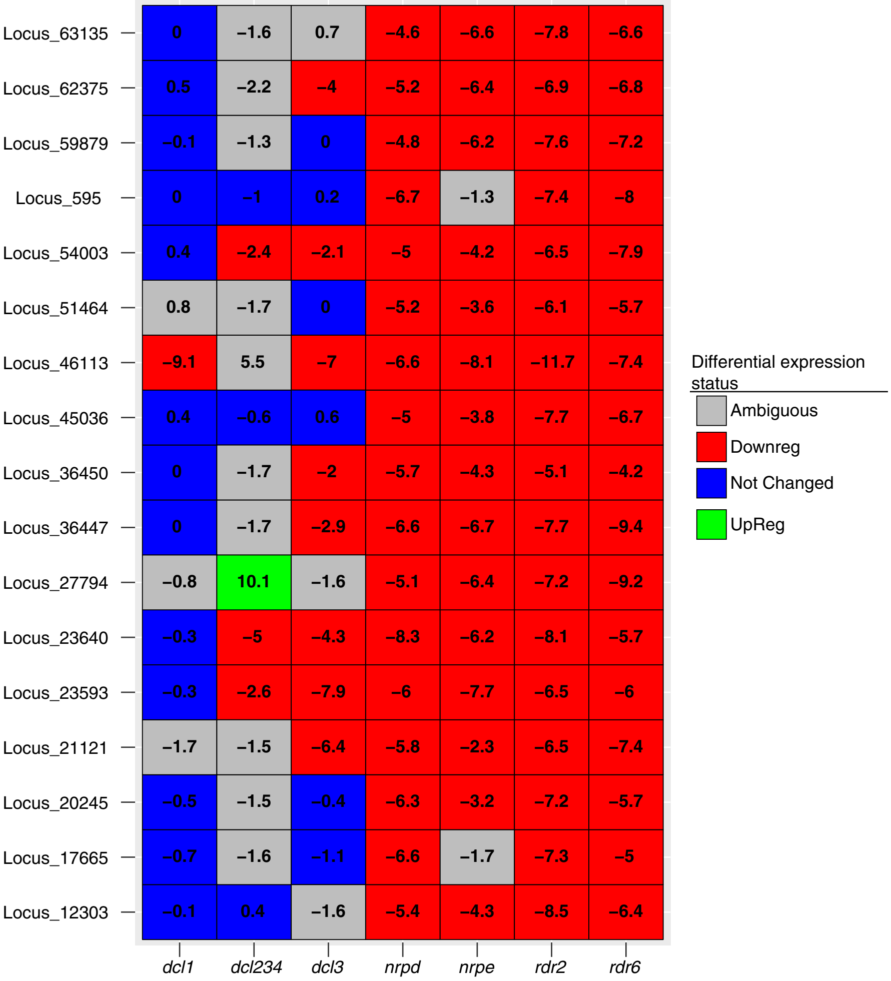


FIGURE 3 The three *PHAS*-test passing small RNA loci have distinct properties compared to other 24 nt-dominated loci. (a) The proportion of multimapped reads in three different types of small RNA loci. For 24 nt-dominated loci that were not called PHAS loci, 10 cohorts comprising 20 randomly selected loci were used as controls. The width of the density plot shows the frequency. The inset boxes show medians (horizontal lines), the 1st–3rd quartile range (boxes), other data out to 1.5 times the interquartile range (whiskers) and outliers (dots). (b) Same as panel a except showing small RNA accumulation. (c) Same as panel a except showing small RNA locus length

We therefore used an alternative strategy to identify possible 24 nt *PHAS* loci by first identifying *rdr6*-dependent, 24 nt-dominated siRNA loci. We found 18 such loci (Supporting information Figure S8; Dataset S1). None of these loci were consistently phased in the eight wild-type, inflorescence *A. thaliana* libraries (Supporting information Table S2) tested according to any of the three algorithms we used (Supporting information Dataset S3). We determined the genetic dependencies of these loci and found that these loci are down-regulated in *nrpd*, *nrpe*, and *rdr2* backgrounds (Figure 4). Furthermore, these loci either overlap transposable elements or are mostly found in otherwise intergenic regions (Supporting information Figure S9a). As these are typical features of hc-siRNAs, it is possible that hc-siRNAs simply erroneously placed at these loci. Again, we examined the proportion of ambiguously mapped reads in these *rdr6*-dependent, 24 nt-dominated loci compared to other 24 nt-dominated loci and found that these loci had low proportions of multimapping reads (Supporting information Figure S9b). Our result argues against this hypothesis.

