BIOINFORMATICS TOOLS TO

CLASSIFY AND CHARACTERIZE PLANT SMALL RNAs

by

Reza Hammond

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 2019

BIOINFORMATICS TOOLS TO

CLASSIFY AND CHARACTERIZE PLANT SMALL RNAs

by

Reza Hammond

Approved: _____

Cathy H. Wu, Ph.D.
Chair of the Department of Bioinformatics and Computational Biology

Approved: _____

Babatunde Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____

Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education and Dean of the Graduate College

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: _____

Randall Wisser, 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: _____

Li Liao, 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: _____

Terry Papoutsakis, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

This dissertation would not at all be possible without the support of so many people in my life. First and foremost, I need to thank Blake Meyers, my advisor throughout both my Masters and Ph.D. Blake took me into his lab despite hardly understanding what a small RNA was and helped hone my skills as a bioinformatician. I certainly have a lot to learn still, but my time training with Blake has made me capable of handling such problems that would not have been possible prior to my training with him. A huge thanks to my committee members: Cathy Wu, Randall Wisser, Li Liao, and Terry Papoutsakis, Your advisement over the last several years have been paramount to my success during my graduate work.

Additionally, I need to thank each and every member of the Meyers' lab. The list is long, but I need to give a special thanks to Atul Kakrana, Saleh Tamim, Parth Patel, Tzuu-fen Lee, Kevin McCormick, Jixian Zhai, Suresh Pokhrel, Kun Huang, Ryan DelPercio, Patricia Baldrich, Josh Rothhaupt, Mayumi Nakano, Guna Gurazada, Caghan Demirci, Feray Demirci, Deepti Ramachandruni, Aleksandra Beric, Alex Harkess, and Chong Teng. This list does not include everybody, but you all deserve a huge and special thanks for your guidance, encouragement, and friendship throughout these years. I can't imagine completing this work without the support from each of you.

To my undergraduate advisors John Cavazos and Jacqueline Washington, thank you so much for your care and development early in my career. I was introduced

to the world of bioinformatics by you both, and you helped me gain the experience required to pursue this PhD. Rest in peace, John.

I need to thank my parents and siblings for their years of support while in graduate school, but especially the many years before. Without your nurture over my life, I cannot imagine having the willpower to complete these years of training and education. Thank you so much.

Finally, I have to give one huge thanks to my wife, Elaina Welch. We started dating shortly after I started graduate school meaning that you have been there through just about everything with me. We have grown so much together during this process, and you made the happiest man alive when you made me your husband. I can’t wait to become a parent with you and introduce our child to this world. Thank you so much. I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Small RNAs (sRNAs) in plants are a highly abundant class of noncoding RNAs that regulate silencing processes, mostly on the basis of sequence complementarity. There are several classes of small RNA that differ in both their biogenesis and mode of silencing. With the advent of next generation sequencing technologies, sRNA presence can be quantified at a relatively low cost and in a short amount of time. Due to the large amount of data however, there are many challenges associated with the classification of individual sRNAs in these datasets. It is for this reason that I have developed several tools to assist in the classification of sRNAs. The first of these tools is *miRador*, a novel micro RNA (miRNA) prediction tool that I designed to predict miRNAs from a set of sRNA libraries. This tool utilizes a newly established set of criteria for annotating plant miRNAs which enable it to predict miRNAs with a high level of accuracy. *sPARTA* is a miRNA target prediction and validation tool which utilizes Parallel Analysis of RNA Ends (PARE) sequencing libraries to validate target cleavage facilitated by a miRNA. Utilizing *sPARTA* in conjunction with *miRador* allows for the prediction and validation of novel miRNAs for a variety of plant species. In this dissertation, I utilize this pipeline on maize sRNA data to predict four novel miRNAs. I also present a clustering script that makes use of the Meyers Lab database systems to quickly cluster a variety of sequencing data. Using this approach, I attempted to characterize 24-nt phasiRNA producing loci for two separate projects, presented as case studies. In the first case study, we identify that

24-nt phasiRNAs in maize may be required for robust fertility at higher growth temperatures. In the second case study, we identify that asparagus generates inverted repeat-derived phasiRNAs from mRNAs lacking a miR2275 target site. Taken together, these tools improve scientists ability to characterize various classes of small RNAs, and their effects on their encoding genomes.

Chapter 1

INTRODUCTION

This section includes a significant portion of a published book chapter within *The Maize Genome* (citation below). I was a co-first author in this book chapter and was responsible for the sections on paramutation, miRNAs, and bioinformatics tools to analyze sRNAs. I was also responsible for much of the editing to have the chapter read as a single voice. Portions of this text have been edited prior to its inclusion here in order to reduce the focus on maize.

Hammond, R., Teng, C., & Meyers, B. C. (2018). Maize Small RNAs as Seeds of Change and Stability in Gene Expression and Genome Stability. In J. Bennetzen, S. Flint-Garcia, C. Hirsch, & R. Tuberosa (Eds.), *The Maize Genome* (pp. 113-127). Cham: Springer International Publishing.

1.1 Overview of Small RNAs

As their name suggests, endogenous small RNAs (sRNA) in plants are short in length, 21 to 24 nucleotides, and thus are unable to encode proteins. Instead, small RNAs guide regulatory silencing processes, mostly on the basis of sequence complementarity, by guiding ARGONAUTE (AGO) proteins to the targets. Small RNAs function at either the transcriptional (DNA) or post-transcriptional (RNA) level, and they are found widely in eukaryotes including yeast, animals, and plants (Borges & Martienssen, 2015; Y. Li, Varala, Moose, & Hudson, 2012). Small RNAs are typically processed from double-stranded or hairpin RNA structures by RNase III enzymes (Dicer-like proteins, or DCLs) from transcripts made by different RNA

polymerases. After processing, the sRNAs are loaded into AGOs to perform their silencing functions. Although sRNAs are indeed small in size, they are complex in their biogenesis and functions, especially in plants (**Figure 1.1**). In *Arabidopsis*, for example, there are three well-characterized categories of small RNAs: microRNA (miRNAs), typically 21 or 22 nucleotides (nt) in length; heterochromatic small interfering RNAs (hc-siRNAs), ~23-24 nt in length; and *trans*-acting small interfering RNAs (tasiRNAs), mainly 21 nt in length. In *Arabidopsis*, each of these sRNA classes is produced and functions via complex and distinct interactions of the existing four Dicer-like proteins (DCLs), ten AGOs, six RNA-dependent RNA polymerases (RDRs) and three RNA polymerase complexes encoded in the *Arabidopsis* genome.

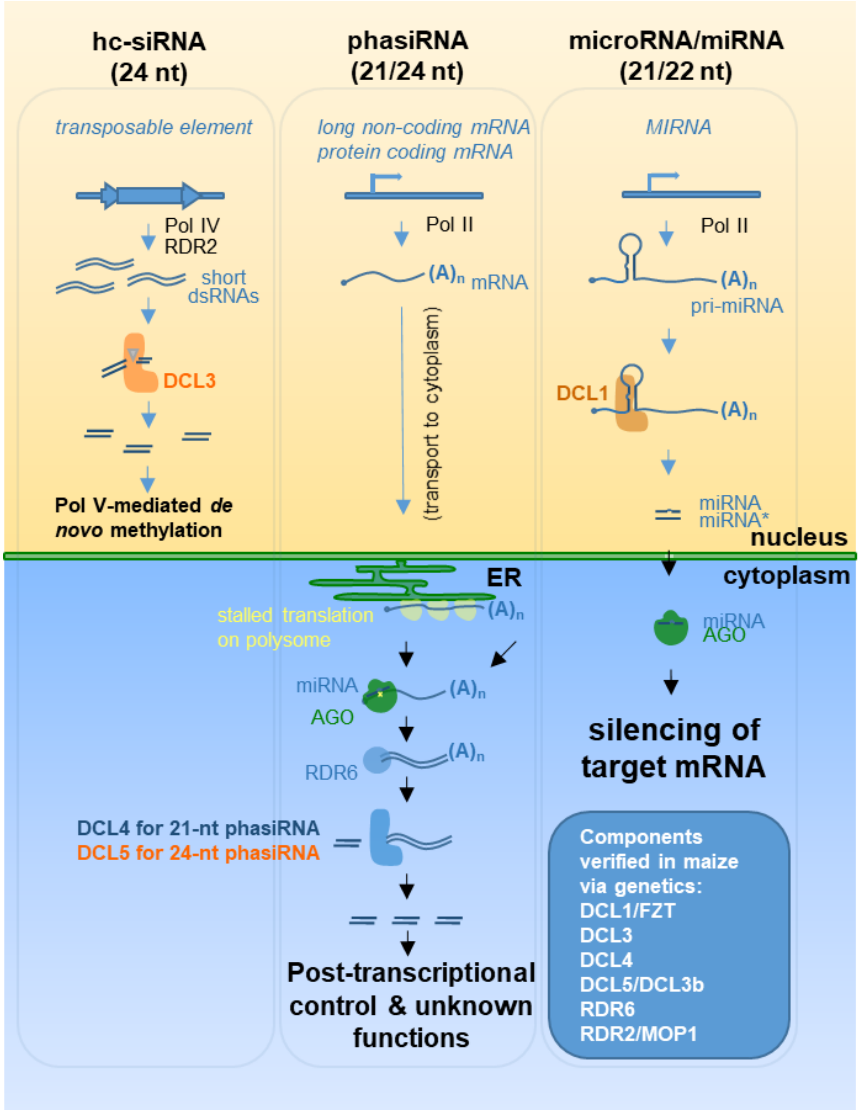

Figure 1.1: Overview of small RNA biogenesis pathways in maize
There are three well-characterized categories of small RNAs: microRNAs (or "miRNAs"), typically 21-22 nucleotides (nt) long; heterochromatic small interfering RNAs (hc-siRNAs), 23-24 nt in length; and phased secondary small interfering RNA (phasiRNAs), mainly 21- or 24-nt in length. hc-siRNAs are diced in the nucleus by DCL3 from short dsRNA precursors, products of Pol IV and RDR2. Then, the hc-siRNAs will direct Pol V-dependent *de novo* methylation at their targeting loci based on sequence complementary. phasiRNAs are generated on membrane-bound polysomes in the cytoplasm from mRNA precursors including long non-coding RNAs or protein-coding mRNAs produced by Pol II. miRNA-AGO complexes initiate a first cut on the phasiRNA precursors, triggering dsRNA conversion by RDR6. Then, DCL4 and (monocot-specific) DCL5 dice the dsRNA into 21- or 24-nt phasiRNAs, respectively. miRNAs are generated as non-coding RNA products of Pol II. The Pol

II-derived precursors miRNA are first folded into imperfect matched hairpin structures, then diced by DCL1 with help from other necessary cofactors (HYL1, SGS3, etc.) yielding miRNA-miRNA* duplexes. Then, the functional miRNAs are loaded into cognate AGO proteins (predominantly AGO1) and translocated into cytosol to perform their function on mRNAs. Several key components mentioned here and in the main text have been validated by years of study in maize mutants, as indicated in the blue box on the bottom right.

The mechanisms of sRNA biogenesis and silencing in *Arabidopsis* have been thoroughly studied over the past two decades. Due to the conservation of many of these pathways across plants, much (but not all) of what has been uncovered in *Arabidopsis* also pertains to sRNAs found in other plant species. While the small size of the *Arabidopsis* genome is sometimes an advantage in sRNA studies, it is not without limitations. In many other plant species, each of the three major classes of sRNAs is more diverse, and some families of enzymes involved in the production of sRNAs have expanded roles. As one example, a consequence of the larger, more repetitive genomes of plants such as maize is that they produce diverse populations of hc-siRNAs from large transposable element populations. Small RNAs in maize also have more diverse functions than in *Arabidopsis*, as they have been implicated in phenomena such as paramutation that are well described in maize (Hollick, 2016). And as studies of sRNAs in maize and other monocots have intensified, additional sRNA-generating loci have been characterized, yielding large sets of long noncoding mRNAs that generate phased secondary small interfering RNAs (phasiRNAs). This led to the identification of at least one additional and seemingly monocot-specific protein, a fifth Dicer (DCL5) with a unique role, demonstrating that there is diversification of sRNA pathways, even within the angiosperms.

1.2 Heterochromatic, Small Interfering RNAs

1.2.1 Overview of Heterochromatic siRNAs

Heterochromatic siRNAs (hc-siRNAs), the most diverse and often most abundant class of sRNAs, were initially characterized from experiments performed mostly in *Arabidopsis*. hc-siRNAs are ~24 nucleotides in length and are typically derived in plant genomes from transposable elements (TEs). hc-siRNAs direct de novo RNA-directed DNA methylation (RdDM), thus influencing chromatin homeostasis and reinforcing heterochromatin formation and maintenance, and, to a lesser extent, regulating gene activity (Law & Jacobsen, 2010; Matzke & Mosher, 2014). The canonical RdDM pathway requires plant-specific RNA polymerase complexes, Pol IV and Pol V, both of which evolved from Pol II. The many components of the hc-siRNA biogenesis and RdDM pathway were recently reviewed in detail (Matzke & Mosher, 2014), so its description is kept to a minimum here.

The RdDM pathway follows three major steps: (i) Pol IV-dependent siRNA biogenesis, (ii) Pol V-mediated de novo DNA methylation, and (iii) chromatin alterations facilitated by Pol V via CG, CHG, and CHH DNA methylation marks (**Figure 1.2**). Briefly, Pol IV is recruited to transposon-like regions marked by H3K9me where it transcribes hc-siRNA precursors; this recruitment occurs with the help of other protein components like SAWADEE HOMEODOMAIN HOMOLOGUE 1 (SHH1) and CLASSY1 (CLSY1). Next, the precursors are converted into double-stranded RNAs (dsRNAs) by RNA-DEPENDENT RNA POLYMERASE 2 (RDR2). These dsRNA molecules are then cleaved by DICER-LIKE 3 (DCL3). The 24-nt siRNA products are then loaded into ARGONAUTE 4 (AGO4) to direct their function. Upon binding the mature 24-nt hc-siRNA products,

AGO4 is recruited by Pol V via proteins that include SUVH2/9, SPT5L/KTF1, IDN2, and the DRD1-DMS3-RDM1 (DDR) complex to target transcripts by sequence homology. When AGO4 finds the target, DOMAINS REARRANGED METHYLASE 2 (DRM2) is recruited by AGO4 and RDM1 to initiate de novo DNA methylation. Finally, the RdDM machinery further influences chromatin as a result of interactions between Pol V scaffold RNAs and the SWI/SNF chromatin remodeling complex, plus an interaction between methylated DNA and histone markers that strengthens gene silencing in a self-reinforcing loop, utilizing the histone-modifying enzymes SUVH4/5/6, UBP26, JMJ14, LDL1/2 and HDA6. Other than this 'canonical RdDM', there are a number of effectors of importance that participate in post-transcriptional gene silencing (PTGS) to also affect RdDM; these include RDR6, Pol II, and NERD, which have roles in cases deemed 'non-canonical RdDM', reviewed in detail in a subsequent section on phasiRNAs (**Section 1.4**). Although only a small proportion of 24 nt hc-siRNAs have been studied in detail (Wang & Axtell, 2017), the following section describes the observation that the absence of hc-siRNAs results in a variety of phenotypic defects, arguably accentuated in maize relative to other plants.

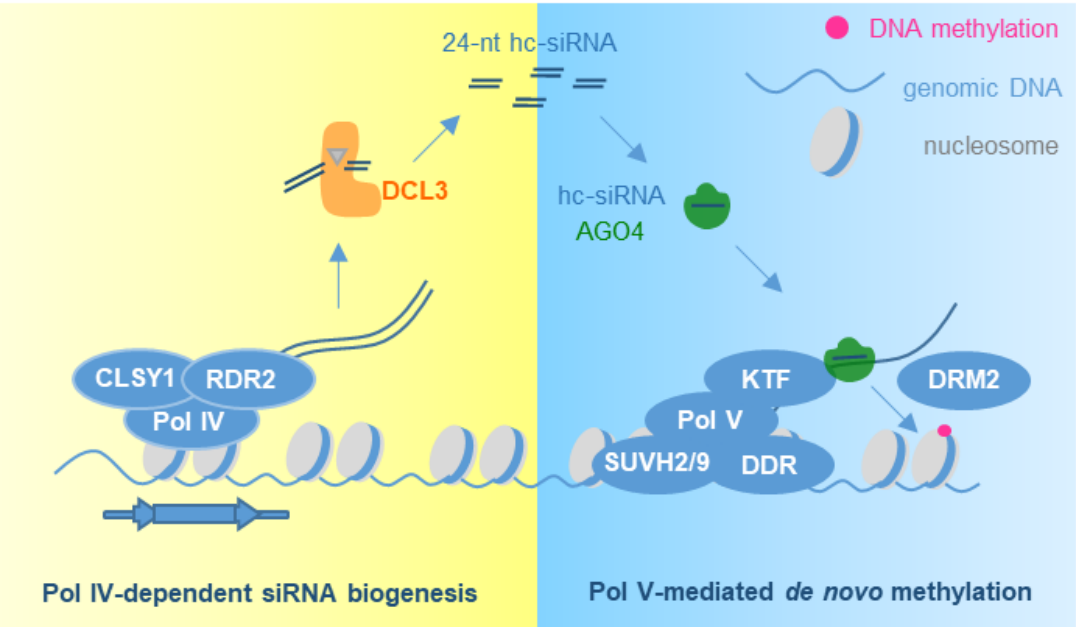

Figure 1.2: Overview of hc-siRNA and RdDM pathway in maize
The primary role of hc-siRNAs is in the RdDM pathway, as outlined here, modified from Matzke and Mosher (2014). hc-siRNA biogenesis begins with the transcription of a transposable repeat locus by Pol IV and other necessary factors. This product is then converted to a dsRNA by RDR2. Then, the dsRNA precursor is cleaved by DCL3 into a 24-nt hc-siRNA. Some hc-siRNAs are bound by AGO4, the major effector in the canonical RdDM pathway. Finally, the siRNA-AGO4 complex binds to Pol V transcripts via sequence complementary to direct *de novo* DNA methylation by DRM2. Analyses of mutants and protein-protein interactions in maize has validated effectors in RdDM, including the following: *RMR1*, *RMR1-Like*, *MOP2*/*RMR7*, *MOP3*/*RMR6*, *RDR2*/*MOP1*, *DCL3*/*RMR5*, *AGO4*/*AGO121*.

1.2.2 Heterochromatic siRNAs in Maize

As the most abundant small RNA species in plants, 24-nt hc-siRNAs have the potential to contribute greatly to phenotypic plasticity. Though not many mutants in RdDM have been isolated in maize, unlike *Arabidopsis*, loss-of-function mutations in maize RDR2 (originally named *MEDIATOR OF PARAMUTATION 1* (*MOP1*) (Dorweiler et al., 2000), *NUCLEAR RNA-DEPENDENT RNA POLYMERASE IV SUBUNIT 1* (*NRPD1*), originally named R*EQUIRED TO MAINTAIN REPRESSION*

6, *RMR6*) (Hollick, Kermicle, & Parkinson, 2005), and *DRD1* (aka *RMR1*) (Hale, Stonaker, Gross, & Hollick, 2007) all substantially impact the production of 24-nt hc-siRNAs, as in *Arabidopsis*. Studies of these maize mutants have revealed roles for hc-siRNAs in paramutation, epiallele regulation, and many other traits, the loss of which is also visually apparent (Hollick, 2016). Therefore, the developmental phenotypic defects of a loss of single components of the hc-siRNA or RdDM pathway are weaker in *Arabidopsis* than in maize, suggesting that hc-siRNAs have more functional significance in maize. Perhaps this is due to the vastly increased repertoire of repetitive elements in maize, comprising the bulk of its large genome. Alternatively, there might be additional required functions for hc-siRNA-dependent pathways in maize that are less important in *Arabidopsis*. Given the variety of phenotypes more evident in maize due to the loss of hc-siRNAs, more questions about 24-nt hc-siRNAs in maize remain to be addressed.

1.3 Paramutation

1.3.1 Overview of Paramutation

Paramutation is the non-Mendelian segregation of alleles due to an epigenetic change induced by one allele in another allele of the same gene. In such an interaction, the paramutagenic allele produces a heritable change in the interaction of the homologous allele, referred to as the ‘paramutable’ allele (Arteaga-Vazquez, Sidorenko, Rabanal, Shrivistava, & Nobuta, 2010; Hollick, 2016). In this scenario, crossing a plant with a paramutagenic allele with a plant carrying a paramutable allele will result in progeny with suppressed expression of the paramutable allele. Outcrossing these hybrids to plants carrying the paramutable allele results in most, if

not all, progeny and successive generations displaying the phenotype of the paramutagenic allele despite not actually carrying this paramutagenic allele (**Figure 1.3**). The phenomenon of paramutation was first described by Alexander Brink in 1956 when he identified the maize *red1* gene, which defies the rules of Mendelian inheritance (Alleman et al., 2006; Brink, 1956). Perhaps one of the more curious features of paramutation is that the paramutable allele becomes paramutagenic upon exposure to another paramutagenic allele in trans (Hollick, 2016).

