INTEGRATED, SCALABLE TOOLS FOR SMALL RNA GENOMICS: NOVEL ALGORITHMS AND THEIR APPLICATION TO CHARACTERIZE GERMLINE-ASSOCIATED sRNA PATHWAYS IN DIVERSE SPECIES

by

Atul Kakrana

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics and Systems Biology

Summer 2017

INTEGRATED, SCALABLE TOOLS FOR SMALL RNA GENOMICS: NOVEL ALGORITHMS AND THEIR APPLICATION TO CHARACTERIZE GERMLINE-ASSOCIATED sRNA PATHWAYS IN DIVERSE SPECIES

by

Atul Kakrana

Approved: _____
Cathy H. Wu, Ph.D.
Chair of Bioinformatics & Computational Biology

Approved: _____
Babatunde Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____
Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Blake Meyers, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Cathy Wu, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Virginia Walbot, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Abhyudai Singh, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

First, I would like to thank my dissertation advisor Dr. Blake Meyers for giving me an opportunity to work in his group and the financial support for my doctoral study. I very much enjoyed working in your group, and you have always been kind, understanding and patient with me and my research. I learned a lot from you, especially on how to stay calm and focused on my goals when things don't work the way we expect. I cannot express what a tremendous impact that Blake has had on my career development.

Second, I would like to express my deepest gratitude to my committee members, Professors Cathy Wu, Virginia Walbot and Abhyudai Singh for and constant moral support and scientific suggestions. I couldn't ask for more caring and supportive committee. Honestly, I feel that I had the best committee among my peers.

I am also thankful to my lab members - Sandra Mathioni, Parth Patel, Reza Hammond, Ayush Dusia, and Kun Huang - for believing in me and supporting me throughout my study. I am particularly grateful to Sandra Mathioni for being the first and the only person in my lab who came forward to generate data for me. I could have achieved significantly more if Sandra had joined the lab earlier. I am also thankful to Delaware Biotechnology staff, particularly Bruce Kingham, Karol Miaskiewicz, and Olga Shevchenko for their consistent support.

I am grateful to Dr. Salil Lachke for guiding me whenever I felt lost and helping me when I needed it most; I will always be indebted to you for your kind favor. I am also thankful to Prof. John McDonald for advising me on statistical

solutions for my experiments and to Dr. Jim Leebens-Mack for sharing the unpublished draft of the *Asparagus* genome assembly.

I thank my parents for their unwavering support. I am very sorry for the lost time, which I should have had spent with you. I will try to make up for that in the coming years. Lastly, I am grateful to my wife; she stole all dull moments and replenished them with beautiful memories.

TABLE OF CONTENTS

Appendix

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Cells associated with the male germline, specifically in rice and maize (grasses), produce diverse and numerous "phased" 21-nt and 24-nt siRNAs. These phased siRNAs (phasiRNAs) show striking similarity to mammalian Piwi-interacting RNAs (piRNAs) in terms of their abundance, biogenesis and timing of accumulation. Both the plant phasiRNA and mammalian piRNA pathways are emerging as factors crucial for reproductive success. However, since the first report of germline-associated plant phasiRNAs, no systematic study of their evolutionary origins has yet been reported; in this context, the meiotic (24-nt) phasiRNAs are particularly interesting, as they have only been described in grasses, a group of monocots that speciated ~71 million years ago (MYA). Grasses include the most important staple crops: rice, maize and wheat. Given the importance of reproductive success to crop yield, a deeper understanding of phasiRNA pathway is crucial.

This dissertation traces the prevalence and origins of phasiRNA pathways in monocot evolution, while simultaneously it addresses a broad range of key computational gaps and algorithmic limitations in leveraging small RNA data for the study of small RNA in plants. First, I present a new set of tools for identifying and validating miRNA targets, and a new suite for computational characterization of phasiRNAs, which together comprise important methods for studies of plant sRNA field. These next generation tools efficiently scale to the increasing volume of high-throughput data, and are fast, sensitive and feature-rich compared to the existing options. Next in my work, I deployed these tools to investigate phasiRNAs in a

recently sequenced genome, that of *Asparagus officinalis.* The common ancestor of asparagus and the grasses diverged approximately 109 MYA. My work then further expanded to study two other non-grass monocots, *Lilium* (*Lilium maculatum)* and daylily (*Hemerocallis lilioasphodelus)*, which diverged from A*sparagus* ~111 MYA. In this dissertation, I demonstrate that both pre-meiotic and meiotic phasiRNAs are prevalent across the monocots that I studied, establishing their origins well before grasses. In addition to male germline, I find evidence for their accumulation in female and somatic tissues, perhaps suggesting that the narrow accumulation of reproductive phasiRNAs in anthers is either not a general characteristic or it is the product of evolutionary refinement in the grasses. I show that the miRNA trigger for pre-meiotic (21-nt) phasiRNAs likely shifted in evolutionary time from targeting pathogen-defense genes to long, non-coding RNAs (observed in grasses) via specialization and sub-functionalization versus neo-functionalization. I also demonstrate that exceptions to the canonical mechanism of biogenesis of phasiRNAs exist in monocot evolution, whereby phasiRNAs are produced apparently without a miRNA trigger. I conclude that plants show substantial variation in their composition and biogenesis of reproductive phasiRNAs, which have broad roles in plant germline development.

Chapter 1

INTRODUCTION

Regulatory small RNAs (sRNAs) are ubiquitous "non-coding" component of the plant transcriptome that function in distinct, yet overlapping, genetic and epigenetic silencing pathways. These play essential regulatory roles in the growth, development, reproduction, genome reprogramming and defense processes. An array of sRNA pathways likely contribute to the phenotypic plasticity in plants as well as animals. Most regulatory sRNAs are 21-, 22- and 24 nt in size, produced as double-stranded duplexes from the helical regions of longer non-coding RNA precursors by endonuclease activity of DICER-LIKE proteins (DCLs). Their biogenesis mechanism relies on the formation of double-stranded RNA (dsRNA) intermediates from either hairpin precursors, formed by the intermolecular hybridization of precursor transcript, or by the synthesis of dsRNA from a single-stranded RNA by RNA-DEPENDENT RNA POLYMERASEs (RDRs). Processed sRNA duplexes load into ARGONAUTE (AGO) proteins to target coding and non-coding RNAs. Depending on the nature of target transcript and the AGO involved, this process leads to target cleavage and degradation, translational repression or recruitment of additional cofactors.

The duplication of genes encoding for DCL and RDR has resulted in extensive diversity of regulatory sRNAs, specifically in terms of their size (Mukherjee, Campos, and Kolaczkowski 2013; Willmann et al. 2011), which along with the diversification of AGO proteins led to the development of distinct gene-silencing processes (Czech and Hannon 2011) based on the differential AGO affinities for sRNA duplexes (Mi et

al. 2008). In plants, endogenous sRNA pathways are divided into five major classes: microRNAs (miRNAs), secondary siRNAs, heterochromatic siRNAs (hc-siRNA), hairpin-derived siRNAs (hp-siRNAs) and natural antisense siRNAs (nat-siRNAs). See **Figure 1.1**, below, for their classification based on mode of biogenesis and nature of precursor.

Among these five classes of sRNAs, the first three comprise the major proportion of regulatory sRNAs, in terms of their abundances and diversity of their functions. In plants, miRNAs are the most well-studied subset of regulatory sRNAs. MicroRNAs execute post-transcription silencing of target genes by precise cleavage or translational repression. In addition, these also trigger secondary siRNA production from RNA polymerase II (Pol II) transcribed coding and non-coding RNAs (Allen et al. 2005; Peragine et al. 2004). miRNAs (21- and 22-nt) usually have a defined set of mRNA targets and both, individual miRNA families and their targets, are mostly conserved over a long evolutionary period (Cuperus, Fahlgren, and Carrington 2011). Secondary siRNAs are generated by the cleavage of specific coding and non-coding precursors by 22-nt miRNAs (Chen et al. 2010; Cuperus et al. 2010). Although 21-nt miRNAs also trigger the production of phasiRNAs via two-hit model (Axtell et al. 2006), the functionally unique "22-nt" class of miRNAs triggers the bulk of secondary siRNAs. Secondary siRNAs, and specifically reproductive phased siRNAs (21- and 24-nt) are the focal point of this dissertation and are described in detail later in this chapter.

In this dissertation, I aimed to fill existing gaps for bioinformatics tools needed for characterizing plant sRNA populations by developing new algorithms and packaging them into industry-grade software's; and finally leveraging these to make

insights into the biogenesis and origins of a particular class of secondary siRNA pathways – “male-germline associated phasiRNAs”. These pathways, referred to as pre-meiotic and meiotic phasiRNAs, are relatively new discoveries, crucial for reproductive success and likely analogs of mammalian PIWI-interaction (piRNAs). Our knowledge of their functional roles as well as evolution is mostly missing. In brief, we first develop necessary tools and then investigate male-germline associated pathways in branches of monocot which represents at least 115 million years of evolution.

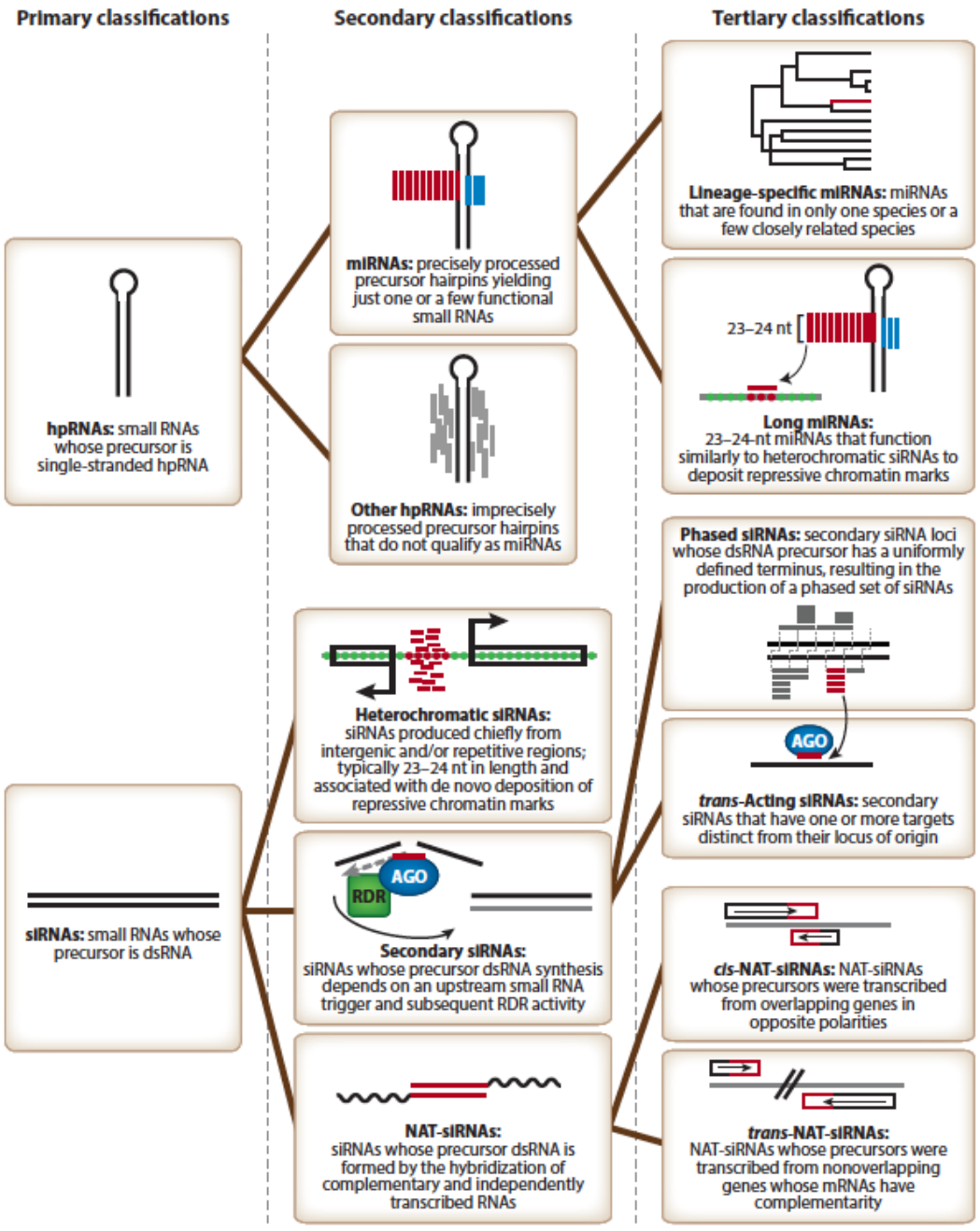


Figure 1.1: **Hierarchical classification system for endogenous plant small RNAs**. Thick black lines indicate hierarchical relationships. Abbreviations: dsRNA, double-stranded RNA; hpRNA, hairpin RNA; miRNA, microRNA; NAT-siRNA, natural antisense transcript small interfering RNA; siRNA, small interfering RNA. Figure from (Axtell et al., 2013), courtesy of Michael Axtell.

1.1 Pre-dissertation state of approaches to discover targets for plant miRNAs

Since the first reports of miRNAs in plants (Llave et al. 2002; Reinhart et al. 2002), there has been a steep escalation in the number of known miRNAs, fueled primarily by concurrent advances in sequencing technologies and computational methodologies. At the time of writing, there are over 7,385 mature miRNAs reported from 72 plant species in miRBASE (version 20). However, identification of a miRNA does not provide insights into its function or regulatory targets, nor is an understanding of the targets part of the process of miRNA identification(Meyers et al. 2008). A key to understanding the biological relevance of a miRNA lies in discovering and validating its targets.

Parallel analysis of RNA ends (PARE) is a high- throughput sequencing technique which profiles uncapped mRNAs, products of cleavage or decay, facilitating studies of miRNA targets (German et al. 2009). Nearly identical techniques have been termed ‘degradome analysis’ or ‘GMUCT’ and they generate equivalent data (Addo-Quaye et al. 2008; Gregory et al. 2008). Because of our role in the development of the technique called PARE, we are partial to this terminology and will use it hereafter. Computational tools to predict miRNA targets and validate those targets using PARE data are limited in both number and functionality. Also, among these tools, there is a divergence in the methodology used to predict targets and assign significance scores. *CleaveLand*, the most-cited and perhaps most commonly-used tool for computational validation of miRNA targets using PARE datasets, presumes that there exists a positive correlation between complementarity at a canonical seed region (2 to 13 nt from the 5, end of the miRNA) of a miRNA::target duplex and probability of actual cleavage (Addo-Quaye, Miller, and Axtell 2009; Fahlgren and Carrington 2010). Therefore, *CleaveLand* implements a ‘seed region’-based target scoring schema along

with a penalty score cutoff of 4, to model the *p-values* for validated interactions. However, cleavage of potential targets can occur even with poor complementarity in the seed region or mismatches at canonical positions (Y. Zheng et al. 2012; Brousse et al. 2014) (**Figure 1.2**).

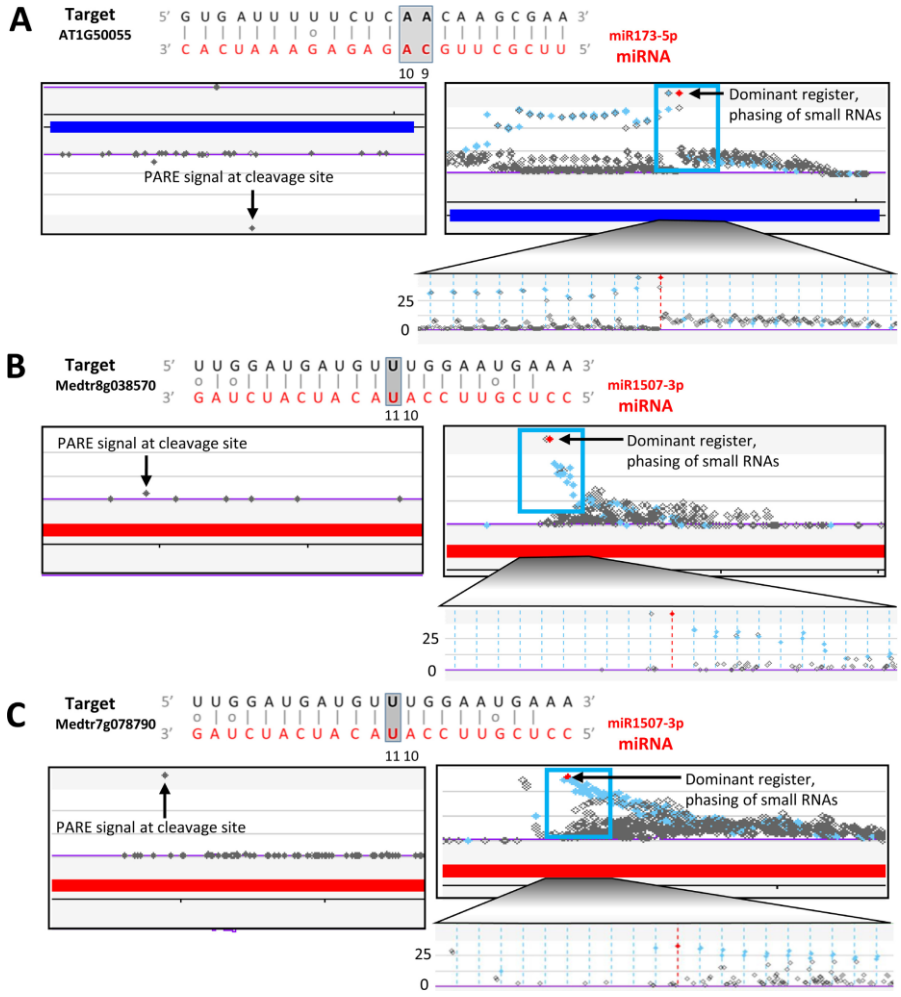


Figure 1.2: **MicroRNA (miRNA) targets with weak or non-canonical interactions are missed by existing PARE-based validation tools**. Each example shows the target-miRNA alignment at the top, with screenshots below from our website (https://mpss.danforthcenter.org); the upper panel shows the PARE data, the middle panel in each case, phased small RNA production from cleavage sites further substantiates the cleavage events. (A) at-miR173–5p cleaves the *Arabidopsis thaliana* TAS1B gene with a penalty score=4.5, and with mismatches at both the 10th and 11th positions. (B) mtr-miR1507 cleaves the *Medicago truncatula* NBS-LRR type disease resistance gene (Medtr8g038570) with a penalty score = 7, and with a mismatch at the 11th position. (C) mtr-miR1507 cleaves the *M. truncatula* NBS-LRR type disease resistance gene (Medtr7g078790) with a penalty score = 7, and with a mismatch at the 11th position.

PAREsnip is an accelerated approach to extend PARE validation of targets from a small set of miRNAs up to a more extensive library of small RNAs (Folkes et al. 2012). Yet, *PAREsnip* suffers from the same inductive bias as *CleaveLand*, which is the assumption that there exists a positive correlation between complementarity in the canonical seed region and probability of actual cleavage. These assumptions about miRNA-target interactions are not easily modified or refined. Furthermore, the target search algorithm implemented in *PAREsnip* expects a perfect match at canonical 10th–11th positions and is dependent on 'seed region'- based rules for its speed. Therefore, both *PAREsnip* and *CleaveLand* tend to bias the results by assigning significant *p-values* to only those interactions that either has a relatively good complementarity in seed regions or to those miRNAs that have limited number of interactions. Another existing tool for PARE-based validation of miRNA targets, *SeqTar* (Y. Zheng et al. 2012), broadens the complementarity-based prediction rules but it is currently moderately slow and therefore best employed for a pre-selected set of miRNAs (or sRNAs) rather than a complex sRNA population.

1.2 Rationale for developing a new PARE-based target discovery tool

Tools that existed for working with PARE data, *Cleaveland*, *PAREsnip* and *SeqTar* (Addo-Quaye, Miller, and Axtell 2009; Folkes et al. 2012; Y. Zheng et al. 2012), before our work in this area not only suffered from inductive bias, but focused exclusively on the annotated portion of the genome, utilizing cDNA sets as their input. Many new genomes are poorly annotated, at least in their initial release, and recent studies indicate that even in well-annotated genomes, target mRNA still remain to be found in unannotated, intergenic regions (IGRs), evidenced in reports of large numbers of miRNA- targeted long, noncoding RNAs in the grasses (Arikit, Zhai, and

Meyers 2013; Johnson et al. 2009; Song, Li, et al. 2012). For example, one report describes these loci (and their miRNA triggers) in unannotated regions of the Brachypodium genome (D. H. Jeong et al. 2013). Such analyses depend on approaches for target identification at a full-genome level, not just using annotated genes. As mentioned above, all existing algorithms to validate miRNA targets from PARE data are built on the assumption that the relevant interactions are within annotated transcripts. Since this is inaccurate, a new approach is required to be able to discover genome-wide targets of miRNAs. Such genome-wide prediction methods are primarily lacking because complete genomes are many times bigger then it's annotated portions, plus the mismatches, wobble pairings and gaps (or bulges) that occur between miRNA and their targets inflates the search space by many magnitudes even for the modest sized genomes.

In the past decade, the increase in the yield-per-dollar cost of sequencing has further democratized the use of high-throughput sequencing technologies. This has initiated a shift in plant genomics, from the study of model plants with modest genome sizes to diverse crops, and now even including species with genomes many times larger than most model genomes. For example, the recently sequenced *Picea abies* and *Triticum aestivum* genomes are both >100 larger than *Arabidopsis thaliana*. Furthermore, an increased DNA sequencing throughput has commoditized the sequencing of RNA samples. A single small RNA or PARE library now includes tens millions of reads that can be analyzed and integrated to predict new miRNAs and their targets. These recent advances and cost reductions in genome and RNA sequencing warrant the development of a PARE validation approach capable of high efficiency to handle large non-model genomes and quickly analyze all likely sRNA-PARE

interactions from multiple libraries. Fortunately, there exist many technologies that could be deployed to meet these needs, via the development of algorithms capable of efficiently exploiting available computing power.

This explosion in genome sequencing has also fueled the sRNA studies. Many groups have developed approaches to identify novel miRNAs and their targets from sequence data. Databases that computationally predict, curate and collect experimentally verified miRNA-target interactions include *TarBase*, *StarBase*, *miRTarBase* and *MiRecords* (Vergoulis et al. 2012; J.-H. Li et al. 2014; Hsu et al. 2014; Xiao et al. 2009). For biologists, the most practical use is to combine the best aspects of different databases and interpret the data using graphical interfaces. However, the database mentioned here lack integrated genome viewers, except *StarBase* (J.-H. Li et al. 2014) which uses the purpose-built *deepView* for visualization of mapped reads, target peaks, and target plots. A limitation for plant biologists is that most of these databases focus on animals with only limited plant data. The extensive availability of data in plants is an opportunity for greater integration of small RNA, PARE, and RNA-seq data. A plant focused, online resource for depositing and curating miRNA-target interactions could advance data- driven small RNA analyses.

1.3 Phased secondary siRNAs are crucial regulators of development, reproduction and plant defense

Phased siRNAs (phasiRNAs) are a major subclass of secondary siRNAs, found extensively in plants (Axtell 2013a). The defining characteristic of phasiRNAs is the DCL-catalyzed processing of double-stranded RNA (dsRNA) precursors, starting from a precisely delimited 5’ terminus and generating regularly spaced 21- or 24-nt populations of siRNAs (Johnson et al. 2009). PhasiRNAs can be further subdivided

into three main categories based on their precursor mRNAs and spatiotemporal patterns of accumulation: i) The first phasiRNAs identified, so-called *trans*-acting siRNAs (tasiRNAs) generated from a small family of long, non-coding mRNAs (lncRNAs) referred to as *TAS* genes (Vazquez et al. 2004; Peragine et al. 2004; Allen et al. 2005), ii) phasiRNAs from protein-coding transcripts, such as NB-LRRs or PPRs (Fei, Xia, and Meyers 2013), and iii) two classes, 21-nt premeiotic or 24-nt meiotic phasiRNAs, highly enriched in reproductive tissues and also produced from lncRNAs, reported in grasses, but with no as-yet reported targets (Johnson et al. 2009; Zhai et al. 2015). Thus, the umbrella name of "phasiRNAs" refers simply to their biogenesis and not their function (unlike the subset of tasiRNAs) because many phasiRNAs lack validated targets, either in *cis* or *trans* (Zhai et al. 2011; Fei, Xia, and Meyers 2013).

The biogenesis of phasiRNAs in plants is dependent on a triggering mechanism that sets the phase of the resulting secondary siRNAs, generated from a specific nucleotide in the mRNA precursor. To date, the only described trigger type is a miRNA, and a breakthrough in our understanding of plant miRNA function came with the observation that all or nearly all 22-nt miRNAs trigger phasiRNA biogenesis from their targets (Chen et al. 2010; Cuperus et al. 2010). The miRNA triggers function via the ARGONAUTE (AGO) proteins into which they are loaded, and since phasiRNA biogenesis requires both SGS3 and RDR6 (Peragine et al. 2004; Vazquez et al. 2004), there may be interactions between these proteins, ultimately recruiting DCL4 or DCL5. SGS3 and RDR6 proteins function in the cytoplasm, forming siRNA bodies (Jouannet et al. 2012). Recent work has identified membrane-bound polysomes in the rough ER as the site where miRNA triggers of phasiRNA accumulate, leading to phasiRNA biogenesis (Shengben Li et al. 2016). miRNA triggers are thus an

important component in the analysis of phasiRNAs, and the identification of specific triggers with specific *PHAS* targets is integral part of studies of phasiRNAs.

Since, the discovery of phasiRNAs, in 2005, *TAS* genes have been relatively well-characterized, especially in Arabidopsis but these represent a miniscule fraction (n=8) of the *PHAS* repertoire found in many plant genomes. In other eudicot genomes, there are many more *PHAS* loci compared to Arabidopsis, the result of up to hundreds of protein-coding genes that are targeted by diverse miRNAs, many of which are lineage-specific (Zhai et al. 2011; D. H. Jeong et al. 2013; Arikit et al. 2014; R. Xia, Ye, et al. 2015). Beyond eudicots, plant genomes contain even more *PHAS* loci. For example, reproductive phasiRNAs number in the hundreds to thousands of loci in maize (Zhai et al. 2015) and rice (Fei et al. 2016), and have yet to be characterized broadly in monocots or other lineages outside of the grasses (described in detail below). These includes, the premeiotic loci that are targeted by miR2118 family members, triggering production of 21-nt phasiRNAs accumulating in early anther development, and the 24-*PHAS* loci that are targeted by miR2275 family members, triggering production of 24-nt phasiRNAs, accumulating in anthers during meiosis (Zhai et al. 2015). Analysis of the spruce genome, a gymnosperm that speciated 329 million years before the evolution of monocots and eudicots have over 2000 *PHAS* loci, most of which are protein-coding genes, including over 750 *NB-LRRs* (R. Xia, Xu, et al. 2015). Thus, plant *PHAS* loci are widely prevalent and highly variable from genome to genome both in the total number and in terms of the types of loci that generate them.

1.4 Rationale for developing new methods for in silico characterization of phased siRNAs

Characterization of *PHAS* loci from each sequenced plant genome will provide insights into this unusual type of post-transcriptional control, its evolution, and diversification. Tools for the *de novo* identification of *PHAS* loci (or genes) to date have required an assembled genome for their discovery and additional experimental data (PARE or degradome libraries) to further identify their miRNA triggers. Integrated tools for discovery and in-depth characterization of *PHAS* genes have not yet been developed; and the existing options are both limited in number and function. These algorithmic limitations and bioinformatic gaps along with increasing depth and volume of sequencing data necessitates a scalable, fast and advanced methods to study this relatively new class of secondary siRNAs that might even be transcending the evolutionary boundaries with parallel pathways existing in mammals (Johnson et al. 2009; D. H. Jeong et al. 2013; Mohn, Handler, and Brennecke 2015; B. W. Han et al. 2015).

1.5 Male-germline associated phased siRNA pathways and their importance

In higher plants, diverse and versatile small RNA (sRNA) pathways are present in reproductive tissues, present presumably to ensure reproductive success (Borges and Martienssen 2015). Two such male germline-associated pathways, generating diverse and abundant phased secondary siRNAs (phasiRNAs) have been described in grasses (Johnson et al. 2009; Zhai et al. 2015; Fei et al. 2016). These phasiRNAs are generated from 5'-capped and polyadenylated, non-coding precursors ("*PHAS*" transcripts) transcribed by RNA polymerase II (Pol II) from non-repetitive loci. Their production is triggered by two 22-nt miRNAs – miR2118 for 21-nt phasiRNAs and miR2275 for 24-nt phasiRNAs – that direct cleavage of *PHAS*

transcripts, setting a consistent 5'-terminus for each *PHAS* precursor. The 3' mRNA fragments are converted to double-stranded RNA by RNA-DEPENDENT RNA POLYMERASE6 (RDR6), and processed by DCL4 and DCL5 to yield phased 21- and 24-nt siRNAs, respectively (Song, Li, et al. 2012).

miR2118-triggered 21-nt phasiRNAs are abundant during the specification of cell fate in anthers, originating from the epidermal layer (Zhai et al. 2015). In contrast, miR2275-triggered 24-nt phasiRNAs accumulate during meiosis in the tapetum and germinal cells, and they persist into the differentiation and maturation of pollen; their production is dependent on normal tapetal cells (Zhai et al. 2015; Nonomura et al. 2007). There are parallels with mammalian PIWI-associated RNAs (piRNAs): both phasiRNAs and mammalian piRNAs originate from non-repetitive loci, are generated in two different size classes with distinct developmental timing, and both generate abundant and diverse siRNAs (Nonomura et al. 2007; Johnson et al. 2009; Zhai et al. 2015). Recently, piRNAs were also shown to be phased (Mohn, Handler, and Brennecke 2015; B. W. Han et al. 2015). Hence, based on their timing and by analogy to the piRNAs, the grass phasiRNAs are referred to as "pre-meiotic" and "meiotic" siRNAs (Axtell 2015)

1.6 Rationale for investigating phasiRNAs beyond grasses

The functions of plant reproductive phasiRNAs are as-yet unknown, but clues about their roles in male reproductive success are emerging. For example, a mutant in MEIOSIS ARRESTED AT LEPTOTENE (MEL1), an Argonaute protein in rice, arrests in early meiosis and has an abnormal tapetum and anomalous pollen mother cells (PMC) (Nonomura et al. 2007; Komiya et al. 2014). MEL1 selectively binds 21-nt phasiRNAs (Komiya et al. 2014), indicating that pre-meiotic phasiRNA functions

are required for male fertility. Various, hypothetical roles of meiotic phasiRNAs in male fertility have also been proposed, potentially including silencing transposable elements, genome imprinting, and chromatin remodeling by facilitating chromosome dynamics, or functions in pairing, synapsis and recombination (Zhai et al. 2015; Dukowic-Schulze et al. 2016)

The evolutionary origins of plant reproductive phasiRNAs are also poorly characterized, with little known outside of their presence in the grasses. In this context, 24-nt meiotic phasiRNAs are particularly interesting as their biogenesis depends on Dicer copy specialized for their production, derived from DCL3 yet absent in eudicots, and described only in grass genomes (Margis et al. 2006). Also, the origins of miR2275, the specialized trigger for these siRNAs, is unknown. The 21-nt, pre-meiotic phasiRNAs have apparently, and unusually, co-opted previously existing components, including as triggers the miR2118/482 superfamily, which in many other plant genomes triggers phasiRNAs from NB-LRR pathogen-defense genes (Zhai et al. 2011; Y. Zhang et al. 2016). The evolutionary route by which miR2118 shifted to targeting non-coding *PHAS* precursors in a highly restricted spatiotemporal manner is also not known. Thus, significant questions remain about when and how in plant evolution reproductive phasiRNAs emerged and were refined to the state in which they've been observed in grasses.

1.7 Overview of dissertation research

This dissertation investigates the origins and biogenesis male-germline associated pathways beyond grasses such as maize and rice; and in process significantly advances the computational methods to leverage RNA-seq data for study of small RNAs. Specifically, by presenting new methods for discovery of miRNA

targets and characterization of phased siRNAs including their discovery and trigger identification. In Chapter 2, I describe a new miRNA target discovery tool small RNA PARE Target Analyzer (*sPARTA*). *sPARTA* was initially imagined to address the algorithmic shortcomings of existing methods, but the explosive growth in number of plant miRNAs, genome and experimental data (PARE and degradome libraries) in last few years along with and the prospects of discovering novel regulatory modules, further, motivated us to develop a complete software package with a built-in, plant focused target prediction module (aka 'miRferno'). The next-generation tool that I describe in Chapter 2 was developed from scratch and is the only existing tool (as of writing this) capable of predicting and validating targets at a whole genome level, even for hundreds to thousands of miRNAs or small RNAs. It is up to 500x faster compared to the most popular alternative. Unlike previous tools, which could only focus on annotated portion of genomes and identify targets for tens to a hundred of miRNAs, *sPARTA* employs true parallel computing to gain significant advances in speed, and it implements a data partitioning scheme for both scalability and to maintain a small memory footprint, which makes it superlatively efficient in handing even the biggest available genomes as well as large input sets of RNA data. sPARTA is freely available to plant researchers, has open-source and released under permissive license.

In Chapter 4, I describe "*PHASIS*", the "first" tool-set for in-depth *in silico* characterization of phasiRNAs. The motivation to develop a feature-rich suite came from the fact that the existing options to study identify phasiRNA are not only limited in number and function but also incompatible or inefficient in handling a large volume of small RNA-seq data. These existing methods do not support the large-scale study of

PHAS pathways, and more importantly, these require a sequenced genome for the discovery of *PHAS* genes and phasiRNAs, and they need additional data such as PARE or degradome to identify triggers. Since I planned to embark on an investigation that transcends evolutionary boundaries, the expectation to have an assembled genome for species of interest was unrealistic. So, to make this study possible, I developed a new computational toolset that we named as "*PHASIS*". *PHASIS* facilitates discovery, quantification, annotation of phasiRNA loci or genes from a few to hundreds of sRNA libraries in a single run, and rapid identification of their miRNA triggers. Benchmarks from five different plant species demonstrate that *PHASIS* is sensitive, scalable and fast. Importantly, *PHASIS* eliminates the requirement of a sequenced genome and PARE/degradome data for discovery of phasiRNAs and their miRNA triggers. Like *sPARTA*, *PHASIS* too is open-source, released under a permissive license. I believe that the algorithmic superiority, flexibility to tailor analysis and the suitability for small to large-scale experiments will make *PHASIS* the *de facto* choice for discovery and study of phased siRNAs in the future.

Chapter 4 focusses on the implementation of next generation tools and methods that I describe in Chapter 2 and 3 for an investigation of phased siRNA pathways beyond the "grasses" such as maize and rice, to check the possibility of their prevalence in other plant species and to gain insights into their origins and evolution. In this chapter, I take advantage of the recently sequenced *Asparagus officinalis* genome (Harkess et al., 2017) for our investigation. The *Asparagus* and grass lineages diverged approximately 114 million years ago (Hedges et al. 2015). In addition to Asparagus, I characterized two representative non-grass monocots, *Lilium* (*Lilium*

maculatum) and daylily (*Hemerocallis lilioasphodelus)*, which diverged >120 MYA from MRCA of grasses (Chase and Reveal 2009). My combined study of Asparagus, *Lilium* and daylily reveal that both pre-meiotic and meiotic pathways outdates speciation of grasses during monocot diversification i.e. these are not specific to grasses, as earlier believed. I also demonstrate the presence of phasiRNAs in female and somatic tissues, perhaps suggesting that the narrow accumulation in anthers is either not universal or the product of refinement in the grasses. I show that the miRNA trigger for pre-meiotic (21-nt) phasiRNAs likely shifted in evolutionary time from targeting pathogen-defense genes to long, non-coding RNAs (observed in grasses) via specialization and sub-functionalization versus neo-functionalization. Finally, I demonstrate that many 24-nt phasiRNAs are produced from precursors lacking miRNA trigger, from long inverted repeats, revealing divergence in both the biogenesis mechanism and the protein factors of male-germline associated phased siRNA pathways.

1.8 Publications from this dissertation

The following publications are direct result of research conducted in this dissertation; and the chapter 2, 3 and 4 directly correspond to paper or manuscript #3, 7 and 9, respectively:

A. Gong L, **Kakrana A**, Meyers BC, Wendel JF (2013) Composition and Expression of Conserved MicroRNA Genes in Diploid Cotton (Gossypium) Species. Genome Biol. Evol. 5, 2449-59

B. Thompson BE, Basham C, Hammond R, **Kakrana A**, et al (2014) The dicer-like1 Homolog fuzzy tassel is required for the regulation of meristem determinacy in the inflorescence and vegetative growth in maize. Plant Cell. 26(12):4702-17

C. **Kakrana A**, Hammond R, Patel P, Nakano M, Meyer BC (2014) sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. Nucleic Acids Res. 42(18):e139 *(*Fastest algorithm (>500x), to date, for plant miRNA target identification and validation*)

D. Arikit S, Xia R, **Kakrana A**, et al (2014) An Atlas of Soybean Small RNAs Identifies Phased siRNAs from Hundreds of Coding Genes. Plant Cell. 26(12):4584-601

E. Patel P, Ramachandruni SD, **Kakrana A,** Nakano M, Meyers BC (2015) miTRATA: a web-based tool for microRNA Truncation and Tailing Analysis. Bioinformatics. 1;32(3):450-2

F. Mathioni, S.M., **Kakrana, A.**, and Meyers, B.C. 2017. Characterization of plant small RNAs by next generation sequencing. Curr. Protoc. Plant Biol. 2:39-63. doi: 10.1002/cppb.20043 (Book Chapter)

G. **Kakrana A**, Li P, Patel P, Hammond R, Mathioni S, Anand D, Meyer BC. *PHAS*IS: A computational suite for de novo discovery and characterization of phased, siRNA-generating loci and their miRNA triggers. (In review)

H. Harkess A, Zhou J, Xu C,…, **Kakrana A**, Meyers BC, Leebens-Mack J (2016) The evolution of sex chromosomes in Asparagus. (accepted in Nature Communications) (*Plant genome sequencing report, led the small RNA study

I. **Kakrana A**, Mathioni S, Huang K, Hammond R, Patel P, Vandiver L, Gregory B, Leebens-Mack J, Meyers BC. Survey of phasiRNA pathway in non-grass monocots, uncovers widespread prevalence and unique plasticity. (Prepared for submission to Genome Research)

Chapter 2

A HIGH-PERFORMANCE APPROACH FOR PREDICTION AND VALIDATION OF miRNA TARGETS

(All of this chapter has been published previously as Kakrana et al. (2014). It has been modified in part to meet the formatting requirements of the dissertation, and to integrate the work together with the rest of the dissertation.)

Parallel analysis of RNA ends (PARE) is a technique utilizing high-throughput sequencing to profile uncapped, mRNA cleavage or decay products on a genome-wide basis. Tools that existed before our work, to validate miRNA targets using PARE data employ only annotated genes, whereas important targets may be found in unannotated genomic regions. To handle such cases and to scale to the growing availability of PARE data and genomes, we developed a new tool, '*sPARTA*' (small RNA-PARE target analyzer) that utilizes a built-in, plant-focused target prediction module (aka 'miRferno'). *sPARTA* not only exhibits an unprecedented gain in speed but also it shows greater predictive power by validating more targets, compared to a popular alternative. In addition, the novel 'seed-free' mode, optimized to find targets irrespective of complementarity in the seed-region, identifies novel intergenic targets. To fully capitalize on the novelty and strengths of *sPARTA*, we developed a web resource, '*comPARE*', for plant miRNA target analysis; this facilitates the systematic identification and analysis of miRNA-target interactions across multiple species, integrated with visualization tools. This collation of high-throughput small RNA and

PARE datasets from different genomes further facilitates reevaluation of existing miRNA annotations, resulting in a 'cleaner' set of microRNAs.

2.1 sPARTA algorithm and workflow

The *sPARTA* algorithm has four main steps that are implemented in series. With the exception of the first step in which user-defined features (gene or intergenic) are extracted and fragmented, the three subsequent steps use single-instruction multiple-data (SIMD) parallel processing via Python (v3.3) *multiprocessing* module. The two most data intensive steps (i) mapping reads from multiple PARE libraries and (ii) the prediction of sRNA or miRNA targets (by *miRferno*), both benefit from two-way SIMD parallelism (**Figure 2.1**).

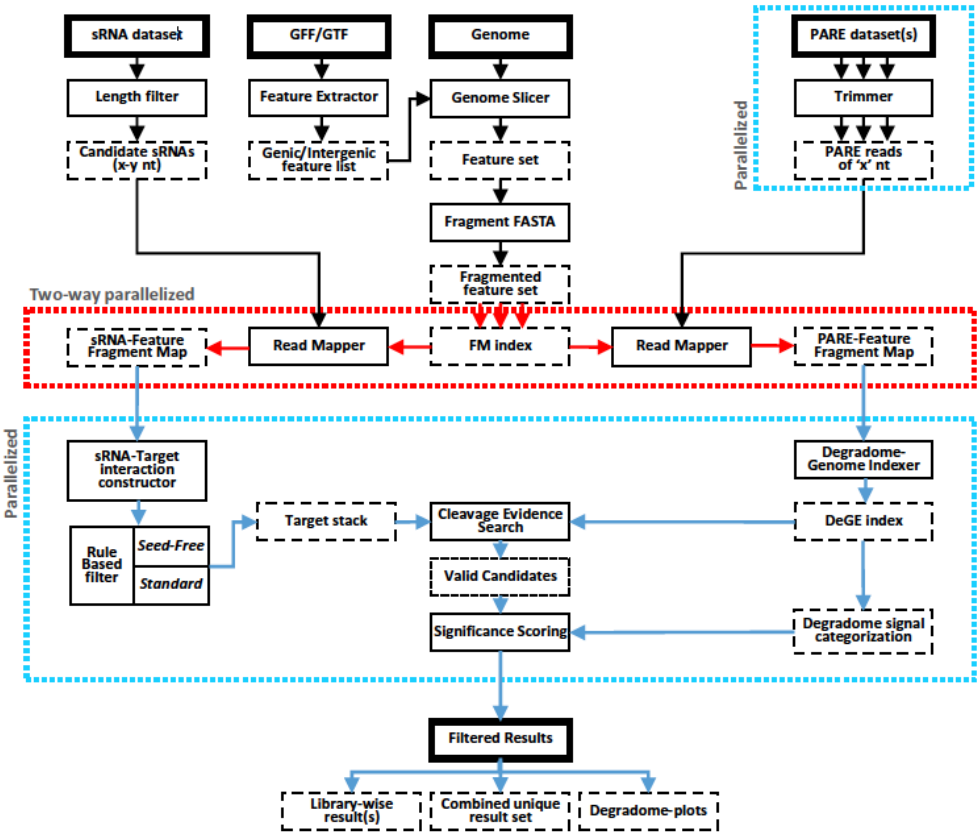


Figure 2.1: ***sPARTA* schematic, showing order of steps in workflow**. Solid boxes represent *sPARTA* functions, dashed boxes represent the product of an applied function. Multiple arrows indicate multiple output files from the preceding function. Steps executed in parallelized environment are enclosed within colored dotted lines.

2.1.1 Feature extraction and input file partitioning

To build a 'feature set' or input library of sequences in which targets will be identified for a species of interest, *sPARTA* starts with a GFF file (Generic Feature Format, version 3) containing gene annotations along with the corresponding genome sequence. In many cases, this is downloaded from Phytozome (Goodstein et al. 2012). These GFF and genome sequence files are used by the built-in *Genome*

Slicer function to extract first the coordinates of selected features (i.e. genic or IGRs) from the GFF files and next to extract the sequences from the genome. These intergenic and genic sequence sets comprise the main feature set, which is further partitioned into different data elements (features) so as to implement data parallelism.