rdr6-dependent, 24 nt-dominated loci are strongly expressed relative to other 24 nt-dominated loci (Supporting information Figure S9c). These *rdr6*-dependent loci have median expression level of nearly 100 RPM compared to the nearly 1 RPM median expression level of the other 24 nt-dominated loci (Supporting information Figure S9c). This result is similar to the three *A. thaliana* loci that passed our *PHAS*-detection algorithms (Figure 3b). However, the lengths of the *rdr6*-dependent, 24 nt-dominated loci are similar to other 24 nt-dominated loci (Supporting information Figure S9d). Overall, while do find evidence for a small number of *rdr6*-dependent, 24 nt-dominated siRNAs in *A. thaliana*, they are not phased,



	dcl1	dcl234	dcl3	nrpd	nrpe	rdr2	rdr6
Locus_63135	0	−1.6	0.7	−4.6	−6.6	−7.8	−6.6
Locus_62375	0.5	−2.2	−4	−5.2	−6.4	−6.9	−6.8
Locus_59879	−0.1	−1.3	0	−4.8	−6.2	−7.6	−7.2
Locus_595	0	−1	0.2	−6.7	−1.3	−7.4	−8
Locus_54003	0.4	−2.4	−2.1	−5	−4.2	−6.5	−7.9
Locus_51464	0.8	−1.7	0	−5.2	−3.6	−6.1	−5.7
Locus_46113	−9.1	5.5	−7	−6.6	−8.1	−11.7	−7.4
Locus_45036	0.4	−0.6	0.6	−5	−3.8	−7.7	−6.7
Locus_36450	0	−1.7	−2	−5.7	−4.3	−5.1	−4.2
Locus_36447	0	−1.7	−2.9	−6.6	−6.7	−7.7	−9.4
Locus_27794	−0.8	10.1	−1.6	−5.1	−6.4	−7.2	−9.2
Locus_23640	−0.3	−5	−4.3	−8.3	−6.2	−8.1	−5.7
Locus_23593	−0.3	−2.6	−7.9	−6	−7.7	−6.5	−6
Locus_21121	−1.7	−1.5	−6.4	−5.8	−2.3	−6.5	−7.4
Locus_20245	−0.5	−1.5	−0.4	−6.3	−3.2	−7.2	−5.7
Locus_17665	−0.7	−1.6	−1.1	−6.6	−1.7	−7.3	−5
Locus_12303	−0.1	0.4	−1.6	−5.4	−4.3	−8.5	−6.4

FIGURE 4 Accumulation of 24 nt-dominated, *rdr6*-dependent small RNA loci in different mutant backgrounds. Numbers represent the log2-transformed ratios of small RNA accumulation in the indicated genotypes over that in corresponding wild-type library as computed by DESeq2. The differential expression status was determined via DESeq2 at an FDR of 0.1

nor do they appear readily discernable from more typical hc-siRNA loci. Thus, we find no evidence of 24 nt-dominated *PHAS* loci in *A. thaliana*.

3.5 | Erroneous detection of 24 nt-dominated *PHAS* loci occurs in other land plant species

We were interested in determining if erroneous annotation of *PHAS* loci was unique to *A. thaliana* small RNAs or if these results could be replicated in other species. We therefore interrogated publicly available *B. rapa*, *C. sativus*, *P. vulgaris*, and *S. tuberosum* small RNA libraries (all eudicots) using the three algorithms, searching for putative 24 nt-dominated *PHAS* loci. We specifically chose these four species because they all lacked a potential *MIR2275* homolog (Figure 1). As miR2275 is the only miRNA known to trigger 24 nt-dominated phasiRNAs, any 24 nt-dominated loci called as *PHAS* loci in these species are likely false positives. Each species had 24 nt-dominated small RNAs that were misannotated as *PHAS* loci (Supporting information Figure S10). We first determined where in the genome the 24 nt-dominated loci are encoded. Like in *A. thaliana*, 24 nt-dominated loci that passed the *PHAS*-detection algorithm seem to come predominantly from transposable elements. This was true in all the species examined except for *P. vulgaris* (Figure 5a). Curiously, the *PHAS*-test passing loci had a slightly higher proportion of ambiguously mapped reads compared to other 24 nt-dominated loci in *B. rapa*, *P. vulgaris*, and *S. tuberosum* (Figure 5b). For *C. sativus*, the proportion of multimapping reads in *PHAS*-test passing 24 nt-dominated loci was substantially higher than other 24 nt-dominated loci (Figure 5b). It is still unlikely that multimapping reads contribute significantly to phasing at these loci as the proportion of ambiguously mapped to both types of 24 nt-dominated loci are similar in three of the four species tested (Figure 5b).