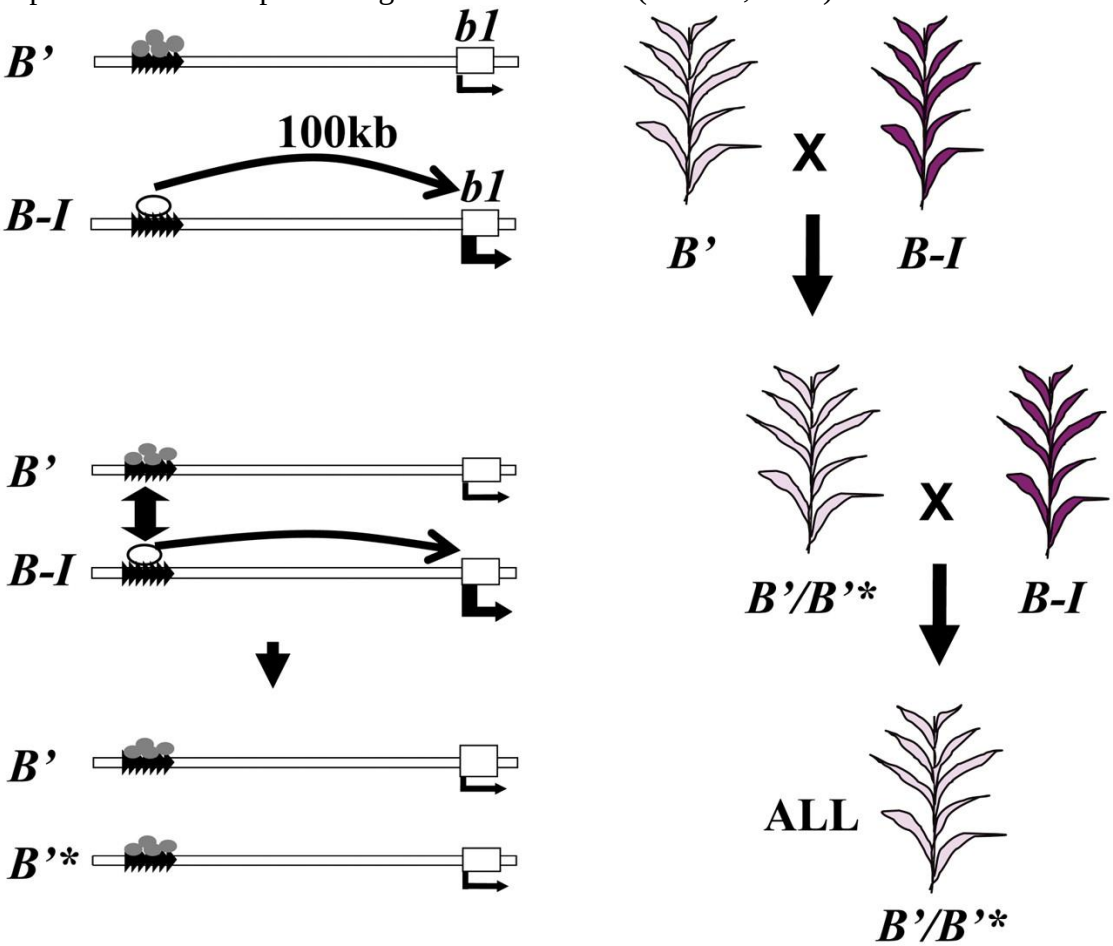

Figure 1.3: Description of paramutation
Paramutation at *b1*. The paramutable (*B-I*) and paramutagenic (*B′*) epialleles have identical DNA sequences, but distinct chromatin structures represented by different proteins (circles) associated with the seven tandem repeats (arrowheads) that mediate paramutation and are located 100 kbp upstream of the coding region (open box). The

arrows below the gene boxes indicate the direction of transcription and the thickness of the arrows indicate the relative transcription levels observed with the two epialleles. The crosses and phenotypes used to monitor paramutation are diagrammed with *B′**, indicating a *B′* allele that was *B-I* in the previous generation. *B′* and *B′** are indistinguishable in their ability to paramutate *B-I*. Figure from (Chandler & Alleman, 2008), courtesy of Vicki Chandler.

Paramutation is best understood in maize as the result of studies of four distinct loci: *booster* (*b*), *red* (*r*), *purple plant1* (*pl1*), and *pericarp color1* (*p1*). Each of these genes encodes a transcription factor that regulates the production of anthocyanin pigments. These loci are studied because they are both susceptible to paramutation and they regulate an easy-to-see phenotype – the coloration of various plant tissues. The combined effects of the *b* and *pl1* loci direct plant color, the combined effects of *pl1* and *r* direct anther color, and the *p1* locus directs pericarp color. Despite the similarities in phenotypic consequences, it appears that different mechanisms control paramutation at each of these loci (Hollick, 2016).

As a result of mutant screens involving these loci, numerous genes required for paramutation have been identified from multiple labs, identified via forward-genetic screens. Mutants identified as *mediator of paramutation* (*mop*) have been isolated in the *b1* system (Arteaga-Vazquez et al., 2010) while several other mutants identified as *required to maintain repression* (*rmr*) have been isolated in the *pl1* system (Hollick & Chandler, 2001). Thus far, all or nearly all genes characterized as required for paramutation apparently encode proteins associated with the biogenesis or function of siRNAs (Table 1), which is why paramutation is an important part of this chapter.

Table 1.1: Loci implicated in paramutation and the mutants identified to suppress paramutation

***Arabidopsis* Gene Name**	**Maize Paramutation Mutant**	**Marker for Loss of Paramutation**	**Reference**

RDR2	*mop1*	*b* and *pl1*	(Dorweiler et al., 2000)
NRPD1	*mop3* and *rmr6*	*b* and *pl1*	(Erhard et al., 2009; Sloan, Sidorenko, & McGinnis, 2014)
NRPD2/E2	*mop2* and *rmr7*	*p1*	(Sidorenko et al., 2009; Stonaker, Lim, Erhard, & Hollick, 2009)
None	*rmr2*	*p1* and r	(Barbour et al., 2012)
SNF2/DRD1	*rmr1*	*p1* and r	(Hale, Erhard, Lisch, & Hollick, 2009)
DCL3	*rmr5*	*b*	(Gabriel et al., 2015)

1.3.2 siRNAs and Paramutation

The first maize gene cloned for which the loss of function eliminates paramutation was *mop1*, subsequently referred to as $rdr2_{mop1}$. $rdr2_{mop1}$ is a mutation in the maize ortholog of *Arabidopsis thaliana RDR2*, a critical protein in the biogenesis pathway of 24-nt heterochromatic siRNAs (Alleman et al., 2006). In the pathway for production of heterochromatic siRNAs, RDR2 is responsible for the biogenesis of the second strand of the dsRNA precursor of hc-siRNAs (Xie et al., 2004). The $rdr2_{mop1}$ mutant impairs this process and thus largely or completely eliminates the production of hc-siRNAs (Nobuta, Lu, Shrivastava, & Pillay, 2008).

Following the identification of $rdr2_{mop1}$, several other mutants were identified in maize that are deficient in paramutation, all of which encode various proteins involved in the heterochromatic siRNAs biogenesis and/or silencing pathways. In *Arabidopsis*, *NRPD1* encodes the largest subunit of DNA-dependent RNA polymerase

IV (Pol IV) (Luo & Hall, 2007), and *NRPD2/E2* (*NUCLEAR RNA-DEPENDENT RNA POLYMERASE IV/V SUBUNIT 2*) encodes the shared second largest subunit of Pol IV and Pol V (Pikaard, Haag, Ream, & Wierzbicki, 2008). Pol IV is responsible for the transcription of the first strand of hc-siRNA precursor molecules, so impairing the maize ortholog of *nrpd1* known as *rmr6/mop3* (the dual name resulting from the identification of the same gene in both screens, rmr from the Hollick lab and mop from the Chandler lab) prevents the accumulation of 24-nt hc-siRNAs, similar to $rdr2_{mop1}$ (Erhard et al., 2009).

The *NRPD2/E2* gene is important to both Pol IV and V. Pol IV is instrumental in the biogenesis of hc-siRNAs while Pol V is instrumental in the de novo methylation of hc-siRNA targets, as mentioned above. With a mutation expected to simultaneously impede the function of two key proteins in this pathway, the anticipated effect of a maize *nrpd2/e2* mutation would be a significant of a reduction of hc-siRNAs as the other mutants deficient in paramutation. Indeed, accumulation of hc-siRNAs was diminished in maize *nrpd2/e2* mutants, but interestingly, the phenotypic defects of these mutants are not nearly as severe as $nrpd1_{rmr6}$ mutants. The hypothesized reason for the lack of severe phenotypic defects is due to the presence of three presumed functional copies of the *NRPD2/E2*-like genes in maize. Both groups that have published *nrpd2/e2* mutants contain mutations in the same, and only one, *NRPD2/E2*-like gene (Sidorenko et al., 2009; Stonaker et al., 2009). If these genes were fully redundant, loss-of-function mutants would be needed for all copies to see mutant effects. However, the fact that these phenotypic defects and loss of 24-nt hc-siRNAs is observed suggests that these gene copies are not entirely redundant (Pikaard & Tucker, 2009).

1.4 Phased, Secondary, Small Interfering Small RNAs (phasiRNAs)

1.4.1 Overview of phasiRNAs

Phased, secondary, small interfering small RNAs (phasiRNAs) are named for their biogenesis, the process which yields a precise head-to-tail arrangement of sRNAs via Dicer processing, starting from the site of cleavage by a miRNA-bound-AGO (Borges & Martienssen, 2015; Fei, Xia, & Meyers, 2013). This pathway is potentially powerful because it represents the amplification of sRNAs, from one miRNA to many phasiRNAs. The well-studied 21- and 22-nt *trans*-acting small interfering RNAs (tasiRNAs) are a good case study (Fei et al., 2013). tasiRNAs were first identified in *Arabidopsis* as secondary products of AGO1/7-bound miRNAs targeting a set of long, non-coding RNAs (lncRNAs) that turned out to be precursor mRNAs. There are four families of tasiRNA-producing loci (*TAS* genes) in *Arabidopsis*, categorized into two types based on differences in biogenesis, known as "one-hit" and "two-hit". In the "one-hit" model, a single targeting site at the 5'-end of the *TAS1*, *TAS2*, or *TAS4* precursor is targeted and sliced by a 22-nt AGO1-bound-miRNA, converted into dsRNA, and diced in a precise head-to-tail arrangement from 5' to 3' by DCL4. In the "two-hit" model of *TAS3*, two target sites for miRNA-bound-AGO7 are required for the processing of *TAS3* precursors, but only the target site on the 3'-end is cleaved, with the tasiRNAs generated in the 3' to 5' direction. A specific *TAS3*-derived tasiRNA, aka tasiARF, targets *AUXIN RESPONSE FACTOR* (*ARF*) genes. Misregulation of tasiARF results in variable developmental defects in several plant species, such as *Arabidopsis* (Hunter et al., 2006), rice (Liu et al., 2007), maize (Dotto et al., 2014), and tomato (Yifhar et al., 2012). The *TAS3* tasiARF is highly conserved

across land plants, and it has co-evolved with the ARF target transcripts (R. Xia, Xu, & Meyers, 2017).

tasiRNAs are so named because tasiARF functions in *trans*, thus the name emphasizes this aspect of their function. Yet, many other phased siRNAs are produced from both lncRNA "*TAS*" loci, and from protein-coding gene families, such as those encoding nucleotide-binding, leucine-rich repeat (NB-LRR)-encoding disease resistance genes, myeloblastotis (MYB) transcription factors and other genes. Thus, tasiRNAs are one subset of the larger group of loci producing phasiRNAs. Although there are only eight lncRNA-derived *TAS* loci in *Arabidopsis*, there are >25 protein-coding sources of phasiRNAs (Howell et al., 2007); in other plant species, hundreds or even thousands of phasiRNA-generating loci have been identified. One particularly interesting group are the abundant reproductive phasiRNAs that have been found to be highly enriched in maize and rice anthers. These reproductive phasiRNAs fall into two classes: miR2118-triggered 21-nt phasiRNAs, enriched in pre-meiotic anther stages, and miR2275-triggered 24-nt phasiRNAs enriched in meiotic anthers (Johnson et al., 2009; Zhai et al., 2015). The accumulation patterns of these reproductive phasiRNAs in very specific stages suggests a critical role of phasiRNAs during the anther developmental process, but more studies are required to characterize exactly what that role is, more on this in **3.5**.

1.4.2 Biogenesis of phasiRNAs

The production of reproductive phasiRNAs starts with mRNA precursors generated by RNA polymerase II (Pol II), yielding mature, polyadenylated mRNAs that are believed to be non-coding. The "one-hit" model describes most cases of phasiRNA biogenesis in which AGO utilizes a 22-nt miRNA to mediate the slicing of

a target. MicroRNAs that are 22-nt (as opposed to most 21-nt miRNAs) typically facilitate the recruitment to targets of the machinery for phasiRNA biogenesis, namely RDR6 and SUPPRESSOR OF GENE SILENCING 3 (SGS3); these proteins convert the single-stranded mRNA precursor into dsRNA. DCL4 then recognizes the dsRNA, perhaps via recruitment by AGO1 or other components, and cleaves or 'dices' the dsRNA into 21-nt phasiRNAs, working in the direction from 5' to 3', downstream of the cleaved target site. Following the conversion to dsRNA by RDR6/SGS3, DCL4 again cleaves the dsRNA into 21-nt tasiRNAs, but operating from the 3' end of the precursor to the 5'-end (in the direction upstream of the cleaved target site). This biogenesis process takes place in the cytosol and requires polysome loading (S. Li et al., 2016).

Of the four canonical Dicer proteins encoded in a typical eudicot genome (like *Arabidopsis*), two are involved in phasiRNA biogenesis. This includes DCL1 for its activity in producing miRNA triggers, and DCL4 for its activity in producing the 21-nt phasiRNAs. However, some monocots have evolved a new Dicer family member, known as DCL5, which emerged from DCL3 at some as-yet-unknown point in the diversification of monocots (Margis et al., 2006). DCL5, formerly known as DCL3b, is the Dicer enzyme that produces the meiotic 24-nt reproductive phasiRNA in grass anthers (Song, Li, et al., 2012). Future work may reveal where and how in the monocot lineage this gene duplication event yielded a novel participant in the production of phasiRNAs.

1.4.3 The Function of Reproductive phasiRNAs

Though reproductive phasiRNAs in maize are quite diverse and numerous, their functional roles are not well elucidated by experimental analysis. However, one

can speculate as to their function based on the activities of plant sRNAs of similar lengths; in other words, 21-nt tasiRNAs may function in post-transcriptional control of targets, while 24-nt hc-siRNAs may direct chromatin modifications at their target loci. In each case, sRNAs could be incorporated into the RNA-induced silencing complex (RISC), with sequence homology directing the silencing of their corresponding targets. Bioinformatics studies have not yet described targets other than in *cis* – the *PHAS* loci themselves, so one possibility is a self-regulatory circuit. One such analysis in rice panicle sRNAs has illustrated the molecular basis that 21-nt phasiRNAs can target in cis to regulate their own precursors (Tamim et al., 2018). The observation of elevated DNA methylation at *PHAS* loci during zygotene in maize anthers provides evidence for this hypothesis (Dukowic-Schulze, Liu, & Chen, 2017). In this context, reproductive 21- and 24-nt phasiRNAs may be functionally analogous to miRNAs in fission yeast or Piwi-interacting RNA (piRNA) in animals - both of which are specific to the process of meiosis. More experimental evidence will be needed to validate this theory. With the advent of site-directed mutation enabled by the technology called Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), it is possible to knock out the key components of the biogenesis pathway to study alterations such as to chromatin structure during meiosis in the absence of phasiRNA activity (**3.5**).

1.5 miRNAs

1.5.1 Overview of Plant miRNAs

MicroRNAs (miRNAs) are a class of small non-coding RNAs in eukaryotes that regulate gene expression via posttranscriptional gene silencing. In plants, miRNAs have been identified to regulate key processes such as development, growth,

and stress response (Budak & Akpinar, 2015). Plant miRNA biogenesis and miRNA-induced silencing have been studied extensively in *Arabidopsis*, and the biogenesis pathway is well conserved in all other plants that have been examined. This biogenesis begins with the transcription of non-coding RNAs by Pol II from miRNA genes (MIRNAs) that were often not annotated in early versions of many plant genomes, since gene-finding software may fail to find lncRNAs. Once a *MIRNA* gene is transcribed by Pol II, the resulting transcript matures as a typical mRNA, with addition of a 5' 7-methylguanosine cap and 3' polyadenylated tail. This product, deemed a primary miRNA transcript (pri-miRNA), folds back onto itself to create the classical hairpin structure of a pri-miRNA. This pri-miRNA is then recognized by DCL1, a member of the Dicer-like family of enzymes, perhaps by recruitment by other protein partners (Budak & Akpinar, 2015; Rogers & Chen, 2013). Through structural features of the pri-miRNA, DCL1 can identify a location near one end of the pri-miRNA for cleavage into a precursor miRNA (pre-miRNA). DCL1 carries out a second and final cleavage of the pre-miRNA to produce the miRNA-miRNA* duplex, a short dsR-NA in which one strand produces the mature miRNA guide and the other produces the passenger miRNA (miRNA*). This dsRNA is characterized by a 2-nucleotide 3' overhang on both strands (Budak & Akpinar, 2015).

The processing of miRNAs takes place in the nucleus, however mature miRNAs must be transported to the cytoplasm prior to functioning in their role in the suppression of protein production from target mRNAs. This transfer of miRNA from the nucleus to the cytoplasm is not fully understood yet. However, data have shown that HASTY (HST), an exportin-related protein, functions as a key enzyme involved in this step (Park et al. 2005), though it has been observed that some miRNAs can still

accumulate in the cytoplasm in hst mutants. To protect the miRNA-miRNA* duplex from degradation via uridylation, it is stabilized with a 3' terminal methyl group by HEN1. It is unclear if this occurs in the nucleus or the cytoplasm (Budak & Akpinar, 2015; Park, Wu, Gonzalez-sulser, Vaucheret, & Poethig, 2005). Once in the cytoplasm, the mature miRNA guide separates from the passenger miRNA and is loaded into an Argonaute protein, typically AGO1, to form the RISC. This RISC, which is comprised of numerous proteins, uses the mature miRNA guide as a template to identify targets to either cleave a mRNA or facilitate translational inhibition (Budak & Akpinar, 2015). Targets of miRNAs are identified by their near perfect complementarity to the miRNA. When suppressing a target via mRNA cleavage, the target site of a miRNA is found on the mRNA while the miRNA is still loaded in the RISC. In plants, cleavage is typically observed between the 10th and 11th positions of the alignment and is facilitated by AGO1 (Huntzinger & Izaurralde, 2011). Products of this cleavage are detectable using Parallel Analysis of RNA Ends (PARE) libraries (German, Luo, Schroth, Meyers, & Green, 2009). Combining these PARE data with miRNA-target prediction tools, such as *sPARTA* (Kakrana et al., 2017) and CleaveLand (Addo-Quaye, Miller, & Axtell, 2009), can validate predicted targets of miRNAs on a genome-wide basis.

Upon miRNA-directed cleavage of a target mRNA, the cleaved mRNA molecule is no longer competent to produce protein products. Disrupting a miRNA would therefore typically result in the over-accumulation of miRNA target gene products, relative to wildtype controls. This phenomenon is visible in numerous maize mutants, some of which were identified long before miRNAs were described. In many plants, miRNAs exist in large numbers (hundreds of distinct loci) and can regulate the

production of numerous proteins. In the following paragraphs, we explore a few cases of maize miRNAs and targets that have been particularly well characterized.

1.5.2 Mutations in miRNA Biogenesis and Silencing Pathways

1.5.2.1 fuzzy tassel

The *fuzzy tassel* (*fzt*) phenotype was identified in EMS-mutagenized plants and it was later shown to result from a mutation in the maize orthologue of *DCL1* (Thompson et al., 2014). The *fzt* mutant plants have numerous developmental and vegetative defects, including the following: reduced plant stature, reduced number of leaves, missing and shorter internodes, reduced leaf surface area, and complete sterility in both ear and tassel due to various developmental defects in both tissues (Thompson et al., 2014). This mutation was mapped, cloned and sequenced, showing that the *fzt* mutation is a single base pair mutation in the RNase IIIa domain of *DCL1*, a domain critical in the cleavage of both pri-miRNA and pre-miRNAs (Thompson et al., 2014). The differential expression analysis of miRNA abundance and their targets supports the hypothesis that the pleiotropic effects of this mutation are caused by the impacted miRNA abundances. Additionally, it was observed that not all known miRNAs had their abundances reduced; it appears that some miRNAs are only moderately affected while others are entirely unaffected (Thompson et al., 2014). The basis of the differential impact on miRNAs is unclear, but could result from variation in how the altered DCL1 protein interacts with different precursors.

1.5.2.2 Corngrass1

Maize *Corngrass1* (*Cg1*) was a mutant first identified in 1947 and named for its closer resemblance to other grasses than to wild type maize (Whaley and Leech

1950). This mutation, deemed a ‘macromutation’ for its profound effect in mak-ing the organism unrecognizable as belonging to its species (Chuck et al. 2007), results in the prolonged juvenile stage of maize plants. During this transition phase of plant development, called heteroblasty, there are numerous changes in leaf morphology and meristem fate. In 2007, researchers from Sarah Hake’s group identified that the *Cg1* mutant encodes a tandem duplicate of miR156 genes, zma-miR156b and zma-miR156c, and the defects in the mutant are due to an overexpression of these miRNAs (Chuck et al. 2007). An analysis of transcript levels was conducted to investigate the expression levels of known targets of zma-miR156 and seven of the 13 predicted target genes, including *tga1*, a gene known to have played a role in the domestication of maize from teosinte, were found to be downregulated in *Cg1* (Chuck et al. 2007).

1.5.2.3 Glossy15

Similar to the *Corngrass1* mutation, the *Glossy15* (*Gl15*) mutation extends the duration of the juvenile-to-adult phase transition in maize; this occurs via the overexpression of the *gl15* gene (Moose and Sisco 1994). Work done in *Arabidopsis* (Aukerman and Sakai 2003; Chen 2004), demonstrated that miR172 can suppress the activity of *APETELA2*, the *Arabidopsis* ortholog of *Gl15*. The *APETELA2* target sequence of miR172 is present in the *gl15* transcript, and thus it was hypothesized that a miR172 homolog could be responsible for the transition from juvenile-to-adult vegetative development in maize. Like *Arabidopsis*, miR172 was found to not be present in early shoot development; it is only during the transition from juvenile to adult that miR172 is detected, coincident with a decline in *gl15* mRNA levels. By increasing gl15 activity, vegetative and reproductive phase changes are delayed while miR172 accumulation is not substantially different from wildtype. It was later found

that the *gl15* mutation contained a mutation in the target site of miR172. This mutation prevents the negative repression of *GL15* transcripts that occurs in wildtype maize during the juvenile-to-adult transition (Lauter et al. 2005). Thus, miR172 promotes the transition to reproductive development by restricting the accumulation of *GL15*, but such a transition can only occur at a threshold of opposing activity of *GL15* and zma-miR172 (Aukerman and Sakai 2003).

1.5.3 Overview of Annotated miRNAs

miRBase (http://www.mirbase.org/) is a public miRNA sequence repository designed to store all published miRNA sequences and their associated annotations and assign consistent identifiers to each published miRNA. As of version 22 (released March, 2018), there are 84 plant species with at least one annotated miRNA and 41 plant species with at least 50 annotated miRNAs. Of all plant species in miRBase, 34 contain genome coordinates for the miRNAs and their precursors. (Dezulian, Palatnik, Huson, & Weigel, 2005; Johnson et al., 2009; Maher, Timmermans, Stein, & Ware, 2004; Thieme, Gramzow, Lobbes, & Theißen, 2011; L. Zhang et al., 2009). Via the use of next generation sequencing, novel bioinformatics tools, and unique mutants, the identification of new miRNAs is becoming increasingly easier. Such use of these tools has been previously utilized in rice in which 76 new mature miRNA sequences were identified using 62 sRNA libraries (Jeong et al., 2011).