2.1.2 PARE data processing and read mapping

The next step in *sPARTA* is to map PARE reads to the feature set. An FM index (Simpson and Durbin 2010) for each component of the feature-set is created using Bowtie (version 2, in the current *sPARTA* implementation) (Langmead and Salzberg 2012) with the default off-rate parameter. PARE reads in tag-count format (a tab-separated file of read sequences and normalized frequencies) from each dataset were then aligned to the partitioned FM indexes using Bowtie, with default end-to-end settings and no mismatch allowed, to generate PARE-fragment maps. PARE datasets in format other than tag-count could be easily converted to tag-count using publicly available Tally (Davis et al. 2013). The PARE mapping step implements SIMD parallel processing on both the involved datasets, i.e. the feature set and the PARE reads. The feature set file size for different species could range up to tens of gigabytes, while the number of reads in a PARE dataset range from millions to hundreds of millions. So, two-way parallelization further enhances both scalability and load balancing, improving the parallel processing efficiency of the *sPARTA* algorithm. The parallelization on the number of reads is achieved by Bowtie's built-in parallel processing function that makes use of the *pthreads* library to distribute reads across concurrent search threads (Langmead et al. 2009).

2.1.3 Prediction of targets using novel miRferno algorithm

In the third major step, targets of small RNAs are identified in the sequences of the feature set. *sPARTA* has a newly developed, built-in target prediction module—miRferno—which has two prediction modes, greedy and exhaustive, described below. In both modes, the miRNA or sRNA sequences used as an input to find targets are mapped to the fragmented features using Bowtie. The advantage of using version 2 of Bowtie is that it allows gapped alignments, and therefore it can find miRNA-target interaction which include gaps and bulges. The inclusion of these mismatches substantially increases the sensitivity of target prediction, but gaps also greatly inflate the size of the search space and slow down the process of finding targets. Prior decomposition of the feature set into smaller partitions (i.e. features) by *sPARTA* reduces the index size and associated search space for gapped alignments. This increases the efficiency of alignments, and in combination with parallelization on the number of PARE reads and genomic partitions (i.e. two-way parallelization), comprises an effective combination of speed, sensitivity and scalability.

The two prediction modes of miRferno allow the user to optimize for time versus sensitivity. The greedy mode is de- signed to be fast but less sensitive. In this mode, multiple seeds are extracted from the miRNA or sRNA sequence. These seeds are 6 nt in length and extracted in 4 nt intervals, and they are aligned to the FM indexes from the partitioned feature set with a maximum allowed mismatch of 1 nt. Matched instances of these seeds are further extended to complete the alignment of the small RNA, unless three consecutive seed extension attempts fail, resulting in the termination of the extension. On other hand, the exhaustive mode is designed for improved sensitivity; it extracts a smaller seed of 4 nt spaced in a 3 nt interval from miRNA or sRNA sequence. The use of multiple 4 nt seeds from a single miRNA or

sRNA along with one allowed mismatch improves the efficiency of finding targets, as the probability is high of at least one seed (out of seven in total, for a 21 nt small RNA) being extracted from the region of the miRNA which binds its target at a region with 3 nt matches. In addition to sensitive mapping parameters, if none of the extracted seeds reports a valid alignment, then a second, 're- seeded' pass is allowed. In second pass, a new set of seeds is generated, slightly offset, and used to search for targets.

miRferno also offers the user two different systems for target scoring, standard and seed-free. Standard scoring pro- vides backward compatibility for earlier miRNA-target pre- diction or validation experiments; in other words, it is based on previously described, complementarity rules based on a seed region (Fahlgren and Carrington 2010). However, we added the seed-free scoring because several recent studies have shown that there exist miRNA-target interactions which deviate from the standard or canonical complementarity rules that utilize a seed region (Y. Zheng et al. 2012; Brousse et al. 2014). Seed-free scoring may have broader utility: several early (Ha, Wightman, and Ruvkun 1996; Wightman, Ha, and Ruvkun 1993) as well as recent studies (Didiano and Hobert 2006; Chi et al. 2009; Chi, Hannon, and Darnell 2012; Z. Xia et al. 2012; Khorshid et al. 2013) from animals also indicate that formation of a functional miRNA-target duplex does not require strict complementarity between a miRNA seed and its target. These non- canonical targets in both plants and animals have been validated and support an 'expanded' range of miRNA-target interactions. Moreover, the targets sites from IGRs are of- ten left unanalyzed because target-prediction tools focus on annotated genes; poorly annotated non-coding RNAs may interact differently with miRNAs in ways that are not yet well defined. So, we wanted

to avoid over-fitting of complementarity rules based on seed regions that might not only restrict our ability to find non-canonical targets but also introduce bias into the results. The seed-free scoring achieved this, based on the assumption that a target site could be functional even with weak seed-region complementarity. Therefore, unlike the standard scoring system, within this region, G:U wobbles, gaps and mismatches have the same penalty score as elsewhere in the miRNA-target pairing. Finally, in the seed-free scoring system, mismatches at the critical 10th and 11th positions are permissible (Y.-F. Li et al. 2010; Nakano et al. 2006). While the seed-free scoring system relaxes many of the conventional miRNA-target interaction constraints, by assigning strong mismatch penalties, it retains a requirement of a correlation between sequence complementarity and cleavage efficacy. Each miRNA-target alignment is scored using following position specific rules, starting from the 5' end of the miRNA:

(i) mismatches at either the 10th or 11th positions carry a penalty of 2.5,

(ii) a wobble with a single flanking mismatch or mismatches on both sides carries a penalty of 1.5 or 2.0, respectively and

(iii) a single gap, mismatch and wobble at any position carries a penalty of 1.5, 1.0 or 0.5, respectively.

Finally, in *sPARTA*, the Bowtie scoring system was modified to reject miRNA-target alignments with more than one gap or six 'edits' (mismatches or G:U wobbles). These settings are user-configurable and can be relaxed using the depth parameter with input values ranging from 0 (default) to 3 (relaxed).

2.1.4 Indexing and prediction of validated interactions

In the final step of *sPARTA*, the PARE read abundances and positions are assessed relative to the predicted miRNA or sRNA targets, with the aim of validating

'real' cleavage events. First, for each PARE library, map files generated for all partitions of the feature set (from the second step of *sPARTA*) are combined and transformed into an index. This PARE-Genome (PAGe) index is specific to PARE libraries and consists of coordinates in which the 5, end of the PARE reads is mapped to the genic or intergenic feature set, along with the read abundance. PAGe indexes are used to classify the mapped reads (the evidence of cleavage at a specific site) into separate classes on the basis of their abundance (the strength of this evidence of cleavage). For a genic feature set, *sPARTA* implements the same signal classification schema described in earlier studies (Addo-Quaye, Miller, and Axtell 2009; Folkes et al. 2012). This schema uses five 'classes' to rank the evidence of cleavage based upon normalized or raw tag count input file; in other words, each PARE read in a gene is assigned to one of the five classes:

Class 0 indicates a PARE signal with abundance greater than one read that is also the maximal signal on the transcript; this is ultimately the most promising site for miRNA-directed cleavage.

Class 1 is similar to class 0 except there exists more than one maximal PARE signal on the transcript with the same abundance.

Class 2 is a PARE read above the median for the gene, and with an abundance of more than one read.

Class 3 is a PARE read below the median, but still with an abundance of more than one read.

Class 4 are PARE reads with an abundance of one, essentially not discernable from 'background'.

IGRs may be more challenging to analyze than genic regions, for the purposes of finding and validating sRNA tar- gets, for several reasons. First and foremost, a single IGR might contain more than one transcript and these transcripts could be coordinately regulated, potentially by one or more miRNAs. In such a scenario, a transcript with the highest expression could impact the signal of a weakly expressed transcript sharing the IGR, pushing it to a lower class (2 or 3) and thus diluting the score and detection of a genuine cleavage event in the weaker transcript. Second, an IGR may contain no transcripts cleaved by an sRNA, with the only mapped PARE signals resulting from decay or non- specific effects; in absence of strong signal from a cleavage event, these low strength signals will be assigned to class 0 or 1 thereby inflating these top categories and confound ing the calculation of the confidence score. Therefore, to fit the variable and difficult-to-assess nature of IGRs, *sPARTA* classifies PARE signals on the basis of the global abundance of PARE reads. For each PARE library, the signals in the bottom 20% (by abundance) are assigned to class 4 and excluded from further calculations. The remaining signals are then classified as follows:

Class 0: > 90th percentile (of all PARE read abundances)

Class 1: 90th percentile PARE read abundance > 75th percentile

Class 2: 75th percentile PARE read abundance > median (50th percentile)

Class 3: median ≥ PARE read abundance

sPARTA calculates the confidence score (*p-value*) as de- fined in *CleaveLand* (v3, or ’CL3’) but with slight modification so as to improve the *p-value* for cases where miRNA- target interactions have weak complementarity or when a single miRNA cleaves hundreds of targets, for example, the miR2118 or miR2275 targets

described previously for rice and maize (13). This *P-value* is further corrected for the noise around the cleavage site. The calculation of the P- value is as follows:

P-value (at least one significant result)= 1 - pbinom (0, trials, probability of success)

Corrected *P-value* = *P-value* of an interaction/signal to noise ratio

Where,

Trials = total number of miRferno predicted targets within a score bracket, i.e. the number of predicted targets with score 5 and <6, instead of cumulative number of predicted targets for a miRNA at specific score as in CL3.

Probability of success = fraction of total (eligible) bases in the feature set occupied by a specific degradome class (7).

And,

P-value of an interaction = PARE-validated interaction with *P-value* <0.25 and signal-to-noise ratio >0.25

Signal to noise ratio = fraction of PARE abundance at cleavage site in a 10 nt window around the cleavage site (5 nt in each the 3, and 5, directions).

This relaxed *p-value* calculation gives more weight to the evidence from PARE data and it yields a greater number of validated targets as compared to CL3, but it could also have a higher proportion of false positives. We believe that this trade-off can be reasonably reduced by either (i) including replicates of PARE datasets (Folkes et al. 2012) or (ii) by establishing the anti-correlation in expression levels between miRNA and their targets.

Finally, for the analyses that we described here, publically available PARE, sRNA and RNA-seq datasets for *A. thaliana*, *Oryza sativa*, *Medicago truncatula* and *Brachypodium distachyon* were downloaded from NCBI GEO (**Table 2.1**). *sPARTA* (in the seed-free mode) was used to generate species-specific sets of PARE-validated miRNA-target interactions. The back-end for the *comPARE* web resource, which

stores the data and perform searches, consists of a relational database implemented with MySQL on CentOS release 6.4. The graphical user interface (GUI) was developed in PHP for seamless integration with our customized genome browser (Nakano et al. 2006) for visualization as well as in-depth exploration of data from different sources such as PARE, small RNA, RNA-seq (when available) integrated with genomic annotations and features.

2.2 *sPARTA* - RESULTS

To assess the performance of *sPARTA* (greedy mode), real datasets (PARE, small RNA, genomes and miRNAs) were used to determine metrics, as it would be in an actual miRNA target identification experiment. Publically- available PARE datasets generated using Illumina sequencing from four different species (*A. thaliana*, *B. distachyon*, *M. truncatula*, and *O. sativa*; (**Table 2.1**) were downloaded from our Massively Parallel Signature Sequencing database (Nakano et al. 2006), and the corresponding genome sequences and annotation information were fetched from their respective repositories (**Table 2.1**). miRNA sequences for all four species were downloaded from miRBASE (version 20). *CleaveLand* (version 3, CL3) and *PAREsnip* (v.2.1) are currently the only publicly-available, command line tools for PARE- based miRNA-target validation. We used CL3 for comparative benchmarking primarily because it's is the most cited.

Table 2.1: **Small RNA and PARE data used in *sPARTA* benchmarks**

Species	miRNAs	Annotation version	PARE datasets	Small RNA datasets
A. thaliana	337	TAIR 10.0	GSM280226 GSM280227	None used.

B. distachyon	882	MIPS 1.0	BDI25, BDI20 (15)	GSM506621,GSM506620
M. truncatula	599	JCVI 3.5	MEDFL3 (35), GSM643818, GSM643817	GSM767273, GSM769274, GSM769275, GSM769276, GSM729277, GSM729279
O. sativa	713	MSU 7.0	GSM476257 GSM434596	None used.

2.2.1 Evaluation of *sPARTA* runtime performance

sPARTA was evaluated on a machine equipped with four 64 bit 8-core 2.4 GHz Intel Xeon (32 cores total) running CentOS release version 6.4. Python 3.3 and R 3.0 (R Development Core Team 2011) were installed 'as is' available from their respective sources. In the comparisons below, the added time to extract features, i.e., genic or intergenic transcripts, from the genome is not included as this feature is not present in any available tools. All the runtimes reflect an average of five independent trials.

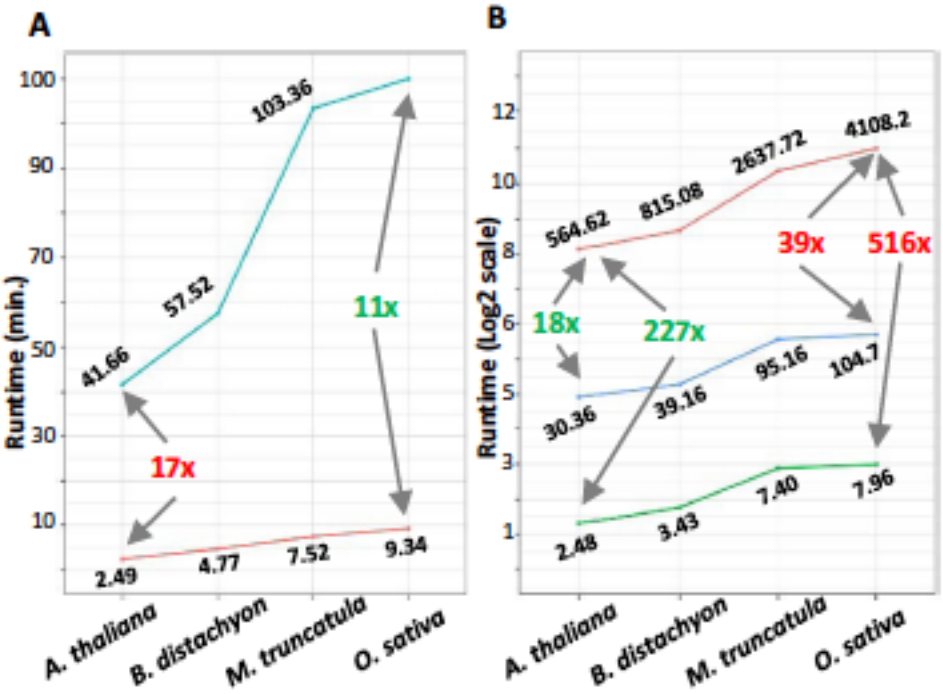


Figure 2.2: **Comparative benchmarking of the *sPARTA* algorithm in parallelized mode and in comparison, to *CleaveLand* version 3 (CL3)**. In both comparisons, four different plant genomes were used, as indicated on the X-axis. In each set of pairwise run comparisons, the minimum fold difference is indicated in green text and the maximum in red text.

A. Runtime comparisons between *sPARTA* in serial (blue line) and parallel (red line) modes exhibit a minimum speed gain of 16.8× and maximum speed gain of 11× for the parallel mode compared to the use of a 28-core single node.

B. *sPARTA* run in parallel mode (green line) is a minimum of 227× and maximum 516× faster than the comparable software package CL3 (red line). Using a single core (blue line), the *sPARTA* package is a minimum of 18× and a maximum 39× faster than CL3.

We first evaluated the total time required by *sPARTA* to predict and validate targets at a whole-genome level for all four species. All available miRNAs for each species were used for target prediction, followed by validation using two separate

PARE libraries (**Table 2.1**). Two different scenarios were tested: sequential and parallel. As the name suggests, the sequential run used just one core for the analysis, whereas the parallel run utilized 85% of the available cores (n = 28). For the total runtimes, we excluded the execution time for the step which maps the PARE dataset to the genome, as mapping step is performed by Bowtie (Langmead and Salzberg 2012), for which the settings can optimize to run using the same number of processors as the *sPARTA* parallel mode. The comparison demonstrated a minimum speed gain of 10.56 with the genic feature set of *O. sativa*, and a maximum speed gain of 22.31 with the intergenic feature set of *A. thaliana*. At a whole-genome level, a maximum speed gain of 16.7x and minimum speed gain of 11.2x was achieved by the parallel mode of *sPARTA* (**Figure 2.2A**).

Next, we compared *sPARTA* performance to *CleaveLand* (CL3) which is the most-widely used tool for the evaluation of plant miRNA targets. CL3 consists of two sequentially executed scripts, requiring input from two third-party tools, *TargetFinder* (Fahlgren and Carrington 2010) and Bowtie (Langmead and Salzberg 2012). To enable a comparison with *sPARTA*, we implemented the CL3-based pipeline using its bundled scripts and required tools, with no modification to those original scripts or settings. For the fairest comparison between algorithms, the PARE mapping step for CL3 was assigned the same number of cores as *sPARTA*. CL3 lacks the functionality to predict intergenic targets, therefore a comparison was made just for the genic feature set. Outperforming CL3, *sPARTA* exhibited a minimum speed gain of 227.39x (564.62 to 2.48 min) with *A. thaliana* and a maximum speed gain of 515.12x (4108 to 7.964 min) with *O. sativa* (**Figure 2.2B**). Even in the serial mode, *sPARTA* was found

to be a minimum 18x (564.62 to 30.36 min) and maximum 39.5x (4108.2 to 104.7 min) faster than CL3 with *A. thaliana* and *O. sativa* respectively (**Figure 2.2B**).

2.2.2 Prediction performance of *sPARTA*

Strong experimental support is required to validate miRNA-target interactions identified by *sPARTA*. Such experimental data may be either modified 5, RACE, applied to individual targets, or genome-level data sets from PARE, an extension of 5, RACE to the genome level. For PARE data, there are a number of earlier miRNA-target validation studies (D. H. Jeong et al. 2013; Gong et al. 2013; Rymarquis, Souret, and Green 2011). Yet, since these earlier studies were also computational (i.e. had their own set of parameters for PARE validation), their sensitivity is unknown and there- fore cannot be used as a 'gold standard' to calculate the degree to which the *sPARTA* predictions generated false positives or false negatives. Moreover, since there is no earlier published approach or tool to cross-validate miRNA tar- gets from IGRs, it is not possible to appraise the sensitivity of these intergenic predictions from *sPARTA*. In the con- text of these limitations, we performed an assessment of the predictive power of *sPARTA* by comparison to CL3.

A subset of plant miRNAs, including miR2118, miR2275, miR173 and miR390 (Johnson et al. 2009; Allen et al. 2005; Axtell et al. 2006) induce the production of secondary siRNAs in a phased arrangement from their target RNA transcript, via the recruitment of RDR6 and DCL4 or DCL5. The start site of the register of this phasing is determined by the position of miRNA-guided cleavage. Since the presence of phased sRNAs (phasiRNAs) from a locus indicates a real miRNA- target interaction occurred, finding a PARE-validated trigger site that was responsible for phasiRNA production further supports the validity of some miRNA-

target inter- action. We used this cross-validation of a computation- ally predicted miRNA-target interaction with phasiRNA ('*PHAS*') loci as the basis to assess and compare the ability of different software tools to identify miRNA target sites.

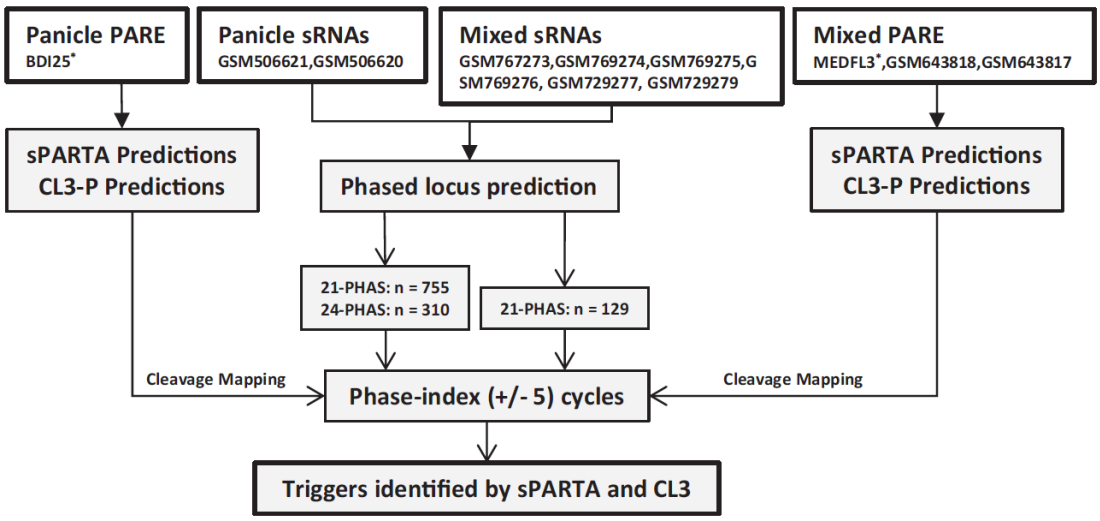


Figure 2.3: **Our approach to assessing the comparative benchmark of the prediction power.** Loci generating phased sRNAs were identified from published small RNA datasets of *B. distachyon* and *M truncatula*, while genome-wide target prediction and validation was performed using their associated PARE datasets against all species-specific miRNAs. GEO accession numbers are indicated in the top row of boxes; asterisks indicate data either from https://mpss.danforthcenter.org /brachy_pare2 or https://mpss.danforthcenter.org/mt_pare/. Triggers of phased sRNA loci validated by *sPARTA* and *CL3* were identified and used for a comparison of predictive power.

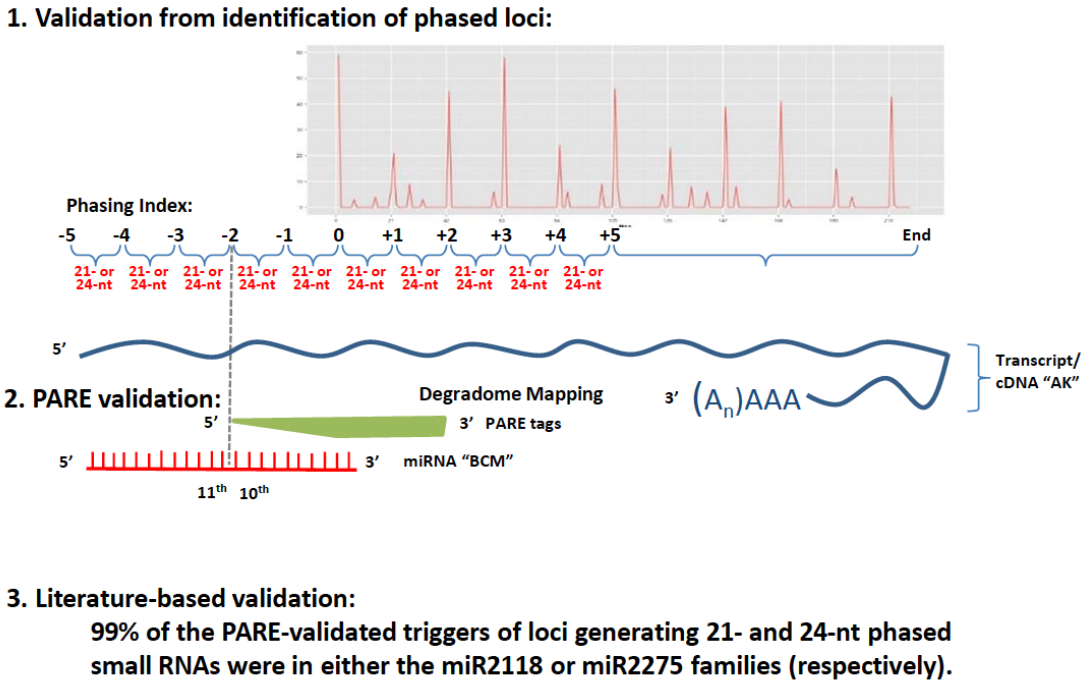


Figure 2.4: **Loci generating phased small RNAs were used in the comparison of predictive power.** The phase-index consisting of 11 coordinates (+/- 5 cycles), corresponding to a phase (21 or 24-nt) periodicity from the initiation site of the phased locus, or site at which the miRNA cleaves to trigger phasiRNA biogenesis. Triggers were identified by searching for miRNA-target interactions with cleavage sites matching a specific phase-index.

Two recent studies have reported many 21-nt phased loci from genic regions of *M. truncatula* and both 21- and 24-nt phased loci from IGRs of *B. distachyon* (D. H. Jeong et al. 2013; Zhai et al. 2011). Using the small RNA data from these studies, we repeated those analyses to identify a total of 310 (24-nt phasing) and 755 (21-nt phasing) *PHAS* loci were identified from IGRs of *B. distachyon*, and 129 (21-nt phasing) *PHAS* loci from genic regions of *M. truncatula* (**Figure 2.3**). For every phased locus, an index of potential miRNA target sites was generated. This index

consisted of 11 coordinates (+/ -5 cycles) in correspondence to the phased (21- or 24-nt) periodicity from the initiation site of phased locus (**Figure 2.4**). PARE datasets from both studies were used to generate a list of PARE-validated targets against all miRNAs for each species, using *sPARTA* and CL3 independently. Though the CL3 functionality is limited to transcriptome or genic regions, our aim was to compare an existing algorithm with *sPARTA* to assess its advantage or disadvantage in its prediction power. For this particular analysis, we rectified one of the main technical shortcomings of CL-the based pipeline by capacitating a parallelized prediction of targets; for this, we developed a parallelized version of *TargetFinder* (Fahlgren and Carrington 2010). No changes were made to the target prediction scoring schema, so as to retain the original approach of CL3. Finally, triggers of phased loci were identified by searching for a match between the *PHAS*-index of an individual locus and validated cleavage sites from both CL3 and *sPARTA*.

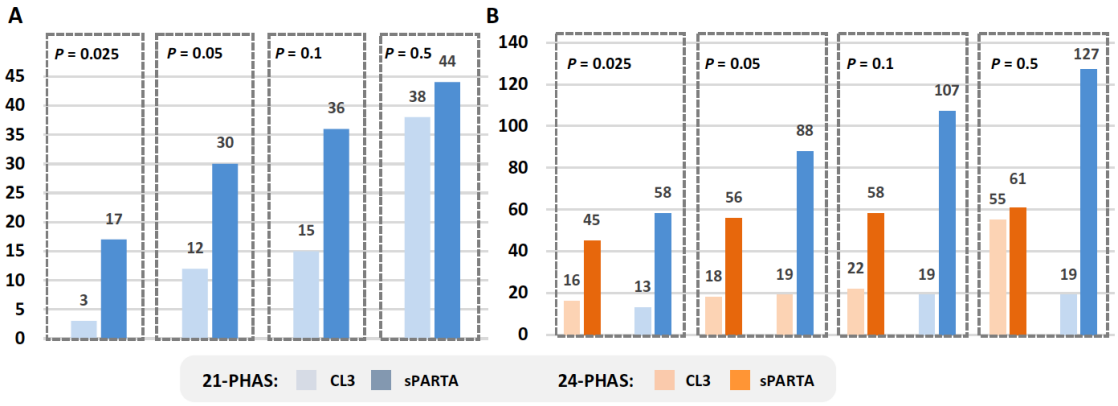


Figure 2.5: ***sPARTA* validates more triggers and exhibit high *p-value* enrichment as compared to *CL3***. We performed comparative benchmarking of the predictive power of *sPARTA*, as outlined in Figure 2.3. (A) In an analysis of only 21-*PHAS* loci from genic regions from *Medicago truncatula*, *sPARTA* identified 2.5 times more miRNA triggers than CL3, with 68% of correct validations under a *p-value* of 0.05. (B) For 21- and 24-*PHAS* loci from intergenic regions of *Brachypodium distachyon*, *sPARTA* identified 3 and 4.5 more miRNA triggers with 70 and 90% of correct predictions under a *p-value* of 0.05, respectively.

sPARTA demonstrated advantages over CL3 by identifying more triggers, as well as by exhibiting a high enrichment in *P-value* of correct predictions. In the case of phased loci from IGR of *B. distachyon*, 3-fold (total 56) and 4.5-fold (total 88) more triggers were validated by *sPARTA* (*P- value* 0.05) for 24- and 21-phased loci respectively (**Figure 2.5A**). Of all the miRNA triggers identified by *sPARTA*, 70% of 21-phased and 90% of 24-phased triggers were predicted under a *p-value* of 0.05. Interestingly, miR2118 was identified as a trigger in 126 out of 127 validations of 21-phased loci whereas, for 24-phased loci, miR2275 was identified as a trigger in all the validations. This is consistent with earlier reports of the miR2118 and miR2275 families (Song, Li, et al. 2012, 4; Zhai et al. 2011) as triggers of reproductive-specific 21- and 24-nt phased loci, respectively. This observation further supports the

robustness of our approach used for the comparative benchmark of predictive power, by showing that PARE-validated triggers of phased loci are not products of chance. For phased loci from genic regions of Medicago, *sPARTA* identified 2.5-fold more triggers under a *p-value* of 0.05 as compared to CL3 (**Figure 2.5B**). As in *B. distachyon*, *sPARTA* exhibited an enrichment of *p-values* by predicting 68% of 21-phased loci triggers under a *P-value* of 0.05.

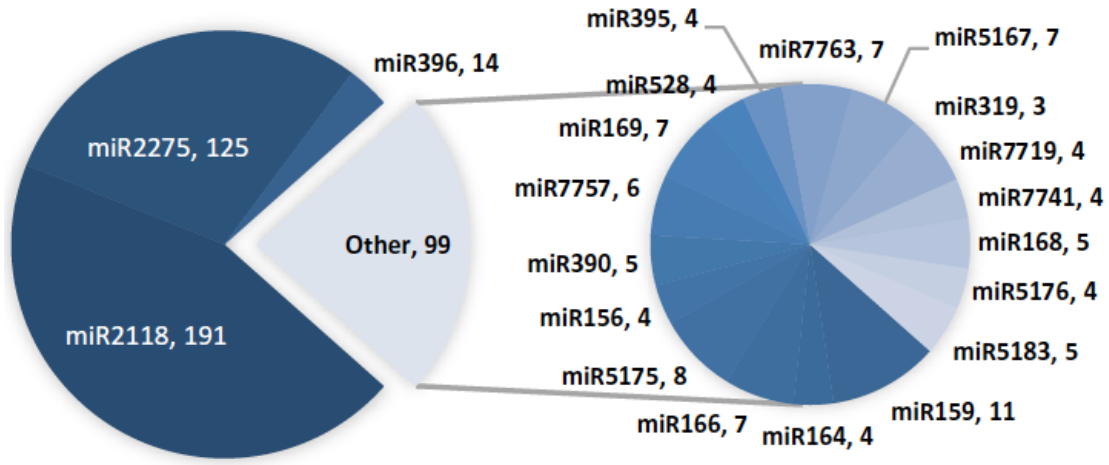


Figure 2.6: **Intergenic targets in *B. distachyon* for 80 different miRNAs**. A total of 506 credible intergenic targets were validated in *B. distachyon* from root, leaf, stem and panicle tissue. miRNAs bdi-miR2118 and bdi-miR2275 accounted for half of the intergenic targets. The pie charts show miRNA families with more than three targets, with the number of targets following the miRNA name.

2.2.3 Targets identified from intergenic regions

A total of 506 credible targets (with a corrected *P-value* ≤ 0.05, degradome signal class ≤3) for 70 different miRNA families (**Figure 2.6**) were identified from the IGRs of B. dis- tachyon, using the published PARE datasets from root, leaf, stem and

panicle tissues (D. H. Jeong et al. 2013). These targets would certainly have been missed by existing PARE validations tools as those tools are limited to analysis of just the annotated genic regions. We also found multiple targets from a single IGR, each with different expression dynamics; our approach for the classification of PARE reads mapped to IGRs was developed with this scenario in mind. From the total set of intergenic targets, the panicle data alone accounted for most validated interactions (n = 344) with 157 and 114 unique cleavages triggered by just two miRNA families, miR2118 and miR2275, respectively. Both miRNA families are known to trigger phasiRNA biogenesis (Johnson et al. 2009). In 48% of cleavage site identified, we found an overlap of the cleavage site within +/-5 phased positions or 'indexes' from the dominant register of phasing, i.e. the position with the highest phasing score. For those cleavage sites which did not match with the phased index, upon inspection, we found presence of a phased locus in close vicinity (250 nt). The reason for this disagreement between the cleavage site and phase index could be the depletion of some sRNAs from a few phasiRNA cycles, consistent with the non-stoichiometric abundances of tasiRNAs from Arabidopsis TAS loci, leading to a shift in the predicted position of the predominant register for the phasiRNAs. We also observed PARE validation of cleavage by miR2118 and miR2275 in leaf with the same cleavage co-ordinates as panicles. For these interactions shared with those that lead to phased sRNA generation in panicles, no associated phased sRNAs were found near the cleavage site in leaf, yet the abundance of PARE reads at the cleavage sites indicates strong expression of the precursors in both panicle and leaf tissues. These observations suggest that there are other factors influencing the production of phased sRNAs.

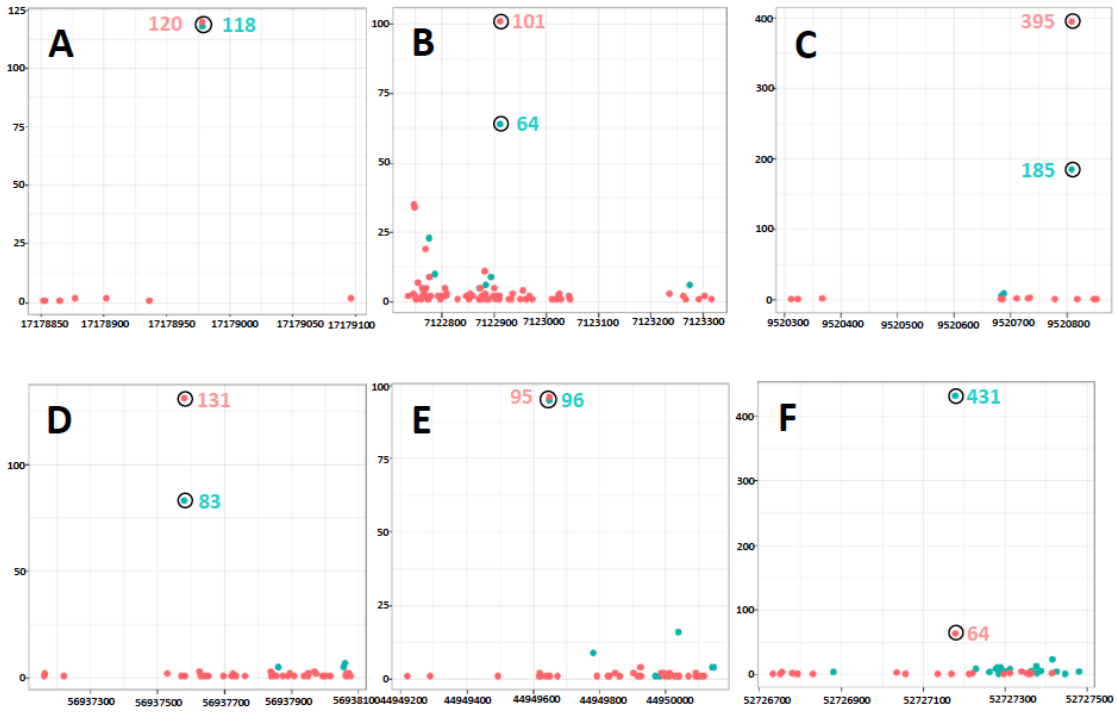


Figure 2.7: **miR396 coordinates cell proliferation in leaf meristem by** regulating transcription factors belonging to the family of GROWTH-REGULATING FACTOR (GRF). Plots of PARE data (D-Plots) mapped to genomic regions with cleavage sites highlighted for genic targets of bdi-miR396 in *B. distachyon*. Green dots indicate PARE reads from leaf libraries, and red dots are from panicle libraries. The numbers indicate abundance of reads (in TP15M). A) Bradi4g16450 (GRF-8 like), B) Bradi1g09900 (GRF-6 like), C) Bradi1g12650 (GRF-9 like) is shared between panicle and leaf. These targets encode proteins in the family of Growth Regulating Factor (GRFs). D, E and F) Examples of novel intergenic targets of miR396 from *B. distachyon* shared between leaf and panicle.

Unlike miR2118 and miR2275 families, whose activity was found to be conserved to leaf and panicle, a few miRNAs like miR396 shared targets across the different combination of tissues. miR396 has been previously demonstrated to coordinate cell proliferation in leaf meristem by regulating transcription factors

belonging to the family of growth- regulating factor (GRF) (Rodriguez et al. 2010). Another transcription factor, bHLH74, crucial to margin and vein pattern formation of Arabidopsis leaves has been found to be a target of miR396 (Debernardi et al. 2012). Recently, it was reported that the miR396 regulatory network and tasiRNA biogenesis pathway synergistically interact to regulate leaf development (Mecchia et al. 2013). We found miR396 to be highly expressed not only in leaf but also seedling, stem and panicle of *B. distachyon* (D. H. Jeong et al. 2013); it is also found in roots but at a comparatively low level. In panicle, a total of nine validated targets of miR396 were identified from genic (n = 4) and IGRs (n = 5). All four genic targets from panicle were found to be a member of GRF family (**Figure 2.7**) like earlier published studies on leaf development. Moreover, the PARE signal at the cleavage site of all four of these targets belonged to class 0, i.e. PARE read abundance 90th percentile of all PARE reads mapped to the genic regions (**Figure 2.7** A, B and C), suggesting moderate expression of cleaved GRF transcripts. There are also several IGRs (**Figure 2.7 D, E** and **F**) with strong signals of miR396 activity, highly enriched in the panicle. These observations indicate that in addition to leaf development, miR396 might also play a role in panicle development.

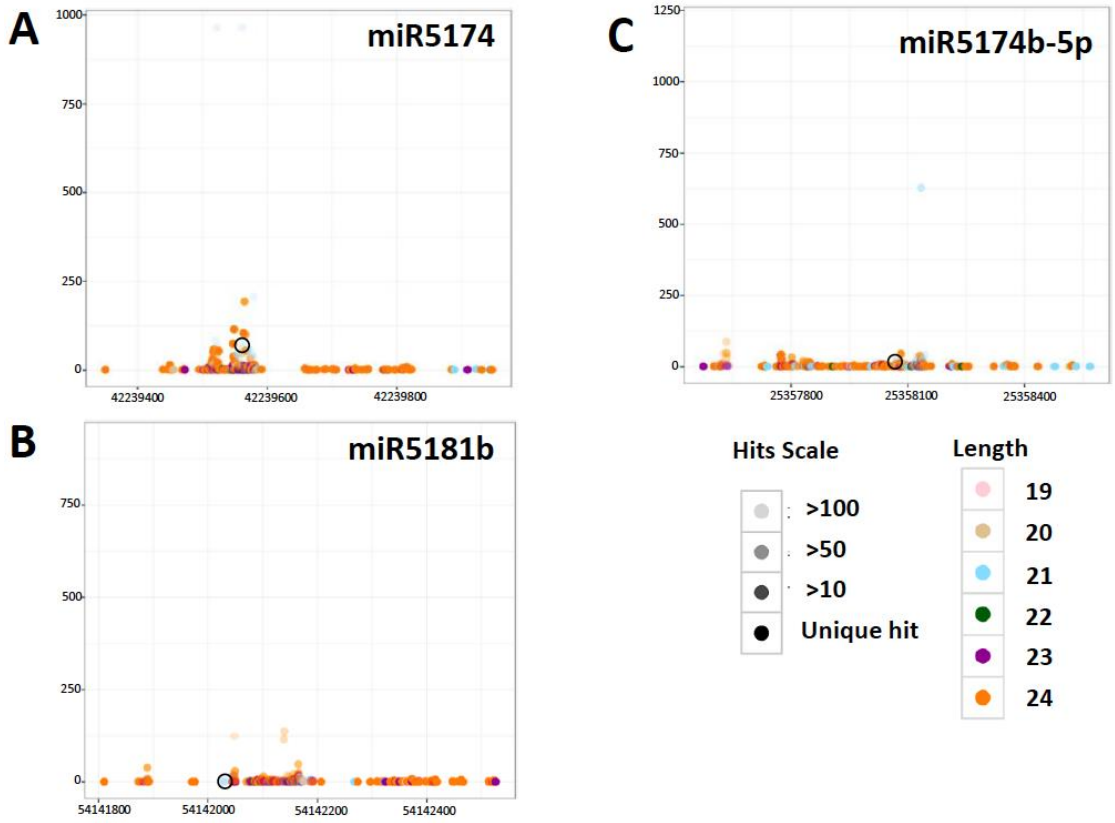


Figure 2.8: **MicroRNAs from families miR5174 and miR5181 originate from repetitive regions, rich with heterochromatic (24-nt) small RNAs**. A resource which allows visualization of miRNAs and their targets in genomic context is sought to allow manual review of miRNAs in online repositories. At the bottom is a legend indicting that the intensity of the fill color indicates the hits (genome matches), while the different colors indicate the small RNA sizes. A) bdi-miR5174, B) bdi-miR5181, C) bdi-miR5174b.

In the process of these analyses, we noted that reliance on annotated miRNAs without their critical assessment can lead to spurious conclusions. As an example, three annotated miRNA families, miR5174, miR5181 and miR5180 (Baev et al. 2011), accounted for the greatest number of validated tar- gets (n 132), after the miR2118 and mi2275 families. Further inspection of these miRNAs revealed that they originate

from repetitive regions, rich in heterochromatic small RNAs (24 nt) and their abundance is quite low in the libraries used for prediction (**Figure 2.8**). The approach (Baev et al. 2011) implemented to annotate these three miRNA families used default Bowtie parameters therefore only first valid sRNA alignment to genome was reported instead of all the mappings of sRNA to the genome which lead to 'clustering' with incomplete set of small RNA mappings, also no hit- or abundance-based filter was applied to re- move lowly expressed or reads with large number of hits to genome. Moreover, through *sPARTA*-based analysis, all of the targets of these three families were found to be in highly repetitive regions. These data strongly suggested that these are incorrectly annotated miRNAs; as miRNAs are largely predicted computationally using different pipelines and parameters, mostly by small RNA sequencing datasets and submitted to miRBASE without experimental validation, researchers need to be wary of such false predictions. The presence of such spurious miRNAs in public repositories suggested the need for a resource which allows visualization of miRNAs and their targets in their genomic contexts, to allow manual inspection.