In the four species tested, the *PHAS*-test passing 24 nt-dominated loci had both significantly greater lengths (Figure 5c) and expression on average (Figure 5d). We observed this same trend in *A. thaliana* (Figure 3b,c), which suggests that long lengths and high expression levels of 24 nt-dominated loci are conducive to *PHAS* loci misannotations, even in other distantly related species.

4 | DISCUSSION

4.1 | *rdr6*-dependent, 24 nt-dominated loci in the *A. thaliana* genome

We found a handful of *rdr6*-dependent, 24 nt-dominated loci to be encoded in the *A. thaliana* genome. However, these loci have

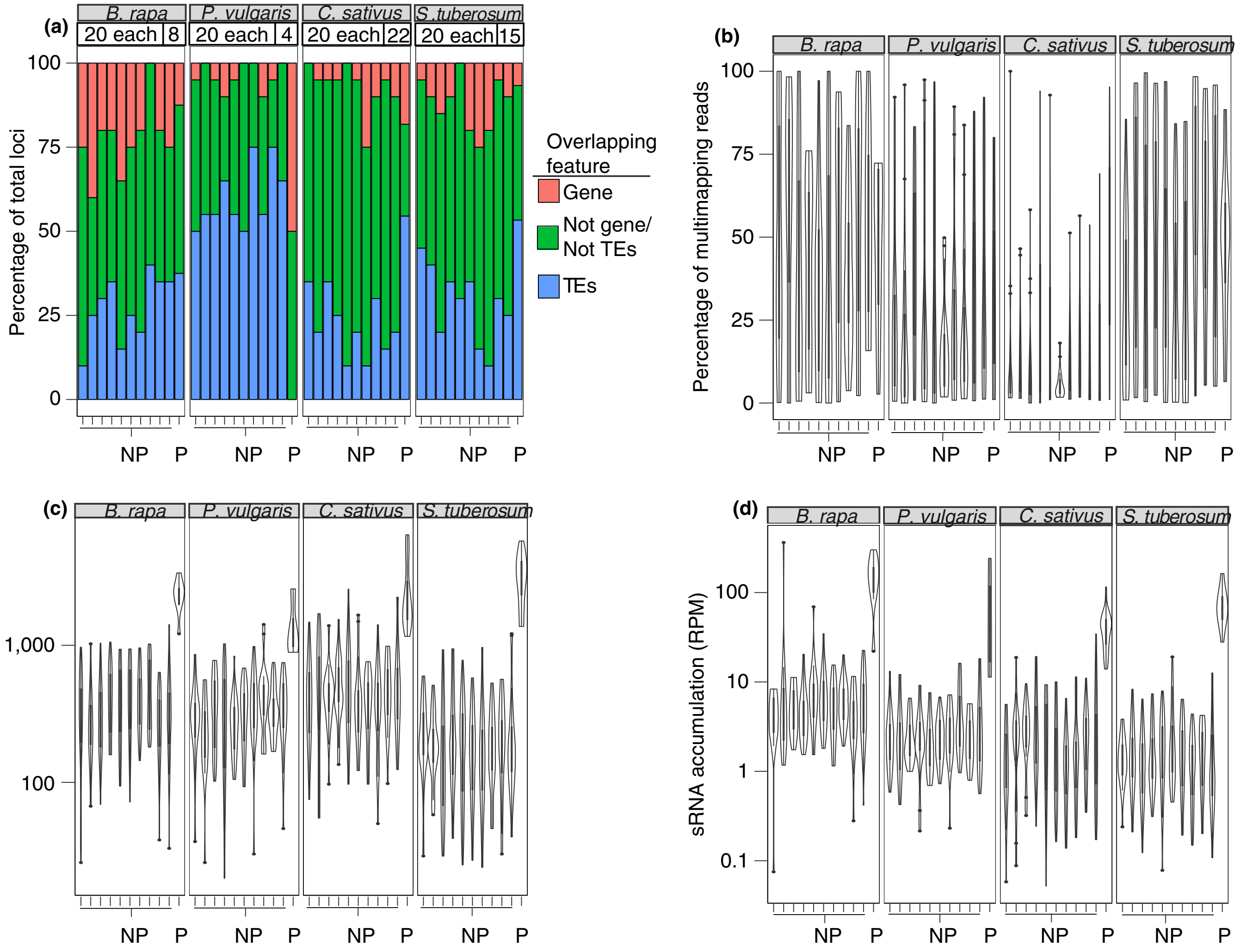


FIGURE 5 Four divergent species also contain 24 nt-dominated loci that passed the three *PHAS*-detecting algorithms. (a) Percentage of 24 nt-dominated Loci overlapping genes, transposable elements, or neither. The species name is shown in the top grey boxes. P: *PHAS*-test passing locus; NP: Not *PHAS*-test passing locus. For NP loci, 10 cohorts of 20 randomly selected loci each were used as negative controls. The number of loci in each category is shown above the bar graphs. (b) Same as panel a except showing the proportion of multimapping reads. (c) Same as panel a except showing length of the small RNA locus. (d) Same as panel a except showing small RNA accumulation