With a centralized miRNA repository like miRBase, it is easy to access and identify miRNAs from diverse species and identify similarities between them. Many miRNA families have been found to be conserved across numerous plant species. Some miRNAs are ancient with origins dating back to mosses, but are still identifiable in any plant due to their fundamental roles in the regulation of transcription that is

required in the basics of plant life (Chávez Montes et al., 2014). Other miRNAs are a lot newer and emerged only within distinct branches of the angiosperms. By tracking the presence of miRNAs in a phylogenetic context, their evolutionary emergence can be discerned. The conserved and divergence of target sequences is also relevant to the study of miRNAs, as evidence of conservation identifies important functional roles that have been selectively maintained, while divergence can yield new regulatory pathways (Axtell, 2013).

1.6 Bioinformatics Analysis of sRNAs

With the advent of next generation sequencing technologies giving rise to rapid, low cost sequencing, many researchers have found it possible to investigate, en masse, the effects that small RNAs have on gene expression. However, due to the large amount of small RNAs and the relative novelty of them, it remains a challenge to effectively and efficiently manage such analyses. Numerous tools have been developed within the Meyers lab as well as outside to perform such analyses.

1.6.1 Select Bioinformatics Tools

1.6.1.1 ShortStack

ShortStack (Shahid & Axtell, 2014) is a comprehensive analytical tool developed by the lab of Michael Axtell with the purpose of analyzing mapped sRNA sequencing data. This tool is a general-purpose analytical tool to classify sRNAs. It requires as input little more than the sRNA data itself and a genome to which to map the sRNAs. In six steps, *ShortStack* works to annotate and quantify sRNA data generated to serve as the input file. These six steps include the following:

a. *De novo* discovery of clusters of sRNAs.

b. Quantification and phasing analysis.

c. Retrieval of genomic sequences for RNA folding.

d. Identification of hairpin structures that may qualify as miRNA precursors.

e. Annotation of hairpin-association and miRNA candidates.

f. Output of organized and summarized results.

Within each of these steps, *ShortStack* makes use of the known biology of various classes of sRNAs to characterize genes producing sRNAs, sRNA clusters, miRNA genes, and phasiRNAs (Axtell, 2013). While running *ShortStack* is generally quite simple, there are many options that may be modified by the user, so it is highly recommended to read all the information provided within the README and even try utilizing some of the test data prior to running new analyses.

1.6.1.2 sPARTA

sPARTA (Kakrana, Hammond, Patel, Nakano, & Meyers, 2014) is a tool developed by the Meyers group for the purpose of rapidly and accurately predicting and validating targets of miRNAs. As discussed previously, miRNAs function to suppress the expression of a gene by matching with high sequence complementarity to an mRNA. When the mRNA target of a miRNA is identified, it is cleaved by AGO1. This cleavage is identifiable via the sequencing of PARE data, mentioned above. By predicting the locations at which a miRNA directs cleavage and by identifying a PARE signal at that same location, cleavage of an mRNA facilitated by a miRNA can be validated. *sPARTA* first predicts targets of miRNAs and then utilizes PARE libraries to validate these miRNA-target interaction (Kakrana et al., 2014). Several alternatives to *sPARTA* have been described, including *CleaveLand* (Addo-Quaye et

al., 2009), *PAREsnip* (Folkes et al., 2012), and *SeqTar* (Zheng, Li, Sunkar, & Zhang, 2012).

1.6.1.3 PHASIS

PHASIS is a suite of tools that functions to rapidly identify *PHAS* loci in plants. This tool, also developed in the Meyers group, has three separate components to discover, annotate, and quantify *PHAS* loci, as well as identify the miRNA triggers for those *PHAS* loci. These components – phasdetect, phasmerge, and phastrigs – each have distinct purposes in the pipeline and were separated to allow users some flexibility in the inputs provided to each. While not required, utilizing PARE data in a *PHASIS* run will allow for experimental validation of predicted cleavage locations, assisting with the identification of miRNA triggers (Kakrana et al., 2017). Some alternatives to *PHASIS* are *PhaseTank* (Guo, Qu, & Jin, 2015) and *unitas* (Gebert, Hewel, & Rosenkranz, 2017).

1.6.1.4 miRBase

As previously mentioned, miRBase is a public miRNA sequence repository designed to store all published miRNA sequences and their associated annotations, as well as assign consistent identifiers to each published miRNA. Consistency amongst identified miRNA names allows comparisons across species. Despite the great utility of miRBase, it is not without issues. For example, version 21 was published in June 2014, and went untouched and unrevised for over four years, when version 22 was published. During this time gap, however, many miRNAs were identified and submitted to miRBase, and even published in peer-reviewed articles, but the lack of updates kept those newly identified miRNAs unnamed or unassigned.

1.7 Research Objectives

The advent of next generation sequencing technologies has provided researchers the access to rapid, low cost sequencing. Over the last decade, much research has been conducted in the realm of small RNAs which resulted in the identification of numerous classes of sRNAs, and the functions of many of these. However, due to the large amount of sRNAs and the relative novelty of them to scientists, it remains a challenge to effectively and efficiently manage the analyses in which we study them. It is the responsibility of bioinformaticians to work with biologists to ensure these large datasets are properly analyzed. Often times, however, the tools that are developed for a single project can be repurposed and reused on future projects. Thus, the goal of this dissertation is to present several bioinformatics tools that I have developed to investigate small RNAs in plants, and showcase several instances in which these tools have been utilized to answer specific biological questions.

The first part of my research focuses on a pipeline that I developed to predict miRNAs in plants (**Chapter 2**). This pipeline utilizes two separate tools that can function independently for separate analyses, but when utilized together, can provide high quality predictions of miRNAs for a given dataset. The first tool in this pipeline, *miRador*, is a standalone miRNA prediction program that I developed to predict miRNAs in a set of small RNA libraries. This program utilizes a newly described set of criteria for annotating plant miRNAs with the intent of preventing false positive miRNA annotations. The second tool in this pipeline, *sPARTA*, is a tool that I co-wrote with Atul Kakrana with the purpose of predicting and validating targets of miRNAs. This tool utilizes computational means to predict targets of a set of user provided miRNAs, and then utilizes Parallel Analysis of RNA Ends (PARE) data to identify

evidence of cleavage of mRNAs at the predicted sites of cleavage. When utilized together, this pipeline can produce high quality miRNA predictions.

The second part of my research discusses a Meyers lab-specific bioinformatics tool that I developed with the purpose of analyzing clusters of small RNAs, with a particular focus on phasiRNAs (**Chapter 3**). These tools were heavily utilized in two separate projects that are presented as case studies. The first case study investigates the impact that mutations in *DCL5* have on the production of 24-nt phasiRNAs, and any potential impact the removal of 24-nt phasiRNAs has on the expression of genes in maize anthers. The second case study investigates the biogenesis of phasiRNAs in asparagus, with an emphasis on the genomic features that these 24-nt phasiRNA producing loci overlap with.

I close my dissertation with an overview of each project. In addition to this summary, I also discuss how I envision each of these tools that I have developed to be utilized after my departure from the Meyers lab. My primary goal is for these tools to be utilized extensively in the future. For such a goal to be achieved, I had to develop these tools to not just be easily used, but also easily understood and easily modified by future developers. I will discuss some coding principles that I utilized as I developed these tools, and some practical project ideas that could be implemented in the future.

Chapter 2

A NOVEL AND HIGH PERFORMANCE APPROACH FOR PREDICTING AND VALIDATING NEW miRNAs

In this chapter, I discuss the development of a novel miRNA prediction tool that I developed, *miRador*. I open this chapter by explaining the motivation for developing this tool. After, I walk through the algorithm of this tool while highlighting a few components that makes it extremely fast relative to other miRNA prediction tools. I then compare the predictive power and runtime of *miRador* to other miRNA prediction tools. I also propose the use of *sPARTA*, a miRNA-target prediction and validation program that I co-developed, in conjunction with *miRador* to provide high quality miRNA predictions. I close this chapter with the identification of four new miRNAs in maize that are not yet present in miRBase.

2.1 Criteria for plant miRNA annotation

As discussed previously, miRNAs regulate gene expression via posttranscriptional gene silencing. Their significance in gene regulation makes them important to study, thus a set of standards were created to allow for the accurate annotation of miRNAs in the organisms in which they were identified. The first effort to define standards for plant miRNA annotation was published in 2003. These criteria primarily relied on the combination of evidence of expression and biogenesis (Ambros et al., 2003).

Evidence of expression include:

- Accumulation of candidate miRNA in gel blots

- Identification of a candidate miRNA in a library of cDNAs made from size-fractionated RNAs

Evidence of biogenesis include:

- Prediction of a fold-back precursor, and the candidate sequence lying on an arm of that hairpin
- Conservation of the candidate miRNA and its predicted precursor secondary structure
- Detection of increased precursor accumulation in *dicer* mutants

While these requirements can be seen as some form of evidence for a candidate miRNA's status, many of these, on their own and by today's standard, should not be viewed to truly validate a candidate miRNA. In particular, the accumulation of a candidate miRNA does nothing to differentiate it from the other many classes of sRNAs, conservation of a known miRNA assumes that the sequence is required to be a miRNA in the newly studied organism and that the original annotation was correct, and reduced accumulation in *dicer* mutants ignores the partial redundancy observed in plant *DCL* genes (Axtell & Meyers, 2018).

In 2008, another community effort was made to redefine the standards of miRNA annotation in order to prevent false-positive miRNA annotations with newly arriving deep sequencing technologies. This criteria required that a candidate miRNA can only be computationally validated if a miRNA:miRNA* duplex could be identified on a hairpin precursor (Axtell & Meyers, 2018; Meyers et al., 2008). These requirements could be summarized into eight specific rules (**Table 2.1**). It is these criteria that served as the basis of numerous first generation plant miRNA prediction tools.

In 2018, Michael Axtell and Blake Meyers updated their plant miRNA annotation criteria to reflect the changes in understanding of plant miRNAs and their biogenesis (**Table 2.1**). The motivation for this update was to utilize our increased understanding of miRNA biogenesis to further minimize false positive miRNA annotations with stricter annotation requirements for candidate miRNAs. This update did more than just make the requirements stricter, however; some rules were relaxed to prevent false negatives in miRNA predictions (Axtell & Meyers, 2018).

Table 2.1: Former and current criteria for plant miRNA annotations

	2008 Criteria	**2018 Criteria**
11	One or more miRNA:miRNA* duplexes with two-nucleotide 3′ overhangs	Add requirements that exclude secondary stems or large loops (larger than 5 nucleotides) in the miRNA:miRNA* duplex and limit precursor length to 300 nucleotides
22	Confirmation of both the mature miRNA and its miRNA*	Disallow confirmation by blot; sRNA-seq only
33	miRNA:miRNA* duplex contains ≤4 mismatched bases	Up to five mismatched positions, only three of which are nucleotides in asymmetric bulges
44	The duplex has at most one asymmetric bulge containing at most two bulged nucleotides	Up to five mismatched positions, only three of which are nucleotides in asymmetric bulges
55	≥75% of reads from exact miRNA or miRNA*	Include one-nucleotide positional variants of miRNA and miRNA* when calculating precision
66	Replication suggested but not required	Required; novel annotations should meet all criteria in at least two sRNA-seq libraries (biological replicates)
77	Homologs, orthologs, and paralogs can be annotated without expression data, provided all criteria met for at least one locus in at least one species	Homology-based annotations should be noted as provisional, pending actual fulfillment of all criteria by sRNA-seq
88	miRNA length not an explicit consideration	No RNAs <20 nucleotide or >24 nucleotides should be annotated as miRNAs. Annotations of 23- or 24-nucleotide miRNAs require extremely strong evidence.

2.2 Overview of existing miRNA prediction tools

2.2.1 *miRDeep-P*

2.2.1.1 *miRDeep-P*

miRDeep is a miRNA prediction tool developed to predict miRNAs in animals (Friedländer et al., 2008). Due to the significant differences in miRNA precursors, however, this tool was unable to be directly used to predict miRNAs in plants. For that reason, *miRDeep-P* was developed with modified plant-specific parameters for the identification of novel miRNAs. The core of *miRDeep* was modified to predict plant miRNA precursors, and these precursors were filtered further utilizing the 2008 annotation criteria (Yang & Li, 2011). As the first computational tool for predicting plant miRNAs from next generation sequencing data, *miRDeep-P* set out with a leg up over eventual competitors and eventually emerged as one of the most highly cited plant miRNA prediction tools.

2.2.1.2 *miRDeep-P2*

In early 2019, *miRDeep-P2* was published as an update to *miRDeep-P* with the intention of improving miRNA predictions. Much of the core of *miRDeep-P* remains, but the scoring parameters were updated to utilize the updated 2018 plant miRNA annotation criteria. In its publication, *miRDeep-P2* is advertised to be superior to other plant miRNA prediction tools in terms of both speed and accuracy of its predictions (Kuang, Wang, Li, & Yang, 2018).

2.2.2 *ShortStack*

ShortStack is a tool that was developed to annotate and quantify small RNA genes. As *ShortStack* identifies these genes, it assesses these genes for their quality as

both *PHAS* and *MIRNA* genes. Thus, miRNA prediction is just one of many functions of *ShortStack*. The first version of *ShortStack* (0.1.0) was published in 2013 and has been consistently updated since its initial publication. Version 3.8.5, the current version of *ShortStack*, was published in January 2018. The maintenance of this code results in a tool that steadily improves over time, but it also means that the initial publication no longer reflects what the code actually does. The *ShortStack* manual indicates that its current version makes a strong effort to prevent or reduce false positives by being highly restrictive of the miRNAs that it predicts, but it does so at the expense of false negatives.

2.3 Limitations of existing miRNA prediction pipelines

A general purpose of miRNA prediction tools is to identify novel miRNAs in a variety of organisms, both well studied and novel. A shortcoming of both *ShortStack* and *miRDeep-P2*, as well as other miRNA prediction tools not mentioned here, is that their output does not provide users with insights as to the novelty of the predicted miRNAs. Specifically, these tools do not indicate to the user if the predicted miRNA is novel or known. Further insights into the novelty of a miRNA could be provided, specifically to highlight if a novel miRNA prediction might be a new member of an existing miRNA family.

Additionally, both of these miRNA prediction tools are advertised as being quick and accurate, and while we will observe that the latter is generally true, I believed that miRNA predictions could be performed faster. For this reason, I decided to develop a novel plant miRNA prediction tool. This tool, *miRador*, is faster than other plant miRNA prediction tools, as accurate as the others, and it provides users with important conservation insights into its predicted miRNAs that other tools do not

provide. In conjunction with *sPARTA*, a target prediction and validation tool that I co-developed, it can generate high quality miRNA predictions in a variety of plant organisms.

2.4 Description of the *miRador* miRNA prediction algorithm

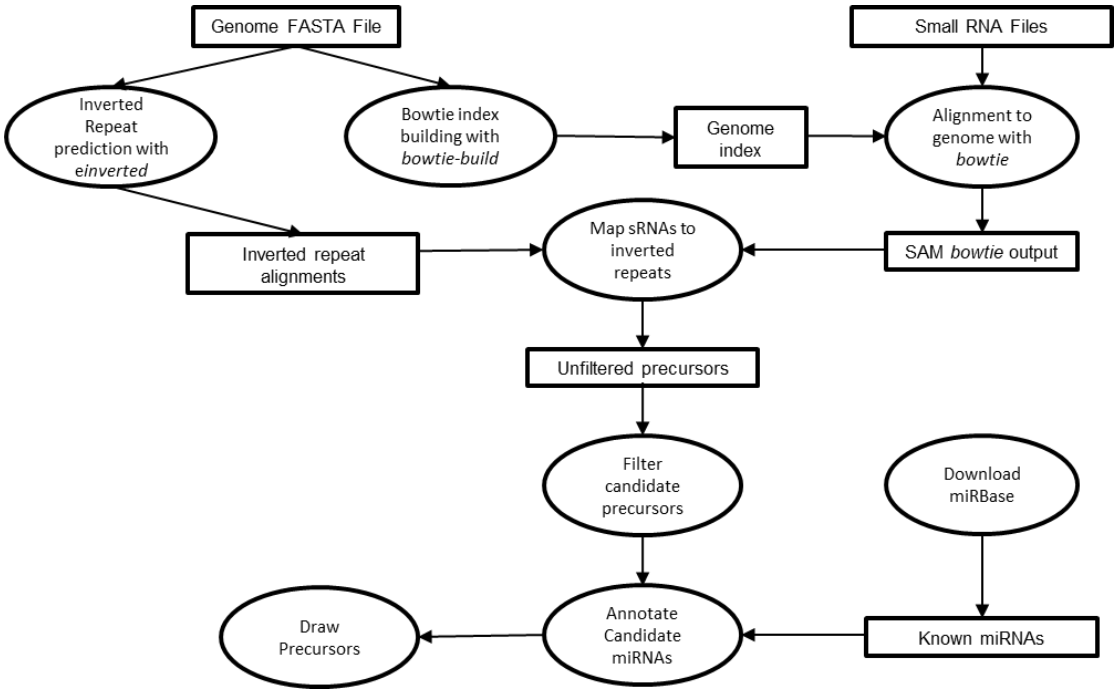

Figure 2.1: Pipeline of *miRador*
miRador requires two sets of input files: 1) Genome FASTA file and 2) sequenced small RNA files. Processing begins with the prediction of inverted repeats on each chromosome with *einverted*. Small RNA files are then mapped to the genome so that they can be mapped to the inverted repeats which are the initial set of "candidate precursor miRNAs." These are then filtered utilizing the 2018 plant miRNA annotation criteria. Finally, these candidate miRNAs are annotated utilizing known plant miRNAs from miRBase, and replication requirements are enforced in the event that a candidate miRNA is novel.

2.4.1 Initializing a preliminary set of candidate miRNA genes

Upon beginning a miRNA prediction run with *miRador*, it will utilize the user-provided genome file to identify inverted repeats within each chromosome with *einverted* (Rice, Longden, & Alan, 2000). *einverted* has several scoring parameters which can be manually set by the user, though *miRador* offers three preset options from which the user can choose to generate a series of inverted repeats. These inverted repeats will serve as the initial set of candidate precursor miRNAs. These inverted repeats are then stored into a Python dictionary for further analysis. *miRador* will then search for a bowtie index for the provided genome. If no such file exists, *bowtie-build* will be run so that sRNA libraries may be mapped to the genome (Langmead, Trapnell, Pop, & Salzberg, 2009). This step concludes the processing of genome files.

2.4.2 Counting sRNA reads and mapping sRNAs to inverted repeats

Small RNA libraries, which can be provided as FASTA, FASTQ, or tag count files, are independently processed from start to finish. The sRNA sequences are read into Python dictionaries with the sequence as a key, and count of each sequence as a value. A unique reads FASTA file is generated during this process which will be utilized when mapping to the genome. This unique reads file significantly reduces the mapping time as each sRNA read will only be processed once during mapping. Python dictionaries are heavily utilized here as these represent the implementation of hash tables in Python (**Figure 2.2**). Unlike an array, dictionaries can be queried on a string of data which makes the accession of a value extremely quick, particularly when the program does not know the index of a particular element. The average complexity of a search for a value in an array is O(n), whereas the average complexity of a search of a value in a hash table is O(1). The real world impact of this decision is reflected in the

lookup times that is handled for the millions of mapped locations for which sRNAs need to be processed.

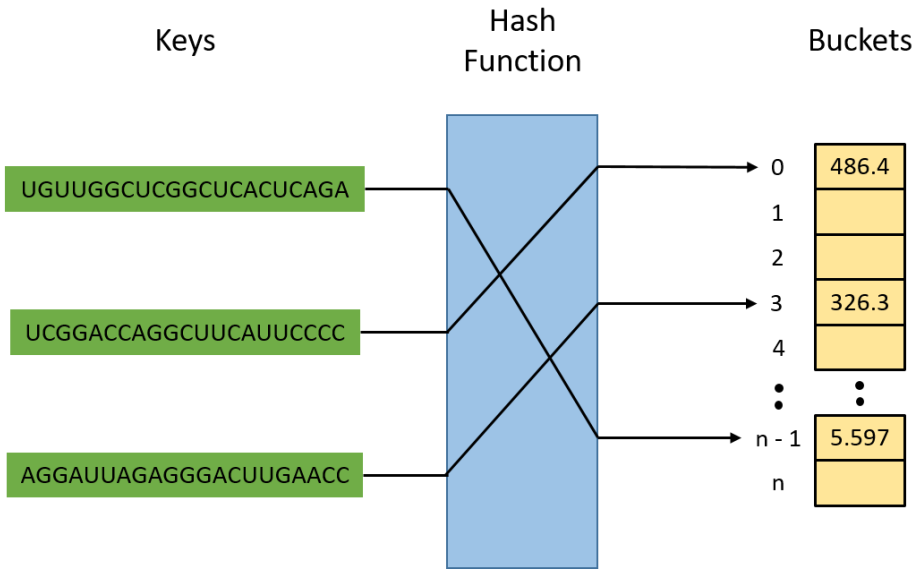

Figure 2.2: Hash table data structure
In a hash table, the key string is processed through a hash function which determines the index of the value at that key. This index is used to identify the actual value for that key, a process which is much quicker than iterating through a list to identify the value for a sequence:abundance pair. Image recreated from Wikipedia

Mapping to the genome is performed utilizing bowtie, and the mapped file is output in SAM format for immediate processing (Langmead et al., 2009). This file is parsed into nested Python dictionaries where the positional coordinates are keys allowing for rapid identification of sRNAs mapping to inverted repeats. Upon the completion of mapping, the number of mapped reads is utilized to normalize the reads. The dictionary containing the sequences and their counts are updated to normalize the counts as reads per million (RPM) mapped reads. These are also hits-normalized to account for reads mapping to numerous locations in the genome. Simply, the

abundance of a read is distributed proportionally to each mapped location in the genome.

For an inverted repeat to be considered as a candidate precursor, a small RNA must map to both arms of the precursor. Thus, *miRador* iterates through each position on each arm of the inverted repeat to identify any sRNA that lies in its entirety on an arm of the inverted repeat. Due to the reuse of the inverted repeats dictionary for each library, a separate dictionary is created to further investigate the inverted repeats for the existence of a candidate miRNA:miRNA* duplex.

2.4.3 Identifying candidate miRNA:miRNA* duplex

Under the current criteria of annotating plant miRNAs, no miRNA can be confirmed without a corresponding miRNA*. Thus, each candidate miRNA mapping to each arm of the inverted repeat is assessed to identify a potential miRNA* sequence on the opposite arm which would complete a miRNA:miRNA* duplex at this precursor. The criteria of a miRNA:miRNA* duplex are the following:

1. 2-nt 3’ overhangs on the alignment of the candidate miRNA and miRNA*
2. Up to 5 mismatched positions, only 3 of which may be nucleotides in asymmetric bulges
   - G-U pairing assessed ½ a mismatch
3. At least 75% of read abundance mapping to precursor miRNA are from 1-nt positional variants of miRNA and miRNA*

miRador assesses the alignment parameters by identifying the sRNA sequences on the *einverted* predicted inverted repeat. If the two alignment criteria are met, then each 1-nt positional variants of both the candidate miRNA and the candidate miRNA*

are identified to pool their abundances. If these abundances exceed 75% of the read abundance mapping to this inverted repeat and the candidate miRNA has an abundance of at least 3 RPM per hit to the genome, then this miRNA:miRNA* duplex will be classified as an candidate miRNA within this library and will be analyzed further in the final step.