2.2.4 The *comPARE* web interface

We developed a web-based tool, which we call '*comPARE*', for two purposes: (i) to serve as a single point of access for plant miRNA-target interactions that we have validated with PARE data, (ii) to facilitate connections of those data to our custom-built genome browser, specialized for small RNA (Nakano et al. 2006). This interface is designed to be easy to use, yet incorporate advanced functionality such as modifiable search parameters, combined searches of sRNA or PARE datasets, and analysis of library-based data. The *comPARE* site is accessible at: https://mpss.danforthcenter.org/tools/mirna_apps/comPARE.php

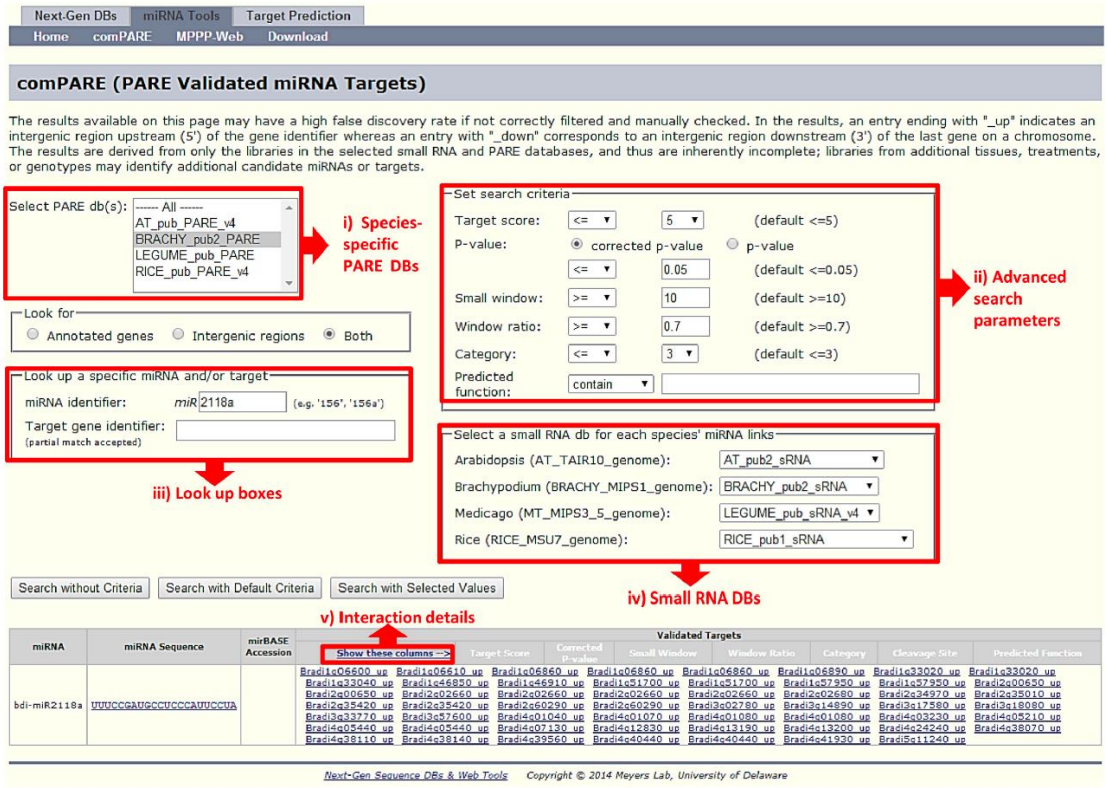


Figure 2.9: **The interface to *comPARE*, web-based access to PARE-validated sets of miRNAs targets**. A screenshot of the *comPARE* web interface. The red boxes highlight different types of user options. For example, in the upper left (i), the user can choose single or multiple species specific PARE databases to search for miRNA-target interactions. In the upper right (ii), in advanced search could be performed by setting the search parameters as per the required confidence level. In the lower left (iii), for a miRNA or target of interest, a search could be executed using a miRNA name and/or genome-specific target identifier as a query. Lower right (iv), if these options are listed, multiple sRNA databases for a species of interest other than the initial selection could be made. Finally (v), at the very bottom, the links, if clicked, display additional information about each interaction.

To use this site, shown in Figure 2.9, first, a user chooses the PARE database for species of interest from the main query page, additional information about the available databases and included libraries are found at our lab's main page

https://mpss.danforthcenter.org/. For specific miRNAs and targets of interest, their identifiers are entered into the respective textboxes on the main query page, generating results by clicking on 'Search with default values'. This then will display all the interactions that pass the criteria, set by default for convenience. A more advanced search could be performed by modifying the values of search parameters, including the miRNA-target complementarity score (Target score), *P-value* cutoff, normalized abundance of the PARE signal at cleavage site (small window) and signal class (see 'Materials and Methods' section); the search is executed by clicking on 'Search with selected values'. The results for both a simple or modified query are presented in a simplified table format consisting of the miRNA name, miRNA sequence and the list of targets. However, a detailed view can be opened by clicking on 'Show extra columns' located in the header of the results table, which displays additional information including the target score, *p-value*, small window (a 1 nt region flanking the cleavage site), large window (a 5 nt region flanking the cleavage site), signal class, the cleavage site coordinates, and the annotated function of the target (if available) for each interaction that passed the selected or default search parameters. A user can also search for all the interactions from the selected species with either modified or default search parameters. The results from the user query in *comPARE* then integrate small RNA and PARE data, layered on an annotated genome. This provides a comprehensive view of cleavage sites, facilitating an in-depth exploration of miRNA-target interactions.

In addition to searches, visualization and exploration of miRNA-target interactions, one of the main strengths of *comPARE* is that it enables the discovery of conserved miRNA targets across different species. This functionality is of high value

for revealing not only the evolutionary con- served targets of specific miRNAs but, most interestingly, the non-conserved targets of different species or libraries. A quick search with 'miR2118' from the main query page shows that its targets are unrelated, genic and intergenic in *M. truncatula* and *B. distachyon* respectively, in consensus with earlier studies (Zhai et al. 2011; D.-H. Jeong et al. 2013). Such cross-species contrasting patterns of miRNA targets are of high biological significance, and *comPARE* could aid in discovering these patterns as it allows identification of genome-wide targets for miRNAs from different species.

2.3 Availability

sPARTA source is freely available under GNU Public License (v3) from our GitHub page: https://github.com/atulkakrana/sPARTA. *sPARTA* is updated periodically and latest release can be downloaded from its 'release' page: https://github.com/atulkakrana/sPARTA/releases. To address user queries and for users to report any issues with *sPARTA*, we also maintain an issue reporting a tracking system here: https://github.com/atulkakrana/sPARTA/issues

2.4 Chapter summary

In this chapter, I

- developed a new miRNA target prediction algorithm "miRferno", to enable genome-wide prediction of hundreds to thousands of miRNAs in a reasonable time
- implemented a novel genome fragmentation approach, to enable SIMD-style parallelization at the most basic level
- implemented two different modes in *miRferno* – Heuristic and Exhaustive to match the user requirements

- implemented a ‘novel’ seed-free approach to *miRferno* to identify targets likely neglected by other tools

- developed a software application that contains *miRferno* and performs validation of targets using experimental PARE or degradome data

- developed a web-based tool “*comPARE*” to visualize miRNA targets integrations in data rich environment using the custom-built genome browser, specialized for small RNA

- used *sPARTA* with real data from three different species and compared the predictions as well as speed with popular alternative

- fixed the algorithmic shortcoming of popular alternative to enable its use at genome-wide scale, just to be able to compare the prediction performance of both tools.

I observed that

- *sPARTA* identifies 2.5-fold more triggers for protein-coding *PHAS* under a *p-value* of 0.05 as compared to CL3 (the popular alternative)

- *sPARTA* identifies 3-and 4.5-fold more triggers for pre-meiotic and meiotic *PHAS* under *p-value* ≤ 0.05 compared to CL3

- *sPARTA* algorithm is minimum 18x times faster than CL3 (the fastest CL version) even at single core

- *sPARTA* is very fast, up to 516x compared to CL3 in our benchmarks on 28-core machine

- *sPARTA* efficiently identifies PARE-supported targets from intergenic regions, which would be missed by other existing options as these can only scan annotated portion of genome i.e. genes for targets

From this work, I concluded that

- *sPARTA* is sensitive, scalable and fast compared to existing tools; it has the fastest and most resource-efficient algorithm to predict targets till date

- *sPARTA* is the “first” and “only available” tools for large scale and genome wide discovery of plant miRNA targets

- *sPARTA* maximizes the parallelization efficacy by optimizing parallelization schema based on genome-size, number of input miRNAs and number of PARE libraries

- *sPARTA* reduces the analysis time from days and hours to minutes and seconds

- *comPARE*, enables discovery, visualization and in-depth exploration of genome wide miRNA-target interactions in heterogeneous yet highly integrative environment

- collation of high-throughput small RNA and PARE datasets from different genomes further facilitates re-evaluation of existing miRNA annotations, resulting in a 'cleaner' set of microRNAs

Chapter 3

A HIGH PERFORMANCE SUITE OF TOOLS FOR IN-DEPTH CHARACTERIZATION OF PHASED siRNAs

Phased siRNAs (phasiRNAs) are secondary siRNAs that are widely prevalent across land plants, generated from both protein-coding transcripts and long, non-coding RNAs and in varying numbers per genome which ranges from tens to thousands. Integrated tools for in-depth characterization of *"PHAS"* loci have not yet been developed; and existing options are not only limited in number and function but also incompatible or inefficient in handling large volume of small RNA-seq data. In this chapter, we describe *PHASIS* suite which provides a complete tool set for discovery, quantification, annotation of phasiRNA loci or genes and rapid identification of their miRNA triggers. Benchmarks from five different species demonstrates that *PHASIS* is sensitive, exceedingly scalable and exceptionally fast software. Importantly, *PHASIS* can be run directly on transcript assembly and predicts miRNA triggers with high accuracy even without the PARE and degradome data, thereby eliminating the crucial requirement of assembled genome and experimental data for discovery of *PHAS* precursors, phasiRNAs and their triggers. The algorithmic novelty, flexibility to tailor analysis and the suitability for small to large-scale experiments makes *PHASIS* a *de facto* choice for discovery and study of phased siRNAs. The name "*PHASIS"* is from the ancient Greek city of Phasis, a destination for Jason and the Argonauts according to Greek mythology; we selected the name as it

links the colloquialism “phasis” as short for phasiRNAs, with the Argonaut proteins that bind them.

3.1 Methods

Lilium genome is not available. So, to use *PHAS*IS on transcriptome we used it mainly due availability of precisely stages samples from our study of phased siRNAs.

3.1.1 Sample Collection and RNA isolation

Flowering *Lilium* plants were purchased from Home Depot (Newark, Delaware). 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 as previously described (Mathioni, Kakrana, and Meyers 2016) . An aliquot of 3 µg of total RNA was used for size selection. Stages were assigned based on the morphology of archesporial (AR) and tapetal cells of *Lilium* anthers.

3.1.2 Single Molecule Real Time (SMRT) sequencing and transcriptome assembly

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, 2.0-4.0 kb, and > 4.0 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 11 SMRT Cells (Pacific Biosciences part # 100-171-800) were generated using the P6C4 polymerase (Pacific Biosciences part #100-372-700), five cells for transcripts < 2KB length, three cells for transcripts between 2KB and 4KB lengths and three cells for all transcripts > 4kb length. The sequencing was performed on *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 *pbscript-tofu* tool set (v2.3.0) using the default settings. 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 the CD-HIT with recommended parameters https://github.com/PacificBiosciences/cDNA_primer/wiki. The matches to protein

transcripts from Uniprot resource were summarized as described here: https://github.com/trinityrnaseq/trinityrnaseq/wiki/Counting-Full-Length-Trinity-Transcripts.

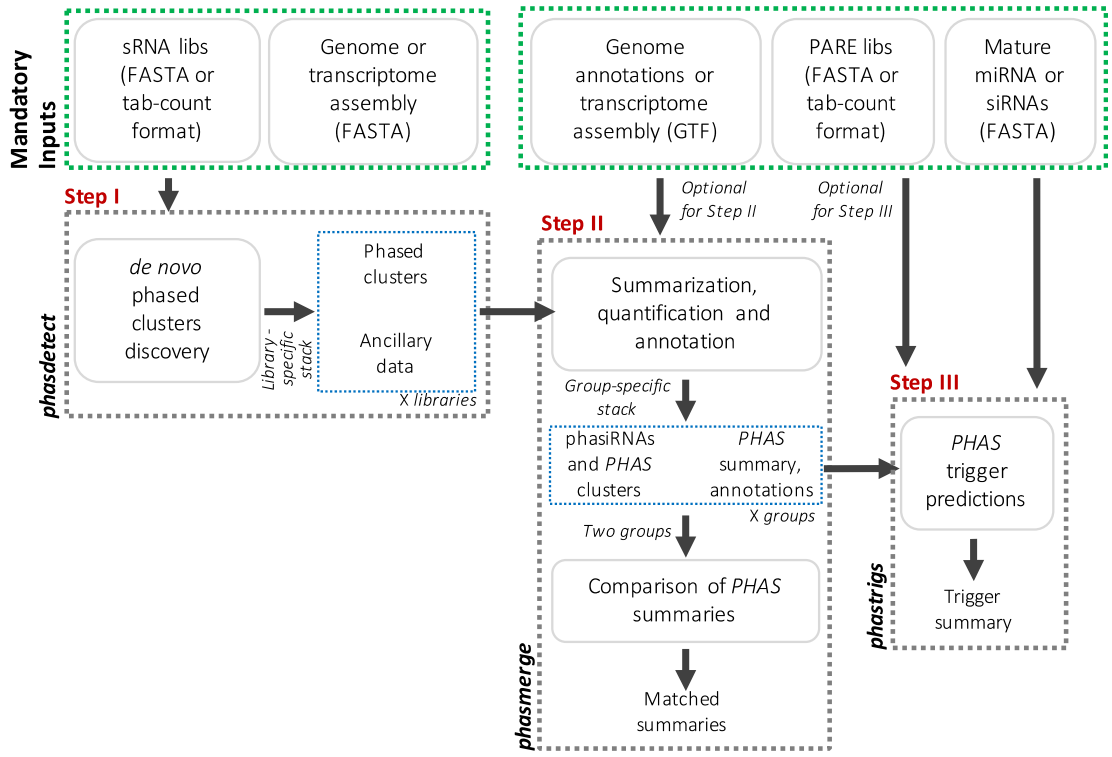


Figure 3.1: ***PHAS* loci or precursors transcripts are predicted through *phasdetect* in the first step**. The library-specific list of *PHAS* predictions can be summarized and annotated through *phasmerge* for libraries of interest into a *PHAS* summary. These summaries from two different groups can also be compared using "compare" mode of *phasmerge*. Triggers for *PHAS* summaries are identified through *phasmerge* either with PARE data in "validation" mode or without any experimental data in "prediction" mode. Selection between these two modes is made automatically based on a PARE library input or the lack of it. All analysis steps are independent and their execution depends upon the requirements of the user.

3.2 Approach and Features

PHASIS comprises three components that together perform *de novo* discovery, annotation, quantification, comparison and trigger identification for *PHAS* loci or precursor transcripts. We chose a modular approach over the single 'one-command' style for the following reasons: i) to maximize the flexibility for specific data or study requirements; ii) to integrate multiple, connected analyses; and, iii) to reduce overall runtime by maximizing phase- and step-specific parallelization. A description of these tools – *phasdetect*, *phasmerge*, *phasmerge* – in order of their utility or phases of study is provided below (see also **Figure 3.1**). *PHASIS* leverages the Python (v3) process-based "threading" interface to achieve efficient scalability and significantly reduce runtimes through parallel computing.

3.2.1 *phasdetect* – scalable and sensitive algorithm for large-scale survey of phasiRNA genes or loci

phasdetect performs *de novo* prediction of *PHAS* loci or precursor transcripts using user-supplied sRNA libraries along with a reference genome or transcriptome. It can efficiently process tens to hundreds of sRNA libraries in parallel, reducing runtimes. *phasdetect* operates via three main steps: i) first, sRNA libraries are normalized and mapped to the reference; ii) second, mapped sRNA reads are scanned to identify regions rich for specific size classes, such as those generated by Dicer activity (typically 21, 22, or 24 nt in plants); 3) finally these regions are stitched into clusters and the phasing of the small RNAs is computed as a *p-value*. We adopted a standard approach to compute *p-values* (Chen, Li, and Wu 2007). Parameters controlling these steps can be modified by users via the setting file "phasis.set", including values for *phase*, *mindepth* and *clustbuffer*; these refer to the phasing periodicity, minimum sRNA abundance to be included for *p-value* computation, and

the minimum distance separating two clusters. These parameters are explained in detail on the *PHASIS* wiki page (https://github.com/atulkakrana/*PHAS*IS/wiki/). The output for *phasdetect* includes library-specific list of *PHAS* loci (or transcripts) at several different confidence levels plus ancillary data, used to reduce runtime for subsequent analyses. For example, in case of a reanalysis after adding new libraries, *phasdetect* checks for any changes in parameters from the earlier analysis, assesses the integrity and compatibility of the ancillary data for, and reuses existing data to avoid repetition. This ancillary data also enables an array for downstream analyses and analysis-specific optimizations directly through *phasdetect*.

3.2.2 *phasmerge* – feature-rich tool to facilitate a tailored analysis, re-analysis and optimizations

phasmerge generates a summarization and performs a comparison between the *PHAS* summaries and annotations using the library-specific *PHAS* lists and ancillary data generated by *phasdetect*. These operations are selected by using the *-mode* option with the *merge* (default) or *compare* values. The *merge* mode prepares a *PHAS* summary for the libraries of interest, or for libraries that belong to different groups based on sample stages, tissues or treatments. The analysis can be tailored to meet the study requirements. For example, to maximize discovery, a user might set a lower confidence level (*p-value*) for summarization and consider all loci with a trigger predicted without the PARE data (identified through *phasmerge*) for downstream analyses. In contrast, a user motivated to maximize the quality might identify *PHAS* loci with the highest confidence level, followed by pruning of results with stringent quality parameters (described on the *phasmerge* wiki), and use *PHAS* loci that have PARE-supported triggers. *PHAS* summaries from different groups of libraries can be

compared using *compare* mode. This is particularly useful to identify intersecting and exclusive *PHAS* loci between different groups of stages, tissues or treatments. In *merge* mode, if an additional annotation file is provided, then merged *PHAS* loci are matched to genome annotations so as to identify coding *PHAS* loci or other available annotations. This function also supports quick discovery of precursor transcripts for summarized *PHAS* loci when provided with a GTF file generated from mapping the transcriptome assembly to genome. Furthermore, *phasmerge* attempts to determine the correct 5' terminus of *PHAS* loci by optimizing for the best 5' or 3' coordinates based on the user's sRNA data – a crucial functionality for determination of the correct miRNA trigger. *phasmerge* benefits from the modular *PHASIS* workflow, allowing users to optimize their results for the study which may vary in purpose, and making *phasmerge* independent from other tools.

The *phasmerge* workflow has three mandatory and two optional steps: i) via *merge* mode, *phasmerge* first generates a unique list of *PHAS* loci (or transcripts) for each user-specified library, by selecting predictions with the highest available confidence score (lowest *p-value*) that pass a user-supplied *p-value* cutoff, after comparing predictions from all available confidence levels; ii) *phasmerge* clusters the "best" candidate loci from specified libraries specific by the user, based on the degree of overlap in phased positions (or 'cycles') to select a representative locus for each cluster; finally, iii) *phasmerge* computes library-specific abundances, a size-class ratio, the maximum to total phasiRNAs abundance ratio, and other quality information. Optional steps include iv) *compare* mode, which first reads *PHAS* loci (or transcripts) from user-supplied summaries (n=2) and then identifies matching *PHAS* pairs based on the overlap in phased positions, to report a combined matrix including

both shared and unique loci in each *PHAS* summary file, and v) *merge* mode; when supplied with annotations, as described above, *phasmerge* matches a merged set of *PHAS* loci with genome annotations or with a genome-matched transcriptome assembly, both provided as GTF file, to report exonic or complete overlaps with annotated transcripts. This step requires prior installation of *SQLite* on user's machine. *phasmerge* generates several reports as output, most importantly, *PHAS* summary for libraries of interest which includes quality parameters (see online wiki for more information), FASTA files for size-specific siRNAs and all the siRNAs from phased positions along with detailed information on phased clusters with phasiRNAs, positions, associated *p-values*, etc.

3.2.3 *phastrigs* – an ultrafast and exhaustive algorithm for discovery of phasiRNA triggers

phasmerge identifies sRNA triggers for *PHAS* loci and precursor transcripts using the *phasmerge* summaries and a user-provided list of miRNAs (or any other small RNA). It was developed with the idea to minimize the requirement of experimental degradome (Addo-Quaye et al. 2008) or PARE (German et al. 2009) libraries. However, if such data ('PARE', henceforth) are provided, then *phasmerge* reports sRNA triggers with experimental support; these may be of higher confidence for some downstream experimental analyses. The strength of *phasmerge* lies in an algorithm designed to be both fast and exhaustive. It uses *miRferno*, an exhaustive target prediction algorithm that we developed (Kakrana et al. 2014) to predict targets sites for user-supplied miRNAs. The speed and precision of *phasmerge* is enhanced by a scan focused on 5' terminus of each *PHAS* locus (5'-end of the first cycle, the P1 position) for the trigger site, which reduces the search space and chance of reporting

false triggers. This 5’ terminus is inferred at the summarization step by *phasmerge* while collating data from different sRNA libraries. In the case of *PHAS* transcripts, only the 5’ terminus of the phased precursor is scanned, while in case of genomic *PHAS* loci, either the 5’ or 3’ end of the phased region is chosen, based on the strand targeted by a specific miRNA. *Phasmerge* analysis is divided into two main steps: i) *PHAS* transcripts or genomic sequences are extracted, and targets for user-supplied miRNAs are predicted; ii) next, a scan of phased positions located at the 5’ or 3’ termini of precursor for a target site that corresponds with the production of phasiRNAs is performed; this scan looks for target sites within ±3 nt of the ‘*PHAS* index’, defined as theoretical phased positions upstream from the 5’ terminus of P1. If PARE data is supplied, then PARE-validated cleavage sites are used for trigger identification. The *Phasmerge* report includes detailed information on miRNA-target interactions, PARE abundances at the predicted cleavage site, and the *PHAS* index of the predicted trigger site relative to the P1 position.

Table 3.1: **Comparison of features from existing tools that can predict phasiRNAs generating loci with the *PHAS*IS suite presented in this study**

Feature Categories	Software	*ShortStack*	*PhaseTank*	*PHASworks*
	Citation	*Axtell MJ (2013)*	*Guo et. al. (2015)*	*Present work*
PHAS prediction-related features	Tool-specific data format requirement?	no	yes	no
	Library wise results?	no	no	yes
	PHAS prediction in w/o genome assembly?	no	no	yes
	Group results based in stage, tissue or treatments?	no	no	yes
	PHAS comparison between groups?	no	no	yes
	PHAS annotation?	yes (from genome GFF only)	no	yes
	PHAS trigger prediction w/o PARE data	no	no	yes
Features unrelated to PHAS prediction	miRNA/hp loci prediction	yes	no	no
	whole genome report of sRNA clusters	yes	no	no

Table 3.2: **Distribution of sRNA and PARE reads along with SMRT sequencing transcripts from Arabidopsis, *Brachypodium*, *Lilium*, rice and maize.** For paired-end mRNA-seq, the number of read pairs listed correspond to read pairs. For SMRT-seq, the polished reads correspond to corrected, high-quality consensus transcripts

Part A. Maize small RNA and PARE data				
Code	**Title**	**Total Sequences**	**Length of Reads**	**Type**
MzRoots	Total RNA from maize root tissues	77,672,260	18-34	sRNA
0.2Fertile_r1	maize fertile anthers, 0.2 mm, rep 1.	12,896,366	18-34	sRNA
0.4Fertile_r1	maize fertile anthers, 0.4 mm, rep 1.	35,286,904	18-34	sRNA
0.7Fertile_r1	maize fertile anthers, 0.7 mm, rep 1.	34,045,564	18-34	sRNA
1.0Fertile_r1	maize fertile anthers, 1.0 mm, rep 1.	34,019,136	18-34	sRNA
1.5Fertile_r1	maize fertile anthers, 1.5 mm, rep 1.	37,521,389	18-34	sRNA
2.0Fertile_r1	maize fertile anthers, 2.0 mm, rep 1.	35,796,937	18-34	sRNA
2.5Fertile_r1	maize fertile anthers, 2.5 mm, rep 1.	55,306,867	18-34	sRNA
3.0Fertile_r1	maize fertile anthers, 3.0 mm, rep 1.	26,285,048	18-34	sRNA
4.0Fertile_r1	maize fertile anthers, 4.0 mm, rep 1.	40,210,297	18-34	sRNA
5.0Fertile_r1	maize fertile anthers, 5.0 mm, rep 1.	29,748,543	18-34	sRNA
PLNFertile_r1	maize mature pollen, rep 1.	33,201,391	18-34	sRNA
1_0w1_5	Mixed sizes of fertile anthers: 1.0 and 1.5 mm	50,968,596	18-20	PARE
2_0w2_5w3_0	Mixed sizes of fertile anthers: 2.0 mm, 2.5 mm and 3.0 mm	38,853,899	18-20	PARE
4_0wPollen	Mixture of fertile 4.0 mm anthers and mature pollen	38,138,281	18-20	PARE
Part B. Rice small RNA and PARE data				
58N_1	Young panicles of Nongken 58, a normal fertile line	17,707,755	18-34	sRNA
58S_1	Young panicles of Nongken 58S, a male-sterile line	19,936,967	18-34	sRNA
WT2003s	WT2003 wildtype leaf library from epigenome study (Stroud et al., 2013)	21,950,616	18-34	sRNA
YL9522_S3_1	Wild type rice, genotype 9522, anther development stage 3 (S3), rep 1.	76,357,541	18-44	sRNA
YL9522_S3_2	Wild type rice, genotype 9522, anther development stage 3 (S3), rep 2.	76,357,845	18-44	sRNA
YL9522_S3_3	Wild type rice, genotype 9522, anther development stage 3 (S3), rep 3,	76,353,223	18-44	sRNA
YL9522_S5_1	Wild type rice, genotype 9522, anther development stage 5 (S5), rep 1.	76,364,888	18-44	sRNA
YL9522_S5_2	Wild type rice, genotype 9522, anther development stage 5 (S5), rep 2.	76,380,262	18-44	sRNA
YL9522_S5_3	Wild type rice, genotype 9522, anther development stage 5 (S5), rep 3.	76,375,868	18-42	sRNA
YL9522_S7_1	Wild type rice, genotype 9522, anther development stage 7 (S7), rep 1.	76,024,297	18-44	sRNA
YL9522_S7_2	Wild type rice, genotype 9522, anther development stage 7 (S7), rep 2.	76,016,776	18-44	sRNA
YL9522_S7_3	Wild type rice, genotype 9522, anther development stage 7 (S7), rep 3.	76,019,313	18-42	sRNA
msp1_S3_1	msp1, genotype AF55, anther development stage 3 (S3), rep 1.	76,359,511	18-44	sRNA
msp1_S3_2	msp1, genotype AF55, anther development stage 3 (S3), rep 2.	76,358,885	18-42	sRNA
msp1_S3_3	msp1, genotype AF55, anther development stage 3 (S3), rep 3.	76,361,304	18-44	sRNA
INF939	Rice wildtype inflorescence degradome/PARE library	74,426,044	18-20	PARE
INF9311a	Rice inflorescence (93-11) wildtype degradome/PARE library	15,360,571	18-20	PARE
Part C. Brachypodium small RNA and PARE data				
BDI08	Root	76,139,430	18-35	sRNA
OBD03	Seedling	72,624,702	18-30	sRNA
BDI04	Leaf 1	76,207,568	18-34	sRNA
BDI09	Leaf 2	77,027,246	18-35	sRNA
BDI06	Stem	78,855,537	18-34	sRNA
OBD02	Leaf and stem	78,603,581	18-30	sRNA
BDI05	Panicle 1	74,944,182	18-34	sRNA
OBD01	Panicle 2	78,992,373	18-30	sRNA
BDI02	Shoot control for stress	72,494,849	18-34	sRNA
BDI21	Leaf	13,772,836	18-20	PARE
BDI25	Panicle	29,026,276	18-20	PARE

Part D. Lilium small RNA data				
Lilium_leaf	*Lilium maculatum* leaf, BM14-190	27,227,974	18-34	sRNA
Lilium_4mm_an	*Lilium maculatum* L-4, 4 mm anthers, BM14-191	35,477,763	18-34	sRNA
Lilium_5mm_an	*Lilium maculatum* L-5, 5 mm anthers, BM14-192	32,039,303	18-34	sRNA
Lilium_6mm_an	*Lilium maculatum* L-6, 6 mm anthers, BM14-193	32,024,530	18-34	sRNA
Lilium_8mm_an	*Lilium maculatum* L-8, 8 mm anthers, BM14-195	32,594,126	18-34	sRNA
Lilium_10mm_a	*Lilium maculatum* L-10, 10 mm anthers, BM14-197	33,533,618	18-34	sRNA
Lilium_leaf	*Lilium maculatum* leaf, BM14-190	68,250,144	2x150	RNA-seq
Lilium_4mm_an	*Lilium maculatum* L-4, 4 mm anthers, BM14-191	70,492,722	2x150	RNA-seq
Lilium_5mm_an	*Lilium maculatum* L-5, 5 mm anthers, BM14-192	67,997,742	2x150	RNA-seq
Lilium_6mm_an	*Lilium maculatum* L-6, 6 mm anthers, BM14-193	62,509,328	2x150	RNA-seq
Lilium_8mm_an	*Lilium maculatum* L-8, 8 mm anthers, BM14-195	66,675,362	2x150	RNA-seq
Lilium_10mm_a	*Lilium maculatum* L-10, 10 mm anthers, BM14-197	68,302,075	2x150	RNA-seq
Part E. Arabidopsis small RNA and PARE data				
AT_Leaf	Leaf wild type col0	21,716,675	18-34	sRNA
Col_2	Inflorescence of Arabidopsis Columbia (Col-0) plants, replication 2	11,396,067	18-34	sRNA
Col_3	Inflorescence of Arabidopsis Columbia (Col-0) plants, replication 3	8,690,024	18-34	sRNA
AtCM	sRNA library from mixed stage inflorescence of Arabidopsis wildtype Col-0	26,084,301	18-34	sRNA
Wt_d	sRNA library from unopened flower buds of Arabidopsis wiltype Col-0	11,808,921	18-30	sRNA
Col_d	sRNA library from unopened flower buds of Arabidopsis wildtype Col-0	24,688,934	18-34	sRNA
TWF	Col-0 inflorescence, control for xrn4; PARE data	17,711,729	20	PARE
Tx4F	xrn4 mutant inflorescence; PARE data	10,643,828	20	PARE
Part F. Maize small RNA data used for comaprision to human-curated data, from Zhai et al., 2015.				
0.2 Fertile_r1	maize fertile anthers, 0.2 mm, rep 1.	12,896,366	18-34	sRNA
0.4 Fertile_r1	maize fertile anthers, 0.4 mm, rep 1.	35,286,904	18-34	sRNA
0.7 Fertile_r1	maize fertile anthers, 0.7 mm, rep 1.	34,045,564	18-34	sRNA
1.0 Fertile_r1	maize fertile anthers, 1.0 mm, rep 1.	34,019,136	18-34	sRNA
1.5 Fertile_r1	maize fertile anthers, 1.5 mm, rep 1.	37,521,389	18-34	sRNA
2.0 Fertile_r1	maize fertile anthers, 2.0 mm, rep 1.	35,796,937	18-34	sRNA
2.5 Fertile_r1	maize fertile anthers, 2.5 mm, rep 1.	55,306,867	18-34	sRNA
3.0 Fertile_r1	maize fertile anthers, 3.0 mm, rep 1.	26,285,048	18-34	sRNA
4.0 Fertile_r1	maize fertile anthers, 4.0 mm, rep 1.	40,210,297	18-34	sRNA
5.0 Fertile_r1	maize fertile anthers, 5.0 mm, rep 1.	29,748,543	18-34	sRNA
PLN Fertile_r1	maize mature pollen, rep 1.	33,201,391	18-34	sRNA
0.2 Fertile_r2	maize fertile anthers, 0.2 mm rep 2.	21,549,901	18-34	sRNA
0.4 Fertile_r2	maize fertile anthers, 0.4 mm rep 2.	36,257,470	18-34	sRNA
0.7 Fertile_r2	maize fertile anthers, 0.7 mm rep 2.	41,837,825	18-34	sRNA
1.0 Fertile_r2	maize fertile anthers, 1.0 mm rep 2.	37,594,678	18-34	sRNA
1.5 Fertile_r2	maize fertile anthers, 1.5 mm rep 2.	27,713,840	18-34	sRNA
2.0 Fertile_r2	maize fertile anthers, 2.0 mm rep 2.	48,664,038	18-34	sRNA
2.5 Fertile_r2	maize fertile anthers, 2.5 mm rep 2.	50,543,592	18-34	sRNA
3.0 Fertile_r2	maize fertile anthers, 3.0 mm rep 2.	28,476,786	18-34	sRNA
4.0 Fertile_r2	maize fertile anthers, 4.0 mm rep 2.	27,693,417	18-34	sRNA
5.0 Fertile_r2	maize fertile anthers, 5.0 mm rep 2	27,562,227	18-34	sRNA
PLN Fertile_r2	maize mature pollen, rep 2	33,953,600	18-34	sRNA
0.4 Fertile, r3	maize fertile anthers, 0.4 mm, rep 3	25,908,576	18-34	sRNA
0.7 Fertile, r3	maize fertile anthers, 0.7 mm, rep 3	22,135,681	18-34	sRNA
1.0 Fertile, r3	maize fertile anthers, 1.0 mm, rep 3.	29,761,645	18-34	sRNA
1.5 Fertile, r3	maize fertile anthers, 1.5 mm, rep 3.	26,530,207	18-34	sRNA
2.0 Fertile_r3	maize fertile anthers, 2.0 mm, rep 3.	23,136,153	18-34	sRNA
2.5 Fertile_r3	maize fertile anthers, 2.5 mm, rep 3.	24,169,553	18-34	sRNA
3.0 Fertile_r3	maize fertile anthers, 3.0 mm, rep 3.	23,155,446	18-34	sRNA
4.0 Fertile_r3	maize fertile anthers, 4.0 mm, rep 3.	23,592,086	18-34	sRNA
5.0 Fertile_r3	maize fertile anthers, 5.0 mm, rep 3.	23,453,249	18-34	sRNA
PollenFertile_r3	maize mature pollen, rep 3.	20,949,777	18-34	sRNA

Code	Title	Reads	polished high-quality isoforms[d]	No. of cells
Lilium.2kb	Lilium, 4-6mm anther, 1-2kb insert length	410,768	54,622	5
Lilium.2-3kb	Lilium, 4-6mm anther, 2-3kb insert length	285,133	44,566	3
Lilium.4kb-abov	Lilium, 4-6mm anther, 3-3kb insert length	305,581	23,591	3

3.3 Results

We first sought to assess the sensitivity and specificity for *PHASIS*; ideally, this would be done with a gold-standard reference set of experimentally-validated *PHAS* loci in plants. While the definition of "gold standard" is as-yet unclear for *PHAS* loci, the recently-described maize loci are among the most exhaustively characterized (Zhai et al. 2015), and thus we used these data below. We also compared *PHASIS* predictions and performance with PhaseTank (Guo, Qu, and Jin 2015). Currently, two computational tools are capable of *de novo* discovery of *PHAS* loci – PhaseTank (Guo, Qu, and Jin 2015) and ShortStack (Axtell 2013b). *PhaseTank* is exclusively build for predicting *PHAS* loci in plants, while *ShortStack* aims to annotate and quantify diverse sRNA-associated genes (or clusters), and it's typically deployed for characterizing miRNAs in plants and animals (Axtell 2013b). A direct comparison between *PHASIS* and *ShortStack* is not possible due to significant differences in their scope, utility and workflow (**Table 3.1**). So, for comparative benchmarking, we chose *PhaseTank*, mainly because of matching objectives and its published superiority over *ShortStack* in predicting *PHAS* loci (Guo, Qu, and Jin 2015). Benchmarking was performed across five plant species – *Arabidopsis thaliana* (Arabidopsis), *Brachypodium distachyon* (Brachypodium), *Oryza sativa* (rice), *Zea mays* (maize) and *Lilium maculatum* (*Lilium*). These species were selected based on availability of high-quality nuclear genome assemblies or anther transcriptomes (in case of *Lilium*), and

deep sRNA libraries from premeiotic and meiotic anther or from at least one of these two stages that should contain many reproductive phasiRNAs (**Table 3.2**). Arabidopsis was included because it was originally used in *PhaseTank* benchmarking (Guo, Qu, and Jin 2015). For *PhaseTank,* the reference genome, transcriptome and sRNA libraries were converted to the appropriate formats, and the time for file conversion process, although complex and lengthy, was not added in the *PhaseTank* runtimes. *PHASIS* and *PhaseTank* use inherently different scoring schemas; because of this difference, we used a conservative *p-value* (1e-05) for *PHASIS* and the recommended score (=15) for *PhaseTank.* All benchmarks were performed on a 28 core, 2.42 GHz machine with 512 GB of RAM, running CentOS 6.6.

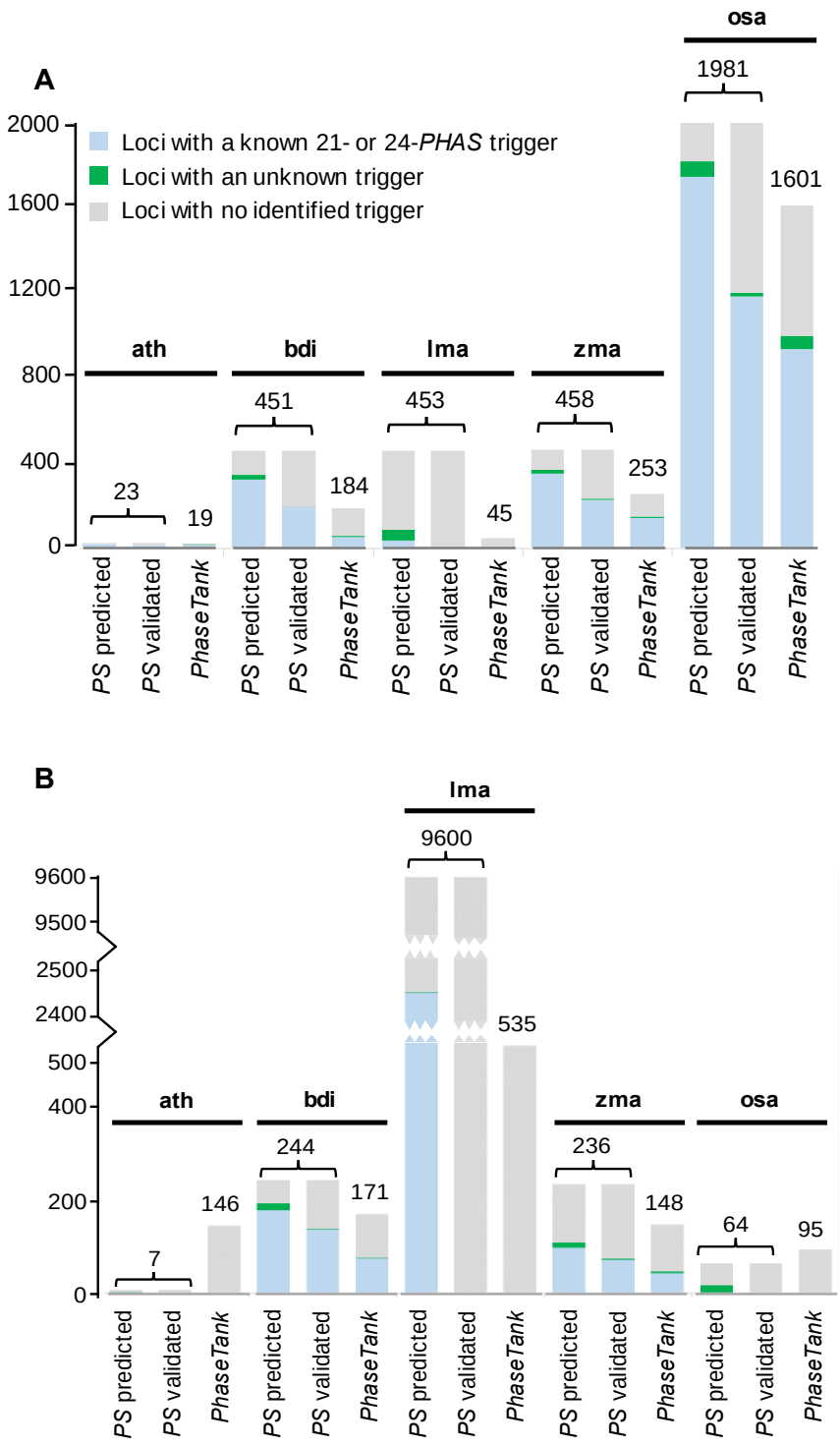

A
Loci with a known 21- or 24-PHAS trigger
Loci with an unknown trigger
Loci with no identified trigger
2000
1600
1200
800
400
0
ath
bdi
lma
zma
osa
23
19
451
184
453
45
458
253
1981
1601
PS predicted
PS validated
PhaseTank
B
9600
9500
2500
2400
500
400
200
0
ath
bdi
lma
zma
osa
7
146
244
171
9600
535
236
148
64
95

Figure 3.2: **Number of *PHAS* loci or transcripts and their trigger predicted by *PHASIS***. *PHASIS* is labelled as 'PS' and it is compared to PhaseTank for benchmarking. A) 21-*PHAS* and B) 24-*PHAS* loci identified by both tools along with their triggers in Arabidopsis (ath), Brachypodium (bdi), *Lilium* (lma), rice (osa) and maize (zma). For *PHASIS* trigger prediction, results from both "validation" and "prediction" mode was included. The bars for *Lilium* 24-*PHAS* loci are split at two different points for display purposes. Triggers assigned to *PHAS* loci that do not match with known or published miRNA triggers were represented as 'unknown' triggers. *PHAS* prediction and runtime performance

3.3.1 *PHAS* prediction and runtime performance

We first compared *PHAS* loci and transcript predictions from *PHASIS* and *PhaseTank*. Since Arabidopsis lacks 24-*PHAS* loci (none have ever been published, nor have we found any), and there are few *TAS* genes (n=8), these were excluded from quantification of prediction and speed comparisons. *PHASIS* demonstrated an edge over *PhaseTank* in *PHAS* predictions: in genomic analyses, it predicted up to 2.5 times more *PHAS* loci, ranging from 73 24-*PHAS* (145% gain) to 380 21-*PHAS* (24% gain) loci in Brachypodium and rice respectively (**Table 3.3**). The biggest gain was observed in an analysis of the *Lilium* transcriptome, in which *PHASIS* predicted ~10 times (n=408) more 21-*PHAS* and 18 times (n=9065) more 24-*PHAS* precursor transcripts compared to *PhaseTank* (**Figure 3.2B**). The specific data format requirements of *PhaseTank* made it difficult to accurately determine the set of common *PHAS* predictions (the 'common *PHAS* pool', hereafter) for transcriptome level analysis, however, by matching the sequences we determined that *PHASIS* captured at least 66% of 21-*PHAS* and 99% of 24-*PHAS* predictions from *PhaseTank*. For genomic analyses, *PHASIS* captured >80% of *PhaseTank* predictions, except in rice and Arabidopsis in which *PhaseTank* predicted additional 24-*PHAS* loci (**Table 3.3**).

Table 3.3: **Comparison of predictions for *PHAS* loci (and precursor transcripts) and their miRNAs triggers between *PHASIS* and its direct competitor *PhaseTank*.** In all comparisons *PHASIS* displays clear superiority of *PhaseTank* except in Arabidopsis 24-*PHAS* where PhaseTank predictions were false-positives and in rice 24-*PHAS* where it predicted loci with weak phased patterns. The weak phased loci from rice were identified by *PHASIS* by running it at a lower *p-value* cutoff. *PhaseTank* also show a little gain in predicting triggers for Arabidopsis 21-*PHAS*, these cases are described in detail in paper and could be identified by relaxing *phasmerge* search-space parameters.