the same genetic dependencies as hc-siRNAs (Figure 4) and are frequently derived from transposable elements (Supporting information Figure S9a). The only other known *rdr6*-dependent, transposon-overlapping small RNA loci encoded in *A. thaliana* are epigenetically activated short interfering RNAs (easiRNAs) (Creasey et al., 2014). easiRNAs are derived from transcriptionally active transposable elements that are hypothesized to be targeted and cleaved by miRNAs (Creasey et al., 2014). Through the biochemical actions of RDR6 and DCL4, the miRNA cleavage product is converted into 21–22 nt double-stranded RNA molecules (Creasey et al., 2014). Furthermore, these easiRNAs are thought to direct initial repression of transposons (He, Yang, Wu, & Zheng, 2015). It is possible that these *rdr6*-dependent, 24 nt-dominated loci could represent a transitory stage of hc-siRNA targeting, in which the genetic machinery of hc-siRNA are used but with a remaining initial dependency on RDR6.

4.2 | Possible losses of the *MIR2275*-generated phasing in eudicots

We examined all Phytozome (ver 12.1)-curated angiosperm genomes (Goodstein et al., 2012) for the presence of *MIR2275* homologs. *MIR2275* shows remarkable conservation in monocots (Figure 1), showing absence only in *Musa accuminata* and *Brachypodium distachyon*. This could be because these are aquatic plants with divergent morphology, and thus there is relaxed selective pressure for *MIR2275*. Another explanation is that the genome assemblies for these species could be incomplete. The lack of

MIR2275 in eudicots was more extensive (Figure 1), with no plants in the Brassicales clade containing a verified *MIR2275* homolog (Figure 1). Interestingly, *Aquilegia coerulea*, a basal eudicot (Sharma & Kramer, 2014), contained a *MIR2275* homolog (Figure 1; Supporting information Figure S1) suggesting that the last common ancestor for eudicots may have contained *MIR2275*, and that the lack of detected putative *MIR2275* homologs in many eudicot plant species could be due to loss of *MIR2275*.

We only examined whether or not a possible *MIR2275* homolog could be detected by homology, if its predicted secondary structure is conducive to *MIRNA* processing (Supporting information Figure S1), and in some species whether or not the putative *MIR2275* homolog was expressed in available small RNA libraries (Figure 1). We did not, however, determine if these species produce true 24 nt-dominated phasiRNAs. It is possible that the potential *MIR2275* homolog in these species is an orphaned *MIRNA* or has taken on a new function. Furthermore, although some putative *MIR2275* homologs had *MIRNA*-like predicted hairpins, we note that they contain mismatches in the stem of the secondary structure which may hinder miRNA biogenesis (Supporting information Figure S1). Further research is necessary to determine if these species truly encode 24 nt-dominated *PHAS* loci.

4.3 | The difficulties of annotating 24 nt-dominated *PHAS* loci

The discovery of 24 nt-dominated *PHAS* loci in maize (Zhai, Bischof, et al., 2015; Zhai, Zhang, et al., 2015), rice (Song et al., 2008), asparagus, daylily, and lily (Kakrana et al., 2018) opened up the possibility that these loci exist in other species, even distantly related eudicots. However, 24 nt-dominated hc-siRNA loci are widespread in land plant genomes (Ghildiyal & Zamore, 2009). This is true even in *A. thaliana*, although 24 nt sRNAs are thought to repress transposable elements (Matzke et al., 2009) and only about 20–30% of the *A. thaliana* genome consists of transposon/repetitive elements (Barakat, Matassi, & Bernardi, 1998). In contrast, other plant genomes consist of as much as 80% transposons/repetitive elements (Springer et al., 2009). The sheer number of 24 nt-dominated sRNA loci could mean that several of them could meet various annotation criteria simply by chance; this is something we have previously noted to occur in natural antisense siRNA annotation (Polydore & Axtell, 2018).