Upon the completion of prediction for each library, each confirmed miRNA:miRNA* duplex on the precursor miRNA is drawn utilizing *RNAFold* (Kerpedjiev, Hofacker, & Hammer, 2015). It is worth noting that the alignments generated by this tool can differ from the *einverted* alignment due to differences in scoring systems of these two tools.

2.4.4 Annotating candidate miRNAs

A novel feature of *miRador* is its annotation of candidate miRNAs. There are five classifications to which a candidate miRNA might be assigned: (1) known, (2) identical to known, (3) new member of existing family, (4) conserved outside the organism of study, and (5) novel. When *miRador* starts, it will automatically download miRBase, if it has not been downloaded already. In the annotation step, each candidate miRNA is analyzed for sequence similarity to any known miRNA via the Basic Local Alignment Search Tool (BLAST) (Altschul, Gish, Miller, Myers, & Lipman, 1990; Altschul et al., 1997; Camacho et al., 2009). Several considerations are made when selecting an annotation classification for a candidate miRNA. If the species being analyzed exists in miRBase, the General Features File (GFF) is analyzed to determine the specific locations that the miRNA and its precursor have been previously identified. A miRNA identified by *miRador* can only be classified as known if the miRNA was identified at the same position that it exists in miRBase. If the sequence is

identical to a known miRNA, but not at the same location of any known miRNA, it is classified as identical to a known miRNA. A sequence will be classified as a new member of an existing family if there are 5 or fewer differences to any known miRNA in the organism of study (a difference is referred to as a gap, mismatch, or bulge). BLAST analysis may not find a similar miRNA within the organism of study, but it might find a match in another plant. If this miRNA is identified as having high similarity to a miRNA outside of this organism, it will be classified as conserved outside the organism of study. Both this classification and a new member of an existing family within the organism of study are not mutually exclusive, though a new member within the organism is given precedence and will appear first in the output file. Finally, if none of these classifications fit the candidate miRNA, it will be classified as novel.

The final step of *miRador* filters candidate miRNAs by ensuring novel miRNA families are independently predicted in multiple sRNA libraries. Upon a candidate miRNA's classification assignment in the annotation step, *miRador* will determine the number of libraries that it was predicted in if the candidate miRNA was not conserved within the organism of study. If the number of libraries predicting this candidate miRNA both exceeds 1 and exceeds 10% of the number of libraries provided for prediction, the candidate miRNA will be confirmed by *miRador*. By requiring a candidate miRNA to be present in at least 10% of libraries, we ensure that *miRador* does not predict false positives when running with a large number of libraries. Maize appears particularly sensitive to predicting what are likely hc-siRNAs as true miRNAs when using a large number of libraries for prediction, so this 10% filter is highly useful in preventing such data from appearing in the results.

Because other miRNA prediction tools do not annotate the miRNAs that they predict, and I believe this information is highly useful to users, I exported the annotation component of *miRador* to annotate the miRNAs predicted by both *ShortStack* and *miRDeep-P2*. This tool is available in a Github repository https://github.com/rkweku/mirnaAnnotation. With this tool, it is also possible to directly compare the number of miRNAs that are accurately predicted by each miRNA prediction tool to assess their prediction capabilities within each classification.

2.5 Comparison of raw predictions to other miRNA prediction tools

To assess the quality of *miRador*, its raw predictions were compared to *miRDeep-P2* and *ShortStack*. miRNAs were predicted in numerous previously published small RNA libraries from *Arabidopsis*, maize, rice, and wheat. These datasets and libraries are detailed in **Table 2.2**-**Table 2.5**

Table 2.2: List of *Arabidopsis* sRNA Libraries

Library	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads	GEO Accession Number
C0F1	Flower	6,963,781	4,877,516	2,135,334	GSM1087973
C0F2	Flower	4,204,455	2,789,465	1,158,635	GSM1087974
d1F1	Flower	6,199,680	3,574,742	2,288,645	GSM1087975
d1F2	Flower	5,989,132	3,705,660	2,077,344	GSM1087976
d234F1	Flower	3,737,661	1,417,116	1,825,447	GSM1087977
d234F2	Flower	6,213,407	3,822,319	1,610,180	GSM1087978
r2F1	Flower	5,994,423	2,195,923	3,148,385	GSM1087979
r2F2	Flower	5,518,093	2,233,522	2,207,895	GSM1087980
r2FSb	Flower	9,128,388	5,728,276	2,478,075	GSM1087981
r2FNaSb	Flower	8,972,532	5,311,047	2,583,042	GSM1087982
r2FNa1	Flower	5,028,092	1,834,656	2,771,970	GSM1087983
r2FNa2	Flower	10,101,244	6,081,068	3,108,458	GSM1087984
r2FL	Flower	3,220,911	1,145,942	1,460,780	GSM1087985

r2FD	Flower	9,605,129	4,928,142	3,492,284	GSM1087986
C0FCSb	Flower	12,306,066	9,399,678	2,566,209	GSM1087987
C0FSb	Flower	11,757,131	9,361,008	2,205,593	GSM1087988
AtCM	Flower	26,084,301	14,488,208	3,312,214	GSM2580302
hen1-8	Flower	11,605,075	5,456,831	3,633,743	GSM1502980
hen1-8 ntp2-1	Flower	12,926,853	6,244,400	3,792,504	GSM1502981
hen1-8 urt1-1	Flower	12,009,796	6,081,292	3,513,881	GSM1502982
hen1-8 ntp4-1	Flower	9,758,076	4,626,878	3,154,794	GSM1502983
hen1-8 ntp5-1	Flower	12,846,863	5,676,073	4,450,328	GSM1502984
hen1-8 ntp6-1	Flower	16,767,218	4,328,718	9,268,517	GSM1502985
hen1-8 ntp7-1	Flower	9,359,591	4,409,080	3,133,092	GSM1502986
hen1-8 ntp8-1	Flower	17,812,005	8,736,008	4,745,988	GSM1502987
hen1-8 ntp10-1	Flower	13,706,298	6,387,459	4,133,831	GSM1502988
hen1-8 r2	Flower	5,895,651	2,874,203	1,833,095	GSM1502989
hen1-8 heso1-1	Flower	9,488,379	6,015,054	2,949,692	GSM1502990
hen1-8 mee44	Flower	7,273,135	3,435,030	2,424,062	GSM1502991
Wt_d	Flower	11,808,921	7,517,450	5,282,881	GSM2580303
Col_d	Flower	24,688,934	11,692,099	11,257,700	GSM2580304
Total		**316,971,221**	**166,374,863**	**104,004,598**	

Table 2.3: List of maize sRNA Libraries

Code	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads	GEO Accession Number
0.2Fertile_r1	Anther	12,896,366	9,791,332	1,437,952	GSM1262524
0.4Fertile_r1	Anther	35,286,904	26,740,108	4,488,851	GSM1262525
0.7Fertile_r1	Anther	34,045,564	28,303,565	2,877,943	GSM1262526
1.0Fertile_r1	Anther	34,019,136	27,917,105	3,073,765	GSM1262527
1.5Fertile_r1	Anther	37,521,389	30,354,875	2,346,580	GSM1262528
2.0Fertile_r1	Anther	35,796,937	27,621,414	3,030,006	GSM1262529
2.5Fertile_r1	Anther	55,306,867	45,996,446	4,263,066	GSM1262530
3.0Fertile_r1	Anther	26,285,048	22,409,073	2,273,035	GSM1262531
4.0Fertile_r1	Anther	40,210,297	35,006,012	2,242,685	GSM1262532
5.0Fertile_r1	Anther	29,748,543	25,965,913	1,963,130	GSM1262533
0.2Fertile_r2	Anther	21,549,901	16,437,169	1,498,768	GSM1262535
0.4Fertile_r2	Anther	36,257,470	27,116,744	1,354,233	GSM1262536
0.7Fertile_r2	Anther	41,837,825	30,925,697	1,803,833	GSM1262537

1.0Fertile_r2	Anther	37,594,678	30,416,815	1,760,781	GSM1262538
1.5Fertile_r2	Anther	27,713,840	20,963,547	1,271,104	GSM1262539
2.0Fertile_r2	Anther	48,664,038	37,454,767	2,255,382	GSM1262540
2.5Fertile_r2	Anther	50,543,592	38,626,298	2,663,562	GSM1262541
3.0Fertile_r2	Anther	28,476,786	22,526,988	1,918,500	GSM1262542
4.0Fertile_r2	Anther	27,693,417	22,530,043	2,244,141	GSM1262543
5.0Fertile_r2	Anther	27,562,227	22,923,492	1,325,922	GSM1262544
0.4Fertile_r3	Anther	25,908,576	18,513,258	4,479,972	GSM1262546
0.7Fertile_r3	Anther	22,135,681	15,832,769	3,216,514	GSM1262547
1.0Fertile_r3	Anther	29,761,645	20,960,885	3,955,800	GSM1262548
1.5Fertile_r3	Anther	26,530,207	17,964,553	2,764,339	GSM1262549
2.0Fertile_r3	Anther	23,136,153	17,117,473	2,115,537	GSM1262550
2.5Fertile_r3	Anther	24,169,553	18,703,941	2,351,628	GSM1262551
3.0Fertile_r3	Anther	23,155,446	19,135,366	2,682,982	GSM1262552
4.0Fertile_r3	Anther	23,592,086	19,363,315	3,576,235	GSM1262553
5.0Fertile_r3	Anther	23,453,249	20,943,566	1,760,722	GSM1262554
Total		**910,853,421**	**718,562,529**	**72,996,968**	

Table 2.4: List of rice sRNA Libraries

Library	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads
58N_D1	Panicle	27,118,307	19,792,069	12,807,932
58N_D2	Panicle	25,823,610	22,162,882	8,711,214
58N_D3	Panicle	22,130,339	16,257,917	8,941,271
58N_D4	Panicle	25,180,026	21,443,524	6,933,374
58N_E1	Panicle	26,272,924	20,889,470	10,929,786
58N_E2	Panicle	26,047,957	20,840,592	10,441,787
58N_E3	Panicle	27,760,572	22,730,798	9,050,095
58N_E4	Panicle	26,287,527	21,797,827	7,628,303
58N_F1	Panicle	26,089,406	19,435,326	11,950,806
58N_F2	Panicle	29,363,491	24,475,366	10,670,198
58N_F3	Panicle	28,069,676	22,871,322	9,190,458
58N_F4	Panicle	26,694,985	21,996,999	8,186,387
Total		**316,838,820**	**254,694,092**	**115,441,611**

Table 2.5: List of wheat sRNA Libraries

Library	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome	GEO Accession Number

				Matched Reads	
Tae_DR2	Spikelet Meristem	16,992,080	6,924,440	3,786,628	SRR5460939
Tae_FM1	Floret Meristem	16,644,411	6,246,273	3,659,216	SRR5460941
Tae_FM2	Floret Meristem	16,500,769	6,369,250	3,548,338	SRR5460949
Tae_AM1	Anther	19,159,860	4,911,224	2,329,600	SRR5460967
Tae_AM2	Anther	17,690,692	5,567,794	2,826,185	SRR5460972
Tae_TS1	Young Floret	13,528,592	3,314,596	1,430,903	SRR5461176
Tae_TS2	Young Floret	14,222,011	3,589,267	1,663,469	SRR5461177
Total		**114,738,415**	**36,922,844**	**19,244,339**	

miRador, *ShortStack*, and *miRDeep-P2* were run on each dataset to independently identify miRNAs. *miRDeep-P2* only performs predictions on single libraries, so its predictions across multiple runs were merged to create a single set of predicted miRNAs for all libraries within a dataset. Additionally, *miRDeep-P2* utilizes an abundance cutoff, but it bypasses this cutoff if the candidate miRNA differs by up to 1 nucleotide from any miRBase miRNA. Because miRBase miRNAs are often used as true positives when assessing the quality of miRNA prediction tools, and will be utilized in this section, this bypass was manually removed. *ShortStack* and *miRDeep-P2* miRNA predictions were also annotated utilizing the modified miRNA annotation component from *miRador* to properly annotate the predicted miRNAs of each tool. This also has the added benefit of enforcing a replication requirement for *miRDeep-P2*'s predictions as there is no utility for this built into *miRDeep-P2* due to it only processing single libraries.

The summary of the predictions made by each tool are detailed for each dataset in **Table 2.6**-**Table 2.9**. Sensitivity in each prediction set was determined by searching the provided sRNA libraries for at least 1 sequenced read of a miRBase annotated miRNA. This extremely low cutoff will result in reduced sensitivities as some weakly expressed miRNAs will be effectively impossible to predict. Additionally, we have no

way to determine the specific locus from which the miRBase miRNA sequences were produced. Thus, paralogous miRNA sequences that are not active will be expected to be identified, reducing sensitivity values even further. Regardless, each tool will suffer the same penalties in these instances, and thus their values are still comparable.

These tables utilize what I deem a naïve metric for determining predictive power of these tools where miRBase annotated miRNAs serve as true positives, and every other predicted candidate is a false positive (we will discuss a different metric in **Section 2.6.2**). In the *Arabidopsis*, maize, and rice predictions (**Table 2.6**-**Table 2.8**), the trends observed are generally the same across each tool. We observe that the number of miRNAs that *ShortStack* predicts in these organisms are much lower than the other two prediction programs. This is, in fact, an intentional feature of *ShortStack* in an effort to prevent false positives in its results. Thus, we observe that the naïve count of false positives predicted in each of these three organisms is the smallest of the three tools. However, we also observe that the sensitivity of *ShortStack* to be between 1/5 and 1/3 of the sensitivity of the other two programs across these three datasets.

The counts of miRBase miRNAs that *miRador* and *miRDeep-P2* identify are very similar in both *Arabidopsis* and maize, though the count of known rice miRNAs that *miRador* identifies does fall a bit behind *miRDeep-P2*. In *Arabidopsis*, *miRador* identified a larger number of novel miRNAs than the other two tools, but *miRDeep-P2* identifies a greater number of novel miRNAs in maize, rice, and wheat (**Table 2.9**). The analysis for wheat was not as comprehensive as the other three organisms, and this is due to miRBase not providing a version of the genome utilized for miRNA mapping. Thus, the wheat miRNAs predicted by each tool have coordinates that I was unable to match with the annotated miRBase wheat miRNAs. For this reason, I was

unable to provide sensitivity and false positive values for the predictions in this organism. Regardless, examining just the counts of miRNAs predicted within each classification, there is the same general trend that *miRDeep-P2* predicts the most miRNAs in each classification group and *ShortStack* predicts the least number of miRNAs.

Table 2.6: Predicted *Arabidopsis* miRNAs by each miRNA prediction tool
Number of miRNAs predicted by each miRNA prediction tool in a dataset of 31 *Arabidopsis* inflorescence small RNA libraries

	miRador	***miRDeep-P2***	***ShortStack***
Known	142	138	43
Identical to Known	2	6	0
New member of existing family	53	59	10
Conserved outside this organism	15	3	2
Novel	27	6	6
Total miRBase miRNAs Expressed	398	398	398
Sensitivity	0.36	0.35	0.11
False Positives - Naïve	97	74	18

Table 2.7: Predicted maize miRNAs by each miRNA prediction tool
Number of miRNAs predicted by each miRNA prediction tool in a dataset of 29 staged maize anther small RNA libraries

	miRador	***miRDeep-P2***	***ShortStack***
Known	59	64	15
Identical to Known	14	18	4
New member of existing family	44	47	10
Conserved outside this organism	2	1	0
Novel	17	95	9
Total miRBase miRNAs Expressed	260	260	260
Sensitivity	0.23	0.25	0.06
False Positives - Naïve	77	161	23

Table 2.8: Predicted rice miRNAs by each miRNA prediction tool
Number of miRNAs predicted by each miRNA prediction tool in a dataset of 12 rice small RNA libraries

	miRador	*miRDeep-P2*	*ShortStack*
Known	59	81	17
Identical to Known	3	4	2
New member of existing family	21	66	8
Conserved outside this organism	11	5	0
Novel	47	111	8
Total miRBase miRNAs Expressed	505	505	505
Sensitivity	0.12	0.16	0.03
False Positives - Naïve	82	186	18

Table 2.9: Predicted wheat miRNAs by each miRNA prediction tool
Number of miRNAs predicted by each miRNA prediction tool in a dataset of 7 wheat small RNA libraries. Due to an inability to determine the genome version used for the coordinates of the known miRNAs in miRBase, it was not possible to properly equate the predicted miRNAs with the known loci from which they are derived. Thus, our best metric for predictability in this case is "identical to known," but it should be noted that some of these miRNAs have numerous matches to the genome and due to our inability to confirm which of these are true miRNA producing loci, these numbers may be slightly inflated.

	miRador	*miRDeeP-P2*	*ShortStack*
Known	0	0	0
Identical to Known	18	43	19
New member of existing family	13	34	13
Conserved outside this organism	52	54	27
Novel	12	119	33

I was also curious to observe the amount of overlap seen by the various miRNA prediction tools. **Figure 2.3**, **Figure 2.5**, **Figure 2.7**, and **Figure 2.9** show Venn diagrams of the miRNAs predicted by each tool in *Arabidopsis*, maize, rice, and wheat, respectively. **Figure 2.4**, **Figure 2.6**, and **Figure 2.8** show Venn diagrams of the known miRBase miRNAs that were predicted by each tool in *Arabidopsis*, maize, and rice, respectively. Unlike issues seen with previous miRNA prediction tools, there is a fairly high levels of overlap between the miRNAs predicted by each tool. The vast majority of *ShortStack* predicted miRNAs, both known and all classifications, are

captured by both *miRador* and *miRDeep-P2*, except in the case of wheat (**Figure 2.9**). There is a large overlap between *miRador* and *miRDeep-P2* total predicted miRNAs, though the overlap is at a greater proportion for known miRNAs. As was observed in the previous tables, *miRDeep-P2* identifies several more miRNAs than the other two tools which does result in the majority of its predictions being unique to *miRDeep-P2*.

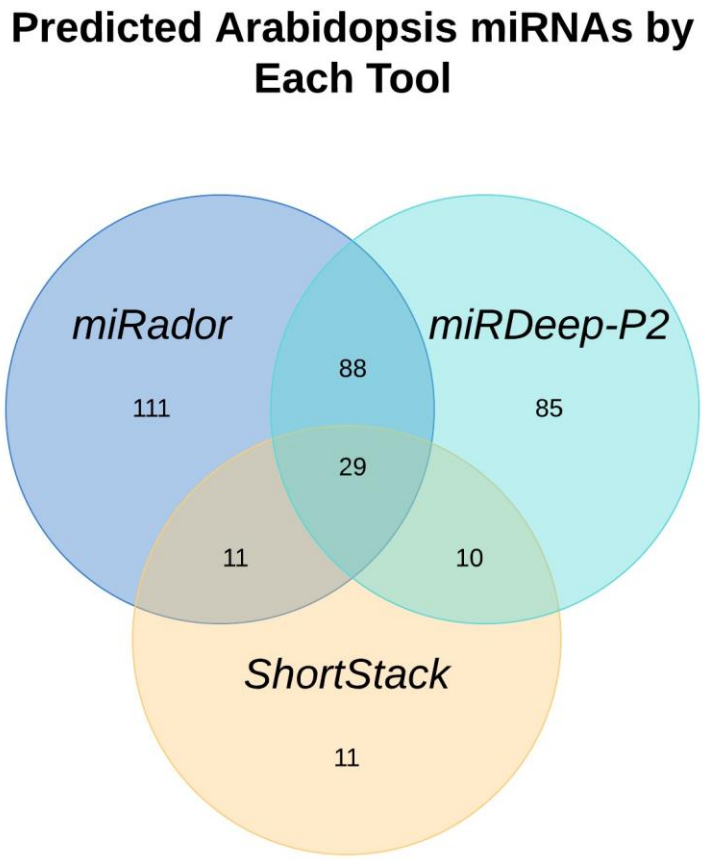

Figure 2.3: Comparison of predicted *Arabidopsis* miRNAs by each tool
Venn diagram showing the counts of *Arabidopsis* miRNAs predicted by each tool, and the counts of those miRNAs that were identified by multiple tools

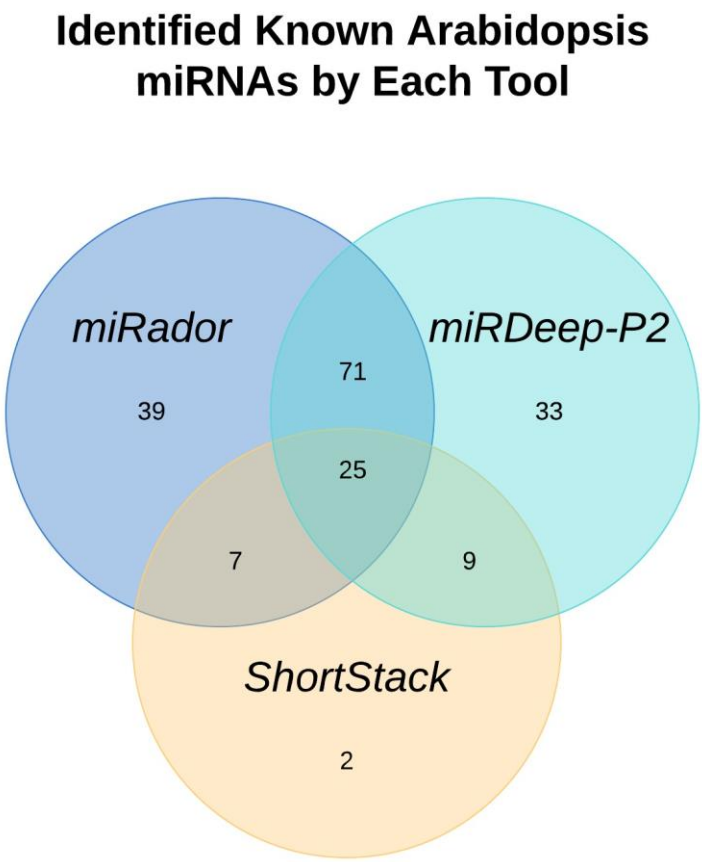

Figure 2.4: Comparison of identified known *Arabidopsis* miRNAs by each tool
Venn diagram showing the counts of known miRNAs identified by each tool, and the counts of those miRNAs that were identified by multiple tools