Species	Type	PHAS locus gain with PHASIS over PhaseTank	Proportion of *PhaseTank* *PHAS loci* *captured by* *PHASIS*	Gain in miRNA triggers: PHASIS (PARE supported) vs. PhaseTank (PARE supported	Gain in miRNA triggers, PHASIS (predicted) vs. PhaseTank (PARE supported)	Gain in miRNA triggers, PHASIS (predicted) vs. PHASIS (PARE supported)
Arabidopsis	21- *PHAS*	21.00%	84.00%	*-54.00%*	-18.18%	45.45%
	24- *PHAS*	*-95.00%*	*1.00%*	No predictions	No predictions	No predictions
Brachypodium	21- *PHAS*	145.00%	79.00%	76.09%	178.26%	81.48%
	24- *PHAS*	49.00%	85.00%	35.90%	69.23%	40.28%
Rice	21- *PHAS*	24.00%	96.63%	4.94%	53.55%	51.25%
	24- *PHAS*	*-33.00%*	*29.00%*	No predictions	N.D	N.D
Maize	21- *PHAS*	81.00%	96.84%	4.11%	54.80%	56.89%
	24- *PHAS*	59.00%	85.81%	9.09%	63.64%	46.67%
Lilium	21- *PHAS*	907.00%	67%*	No PARE data	No PARE data	No PARE data
	24- *PHAS*	1694.00%	94%*	No PARE data	No PARE data	No PARE data

The additional 24-*PHAS* loci predicted by *PhaseTank* in rice and Arabidopsis all had significantly lower quality scores (from *PhaseTank*) compared to the common *PHAS* pool, as did the *PhaseTank*-exclusive 21- and 24-*PHAS* predictions from other species. The average quality scores computed for each species were 1.7 to 7.8 times lower compared to the common *PHAS* pool (*p-value* < 0.001, t-test); the predictions exclusive to *PhaseTank* are likely unphased and a misinterpretation of loci yielding profuse heterochromatic siRNAs (hc-siRNAs). This may explain the 24-*PHAS* predictions in Arabidopsis by *PhaseTank* (**Figure 3.2B and Table 3.3**), as 24-nt phasiRNAs have not been reported in Arabidopsis despite exhaustive analyses (Axtell 2013a). Nonetheless, considering that these *PhaseTank* predictions could represent

weak *PHAS* loci, we attempted to capture them by running *PHASIS* at lower *p-value* cutoff (1e-03) but failed to detect >96% of them. Manual investigation of a portion of these *PHAS* loci using our custom sRNA browser which uses a slightly different *PHAS* scoring schema (Allen and Howell 2010), revealed that these are indeed either unphased or show typical characteristics of hc-siRNA loci, i.e. are false positives predicted by *PhaseTank* (**Figure 3.3A**). However, we could detect 70% (n=67) of the total 24-*PHAS PhaseTank* predictions in rice at the lower *p-value* cutoff (1e-03) of *PHASIS*, and a majority of these showed weak phasing patterns (**Figure 3.3B**), suggesting that *PHASIS* missed these at the selected cutoff. However, the count of 24-*PHAS* loci predicted in rice by both tools in these libraries from a recent study (Fei et al. 2016), was lower than earlier estimates (Johnson et al. 2009), indicating that the libraries likely missed meiotic peak of accumulation. These contrasting observations – Arabidopsis, in which *PHASIS* correctly excluded 24-*PHAS* predictions even at relaxed cutoff, versus rice, in which it correctly captured 70% of weakly phased 24-*PHAS* loci – highlights differences in scoring in the two tools, with the default *PHASIS p-value* cutoff (1e-05) more stringent than that of *PhaseTank* (score=15). Using a lower *p-value* cutoff for *PHASIS* could further increase the gain in *PHAS* predictions over *PhaseTank* without adding much noise.

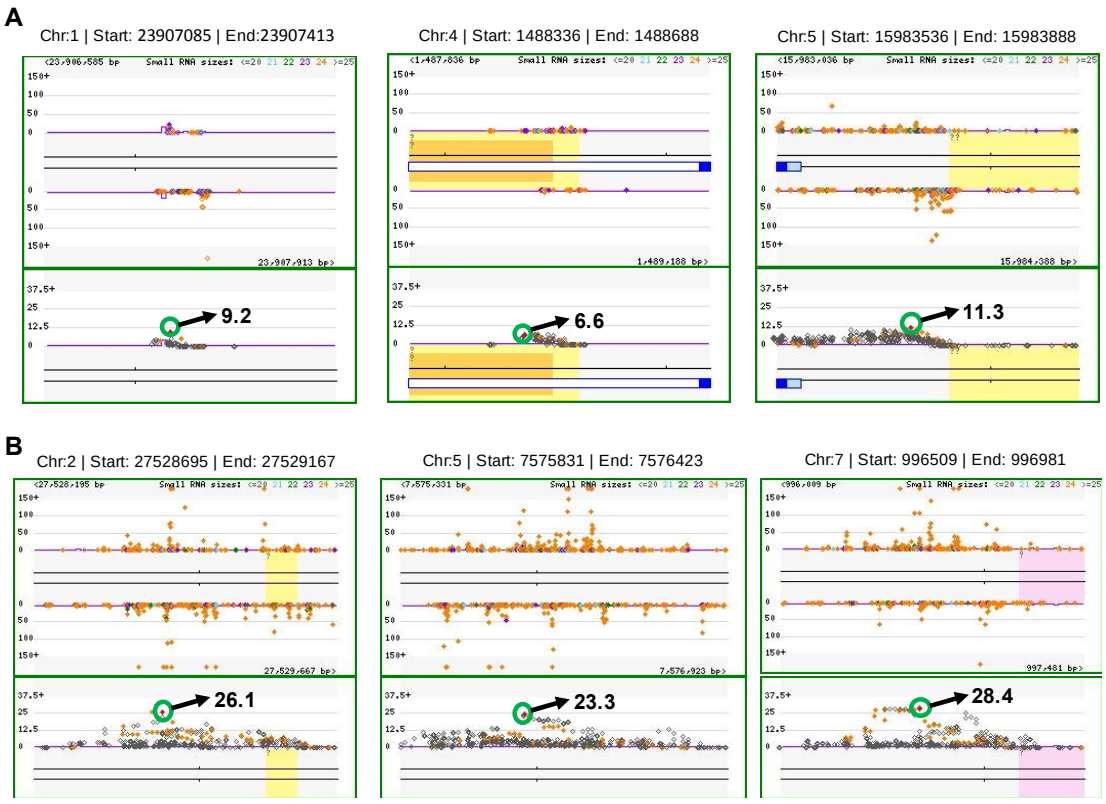


Figure 3.3: **Snapshots of genomic loci with evidence of phasing.** A) Examples of 24-*PHAS* loci predicted by *PhaseTank* in Arabidopsis. These are either un-phased or display characteristics typical of heterochromatic siRNA-associated regions. B) Rice 24-*PHAS* loci predicted by PhaseTank and rescued in *PHASIS* by using a lower *p-value* cutoff display. Most of these had weak phasing scores but display characteristics typical of phased loci described in maize (Zhai et al., 2015). Phased scores for all the loci were computed as described by Allen et al., 2007.

We manually investigated 21- and 24-*PHAS* predictions that are exclusive to *PHASIS*, using the Meyers lab sRNA viewer. The majority of these displayed characteristics matching those of the canonical 21- and 24-*PHAS* loci reported in maize (Zhai et al. 2015) (**Figure 3.4**). Moreover, a major proportion of these *PHASIS*-exclusive predictions had PARE-validated miRNA triggers, matching to the earlier

reports from maize, rice and Brachypodium (Johnson et al. 2009; Zhai et al. 2015; D. H. Jeong et al. 2013).

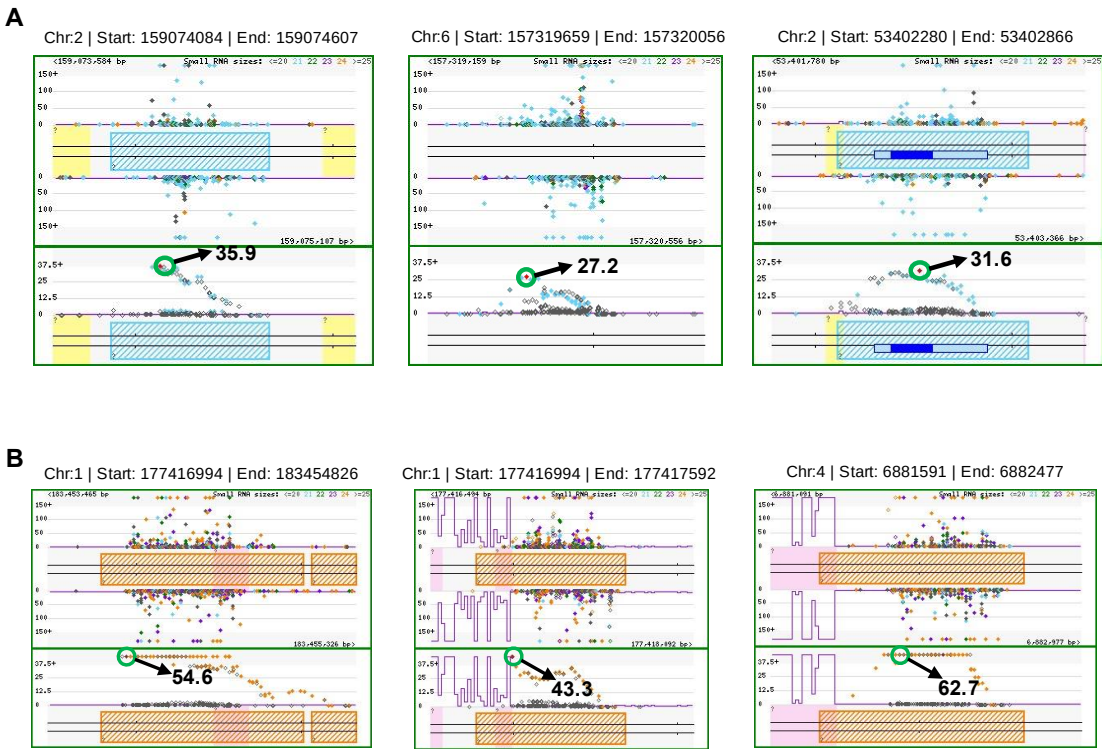


Figure 3.4: **Snapshots of genomic loci from maize with evidence of phasing.** Examples of A) 21-*PHAS* and B) 24-*PHAS* loci identified in maize by *PHASIS* and missed by the PhaseTank. Our small RNA genome browser displays robust phasing scores at these loci suggesting that these are indeed true phased loci. In 24-*PHAS* snapshots 24-nt sRNAs (orange diamonds) are shadowed by 23-nt sRNAs (violet diamonds) if these have close 5' ends. Blue or orange cross-hatched boxes in were annotated as 21- or 24-*PHAS* loci by Zhai et al. (2015).

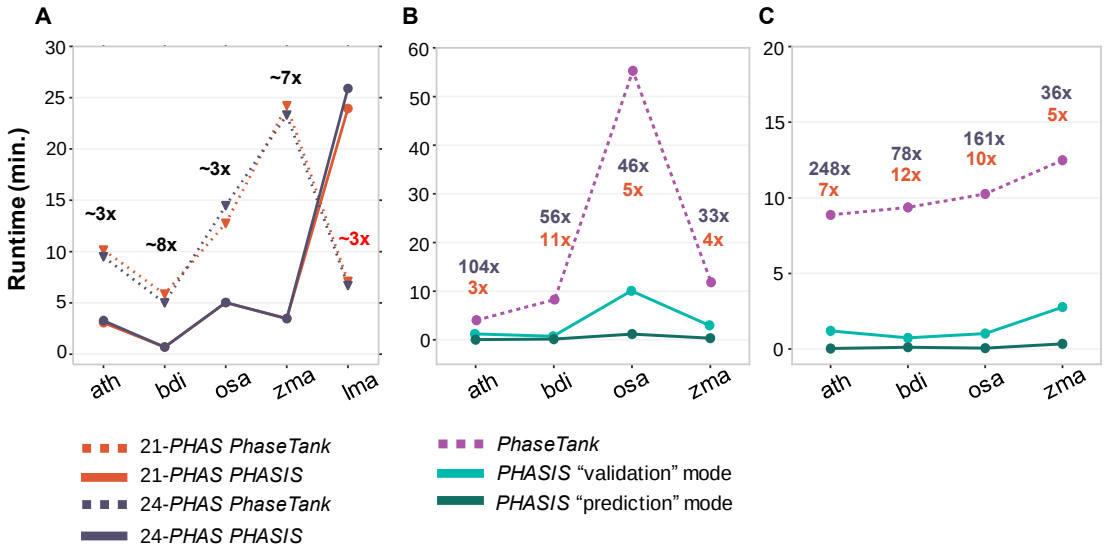


Figure 3.5: **Runtime comparisons between *PHASIS* and *PhaseTank*. A)** Time taken by both tools in prediction of 21- and 24-*PHAS* loci or precursors transcripts. Speed gain displayed by *PHASIS* over *PhaseTank*, approximated for both size classes, is individually marked for each species. **B)** and **C)** Time taken by both tools in predicting 21- and 24-*PHAS* triggers, respectively. Speed gain displayed by *PHASIS* in "validation" and "prediction" mode over *PhaseTank* is displayed in blue and orange colors respectively. In all comparisons, *Arabidopsis* is marked as "ath", *Brachypodium* as "bdi", rice as "osa", maize as "zma" and *Lilium* as "lma".

Next, we compared prediction runtimes of *PHASIS* and *PhaseTank* from genome- and transcriptome-level experiments. To get the correct runtimes for both tools, we excluded the execution time for a common step performed by an external tool (Bowtie, version 1) that prepares the index for the reference genome or transcriptome. For genome-level experiments, *PHASIS* displayed a minimum speed gain of 3x in Arabidopsis and rice and a maximum speed gain of 7x in maize (**Figure 3.5**). In transcriptome-level experiments, both tools took almost equal time (**Figure 3.5**). However, *PHASIS* yielded 10x (n=408) to 17x (n=9065) more *PHAS* predictions

for 21- and 24-*PHAS* loci, respectively (**Table 3.3, Figure 3.6**), compared to *PhaseTank*, which means that *PHASIS* processed a high number of *PHAS* transcripts in the same runtime. Moreover, the time and effort required to convert the reference genome as well as the sRNA libraries to meet *PhaseTank* input requirements were not included in these runtime comparisons. Lastly, it should be noted that *PHASIS* takes significantly less time for any subsequent analyses in these species because of its unique ability to systematically store ancillary data in the first run, check data integrity and compatibility with parameters for subsequent runs, and avoid redoing the slowest steps, such as reference pre-processing, index preparation, etc.

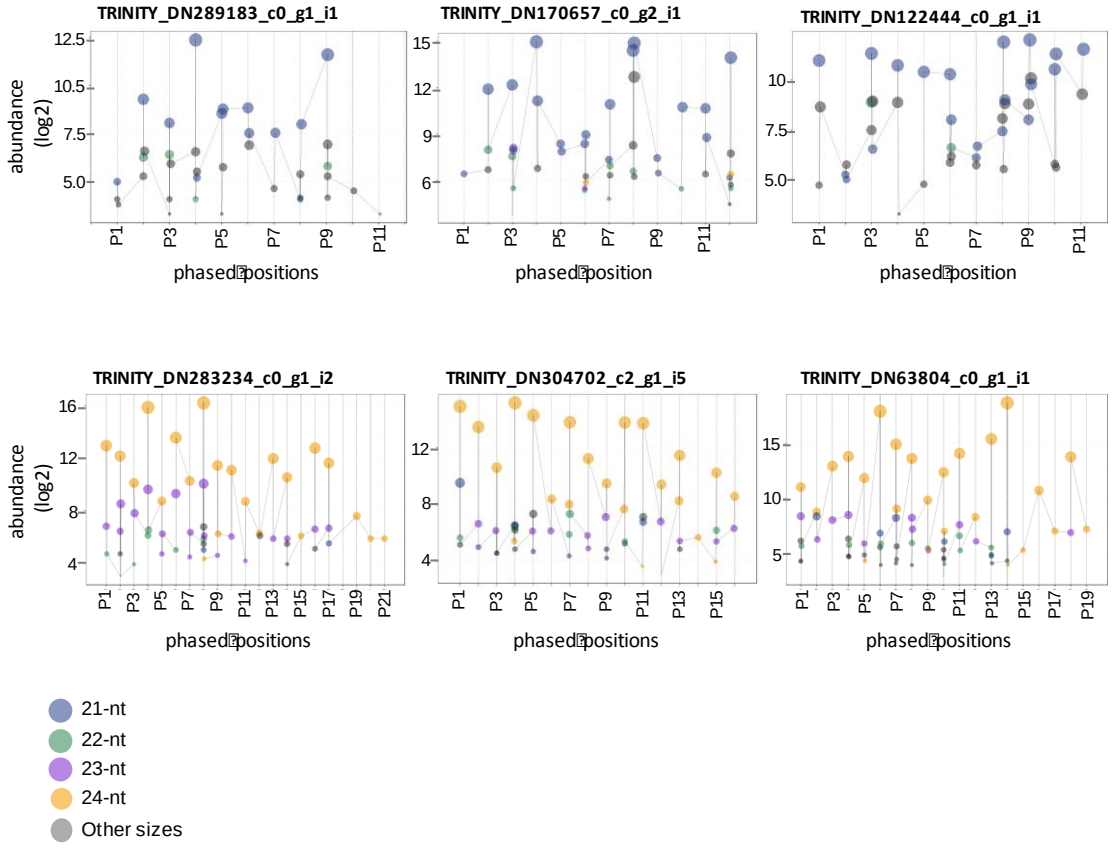


Figure 3.6: **sRNA abundance plot for *Lilium PHAS* precursor transcripts.** Examples of A) 21-*PHAS* and B) 24-*PHAS* precursor transcripts in *Lilium* that were missed by PhaseTank but identified by *PHASIS*. Both the position and abundance of sRNAs generated from these precursors display characteristics typical of reproductive phased loci described in rice (Johnson et al., 2009). Gridlines on the x-axis represent a 21- or 24-nt phased position starting from the 5' end of first phased cycle. The x-axis represents abundances for sRNAs in log2 scale.

3.3.2 Comparison of *PHASIS* predictions with manually-curated data

We next wanted to address how well the predictions from *PHASIS* compare with a set of manually-curated *PHAS* loci. We and collaborators curated a set of 21- and 24-*PHAS* (n= 463 and 163) loci from precisely-staged, premeiotic and meiotic maize anthers (Zhai et al. 2015). This curated set was prepared by first combining all

libraries from the sampled premeiotic and meiotic stages into a single file, followed by genome wide scans to identify phasiRNA generating loci using a score-based approach (Allen et al. 2005) and finally curating each *PHAS* locus to exclude those that overlap with repeat-associated regions or display sRNA distribution atypical of hc-siRNA generating loci (Zhai et al. 2015). *PHASIS* processes each library separately mainly to a) detect phased patterns independently in at least one of the input sRNA libraries, b) minimize any noise that could be added by combining sRNAs from multiple stages, tissues or treatments, and c) infer the correct 5'-end of *PHAS* loci by collating data from different libraries. Therefore, unlike the original analysis, we did not combine the 32 libraries (see **Table 3.2**) for predictions by *PHASIS*. Furthermore, to emulate 'real world' conditions in which *PHASIS* would be used by non-experts, we did not provide a confidence cutoff - i.e. *PHASIS* was run in the default mode. Of the manually-curated 463 21-*PHAS* and 178 24-*PHAS* loci, *PHASIS* capturing 89.0% (n=411) and 85.79% (n=151) (**Table 3.4**). The majority of those missed either lacked continuous phased positions or had a very low abundance across all sRNA libraries, and some had a single sRNA read accounting for the major proportion (>90%) of the abundance at the *PHAS* locus. The average abundance of siRNAs in the 'missed' 21- and 24-*PHAS* set was ~12- and 252-times lower compared to the common pool ($p < 1.02e-09$), supporting the observation that those missed by *PHASIS* were weakly phased loci; a portion of these could be captured with a relaxed cutoff. Nonetheless, these results demonstrate that *PHASIS* predictions are largely consistent with the manually-curated data, and for most studies, the use of *PHASIS* may ameliorate the need to manually curate *PHAS* locus predictions, an otherwise complex and cumbersome task especially when *PHAS* loci number in the hundreds to thousands, as

reported in many plant genomes (Johnson et al. 2009; Zhai et al. 2011; D. H. Jeong et al. 2013; Arikit et al. 2014; R. Xia, Xu, et al. 2015; Zhai et al. 2015).

Table 3.4: **Comparison of *PHASIS* predictions with the published and manually-curated data from maize anthers.** *abundances in the last two columns are the trimmed mean of abundances.

	PHASIS predicted	Number of curated *PHAS* loci (Zhai et al., 2015)	Number of curated *PHAS* captured	Proportion of curated *PHAS* captured	Common set of loci (PNAS & *PHASIS*)*	Missed by *PHASIS**
21-PHAS	488	463	411	88.98%	30670	2595
24-PHAS	178	176	151	85.79%	134478	534

3.3.3 Trigger prediction and runtime performance

The identification of the miRNA triggers of *PHAS* loci is important for understanding their potential roles, classification and for discovery of secondary siRNA cascades. In addition, a set of *PHAS* loci or transcripts when combined with the trigger identity, may serve as a gold-standard reference set for downstream experimental and bioinformatic studies. Given the importance of triggers identification, we compared the trigger prediction performance of *PHASIS* in 'validation' mode with *PhaseTank*. The *PHASIS* 'validation' mode will identify triggers for *PHAS* loci or transcripts using experimental data such as PARE, degradome or GMUCT libraries. *PhaseTank* by default predicts triggers in 'validation' mode, i.e. experimental data is required. Since, *PHASIS* predicted more *PHAS* loci compared to *PhaseTank*, the number of *PHAS* loci (and transcripts) with the predicted triggers by *PHASIS* were higher too. So, for a fair comparison, we used only the common pool of *PHAS* loci to evaluate the trigger prediction performances.

Table 3.5: **Accuracy of triggers predicted by *PHASIS* in prediction mode. *PHASIS* 'prediction' mode as analysis to predict triggers for *PHAS* loci or transcripts without any supporting experimental data such as PARE, degradome or GMUCT libraries.** Accuracy was computed as the proportion of triggers out of total that match to known triggers of phasiRNAs and tasiRNAs, as described in earlier studies. N.D - No triggers were identified for rice 24-*PHAS* loci, as samples for sRNA libraries didn't correspond to the precise meiotic stage at which 24-nt phasiRNAs accumulate. [#]A major proportion of 21-*PHAS* loci unexpectedly had miR2275 triggers, the known trigger typically of reproductive 24-*PHAS* loci.

Species	Type	Trigger prediction accuracy (PARE support)	Trigger prediction accuracy (prediction mode)
Arabidopsis	21-*PHAS*	0.99	0.99
	24-*PHAS*	No Prediction	No Prediction
Brachypodium	21-*PHAS*	0.995	0.924
	24-*PHAS*	0.993	0.923
Rice	21-*PHAS*	0.986	0.961
	24-*PHAS*	N.D	N.D
Maize	21-*PHAS*	0.966	0.951
	24-*PHAS*	0.96	0.891
Lilium	21-*PHAS*	No PARE data	0.439[#]
	24-*PHAS*	No PARE data	0.999

PHASIS displayed a gain of up to 76.0% in predicted triggers, except for 21-*PHAS* loci in Arabidopsis (**Figure 3.2A and B**), with a minimum accuracy of 96.0% for 24-*PHAS* maize loci and maximum accuracy of 99.5% in Brachypodium 21-*PHAS* loci (**Table 3.5**). This accuracy was computed as the proportion of triggers (out of the total) that match to known triggers of phasiRNAs and tasiRNAs described in earlier studies (Allen et al. 2005; Fei, Xia, and Meyers 2013; Axtell et al. 2006; Allen and Howell 2010; Zhai et al. 2011; Y. Zheng et al. 2012; Johnson et al. 2009). These

estimates of accuracy are likely conservative, given that there might be a few new and unknown triggers that we counted as false positives in our accuracy computations. We excluded rice 24-*PHAS* loci from our comparisons because both the tools failed to report triggers for these loci, likely due to sRNA libraries that were not precisely staged relative to the accumulation of 24-nt phasiRNAs and thereby making it difficult to capture the 5' and 3' ends of *PHAS* loci – information crucial to the identification of correct triggers. *Lilium* 21- and 24-*PHAS* transcripts were also excluded from the comparisons because of a lack of PARE data from the corresponding anther stages, data required by *PhaseTank* to predict triggers. Likewise, Arabidopsis 24-*PHAS* couldn't be included in our comparison as *PhaseTank* predicted loci (n=146) were false positives, and there were no overlapping loci with *PHASIS*.

We noticed a decline in number of predicted triggers by *PHASIS* for 21-*PHAS* loci in Arabidopsis, compared to those predicted by *PhaseTank* (**Figure 3.2A**). This decline in predicted triggers was traced to seven phased loci corresponding to the pentatricopeptide repeat (PPR) gene family, triggered by miR161. We found *PhaseTank* predicted trigger sites for five of these loci that were located towards the middle of *PHAS* loci, 214 nt to 310 nt from the first or last phased cycle. Since, *phasmerge* (the trigger discovery tool of *PHASIS*) is built with the aim to eliminate the need for experimental data and because trigger sites are expected to overlap with 5' or 3' ends of the phased region, it uses a narrow search space at the 5' and 3' ends to search for triggers. Hence, these particular trigger sites for miR161 were missed by *PHASIS*. In *phasmerge*, the search space to identify triggers is defined by the number of phased positions (*PHAS*-index) on either side of 5' and 3' ends of phased regions, and by default it is set to ±3 *PHAS*-index for both ends. The *PHAS*-index setting to

expand or the narrow search space for triggers is user tunable and can be adjusted to capture such cases. Nonetheless, these 21-*PHAS* loci from Arabidopsis support our estimates that trigger identification by *phasmerge* is conservative, and relaxing the *phasmerge* search parameters could further increase the gain in predicted triggers compared to *PhaseTank.*

3.3.4 Identifying *PHAS* triggers without additional experimental data

We next evaluated the trigger prediction performance of *PHASIS* in 'prediction' mode by comparing it with *PhaseTank* and *PHASIS* in the 'validation 'mode. We define *PHASIS* 'prediction' mode as an analysis to predict triggers for *PHAS* loci or transcripts without any supporting experimental data such as PARE, degradome or GMUCT libraries. *Lilium* was excluded from the comparison of predicted triggers due to the lack of PARE data, which is compulsory for *PhaseTank* to predict triggers and required by *PHASIS* in 'validation' mode. Also, for reasons mentioned above, 24-*PHAS* loci from Arabidopsis and rice were excluded from the comparisons. *PHASIS* displayed a minimum gain of 40.3% and maximum gain of 178.3% over *PhaseTank* in predicting triggers for 21-*PHAS* and 24-*PHAS* loci from Brachypodium, respectively (**Figure 3.2 and Table 3.3**). The gain in number of triggers ranged from a minimum of 35 for maize 24-*PHAS* loci to a maximum of 611 for rice 21-*PHAS* loci. In addition to the gain in trigger prediction, *PHASIS* also displayed significant accuracy in prediction mode, with a minimum accuracy of 89.9% in predicting triggers for 24-*PHAS* loci from maize and maximum accuracy of 99.9% in the case of *Lilium* 24-*PHAS* precursor transcripts, however, with an exception to *Lilium* 21-*PHAS* triggers. The accuracy of predicted triggers of *Lilium* 21-*PHAS* loci was significantly lower (43.9%) compared to the other species (**Table 3.5**). For

Lilium, we used miRNAs from well-characterized monocots like rice and maize because a complete set of miRNAs were not available due to the absence of a sequenced genome. Surprisingly, we found that for *Lilium* 21-*PHAS* transcripts a majority of triggers corresponded to miR2275 instead of miR2118; this observation was puzzling because miR2275 is known to trigger 24-nt phasiRNAs in the grasses, and it was the basis for the low recorded accuracy in predicting *Lilium* 21-*PHAS* triggers; we did not further investigate the miR2275-triggered 21-*PHAS* transcripts. We also noticed that the proportion of 21- and 24-*PHAS* precursors for which triggers could be identified in *Lilium*, 18.1% and 25.9% respectively (**Figure 3.2 and Table 3.3**), was substantially lower compared to the overall average of 73.8% in other species for which genomic analysis was performed. Plant *PHAS* precursor transcripts are typically cleaved by the miRNA trigger, converted to dsRNA by an RNA-dependent RNA polymerase, and then successively diced by a Dicer enzyme. Since no data on transcriptional rate, stability and half-life of phasiRNA precursors are available, we speculated that a portion of the *Lilium PHAS* precursor transcripts were shortened by processing from the 5' end, removing the trigger target sites. Identifying triggers from such "processed" precursor transcripts is not possible because the P1 site corresponding to the first phasiRNA (at the 5' terminus) could be missing from the transcript. In addition, the presence of already-processed mRNAs will confound the *de novo* assembly of precursor transcripts from short-reads.

To test whether the low yield of triggers by *phasmerge* resulted from our use of processed precursor transcripts and not a technical shortcoming of *PHASIS*, we generated Single Molecule Real Time (SMRT) PacBio sequencing data from *Lilium* anthers 4 mm to 6 mm in length. These sizes represented premeiotic and meiotic

stages of anther development (**see methods**) and were selected based on the availability of the samples. Capturing *PHAS* precursors is complex, not just because these are targets of miRNAs presumably rapidly processed by a Dicer, but reproductive phasiRNAs are ephemeral in development and thus not easily captured (Zhai et al. 2015). SMRT-seq produced 425,897 full-length transcripts for 176,373 unique isoforms, which were pre-processed to generate 122,779 high quality (polished) transcripts. This set had 5,131 unique proteins covered by more than 80% protein length, relative to the Uniprot protein-sequence resource, thereby suggesting a reasonable assembly of the anther transcriptome. *PHASIS* identified 87 21-*PHAS* and 175 24-*PHAS* transcripts respectively. This low yield of *PHAS* transcripts was expected, though not to such a degree, because of the combination of the following: a) low read counts for SMRT-seq compared to the deep RNA-seq data, b) the coverage-based error correction algorithm - 'Quiver' implemented in the IsoSeq protocol (SMRT Analysis software version 2.3, Pacific Biosciences) which filters out transcripts with insufficient coverage, i.e. those that cannot be confidently corrected, and c) the aforementioned processive cleavage of *PHAS* precursors by Dicer *phasmerge* could identify triggers for only 21.8% (n=19) of 21-*PHAS* precursors, a slight increase compared to 18.1% in the RNA-seq assembly, and these triggers included miR2275, miR2118 and miR390. This low proportion of triggers detected for 21-*PHAS* could result from missing the precise stage at which 21-*PHAS* precursors accumulate in the *Lilium* samples. However, *phasmerge* could identify triggers for 54.2% of the 24-*PHAS* precursors, a significant increase over the 25.9% in the RNA-seq assembly, supporting our premise about the completeness of the *PHAS* precursor transcripts. The processed precursors were likely collapsed into the full-length or the

longest transcript in SMRT-seq assembly, thereby enriching the proportion of uncleaved precursor transcripts. Hence, it should be noted that neither the precursors from neither RNA-seq nor SMRT-seq may accurately represent the true total count of *PHAS* loci in *Lilium*.

Lastly, we compared runtimes for both tools for miRNA trigger prediction of *PHAS* loci and transcripts. *PHASIS* showed a minimum speed gain of 3.3x and a maximum speed gain of 12.6x over *PhaseTank* in 'validation' mode (**Figure 3.5**). In 'prediction' mode, *PHASIS* was at least 5.0x and at most 31.2x faster compared to its own 'validation' mode without any significant loss in accuracy (**Table 3.5**). *PhaseTank* requires PARE data to predict triggers, and lacks a functionality equivalent to *PHASIS* 'prediction' mode, but since *PHASIS*, even without the additional experimental data (like PARE) displays >89.9% accuracy in trigger prediction, we decided to compare runtimes for both. *PHASIS* in 'prediction' mode displayed a minimum speed gain of 33.3x and a maximum gain of 104.3x for Arabidopsis 21-*PHAS* loci (**Figure 3.5**). The trigger predictions for 24-*PHAS* loci from Arabidopsis and rice, which displayed even higher speed gains, were excluded from the runtime comparisons due to the reasons described above. This gain in *PHAS* trigger identification demonstrates the capacity of *PHASIS* to predict triggers without experimental data. This functionality will save time and the cost of preparing PARE libraries; it will also reduce the amount of sample required for phasiRNA analysis. Protocols for preparing PARE libraries requires comparatively more input RNA relative to RNA-seq or sRNA-seq (Zhai et al. 2014).

3.4 Availability

The methods and algorithm described in this article, implemented as *PHASIS* suite of tools for *PHAS* discovery, are freely available from https://github.com/atulkakrana/*PHAS*IS. *PHASIS* is released under the OSI Artistic License 2.0. Tools and Perl libraries required to use *PHASIS* along with the instructions to install and usage of individual tools is provided in detail in the *PHASIS* wiki (https://github.com/atulkakrana/*PHAS*IS/wiki/)

3.5 Chapter summary

In this chapter, I

- present a new suite for discovery and in-depth characterization of phasiRNAs, this suite "*PHASIS*" includes three independent tools – *phasdetect*, *phasmerge* and *phasmerge*
- developed *phasdetect*, which performs *de novo* prediction of *PHAS* loci or precursor transcripts using user-supplied sRNA libraries along with a reference genome or transcriptome, by efficiently processing tens to hundreds of sRNA libraries in parallel, reducing runtimes
- developed *phasmerge*, which generates a summarization and performs a comparison between the *PHAS* summaries and annotations using the library-specific *PHAS* lists and ancillary data generated by *phasdetect*
- developed *phasdetect*, which identifies sRNA triggers for *PHAS* loci and precursor transcripts using the *phasmerge* summaries and a user-provided list of miRNAs
- performed comparative benchmarking of results and runtimes between *PHASIS* and its direct competitor, on five different plant species
- compared the predictions from *PHASIS* with human-curated set of *PHAS* loci from maize
- tested the hypothesis that in transcriptome assembly, generated to compensate for absence of genome, a significant portion of the *PHAS*

precursor transcripts were shortened by processing from the 5' end, removing the trigger target sites

I observed that

- *PHASIS* predicts up to 2.5 times more *PHAS* loci compared to its competitor in genome-level experiments - ranging from 73 24-*PHAS* (145% gain) to 380 21-*PHAS* (24% gain) loci in Brachypodium and rice respectively

- *PHASIS* predicts ~10 times (n=408) more 21-*PHAS* and 18 times (n=9065) more 24-*PHAS* precursor transcripts, compared to its competitor in transcriptome-level experiments

- *PHASIS* captures 66% of 21-*PHAS* and 99% of 24-*PHAS* predictions from *PhaseTank* in transcriptome-level analysis*; and* >80% of PhaseTank predictions in genome-level analysis

- *PHASIS* is 2.5x to 7x faster in prediction *PHAS* loci compared to its competitor, and these speed gains reflect time consumed by *PHAS*IS to process each library individually, unlike its competitor which requires a non-redundant set of reads from all libraries in a single file

- *PHASIS* displays a minimum gain of 40.3% and maximum gain of 178.3% over its competitor in predicting triggers for 21-*PHAS* and 24-*PHAS* loci from Brachypodium, respectively, with numbers from minimum of 35 for maize 24-*PHAS* loci to a max. of 611 for rice 21-*PHAS* loci.

- *PHASIS* is 3x to 12x faster in predicting miRNA triggers in 'validation' mode, in which both tools were supplied with additional PARE libraries

- *PHASIS* predicts triggers, with a minimum accuracy of 96.0% for 24-*PHAS* maize loci and maximum accuracy of 99.5% in Brachypodium 21-*PHAS* loci, when provided with additional experimental data PARE or degradome libraries

- *PHASIS* is 33x to 104x faster in 'prediction' mode i.e. *de novo* prediction of triggers without any experimental data, such functionality is exclusive to *PHASIS*

- *PHASIS* display >89% accuracy in identifying triggers even without the experimental data

- *PHASIS* captured >86% of human curated *PHAS* predictions, and those missed have 11.82 and 252.30 times lower average abundance compared to those captures 21- and 24-*PHAS* loci respectively ($p < 1.02e-09$)

From this, I conclude that

- *PHASIS* is the "first" and "only" suite that enables large-scale survey of tens to hundreds of sRNA libraries for discovery, annotation, quantification and to identify their miRNA triggers

- *PHASIS* simplifies the study of phasiRNAs by waiving off the need for additional experimental data for discovery of miRNA triggers, through its trigger 'prediction' mode

- The novel trigger 'prediction' mode will save time and the cost of preparing PARE libraries; it will also reduce the amount of sample required for phasiRNA analysis

- *PHASIS* also waives the crucial requirement of assembled genome for discovery of *PHAS* transcripts and phasiRNAs

- *PHASIS* may ameliorate the need to manually curate *PHAS* locus predictions, an otherwise complex and cumbersome task especially when *PHAS* loci number in the hundreds to thousands

- *PHASIS* enables characterization of phasiRNAs in evolutionarily diverse plant genomes, which will advance our understanding of phasiRNA function and the adaptation of the pathway, and it may yet discover new classes of *PHAS* genes

- SMRT-Seq for full length transcripts is not an appropriate approach to capture *PHAS* precursors

Chapter 4

DISCOVERING GERMLINE-ASSOCIATED PHASED siRNA PATHWAYS ACROSS MONOCOT EVOLUTION

In grasses, two pathways generate diverse and numerous 21-nt (pre-meiotic) and 24-nt (meiotic) phased siRNAs highly enriched in anthers. These "phasiRNAs" are analogous to mammalian piRNAs, yet their functions remain largely unknown, as are their evolutionary origins. The 24-nt meiotic phasiRNAs are as-yet only described in grasses (Poaceae), in which their biogenesis is dependent on a specialized Dicer (DCL5). To assess the evolutionary path that gave rise to this pathway, we examined reproductive phasiRNA pathways in garden asparagus (*Asparagus officinalis)*, a non-grass monocot that speciated ~63 mya from MRCA of grasses, and in lily (*Lilium maculatum)* and daylily (*Hemerocallis lilioasphodelus)*, that diverged approximately 117 mya from *Asparagus*. We demonstrate that both pre-meiotic and meiotic phasiRNAs are prevalent across the monocots included in this study, establishing their origins well before grasses. In addition to male germline, we find evidence for their accumulation in female and somatic tissues, perhaps suggesting that the narrow accumulation of reproductive phasiRNAs in anthers is either not a general characteristic or it is the product of evolutionary refinement in the grasses. We also show that the miRNA trigger for pre-meiotic (21-nt) phasiRNAs likely shifted in evolutionary time from targeting pathogen-defense genes to long, non-coding RNAs (observed in grasses) via specialization and sub-functionalization versus neo-functionalization. Finally, we demonstrate that exceptions to the canonical mechanism

of biogenesis of phasiRNAs exist in monocot evolution, whereby phasiRNAs are produced apparently without a miRNA trigger. I conclude that plants show substantial variation in their composition and biogenesis of reproductive phasiRNAs, which have broad roles in plant germline development.

4.1 Methods

Here we summarize the experimental methods for data generation and computational approaches.

4.1.1 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.

4.1.2 Anther stage: 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.

4.1.3 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.

4.1.4 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/helper.github) with default settings as described earlier (Patel et al. 2016) and as described by Mathioni et al. (2017). 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.

4.1.5 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 pbscript-tofu tool set (v2.3.0) using the default settings. 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.

4.1.6 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 (D.-H. Jeong et al. 2013). Steps in both pipelines involved processing using perl scripts as described earlier (D.-H. 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 (D.-H. 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.

4.1.7 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.

4.1.8 PhasiRNA prediction and trigger identification

Phased siRNA generating (*PHAS*) loci or precursors were identified using the purpose-built tool ‘*PHASIS*’ (https://github.com/atulkakrana/PHASIS) (Kakrana et al.

2017). The *PHAS* loci (or precursors), predicted from different sRNA libraries can have different coordinates primarily due to differences in sRNA population and abundances. These library-specific lists of *PHAS* loci were collapsed to a non-redundant set by selecting the coordinates for sRNA library with most abundant phased siRNAs. Triggers for these *PHAS* loci (or precursors) were further identified using the '*revFerno*' script (https://github.com/atulkakrana/*PHASIS*), developed as an extension to the *sPARTA* miRNA prediction and validation tool (Kakrana et al. 2014). *revFerno* requires two input files 1) phased-siRNA prediction results from *PHASIS* and predicted targets for a set of candidate miRNAs from *sPARTA*. Using these files *revFerno* performs a head-scan (from 5'-end) and/or tail-scan (from 3' end) of *PHAS* loci (or precursor) to identify predicted cleavage sites that match with (+3 to -2) phased-siRNA registers, and selects the cleavage site either matching the first register or next closest register. We used revFerno with default settings, that is accounted for (+1/-1-nt) dicer offset, and (+2/-2-nt) strand-offset for *PHAS* loci predicted using genome. As a control for *revFerno* predictions, we first predicted *PHAS* loci in maize using publically available sRNA libraries (Zhai et al. 2015), and then tested *revFerno* using genome-wide target sites predicted by *sPARTA*. 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 (S Mathioni, A Kakrana and B. Meyers, in preparation).

4.1.9 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 (L. 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 ("RNAcentral: An International Database of ncRNA Sequences" 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.

4.1.10 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-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 (B. Li and Dewey 2011) with default settings, as implemented in the Trinity platform.

4.1.11 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, Jagannadham, and Pandit 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: TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACnnnnnnATCTCGTATGCCG TCTTCTGCTTG). 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.

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

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 2013b). All the sRNA libraries (**Table 4.1**) 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 (F. 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. 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 were used to plot the consensus.

4.1.12 Identification of isomiRs and putative miRNA loci in sequenced genomes

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

4.1.13 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, Higgins, and Heringa 2000) in 'accurate' mode. Finally, a phylogenetic tree was generated using PhyML (Guindon et al. 2010) with default parameters that 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.

4.1.14 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), and then with donkey anti-sheep antibodies conjugated to AlexaFluor647 (1/1000). 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.).

Table 4.1: **Summary of probes used for *in situ* experiments**

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

4.1.15 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

4.1.16 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.

Table 4.2: **Samples used for quantitative RT PCR for probing expression of Asparagus DCL5 in vegetative and reproductive tissues**

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

Table 4.3: **Primers used for quantitative RT PCR for probing expression of Asparagus DCL5 in vegetative and reproductive tissues.**

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

4.2 Results

To characterize phasiRNAs in *Asparagus*, we generated sRNA libraries from sequential stages of reproductive tissues, plus a number of vegetative tissues (**Table 4.1**). These included stages of pre- to post-meiotic anthers, as one of our aims was to assess the specificity of accumulation in anthers, in comparison to reproductive phasiRNAs in grasses (Zhai et al. 2015). The leaf and shoot samples were generated to validate patterns of tissue enrichment. In total, we generated 23 sRNA libraries for this study, and we combined these with 15 libraries we recently published but had not extensively characterized (Harkess et al., 2017). In that work, we annotated 166 miRNA precursors in *Asparagus*, generating 105 unique miRNAs from 78 families, including both known/conserved miRNAs and some novel miRNAs (Harkess et al., 2017). Since a saturating level of miRNA identification requires the characterization of many tissues and stages, we investigated the newly-generated sRNA libraries for miRNAs, using two different set of criteria, followed by manual curation of structures (see methods). We identified 15 conserved and 46 lineage- or species-specific 21-nt and 22-nt miRNAs, excluding those miRNAs we recently reported. The temporal

expression patterns of complete repertoire of miRNAs identified in this study is presented in **Figure 1.1**.