The supposed 24 nt-dominated *PHAS* loci examined in this study consistently showed higher levels of expression than other 24 nt-dominated sRNA loci and derived from very long loci. This is not particularly surprising because the accumulation of reads (in- and out-of-phase) is a factor in most *PHAS*-detection algorithms (Supporting information Figure S3). However, it seems that particularly highly expressed 24 nt-dominated small RNAs are able to consistently pass *PHAS*-detection algorithms because of this. It's possible that the sheer number of reads produced at these loci means that these loci produce enough “in-phase” reads by chance to score highly in *PHAS*-detecting algorithms.

4.4 | Annotating 24 nt-dominated *PHAS* loci in the future

Our study demonstrates that when annotating novel 24 nt-dominated *PHAS*, more than utilizing *PHAS*-detecting algorithms is necessary for robust annotation. First, examining different available mutants, especially of genes involved in small RNA biogenesis, can be critical in determining type of small RNA in question (Figure 2c). All types of phasiRNAs are known to be reliant on the biochemical activity of RDR6 and Pol II (Cuperus et al., 2010; Song et al., 2008; Zhai, Bischof, et al., 2015; Zhai, Zhang, et al., 2015). While it is entirely possible that non-canonical phasiRNAs that are reliant on RDR2 and Pol IV may be described eventually, such a study must verify with robust methodologies that the *PHAS* loci are not false positives due to the sheer number of Pol IV/RDR2 –dependent reads produced in the land plant genome.

We searched for *MIR2275* homologs because miR2275 is the only miRNA known to trigger 24 nt phasiRNAs. If a species lacks a *MIR2275* homolog, then one should be skeptical of any 24 nt-dominated small RNA locus that is annotated as *PHAS* locus in that specie. An organism producing a mature miR2275 small RNA homolog is not sufficient to show that 24 nt-dominated *PHAS* loci are produced in that species. One should also take care to ascertain if any 24 nt-dominated small RNA locus called as *PHAS* loci also contains a miR2275 target site that is “in phase” with the phasiRNAs produced from the transcript. miR2275 is also very specifically expressed in the tapetum of male floral tissue. In asparagus, 24 nt-dominated phasiRNAs have also been shown to be expressed in female floral tissue. However, 24 nt-dominated phasiRNAs have not been demonstrated to be expressed outside of reproductive tissue as of yet. Therefore, any 24 nt-dominated *PHAS* loci annotated in libraries not produced from reproductive tissues are more likely to be mis-annotations.

Despite the fact that several small RNA loci in *A. thaliana* are able to consistently pass the *PHAS*-detecting algorithms in different libraries, reproducibility among distinct small RNA libraries is of utmost importance. It could also be useful to employ several different *PHAS*-test algorithms as well. In *A. thaliana*, three loci were able to consistently pass our *PHAS*-detecting algorithms (Supporting information Figure S5), but the number of loci that each algorithm detected on its own was generally much higher (Figure 2a). Had we not used stringent *PHAS*-detecting methods, it would've been plausible to assume we had found 24 nt-dominated *PHAS* loci in *A. thaliana* based on the number of loci that consistently pass alone. Utilizing multiple small RNA libraries should become easier to accomplish as more and more libraries in different treatments/genotypes become available for socioeconomically relevant species. Certain *PHAS*-detecting programs, such as *PHASIS* (Kakrana et al., 2017), automatically evaluate potential small RNA loci in different libraries individually before merging the results of each library.

It is not always possible to determine the small RNA that targets the phasiRNA precursor transcript. Indeed, several reads may be predicted to target a certain transcript by chance due to the sheer number of unique reads in a small RNA library. However, due to hydrolysis of the precursor transcript following small RNA-mediated targeting, phasiRNAs have well-defined termini from which they are produced (Axtell et al., 2006; Cuperus et al., 2010). Therefore, a true *PHAS* locus should have a large proportion of its reads reproducibly falling into a particular phase register. This is what we observed with well-characterized *PHAS* locus, *TAS2*, but not with the three loci that pass our rigorous *PHAS*-annotation regime in *A. thaliana* (Supporting information Figure S6).

As 21–22 nt-dominated *PHAS* loci are far more common in land plants, especially outside of the monocots, one can conservatively limit their discovery of new *PHAS* loci to 21–22 nt-dominated small RNAs and not employ as rigorous methodology as the ones used in this study. However, one should still employ post *PHAS*-discovery quality controls to ensure these 21–22 nt-dominated *PHAS* loci are genuine (such as determining if the predominant phase registers are reproducibly dominant at these loci (Supporting information Figure S6)). 24 nt-dominated *PHAS* loci on the other hand seem to be have undergone loss in many angiosperms. However, even in the species in which they are conserved, they seem to play very specific, reproductive-associated roles as evidenced by their expression patterns. Great caution should be used for annotating 24 nt-dominated *PHAS* loci in the future.