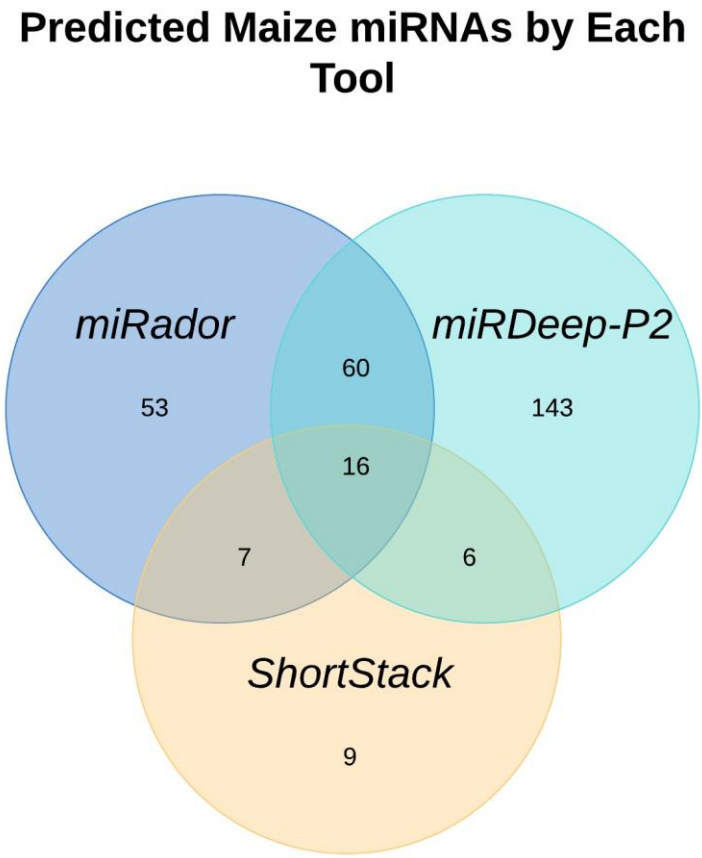

Figure 2.5: Comparison of predicted maize miRNAs by each tool
Venn diagram showing the counts of maize miRNAs predicted by each tool, and the counts of those miRNAs that were identified by multiple tools

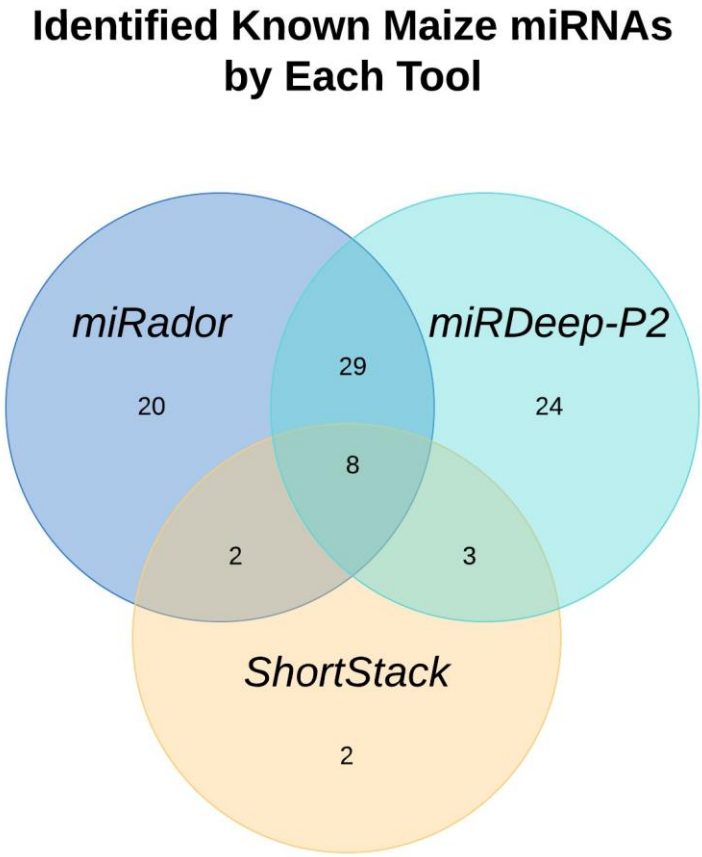

Figure 2.6: Comparison of identified known maize miRNAs by each tool
Venn diagram showing the counts of known maize miRNAs identified by each tool, and the counts of those miRNAs that were identified by multiple tools

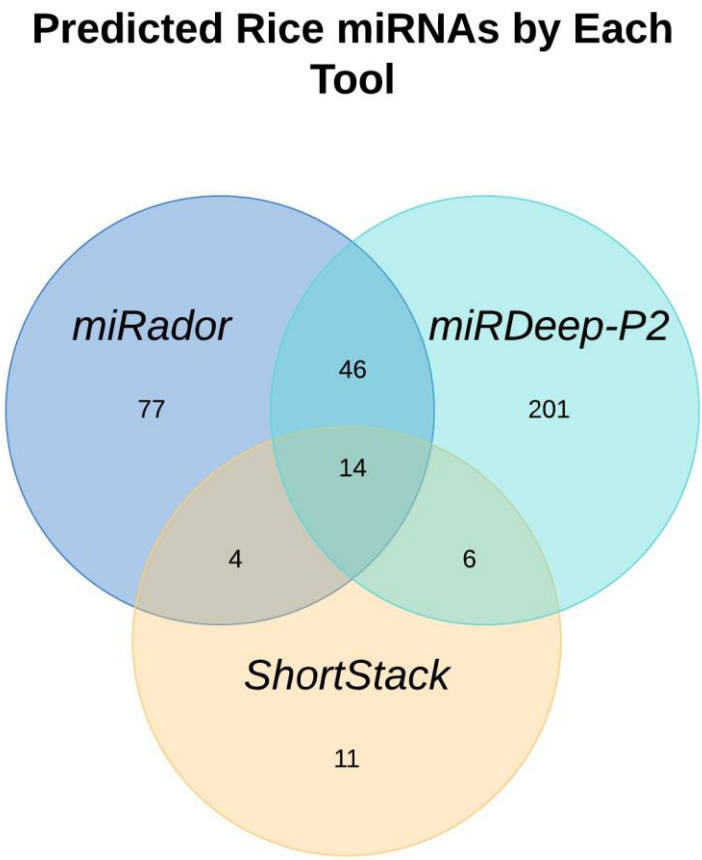

Figure 2.7: Comparison of predicted rice miRNAs by each tool
Venn diagram showing the counts of miRNAs predicted by each tool, and the counts of those miRNAs that were identified by multiple tools

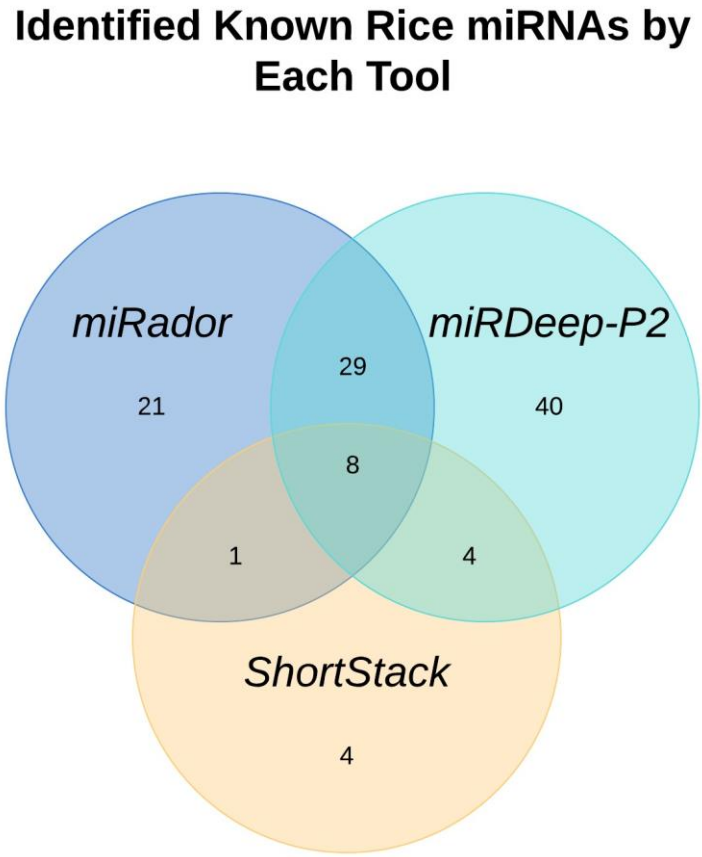

Figure 2.8: Comparison of known rice miRNAs identified by each tool
Venn diagram showing the counts of known maize miRNAs identified by each tool, and the counts of those miRNAs that were identified by multiple tools

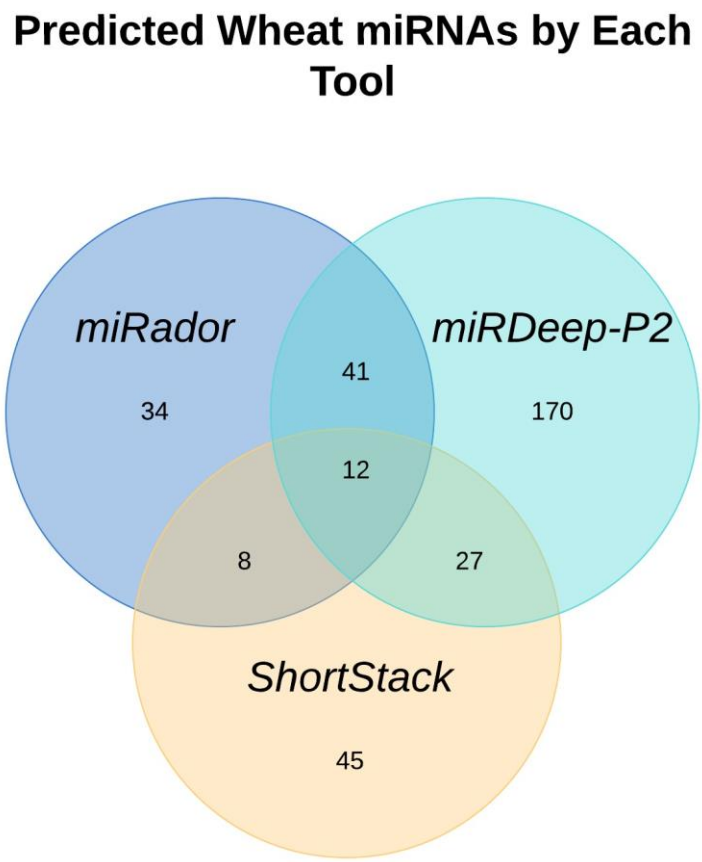

Figure 2.9: Comparison of predicted wheat miRNAs by each tool
Venn diagram showing the counts of miRNAs predicted by each tool, and the counts of those miRNAs that were identified by multiple tools

A major issue with the analysis of the predictions made by each of these tools in this section is that the true positive set includes only miRNAs annotated by miRBase. This metric, however, assumes that all of miRNAs in miRBase are in fact real, and possibly a more egregious assumption is that anything that does not exist in miRBase is a false positive. In order to better assess the quality of these tools, I decided to utilize PARE (Parallel Analysis of RNA Ends) libraries, as referenced in **Section 1.5.2,** to determine a better set of true positives within each prediction set. With this data, we can better assess the quality of the predictions made by each of the

miRNA prediction tools, especially outside of the class of known miRBase miRNAs. To do this, I made use of *sPARTA*, a miRNA-target prediction and target validation tool that I co-developed.

2.6 Utilizing *sPARTA* miRNA-target validation to validate novel miRNAs

2.6.1 Description of *sPARTA*

This section includes work that was published in *Nucleic Acids Research* (citation below). My primary contributions to this work was the development of the miRNA target validation component utilizing processed PARE data. The text in this section only features information from this paper that is relevant for the background of this chapter.

Kakrana, A., Hammond, R., Patel, P., Nakano, M., & Meyers, B. C. (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), 2973-2981. doi:10.1093/nar/gku693

Parallel analysis of RNA ends (PARE) is a high-throughput sequencing technique which profiles uncapped mRNAs, those being products of cleavage or decay, which can be utilized to study miRNA targets (German et al., 2009). *sPARTA* is a powerful tool developed by the Meyers lab to predict targets of plant miRNAs and validated those targets with PARE. In addition to searching for targets in annotated genomic regions, it can also search for targets in unannotated genomic regions, which is useful to discover novel regulatory modules, independent of genome annotations that may be incomplete. Through the use of PARE data and target validation with

sPARTA, it is possible to identify evidence of action of candidate miRNAs that have been predicted by *miRador*.

2.6.1.1 Overview of *sPARTA*

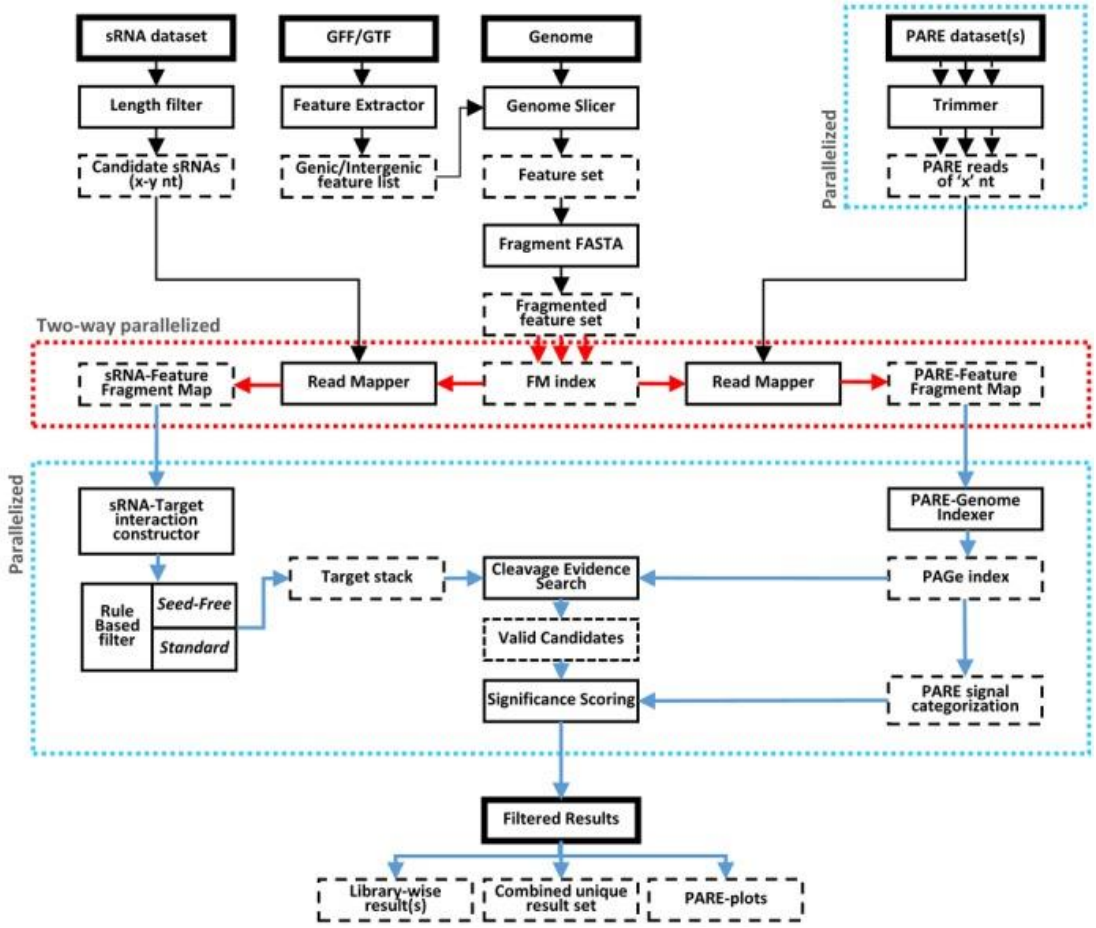

Figure 2.10: *sPARTA* schematic, showing order of steps in workflow
Solid boxes represent *sPARTA* functions, dashed boxes represent the product of an applied function. Multiple arrows indicate multiple output files from the preceding function. Steps executed in parallelized environment are enclosed within colored dotted lines.

sPARTA has four primary steps, two of which are data and time intensive. *miRferno*, the target prediction component, utilizes version 2 of *bowtie* to predict targets of miRNAs. The second component, the component that I was primarily

involved in the development of, utilizes PARE read abundances and positions to assess the validity of the *miRferno* predicted targets. The entire *sPARTA* pipeline may be viewed in **Figure 2.10**.

2.6.1.2 Target prediction with *sPARTA*

miRferno offers two prediction modes, greedy and exhaustive, which differ primarily in terms of the arguments that are passed to *bowtie2*. *bowtie2* utilizes a multiseed heuristic to narrow the number of possible alignments that must be considered which can greatly reduce its runtime and memory footprint (Langmead & Salzberg, 2012). Greedy mode utilizes a seed length of 6-nt and are extracted at 4-nt intervals. Exhaustive mode utilizes a seed length of 4-nt and are extracted at 3-nt intervals. Additionally, if no extracted seed reports an alignment, a re-seeding pass is allowed where seeds are offset and re-searched for potential targets. In both modes, matched instances of these seeds are further investigated to determine the full alignment between the miRNA and its potential target. In addition to these two modes, *miRferno* offers two different scoring systems for target scoring. Standard scoring utilizes a previously established set of rules for target scoring which emphasizes strict complementarity between a miRNA's seed (positions 2-13) and the miRNA target (Fahlgren & Carrington, 2010). The seed free scoring system does away with the notion of a strict complementarity between a miRNA seed and its target due to the recent identification of miRNA-target interactions that deviate from this strict scoring system (Brousse et al., 2014; Zheng et al., 2012). Under the seed-free scoring system, each miRNA-target alignment is scored using the following position specific rules, starting from the 5' end of the miRNA:

1. Mismatches at either the 10th or 11th positions carry a penalty of 2.5.

2. A wobble with a single flanking mismatch or mismatches on both sides carries a penalty of 1.5 or 2.0, respectively.
3. A single gap, mismatch and wobble at any position carries a penalty of 1.5, 1.0 or 0.5, respectively.

2.6.1.3 Target validation with *sPARTA*

In this step, the PARE read abundances and positions are analyzed 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 are 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 et al., 2009; Folkes et al., 2012). This schema uses five 'classes' to rank the evidence of cleavage based on abundance; 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.

Lastly, class 4 are PARE reads with an abundance of one, essentially non-discernable from ‘background’.

sPARTA calculates the confidence score (P-value) as defined in Cleaveland (v3). but with a slight modification to improve the P-value in cases where miRNA-target interactions have weak complementarity, or when a single miRNA cleaves hundreds of targets. 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 *Cleaveland3*.

probability of success = fraction of total (eligible) bases in the feature set occupied by a specific PARE signal class (Addo-Quaye et al., 2009).

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 *Cleaveland3*, 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.

2.6.2 Comparison of *miRador* PARE validated miRNAs with *miRDeep-P2* and *ShortStack*

As discussed previously, many miRNA prediction programs assess their predictive power by utilizing miRBase miRNAs as a true positive dataset. This method, however, results in real miRNAs that are not yet known as being classified as false positives, which negatively impacts the predictive score of these miRNA prediction programs. Additionally, it has been shown that several miRNAs in miRBase may be false positives (Meng, Shao, Wang, & Chen, 2012; Taylor, Tarver, Foroozani, & Donoghue, 2017; Taylor, Tarver, Hiscock, & Donoghue, 2014). To alleviate these problems, I sought to validate the action of predicted miRNAs by identifying PARE signals at *sPARTA* predicted targets rather than depending on the presence of these predicted miRNAs in miRBase. Through this method, we are able identify a set of true positives as predicted miRNAs that have a PARE signal at predicted cleavage sites with a corrected p-value ≤ 0.05. Any predicted miRNA that fails to meet this requirement is then tagged as false positive. With this data, we can calculate the positive predictive value of these tools. *Arabidopsis*, maize, and rice PARE libraries were used for this analysis utilizing the same tissues used in miRNA prediction. These datasets are detailed in **Table 2.10**-**Table 2.12**.

Table 2.10: List of *Arabidopsis* PARE libraries utilized for miRNA validation of the miRNAs predicted by each miRNA prediction tool

Code	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads	GEO Accession Number
AxIDT	Inflorescence	1,371,358	961,038	457,398	GSM278334
AxIRP	Inflorescence	7,179,831	2,456,737	2,968,918	GSM278335
Col	Inflorescence	4,711,675	2,242,387	564,328	GSM284751
TWF	Inflorescence	17,711,729	12,870,571	4,979,361	GSM280226
Tx4F	Inflorescence	10,643,828	7,513,966	5,030,109	GSM280227
Total		**798,702,483**	**621,500,243**	**14,000,114**	

Table 2.11: List of maize PARE libraries utilized for miRNA validation of the miRNAs predicted by each miRNA prediction tool

Code	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads	Accession Number
ant_1mm_1	Anther	40,197,934	30,461,899	5,341,797	
ant_1mm_2	Anther	40,836,565	31,040,252	4,934,942	
ant_1mm_3	Anther	43,623,038	32,586,288	5,266,629	
ant_15mm_1	Anther	41,335,830	31,421,753	4,417,321	
ant_15mm_2	Anther	44,661,050	33,921,566	4,757,241	
ant_15mm_3	Anther	34,987,962	27,106,460	4,823,546	
ant_2mm_1	Anther	40,291,178	31,263,978	5,592,297	
ant_2mm_2	Anther	48,670,625	37,629,197	6,064,297	
ant_2mm_3	Anther	36,212,687	28,638,538	4,425,117	
ant_25mm_1	Anther	33,773,432	26,317,098	4,808,669	
ant_25mm_2	Anther	41,864,996	32,794,371	5,098,239	
ant_25mm_3	Anther	40,149,087	31,618,951	5,180,878	
1_0w1_5	Anther	50,968,596	39,492,311	8,494,838	GSM1262610
2_0w2_5w3_0	Anther	38,853,899	29,536,603	7,099,772	GSM1262611

Table 2.12: List of rice PARE libraries utilized for miRNA validation of the miRNAs predicted by each miRNA prediction tool

Code	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads	GEO Accession Number
1ldNILs	Panicle	34,120,199	30,548,637	6,247,885	GSM2253138
1ld58Ss	Panicle	42,420,515	38,673,590	7,263,381	GSM2253139
1sdNILs	Panicle	35,826,997	32,282,105	6,305,172	GSM2253140
1sd58Ss	Panicle	35,575,232	32,607,464	6,762,129	GSM2253141
Total		**147,942,943**	**134,111,796**	**26,578,567**	

In this analysis, we generate true positives across all classes of miRNA predictions for *Arabidopsis*, maize, and rice that is independent for each tool (**Table 2.13-Table 2.15**). Of brief note, while the PARE data that was used for each of these organisms matches the tissue that was used for the miRNA predictions, the actual samples in *Arabidopsis* and rice do not align as well as they do for the maize data. Both maize sRNA and PARE data were generated from the same stages of tissue in wildtype plants whereas the *Arabidopsis* and rice libraries are the same tissue, but are not staged in the same manner, and in some cases, come from mutant plants. This appears to be reflected by the high positive predictive value identified in maize for each of the miRNA prediction tools. Several miRNAs in the class of "new members of existing family" were identified to be validated by PARE for having the same target as known miRBase miRNAs that they had high sequence similarity to. That is not to say that these candidate miRNAs aren't real, but to take another step in assessing the quality of these tools, I decided to generate positive predictive values for only the class of novel miRNAs. These miRNAs are not at all similar to any known miRNA, and thus their validating targets should be unique to them, and any other members of this novel miRNA family.