Table 4.4: **Summary statistics of garden asparagus (*Asparagus officinalis*) sequencing libraries used in this study.**

Part A. Asparagus small RNA data				
Code	**Title**	**Total Sequences**	**Genome Matched Reads**	**Distinct Genome Matched Reads**
leaf_r1	Asparagus leaf, rep1, BM14-72	30,024,542	26,088,582	2,796,076
ant1_07mm_r1	Asparagus anther 0.7mm, male1, rep1, BM14-48	4,603,029	37,868,769	5,892,680
ant2_11mm_r1	Asparagus anther 1.1mm, male2, rep1, BM14-50	51,259,921	43,323,163	7,446,886
ant3_15mm_r1	Asparagus anther 1.5mm, male3, rep1, BM14-52	42,990,118	36,641,232	6,555,377
ant4_18mm_r1	Asparagus anther 1.8mm, male4, rep1, BM14-54	53,728,958	43,957,932	7,414,084
ant5_20mm_r1	Asparagus anther 2.0mm, male5, rep1, BM14-56	30,392,916	25,408,903	3,691,026
pis_ma1_05mm_r1	Asparagus pistil 0.5mm, male1 rep1, BM14-58	39,685,208	34,496,108	4,344,113
pis_ma2_07mm_r1	Asparagus pistil 0.7mm, male2 rep1, BM14-60	28,875,412	24,855,917	3,889,869
pis_ma3_11mm_r1	Asparagus pistil 1.1mm, male3 rep1, BM14-62	31,683,580	27,569,856	3,411,535
pis_ma4_15mm_r1	Asparagus pistil 1.5mm, male4, rep1, BM14-64	30,283,321	25,730,428	3,707,690
pis_ma5_18mm_r1	Asparagus pistil 1.8mm, male5 rep1, BM14-66	29,394,111	24,907,377	3,319,557
pis_fe1_15mm_r1	Asparagus pistil 1.5mm, female1, rep1, BM14-74	28,794,913	24,810,788	3,503,894
pis_fe2_20mm_r1	Asparagus pistil 2.0mm, female2, rep1, BM14-76	31,706,438	27,193,425	3,437,506
pis_fe3_28mm_r1	Asparagus pistil 2.8mm, female3, rep1, BM14-78	28,882,041	24,124,807	3,750,256
pis_fe4_35mm_r1	Asparagus pistil 3.5mm, female4, rep1, BM14-80	33,046,715	27,551,793	4,760,947
pis_fe5_42mm_r1	Asparagus pistil 4.2mm, female5, rep1, BM14-82	32,476,318	27,107,377	3,298,178
pis_fe6_45mm_r1	Asparagus pistil 4.5mm, female6, rep1, BM14-84	39,544,560	32,972,952	4,612,981
shoot_old_r1	Asparagus shoot, old spear, rep1, BM14-68	33,829,729	28,964,427	3,122,177
shoot_you_r1	Asparagus shoot, young plant, rep1, BM14-70	30,917,138	26,512,078	2,830,267
88_Fs	Spears from 3 female plant, line 88	51,071,610	44,361,206	12,174,564
88_Ms	Spears from 3 male plant, line 88	47,521,321	41,489,178	11,116,150
88_SMs	Spears from 3 super male plant, line 88	53,130,780	46,200,778	11,821,295
89_Fs	Spears from 3 female plant, line 89	29,631,096	25,785,176	9,445,138
89_Ms	Spears from 3 male plant, line 89	24,181,331	37,349,173	9,729,097
89_SMs	Spears from 3 super male plant, line 89	28,687,896	40,755,331	9,644,782
103_Fs	Spears from 3 female plant, line 103	53,644,329	44,843,564	10,916,426
103_Ms	Spears from 3 male plant, line 103	51,686,019	43,075,904	9,482,442
St_Aspa	Asparagus shoot	28,684,914	39,392,106	8,431,386
Lf_Aspa	Asparagus leaf	24,967,647	37,868,082	6,553,948
Rt_Aspa	Asparagus root	28,265,822	15,523,108	2,845,038
MFE_Aspa	Asparagus male flower early stage	20,926,331	33,278,523	8,930,903
MFM_Aspa	Asparagus male flower mid stage	56,362,666	46,066,266	10,196,795
FeFE_Aspa	Asparagus female flower early stage	23,941,398	35,592,215	9,263,600
FeFM_Aspa	Asparagus female flower mid stage	58,262,210	48,153,462	9,493,296
Asp_0_5_ant_bud	Asparagus, <0.5mm anthers (whole buds), BM14-181	36,110,327	27,770,843	7,025,326
Asp_1_ant_bud	Asparagus, 0.5-1.0mm anthers (whole buds), BM14-182	31,534,472	23,350,045	5,321,879
Asp_1_5_ant_bud	Asparagus, 1.0-1.5mm anthers (whole buds), BM14-183	41,127,611	33,446,841	4,022,099
Asp_1_0_ant	Asparagus, 0.5-1.0mm anthers, BM14-184	33,450,319	26,021,154	4,071,462
Asp_1_5_ant	Asparagus, 1.0-1.5mm anthers, BM14-185	37,476,488	29,030,034	7,179,341

Part B. Asparagus RNA-seq data

Code	Title	Total Sequences[a]	Genome Matched Reads	Distinct Genome Matched Reads
Asp_leaf	Asparagus, leaf, BM14-72	31,990,892	14,690,194	3,111,600
Asp_0_5_ant_budr	Asparagus, <0.5mm anthers (whole buds), BM14-181	26,788,747	10,593,969	2,705,490
Asp_1_ant_budr	Asparagus, 0.5-1.0mm anthers (whole buds), BM14-182	24,637,152	10,657,889	2,806,443
Asp_le	Asparagus, leaf, BM14-72	34,433,768	15,163,706	2,791,096
Asp_0_5_ant_b	Asparagus, <0.5mm anthers (whole buds), BM14-181	30,330,547	11,162,814	2,399,093
Asp_1_ant_b	Asparagus, 0.5-1.0mm anthers (whole buds), BM14-182	30,145,994	12,319,937	2,626,547

Part C. Asparagus RNA Structure data

Code	Title	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads
05_ant_bud_o	Asparagus dsRNA, <0.5mm anthers (whole buds), BM14-181, Rnase ONE	12,342,640	6,897,413	294,724
05_10_ant_bud	Asparagus dsRNA, 0.5-1.0mm anthers (whole buds), BM14-182, Rnase ONE	12,111,370	7,107,136	255,080
05_ant_bud_v	Asparagus dsRNA, <0.5mm anthers (whole buds), BM14-181, Rnase V1	20,531,053	15,283,502	215,299
05_10_ant_bud_v	Asparagus dsRNA, 0.5-1.0mm anthers (whole buds), BM14-182, Rnase V1	13,925,322	9,800,453	263,252

Part D. Asparagus PacBio "SMRT" data

Code	Title	Total Sequences	Full-length[b]	high-quality transcripts[c]
Aspa-denovo-A-2kb	Asparagus, <=0.5 and 0.5-1.0mm anther, <2kb insert length (3 SMRT cells)	213,951	120,540	30,059
Aspa-denovo-B-2kb-3kb	Asparagus, <=0.5 and 0.5-1.0mm anther, 2-3kb insert length (3 SMRT cells)	294,820	106,233	29,932
Aspa-denovo-C-3kb	Asparagus, <=0.5 and 0.5-1.0mm anther, >3kb insert length (3 SMRT cells)	240,945	48,792	13,083

Part E. *Asparagus* PARE data

Code	Title	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads
88F_d	Spears from a female plant, line 88	[redacted]	[redacted]	[redacted]
88M_d	Spears from a male plant, line 88	[redacted]	[redacted]	[redacted]
88SM_d	Spears from a super male plant, line 88	[redacted]	[redacted]	[redacted]
89F_d	Spears from a female plant, line 89	[redacted]	[redacted]	[redacted]
89M_d	Spears from a male plant, line 89	[redacted]	[redacted]	[redacted]
89SM_d	Spears from a super male plant, line 89	[redacted]	[redacted]	[redacted]
103F_d	Spears from a female plant, line 103	[redacted]	[redacted]	[redacted]
103M_d	Spears from a male plant, line 103	[redacted]	[redacted]	[redacted]
AspM_mid	Male flower mid stage	[redacted]	[redacted]	[redacted]
AspM_ear	Male flower early stage	[redacted]	[redacted]	[redacted]
AspFM_mid	Female flower mid stage	[redacted]	[redacted]	[redacted]
AspFM_ear	Female flower early stage	[redacted]	[redacted]	[redacted]
Asp_shoot	Shoot	[redacted]	[redacted]	[redacted]
Asp_Lf	Leaf	[redacted]	[redacted]	[redacted]

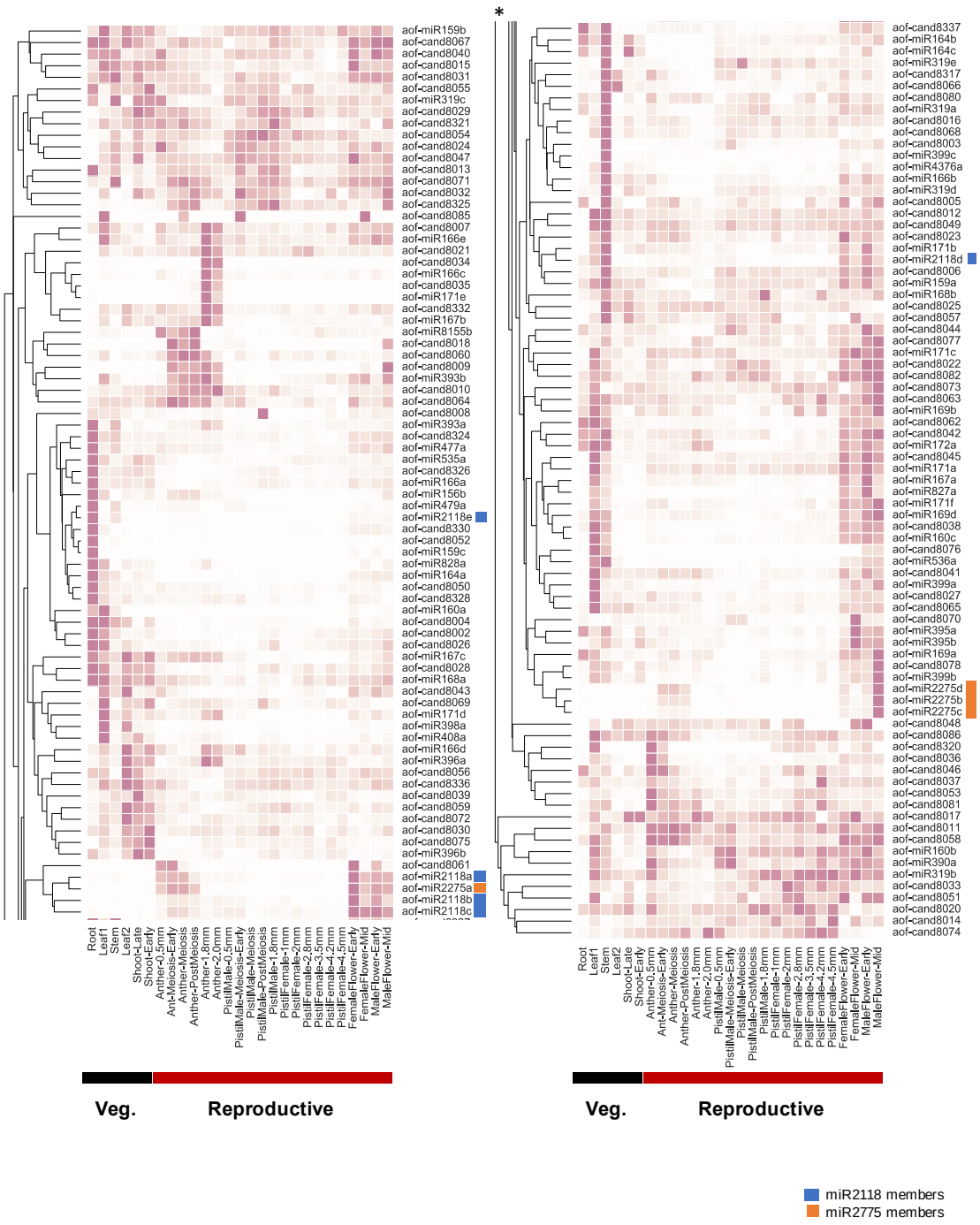

*
aof-miR159b
aof-cand8067
aof-cand8040
aof-cand8015
aof-cand8031
aof-cand8055
aof-miR319c
aof-cand8029
aof-cand8321
aof-cand8054
aof-cand8024
aof-cand8047
aof-cand8013
aof-cand8071
aof-cand8032
aof-cand8325
aof-cand8085
aof-cand8007
aof-miR166e
aof-cand8021
aof-cand8034
aof-miR166c
aof-cand8035
aof-miR171e
aof-cand8332
aof-miR167b
aof-miR8155b
aof-cand8018
aof-cand8060
aof-cand8009
aof-miR393b
aof-cand8010
aof-cand8064
aof-cand8008
aof-miR393a
aof-cand8324
aof-miR477a
aof-miR535a
aof-cand8326
aof-miR166a
aof-miR156b
aof-miR479a
aof-miR2118e
aof-cand8330
aof-cand8052
aof-miR159c
aof-miR828a
aof-miR164a
aof-cand8050
aof-cand8328
aof-miR160a
aof-cand8004
aof-cand8002
aof-cand8026
aof-miR167c
aof-cand8028
aof-miR168a
aof-cand8043
aof-cand8069
aof-miR171d
aof-miR398a
aof-miR408a
aof-miR166d
aof-miR396a
aof-cand8056
aof-cand8336
aof-cand8039
aof-cand8059
aof-cand8072
aof-cand8030
aof-cand8075
aof-miR396b
aof-cand8061
aof-miR2118a
aof-miR2275a
aof-miR2118b
aof-miR2118c
Root
Leaf1
Stem
Leaf2
Shoot-Late
Shoot-Early
Anther-0.5mm
Ant-Meiosis-Early
Anther-Meiosis
Anther-PostMeiosis
Anther-1.8mm
Anther-2.0mm
PistilMale-0.5mm
PistilMale-Meiosis-Early
PistilMale-Meiosis
PistilMale-PostMeiosis
PistilMale-1.8mm
PistilFemale-1mm
PistilFemale-2mm
PistilFemale-2.8mm
PistilFemale-3.5mm
PistilFemale-4.2mm
PistilFemale-4.5mm
FemaleFlower-Early
FemaleFlower-Mid
MaleFlower-Early
MaleFlower-Mid
Veg.
Reproductive
aof-cand8337
aof-miR164b
aof-miR164c
aof-miR319e
aof-cand8317
aof-cand8066
aof-cand8080
aof-miR319a
aof-cand8016
aof-cand8068
aof-cand8003
aof-miR399c
aof-miR4376a
aof-miR166b
aof-miR319d
aof-cand8005
aof-cand8012
aof-cand8049
aof-cand8023
aof-miR171b
aof-miR2118d
aof-cand8006
aof-miR159a
aof-miR168b
aof-cand8025
aof-cand8057
aof-cand8044
aof-cand8077
aof-miR171c
aof-cand8022
aof-cand8082
aof-cand8073
aof-cand8063
aof-miR169b
aof-cand8062
aof-cand8042
aof-miR172a
aof-cand8045
aof-miR171a
aof-miR167a
aof-miR827a
aof-miR171f
aof-miR169d
aof-cand8038
aof-miR160c
aof-cand8076
aof-miR536a
aof-cand8041
aof-miR399a
aof-cand8027
aof-cand8065
aof-cand8070
aof-miR395a
aof-miR395b
aof-miR169a
aof-cand8078
aof-miR399b
aof-miR2275d
aof-miR2275b
aof-miR2275c
aof-cand8048
aof-cand8086
aof-cand8320
aof-cand8036
aof-cand8046
aof-cand8037
aof-cand8053
aof-cand8081
aof-cand8017
aof-cand8011
aof-cand8058
aof-miR160b
aof-miR390a
aof-miR319b
aof-cand8033
aof-cand8051
aof-cand8020
aof-cand8014
aof-cand8074
Root
Leaf1
Stem
Leaf2
Shoot-Late
Shoot-Early
Anther-0.5mm
Ant-Meiosis-Early
Anther-Meiosis
Anther-PostMeiosis
Anther-1.8mm
Anther-2.0mm
PistilMale-0.5mm
PistilMale-Meiosis-Early
PistilMale-Meiosis
PistilMale-PostMeiosis
PistilMale-1.8mm
PistilFemale-1mm
PistilFemale-2mm
PistilFemale-2.8mm
PistilFemale-3.5mm
PistilFemale-4.2mm
PistilFemale-4.5mm
FemaleFlower-Early
FemaleFlower-Mid
MaleFlower-Early
MaleFlower-Mid
Veg.
Reproductive
miR2118 members
miR2775 members

Figure 4.1: **Heat maps showing normalized expression of conserved and species-specific miRNAs in Asparagus**. miRNA abundances were assessed using the small RNA data from vegetative tissues, male flowers, female flowers, anthers, degenerate pistils from male flower, and fertile pistils from female flowers; all libraries were normalized to transcripts per 20 million reads (TP20M). Lineage- or species-specific miRNA candidates have the "cand" prefix in their names. Reproductive phasiRNA triggers miR2118 and miR2275 are highlighted by blue and orange sidebars. All miRNAs are hierarchically clustered based on abundances across tissues as indicated by the tree at the left (split across the two portions) using the "single" method and "Euclidean" distances.

4.2.1 Presence of miR2118 and miR2275, triggers of reproductive phasiRNAs, in Asparagus

We next sought to examine miRNAs in *Asparagus* flowers. *Asparagus* is dioecious, and its flowers develop initially as hermaphrodites, but later matures into a female or male, primarily due to degeneration of the other sex organ (Harkess et al., 2017). As a result, male flowers include both an aborted pistil and the fertile anther. We collected male flowers, six stages of fertile anthers, and five stages of aborted pistils for small RNA analysis. Comparing 105 conserved miRNAs from 78 families, we observed a high degree of similarity between fertile anthers and aborted pistils (**Figure 4.2A**), indicating that events leading to cessation of pistil development in male flowers might not include significant variation in sRNA regulation or could have taken place prior to initiation of stamen primordia. We observed significant disparity between pre-meiotic and meiotic anthers, primarily due to seven miRNA families that were enriched (fold change ≥ 4, $p \leq 0.05$) relative to post-meiotic anthers. These miRNAs included miR160, miR166, miR171, miR319 and miR390, all known to play

roles in regulation of flowering time (Schommer et al. 2012; Rubio-Somoza and Weigel 2011), floral organ formation (A. C. Mallory et al. 2004; Nagasaki et al. 2007) and vegetative-to-reproductive phase transition (Curaba et al. 2013). In addition to these miRNAs, two triggers of reproductive phasiRNAs – miR2118 and miR2275 – were preferentially expressed in all the reproductive libraries, including in pistils from female flowers, yet also highly enriched in pre-meiotic and meiotic stages (**Figure 4.1**).

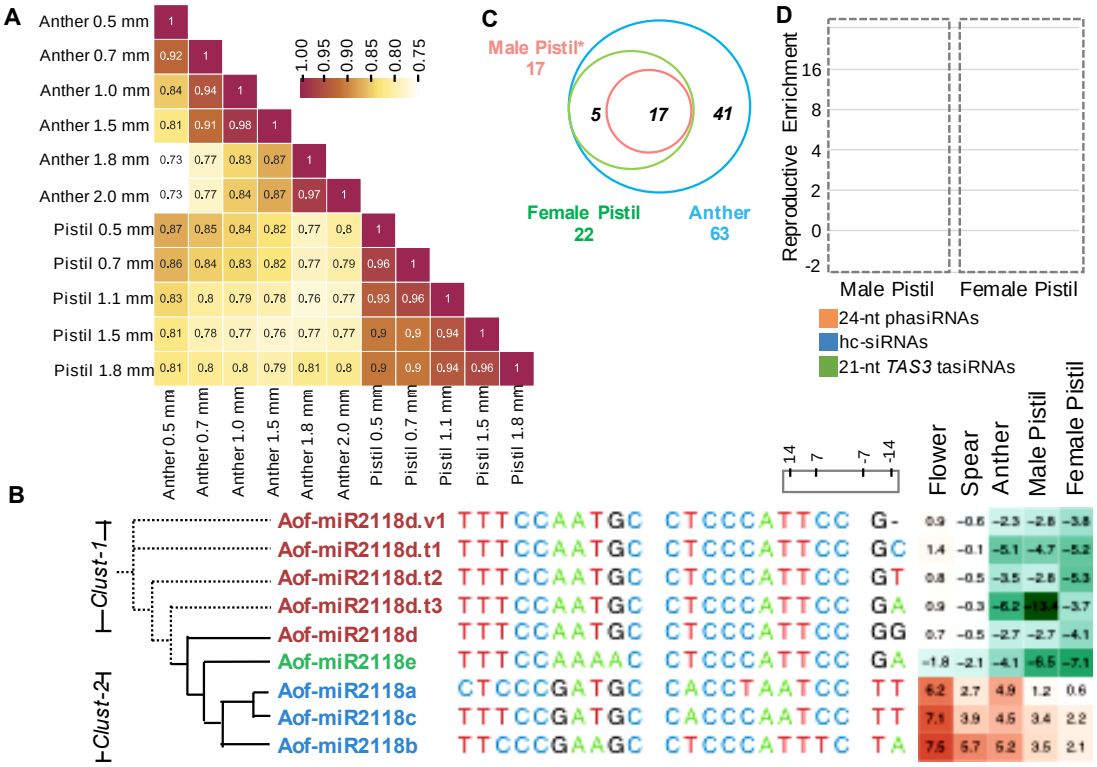


Figure 4.2: **miRNA abundances in Asparagus flowers, and phasiRNAs in female pistils.** (A) Heat map representing the Pearson's correlation values for an all-versus-all comparison of miRNA abundance levels in developmental stages of anther and degenerated pistils from male flowers. The pistil length corresponds to the stage of the anther of that specific length. (B) The miR2118 family in Asparagus, with heat-map showing enrichment or depletion in reproductive tissues relative to the leaf samples; variants of miR2118d are described in the main text. The numbers represent enrichment level in log (2) scale, as indicated above. Solid lines in phylogenetic tree represents genomic variants of miR2118 family while the dotted lines represent transcriptional variants of miR2118d found in sRNA libraries. (C) Venn diagrams show counts of 24-nt *PHAS* loci identified in aborted male pistils and fertile female pistils, and their overlap with the set of 24-nt *PHAS* loci from anthers of male flowers. (D) Bar plots showing enrichment of 24-nt phasiRNAs, tasiRNAs, and hc-siRNAs in fertile pistils, represented in a log (2) scale.

miR2118 and miR2275 both trigger reproductive phasiRNAs and thus we analyzed these miRNAs and their precursors in *Asparagus.* In grasses, miR2118 triggers reproductive 21-nt phasiRNAs (Zhai et al. 2015; Johnson et al. 2009), while in eudicots, miR2118 members target *NB-LRR* disease resistance genes (Shivaprasad et al. 2012). In *Asparagus*, five miR2118 members are generated from three loci (**Figure S3**); two members (miR2118d/e) accumulate in both vegetative and reproductive stages (**Figure 4.1 and 4.2B**), while the other three members (miR2118a/b/c) from a genomic cluster (**Figure 4.3**) displayed reproductive enrichment that peaks in pre-meiotic anthers (**Figure 4.1 and 4.3B**), similar to the Poaceae (D. H. Jeong et al. 2013; Zhai et al. 2013; Song, Li, et al. 2012). This dichotomy in tissue specificity and temporal dynamics may suggest that miR2118 shifted in evolutionary time from targeting NB-LRRs (observed in mosses and later-diverged species) to long, non-coding RNAs (observed in grasses) via specialization and subfunctionalization versus neo-functionalization.

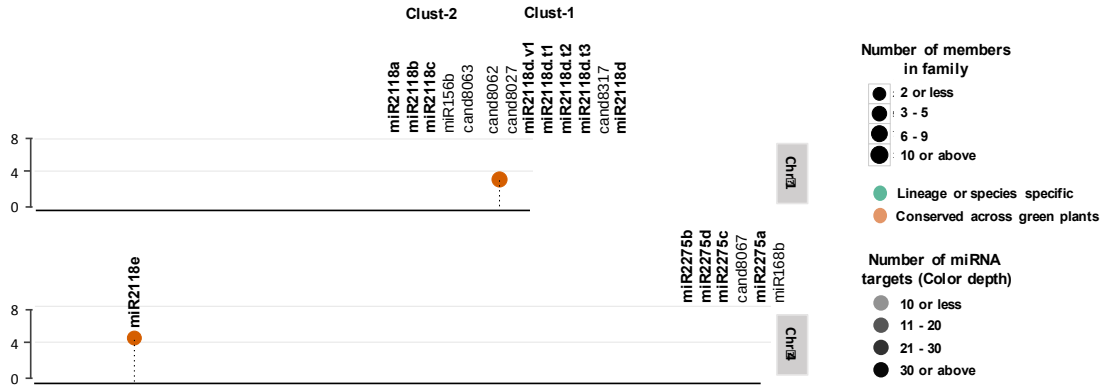


Figure 4.3: **Genomic organization and abundances of known and novel miRNAs in chromosomes 1 and 4 of the Asparagus genome**. miRNAs were mapped to the Asparagus genome with chromosomes as indicated at right, and the abundance of each miRNA is displayed in a dot with the size indicated in a Log10 scale. The miR2118 family is encoded at three loci and miR2275 family is encoded in a single cluster on chromosome 4. The Y-axis is a representation of genomic positions of miRNAs.

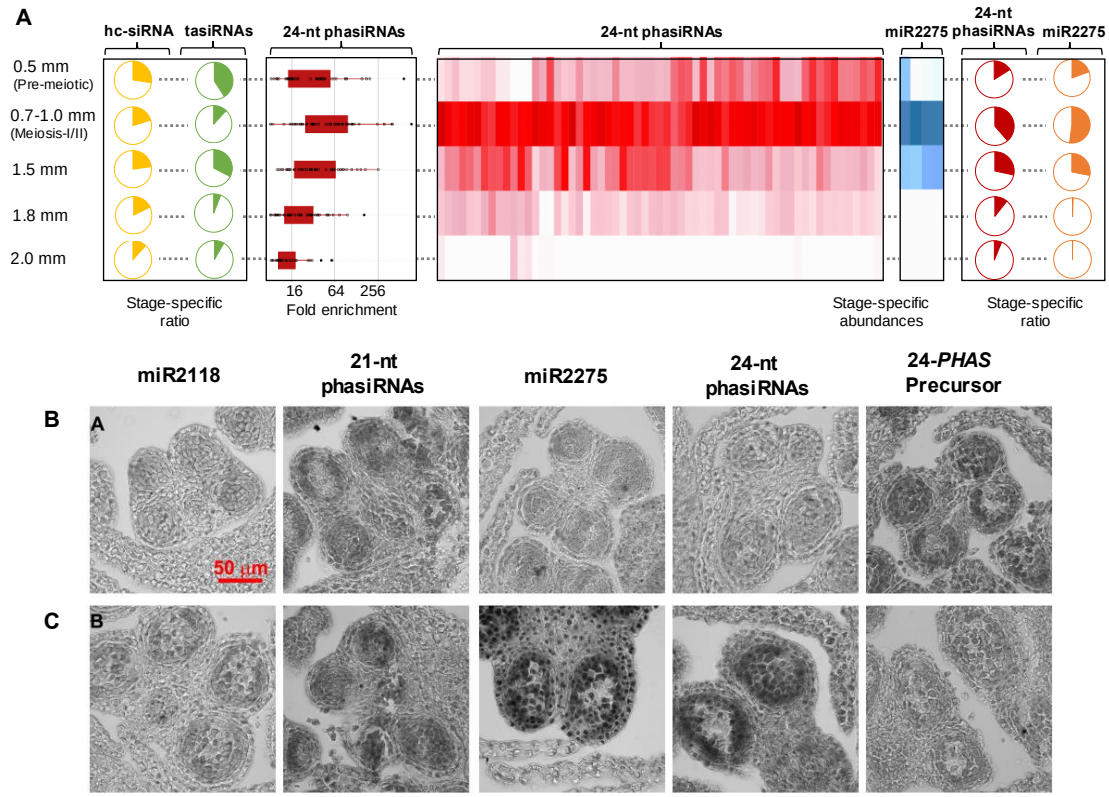


Figure 4.4: **Reproductive phasiRNAs and their triggers in *Asparagus*.** Heat maps depicting abundance of 24-nt phasiRNAs (in red) and their triggers, miR2275 (in blue), in developing anthers. Both heat maps are clustered on their similarity of expression. Pie charts at left or right represent the proportion of all small RNA abundances comprised by the 24-nt phasiRNAs (in red), miR2275 (in orange), hc-siRNAs (in yellow) and *TAS3* tasiRNAs (in green) across anther developmental stages. Box-whisker plots indicate enrichment (log2) of *Asparagus* 24-nt phasiRNAs abundance from all *PHAS* loci in the meiotic anther compared to the vegetative sample (leaf). *(B)* and *(C)* Small RNA *in situ* hybridization with probes for the following, from left to right: (i) miR2118, (ii) a pre-meiotic phasiRNA from locus 21-*PHAS*-4, (iii) miR2275, and (iv) a meiotic IR-related phasiRNA from locus 24-*PHAS*-31. The right-most images show mRNA *in situ* hybridizations with probes for the 24-*PHAS*-31 precursor. The scale bar indicates 50 µm, for all images. (*B*) Images from pre-meiotic anthers. (*C*) Images from meiotic-stage anthers.

4.2.2 PhasiRNAs in Asparagus

We next examined loci generating phasiRNAs in *Asparagus*, coupling the small RNA analysis with the identification of the potential miRNA triggers (see methods). From the sRNA data described above, we identified 29 loci generating 21-nt phasiRNAs and 42 loci generating 24-nt phasiRNAs. Of the 29 "21-*PHAS*" loci, 23 were vegetatively expressed and 15 (~55%) overlapped annotated protein-coding genes. Only three reproductive-enriched 21-*PHAS* loci were in intergenic regions, presumably long, non-coding RNAs (lncRNAs), substantially lower than the hundreds or thousands seen in maize or rice (Zhai et al. 2015; Fei et al. 2016). In contrast, all 42 24-*PHAS* loci were enriched in reproductive tissues (by 6- to 380-fold), peaked in abundance at meiotic stages (t-test, $p \leq 0.05$), and correlated with miR2275 abundances (**Figure 4.4A**). This spatiotemporal pattern, similar to those of the grasses (Komiya et al. 2014; Zhai et al. 2015), is distinct from both *TAS3* tasiRNAs and heterochromatic siRNAs (hc-siRNAs) (**Figure 4.4A**). Unlike the lncRNA precursors of grasses (Johnson et al. 2009; Zhai et al. 2015), we found 30% (n = 18) of 24-*PHAS* loci overlap annotated protein-coding genes. We also noted that some 24-nt phasiRNAs were abundant in the data from aborted pistils from male flowers; prior work had associated these phasiRNAs only with the male reproductive organs (Zhai et al. 2015). In order to explore the presence of 24-nt phasiRNA in female pistils, we examined small RNAs from six sequential stages of fertile pistils from female flowers (**Table 4.1**). Enriched in these data, we identified 22 24-*PHAS* loci, a subset of those identified from anthers and aborted male pistils (**Figure 4.2C and 4.2D**). Unlike anthers, the 24-nt phasiRNAs in fertile pistils lacked a clear peak in temporal abundance, most likely due to sampled stages that varied relative to those in anthers.

To investigate and validate the 24-*PHAS* loci that overlap with protein-coding genes, we sequenced from meiotic *Asparagus* anthers a set of 275,565 putative full-length, polyadenylated transcripts using single molecule real time (SMRT) sequencing. These data included 5,671 annotated genes, at least 23,797 novel isoforms, and 1,771 unannotated loci. After updating the genome annotations with these data, we identified transcripts overlapping 68% (n = 30) of 24-*PHAS* clusters; in most cases, 24-*PHAS* loci overlap intronic regions, except for three loci (7%) where at least one transcript with an exonic overlap (≤ 50% of *PHAS* length) was detected. We also identified transcripts corresponding to 88% of 21-*PHAS* loci, and 54% of these overlapping transcript(s) showed overlap with exonic regions.

To better evaluate the coding potential of transcripts overlapping 21- and 24-*PHAS* loci, we built a classifier that categorizes transcripts as coding, non-coding or of unknown coding potential (TUCP) by using Coding Potential Calculator (Kong et al. 2007) and Coding-Potential Assessment Tool (L. Wang et al. 2013) scores to integrate sequence- and similarity-based features (see methods). Over 62% of 21-*PHAS* loci and 42% of 24-*PHAS* loci overlapped predicted protein-coding transcripts. The 24-*PHAS* precursors in grasses are thought to be lncRNAs (Johnson et al. 2009; Zhai et al. 2015). The production of 21-nt phasiRNAs from protein-coding genes is well-described for vegetative tissues in both eudicots and gymnosperms (R. Xia, Xu, et al. 2015; Shivaprasad et al. 2012; Zhai et al. 2011; Arikit et al. 2014). In *Asparagus*, these 21-*PHAS* transcripts included five *NB-LRR*s (disease resistance genes), *DCP* (encoding an mRNA de-capping enzyme), *SGS3* (*SUPPRESSOR OF GENE SILENCING 3*), and other gene families.

In maize, reproductive phasiRNAs display distinct spatiotemporal patterns of accumulation (Zhai et al. 2015); therefore, we decided to assess whether this is also the case in *Asparagus*. We performed small RNA *in situ* hybridizations using probes for the following: (i) miR2118, (ii) a pre-meiotic 21-nt phasiRNA, (iii) miR2275, and (iv) a meiotic IR-derived 24-nt phasiRNA. All of these *Asparagus* small RNAs displayed distinct spatiotemporal patterns (**Figure 4.1B and 4.1C**). miR2118, the trigger of limited reproductive-enriched 21-*PHAS* loci, co-localized with phasiRNAs in the middle layer, tapetum and archesporial (AR) cells. Like the maize meiotic 24-nt phasiRNAs that require normal tapetal differentiation and localize in the tapetum and germinal cells, *Asparagus* 24-nt phasiRNAs were enriched in meiotic stages and predominantly localized in the tapetum and meiocytes, precisely where miR2275 was present (**Figure 4.2B and C**).

4.2.3 MicroRNA triggers and biogenesis components of 21- and 24-nt phasiRNAs in Asparagus

We next investigated the miRNA triggers for phasiRNAs, focusing on reproductive 21-*PHAS* and 24-*PHAS* loci. For this, we first generated a list of computationally predicted and PARE-supported target sites using *sPARTA* (Kakrana et al. 2014), followed by an exhaustive search for miRNA triggers using the '*revFerno*' extension (see methods). Of 29 21-*PHAS* loci, we found triggers for 16 (57%), most (n = 14) triggered by miR2118 family members. These loci included the three reproductive-specific 21-*PHAS* loci mentioned above, plus seven protein-coding loci. Isoforms (22-nt) of miR167 and miR390 triggered 21-nt phasiRNAs from an *AUXIN-RESPONSE FACTOR* (*ARF*) gene and an intergenic locus respectively, the latter likely a *TAS3* locus (Axtell et al. 2006). The numbers of "*pNLs*" (phased *NB-LRRs*) (n

= 5) and *TAS3* (n = 1) loci identified in *Asparagus* are substantially lower than in Norway spruce or eudicots (R. Xia, Xu, et al. 2015; Shivaprasad et al. 2012; Zhai et al. 2011; Arikit et al. 2014). The reduction in *pNLs* may reflect a relatively small compositions of *NB-LRR* genes in *Asparagus* (n = 46) (Y. Zhang et al. 2016), similar to *Zostera* and *Amborella* with only 44 (Olsen et al. 2016) and 45 members reported respectively (Y. Zhang et al. 2016). The low numbers or absence of *pNLs* in monocots and the concordant presence of miR2118-triggered, reproductive 21-phasiRNAs relative to eudicots suggests divergent paths of miR2118 function.

For the reproductive 24-*PHAS* loci, our analyses failed to identify a miRNA trigger, including miR2275. To date, all 24-nt phasiRNAs were reported in grasses and triggered by miR2775 family members (Zhai et al. 2015; Johnson et al. 2009). We did identify 15 intergenic and 10 protein-coding targets, supported by PARE data, for miR2275 in *Asparagus*; while none of these targets corresponded to 24-*PHAS* loci, weakly abundant, anther-enriched 22-nt phasiRNAs were found at three intergenic targets. We attempted to find miR2275-triggered 22-*PHAS* loci in maize and rice, in both reproductive and vegetative tissues, using published data (Zhai et al. 2015; Fei et al. 2016), but none were found. Next, we tested the remote possibility of 22-nt phasiRNAs (secondary siRNAs) triggering 24-nt phasiRNAs in a tertiary cycle, perhaps analogous to piRNAs production in insects (Brennecke et al. 2007). For 101 22-nt phasiRNAs, 29,594 targets sites (score < 4) were identified, among which 1,430 had matched PARE reads at any abundance level, but only one target site passed stringent filters (abundance > 5TP30M in at least one library and the ratio of PARE signal to a ± 10-nt window ≥ 0.75). This target site and others found using relaxed filters showed no evidence of 21-, 22- and 24-nt phasiRNAs. In addition, no 22-nt

phasiRNAs were assigned as triggers to 24-*PHAS* loci using *revFerno* (see methods). These observations left us with no clear functional explanation for the 22-*PHAS* loci, and perhaps more importantly, there were evidently no miRNA triggers of the 42 reproductive 24-*PHAS* loci identified in *Asparagus*.

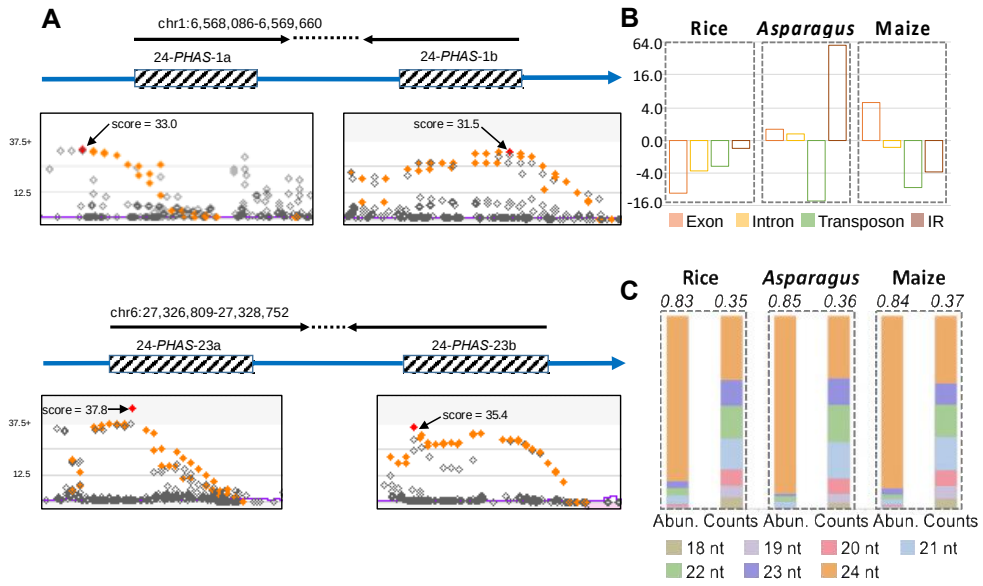


Figure 4.5: **Many *Asparagus* 24-*PHAS* loci are derived from inverted repeats.** *(A)* Genomic organization of two representative IR-type *PHAS* loci, overlapping with 5' and 3'-arm of inverted repeats. Phasing scores are presented as shown in our custom web viewer. *(B)* Fold-change representing enrichment or depletion of overlap of 24-*PHAS* loci from rice, *Asparagus* and maize, with exons, introns, transposons and inverted repeats, against the random chance. *(C)* Comparison of abundances and counts of sRNAs produced from 24-*PHAS* loci from rice, *Asparagus* and maize. The values on top represent 2.5% trimmed mean of ratio of abundances or counts of 24-nt phasiRNAs from all 24-*PHAS* loci of corresponding species.

During inspection of the 24-*PHAS* loci in our sRNA browser, we observed that many showed a substantial strand bias, which is inconsistent with the *RDR6*-dependent biogenesis of *PHAS* precursors in grasses (Song, Wang, et al. 2012). Moreover, we also noticed that these corresponded to inverted repeats ("IRs"), in other plants like Arabidopsis processed by DCL4 or even DCL1 into 21-nt phasiRNAs or miRNAs. We performed a genome-wide analysis, and 90% of *Asparagus* 24-*PHAS* loci corresponded to inverted repeats (IRs); based on their overlap to 5' or 3' arms of an IR (**Figure 4.5A and 4.6**). To test the statistical significance of this overlap, we computed an enrichment or depletion of 24-*PHAS* overlap with exons, introns, TE-related regions and IRs, versus random chance; we also assessed 24-*PHAS* loci from maize and rice using 176 and 111 24-*PHAS* loci, respectively. Both grasses had few 24-*PHAS* loci overlapping IRs (**Figure 4.5B**), particularly in contrast to the ~55-fold enrichment observed in *Asparagus*. All three species displayed a relative paucity of 24-*PHAS* loci in TE-related regions (**Figure 4.5B**), distinguishing these small RNAs from hc-siRNAs. 24-*PHAS* loci from all three species, whether from IRs or not, displayed similar distributions of sizes (**Figure 4.5C**), suggesting similar efficiencies of Dicer processing.

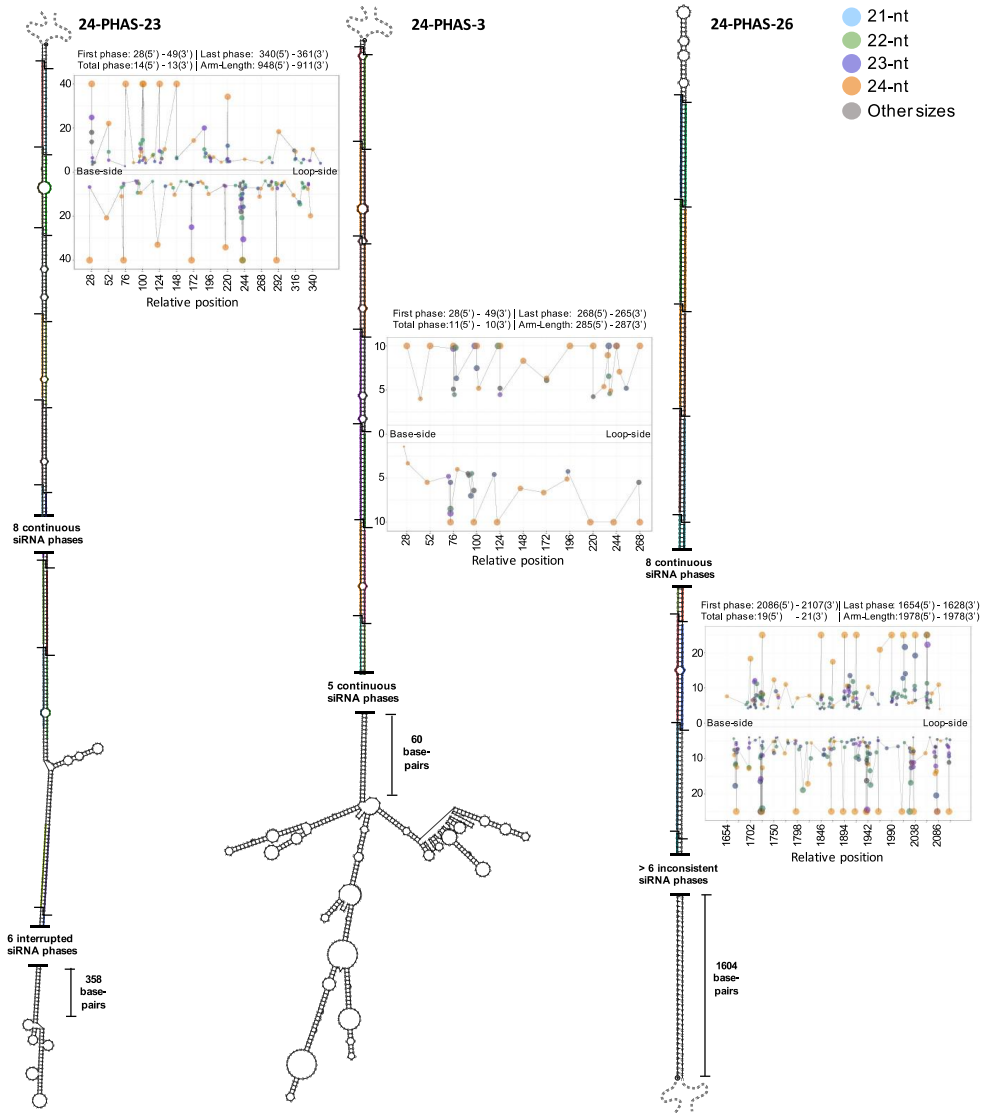


Figure 4.6: **Secondary structure and small RNA abundance plots of three representative hairpin *PHAS* loci from Asparagus.** Foldbacks from unspliced genomic sequence display 24-nt siRNAs from both arms, at 24-nt intervals, a processive signature of Dicer activity. Inset scatterplots depict the sRNA distribution on *PHAS* transcripts, starting from the 5'-most 24-nt phasiRNA. The abundance is indicated on the Y-axis, shown in log2 scale, and axis limits set to 40, 10 and 20 for 24-*PHAS*-23, 24-*PHAS*-3 and 24-*PHAS*-26 respectively. The position of the first and last phasiRNAs for the 5'- and 3' arms, along with the total phases and arm lengths, are described in the header of each scatterplot. The dot colors and sizes represent sRNA sizes and abundances, respectively.