ACKNOWLEDGMENTS

This research was funded by NSF Award 1339207 to MJA, the NIH Bioinformatics and Statistics (CBIOS) Predoctoral Training Program (1T32GM102057- 0A1) to SP, and the Henry W. Popp Graduate Assistantship Fund to SP.

AUTHOR CONTRIBUTIONS

SP conceived of the project. AL aligned all the small RNA data other than the *A. thaliana* small RNA libraries. SP performed all the other experiments and analyses. MJA and SP wrote the manuscript.

REFERENCES

Ahmed, I., Sarazin, A., Bowler, C., Colot, V., & Quesneville, H. (2011). Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. *Nucleic Acids Research*, *39*(16), 6919–6931. https://doi.org/10.1093/nar/gkr324

Axtell, M. J. (2010). A method to discover phased siRNA loci. In C. M., Blake & J. G., Pamela (Eds.), *Plant MicroRNAs* (pp. 59–70). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-60327-005-2_5

Axtell, M. J. (2013). Classification and comparison of small RNAs from plants. *Annual Review of Plant Biology*, *64*(1), 137–159. https://doi.org/10.1146/annurev-arplant-050312-120043

Axtell, M. J., Jan, C., Rajagopalan, R., & Bartel, D. P. (2006). A two-hit trigger for siRNA biogenesis in plants. *Cell*, *127*(3), 565–577. https://doi.org/10.1016/j.cell.2006.09.032

Baev, V., Naydenov, M., Apostolova, E., Ivanova, D., Doncheva, S., Minkov, I., & Yahubyan, G. (2010). Identification of RNA-dependent DNA-methylation regulated promoters in Arabidopsis. *Plant Physiology and Biochemistry*, *48*(6), 393–400. https://doi.org/10.1016/j.plaphy.2010.03.013

Barakat, A., Matassi, G., & Bernardi, G. (1998). Distribution of genes in the genome of Arabidopsis thaliana and its implications for the genome organization of plants. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(17), 10044–10049.

Baulcombe, D. (2004). RNA silencing in plants. *Nature*, *431*(7006), 356–363. https://doi.org/10.1038/nature02874

Baumberger, N., & Baulcombe, D. C. (2005). Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proceedings of the National Academy of Sciences, USA*, *102*(33), 11928–11933. https://doi.org/10.1073/pnas.0505461102

Boualem, A., Laporte, P., Jovanovic, M., Laffont, C., Plet, J., Combier, J.-P., & Frugier, F. (2008). MicroRNA166 controls root and nodule development in Medicago truncatula. *Plant Journal: For Cell and Molecular Biology*, *54*(5), 876–887. https://doi.org/10.1111/j.1365-313X.2008.03448.x

Chapman, E. J., & Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. *Nature Reviews Genetics*, *8*(11), 884–896. https://doi.org/10.1038/nrg2179

Creasey, K. M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B. C., & Martienssen, R. A. (2014). miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. *Nature*, *508*(7496), 411–415. https://doi.org/10.1038/nature13069

Cuperus, J. T., Carbonell, A., Fahlgren, N., Garcia-Ruiz, H., Burke, R. T., Takeda, A., & Carrington, J. C. (2010). Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in *Arabidopsis*. *Nature Structural & Molecular Biology*, *17*(8), 997–1003. https://doi.org/10.1038/nsmb.1866

Dotto, M. C., Petsch, K. A., Aukerman, M. J., Beatty, M., Hammell, M., & Timmermans, M. C. P. (2014). Genome-wide analysis of leafbladeless1-regulated and phased small RNAs underscores the importance of the TAS3 ta-siRNA pathway to maize development. *PLOS Genetics*, *10*(12), e1004826. https://doi.org/10.1371/journal.pgen.1004826

Ender, C., & Meister, G. (2010). Argonaute proteins at a glance. *Journal of Cell Science*, *123*(11), 1819–1823. https://doi.org/10.1242/jcs.055210

Fei, Q., Xia, R., & Meyers, B. C. (2013). Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell Online*, *25*(7), 2400–2415. https://doi.org/10.1105/tpc.113.114652

Gasciolli, V., Mallory, A. C., Bartel, D. P., & Vaucheret, H. (2005). Partially redundant functions of Arabidopsis DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. *Current Biology*, *15*(16), 1494–1500. https://doi.org/10.1016/j.cub.2005.07.024

Ghildiyal, M., & Zamore, P. D. (2009). Small silencing RNAs: An expanding universe. *Nature Reviews. Genetics*, *10*(2), 94–108. https://doi.org/10.1038/nrg2504