We observe strong positive predictive values for all *ShortStack* predictions (**Table 2.13-Table 2.15**). This identification remains consistent with the objective of *ShortStack* in preventing false positive annotations. This proves true even in rice where both other prediction tools have much lower positive predictive values. However, *ShortStack's* ability to identify novel miRNAs is severely impacted by its strict parameters.

miRador has the lowest positive predictive value in *Arabidopsis*, though its predictive power for the class of novel miRNAs is higher than *miRDeep-P2* (**Table 2.13-Table 2.15**). Despite this, the total number of novel miRNAs that were identified by *miRador* was surprisingly high. These candidate miRNAs will be investigated at a later time. In maize, all prediction tools had strong positive predictive values. *miRDeep-P2* predicted a large number of novel maize miRNAs, but 26% failed to be validated. This was the lowest of the three prediction tools.

Finally, the rice predictions for both *miRador* and *miRDeep-P2* were the lowest of any organism. The positive predictive value for novel miRNAs as predicted by *miRDeep-P2* was quite low, with just 52 of its predicted 111 miRNAs having a PARE signal. One hypothesis for why these values are lower than the other organisms is that the PARE resolution was much lower in rice as compared to *Arabidopsis* and maize. We only had access to four PARE libraries for this organism, and not all of these directly aligned with the sRNA data that we used to predict these miRNAs. Thus, some of these candidate miRNAs could be found to be real, but we do not have the data to show that at this time.

Table 2.13: PARE validated *Arabidopsis* miRNAs predicted by each miRNA prediction tool

Number of miRNAs predicted by each miRNA prediction tool that had at least one PARE validated target as predicted by sPARTA ($P \leq 0.05$).

	miRador	miRDeep-P2	ShortStack
Known	109	115	40
Identical to Known	1	5	0
New member of existing family	45	54	8
Conserved outside this organism	7	3	1
Novel	21	4	5
Total Predicted miRNAs Confirmed by sPARTA	183	181	54
Positive predictive value - sPARTA	76.57%	85.38%	88.52%
Positive predictive value of novel - sPARTA	77.78%	66.67%	83.33%
False Positives - Confirmed by sPARTA	56	31	7

Table 2.14: PARE validated Maize miRNAs predicted by each miRNA prediction tool
Number of miRNAs predicted by each miRNA prediction tool that had at least one PARE validated target as predicted by sPARTA ($P \leq 0.05$).

	miRador	miRDeep-P2	ShortStack
Known	56	64	15
Identical to Known	14	18	4
New member of existing family	42	45	9
Conserved outside this organism	2	1	0
Novel	14	70	8
Total Predicted miRNAs Confirmed by sPARTA	128	198	36
Positive predictive value - sPARTA	94.12%	88.00%	94.74%
Positive predictive value of novel - sPARTA	82.35%	73.68%	88.89%
False Positives - Confirmed by sPARTA	8	27	2

Table 2.15: PARE validated Rice miRNAs predicted by each miRNA prediction tool
Number of miRNAs predicted by each miRNA prediction tool that had at least one PARE validated target as predicted by sPARTA ($P \leq 0.05$).

	miRador	miRDeep-P2	ShortStack
Known	38	71	14
Identical to Known	2	3	1
New member of existing family	11	48	7
Conserved outside this organism	2	1	0
Novel	31	52	7
Total Predicted miRNAs Confirmed by sPARTA	84	175	29
Positive predictive value - sPARTA	59.57%	65.54%	82.86%

Positive predictive value of novel - sPARTA	65.96%	46.85%	87.50%
False Positives - Confirmed by sPARTA	57	92	6

The final analysis involved a comparison of the validated miRNAs identified by each tool across the three organisms (**Figure 2.11**-**Figure 2.13**). The trends identified in the analysis of total predicted miRNAs and their overlaps remain consistent here (**Figure 2.3**, **Figure 2.5**, **Figure 2.7**, and **Figure 2.9**). *ShortStack* identifies the fewest miRNAs, *miRDeep-P2* generally identifies the most miRNAs, and *miRador* is in between the two, though much closer to the count observed by *miRDeep-P2*. There is a large overlap in the miRNAs that are predicted, though *miRador* and *miRDeep-P2* have a sizeable number of candidate miRNAs that the other prediction tools do not identify. Because *miRDeep-P2* predicts the greatest number of miRNAs, the majority of its predicted miRNAs are unique to *miRDeep-P2*.

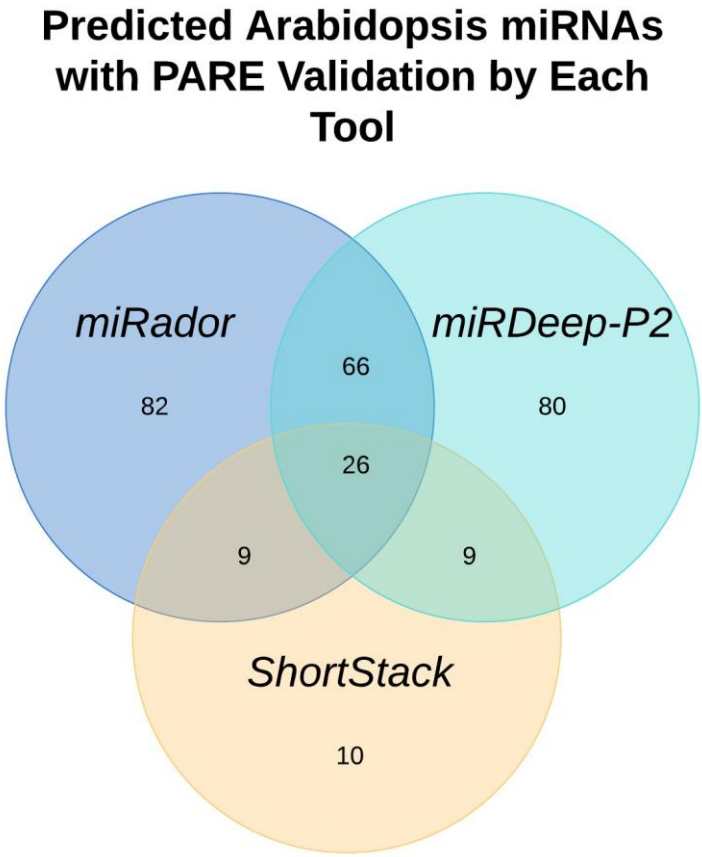

Figure 2.11: Comparison of *Arabidopsis* miRNAs that were PARE validated for each tool
Venn diagram showing the counts of *Arabidopsis* miRNAs that were predicted by each tool that could be PARE validated by *sPARTA*

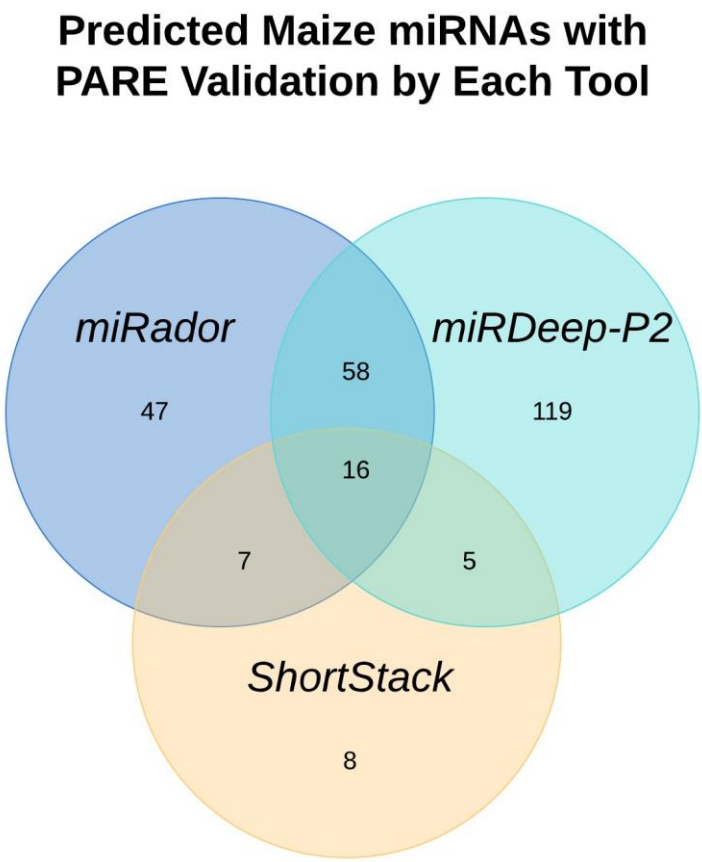

Figure 2.12: Comparison of maize miRNAs that were PARE validated for each tool Venn diagram showing the counts of maize miRNAs that were predicted by each tool that could be PARE validated by *sPARTA*

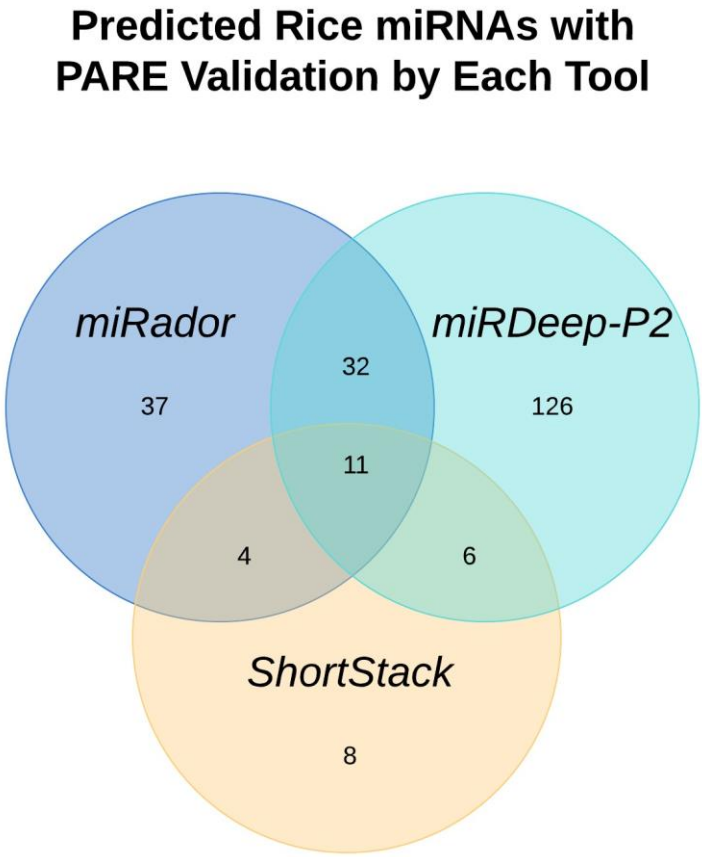

Figure 2.13: Comparison of rice miRNAs that were PARE validated for each tool Venn diagram showing the counts of rice miRNAs that were predicted by each tool that could be PARE validated by *sPARTA*

2.7 Evaluation of runtime performance

Evaluating runtime performance was determined by timing the runs of each program during the predictions performed in the previous section. Due to variability in available computational resources at any given time, especially on communal servers, I tested runtime performance utilizing three separate runs of each tool with each dataset on servers that see minimal use by other users in every case except for wheat. Due to the size of the wheat genome and the runtime its extended runtime, its results

are limited to just a single run. By utilizing multiple runs, we can identify atypical runtimes that might not be representative of a tool's runtime with a given dataset.

In this analysis, we observe that the runtime of *miRador* is significantly shorter than all other miRNA prediction pipelines across each dataset. *miRDeep-P2* exhibited prediction times between *miRador* and *ShortStack* for *Arabidopsis*, maize, and rice predictions. Curiously, the runtime of *ShortStack* in wheat was the lowest across all of the tools. It is difficult to discern exactly why this may be, but it appears that it might have to do with the command to *bowtie* that is issued for mapping of the sRNAs that is specific to larger genomes, including wheat. miRNA predictions were performed with default parameters for all prediction tools. While both *miRador* and *miRDeep-P2* do not allow mismatches in their mapping of sRNAs, *ShortStack* typically allows 1 mismatch. With the wheat genome however, *ShortStack* allows no mismatches due to a bug in *bowtie* that is specific to large genomes. With this change, the bowtie mapping of each library, a process that is cumulatively the most time intensive step for *ShortStack*, becomes extremely quick.

In analyzing the runtime of *miRDeep-P2*, I found that processing of each *Arabidopsis* library took between 10 and 15 minutes, the processing of each maize library took between 1 and 2 hours, and the processing time of each rice library took between 30 and 45 minutes. These times seem to align well with the publication. but the processing time of wheat miRNAs displayed massive differences from the published times. The published results for *miRDeep-P2* show that the runtime for predicting miRNAs in wheat to be around 2 hours. I found that each wheat library took between 6 and 12 hours to completely process. The runtimes that I identify seem to scale appropriate with the size of the provided genomes (*Arabidopsis*: 116 MB,

maize: 2.1 GB, rice: 364 MB, wheat: 14 GB). I suspect that parameters may have been made for the processing of wheat data, but I was unable to identify any indication to this in the publication. It is also possible that the differences in timing were due to the differences in the processing computers, though I suspect some underlying issue is at play due to the massive differences seen in wheat that were not observed in *Arabidopsis*, rice, or maize.

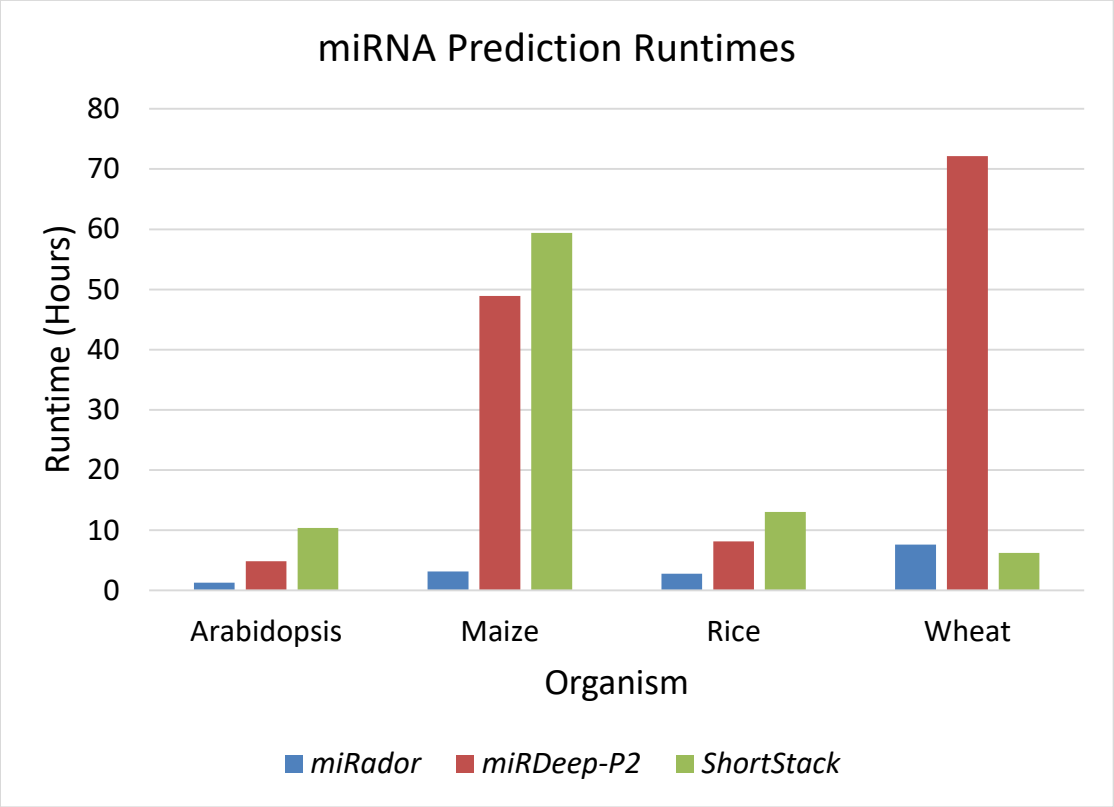

Figure 2.14: Runtime comparison of the various miRNA prediction tools across the different datasets

2.8 Identification of new miRNAs with *miRador*

The original motivation for the development of this miRNA prediction pipeline stemmed from the desire to predict novel maize miRNAs utilizing a series of sRNA

libraries provided by Marja Timmermans and Jay Hollick. These included sRNA libraries from seedling, SAM, and tassel, and these were generated from a variety of mutants including *hen1*, *dcl1*, *dcl4*, and *rmr5*. These mutant libraries have varying effects on sRNA accumulation, so we hypothesized that we could utilize impacted accumulation as further evidence of a candidate miRNA's status. By utilizing these maize data, as well as several other public maize datasets, I attempted to identify novel maize miRNAs with *miRador*. I utilized PARE data in an attempt to validate the action of these candidate miRNAs to provide further evidence of their status. The sRNA libraries used are detailed in **Table 2.16** and PARE libraries used are detailed in **Table 2.17**.

Table 2.16: List of maize sRNA libraries utilized for maize-specific miRNA identification

Code	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads	GEO Accession Number
0.2Fertile_r1	Anther	12,896,366	9,791,332	1,437,952	GSM1262524
0.4Fertile_r1	Anther	35,286,904	26,740,108	4,488,851	GSM1262525
0.7Fertile_r1	Anther	34,045,564	28,303,565	2,877,943	GSM1262526
1.0Fertile_r1	Anther	34,019,136	27,917,105	3,073,765	GSM1262527
1.5Fertile_r1	Anther	37,521,389	30,354,875	2,346,580	GSM1262528
2.0Fertile_r1	Anther	35,796,937	27,621,414	3,030,006	GSM1262529
2.5Fertile_r1	Anther	55,306,867	45,996,446	4,263,066	GSM1262530
3.0Fertile_r1	Anther	26,285,048	22,409,073	2,273,035	GSM1262531
4.0Fertile_r1	Anther	40,210,297	35,006,012	2,242,685	GSM1262532
5.0Fertile_r1	Anther	29,748,543	25,965,913	1,963,130	GSM1262533
0.2Fertile_r2	Anther	21,549,901	16,437,169	1,498,768	GSM1262535
0.4Fertile_r2	Anther	36,257,470	27,116,744	1,354,233	GSM1262536
0.7Fertile_r2	Anther	41,837,825	30,925,697	1,803,833	GSM1262537
1.0Fertile_r2	Anther	37,594,678	30,416,815	1,760,781	GSM1262538
1.5Fertile_r2	Anther	27,713,840	20,963,547	1,271,104	GSM1262539
2.0Fertile_r2	Anther	48,664,038	37,454,767	2,255,382	GSM1262540
2.5Fertile_r2	Anther	50,543,592	38,626,298	2,663,562	GSM1262541

3.0Fertile_r2	Anther	28,476,786	22,526,988	1,918,500	GSM1262542
4.0Fertile_r2	Anther	27,693,417	22,530,043	2,244,141	GSM1262543
5.0Fertile_r2	Anther	27,562,227	22,923,492	1,325,922	GSM1262544
0.4Fertile_r3	Anther	25,908,576	18,513,258	4,479,972	GSM1262546
0.7Fertile_r3	Anther	22,135,681	15,832,769	3,216,514	GSM1262547
1.0Fertile_r3	Anther	29,761,645	20,960,885	3,955,800	GSM1262548
1.5Fertile_r3	Anther	26,530,207	17,964,553	2,764,339	GSM1262549
2.0Fertile_r3	Anther	23,136,153	17,117,473	2,115,537	GSM1262550
2.5Fertile_r3	Anther	24,169,553	18,703,941	2,351,628	GSM1262551
3.0Fertile_r3	Anther	23,155,446	19,135,366	2,682,982	GSM1262552
4.0Fertile_r3	Anther	23,592,086	19,363,315	3,576,235	GSM1262553
5.0Fertile_r3	Anther	23,453,249	20,943,566	1,760,722	GSM1262554
ht_rmr5_2	Ear	27,918,595	19,295,087	7,221,909	
ho_rmr5_2	Ear	22,984,093	16,172,755	6,018,959	
Zma2	Ear	4,644,825	3,909,378	2,064,083	GSM433621
ht_rmr2_1	Ear	33,671,756	18,318,136	6,335,563	GSM913702
ho_rmr2_1	Ear	6,521,543	5,140,583	1,672,750	GSM913701
Out_ear_Unfrt	Ear	49,260,408	44,841,158	12,521,696	SRX120259
Ear1_*mop1*	Ear	3,309,542	2,476,630	607,963	SRX708787
Ear2_*mop1*	Ear	11,158,292	8,300,865	1,510,278	SRX708788
Ear1_Control	Ear	2,586,754	2,084,981	1,014,119	SRX708789
Ear2_Control	Ear	15,974,948	12,857,863	4,284,875	SRX708790
S*mop1*	Ear	7,265,148	5,250,202	1,268,157	GSM306488
Swt	Ear	5,768,313	3,618,896	1,854,497	GSM306487
MzEar	Ear	5,110,322	4,522,193	3,173,811	GSM448853
PLN_Fertile_r1	Pollen	33,201,391	28,333,867	2,322,587	GSM1262534
PLN_Fertile_r2	Pollen	33,953,600	27,712,546	1,464,303	GSM1262545
PollenFertile_r3	Pollen	20,949,777	18,715,984	1,362,243	GSM1262555
MzPollen	Pollen	8,665,587	7,067,189	3,623,531	GSM448854
ht_rmr5_2_SAM1	SAM	28,106,096	22,739,358	4,191,938	
ht_rmr5_2_SAM2	SAM	27,492,997	22,847,162	3,682,476	
ht_rmr5_2_SAM3	SAM	26,785,804	21,639,355	3,893,125	
ho_rmr5_2_SAM1	SAM	30,227,898	24,535,811	5,149,225	
ho_rmr5_2_SAM2	SAM	28,763,183	25,126,043	2,865,313	
ho_rmr5_2_SAM3	SAM	25,824,037	21,802,137	3,117,043	
WT_h1_rp1	Seedling	48,831,520	38,943,929	11,877,859	GSM1178887
WT_h1_rp2	Seedling	22,534,155	16,657,415	4,779,483	
WT_h1_rp3	Seedling	22,701,429	17,084,728	4,668,610	
hen1_rp1	Seedling	33,477,122	24,216,068	7,319,921	GSM1178886

hen1_rp2	Seedling	12,843,660	8,534,877	2,473,963	
hen1_rp3	Seedling	24,360,691	16,989,186	3,581,789	
WT_d1_rp1	Seedling	39,837,626	34,954,395	11,290,431	
WT_d1_rp3	Seedling	18,023,412	14,498,549	4,020,681	
dcl1_rp1	Seedling	31,530,325	26,636,297	10,278,404	
dcl1_rp3	Seedling	28,533,217	22,435,430	5,323,486	
WTtassel3_2cm	Tassel	29,994,483	25,829,115	6,703,922	
dcl4tassel4_0cm	Tassel	33,852,863	28,922,877	6,675,010	
WTtassel2_5cm	Tassel	33,105,010	28,407,232	6,872,976	
dcl4tassel2_7cm	Tassel	44,845,740	38,347,354	8,742,095	
Zma3	Tassel	4,207,601	3,663,843	2,026,537	GSM433622
rmr5_florets_1	Tassel	26,032,150	21,287,465	2,360,629	
rmr5_florets_2	Tassel	21,808,429	17,919,342	1,852,785	
MzTassel	Tassel	8,207,235	7,181,656	4,341,950	GSM448857
rmr5_3cm	Tassel	37,449,795	30,848,081	4,390,135	
rmr5_15cm	Tassel	25,889,323	19,325,195	2,930,347	
Total		**1,919,064,116**	**1,528,553,742**	**266,728,425**	