4.2.4 Inverted repeat precursors of 21- and 24-nt phasiRNAs in Asparagus

IR-derived 24-*PHAS* loci were not identified in our earlier work in maize (Zhai et al. 2015), so we sought to better characterize them. We found four loci corresponding to two clusters (clust-19 and clust-125) from the 5' and 3' arms of two long inverted repeats of length 1,188 nt and 9,433 nt (**Figure 4.7**). Both clusters had a single miR2275 target site in their 5' arm. In both cases, the phasiRNAs were precisely spaced at the predicted base of the putative foldback, distal to the loop. In this regard, clust-125 was more unusual, as it also has a tandem repeat of the IR – i.e. two more 24-*PHAS* loci flanking the foldback, with sequence similarity (>94%) (**Figure 4.7**).

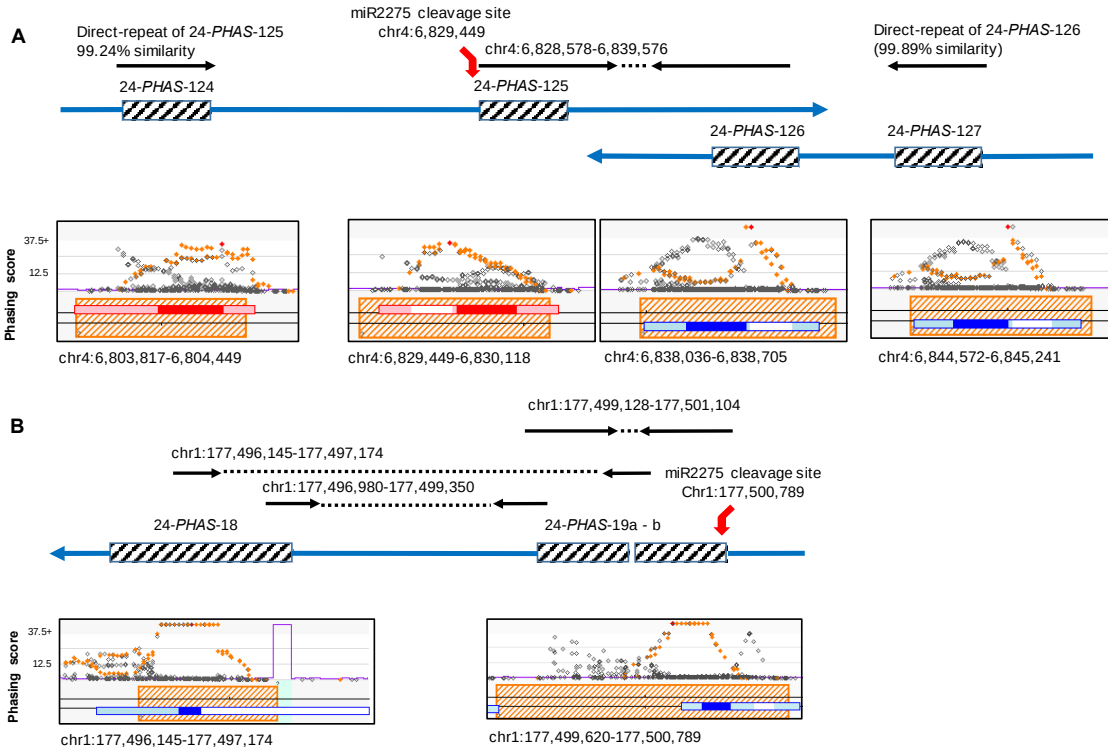


Figure 4.7: **24-nt *PHAS* loci in maize derived from inverted repeats.** (A) Maize cluster-125, with two 24-nt *PHAS* loci precisely located at edges of the 5' and 3' arms of a 9433-nt inverted repeat, and flanked by 24-nt loci that are direct repeats. Inset images are screenshots of our browser showing the phasing scores of 24-nt sRNAs from this region with the red dot indicating the maximum score and orange dots are sRNAs in phase (grey are out of phase). Red or blue boxes are annotated genes on the top or bottom strand; orange cross-hatched boxes indicate that we have marked this as a 24-nt *PHAS* locus. Positions are from version 2 of the maize genome. (B) Maize cluster-19 is a *PHAS* locus with an internal foldback structure, but flanked by another 24-nt *PHAS* locus on left, both are located in the 5' and 3' arms of a fragmented but longer inverted repeat. The distance between *PHAS* loci in (A) and sequence similarity between the 5' and 3' arms of a longer inverted-repeat in (B) suggest that these longer inverted repeats are likely disrupted during evolution. Small RNA libraries for maize meiotic anthers from Zhai et al., 2016 were used for these plots.

Finally, to validate the foldback structure of the 24-*PHAS* precursors in *Asparagus*, we sequenced libraries of mRNA from pre-meiotic and meiotic anthers, and from leaf samples (**Table 4.1**). For analysis, we assembled these data into a transcriptome to accommodate assembly errors (gaps and false joins). The assembly was done by two different approaches: (i) a genome-based assembly using a step-wise Cufflinks protocol (see methods) (Trapnell et al. 2012), and (ii) a *de novo* hybrid assembly using RNA-seq and SMRT libraries (see methods) (Grabherr et al. 2011). The genome-based assembly yielded 46,698 transcripts from 26,687 transcriptional loci, which included 19,660 transcripts from annotated genes, 17,437 new isoforms and 9,601 new transcriptional loci. The *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.

To identify the 21- and 24-*PHAS* precursors from the assembled transcripts, we first mapped transcripts to the genome, and identified those that overlap with predicted *PHAS* loci. For these mapped transcripts, a foldback potential (FP) was computed using *einverted* (Rice, Longden, and Bleasby 2000), which along with the minimum length of overlap with genomic *PHAS* loci allowed us to divide precursors into four categories: **I.** Phased transcripts with full-length or near full-length precursors (> 85% coverage of a *PHAS* locus), that formed at least 240 nt of foldback (FP $\geq$ 500). **II.** Processed precursors with $\leq$ 85% *PHAS* coverage, and high foldback potential (FP $\geq$ 500). **III.** Near full-length precursors (> 85% *PHAS* coverage) with low foldback potential (FP < 300). **IV.** All other *PHAS*-matched transcripts, i.e.

processed or incomplete precursors. This analysis identified 74 precursors from 29 unique 24-*PHAS* loci, 11 represented by at least one category I or II precursors. All of these eleven precursors were non-coding (one had weak coding potential); these had an average arm length of 510 nt and terminal loop of 115 nt. Category I/II precursors comprised 38% of the total 24-*PHAS* precursors, whereas category IV comprised 48%, suggesting a high degree of processed transcripts in our data. Only 13.7% of precursors were in category III, unstructured or of unknown secondary structure. These 24-*PHAS* precursors all lacked miR2275 trigger sites. In contrast to 24-*PHAS* precursors, 83% of 21-*PHAS* precursors (n=103 from 22 loci) corresponded to category III, a clear absence of secondary structure. However, we found two 2,200 nt category I precursors for one 21-*PHAS* locus, with a PARE-supported miR2118 trigger site matching the first phasiRNA position in the 5'-arm. The presence of a miRNA target site in an IR 21-*PHAS* precursor is puzzling, as foldback RNAs are substrates for Dicer even without RDR6 activity (Henderson et al. 2006; Dunoyer, Himber, and Voinnet 2005), yet a miRNA trigger is required to initiate phasing (Arribas-Hernández et al. 2016). Coupled with the observation (above) that *Asparagus* 24-*PHAS* loci lack miR2275 trigger sites, including IR precursors, we found remarkable diversity in the reproductive phasiRNA precursors of *Asparagus*

We next examined RNA secondary structure in the meiotic anther transcriptome to confirm if these precursors were indeed processed as foldback dsRNAs. For this, we used a structure-mapping approach (Q. Zheng et al. 2010) to identify the paired (dsRNA) and unpaired (single-stranded RNA, ssRNA) components in meiotic anthers (0.7 to 1.1 mm). This method generated ~32 million dsRNA and ~37 million ssRNA genome-mapped reads (**Table 4.1**). We focused on 24-*PHAS* loci,

finding 23 of 42 24-*PHAS* loci with sufficient coverage (see methods); 22 showed a significant ($p \leq 0.05$) structure score (log-odds ratio of base-pairing probability > 2.5), consistent with base pairing, while 13 of these 22 showed significant ($p \leq 0.05$) strand specificity (log-odds ratio of base-pairing probability > 2.5). A higher proportion might be found with less stringent filters, or by allowing lower identity between paired arms of the IRs. Comparing these loci with those associated with miRNAs, hc-siRNAs, tasiRNAs, or the IR-derived 21-*PHAS* locus (above) revealed varying types of RNA structure. miRNAs exhibited strong secondary-structure and strand bias indicative of intra-molecular interactions (**Figure 4.8A**); 21-*PHAS* loci exhibited strong secondary structure with low strand specificity. The 24-*PHAS* loci showed intramolecular structure whereas hc-siRNAs showed an intermolecular structure (**Figure 4.8A**) consistent with earlier reports (F. Li et al. 2012) and a role of RDR2 in hc-siRNA biogenesis. For the most structured 24-*PHAS* loci, we observed two significant structure 'peaks' separated by a 'valley', supporting formation of the foldback structure from 24-*PHAS* loci (**Figure 4.8B**).

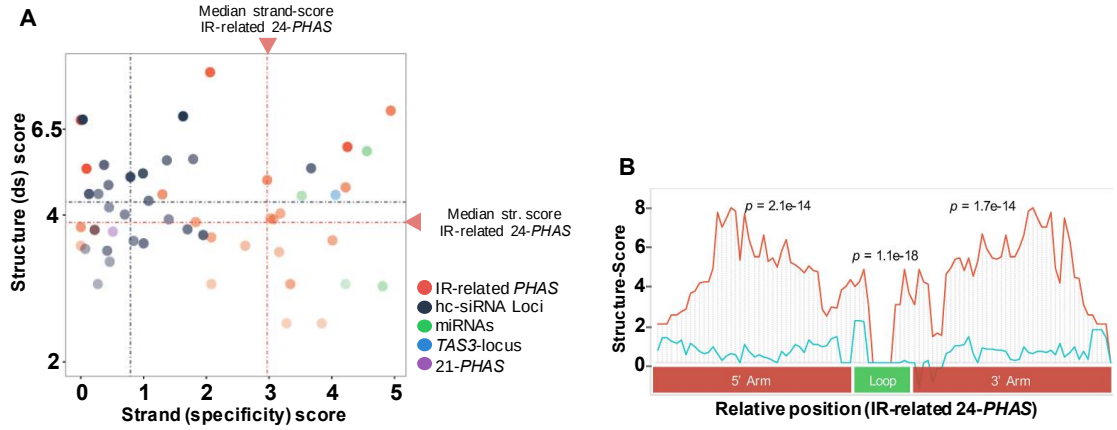


Figure 4.8: **Intra-molecular secondary structure at IR-related 24-nt *PHAS* loci in Asparagus**. (A) Scatter plots showing secondary structure scores as function of strand-specificity scores for IR-related loci along with randomly selected hc-siRNA, miRNA, tasiRNA and IR-related 21-nt loci that passed the coverage cutoff. Dotted line represents score medians, red for 24-nt *PHAS* loci and blue for hc-siRNAs. (B) Consensus of dsRNA structure scores (red) from five IR-based *PHAS* loci show two statistically significant peaks of paired nucleotides and a "valley" (loop, in green) of unpaired nucleotides validating formation of stem-loop structure from these IR-related *PHAS* transcripts. The five loci for this figure were selected based on high coverage and similar lengths and loop sizes. The control (blue line) represents the mean score from shuffled controls.

4.2.5 The 24-nt phasiRNA pathway exists more broadly in monocots

We sought to further trace the evolutionary origins of these reproductive-specific phasiRNA pathways, and thus we extended our study to two more species – daylily (*H. lilioasphodelus*) and *Lilium* (*L. maculatum*). These monocots were selected based on their availability, ease of dissection of anthers, and most importantly, due to their evolutionary distance from *Asparagus*. Daylily, another member of the *Asparagales*, diverged ~66 mya from the MRCA of *Asparagus* (Hedges et al. 2015), while *Lilium* diverged ~117 mya from the MRCA of the *Asparagales* (Hedges et al.

2015). We first established a size-stage correlation for developing anthers of both species (see methods and **Figure 4.9**), and then performed large-scale transcriptome sequencing of staged anthers and leaf samples, to compensate for the absence of a genome for both species. The deep, paired-end libraries from *Lilium* generated a total of 404 million reads with an average of 130 to 140 million reads for pre-meiotic and meiotic stages; for the same developmental stages in daylily, a combination of single- and paired-end Illumina RNA-seq plus SMRT sequencing generated a total of 140 and 125 million reads, and 324,879 full-length transcripts respectively (**Table 4.2**). The assemblies for both were generated using Trinity (Haas et al. 2013), including a hybrid assembly for daylily which included full length transcripts. This yielded 157,913 and 182,225 transcripts with normalized expression greater than 1TPM in at least one library, for *Lilium* and daylily respectively. Assemblies for both species displayed a significant Ex90N50 statistic (**Figure 4.10**) 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). To identify phasiRNAs, sRNA data were generated from the same stages (**Table 4.2**).

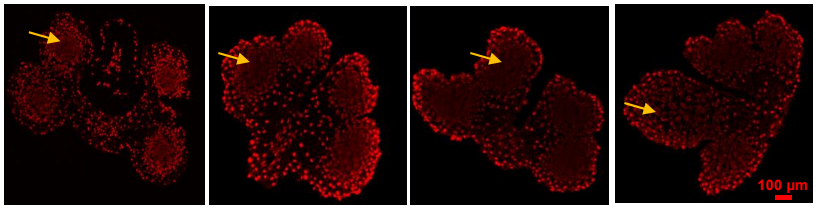
Table 4.5: **Summary statistics of daylily and *Lilium* sequencing libraries used in this analysis**. [a]Genome-matched reads not available due to absence of sequenced genomes for *Lilium* and daylily. [b]Number of read pairs listed for paired-end data. [c]Full-length non chimeric isoforms. [c]Polished (corrected) high-quality consensus transcripts.

Part A. Daylily small RNA data				
Code	**Title**	**Total Sequences**	**Len. of Reads**	**Technology**
Daylily_leaf_	Daylily leaf rep1, BM14-01	39,246,683	34	single-end
Daylily_wbud	Daylily, whole bud, no sepals, with <1.0 mm anthers, BM14-180	37,627,819	34	single-end
Daylily_01mm	Daylily, <1.0 mm anthers, BM14-177	52,380,062	34	single-end
Daylily_02mm	Daylily, 1.5-2.0 mm anthers, BM14-178	30,544,192	34	single-end
Daylily_03mm	Daylily, 2.5-3.0 mm anthers, BM14-179	41,155,375	34	single-end
Daylily_ant_0	Daylily 3.3mm anther stage 0 rep1, BM14-04	48,546,267	34	single-end
Daylily_ant_1	Daylily 4.1mm anther stage 1 rep1, BM14-08	45,773,743	34	single-end
Daylily_ant_2	Daylily 5.8mm anther stage 2 rep1, BM14-12	32,982,067	34	single-end
Daylily_fem_	Daylily female stage 0 rep1, BM14-06	45,534,573	34	single-end
Daylily_fem_	Daylily female stage 1 rep1, BM14-10	50,177,784	34	single-end
Daylily_fem_	Daylily female stage 2 rep1, BM14-14	35,155,630	34	single-end
Daylily_fem_	Daylily female stage 3 rep1, BM14-18	68,247,257	34	single-end
Daylily_fem_	Daylily female stage 4 rep1, BM14-22	55,651,517	34	single-end
Part B. Lilium small RNA data				
Lilium_leaf	Lilium maculatum leaf, BM14-190	27,227,974	34	single-end
Lilium_4mm_	Lilium maculatum L-4, 4 mm anthers, BM14-191	35,477,763	34	single-end
Lilium_5mm_	Lilium maculatum L-5, 5 mm anthers, BM14-192	32,039,303	34	single-end
Lilium_6mm_	Lilium maculatum L-6, 6 mm anthers, BM14-193	32,024,530	34	single-end
Lilium_8mm_	Lilium maculatum L-8, 8 mm anthers, BM14-195	32,594,126	34	single-end
Lilium_10mm	Lilium maculatum L-10, 10 mm anthers, BM14-197	33,533,618	34	single-end
Part C. Daylily RNA-seq data				
Daylily_leafr	Daylily leaf rep1, BM14-01	42,572,707	100	single-end
Daylily_01mm	Daylily, <1.0 mm anthers, BM14-177	29,999,859	100	single-end
Daylily_02mm	Daylily, 1.5-2.0 mm anthers, BM14-178	27,877,362	100	single-end
Daylily_03mm	Daylily, 2.5-3.0 mm anthers, BM14-179	37,901,322	100	single-end
Daylily_wbud	Daylily, whole bud, no sepals, with <1.0 mm anthers, BM14-180	38,943,956	100	single-end
Daylily_le	Daylily leaf rep1, BM14-01	33,520,975	150	paired-end
Daylily_01	Daylily, <1.0 mm anthers, BM14-177	33,705,644	150	paired-end
Daylily_02	Daylily, 1.5-2.0 mm anthers, BM14-178	29,568,897	150	paired-end
Daylily_03	Daylily, 2.5-3.0 mm anthers, BM14-179	30,750,435	150	paired-end
Daylily_wb	Daylily, whole bud, no sepals, with <1.0 mm anthers, BM14-180	31,514,569	150	paired-end
Part D. Lilium RNA-seq data				
Lilium_leaf	Lilium maculatum leaf, BM14-190	68,250,144	150	paired-end
Lilium_4mm_	Lilium maculatum L-4, 4 mm anthers, BM14-191	70,492,722	150	paired-end
Lilium_5mm_	Lilium maculatum L-5, 5 mm anthers, BM14-192	67,997,742	150	paired-end
Lilium_6mm_	Lilium maculatum L-6, 6 mm anthers, BM14-193	62,509,328	150	paired-end
Lilium_8mm_	Lilium maculatum L-8, 8 mm anthers, BM14-195	66,675,362	150	paired-end
Lilium_10mm	Lilium maculatum L-10, 10 mm anthers, BM14-197	68,302,075	150	paired-end

Part E. Asparagus PacBio "SMRT" data				
Code	**Title**	**Total Sequences**	**high-quality transcripts[d]**	**Technology**
DayLily-A-1K	Daylily, 0.5-1.0mm anther, <2kb insert length (135 SMRT cells)	232,523	35,542	SMRT-seq
DayLily-C-2-3	Daylily, 0.5-1.0mm anther, 2-3kb insert length (135 SMRT cells)	290,648	8,436	SMRT-seq
DayLily-B-3K	Daylily, 0.5-1.0mm anther, >3kb insert length (135 SMRT cells)	212,767	38,539	SMRT-seq

A

Size	4 mm	5 mm	6 mm	7 mm
Stage	Pre-meiosis	Pre-meiosis	Meiosis I	Meiosis II
Number of cell layers	NA	5	5	5
Tapetum	NA	Formed, intact	Intact	Start to thin out



B

Size	1 mm	2 mm	3 mm
Stage	Pre-meiosis	Post-meiosis (indicated by filled anthers)	Pollen developed
Number of cell layers	5	4	4
Tapetum	Formed	Thinning	Thinned out

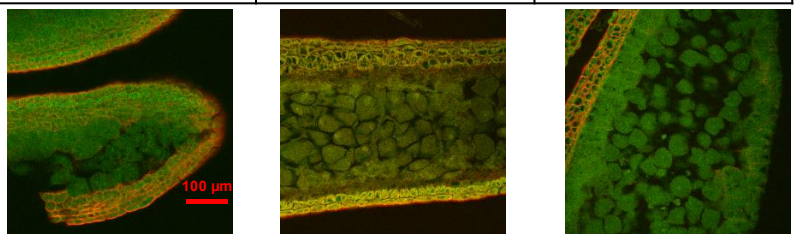


Figure 4.9: **Anther stage and size correlations capture pre-meiotic and meiotic anther stages for *Lilium* and daylily.** (A) Paraffin-embedded *Lilium* samples, cross-sectioned and stained with propidium iodide. Histology and cell divisions were examined for determination of the cell stages using confocal microscopy. Based on the morphology of archesporial cells (yellow arrows), 4 mm and 5 mm anthers corresponded to pre-meiotic stages. The 6 mm and 7 mm anthers were undergoing meiosis, and displayed a well-developed tapetum. (B) For daylily, anthers were treated with ScaleP clearing buffer for 1 week (see methods), and imaged using confocal microscopy. Histology and cell divisions in the longitudinal images of anthers were examined for determination of stages; the 1 mm anther was at a pre-meiotic stage, while 2 mm and 3 mm anthers were past meiosis and the tapetum was starting to thin out. Scale bars = 100 µm for all images.

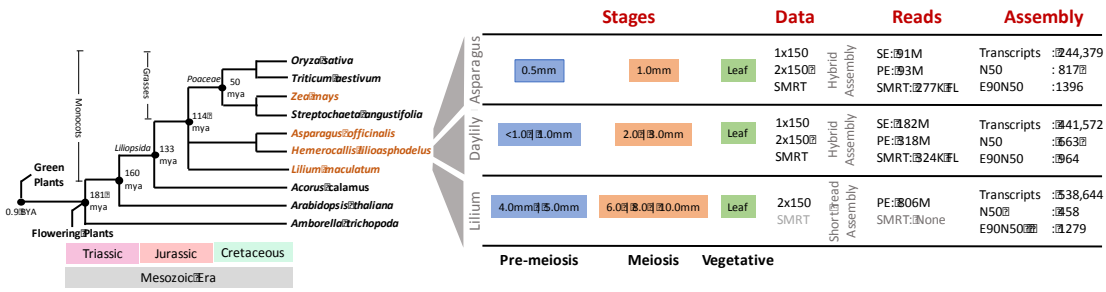


Figure 4.10: **Transcriptome and hybrid assemblies developed for Asparagus, *Lilium* and daylily.** Precisely-staged pre-meiotic, meiotic anther and leaf samples were used to generate transcriptome assemblies for *Lilium* and hybrid assemblies for Asparagus and daylily; a phylogeny of species is at left and data types and metrics at right. For single- and paired-end libraries, reads are represented in million(s), and for SMRT libraries processed full-length transcripts are represented in thousands. The E90N50 metric signifies the N50 statistic for transcripts in the 90th percentile of normalized expression. Phylogeny of plant species is for indicative purpose only and it is derived by comparing (median) divergence times from.timetree.org.

From the sRNA data, we identified triggers of reproductive phasiRNAs in daylily and *Lilium*. In both species, we found miR2118 members that were highly enriched in reproductive tissues (**Figure 4.11A**). We identified at least 19 miR2275 members in *Lilium*, and three miR2275 members in daylily, peaking in abundance at pre-meiotic and meiotic stages (**Figure 4.11a and Figure S9**). Both miR2118 and miR2275 were measurable in *Lilium* pistils. Neither family was characterized in earlier studies of the basal angiosperm *Amborella* (Albert et al. 2013) and the early diverged monocot *Zostera* (Olsen et al. 2016), so we reanalyzed the genomes and published sRNA data for these two species. By comparing mature miRNA sequences from maize and rice to find isomiRs of both families in *Zostera* and *Amborella*, we identified at least one candidate locus for both miRNA families. Based on presence of

isomiRs and presence of candidate loci, we concluded that miR2275 is likely found throughout the monocots; miR2118 has previously been shown to have ancient origins (Y. Zhang et al. 2016).

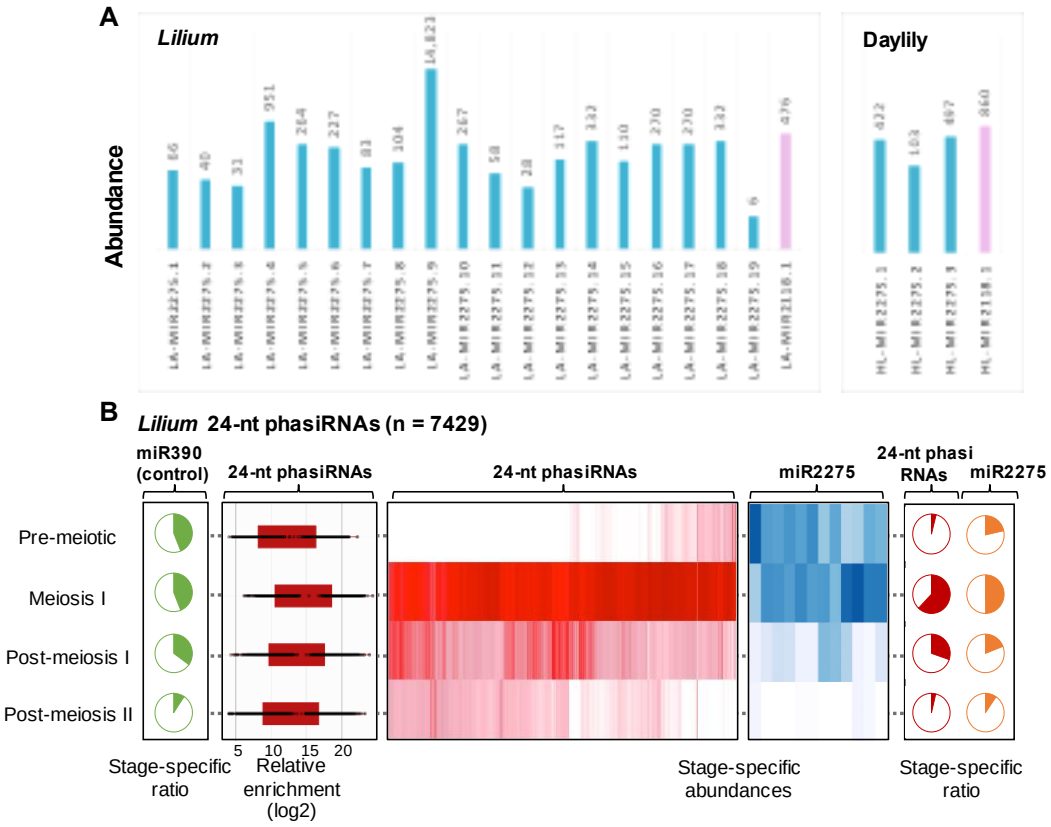


Figure 4.11: **Reproductive *PHAS* triggers and 24-nt phasiRNAs in *Lilium*.** (A) miR2118 (violet) and miR2275 (blue) family members identified in *Lilium* and daylily by comparing mature sRNA sequences to members in miRBASE (v.21); matches with total variance ≤ 4 were considered as valid candidates. Values on top of bars represent their total abundance (TP30M) in anthers. (B) Heat maps depicting abundance of *Lilium* 24-nt phasiRNAs (in red) and miR2275-triggers (in blue) in developing anthers. Both heat-maps are clustered on similarity of expression. Pie charts represent the proportion of stage-specific abundances for 24-nt phasiRNAs (in red), miR2275 (in orange) and miR390 (in green) the trigger of tasiRNAs across different anther developmental stages that are included in this study. Box-whisker plot shows enrichment (log2) of *Lilium* 24-nt phasiRNAs abundance from all *PHAS* loci in the meiotic anther compared to the vegetative sample (leaf).

What do miR2118 and miR2275 target in *Lilium* and daylily? We identified 6,277 and 392 24-*PHAS* transcripts, and 158 and six 21-*PHAS* transcripts in *Lilium* and daylily, respectively. The extraordinarily high number of 24-*PHAS* transcripts in *Lilium* matches the expanded, 18-member miR2275 family. The low numbers of 21-*PHAS* transcripts might reflect a sampling bias against pre-meiotic stages. Nonetheless, in *Lilium*, 21-nt phasiRNAs peaked in 5 mm anthers, and we infer that *Lilium* has an ~16-fold smaller 21-*PHAS* repertoire compared to the 24-*PHAS* loci - the opposite ratio relative to maize and rice (Zhai et al. 2015; Fei et al. 2016). As in the grasses, the precursors appeared to be mainly non-coding, and enriched in pre-meiotic and meiotic anther stages respectively (**Figure 4.11B and 4.12**), with the exception of eight 21-*PHAS* transcripts from *Lilium* in the post-meiotic stages (8 to 10 mm) with significant coding potential. We also identified 25 24-*PHAS* transcripts in developing daylily pistils that largely overlapped the anther 24-*PHAS* repertoire and were mostly low abundance.

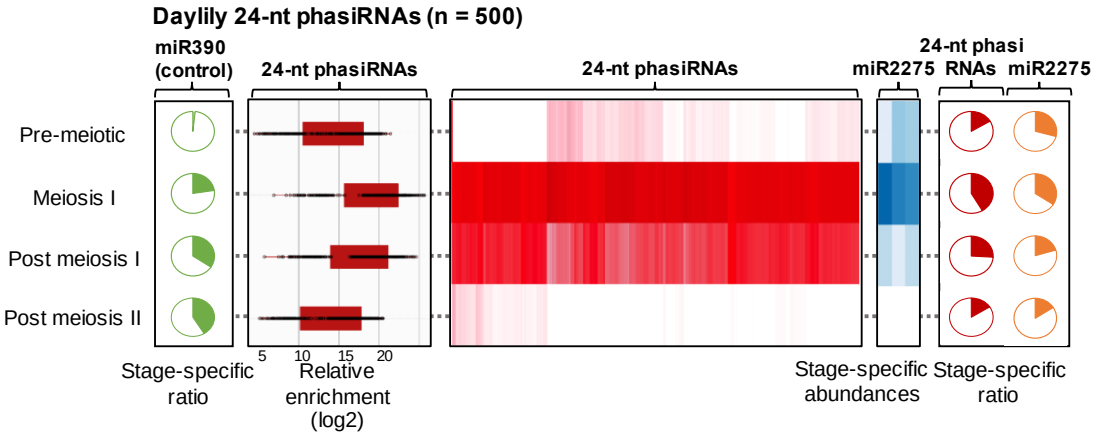


Figure 4.12: **Daylily 24-nt phasiRNAs and miR2275 are abundant in meiotic-stage anthers.** Heat maps depicting abundance of daylily 24-nt phasiRNAs (in red) and the miR2275 trigger family (in blue) in developing anther. Both heat maps are clustered on similarity of expression. Pie charts represent the proportion of stage-specific abundances for 24-nt phasiRNAs (in red), miR2275 (in orange) and and miR390 (the trigger of TAS3 tasiRNAs, in green) across different anther developmental stages that are included in this study. The box-whisker plot shows the enrichment (log2) of daylily 24-nt phasiRNA abundance from all *PHAS* loci in the meiotic anther compared to the vegetative sample (leaf).

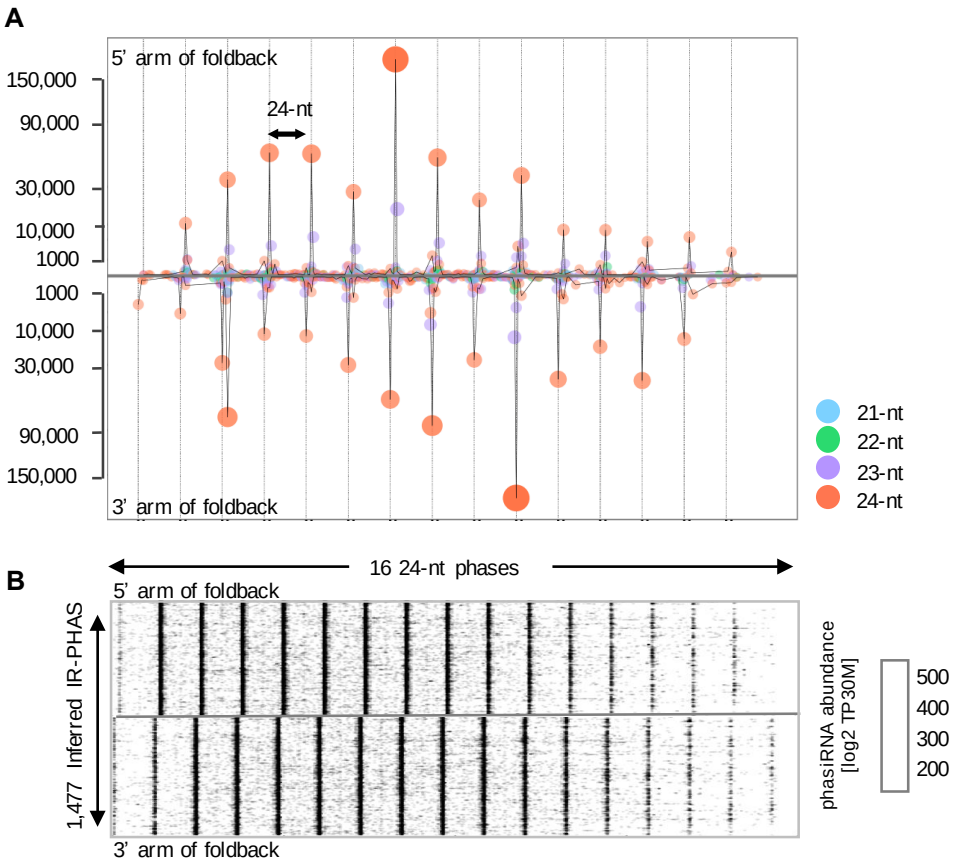


Figure 4.13: **Distribution of sRNAs in hairpin *PHAS* (hp-*PHAS*) and inferred inverted-repeat (IR-*PHAS*) precursor transcripts.** (A) Summed sRNA abundances from 5' and 3' arms of 50 hp-*PHAS* transcripts show a clear 24-nt phasing with a 2-nt overhang. Representative hp-*PHAS* with foldback score > 500, arm length > 384 (8 or more phases) were used to generate this distribution plot. (B) Scatterplot of sRNA abundances from 5' and 3' arm of inferred IR-related *PHAS*-transcripts (n= 1,477) show a strong 24-nt phasing pattern with a 2-nt overhang between the paired arms.

Knowing that *Asparagus* produces many 24-nt phasiRNAs from IRs, we next examined IR *PHAS* precursors in daylily and *Lilium*, looking particularly for the characteristic (of Dicer processing) 2-nt 3' overhang of phasiRNAs from different

arms of the IR. We found at least 131 24-*PHAS* transcripts with strong propensity to form a long foldback (≥ 271 bp) (referred to as foldback-*PHAS* hereafter) with high complementarity (≥ 99%), yielding sRNAs with a 2-nt 3' overhang (**Figure 4.13A and 4.14**). If the precursors are rapidly processed, full-length mRNAs may be rare in our data (like category III and IV RNAs, above), obfuscating the detection of intramolecular secondary structures. To account for this possibility, we computationally inferred pairs of *PHAS* transcripts that forms a stem-loop structure when docked in correct order, and exhibit a 2-nt overhang of overlapped phasiRNAs from different arms (see methods). A total of 2,888 (46.6%) and 87 (22.2%) of 24-*PHAS* transcripts from *Lilium* and daylily, respectively, matched these criteria i.e. displayed characteristic matching IR-type *PHAS* from *Asparagus* (**Figure 4.13B**). Collapsing the entire set of precursors based on sequence similarity and degree of overlap, and comparing the final tally to maize, *Lilium* displays a substantially larger (>25-fold) set of meiotic 24-*PHAS* precursors (n=3,394), with the *Lilium* 24-nt phasiRNAs accounting for more than 55% of total 24-nt sRNAs at their peak, in developing anthers. We again used sRNA fluorescent *in situ* hybridization to examine their spatial distribution in anther cells. Both types of precursors (IR-type and foldback) and their 24-nt phasiRNAs localized in tapetal and archesporial cells, with 24-nt phasiRNAs enriched at meiosis (**Figure 4.15**).

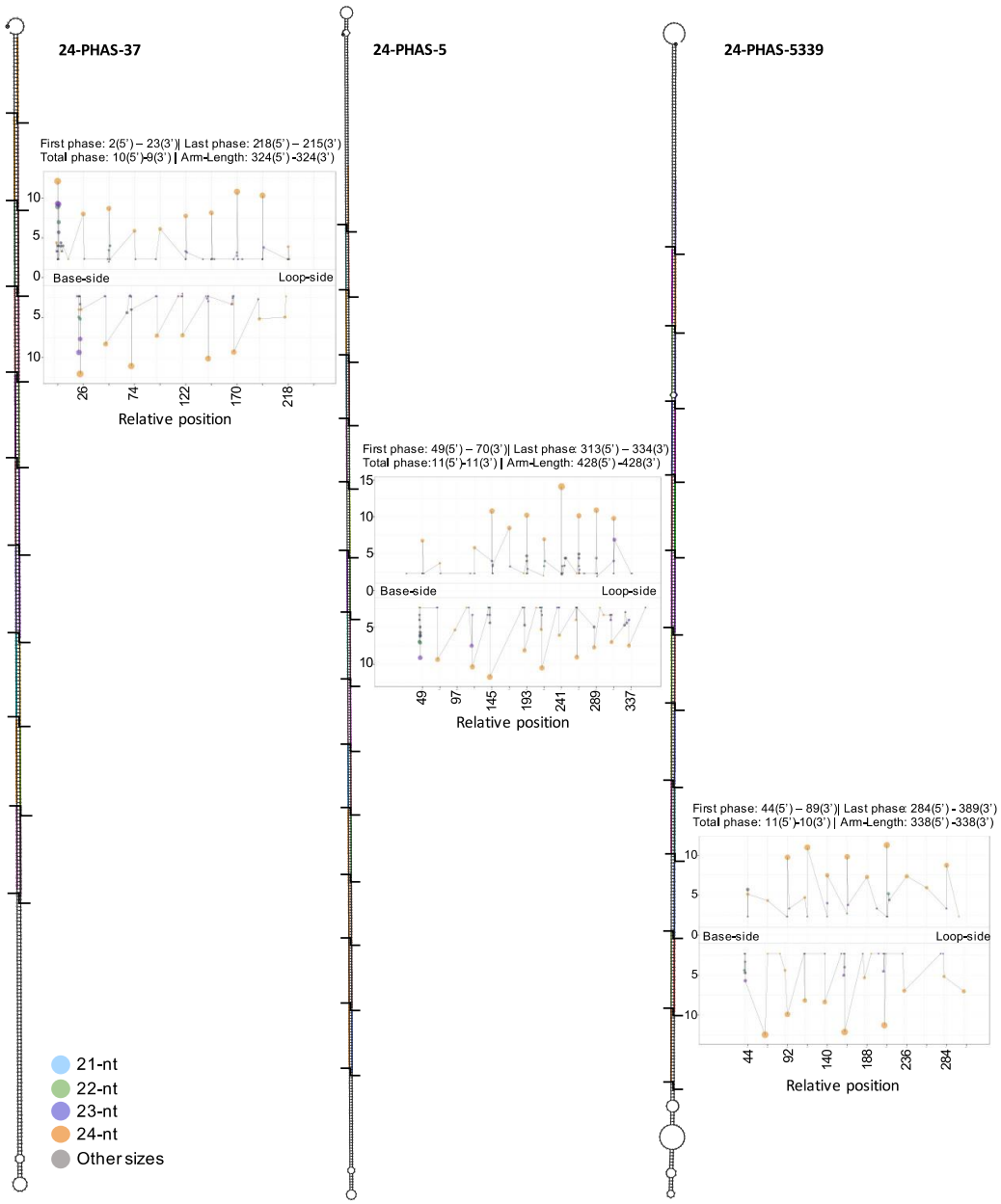

24-PHAS-37
First phase: 2(5') – 23(3')| Last phase: 218(5') – 215(3')
Total phase: 10(5')-9(3') | Arm-Length: 324(5')-324(3')
Base-side
Loop-side
Relative position
24-PHAS-5
First phase: 49(5') – 70(3')| Last phase: 313(5') – 334(3')
Total phase:11(5')-11(3') | Arm-Length: 428(5')-428(3')
Base-side
Loop-side
Relative position
24-PHAS-5339
First phase: 44(5') – 89(3')| Last phase: 284(5') - 389(3')
Total phase: 11(5')-10(3') | Arm-Length: 338(5')-338(3')
Base-side
Loop-side
Relative position
21-nt
22-nt
23-nt
24-nt
Other sizes

Figure 4.14: **Secondary structure and sRNAs for three representative hairpin (hp-) *PHAS* precursors from *Lilium*.** Precursors display consistent production of 24-nt long siRNAs from both arms, at 24-nt intervals, a processive signature of DCL5 activity. Scatter-plot depicts sRNA distribution on *PHAS* precursor transcripts, starting from the first detected 24-nt phasiRNAs. The abundance, on Y-axis, is shown in log2 scale. Position of first and last phasiRNAs for 5'- and 3'-arm along with the total phases and arm lengths are described in header of each scatter plot. The colors and size, in scatter plot, represent sRNA size class and abundance respectively.

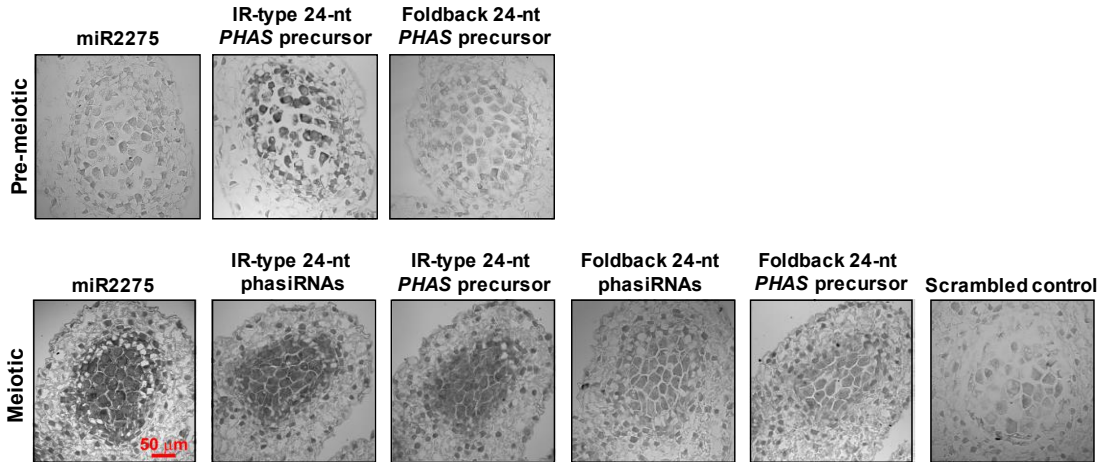


Figure 4.15: **Localization of 24-nt phasiRNA components in premeiotic (~4mm) and meiotic (~5 mm) anthers of *Lilium*.** Small RNA in situ hybridizations in pre-meiotic and meiotic anthers of *Lilium*, using probes for miR2275, meiotic phasiRNAs from IR locus 24-*PHAS*-5505 and hp-*PHAS*-5843. These phasiRNAs were not detected in pre-meiotic stages. Meiotic anthers were used for these in situ hybridizations. 24-nt phased siRNAs were not detected at pre-meiotic stage.