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., … Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(Database issue), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Grant-Downton, R., Hafidh, S., Twell, D., & Dickinson, H. G. (2009). Small RNA pathways are present and functional in the angiosperm male gametophyte. *Molecular Plant*, *2*(3), 500–512. https://doi.org/10.1093/mp/ssp003

Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2008). miRBase: Tools for microRNA genomics. *Nucleic Acids Research*, *36*(suppl_1), D154–D158. https://doi.org/10.1093/nar/gkm952

Guo, Q., Qu, X., & Jin, W. (2015). PhaseTank: Genome-wide computational identification of phasiRNAs and their regulatory cascades. *Bioinformatics*, *31*(2), 284–286. https://doi.org/10.1093/bioinformatics/btu628

Hardigan, M. A., Crisovan, E., Hamiltion, J. P., Kim, J., Laimbeer, P., Leisner, C. P., … Buell, C. R. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *Plant Cell*, *28*, 388–405. https://doi.org/10.1105/tpc.15.00538

He, H., Yang, T., Wu, W., & Zheng, B. (2015). Small RNAs in pollen. *Science China Life Sciences*, *58*(3), 246–252. https://doi.org/10.1007/s11427-015-4800-0

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., & Li, S. (2009). The genome of the cucumber, *Cucumis sativus* L. *Nature Genetics*, *41*(12), 1275–1281. https://doi.org/10.1038/ng.475

Johnson, N. R., Yeoh, J. M., Coruh, C., & Axtell, M. J. (2016). Improved placement of multi-mapping small RNAs. *G3: Genes, Genomes, Genetics*, *6*(7), 2103–2111. https://doi.org/10.1534/g3.116.030452

Kakrana, A., Li, P., Patel, P., Hammond, R., Anand, D., Mathioni, S., & Meyers, B. (2017). PHASIS: a computational suite for de novo discovery and characterization of phased, siRNA-generating loci and their miRNA triggers. *bioRxiv*, 158832. https://doi.org/10.1101/158832

Kakrana, A., Mathioni, S. M., Huang, K., Hammond, R., Vandivier, L., Patel, P., … Meyers, B. C. (2018). Plant 24-nt reproductive phasiRNAs from intramolecular duplex mRNAs in diverse monocots. *Genome Research*, *28*, 1333–1344. https://doi.org/10.1101/gr.228163.117

Komiya, R., Ohyanagi, H., Niihama, M., Watanabe, T., Nakano, M., Kurata, N., & Nonomura, K.-I. (2014). Rice germline-specific Argonaute MEL1 protein binds to phasiRNAs generated from more than 700 lincRNAs. *Plant Journal*, *78*(3), 385–397. https://doi.org/10.1111/tpj.12483

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Kutter, C., Schöb, H., Stadler, M., Meins, F., & Si-Ammour, A. (2007). MicroRNA-mediated regulation of stomatal development in Arabidopsis. *Plant Cell*, *19*(8), 2417–2429. https://doi.org/10.1105/tpc.107.050377

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Laufs, P., Peaucelle, A., Morin, H., & Traas, J. (2004). MicroRNA regulation of the CUC genes is required for boundary size control in Arabidopsis meristems. *Development*, *131*(17), 4311–4322. https://doi.org/10.1242/dev.01320

Matzke, M., Kanno, T., Daxinger, L., Huettel, B., & Matzke, A. J. (2009). RNA-mediated chromatin-based silencing in plants. *Current Opinion in Cell Biology*, *21*(3), 367–376. https://doi.org/10.1016/j.ceb.2009.01.025

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., & Qi, Y. (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5′ terminal nucleotide. *Cell*, *133*(1), 116–127. https://doi.org/10.1016/j.cell.2008.02.034

Millar, A. A., & Waterhouse, P. M. (2005). Plant and animal microRNAs: Similarities and differences. *Functional & Integrative Genomics*, *5*(3), 129–135. https://doi.org/10.1007/s10142-005-0145-2

Ono, S., Liu, H., Tsuda, K., Fukai, E., Tanaka, K., Sasaki, T., & Nonomura, K.-I. (2018). EAT1 transcription factor, a non-cell-autonomous regulator of pollen production, activates meiotic small RNA biogenesis in rice anther tapetum. *PLOS Genetics*, *14*(2), e1007238. https://doi.org/10.1371/journal.pgen.1007238

Onodera, Y., Haag, J. R., Ream, T., Nunes, P. C., Pontes, O., & Pikaard, C. S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell*, *120*(5), 613–622. https://doi.org/10.1016/j.cell.2005.02.007