Table 2.17: List of maize PARE libraries utilized for miRNA validation

Code	Tissue	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads	Accession Number
ant_1mm_1	Anther	40,197,934	30,461,899	5,341,797	
ant_1mm_2	Anther	40,836,565	31,040,252	4,934,942	
ant_1mm_3	Anther	43,623,038	32,586,288	5,266,629	
ant_15mm_1	Anther	41,335,830	31,421,753	4,417,321	
ant_15mm_2	Anther	44,661,050	33,921,566	4,757,241	
ant_15mm_3	Anther	34,987,962	27,106,460	4,823,546	
ant_2mm_1	Anther	40,291,178	31,263,978	5,592,297	
ant_2mm_2	Anther	48,670,625	37,629,197	6,064,297	
ant_2mm_3	Anther	36,212,687	28,638,538	4,425,117	
ant_25mm_1	Anther	33,773,432	26,317,098	4,808,669	
ant_25mm_2	Anther	41,864,996	32,794,371	5,098,239	
ant_25mm_3	Anther	40,149,087	31,618,951	5,180,878	
1_0w1_5	Anther	50,968,596	39,492,311	8,494,838	GSM1262610
2_0w2_5w3_0	Anther	38,853,899	29,536,603	7,099,772	GSM1262611
B73_ear_I	Ear	13,803,438	10,924,196	1,177,196	SRX300975

B73_ear_II	Ear	18,669,809	15,210,952	2,001,585	SRX300976
B73_ear_III	Ear	9,720,496	7,798,033	441,577	SRX300977
B73_ear_IV	Ear	8,603,952	6,768,236	1,770,874	SRX300978
B73p1_pln	Pollen	63,454,337	57,659,581	3,193,711	GSM1262608
*mop1*_pln	Pollen	88,342,367	66,045,008	4,744,142	GSM1262609
4_0wPollen	Pollen	38,138,281	30,325,034	5,729,488	GSM1262612
WT_h1_rep1	Seedling	19,204,263	15,339,511	6,724,500	
WT_h1_rep3	Seedling	20,929,612	17,041,157	6,651,336	
hen1_rep1	Seedling	11,269,141	9,312,134	4,551,210	
hen1_rep3	Seedling	17,833,655	14,518,424	6,111,477	
WT_dc1_rep1	Seedling	15,995,927	13,584,335	5,560,917	
WT_dc1_rep3	Seedling	16,075,951	13,273,802	6,034,179	
dcl1_rep1	Seedling	46,892,795	38,008,816	12,974,953	
dcl1_rep3	Seedling	11,171,463	9,294,215	4,570,905	
DCL3a_1P	SAM	39,389,706	28,991,832	8,295,318	
DCL3a_2P	SAM	31,059,936	23,193,902	7,618,204	
DCL3a_3P	SAM	35,283,156	26,468,321	9,717,355	
hodcl3a_1P	SAM	33,162,691	25,290,178	8,875,118	
hodcl3a_2P	SAM	31,475,601	23,386,266	6,904,363	
hodcl3a_3P	SAM	37,196,131	28,403,737	7,512,502	
dWT_tassel_3_2cm	Tassel	37,531,229	32,133,936	11,461,851	
d_dcl4tassel_4_0	Tassel	34,990,260	29,740,235	11,175,374	
dWT_tassel_2_5cm	Tassel	33,235,745	28,402,052	10,521,477	
d_dcl4tassel_2_7	Tassel	40,218,976	34,267,089	12,179,100	
Total		**1,330,075,797**	**1,049,210,247**	**242,804,295**	

miRNAs were separated by tissue and identified with *miRador*. **Table 2.14** shows the counts of the various miRNA classifications identified in each tissue. The predicted miRNAs of each tissue type were kept separate. Targets of these candidate miRNAs were then predicted and validated within *sPARTA* utilizing PARE data for the corresponding tissues. A target was only considered real if its p-value, as determined by *sPARTA*, was ≤ 0.05. The counts of miRNAs validated in each of these tissues is described in **Table 2.15**.

Table 2.18: Counts of maize predicted miRNAs in each tissue

Counts of miRNA Classifications						
	Anther	Ear	Pollen	SAM	Seedling	Tassel
Known	59	68	20	41	45	45
Identical to known	14	14	4	13	13	12
New member existing family	44	34	9	23	28	32
Conserved outside Organism	2	0	0	0	0	0
Novel	17	5	1	3	18	6
Total	**136**	**121**	**34**	**80**	**104**	**95**

Table 2.19: Counts of maize miRNAs that had at least one predicted target with one PARE signal

Counts of miRNA Classifications - PARE Validation						
	Anther	Ear	Pollen	SAM	Seedling	Tassel
Known	56	52	15	35	36	31
Identical to known	14	12	2	13	12	10
New member existing family	42	29	5	20	23	23
Conserved outside Organism	2	0	0	0	0	0
Novel	14	0	0	2	16	2
Total	**128**	**93**	**22**	**70**	**87**	**66**

While focusing on just the novel miRNAs, 28 novel miRNA producing loci were predicted to create 14 distinct mature miRNA sequences. These miRNAs were then manually analyzed to assess their quality. Of the 28 miRNA producing loci, 21 were removed upon manual analysis leaving 7 distinct miRNA producing loci (**Table 2.16**). These 7 loci produced 5 distinct mature miRNA sequences. The manual curation of this data involved a combination of several characteristics of these miRNAs including the prediction across a variety of libraries, the lack of accumulation in *dcl1* and *hen1* libraries, analysis of the PARE signals and quantity of targets, number of mapped locations to the genome, sequence length, predicted precursor image, and analysis of the candidate miRNA in relation to other reads mapping to the same region.

Table 2.20: List of identified maize miRNAs

miR Name	Chr	Strand	miR Position	miR Sequence	Tissue Identified in
candidate-714290_1-5p	6	w	6798215	AGAAUUGGAUGGAACCAAAUC	Anther
candidate-714288_2-3p	6	w	6798088	UUCAGUUUUCUCCAACAUCUUA	Anther
candidate-136244_2-3p	1	w	237845400	UUCAGUUUUCUCCAACAUCUUA	Anther
candidate-612119-3p	5	c	45134052	CGGGAACUGGAGAUGCUACUC	Anther
candidate-1187004-5p	10	w	149277488	AGGACUGGAUCGCCGGAGGGU	Anther SAM Seedling Tassel
candidate-1187005-3p	10	c	149277612	AGGACUGGAUCGCCGGAGGGU	Anther SAM Seedling Tassel
candidate-585842-5p	5	w	5449517	UCCUCUCCCCUGCUGCCACGA	Seedling

We then searched the Internet for these mature miRNA sequences to determine if these candidates have been previously identified, but not present in miRBase yet. The sequence of candidate-714290_1-5p was identified in *Sorghum bicolor* as a member of the miR2275 family (Taylor et al., 2014). This family is present in maize, but an analysis of this showed that it differs by 5 nucleotides from a copy of a member of the maize miR2275 family. The cutoff for a new member of an existing family is 4 nucleotides, so this candidate missed that particular threshold. While I don't believe the threshold for similarity should be changed, this particular candidate does appear to belong to the miR2275 family and thus it is not an entirely novel miRNA.

Candidate-714288_2-3p and candidate-136244_2-3p produce the same mature miRNA sequence. Searches for this candidate miRNA resulted in the identification of another miRNA that was not annotated in miRBase, though it was found to have been previously identified in large scale miRNA analyses in wheat and *Dendrocalamus*

latiflorus (sweet bamboo) (Sun et al., 2014; Zhao et al., 2015). These sequences were not identified as a known miRNA family in either of these studies, and *miRador*'s annotation component was unable to identify this family, thus it is believed that this is a truly novel miRNA for maize. The identification in wheat and sweet bamboo should be seen as supporting evidence of this classification, though further analysis will be done prior to making a final assessment of this sequence.

The search for candidate-612119-3p resulted in the identification of a candidate miRNA that had previously been identified in maize, though it is not present in miRBase. This candidate miRNA, along with 212 other candidates, were identified in maize in response to *Sugarcane mosaic virus* infection (Z. Xia et al., 2018). It is unclear if this sequence has already been submitted to miRBase as the study identifying this candidate was published after the most recent major update to miRBase.

candidate-1187004-5p and candidate-1187005-3p are tandem duplicates that both produce the same mature miRNA sequence. Unlike all of the other identified candidate miRNAs, this was the only candidate miRNA to be identified in multiple tissues. The search for this sequence only found this sequence in a published dissertation (Wu, 2014), but no evidence of this sequence was found anywhere else.

The final candidate miRNA, candidate-585842-5p, found no results and thus it appears to be a truly novel miRNA.

Further analysis will be performed for each of these candidate miRNAs in a variety of other datasets. I hope to utilize RNA-seq transcript data to determine differential expression in mutant datasets where these candidate miRNAs have their production pathways interrupted, thus affecting their total accumulation.

2.9 Summary

In this chapter, I presented a novel plant miRNA prediction tool, *miRador*. One of the inspirations for this tool's development was the publication of an updated set of plant miRNA annotation criteria that no tool utilized when this project was initiated. By this tool's completion, existing tools were updated to make of use those criteria, but there are still several features that set *miRador* apart from the competition. In addition to being faster than the other tools, *miRador* annotates the predicted miRNAs to provide users with insights into the novelty of the predicted miRNAs, a feature that other miRNA prediction pipelines do not utilize.

I compared the quality of results of *miRador* predicted miRNAs to the results of two other miRNA prediction programs. The results of these comparisons show that while there are differences between the various miRNA prediction programs, it is difficult to state which is the optimal program. Each of these miRNA prediction tools have high predictive capabilities, and thus making a recommendation on the best is difficult. While my bias for *miRador* exists, I chose to export the most novel component of *miRador*, the miRNA annotation script, out of the program to be usable on the predictions produced by both *ShortStack* and *miRDeep-P2*. This can be further modified and extended to process predictions of other miRNA tools. By utilizing any of these tools with this standalone annotation tool, researchers can be better informed as to what miRNAs are present in their datasets and from which loci they come from while also identifying novel miRNAs. Extending this pipeline further to utilize PARE validation of predicted targets of candidate miRNAs with *sPARTA* will greatly increase the predictive power of these pipelines.

Finally, utilizing *miRador* and *sPARTA*, I was able to identify four novel miRNAs in maize. Further analysis will be performed on these sequences to confirm their status.

Chapter 3

UTILIZING CLUSTERING AND ABUNDANCE TO CHARACTERIZE DIFFERENTIALLY REGUALTED sRNA LOCI

In this chapter, I discuss the significance and concept of the clustering of small RNAs, particularly within the Meyers Lab database system. I present a clustering tool that I developed to be utilized within the Meyers Lab, and discuss two published projects in which these tools were utilized. Unfortunately, because this tool utilizes the Meyers Lab database system, this tool is currently not generalizable to a 'generic' genomic system. However, through the diversity of the queries that are made possible by MySQL, this clustering script is capable of being used in a variety of analyses and datatypes that are not discussed in this dissertation. The projects in which this tool was utilized are discussed as independent case studies. Both of these projects investigated the production and significance of 24-nt phasiRNAs, the first in maize, and the second in asparagus.

3.1 Small RNA clustering

miRNAs are the most extensively studied class of small RNAs, but they are not the most abundant class of sRNA in plants. Short interfering RNAs (siRNAs), though not as well studied as miRNAs, hold the title as most abundant class of sRNAs in plants. siRNAs are split into three classes: heterochromatic siRNAs (hc-siRNAs, phased siRNAs (phasiRNAs), and *trans*-acting RNAs (tasiRNAs). When we study siRNAs and their effects, we typically do not focus on individual siRNA sequences as we do with miRNAs. Rather, we focus on groups of siRNAs that map to the genome

in densely packed siRNA producing loci. siRNAs that are closely related in function or in biogenesis are typically found in densely packed clusters.

An example of the importance of siRNA clustering can be found in analyses of hybrid vigor in *Arabidopsis thaliana*. F_1 hybrids between Columbia (Col) and Landsberg *erecta* (Ler) lines of *Arabidopsis thaliana* display strong levels of hybrid vigor. These F_1 hybrids display a variety of traits that are superior to each parent separately, including stress tolerance, growth rate, and biomass. These hybrids differ greatly in sRNA populations. sRNA clustering analysis of these plants have identified clusters of siRNAs within transposable elements in F_1 hybrids that appear to be inherited in an additive manner (i.e. siRNA clusters from each parents are observed in the F_1 population). Genic derived siRNAs, on the other hand, are often reduced in the hybrids to the level of parent with lower siRNA expression. At transposable element loci, hybrids showed alterations in RNA-directed DNA methylation (RdDM) with decreases in CHH methylation (Groszmann et al., 2011; Y. Li et al., 2012).

3.2 Overview of Meyers lab database systems

The Meyers group utilizes a proprietary MySQL database schema for the storage of genomic and RNA expression data. This method of storage allows for rapid querying of mapped and abundance information for every library ever processed by the Meyers group. Initial analyses of these processed data is quick through this format, but the true benefit exists upon successive analyses of these previously processed datasets. While the schema is proprietary, I can discuss portions in order to best explain the significance of the analytical scripts described in this chapter.

The Meyers lab separates sRNA expression data from genomic data in MySQL databases. Within a genome database, we store genome sequences, features and their

coordinates. Information such as exons, introns, inverted repeats, transposons, retrotransposons, and more are accessible within this database. RNA expression databases typically store reads, their abundances, and the locations that each read maps to the genome. With these databases, the Meyers lab builds websites that can be utilized to observe where sRNAs map within the genome. While the scripts that I have created do not depend on these visual representations of the data, these resources do allow us to visualize how these clustering scripts function, on a small scale. **Figure 3.1** shows the visual representation of the data contained within the Meyers lab database systems by showing data generated from the combined data of an sRNA expression database and a genome database.

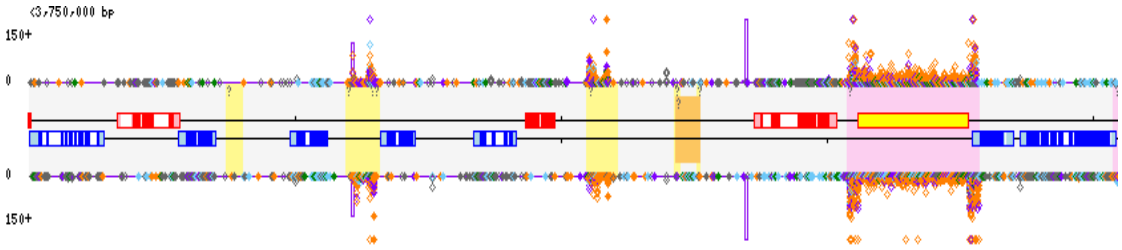

Figure 3.1: Read distribution on genome using Meyers lab web viewer
Small RNAs distributed across a short segment of the genome with several genomic features within. Red and blue rectangles represent genes on the Watson and Crick strands, respectively. White gaps within these regions represent introns, solid regions are exons, and lighter blocks are UTRs. The yellow, pink, and orange shaded boxes are unannotated repeats predicted by *RepeatMasker*, and *einverted*. Each diamond along the plot represents a single sRNA mapping to that location, where color represents a specific sequence length of sRNA. Filled diamonds represent sRNAs with unique mappings to the genome and unfilled diamonds are that non-uniquely mapped sRNAs. The y-axis represents the normalized expression of the reads.

There are a variety of clustering tools that exist, but many require the data to be stored in a specific file format to process it. Data within these databases could be exported into these file formats for clustering, but this is not an ideal solution. By utilizing clustering techniques in conjunction with these databases, we can cluster data

utilizing largely the same algorithms, but in shorter runtimes. Unfortunately, this makes the tools described in this chapter proprietary and unable to be released outside of the lab. However, these techniques have been extensively useful in analyses in several projects, some of which are described below. These tools have already been made available to members of the Meyers lab and will likely have further use beyond the case studies described below.

3.3 Techniques for clustering small RNAs

3.3.1 Clustering by static window sizes

One of the most basic versions of clustering that can be conducted will simply segment the entire genome into set window sizes (**Figure 3.2**). Reads mapping to each cluster are hits-normalized, meaning that the total abundance of each small RNA is distributed evenly across all mapped locations. These hits-normalized abundances are then summed for each size class to create a representative abundance for each cluster which could then be compared across multiple libraries utilizing differential expression techniques.

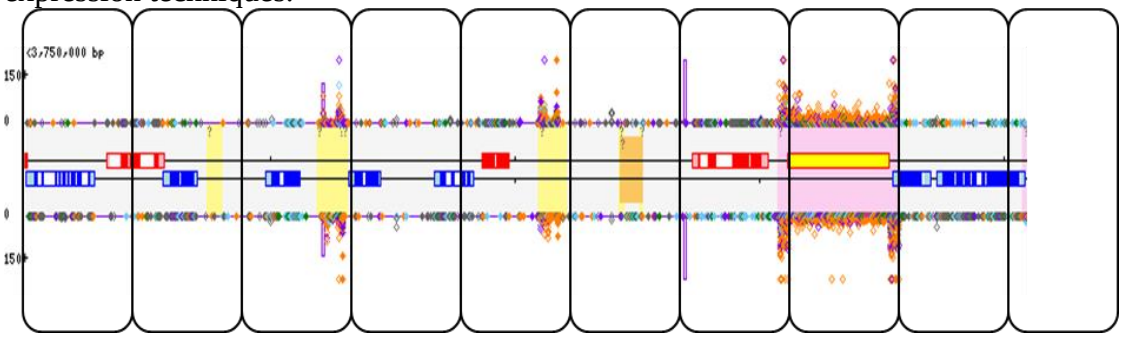

Figure 3.2: Example of static clustering by set window sizes
The clustering of small RNAs utilizing set window sizes across the entire genome. Hits-normalized abundances of all sRNAs mapping to each cluster are summed together to create a representative abundance for each cluster.

One issue with this method of clustering is that many sRNAs that are biologically linked and thus should be clustered together can become separated by the arbitrary window sizes. Dynamic clustering would alleviate this issue by automatically grouping sRNAs into clusters based on their abundance and general mapped location. This, however, can be an extremely time consuming process. While there is significant utility with such a technique, it was not utilized in this dissertation due to better alternatives.

3.3.2 Clustering by genomic features

Because we know that specific classes of siRNAs cluster together on various genomic features like genes (**Figure 3.3**), repetitive elements (**Figure 3.4**), or even *PHAS* loci, we have identified that it is possible to simply cluster sRNAs on those specific genomic features. This significantly reduces the runtime compared to both dynamic clustering and static clustering because each sRNA does not need to be processed as only sRNAs mapping to locations of interest need to be analyzed. In the case of dynamic clustering, multiple passes are not required to group sRNAs into their final clusters. Representation across these features can be compared to identify potential sRNA impact on these specific genomic features, or they can be compared across libraries by utilizing differential expression techniques. This is the technique that was utilized in the case studies below.

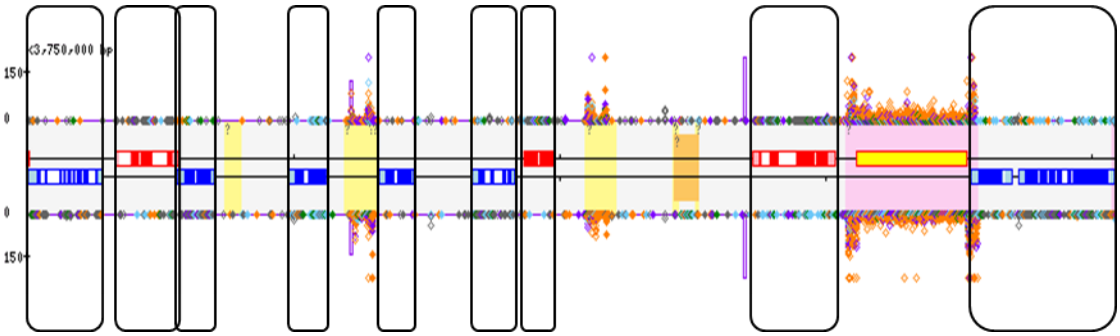

Figure 3.3: Example of gene based clustering
Clustering of small RNAs utilizing annotated genes as cluster windows

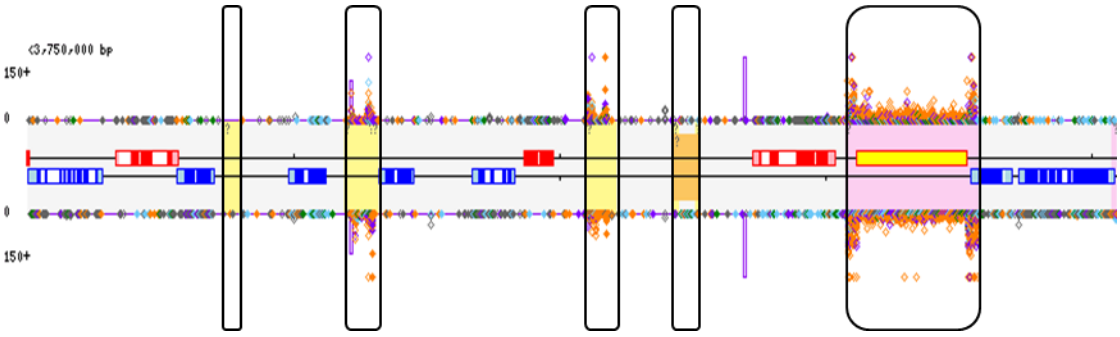

Figure 3.4: Example of repeat based clustering
Clustering of small RNAs utilizing predicted repeats as cluster windows

3.4 Description of algorithms

The tool that was developed for this method of clustering has been utilized in multiple projects, with a few variations within. The original purpose of this program was to identify representative abundances of small RNAs mapping to specific genomic features. In one particular case (**Section 3.6**), however, we were instead interested in identifying the genomic features that overlapped with a given set of loci. In this case, the cluster abundances that were tracked rather than the number of nucleotides of the loci of interest that overlapped with the various genomic features. I will detail this as a separate variation of the algorithm, but the general clustering concepts discussed previously will still apply for this variation. This specific case also highlights the

modularity of the clustering that can be achieved utilizing the Meyers lab database system.