An investigation of the *PHAS* precursors from daylily and *Lilium* provided insights into the processing and miRNA triggers. For these precursors with no intrinsic foldback potential, we predicted miR2118 and miR2275 target sites using *revFerno* (component of *PHASIS* suite, see methods), and then validated the triggers using

sPARTA (Kakrana et al. 2014). Through this analysis, we could identify triggers for 1,098 (32.3%) and 29 (11.8%) of 24-*PHAS* precursors in *Lilium* and daylily, respectively, and for 22 (23.9%) of 21-*PHAS* precursors in *Lilium*. The miRNA triggers for seven of the 22 *Lilium* 21-*PHAS* transcripts were, unexpectedly, miR2275, reminiscent of the unusual miR2275-triggered 22-*PHAS* transcripts in *Asparagus*, all of which (from both species) peaked in abundance at pre-meiotic stages. In these 21-*PHAS* precursors, miR2275 target site occurred between 56 and 522 nt inside of the transcript 5' end, indicating that at least a few of these were captured unprocessed and likely lack the anticipated miR2118 target site. The predicted miR2275 cleavage sites were 'in phase' with the first phasiRNA, and had reasonable complementarity (score of 1.5 to 4.5). Perhaps the DCL4 vs DCL5 specificity of 21- vs 24-nt phasiRNA processing depends on distinct spatiotemporal boundaries of precursor expression, and not on recruitment by miR2118 or miR2275.

We examined IR 24-*PHAS* precursors with 5' miR2275 target sites; these were puzzling as one role of miR2275 targeting is to recruit RDR6, unnecessary for IR processing, and another role is to mark the phasing of sRNAs. As in *Asparagus*, a large number of such transcripts in both *Lilium* and daylily (530 or 35.8% of IR-type and 32 or 27.5% of foldback type in *Lilium*; and 9 or 18.7% in daylily), had miR2275 trigger sites, typically located +45 to -18 nt from the ssRNA-dsRNA junction. We manually examined a subset of IR precursors (*einverted* score $\geq$ 600), these lacked sRNAs upstream of the miR2275 target site, consistent with the functionality of the site. Precursors with the target site on the stem showed sRNA variability (18- to 23-nt) of the first phasiRNA cycle, especially on the 3' arm, lacking dominant 24-nt phasiRNAs (**Figure 4.16A, B and 4.17**). These siRNAs were largely absent from

precursors with miR2275 sites overlapping the dsRNA-ssRNA junction (**Figure 4.16C**), as were siRNAs upstream of the trigger site on the paired 5’ and 3’ arms. These observations suggest that 5’ and 3’ unpaired (ssRNA) ends are removed, either together – perhaps in absence of miRNA trigger like pri-miRNA processing – or sequentially, with the 5’ arm first removed via miR2275-directed cleavage, followed by trimming of the 3’ arm by an unknown mechanism, consequently releasing the stem-loop structure for subsequent processing by DCL5. Perhaps trimming could occur as with metazoan miRNA precursors, i.e. recognition of the ssRNA-dsRNA junction (J. Han et al. 2006; Kim, Han, and Siomi 2009), or as with piRNAs, in which a 3'-5' exonuclease trims excessive nucleotides (Izumi et al. 2016; Tang et al. 2016).

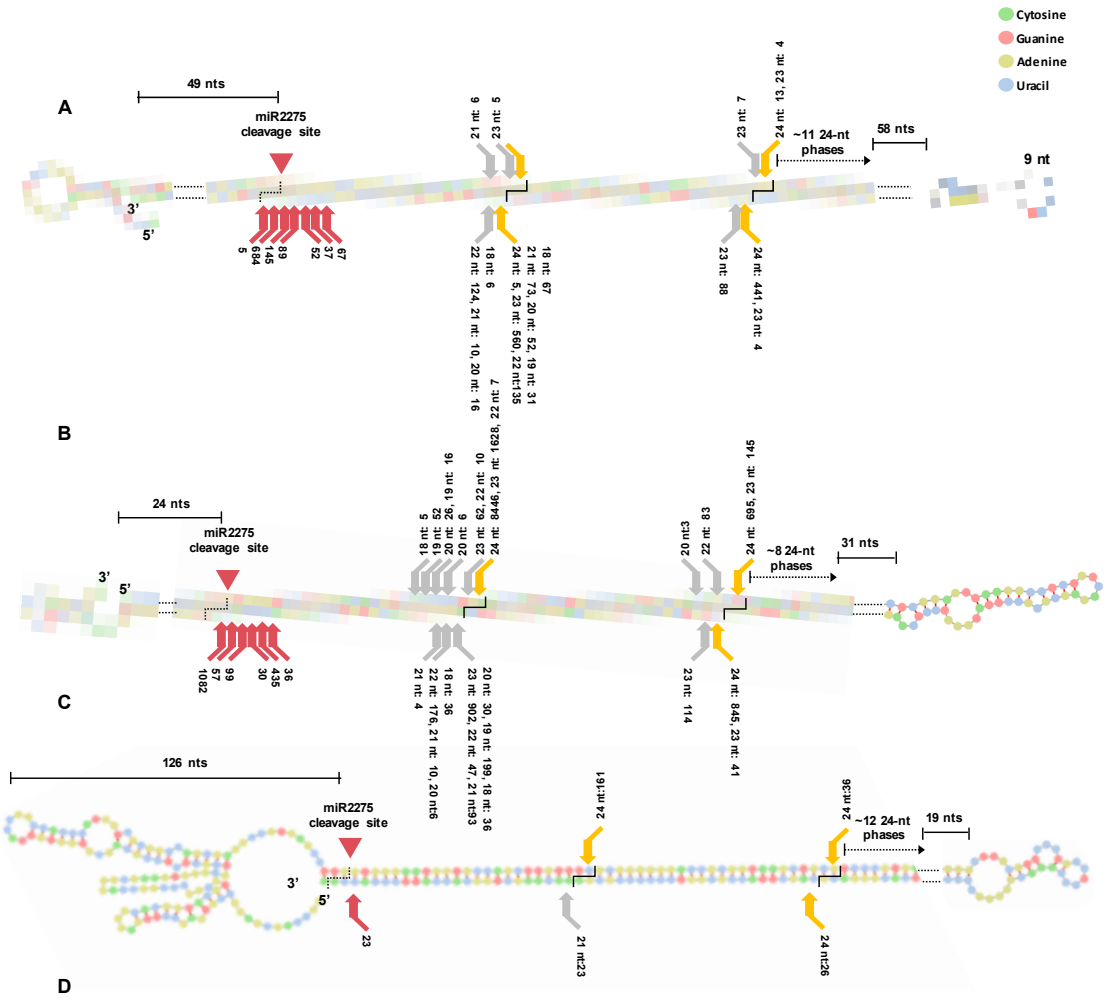


Figure 4.16: **Processing of miR2275-triggered hairpin *PHAS* precursors in *Lilium*. (A,B)** Foldbacks of two representative miR2275 triggered hp-*PHAS* precursor transcripts in *Lilium*, 24-*PHAS*-5 and 24-*PHAS*-1681. (C) Precursor for hp-*PHAS*-2398 with no unpaired 3'-arm. (D) Precursor for hp-*PHAS*-4395 putatively processed from loop-to-base. The cuts leading to release of 24-nt phased siRNAs are shown as orange arrows while those that generate siRNAs of other sizes are indicated as grey arrows. Counts represents cut frequencies computed from sRNA data. Red arrows indicate 5-termini of sRNAs of different sizes at non-triggered end along with their prevalence as indicated by sRNA data. In (A) and (B) the miR2275 cleavage site is 49 and 24 nucleotides inside the dsRNA region, while in (C) the cleavage site is 126 nucleotides from the 5'-terminus of the precursor.

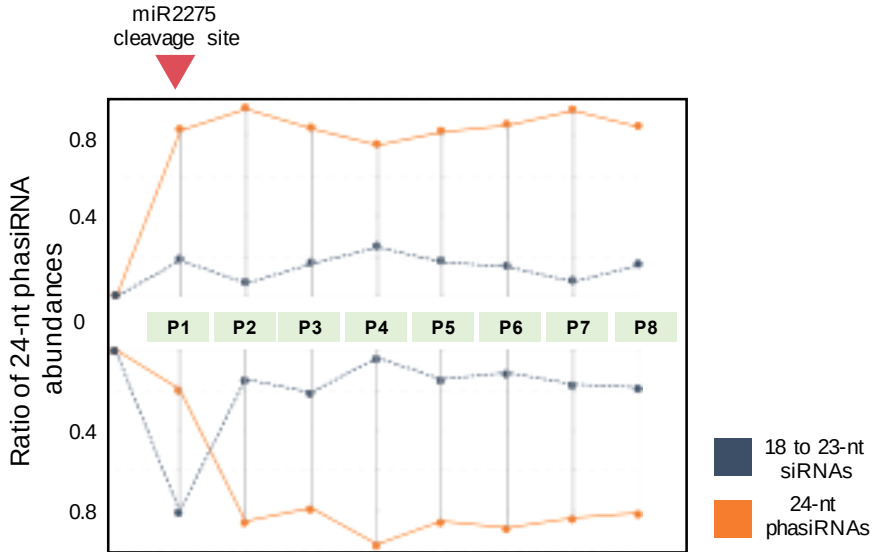


Figure 4.17: **Ratio of 24-nt phasiRNAs abundances in triggered foldback *PHAS* precursor transcripts.** Phased siRNAs (24-nt) (orange) and other small RNA size classes (grey) in miR2275 triggered foldback *PHAS* precursor transcripts. P1 to P8 represents first eight phasiRNA sites on precursor. Foldback precursors with miR2275 trigger site predicted precisely at P1, i.e. phase index = 0 (n=18), were used as a representative set. P1 is critical in this analysis, and any precursor with trigger site predicted 1 or 2 (24-nt) phases to left or right of P1 is most likely missing the first phase cycle, and therefore cannot be used in this particular analysis..

DCL1 functions to cut unpaired ends of miRNA foldback precursors (Cuperus, Fahlgren, and Carrington 2011; Bologna et al. 2013), and with this in mind, we analyzed foldback *PHAS* precursors for their inferred direction of processing. Based on phasiRNA abundances, we identified a representative set (n=10) displaying processing signatures consistent with loop-to-base processing (see supplemental methods). These precursors lacked miR2275 target site and showed no major raggedness in processing of the first loop-side phasiRNA (**Figure 4.16D**), compared to foldback-precursors likely processed base-to-loop from a miR2275 site ("miR2275-to-loop"), consistent with the idea that ragged processing could be due to inconsistent

trimming of unpaired ends. This leaves unaddressed the enzyme, which makes this first, loop-side cut in processing of these reproductive 24-*PHAS* foldback-precursors. It could be any one of the five monocot Dicers, with subsequent processing presumably handed off to the DCL5.

Which Dicer might make this first cut in an IR transcript? We used public Arabidopsis sRNA data (Lee et al. 2012; D.-H. Jeong et al. 2013; Shaofang Li et al. 2015), and focused on two known IR loci, *IR71* and *IR2039* (Henderson et al. 2006). A clear distribution of highly abundant sRNAs occurred at both loci, consistent with intramolecular folding and subsequent dicing. In wild-type, the IR-transcripts were mainly processed into 21-, 22- and 24-nt sRNA species (**Figure 4.18A**), products of DCL4, DCL2 and DCL3 respectively. In *dcl3* and *dcl2/3/4* mutants, 24-nt sRNAs were largely absent, with both backgrounds displaying slight accumulation of 21-nt sRNAs (**Figure 4.18B**), suggesting processing of these foldbacks by DCL4 and DCL1 in the absence of DCL3 and DCL2. The 24-nt sRNAs were not impacted in *nrpe1*, *nrpd1* and *rdr2* indicating their independence of the RdDM pathway (**Figure 4.18B**). *dcl1* showed a strong reduction (3- to 1400-fold) in levels of all sRNA size classes (**Figure 4.18B**), suggesting a primary role of DCL1 in facilitating the production of siRNAs from these IR loci. This reduced processing of fold-backs in *dcl1* could be related to its activity in cleaving the pri-miRNA stem-loops to release mature miRNA duplexes. These observations suggest hierarchical processing of at least some IR loci in Arabidopsis, first by DCL1 and subsequently by DCL2, DCL3 and DCL4. We conclude that DCL1 might similarly initiate some DCL5-processed IR transcripts that yield 24-nt reproductive phasiRNAs.

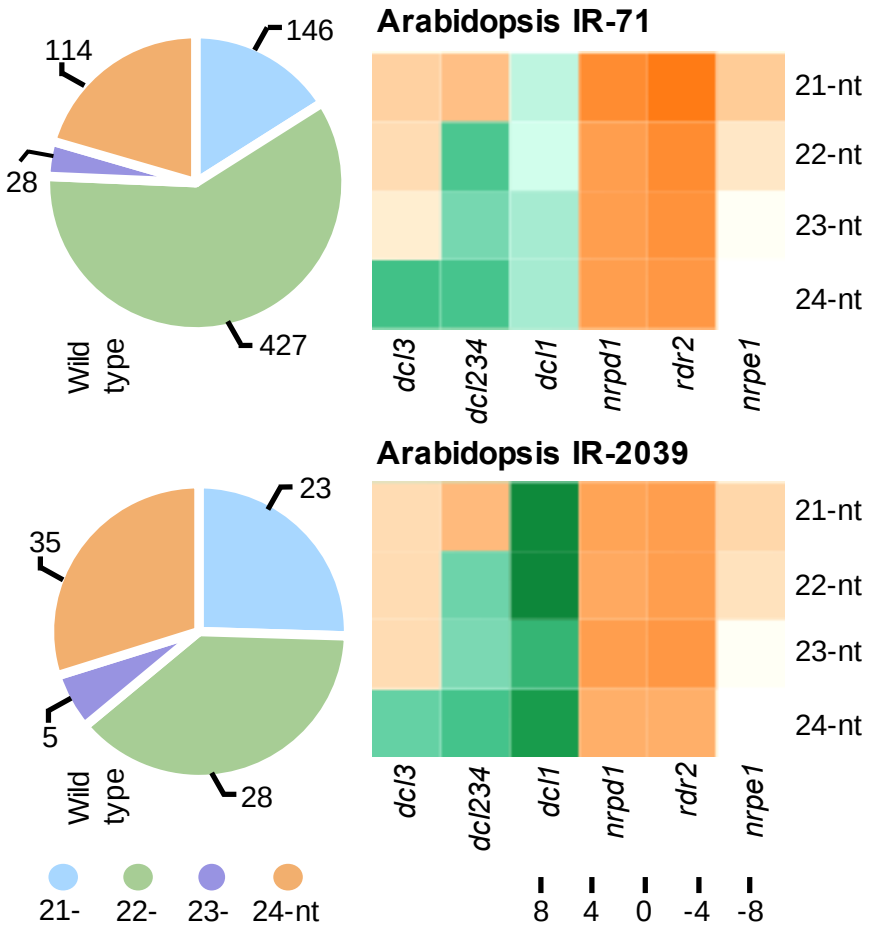
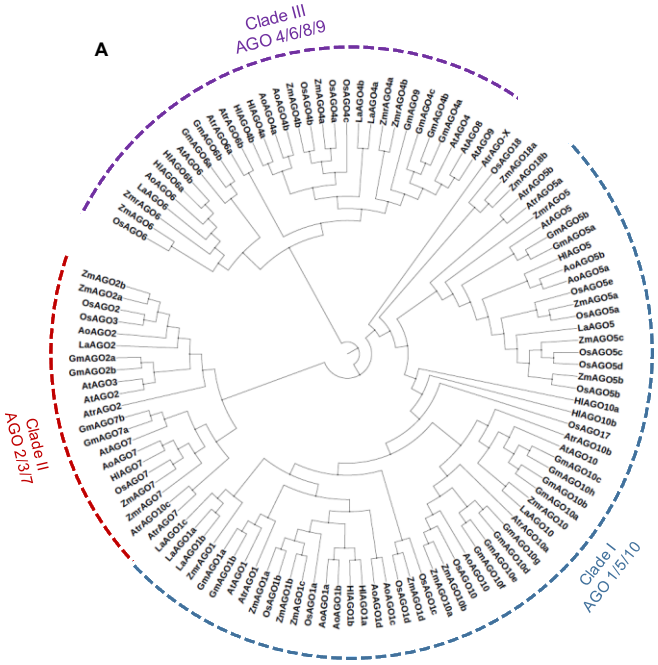


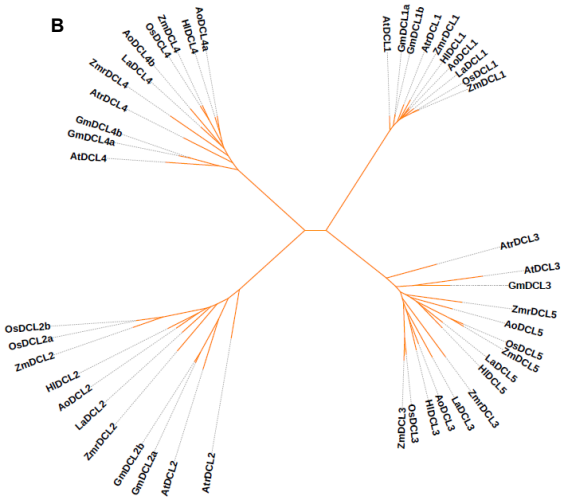
Figure 4.18: **Protein partners involved in processing of two endogenous inverted-repeats in Arabidopsis**. (A) Pie-charts represent sRNAs of 21- to 24-nt sizes derived from IR-71 and IR-2039 endogenous IR loci. Counts represents the normalized abundance in thousands. (B) Heat maps representing differential abundance of 21-, 22- 23- and 24-nt sRNAs in Arabidopsis dcl3, dcl2/3/4, dcl1, nrpd1, rdr6 and nrpe1 mutants against wild-type.

4.2.6 Protein-partners of the 24-nt phasiRNA pathway: grass AGO proteins are not entirely representative of monocots

We investigated the Argonaute family members, binding partners of small RNAs, identifying 12, nine and eight AGO proteins for *Asparagus*, daylily and *Lilium*

respectively (**Figure 4.19**). AGO candidates known for roles in reproductive phasiRNA functions are AGO5, a homologue of rice MEL1 which selectively recruits 21-nt phasiRNAs (Komiya et al. 2014), and AGO1d, AGO2b, and AGO18 for 24-nt phasiRNAs based on transcriptome profiling and spatial localization in maize and rice (Zhai et al. 2015; Fei et al. 2016). Among these, AGO5 members were present and consistently enriched in pre-meiotic or meiotic anthers of *Asparagus*, *Lilium* and daylily, matching earlier described reproductive enriched expression. AGO5 members are also present in the *Z. marina* and *A. trichopoda* genomes. In contrast, AGO18 was missing from the genome and transcriptome assemblies of all three species plus *Z. marina* and *A. trichopoda*, consistent with its possible emergence in grasses (H. Zhang et al. 2015). Strikingly, we found AGO4 which shares 24 nt spectra with AGO18 and predominantly recruits 1A-siRNAs (H. Wang et al. 2011; A. Mallory and Vaucheret 2010), which is the enriched class in 24-nt phasiRNAs, to be robustly expressed and enriched at pre-meiotic and meiotic anther compared to vegetative tissues (**Figure 4.20A**). Reproductive enriched AGO1 family members were also identified in study, consistent with reports of their potential association with 24-nt phasiRNAs (Fei et al. 2016), likely due to the diversity in 5'-nucleotide of 24-nt phasiRNAs, suggesting possibility of recruitment of multiple AGO members downstream of 24-nt phasiRNA production.


A
Clade III
AGO 4/6/8/9
Clade I
AGO 1/5/10
Clade II
AGO 2/3/7


B

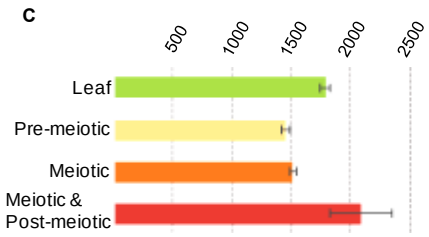

C
500
1000
1500
2000
2500
Leaf
Pre-meiotic
Meiotic
Meiotic &
Post-meiotic

Figure 4.19: **Dicer-like (DCL) and Argonaute gene family members in Asparagus, daylily and *Lilium*.** (A) Phylogenetic tree of AGO members from Asparagus (Ao), Daylily (Hl) and *Lilium* (La) identified in this study along with four representative species – Arabidopsis (At), rice (Os), maize (Zm) and soybean (Gm). AGO9 was renamed to AGO4 family because these are closely related in many plants. (B) DCL phylogeny with members from Asparagus (Ao), daylily (Hl) and *Lilium* (La) identified in this study along with four representative species – Arabidopsis (At), rice (Os), maize (Zm) and soybean (Gm). (C) Bar plots representing the relative expression of DCL5 in Asparagus pre-meiotic & meiotic anthers, and leaves, as measured by quantitative, real-time PCR.

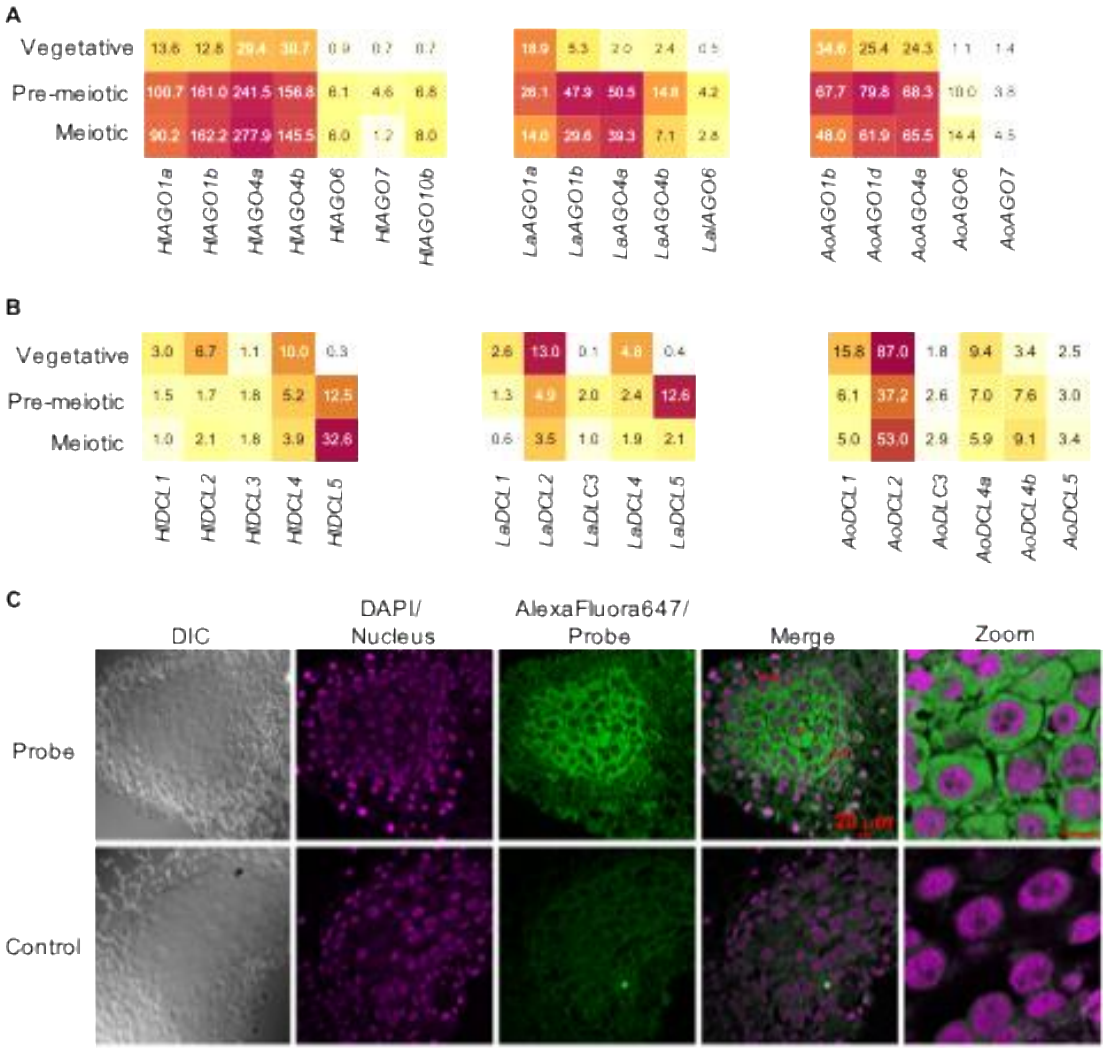


Figure 4.20: **Dicer-like (DCL) gene family and expression in Asparagus, daylily and *Lilium*.** (A) Heat-map representing expression profile of Asparagus, daylily and *Lilium* AGO members. Phylogeny of AGO members is provided in Figure S13. (B) Heat map of DCL abundances for three monocots, that were reliably detected (>1 FPKM) in one of three anther stages or the vegetative material. Phylogeny of DCL members is provided in **Figure 4.19**. (C) FISH localizing DCL5 transcripts in the cytoplasmic area of the tapetum and archesporial cells in meiotic-stage anthers from *Lilium*. AF647 (green) indicates the DCL5 mRNA localization. DAPI (pink) shows the stained nucleus. Scale bar = 20 μm for all images.

Finally, we examined *Dicer-like* (*DCL*) genes. We identified *DCL4* and *DCL5* in all three species, with an additional *DCL4* in *Asparagus* (**Figure 4.19B**). The latter, showed higher abundance in anthers compared to leaves but lacked a 5'-helicase domain. *DCL5* abundance patterns varied among the three species: daylily *DCL5* displayed reproductive-specific expression peaking at meiosis; *Lilium DCL5* abundance peaked at a pre-meiotic stage, although measured with limited meiotic stages, and via FISH, co-localized with the 24-*PHAS* precursors and phasiRNAs in tapetal and archesporial cells (**Figure 4.15 and 4.20C**), matching the patterns described in maize (Zhai et al. 2015). We could not detect *Asparagus DCL5* using *in situ* localization experiments, and it was absent from the high-quality full-length transcripts from SMRT-sequencing, presumably due to its low expression levels in *Asparagus* (**Figure 4.20B**). Given this low abundance, we validated *DCL5* presence in *Asparagus* using quantitative RT-PCR (**Figure 4.19C**). Surprisingly, its expression in *Asparagus*, as indicated by RNA-seq. and RT-qPCR, was not restricted to anthers, and it accumulated to similar levels in the vegetative (leaf) samples as in reproductive tissues.

4.3 Chapter summary

In this chapter (and with collaborators), I

- generated extensive sRNA data (n=38 sRNA libraries) for asparagus, a non-grass monocot whose genome has been recently sequenced, corresponding to 12 different tissues
- identified 105 miRNAs from 78 distinct families, and performed a comparison with miRNA repertoire from all other monocots and eudicots species to identify the loss, gain or expansion of miRNA families in asparagus, which diverged 114 and 160 MYA from most well characterized monocots and dicots, respectively

- discovered vegetative- and reproductive-enriched copies for 21-*PHAS* trigger miR2118 and reproductive-specific copies of 24-*PHAS* trigger miR2275 members

- discovered 21-nt phased siRNA generating loci in male-germline tissue and 24-nt phasiRNAs in both male and female-germline tissues, using *PHASIS* suite

- probed the secondary structure of phasiRNA generating loci along with miRNA and hc-siRNA loci as control through double-stranded RNA seq

- established stage-size correlation for anthers from *Lilium* and daylily, two new species based on the morphology of archesporial (AR) and tapetal cells of anthers

- generated deep short-read RNA-seq and single molecule real time seq data from sequential cohorts of staged anthers plus dayily pistil, and produced 'hybrid' assembly for both to compensate for absence of sequenced genome

- discovered 21- and 24-*PHAS* trigger miRNA families, precursors and phasiRNAs from *Lilium* and daylily and established their temporal expression patterns across developing anther

- identified the inverted-repeat related population of 24-*PHAS* precursors and inferred their proportion by considering the fact that a portion of precursors are processed, therefore cannot be assembled as complete transcript

- probed the spatial localization of IR-related phasiRNAs, miR2275 using *in-situ* hybridizations

- studied the processing mechanism of these IR-related precursors, because presence of trigger site and the foldback tendency is redundant to the processing of precursors to generate phasiRNAs

- probed sRNA data from loss-of-mutant lines of key protein factor involved in secondary siRNAs and hc-siRNA pathways to study their roles in processing of inverted-repeat transcripts

- identified key protein partners of phasiRNA pathways – DCL, RDR and AGO family members in Asparagus, *Lilium* and daylily

- generated phylogenies of these newly identified members with other well-studied monocots species, and established the temporal expression patterns in all three species.

- probed the DCL5 expression levels and spatial localization using qRT-PCR and fluorescent in situ hybridizations, respectively.

I observed that

- miR2118 family members display a dichotomy in tissue-specificity and temporal expression, correlated to loci of origin, and triggers both *pNLs* (n=12) and pre-meiotic *PHAS* (n=3) precursors in Asparagus.

- miR2274 likely targets, a few, "non-phased" genic and intergenic transcripts, unlike grasses where it triggers meiotic phasiRNAs

- 24-*PHAS* loci has statistically significant overlap with inverted-repeats (IRs), >64 fold compared to random chance and the size-distribution of sRNAs at *PHAS* loci matches to those of canonical *PHAS* loci reported in maize and rice.

- 24-nt phased siRNAs in Asparagus have spatio- and temporal-accumulation pattern as canonical phasiRNAs

- 24-*PHAS* loci display statistically significant base-pairing; and the comparison of 24-*PHAS* with het-siRNA loci, shows distinct differences; with *PHAS* loci displaying intramolecular secondary structure while latter shows intermolecular secondary structures

- 24-*PHAS* loci identified in female germline tissues has a complete overlap with those identified in male germline i.e. no loci specific to female germline identified in Asparagus

- *Lilium* has a 5x and 25x expansion in miR2275 members and *PHAS* repertoire, respectively

- the proportion of 24-*PHAS* precursors forming a foldback is very little compared to the complete repertoire

- both type of precursors (fold-back and inferred IR-type), their corresponding phasiRNAs and miR2275 localizes in tapetal and AR cells of *Lilium* anther, matching to observation in Asparagus and grasses

- miR2275 triggered foldback precursors, which are unique to *Lilium* and Maize, show sRNA variability (18- to 23-nt) in first phasiRNA cycle, especially on the 3' arm, lacking dominant 24-nt phasiRNAs

- DCL1 loss-of-function mutant, *dcl1* in *A. thaliana* Col0 background, show a strong reduction (3- to 1400-fold) in levels of all sRNA size classes generated from two well-characterized IRs

- protein partners of pre-meiotic phasiRNA pathway – DCL4, RDR6 and AGO5 are well conserved across all species, matching canonical pathway in grasses

- protein partners of mei-phasiRNA pathway – DCL5 and RDR6 are conserved, however, AGO18 is missing from all species included in this study, plus Z. Marina (another monocot that diverged even earlier)

- DCL5 is co-localizes with IR-related phased siRNAs components in Asparagus

- AGO1 and AGO4 members, likely homologs of rice AGO1d and maize AGO4d, were consistently enriched in meiotic anther of Asparagus, *Lilium* and Daylily

From this, I conclude that

- contrary to earlier beliefs, both pre-meiotic and meiotic phasiRNA pathways are widely prevalent beyond grasses, which pushes their origins into the monocots

- presence of meiotic phasiRNA pathways in pistils, the female-reproductive organ, identified in this study suggests wider roles for these pathways, which might not be restricted to male-germline

- elaborate pathogen defense regulatory network of NBS-LRR genes prevailed in earlier monocots unlike the grasses, displaying an increasingly specialized role of miR2118 in *Asparagales*

- specialized triggers for both pre-meiotic and meiotic phasiRNA pathways emerged before the divergence of *Lilium* and Asparagus from MRCA of Maize, respectively

- factors for pre-meiotic phasiRNA pathways are well conserved across all the species included in this study, supporting their continuing functional association across monocots

- characterization of maize AGO4d or its homologs, which might serve as effector molecule for meiotic phasiRNA in non-grass monocots, holds significant potential to provide crucial insights into the functions of meiotic phasiRNAs

- absence of IR-related phasiRNA loci in other species, the parallels in spatial localization and temporal dynamics are not sufficient to establish these as functional analogues to meiotic phasiRNAs

- discovery of phasiRNA pathways, in their canonical forms, beyond the grasses along with the presence of meiotic phasiRNAs in female reproductive tissues suggests a broader role of pre-meiotic and meiotic phasiRNA pathways in germline development

- phasiRNA pathways, especially mei-phasiRNAs, are not dependent upon grass-specific components, therefore, their existence even beyond the monocots.

Chapter 5

DISCUSSION AND CONCLUSION

High-throughput sequencing (HTS) has revolutionized the field of life sciences research. Multiple variations of HTS exists today, and these are being continually upgraded to improve the resolution, reduce cost and to expand the applications. The capacity to sequence hundreds to millions of miRNAs and siRNAs in a single library has substantially advanced the field of small RNA biology. Ever-decreasing sequencing costs further expanded the scale of experiments. Since each sequencing library that we generate can be reused to investigate a related, connected or an entirely different question, this massive and continual pile-up of sequencing data necessitate the development of i) new scalable algorithms that can efficiently mine large volume of data and ii) new approaches to leverage from expanding applications of sequencing technology. This tectonic shift towards data-based research has brought the inherently different fields of computational and wet-lab research together, making both inseparable and complementing each other. In this dissertation, I have aimed to maintain a balance between both computational and experimental aspects of plant research. The focal point of this dissertation is to investigate the origins and biogenesis of germline-associated class of phased secondary siRNA (phasiRNA) pathways. However, to research this topic, I required a new set of computational tools, as the available tools were insufficient to support my studies.

5.1 Advanced algorithm for miRNA target prediction and new software for leveraging experimental data for targets discovery

I began by addressing significant bioinformatic gaps in the discovery of miRNA targets. Methods that existed before I started my work in this field were developed for smaller genomes like *A. thaliana* (a model eudicot) and focused exclusively on annotated portion of the genome. Since annotations for most plant genomes are incomplete with target mRNAs remain to be found in intergenic, unannotated regions, these tools limited the scope of the study to a small and usually incomplete "annotated" portion of the genome. These existing tools also suffered from an inductive bias, which is introduced by pre-selecting the potential targets based on the degree of sequence complementarity and, especially in the seed regions of miRNAs, rather than giving more weight to the experimental data for finding miRNA targets. Although, such approaches reduce the search space and accelerate the analysis, but most likely miss the actual targets that have comparatively lower sequence based complementarity with miRNAs. Another flaw that existed in the scoring system of these tools, an extension of issue related to sequence-based complementarity scores is that targets for large miRNAs families that range from tens to hundreds will get low confidence scores due to the presence of large number of binding sites in an (annotated) genome. In addition to these shortcomings, the diminishing cost of sequencing had led to sequencing of more and more genomes =some even >100 times bigger than *A. thaliana*. These larger genomes make the process of target discovery even more challenging and if factor in the nature of miRNA-target interaction, which includes bulges and mismatches, the genome-level discovery of miRNA targets is a virtually impossible by tools that existed at the beginning of my dissertation.

I realized that these issues could only be addressed by developing a new tool, with a new scoring system, target scan algorithm, and capability to leverage raw computing power. So, I developed "*sPARTA*", a powerful tool for plant miRNA target prediction and PARE-based validation. *sPARTA* can also search for targets in unannotated genomic regions, which is useful to discover novel regulatory modules, completely independent of genome annotations. Earlier tools like *PAREsnip* use seed-region complementarity rules to accelerate the analysis, whereas *sPARTA* introduced a novel 'seed-free' mode is based on empirical observations regarding miRNA-target interactions, and it identifies targets with weak seed-region complementarities or mismatches at canonical positions. To fasten the analysis, *sPARTA* implements true parallelization reducing runtime from hours or days to minutes and seconds. These significantly reduced runtimes and *sPARTA's* capability to efficiently handle large genomes, with hundreds to thousands of sRNA/miRNAs and as many as PARE libraries facilitates the data-intensive, genome-level scans of miRNA targets without comprising on sensitivity.

sPARTA also forms the core of *comPARE*, a web resource that allows the discovery, visualization and in-depth exploration of genome-wide miRNA-target interactions in the heterogeneous yet highly integrative environment. *comPARE* was developed to serve as a repository of our validated miRNA interactions, collating small RNA and PARE datasets along with their genomic context. In essence, this study presents three novel tools: miRferno for target prediction, *sPARTA* for PARE-based target validation and *comPARE* for visualization, exploration and comparative analysis of miRNAs targets. These three tools empowered me to effectively exploit

advanced computing power, discover novel regulatory modules and deliver datasets containing high-quality, well-supported predictions of miRNA-target interactions.

5.2 New suite for discovery and in-depth characterization of phased siRNAs

Next, I developed a new tool set for discovery and in-depth characterization of phased siRNA loci (or genes). The reason I had to undertake this work was because integrated tools for discovery and computational characterization of phasiRNAs did not existed at the time I began my work. Existing options were not only limited in number and function but also incompatible or inefficient in handling large volume of small RNA-seq data. Importantly, these existing tools for the *de novo* identification of *PHAS* genes (or loci) required an assembled genome for their discovery and additional experimental data (PARE or degradome libraries) to further identify their miRNA triggers. These algorithmic limitations restricted the study of phasiRNA pathways to species that have assembled genome, therefore did not support my aim to discover and trace the origins of male germline pathways beyond the well-studied species like maize and rice. So, I decided to develop an advanced computational suite, which we call "*PHASIS*".

Loci generating phased siRNAs (21- and 24-nt) are widely prevalent across land plants (Allen et al. 2005; Shivaprasad et al. 2012; Johnson et al. 2009; Zhai et al. 2011; Arikit et al. 2014; R. Xia, Xu, et al. 2015; Fei et al. 2016), varying in numbers per genome from tens to thousands, displaying diverse spatial and temporal expression patterns, and participating in an array of different functions (Allen et al. 2005; Shivaprasad et al. 2012; Zhai et al. 2011, 2015; Dukowic-Schulze et al. 2016). Recently, piRNAs in *Drosophila* too were reported to be phased, generating 'trailer' piRNAs in 27-nt intervals after cleavage by secondary siRNA and Zucchini-dependent

processing of cleaved transcript (Mohn, Handler, and Brennecke 2015; B. W. Han et al. 2015). Given the wide prevalence of phasiRNAs and the rate of genome sequencing, it is likely that they will be better characterized and studied in the coming years.

The *PHASIS* suite that I developed provides an integrated solution for the large-scale survey of tens to hundreds of sRNA libraries for the following applications: a) *de novo* discovery of *PHAS* loci and precursor transcripts, b) a summarization of *PHAS* loci or precursor transcripts (referred to as *PHAS* loci hereafter) from specific groups of sRNA libraries, c) a comparison of *PHAS* summaries between groups corresponding to samples from different stages, tissues and treatments, d) quantification and annotations of *PHAS* loci, and e) discovery of their miRNA triggers. *PHASIS* generates easily parsed output files for downstream bioinformatics analysis, formatted result files for immediate consumption and organized ancillary data to facilitate optimizations like a re-summarization to exclude or include libraries. I benchmarked *PHASIS* on five different plant species and compared its performance with its direct competitor, and further compared *PHAS* predictions with a human curated set. The comparative benchmarking with other tool, showed that *PHASIS* is superior to existing alternative in accuracy, yield and speed of predictions. It captured more than 86% of manually-curated set of 21- and 24-*PHAS* loci in default mode, i.e. without providing any additional setting to increase or decrease the quality of prediction, as it would be used by non-experts.

PHASIS facilitates the discovery of phasiRNAs and their precursors, and the identification of their triggers by eliminating the requirement of a genome assembly and experimental PARE/degradome data. It offers flexibility to users to tailor analyses

for their own goals and it integrates an array for functions in one package. In essence, the *PHASIS* suite developed in this dissertation is the "first" suite for the in-depth characterization of phasiRNAs, their loci or precursors and discovery of triggers; it is sensitive, exceedingly scalable and exceptionally fast software. Together, *PHASIS* and *sPARTA* provided us vital tools to attempt or investigation of phasiRNA pathways.

5.3 Insights onto the evolution of phasiRNA pathways

Phased siRNA pathways, associated with the male-germline, were first described in rice (Johnson et al. 2009). Since their discovery, large scale sRNA profiling along with genetics and biochemical studies from rice and maize have elucidated their precise temporal expression, spatial niches and pathway-specific factors (Song, Wang, et al. 2012; Song, Li, et al. 2012; Komiya et al. 2014; Zhai et al. 2015; Fei, Xia, and Meyers 2013; Dukowic-Schulze et al. 2016). However, their evolutionary origins and penetration into other non-grass clades are yet to characterized, primarily because of their restricted presence in male reproductive tissues, narrow window of accumulation of and lack of information on the precise stages at which phasiRNAs peak or appear. In this investigation and together with my laboratory peers, we first established the size-stage correlations using a combination of microscopy imaging techniques and then used precisely staged anthers from asparagus, daylily (from order *Asparagales*) and *Lilium*, which diverged 120 and 121 MYA from MRCA of grasses respectively (Chase and Reveal 2009), to reveal that both pre-meiotic and meiotic phasiRNA pathways are widely prevalent beyond grasses. This discovery of phasiRNAs beyond grasses pushes their origins into the monocots, at least 70 million years before earlier estimates. In addition, I found that meiotic phasiRNAs are present in pistils, the female-reproductive organ, which

suggests the wider roles for these pathways, and these might not be restricted to male-germline contrary to earlier reports. Our discovery is consistent with the earlier report from rice *MEL1* mutant, which binds with the 21-nt phasiRNAs, and its knockdown impacts female germline development. However, the precise stages and cells in which 24-nt phasiRNAs peak in abundance in female organs will require additional experimentation. Nonetheless, the overlap of 24-nt *PHAS* repertoire with anthers, along with presence of abundant pistil-enriched 24-*PHAS* loci demonstrates that 24-nt phasiRNAs are not restricted to male organs.

This presence of phasiRNA pathways outside the grasses also provided novel insights into the emergence and primitive roles of miRNA triggers, miR2118 and miR2275. The trigger for the pre-meiotic phasiRNA pathway showed a dramatic functional divergence among eudicots, from regulating the *NBS-LRR* network to targeting non-coding *PHAS* precursors in pre-meiotic anthers. The timing of the switch from the constitutively expressed to a specialized pre-meiotic phasiRNA trigger is an open question. Similarly, the origin of the miR2275 family, which was previously reported only in grasses and specifically targets reproductive-enriched *PHAS* precursors in meiotic-stage anthers, is yet unknown. The dichotomy in tissue-specificity and temporal dynamics, correlating with the clustered precursor loci in the asparagus genome, and the presence of miR2275 members perpetuating grass-like functions in *Lilium* showed that specialized triggers for both pre-meiotic and meiotic phasiRNA pathways emerged before the divergence of *Lilium* and asparagus from MRCA of maize, respectively. The co-occurrence of temporally distinct reproductive- and vegetative miR2118 members targeting both *21-PHAS* precursors and *NBS-LRRs* demonstrates that the elaborate pathogen defense regulatory network of *NBS-LRR*

genes was prevalent in earlier monocots than the grasses, perhaps demonstrating an increasingly specialized role of miR2118 in the *Asparagales*. In contrast, miR2275 display a non-canonical role, specifically in Asparagus, triggering 22-nt instead of 24-nt phasiRNAs which co-localizes with miR2275 in pre-meiotic stages and diffuses to all cell layers of meiotic anther. These 22-nt phasiRNAs lack cleaved targets, and show no role in reinforcing meiotic phasiRNA pathways, leaving a gap in our understanding of their roles.