Polydore, S., & Axtell, M. J. (2018). Analysis of RDR1/RDR2/RDR6-independent small RNAs in Arabidopsis thaliana improves MIRNA annotations and reveals unexplained types of short interfering RNA loci. *Plant Journal: For Cell and Molecular Biology*, *94*(6), 1051–1063. https://doi.org/10.1111/tpj.13919

Qi, Y., Denli, A. M., & Hannon, G. J. (2005). Biochemical specialization within Arabidopsis RNA silencing pathways. *Molecular Cell*, *19*(3), 421–428. https://doi.org/10.1016/j.molcel.2005.06.014

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., & Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics*, *46*(7), 707–713. https://doi.org/10.1038/ng.3008

Sharma, B., & Kramer, E. M. (2014). The MADS-box gene family of the basal eudicot and hybrid aquilegia coerulea "origami" (Ranunculaceae). *Annals of the Missouri Botanical Garden*, *99*(3), 313–322.

Song, X., Li, P., Zhai, J., Zhou, M., Ma, L., Liu, B., & Cao, X. (2012). Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. *Plant Journal*, *69*(3), 462–474. https://doi.org/10.1111/j.1365-313X.2011.04805.x

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., & Schnable, P. S. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLOS Genetics*, *5*(11), e1000734. https://doi.org/10.1371/journal.pgen.1000734

Sugiyama, T., Cam, H., Verdel, A., Moazed, D., & Grewal, S. I. S. (2005). RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(1), 152. https://doi.org/10.1073/pnas.0407641102

Tamim, S., Cai, Z., Mathioni, S., Zhai, J., Teng, C., Zhang, Q., & Meyers, B. C. (2018). Cis-directed cleavage and nonstoichiometric abundances of 21-nt reproductive phasiRNAs in grasses. *bioRxiv*, 243907. https://doi.org/10.1101/243907

Tolia, N. H., & Joshua-Tor, L. (2007). Slicer and the Argonautes. *Nature Chemical Biology*, *3*(1), 36–43. https://doi.org/10.1038/nchembio848

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., & Zhang, Z. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*, *43*(10), 1035–1039. https://doi.org/10.1038/ng.919

Williams, L., Grigg, S. P., Xie, M., Christensen, S., & Fletcher, J. C. (2005). Regulation of Arabidopsis shoot apical meristem and lateral organ formation by microRNA miR166g and its AtHD-ZIP target genes. *Development*, *132*(16), 3657–3668. https://doi.org/10.1242/dev.01942

Willmann, M. R., Endres, M. W., Cook, R. T., & Gregory, B. D. (2011). The functions of RNA-dependent RNA polymerases in Arabidopsis. *Arabidopsis Book*, *9*, e0146. https://doi.org/10.1199/tab.0146

Ye, R., Chen, Z., Lian, B., Rowley, M. J., Xia, N., Chai, J., & Qi, Y. (2016). A dicer-independent route for biogenesis of siRNAs that direct DNA methylation in Arabidopsis. *Molecular Cell*, *61*(2), 222–235. https://doi.org/10.1016/j.molcel.2015.11.015

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T., Teng, C., & Jacobsen, S. E. (2015). One precursor one siRNA model for Pol IV- dependent siRNAs biogenesis. *Cell*, *163*(2), 445–455. https://doi.org/10.1016/j.cell.2015.09.032

Zhai, J., Zhang, H., Arikit, S., Huang, K., Nan, G.-L., Walbot, V., & Meyers, B. C. (2015). Spatiotemporally dynamic, cell-type–dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proceedings of the National Academy of Sciences*, *USA*, *112*(10), 3146–3151. https://doi.org/10.1073/pnas.1418918112

Zhang, H., Xia, R., Meyers, B. C., & Walbot, V. (2015). Evolution, functions, and mysteries of plant ARGONAUTE proteins. *Current Opinion in Plant Biology*, *27*, 84–90. https://doi.org/10.1016/j.pbi.2015.06.011

Zheng, Yun, Wang, S., & Sunkar, R. (2014). Genome-wide discovery and analysis of phased small interfering RNAs in Chinese sacred lotus. *PLoS ONE*, *9*(12), e113790. https://doi.org/10.1371/journal.pone.0113790

Zheng, Yi, Wang, Y., Wu, J., Ding, B., & Fei, Z. (2015). A dynamic evolutionary and functional landscape of plant phased small interfering RNAs. *BMC Biology*, *13*, 32. https://doi.org/10.1186/s12915-015-0142-4

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, *31*(13), 3406–3415.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Polydore S, Lunardon A, Axtell MJ. Several phased siRNA annotation methods can frequently misidentify 24 nucleotide siRNA-dominated *PHAS* loci. *Plant Direct*. 2018;2:1–13. https://doi.org/10.1002/pld3.101