5.4 Status of phasiRNA components and variation of meiotic phasiRNA pathways in monocots

I found that factors for pre-meiotic phasiRNA pathways are well conserved across all the species included in this study, spanning 128 million years of evolution from *Z. marina* to maize, thereby supporting their continuing functional association across monocots. However, AGO18 the proposed effector molecule for meiotic phasiRNAs was absent from *Asparagales*, *Lilium*, and *Zostera*. Instead, I found members of AGO1 and AGO4, likely homologs of rice AGO1d and maize AGO4d (also referred to as AGO104 and AGO9), consistently enriched in meiotic anther across all three species – Asparagus, *Lilium* and daylily. Both AGO1d and AGO4d, were recently proposed by two independent studies in rice (Fei et al. 2016) and maize (Dukowic-Schulze et al. 2016) respectively to load meiotic phasiRNAs, in addition to the AGO18 proposed earlier (Zhai et al. 2015). Although these associations still need to be validated, the diversity of 5'-nucleotide along with results from an earlier study where *ago18* loss of function mutant shows no obvious developmental defects (Wu et al. 2015) supports functional redundancy or co-operative roles of AGOs in phasiRNA pathways. Meiotic-enriched expression of AGO1 and AGO4 homologs in non-grass

monocots strongly support their candidacy as effector molecules for meiotic phasiRNAs. In addition, AGO4 shares the same spectrum of preference for the 24-nt size (H. Wang et al. 2011) and is phylogenetically close to AGO18 (phylogram not shown here). Although such phylogenetic relationships do not necessarily imply functional redundancy, given the importance of maize AGO4d in male and female meiosis (Singh et al. 2011), it is a promising candidate to act in meiotic phasiRNA pathways in the absence of an AGO18 member. Furthermore, maize AGO4d functions in heterochromatic CHH and CHG methylation (Singh et al. 2011). Its homolog in Arabidopsis, AtAGO9 is a key component of the RdDM pathway (Matzke and Mosher 2014); meiotic phasiRNAs too were reported to play role in *cis* DNA methylation (Dukowic-Schulze et al. 2016). This functional overlap between AGO4d and meiotic phasiRNAs, although weak at this point but given its enrichment in meiotic anther which is consistent across non-grass monocots and phylogenetic closeness to AGO18 cannot be ignored as a mere coincidence. Further characterization of maize AGO4d or its homologs in species included in this study holds significant potential to provide crucial insights into the functions of meiotic phasiRNAs. Nonetheless, data from this study demonstrates that meiotic phasiRNA pathway is independent of grass-specific components, therefore could even extend beyond the monocots.

This study also revealed substantial variation in phasiRNA pathways compared to those described in grasses. Three forms of meiotic phasiRNAs emerge in this study – a) the canonical matching grasses, b) phased siRNAs derived from IRs (pird-siRNAs), triggered by miR2275 and c) phased- (but not secondary) IR-derived sRNAs, lacking miR2275 trigger (pird-sRNAs). The pird-siRNAs were detected along with canonical phasiRNAs in maize and *Lilium*, and represent a very small fraction of

repertoire. Foldback-*PHAS* in *Lilium*, lacking miR2275 target site could also be pird-siRNA transcripts with processed 5'-end. Considering this possibility, pird-sRNAs are identified only in Asparagus. Although, there are a number of 24-nt phased loci reported in maize that lacks miR2275 trigger site, but these are not associated with inverted repeats (Zhai et al. 2015). In the absence of pird-sRNA loci in other species, the parallels in spatial localization and temporal dynamics are not sufficient to establish these as functional analogs to meiotic phasiRNAs. The precise nature of signals (such as a sequence or structural motif) that guides the processing of pird-sRNA precursors and mechanism through which these are shuttled to the cell layer needs to be determined in future research. Furthermore, genetic characterization of loci generating pird-sRNAs in Asparagus and loci generating canonical phasiRNAs in maize is required could provide an answer to these being functional analogs.

5.5 Summary

This dissertation provides a new set of computational methods and algorithms for the field of plant sRNA biology, allowing researchers to accelerate their work to catch up with the progress in sequencing technologies. The high-performance, feature-rich and next-generation software developed here addresses critical bioinformatic gaps and significantly expands the capacity of plant researchers in mining small RNA sequencing data to characterize miRNA and secondary siRNA pathways. Both the *PHASIS* (Kakrana et al. 2017) suite and *sPARTA* (Kakrana et al. 2014) are open-source, released under permissive free software license; they are hosted on GitHub with wiki pages, an issue reporting system and progress tracker, to ensure a long-term support to the community.

My research on phased siRNA pathways, described in this dissertation provides substantial new insights into their functional roles, evolution and mechanisms of biogenesis. Cells associated with the male germline, in rice and maize (grasses), produce massive amounts of reproductive-enriched phased siRNAs (21- and 24-nt). Since their first report in rice, parallels with mammalian PIWI-associated RNAs (piRNAs) have been consistently highlighted. Earlier reports have shown that these siRNAs are a general property of the grasses, specific to male germline and critical for reproductive success. My study reports their presence, in canonical forms, beyond the grasses, along with presence of meiotic phasiRNAs in female reproductive tissues, thereby indicating a broader role of pre-meiotic and meiotic phasiRNA pathways in germline development, which coupled with the observation that these pathways are not dependent upon grass-specific components proposes their existence even beyond the monocots. These novel insights on germline-associated phasiRNA pathways warrant a much deeper evolutionary and mechanistic investigation of phasiRNAs in a broader range of angiosperms and even eudicots.

REFERENCES

Addo-Quaye, Charles, Tifani W. Eshoo, David P. Bartel, and Michael J. Axtell. 2008. "Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome." *Current Biology* 18: 758–762. doi:10.1016/j.cub.2008.04.042.

Addo-Quaye, Charles, Webb Miller, and Michael J. Axtell. 2009. "CleaveLand: A Pipeline for Using Degradome Data to Find Cleaved Small RNA Targets." *Bioinformatics* 25 (1): 130–31. doi:10.1093/bioinformatics/btn604.

Albert, Victor a., W. Bradley Barbazuk, Claude W. DePamphilis, Joshua P. Der, James Leebens-Mack, Hong Ma, Jeffrey D. Palmer, et al. 2013. "The Amborella Genome and the Evolution of Flowering Plants." *Science* 342 (December): 1241089. doi:10.1126/science.1241089.

Allen, Edwards, and Miya D. Howell. 2010. "miRNAs in the Biogenesis of Trans-Acting siRNAs in Higher Plants." *Seminars in Cell & Developmental Biology*, Plant MicroRNAs and Development and Disease, 21 (8): 798–804. doi:10.1016/j.semcdb.2010.03.008.

Allen, Edwards, Zhixin Xie, Adam M. Gustafson, and James C. Carrington. 2005. "microRNA-Directed Phasing during Trans-Acting siRNA Biogenesis in Plants." *Cell* 121: 207–221. doi:10.1016/j.cell.2005.04.004.

Arikit, Siwaret, Rui Xia, Atul Kakrana, Kun Huang, Jixian Zhai, Zhe Yan, Oswaldo Valdés-López, et al. 2014. "An Atlas of Soybean Small RNAs Identifies Phased siRNAs from Hundreds of Coding Genes." *The Plant Cell*, December. doi:10.1105/tpc.114.131847.

Arikit, Siwaret, Jixian Zhai, and Blake C. Meyers. 2013. "Biogenesis and Function of Rice Small RNAs from Non-Coding RNA Precursors." *Current Opinion in Plant Biology* 16 (2): 170–79. doi:10.1016/j.pbi.2013.01.006.

Arribas-Hernández, Laura, Antonin Marchais, Christian Poulsen, Bettina Haase, Judith Hauptmann, Vladimir Benes, Gunter Meister, and Peter Brodersen. 2016. "The Slicer Activity of ARGONAUTE1 Is Required Specifically for the Phasing, Not Production, of Trans-Acting Short Interfering RNAs in Arabidopsis." *The Plant Cell*, June. doi:10.1105/tpc.16.00121.

Axtell, Michael J. 2013a. "Classification and Comparison of Small RNAs from Plants." *Annual Review of Plant Biology* 64 (1): 137–59. doi:10.1146/annurev-arplant-050312-120043.

———. 2013b. "ShortStack: Comprehensive Annotation and Quantification of Small RNA Genes." *RNA* 19 (6): 740–51. doi:10.1261/rna.035279.112.

———. 2015. “Non-Coding RNAs: The Small Mysteries of Males.” *Nature Plants* 1 (5): 15055. doi:10.1038/nplants.2015.55.

Axtell, Michael J., Calvin Jan, Ramya Rajagopalan, and David P. Bartel. 2006. “A Two-Hit Trigger for siRNA Biogenesis in Plants.” *Cell* 127 (3): 565–77. doi:10.1016/j.cell.2006.09.032.

Baev, Vesselin, Ivan Milev, Mladen Naydenov, Elena Apostolova, Georgi Minkov, Ivan Minkov, and Galina Yahubyan. 2011. “Implementation of a de Novo Genome-Wide Computational Approach for Updating Brachypodium miRNAs.” *Genomics* 97: 282–293. doi:10.1016/j.ygeno.2011.02.008.

Bologna, Nicolás G., Arnaldo L. Schapire, Jixian Zhai, Uciel Chorostecki, Jerome Boisbouvier, Blake C. Meyers, and Javier F. Palatnik. 2013. “Multiple RNA Recognition Patterns during microRNA Biogenesis in Plants.” *Genome Research* 23 (10): 1675–89. doi:10.1101/gr.153387.112.

Borges, Filipe, and Robert A. Martienssen. 2015. “The Expanding World of Small RNAs in Plants.” *Nature Reviews Molecular Cell Biology* 16 (12): 727–41. doi:10.1038/nrm4085.

Brennecke, Julius, Alexei A. Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J. Hannon. 2007. “Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila.” *Cell* 128 (6): 1089–1103. doi:10.1016/j.cell.2007.01.043.

Brousse, Cécile, Qikun Liu, Linda Beauclair, Aurélie Deremetz, Michael J Axtell, and Nicolas Bouché. 2014. “A Non-Canonical Plant microRNA Target Site.” *Nucleic Acids Research*, February, gku157–. doi:10.1093/nar/gku157.

Chase, Mark W., and James L. Reveal. 2009. “A Phylogenetic Classification of the Land Plants to Accompany APG III.” *Botanical Journal of the Linnean Society* 161 (2): 122–27. doi:10.1111/j.1095-8339.2009.01002.x.

Chen, Ho-Ming, Li-Teh Chen, Kanu Patel, Yi-Hang Li, David C. Baulcombe, and Shu-Hsing Wu. 2010. “22-Nucleotide RNAs Trigger Secondary siRNA Biogenesis in Plants.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (34): 15269–74. doi:10.1073/pnas.1001738107.

Chen, Ho-Ming, Yi-Hang Li, and Shu-Hsing Wu. 2007. “Bioinformatic Prediction and Experimental Validation of a microRNA-Directed Tandem Trans-Acting siRNA Cascade in Arabidopsis.” *Proceedings of the National Academy of Sciences* 104 (9): 3318–23. doi:10.1073/pnas.0611119104.

Chi, Sung Wook, Gregory J. Hannon, and Robert B. Darnell. 2012. “An Alternative Mode of microRNA Target Recognition.” *Nature Structural & Molecular Biology* 19 (3): 321–27. doi:10.1038/nsmb.2230.

Chi, Sung Wook, Julie B. Zang, Aldo Mele, and Robert B. Darnell. 2009. “Argonaute HITS-CLIP Decodes microRNA-mRNA Interaction Maps.” *Nature* 460 (7254): 479–86. doi:10.1038/nature08170.

Cuperus, Josh T, Alberto Carbonell, Noah Fahlgren, Hernan Garcia-Ruiz, Russell T Burke, Atsushi Takeda, Christopher M Sullivan, Sunny D Gilbert, Taiowa A Montgomery, and James C Carrington. 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. doi:10.1038/nsmb.1866.

Cuperus, Josh T., Noah Fahlgren, and James C. Carrington. 2011. "Evolution and Functional Diversification of MIRNA Genes." *The Plant Cell* 23 (2): 431–42. doi:10.1105/tpc.110.082784.

Curaba, Julien, Mark Talbot, Zhongyi Li, and Chris Helliwell. 2013. "Over-Expression of microRNA171 Affects Phase Transitions and Floral Meristem Determinancy in Barley." *BMC Plant Biology* 13: 6. doi:10.1186/1471-2229-13-6.

Czech, Benjamin, and Gregory J. Hannon. 2011. "Small RNA Sorting: Matchmaking for Argonautes." *Nature Reviews. Genetics* 12 (1): 19–31. doi:10.1038/nrg2916.

Davis, Matthew P A, Stijn van Dongen, Cei Abreu-Goodger, Nenad Bartonicek, and Anton J. Enright. 2013. "Kraken: A Set of Tools for Quality Control and Analysis of High-Throughput Sequence Data." *Methods* 63: 41–49. doi:10.1016/j.ymeth.2013.06.027.

Debernardi, Juan M., Ramiro E. Rodriguez, Martin A. Mecchia, and Javier F. Palatnik. 2012. "Functional Specialization of the Plant miR396 Regulatory Network through Distinct microRNA-Target Interactions." *PLoS Genetics* 8. doi:10.1371/journal.pgen.1002419.

Didiano, Dominic, and Oliver Hobert. 2006. "Perfect Seed Pairing Is Not a Generally Reliable Predictor for miRNA-Target Interactions." *Nature Structural & Molecular Biology* 13: 849–851. doi:10.1038/nsmb1138.

Dukowic-Schulze, Stefanie, Anitha Sundararajan, Thiruvarangan Ramaraj, Shahryar Kianian, Wojciech P. Pawlowski, Joann Mudge, and Changbin Chen. 2016. "Novel Meiotic miRNAs and Indications for a Role of PhasiRNAs in Meiosis." *Plant Genetics and Genomics*, 762. doi:10.3389/fpls.2016.00762.

Dunoyer, Patrice, Christophe Himber, and Olivier Voinnet. 2005. "DICER-LIKE 4 Is Required for RNA Interference and Produces the 21-Nucleotide Small Interfering RNA Component of the Plant Cell-to-Cell Silencing Signal." *Nature Genetics* 37 (12): 1356–60. doi:10.1038/ng1675.

Fahlgren, Noah, and James C. Carrington. 2010. "miRNA Target Prediction in Plants." *Methods in Molecular Biology (Clifton, N.J.)* 592: 51–57. doi:10.1007/978-1-60327-005-2_4.

Fei, Qili, Rui Xia, and Blake C. Meyers. 2013. "Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks[OPEN]." *The Plant Cell* 25 (7): 2400–2415. doi:10.1105/tpc.113.114652.

Fei, Qili, Li Yang, Wanqi Liang, Dabing Zhang, and Blake C. Meyers. 2016. "Dynamic Changes of Small RNAs in Rice Spikelet Development Reveal Specialized Reproductive phasiRNA Pathways." *Journal of Experimental Botany* 67 (21): 6037–49. doi:10.1093/jxb/erw361.

Folkes, Leighton, Simon Moxon, Hugh C. Woolfenden, Matthew B. Stocks, Gyorgy Szittya, Tamas Dalmay, and Vincent Moulton. 2012. “PAREsnip: A Tool for Rapid Genome-Wide Discovery of Small RNA/Target Interactions Evidenced through Degradome Sequencing.” *Nucleic Acids Research* 40 (13): e103. doi:10.1093/nar/gks277.

German, Marcelo A., Shujun Luo, Gary Schroth, Blake C. Meyers, and Pamela J. Green. 2009. “Construction of Parallel Analysis of RNA Ends (PARE) Libraries for the Study of Cleaved miRNA Targets and the RNA Degradome.” *Nature Protocols* 4 (3): 356–62. doi:10.1038/nprot.2009.8.

Gong, Lei, Atul Kakrana, Siwaret Arikit, Blake C. Meyers, and Jonathan F. Wendel. 2013. “Composition and Expression of Conserved MicroRNA Genes in Diploid Cotton (Gossypium) Species.” *Genome Biology and Evolution* 5 (12): 2449–59. doi:10.1093/gbe/evt196.

Goodstein, David M, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, et al. 2012. “Phytozome: A Comparative Platform for Green Plant Genomics.” *Nucleic Acids Research* 40: 1178–1186. doi:10.1093/nar/gkr944.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome.” *Nature Biotechnology* 29 (7): 644–52. doi:10.1038/nbt.1883.

Gregory, Brian D., Ronan C. O'Malley, Ryan Lister, Mark A. Urich, Julian Tonti-Filippini, Huaming Chen, A. Harvey Millar, and Joseph R. Ecker. 2008. “A Link between RNA Metabolism and Silencing Affecting Arabidopsis Development.” *Developmental Cell* 14: 854–866. doi:10.1016/j.devcel.2008.04.005.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.” *Systematic Biology* 59 (3): 307–21. doi:10.1093/sysbio/syq010.

Guo, Qingli, Xiongfei Qu, and Weibo Jin. 2015. “PhaseTank: Genome-Wide Computational Identification of phasiRNAs and Their Regulatory Cascades.” *Bioinformatics* 31 (2): 284–86. doi:10.1093/bioinformatics/btu628.

Ha, I, B Wightman, and G Ruvkun. 1996. “A Bulged Lin-4/Lin-14 RNA Duplex Is Sufficient for Caenorhabditis Elegans Lin-14 Temporal Gradient Formation.” *Genes & Development* 10: 3041–3050. doi:10.1101/gad.10.23.3041.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. “De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis.” *Nature Protocols* 8 (8): 1494–1512. doi:10.1038/nprot.2013.084.

Han, Bo W., Wei Wang, Chengjian Li, Zhiping Weng, and Phillip D. Zamore. 2015. “Noncoding RNA. piRNA-Guided Transposon Cleavage Initiates Zucchini-

Dependent, Phased piRNA Production." *Science (New York, N.Y.)* 348 (6236): 817–21. doi:10.1126/science.aaa1264.

Han, Jinju, Yoontae Lee, Kyu-Hyeon Yeom, Jin-Wu Nam, Inha Heo, Je-Keun Rhee, Sun Young Sohn, Yunje Cho, Byoung-Tak Zhang, and V. Narry Kim. 2006. "Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex." *Cell* 125 (5): 887–901. doi:10.1016/j.cell.2006.03.043.

Hedges, S. Blair, Julie Marin, Michael Suleski, Madeline Paymer, and Sudhir Kumar. 2015. "Tree of Life Reveals Clock-Like Speciation and Diversification." *Molecular Biology and Evolution* 32 (4): 835–45. doi:10.1093/molbev/msv037.

Henderson, Ian R., Xiaoyu Zhang, Cheng Lu, Lianna Johnson, Blake C. Meyers, Pamela J. Green, and Steven E. Jacobsen. 2006. "Dissecting Arabidopsis Thaliana DICER Function in Small RNA Processing, Gene Silencing and DNA Methylation Patterning." *Nature Genetics* 38 (6): 721–25. doi:10.1038/ng1804.

Hsu, Sheng-Da, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel, Chih-Hung Chou, Chao-Fang Chu, et al. 2014. "miRTarBase Update 2014: An Information Resource for Experimentally Validated miRNA-Target Interactions." *Nucleic Acids Research* 42 (Database issue): D78–85. doi:10.1093/nar/gkt1266.

Izumi, Natsuko, Keisuke Shoji, Yuriko Sakaguchi, Shozo Honda, Yohei Kirino, Tsutomu Suzuki, Susumu Katsuma, and Yukihide Tomari. 2016. "Identification and Functional Analysis of the Pre-piRNA 3′ Trimmer in Silkworms." *Cell* 164 (5): 962–73. doi:10.1016/j.cell.2016.01.008.

Javelle, Marie, and Marja C. P. Timmermans. 2012. "In Situ Localization of Small RNAs in Plants by Using LNA Probes." *Nature Protocols* 7 (3): 533–41. doi:10.1038/nprot.2012.006.

Jeong, Dong H, Skye A Schmidt, Linda A Rymarquis, Sunhee Park, Matthias Ganssmann, Marcelo A German, Monica Accerbi, et al. 2013. "Parallel Analysis of RNA Ends Enhances Global Investigation of microRNAs and Target RNAs of Brachypodium Distachyon." *Genome Biology* 14 (12): R145. doi:10.1186/gb-2013-14-12-r145.

Jeong, Dong-Hoon, Sunhee Park, Jixian Zhai, Sai Guna Ranjan Gurazada, Emanuele De Paoli, Blake C. Meyers, and Pamela J. Green. 2011. "Massive Analysis of Rice Small RNAs: Mechanistic Implications of Regulated microRNAs and Variants for Differential Target RNA Cleavage." *The Plant Cell* 23 (12): 4185–4207. doi:10.1105/tpc.111.089045.

Jeong, Dong-Hoon, Shawn R Thatcher, Rebecca S H Brown, Jixian Zhai, Sunhee Park, Linda a Rymarquis, Blake C Meyers, and Pamela J Green. 2013. "Comprehensive Investigation of microRNAs Enhanced by Analysis of Sequence Variants, Expression Patterns, ARGONAUTE Loading, and Target Cleavage." *Plant Physiology* 162 (3): 1225–45. doi:10.1104/pp.113.219873.

Johnson, Cameron, Anna Kasprzewska, Kristin Tennessen, John Fernandes, Guo-Ling Nan, Virginia Walbot, Venkatesan Sundaresan, Vicki Vance, and Lewis H. Bowman. 2009. "Clusters and Superclusters of Phased Small RNAs in the Developing Inflorescence of Rice." *Genome Research* 19 (8): 1429–40. doi:10.1101/gr.089854.108.

Jouannet, Virginie, Ana Beatriz Moreno, Taline Elmayan, Hervé Vaucheret, Martin D. Crespi, and Alexis Maizel. 2012. "Cytoplasmic Arabidopsis AGO7 Accumulates in Membrane-Associated siRNA Bodies and Is Required for Ta-siRNA Biogenesis." *The EMBO Journal* 31 (7): 1704–13. doi:10.1038/emboj.2012.20.

Kakrana, Atul, Reza Hammond, Parth Patel, Mayumi Nakano, and Blake C. Meyers. 2014. "sPARTA: A Parallelized Pipeline for Integrated Analysis of Plant miRNA and Cleaved mRNA Data Sets, Including New miRNA Target-Identification Software." *Nucleic Acids Research* 42 (18): e139–e139. doi:10.1093/nar/gku693.

Kakrana, Atul, Pingchuan Li, Parth Patel, Reza Hammond, Deepti Anand, Sandra Mathioni, and Blake Meyers. 2017. "PHASIS: A Computational Suite for de Novo Discovery and Characterization of Phased, siRNA-Generating Loci and Their miRNA Triggers." *bioRxiv*, July, 158832. doi:10.1101/158832.

Khorshid, Mohsen, Jean Hausser, Mihaela Zavolan, and Erik van Nimwegen. 2013. "A Biophysical miRNA-mRNA Interaction Model Infers Canonical and Noncanonical Targets." *Nature Methods* 10: 253–5. doi:10.1038/nmeth.2341.

Kim, V. Narry, Jinju Han, and Mikiko C. Siomi. 2009. "Biogenesis of Small RNAs in Animals." *Nature Reviews Molecular Cell Biology* 10 (2): 126–39. doi:10.1038/nrm2632.

Komiya, Reina, Hajime Ohyanagi, Mitsuru Niihama, Toshiaki Watanabe, Mutsuko Nakano, Nori Kurata, and Ken-Ichi Nonomura. 2014. "Rice Germline-Specific Argonaute MEL1 Protein Binds to phasiRNAs Generated from More than 700 lincRNAs." *The Plant Journal* 78 (3): 385–97. doi:10.1111/tpj.12483.

Kong, Lei, Yong Zhang, Zhi Qiang Ye, Xiao Qiao Liu, Shu Qi Zhao, Liping Wei, and Ge Gao. 2007. "CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine." *Nucleic Acids Research* 35 (SUPPL.2): 345–349. doi:10.1093/nar/gkm391.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–9. doi:10.1038/nmeth.1923.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.

Lee, Tzuu-fen, Sai Guna Ranjan Gurazada, Jixian Zhai, Shengben Li, Stacey A. Simon, Marjori A. Matzke, Xuemei Chen, and Blake C. Meyers. 2012. "RNA Polymerase V-Dependent Small RNAs in Arabidopsis Originate from Small, Intergenic Loci Including Most SINE Repeats." *Epigenetics* 7 (7): 781–95. doi:10.4161/epi.20290.

Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12 (1): 323. doi:10.1186/1471-2105-12-323.

Li, Fan, Qi Zheng, Paul Ryvkin, Isabelle Dragomir, Yaanik Desai, Subhadra Aiyer, Otto Valladares, et al. 2012. “Global Analysis of RNA Secondary Structure in Two Metazoans.” *Cell Reports* 1 (1): 69–82. doi:10.1016/j.celrep.2011.10.002.

Li, Jun-Hao, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. 2014. “starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data.” *Nucleic Acids Research* 42 (Database issue): D92–7. doi:10.1093/nar/gkt1248.

Li, Shaofang, Lee E. Vandivier, Bin Tu, Lei Gao, So Youn Won, Shengben Li, Binglian Zheng, Brian D. Gregory, and Xuemei Chen. 2015. “Detection of Pol IV/RDR2-Dependent Transcripts at the Genomic Scale in Arabidopsis Reveals Features and Regulation of siRNA Biogenesis.” *Genome Research* 25 (2): 235–45. doi:10.1101/gr.182238.114.

Li, Shengben, Brandon Le, Xuan Ma, Shaofang Li, Chenjiang You, Yu Yu, Bailong Zhang, et al. 2016. “Biogenesis of Phased siRNAs on Membrane-Bound Polysomes in Arabidopsis.” *eLife* 5 (December): e22750. doi:10.7554/eLife.22750.

Li, Yong-Fang, Yun Zheng, Charles Addo-Quaye, Li Zhang, Ajay Saini, Guru Jagadeeswaran, Michael J Axtell, Weixiong Zhang, and Ramanjulu Sunkar. 2010. “Transcriptome-Wide Identification of microRNA Targets in Rice.” *The Plant Journal : For Cell and Molecular Biology* 62: 742–759. doi:10.1111/j.1365-313X.2010.04187.x.

Llave, Cesar, Kristin D Kasschau, Maggie A Rector, and James C Carrington. 2002. “Endogenous and Silencing-Associated Small RNAs in Plants.” *The Plant Cell* 14: 1605–1619. doi:10.1105/tpc.003210.

Mahalakshmi, Y. V., M. V. Jagannadham, and M. W. Pandit. 2000. “Ribonuclease from Cobra Snake Venom: Purification by Affinity Chromatography and Further Characterization.” *IUBMB Life* 49 (4): 309–16. doi:10.1080/15216540050033186.

Mallory, Allison C., Diana V. Dugas, David P. Bartel, and Bonnie Bartel. 2004. “MicroRNA Regulation of NAC-Domain Targets Is Required for Proper Formation and Separation of Adjacent Embryonic, Vegetative, and Floral Organs.” *Current Biology* 14 (12): 1035–46. doi:10.1016/j.cub.2004.06.022.

Mallory, Allison, and Hervé Vaucheret. 2010. “Form, Function, and Regulation of ARGONAUTE Proteins.” *The Plant Cell* 22 (12): 3879–89. doi:10.1105/tpc.110.080671.

Margis, Rogerio, Adriana F. Fusaro, Neil A. Smith, Shaun J. Curtin, John M. Watson, E. Jean Finnegan, and Peter M. Waterhouse. 2006. “The Evolution and Diversification of Dicers in Plants.” *FEBS Letters* 580 (10): 2442–50. doi:10.1016/j.febslet.2006.03.072.

Mathioni, Sandra M., Atul Kakrana, and Blake C. Meyers. 2016. "Characterization of Plant Small RNAs by Next Generation Sequencing." In *Current Protocols in Plant Biology*. John Wiley & Sons, Inc. doi:10.1002/cppb.20043.

Matzke, Marjori A, and Rebecca A Mosher. 2014. "RNA-Directed DNA Methylation: An Epigenetic Pathway of Increasing Complexity." *Nature Reviews. Genetics* 15 (6): 394–408. doi:10.1038/nrg3683.

Mecchia, Martin A., Juan M. Debernardi, Ramiro E. Rodriguez, Carla Schommer, and Javier F. Palatnik. 2013. "MicroRNA miR396 and RDR6 Synergistically Regulate Leaf Development." *Mechanisms of Development* 130: 2–13. doi:10.1016/j.mod.2012.07.005.

Meyers, Blake C, Michael J Axtell, Bonnie Bartel, David P Bartel, David Baulcombe, John L Bowman, Xiaofeng Cao, et al. 2008. "Criteria for Annotation of Plant MicroRNAs." *The Plant Cell* 20: 3186–3190. doi:10.1105/tpc.108.064311.

Mi, Shijun, Tao Cai, Yugang Hu, Yemiao Chen, Emily Hodges, Fangrui Ni, Liang Wu, et al. 2008. "Sorting of Small RNAs into Arabidopsis Argonaute Complexes Is Directed by the 5' Terminal Nucleotide." *Cell* 133 (1): 116–27. doi:10.1016/j.cell.2008.02.034.

Mohn, Fabio, Dominik Handler, and Julius Brennecke. 2015. "piRNA-Guided Slicing Specifies Transcripts for Zucchini-Dependent, Phased piRNA Biogenesis." *Science* 348 (6236): 812–17. doi:10.1126/science.aaa1039.

Mukherjee, Krishanu, Henry Campos, and Bryan Kolaczkowski. 2013. "Evolution of Animal and Plant Dicers: Early Parallel Duplications and Recurrent Adaptation of Antiviral RNA Binding in Plants." *Molecular Biology and Evolution* 30 (3): 627–41. doi:10.1093/molbev/mss263.

Nagasaki, Hiroshi, Jun-ichi Itoh, Katsunobu Hayashi, Ken-ichiro Hibara, Namiko Satoh-Nagasawa, Misuzu Nosaka, Motohiro Mukouhata, et al. 2007. "The Small Interfering RNA Production Pathway Is Required for Shoot Meristem Initiation in Rice." *Proceedings of the National Academy of Sciences of the United States of America* 104 (37): 14867–71. doi:10.1073/pnas.0704339104.

Nakano, Mayumi, Kan Nobuta, Kalyan Vemaraju, Shivakundan Singh Tej, Jeremy W Skogen, and Blake C Meyers. 2006. "Plant MPSS Databases: Signature-Based Transcriptional Resources for Analyses of mRNA and Small RNA." *Nucleic Acids Research* 34: D731–D735. doi:10.1093/nar/gkj077.

Nonomura, Ken-Ichi, Akane Morohoshi, Mutsuko Nakano, Mitsugu Eiguchi, Akio Miyao, Hirohiko Hirochika, and Nori Kurata. 2007. "A Germ Cell–Specific Gene of the ARGONAUTE Family Is Essential for the Progression of Premeiotic Mitosis and Meiosis during Sporogenesis in Rice." *The Plant Cell* 19 (8): 2583–94. doi:10.1105/tpc.107.053199.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment." *Journal of Molecular Biology* 302 (1): 205–17. doi:10.1006/jmbi.2000.4042.

Olsen, Jeanine L., Pierre Rouzé, Bram Verhelst, Yao-Cheng Lin, Till Bayer, Jonas Collen, Emanuela Dattolo, et al. 2016. "The Genome of the Seagrass Zostera

Marina Reveals Angiosperm Adaptation to the Sea." *Nature* advance online publication (January). doi:10.1038/nature16548.

Patel, Parth, S. Deepthi Ramachandruni, Atul Kakrana, Mayumi Nakano, and Blake C. Meyers. 2016. "miTRATA: A Web-Based Tool for microRNA Truncation and Tailing Analysis." *Bioinformatics (Oxford, England)* 32 (3): 450–52. doi:10.1093/bioinformatics/btv583.

Peragine, Angela, Manabu Yoshikawa, Gang Wu, Heidi L Albrecht, and R Scott Poethig. 2004. "SGS3 and SGS2/SDE1/RDR6 Are Required for Juvenile Development and the Production of Trans-Acting siRNAs in Arabidopsis." *Genes & Development* 18 (19): 2368–79. doi:10.1101/gad.1231804.

R Development Core Team, R. 2011. *R: A Language and Environment for Statistical Computing*. Edited by R Development Core Team. Vol. 1. R Foundation for Statistical Computing, 2.11.1. R Foundation for Statistical Computing. http://www.r-project.org.

Reinhart, Brenda J, Earl G Weinstein, Matthew W Rhoades, Bonnie Bartel, and David P Bartel. 2002. "MicroRNAs in Plants." *Genes & Development* 16: 1616–1626. doi:10.1101/gad.1004402.

Rice, P., I. Longden, and A. Bleasby. 2000. "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics: TIG* 16 (6): 276–77.

"RNAcentral: An International Database of ncRNA Sequences." 2015. *Nucleic Acids Research* 43 (Database issue): D123–29. doi:10.1093/nar/gku991.

Rodriguez, Ramiro E, Martin A Mecchia, Juan M Debernardi, Carla Schommer, Detlef Weigel, and Javier F Palatnik. 2010. "Control of Cell Proliferation in Arabidopsis Thaliana by microRNA miR396." *Development (Cambridge, England)* 137: 103–112. doi:10.1242/dev.043067.

Rubio-Somoza, Ignacio, and Detlef Weigel. 2011. "MicroRNA Networks and Developmental Plasticity in Plants." *Trends in Plant Science* 16 (5): 258–64. doi:10.1016/j.tplants.2011.03.001.

Rymarquis, Linda A., Frederic F. Souret, and Pamela J. Green. 2011. "Evidence That XRN4, an Arabidopsis Homolog of Exoribonuclease XRN1, Preferentially Impacts Transcripts with Certain Sequences or in Particular Functional Categories." *RNA* 17 (3): 501–11. doi:10.1261/rna.2467911.

Sato, Kengo, Michiaki Hamada, Kiyoshi Asai, and Toutai Mituyama. 2009. "CENTROIDFOLD: A Web Server for RNA Secondary Structure Prediction." *Nucleic Acids Research* 37 (Web Server issue): W277-280. doi:10.1093/nar/gkp367.

Schommer, Carla, Edgardo G. Bresso, Silvana V. Spinelli, and Javier F. Palatnik. 2012. "Role of MicroRNA miR319 in Plant Development." In *MicroRNAs in Plant Development and Stress Responses*, edited by Ramanjulu Sunkar, 29–47. Signaling and Communication in Plants 15. Springer Berlin Heidelberg. doi:10.1007/978-3-642-27384-1_2.

Shivaprasad, Padubidri V., Ho-Ming Chen, Kanu Patel, Donna M. Bond, Bruno A. C. M. Santos, and David C. Baulcombe. 2012. "A MicroRNA Superfamily

Regulates Nucleotide Binding Site–Leucine-Rich Repeats and Other mRNAs." *The Plant Cell* 24 (3): 859–74. doi:10.1105/tpc.111.095380.

Simpson, Jared T, and Richard Durbin. 2010. "Efficient Construction of an Assembly String Graph Using the FM-Index." *Bioinformatics (Oxford, England)* 26: i367–i373. doi:10.1093/bioinformatics/btq217.

Singh, Manjit, Shalendra Goel, Robert B. Meeley, Christelle Dantec, Hugues Parrinello, Caroline Michaud, Olivier Leblanc, and Daniel Grimanelli. 2011. "Production of Viable Gametes without Meiosis in Maize Deficient for an ARGONAUTE Protein[W]." *The Plant Cell* 23 (2): 443–58. doi:10.1105/tpc.110.079020.

Song, Xianwei, Pingchuan Li, Jixian Zhai, Ming Zhou, Lijia Ma, Bin Liu, Dong-Hoon Jeong, et al. 2012. "Roles of DCL4 and DCL3b in Rice Phased Small RNA Biogenesis." *The Plant Journal* 69 (3): 462–74. doi:10.1111/j.1365-313X.2011.04805.x.

Song, Xianwei, Dekai Wang, Lijia Ma, Zhiyu Chen, Pingchuan Li, Xia Cui, Chunyan Liu, et al. 2012. "Rice RNA-Dependent RNA Polymerase 6 Acts in Small RNA Biogenesis and Spikelet Development." *The Plant Journal* 71 (3): 378–89. doi:10.1111/j.1365-313X.2012.05001.x.

Tang, Wen, Shikui Tu, Heng-Chi Lee, Zhiping Weng, and Craig C. Mello. 2016. "The RNase PARN-1 Trims piRNA 3′ Ends to Promote Transcriptome Surveillance in C. Elegans." *Cell* 164 (5): 974–84. doi:10.1016/j.cell.2016.02.008.

The UniProt Consortium. 2015. "UniProt: A Hub for Protein Information." *Nucleic Acids Research* 43 (D1): D204–12. doi:10.1093/nar/gku989.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks." *Nature Protocols* 7 (3): 562–78. doi:10.1038/nprot.2012.016.

Vazquez, Franck, Hervé Vaucheret, Ramya Rajagopalan, Christelle Lepers, Virginie Gasciolli, Allison C. Mallory, Jean-Louis Hilbert, David P. Bartel, and Patrice Crété. 2004. "Endogenous Trans-Acting siRNAs Regulate the Accumulation of Arabidopsis mRNAs." *Molecular Cell* 16 (1): 69–79. doi:10.1016/j.molcel.2004.09.028.

Vergoulis, Thanasis, Ioannis S Vlachos, Panagiotis Alexiou, George Georgakilas, Manolis Maragkakis, Martin Reczko, Stefanos Gerangelos, Nectarios Koziris, Theodore Dalamagas, and Artemis G Hatzigeorgiou. 2012. "TarBase 6.0: Capturing the Exponential Growth of miRNA Targets with Experimental Support." *Nucleic Acids Research* 40: D222–9. doi:10.1093/nar/gkr1161.

Wang, Huan, Xiuren Zhang, Jun Liu, Takatoshi Kiba, Jongchan Woo, Tolulope Ojo, Markus Hafner, Thomas Tuschl, Nam-Hai Chua, and Xiu-Jie Wang. 2011. "Deep Sequencing of Small RNAs Specifically Associated with Arabidopsis AGO1 and AGO4 Uncovers New AGO Functions." *The Plant Journal* 67 (2): 292–304. doi:10.1111/j.1365-313X.2011.04594.x.

Wang, Liguo, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. 2013. "CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model." *Nucleic Acids Research* 41 (6): e74. doi:10.1093/nar/gkt006.

Warner, Cherish A., Meredith L. Biedrzycki, Samuel S. Jacobs, Randall J. Wisser, Jeffrey L. Caplan, and D. Janine Sherrier. 2014. "An Optical Clearing Technique for Plant Tissues Allowing Deep Imaging and Compatible with Fluorescence Microscopy." *Plant Physiology* 166 (4): 1684–87. doi:10.1104/pp.114.244673.

Wightman, B., I. Ha, and G. Ruvkun. 1993. "Posttranscriptional Regulation of the Heterochronic Gene Lin-14 by Lin-4 Mediates Temporal Pattern Formation in C. Elegans." *Cell* 75: 855–862. doi:10.1016/0092-8674(93)90530-4.

Willmann, Matthew R., Matthew W. Endres, Rebecca T. Cook, and Brian D. Gregory. 2011. "The Functions of RNA-Dependent RNA Polymerases in Arabidopsis." *The Arabidopsis Book* 9: e0146. doi:10.1199/tab.0146.

Wu, Jianguo, Zhirui Yang, Yu Wang, Lijia Zheng, Ruiqiang Ye, Yinghua Ji, Shanshan Zhao, et al. 2015. "Viral-Inducible Argonaute18 Confers Broad-Spectrum Virus Resistance in Rice by Sequestering a Host microRNA." *eLife* 4. doi:10.7554/eLife.05733.

Xia, Rui, Jing Xu, Siwaret Arikit, and Blake C. Meyers. 2015. "Extensive Families of miRNAs and PHAS Loci in Norway Spruce Demonstrate the Origins of Complex phasiRNA Networks in Seed Plants." *Molecular Biology and Evolution*, August, msv164. doi:10.1093/molbev/msv164.

Xia, Rui, Songqing Ye, Zongrang Liu, Blake C. Meyers, and Zhongchi Liu. 2015. "Novel and Recently Evolved MicroRNA Clusters Regulate Expansive F-BOX Gene Networks through Phased Small Interfering RNAs in Wild Diploid Strawberry1[OPEN]." *Plant Physiology* 169 (1): 594–610. doi:10.1104/pp.15.00253.

Xia, Zhen, Peter Clark, Tien Huynh, Phillipe Loher, Yue Zhao, Huang-Wen Chen, Isidore Rigoutsos, and Ruhong Zhou. 2012. "Molecular Dynamics Simulations of Ago Silencing Complexes Reveal a Large Repertoire of Admissible 'seed-Less' Targets." *Scientific Reports* 2 (August). doi:10.1038/srep00569.

Xiao, Feifei, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. 2009. "miRecords: An Integrated Resource for microRNA-Target Interactions." *Nucleic Acids Research* 37: D105–D110. doi:10.1093/nar/gkn851.

Yi, Xin, Zhenhai Zhang, Yi Ling, Wenying Xu, and Zhen Su. 2015. "PNRD: A Plant Non-Coding RNA Database." *Nucleic Acids Research* 43 (Database issue): D982-989. doi:10.1093/nar/gku1162.

Zhai, Jixian, Siwaret Arikit, Stacey A. Simon, Bruce F. Kingham, and Blake C. Meyers. 2014. "Rapid Construction of Parallel Analysis of RNA End (PARE) Libraries for Illumina Sequencing." *Methods (San Diego, Calif.)* 67 (1): 84–90. doi:10.1016/j.ymeth.2013.06.025.

Zhai, Jixian, Dong-Hoon Jeong, Emanuele De Paoli, Sunhee Park, Benjamin D. Rosen, Yupeng Li, Alvaro J. González, et al. 2011. "MicroRNAs as Master Regulators of the Plant NB-LRR Defense Gene Family via the Production of Phased, Trans-Acting siRNAs." *Genes & Development* 25 (23): 2540–53. doi:10.1101/gad.177527.111.

Zhai, Jixian, Han Zhang, Siwaret Arikit, Kun Huang, Guo-Ling Nan, Virginia Walbot, and Blake C Meyers. 2015. "Spatiotemporally Dynamic, Cell-Type-Dependent Premeiotic and Meiotic phasiRNAs in Maize Anthers." *Proceedings of the National Academy of Sciences of the United States of America* 112 (10): 3146–51. doi:10.1073/pnas.1418918112.

Zhai, Jixian, Yuanyuan Zhao, Stacey A Simon, Sheng Huang, Katherine Petsch, Siwaret Arikit, Manoj Pillay, et al. 2013. "Plant microRNAs Display Differential 3' Truncation and Tailing Modifications That Are ARGONAUTE1 Dependent and Conserved across Species." *The Plant Cell* 25 (7): 2417–28. doi:10.1105/tpc.113.114603.

Zhang, Han, Rui Xia, Blake C Meyers, and Virginia Walbot. 2015. "Evolution, Functions, and Mysteries of Plant ARGONAUTE Proteins." *Current Opinion in Plant Biology*, Cell signalling and gene regulation, 27 (October): 84–90. doi:10.1016/j.pbi.2015.06.011.

Zhang, Yu, Rui Xia, Hanhui Kuang, and Blake C. Meyers. 2016. "The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them." *Molecular Biology and Evolution*, August, msw154. doi:10.1093/molbev/msw154.

Zheng, Qi, Paul Ryvkin, Fan Li, Isabelle Dragomir, Otto Valladares, Jamie Yang, Kajia Cao, Li-San Wang, and Brian D. Gregory. 2010. "Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in Arabidopsis." *PLoS Genet* 6 (9): e1001141. doi:10.1371/journal.pgen.1001141.

Zheng, Yun, Yong-Fang Li, Ramanjulu Sunkar, and Weixiong Zhang. 2012. "SeqTar: An Effective Method for Identifying microRNA Guided Cleavage Sites from Degradome of Polyadenylated Transcripts in Plants." *Nucleic Acids Research* 40 (4): e28. doi:10.1093/nar/gkr1092.

Appendix

PERMISSIONS

At the time of submission of this dissertation the research work described in chapter 3 is being prepared for publication and the research described in chapter 4 is being reviewed for publication in Genome Research journal (Cold Spring Harbor Press). The content included in chapter 3 and 4 of this dissertation is part of authors copy of research work. The published copy, if research is published in Genome Research or any other journal, would be property of respective journal or press