3.4.1 Creating a set of empty clusters

One of the goals of this tool was to identify genomic features that are overrepresented with clusters of specific types small RNAs. However, rather than performing a dynamic clustering of all sRNAs, we utilize either annotated genomic features or user provided features to determine regions that may be impacted by the sRNAs that map to these regions.

In the case of annotated genomic features, the Meyers lab stores these coordinates in MySQL databases for each annotated genome. These features include: exons, introns, transposons, retrotransposons, and inverted repeats. These coordinates, including the chromosome, strand, start position, and end position, are queried into Python lists. In the case of the user defined features, these are simply pulled from a provided CSV file and stored into Python lists. At this stage, these list of coordinates are effectively the empty clusters which will become populated with the relevant data in the next step.

3.4.2 Generating representative abundances for a cluster

3.4.2.1 Counting sRNAs that map to clusters

A case study utilizing this version of this tool is described in **Section 3.6**. In this case, the abundance of a cluster is comprised of the sum of abundances of all sRNAs that map to it. As was discussed in **Section 3.3.1**, allocating the full abundance of a multi-mapped sRNA read (a read that maps to multiple locations in the genome) to each mapped location will result in that read having an effective abundance of N

times greater than what was actually sequenced, where N is the number of locations that the sRNA maps to the genome. In order to prevent this from happening, we utilize a hits-normalized abundance approach, where every mapped location of an sRNA will receive an equal proportion of that read's total abundance. We refer to this as hits-normalized abundance, or abundance per hit, where a "hit" refers to any location that a read maps to.

To determine the abundance of the empty clusters created in the previous step, a second database is utilized. This database stores sRNA mapping and abundance information for a group of sRNA libraries. For each cluster, all sRNAs that map to within the coordinates of that cluster are queried to determine both their abundance and multi-mapping rate in order to identify the hits-normalized abundance of each read (abundance / # of hits to the genome). The hits-normalized abundance is then summed across all reads mapping to that cluster to generate the representative expression abundance for that cluster.

Each of these tables has an index on the sRNA sequences and their coordinates, so these data accession queries are exceptionally quick. These indexes are akin to those used in the discussion of Python dictionaries presented in **Figure 2.2** where the accession of this data utilizes a specific hash function to identify the location in a table in which the data is stored. These clusters and their representative abundances are then stored into CSV files, as well as another MySQL table so that the user has the power to create their own queries to access only clusters meeting specific requirements.

In addition to working with all sRNAs that map to these positions, these data accession queries are also modifiable via user provided options, or even a direct

modification to the query itself. The modularity of MySQL provides the user with significant power to determine what should be included in a cluster without breaking the script. Such modifications can include the size of sRNAs that should be included, or even by adjusting the cluster regions to identify clusters of sRNAs that map a specific distance from the original coordinates provided. These techniques have been utilized extensively within the Meyers lab. This means that data that may have never been intended to be clustered upon the inception of this tool could be clustered without modifying anything beyond the base MySQL queries.

3.4.3 Identifying overlaps of user-defined loci with genomic features

A set of results utilizing this version of clustering is described in **Section 3.6**. In this case, rather than investigating the sRNAs that map to various genomic features, we investigate the level of overlap that exists between a set of user defined loci with the various annotated genomic features. The clusters in this case are the user-defined loci, and the goal is to identify the number of nucleotides of a cluster that overlap with a genomic feature. This requires that each cluster be duplicated for as many genomic features are included in the analysis.

This analysis can be thought of as the identification of overlapping clusters from two separate sets. The core of this function loops through the user defined loci, and for each locus, it scans each genomic feature to identify any overlapping coordinates. When an overlap is found, a counter is increased for each nucleotide that overlaps with the corresponding genomic feature. This counter of overlaps serves as the ‘cluster abundance’. Thus, by the end of the analysis, each cluster will have abundances for each genomic feature. As with the previous case, these clusters and their overlapping nucleotide counts are stored into a MySQL table so that the user has

the power to create their own queries to access only clusters meeting specific requirements.

In this method of clustering, expression data is not utilized directly. Typically, the user defined loci is generated by a separate program that utilizes the expression data. In the case of **Section 3.6**, the data was generated utilizing a tool that I assisted in the development of, *PHASIS* (Kakrana et al., 2017).

3.5 Case study: Impact of the *DCL5* mutation on 24-nt phasiRNAs and phasiRNA producing loci

This section includes work that has been published in *bioRxiv* (citation below), and is currently in revision with *Nature Communications*. My primary contributions to this work include all sRNA and RNA-seq analysis, including sRNA profile analysis, phasiRNA expression analysis, RNA-seq transcript analysis, and all differential expression analysis. I have only included portions of this paper that is relevant to this specific chapter, and thus certain experiments and conclusions were removed before being included in this chapter.

Teng, C., Zhang, H., Hammond, R., Huang, K., Meyers, B., & Walbot, V. (2018). *Dicer-like 5* deficiency confers temperature-sensitive male sterility in maize. *bioRxiv*, 498410. doi:10.1101/498410

3.5.1 Background

From extensive sRNA sequencing in plants, numerous loci generating phasiRNAs have been reported; in grasses, phasiRNAs are enriched in flowers, particularly male reproductive organs (Johnson et al., 2009; Song, Li, et al., 2012; Song, Wang, et al., 2012; Zhai et al., 2015). PhasiRNA production is initiated by miRNA-mediated cleavage of RNA polymerase II transcripts of two classes of *PHAS*

loci. Subsequently, the 3' portion of cleaved transcripts is converted to double-stranded RNA, a substrate for precise chopping by DICER-LIKE 4 (DCL4) yielding 21-nt products; a distinct, proposed role for DCL5 is in the generation of 24-nt phasiRNAs (Johnson et al., 2009; Song, Li, et al., 2012; Song, Wang, et al., 2012). The 21-nt phasiRNAs are highly abundant during initial cell fate setting in maize anthers (Zhai et al., 2015) and are important for male fertility in rice (Komiya et al., 2014; Nonomura et al., 2007). The 24-nt phasiRNAs accumulate coincident with the start of meiosis, peak during meiosis, and persist at lower levels afterwards; this pattern has generated speculation that they regulate meiosis (Nonomura, 2018; Ono et al., 2018).

In angiosperms, five DCLs have been described and partially characterized (Margis et al., 2006). DCL1 is important for miRNA biogenesis, as illustrated by regulation of meristem determinacy in maize inflorescences (Thompson et al., 2014). In *Arabidopsis thaliana*, DCL2 processes viral- and transgene-derived siRNAs (Parent, Bouteiller, Elmayan, & Vaucheret, 2015); DCL3 produces 24-nt hc-siRNAs, which then direct DNA methylation of target loci (Daxinger et al., 2009). In many plants including *Oryza sativa* (rice), DCL4 generates 21-nt *trans*-acting siRNAs and phasiRNAs (Song, Li, et al., 2012). Although the functions of these four *DICER-LIKE* genes are well conserved in flowering plants, these genes have partially overlapping functions (Gasciolli, Mallory, Bartel, & Vaucheret, 2005; Parent et al., 2015). A fifth, more recently discovered gene, *Dcl5*, is monocot-specific (Margis et al., 2006).

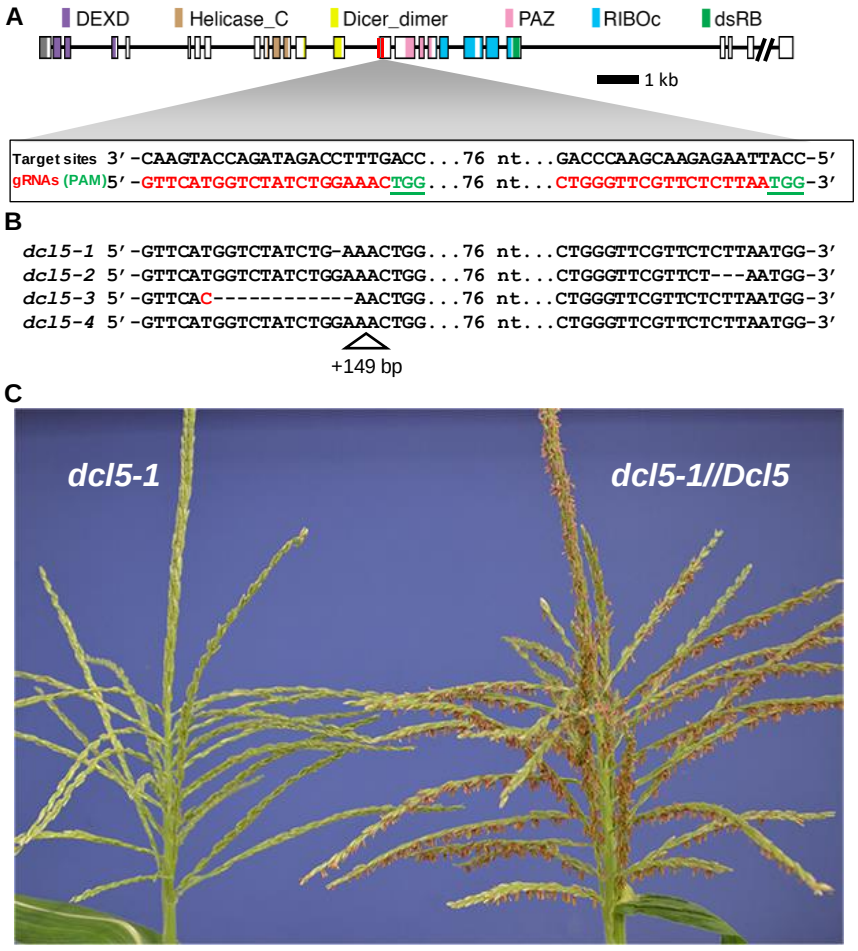

Figure 3.5: CRISPR-Cas9 was used in maize to mutate *Dcl5*.
(A) Schematic diagram of the *Dcl5* gene model. Conserved domains were identified using the NCBI conserved domain search tool against database CDD v3.16. All domains were found with extremely high confidence except the double stranded RNA-binding (dsRB) domain, which had an e-value of 0.05. (**B)** Sequence of the four mutant alleles selected for this study. (**C)** A *dcl5-1* tassel lacking any exerted anthers (left) and a sibling heterozygous *dcl5-1//Dcl5* plant at peak pollen shed with hundreds of exerted anthers (right).

Four *dcl5* mutant alleles were utilized throughout this analysis that displayed male sterility (**Figure 3.5**). *dcl5-1* and *dcl5-4* are frameshift mutants with *Dcl5* transcript levels down-regulated to about a third of their wild type siblings; in contrast, *dcl5-2* (3 bp deletion) and *dcl5-3* (12 bp deletion, 1 bp substitution) have similar or higher transcript levels compared to control siblings (**Figure 3.6**). Wild type and

heterozygous *dcl5-1//Dcl5* plants greenhouse-grown at 32° C maximum day/21° C minimum night were identical in whole plant architecture, anther morphology, and fertility. Under the same conditions, homozygous *dcl5-1* plants were male sterile. Compared to their fertile siblings, sterile *dcl5-1* plants lacked visible differences in tassels and spikelets, however, the sterile anthers were shorter, contained shrunken pollen, and did not exert from the spikelets. Sporadically, a few anthers exerted and shed viable pollen. Under field conditions, all four alleles showed equivalent levels of male sterility. We concluded that *Dcl5* is required for robust male fertility. To better understand the function of *Dcl5*, its proposed role in 24-nt phasiRNA biogenesis, and potential role of 24-nt phasiRNAs, we analyzed both sRNA and RNA-seq expression datasets.

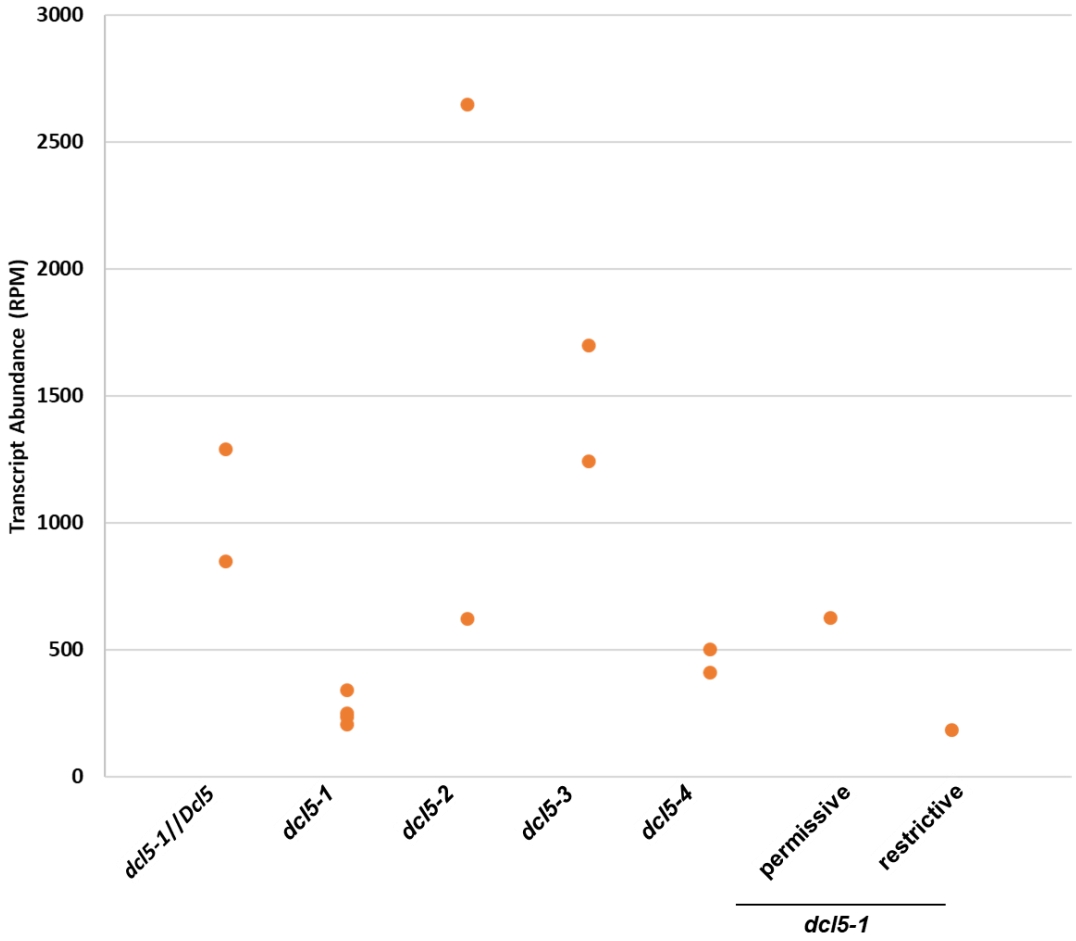

Figure 3.6: *Dcl5* mRNA levels in 2.0 mm anthers of control and mutant plants *Dcl5* transcript levels from the RNA-seq data are indicated for each genotype. *dcl5-1* and *dcl5-4* are frameshift mutants with transcripts levels about a third of their wild type siblings at 2.0 mm. In contrast, *dcl5-2* and *dcl5-3* contain short in-frame deletions and have similar or higher transcript levels compared to control (*dcl5-1//Dcl5*). For the *dcl5-1* plants grown under "restrictive" conditions, due to developmental differences, spikelets equivalent to 2.0 mm anthers were analyzed, which contains the same transcripts level as the other *dcl5-1* samples under the field and "permissive" conditions.

3.5.2 Results

Numerous sRNA and RNA-seq datasets were generated to determine the role that *Dcl5* plays in sRNA accumulation as well as any potential transcriptional effects

caused within these *Dcl5* mutants (**Table 3.1** and **Table 3.2**). These libraries are generated from 1.5 mm and 2.0 mm anthers and their temperatures were controlled during its growth period. These anther stages were selected as the 24-nt phasiRNA expression peaks from 1.5 to 2.0 mm, coinciding with meiotic progression through prophase I to metaphase I, and when somatic cells continue differentiating for postmeiotic supporting roles. At this peak, 24-nt phasiRNAs reach approximately 60% of all 24-nt RNAs in normal controls (Zhai et al., 2015).

Table 3.1: Summary of sRNA libraries

Stage	Genotype	Temperature	Total Sequences [a]	Genome Matched Reads [b]
1.5 mm Anther	Fertile (*dcl5-1//Dcl5*)	28 °C/ 22 °C	24,064,348	13,089,838
2.0 mm Anther	Fertile (*dcl5-1//Dcl5*)	28 °C/ 22 °C	43,109,925	21,568,620
1.5 mm Anther	*dcl5-1*	28 °C/ 22 °C	22,909,889	8,085,616
2.0 mm Anther	*dcl5-1*	28 °C/ 22 °C	14,688,166	3,672,763
1.5 mm Anther	*dcl5-2*	28 °C/ 22 °C	36,666,135	21,891,746
2.0 mm Anther	*dcl5-2*	28 °C/ 22 °C	13,636,538	5,751,885
1.5 mm Anther	*dcl5-3*	28 °C/ 22 °C	24,861,769	13,078,105
2.0 mm Anther, rep1	*dcl5-3*	28 °C/ 22 °C	13,250,979	4,532,956
2.0 mm Anther, rep2	*dcl5-3*	28 °C/ 22 °C	30,523,224	12,988,280
1.5 mm Anther	*dcl5-4*	28 °C/ 22 °C	29,017,229	14,575,689
2.0 mm Anther	*dcl5-4*	28 °C/ 22 °C	20,242,459	9,726,062
Spikelets with 1 mm and 2 mm Anther	*dcl5-1*	22 °C/ 20 °C	29,794,869	21,328,316
Spikelets with 1 mm and 2 mm Anther	*dcl5-1*	28 °C/ 22 °C	37,967,130	27,402,246
Spikelets with 2 mm and 3 mm Anther	*dcl5-1*	22 °C/ 20 °C	24,254,347	17,295,888
Spikelets with 2 mm and 3 mm Anther	*dcl5-1*	28 °C/ 22 °C	28,087,677	20,283,182
	Total		**393,074,684**	**215,271,192**

Table 3.2: Summary of RNA-seq libraries

Stage	Genotype	Temperature	Total Sequences [a]	Genome Matched Reads [b]
2.0 mm Anther, rep1	Fertile (*dcl5-1//Dcl5*)	28 °C/ 22 °C	25,852,398	22,631,109

2.0 mm Anther, rep2	Fertile (*dcl5-1//Dcl5*)	28 °C/ 22 °C	22,707,551	18,660,169
2.0 mm Anther, rep1	*dcl5-1*	28 °C/ 22 °C	26,177,644	21,876,048
2.0 mm Anther, rep2	*dcl5-1*	28 °C/ 22 °C	23,154,299	20,200,061
2.0 mm Anther, rep3	*dcl5-1*	28 °C/ 22 °C	26,964,117	23,166,502
2.0 mm Anther, rep4	*dcl5-1*	28 °C/ 22 °C	26,825,344	23,890,823
2.0 mm Anther, rep1	*dcl5-2*	28 °C/ 22 °C	25,002,197	21,429,646
2.0 mm Anther, rep2	*dcl5-2*	28 °C/ 22 °C	28,568,522	25,349,114
2.0 mm Anther, rep1	*dcl5-3*	28 °C/ 22 °C	24,497,165	18,778,333
2.0 mm Anther, rep2	*dcl5-3*	28 °C/ 22 °C	25,494,811	16,463,390
2.0 mm Anther, rep1	*dcl5-4*	28 °C/ 22 °C	25,391,141	20,232,690
2.0 mm Anther, rep2	*dcl5-4*	28 °C/ 22 °C	22,897,348	20,646,852
Spikelets with 2 mm and 3 mm Anther	*dcl5-1*	22 °C/ 20 °C	29,887,719	23,216,420
Spikelets with 2 mm and 3 mm Anther	*dcl5-1*	28 °C/ 22 °C	22,928,299	16,326,032
	Total		**356,348,555**	**292,867,189**

[a] Total small RNAs after filtering bad reads and trimming adapters.
[b] Numbers determined by mapping to the maize B73 genome, version 4.

3.5.2.1 24-nt phasiRNA abundance reduced in *dcl5* anthers

A profile of sRNA expression was created utilizing a relatively simple MySQL query. 24-nt phasiRNA expression, on the other hand, was generated utilizing the previously discussed clustering script. 24-nt phasiRNAs can be identified utilizing the previously published 176 *24-PHAS* loci (Zhai et al., 2015). By clustering 24-nt sRNAs on these 176 loci, we can identify the representative abundance of 24-nt phasiRNAs. Thus, by combining the total sRNA expression data with the 24-nt phasiRNA expression data, I was able to generate a profile of the effects that each *dcl5* allele has on phasiRNA accumulation in 1.5 and 2.0 mm anthers (**Figure 3.7**). In all *dcl5* libraries, 24-nt phasiRNAs were dramatically reduced. In this analysis, we also examined 21-nt phasiRNAs to confirm no potential effects due to the lack of *Dcl5*. 21-nt phasiRNAs do appear slightly elevated, though the expression is not much higher than the control libraries. Such an effect has been seen previously when a significant class of sRNA, 24-nt hc-siRNAs, were eliminated resulting in the increased accumulation of miRNAs and other 21-22 nt sRNAs (Nobuta et al., 2008). To confirm

this as the case here, I investigated the ratio increase of 21-nt phasiRNAs relative to 21-nt non-phasiRNAs. This ratio was increased in the various alleles of the *dcl5* mutant libraries, though this increase was slight and did not appear to be biologically relevant.

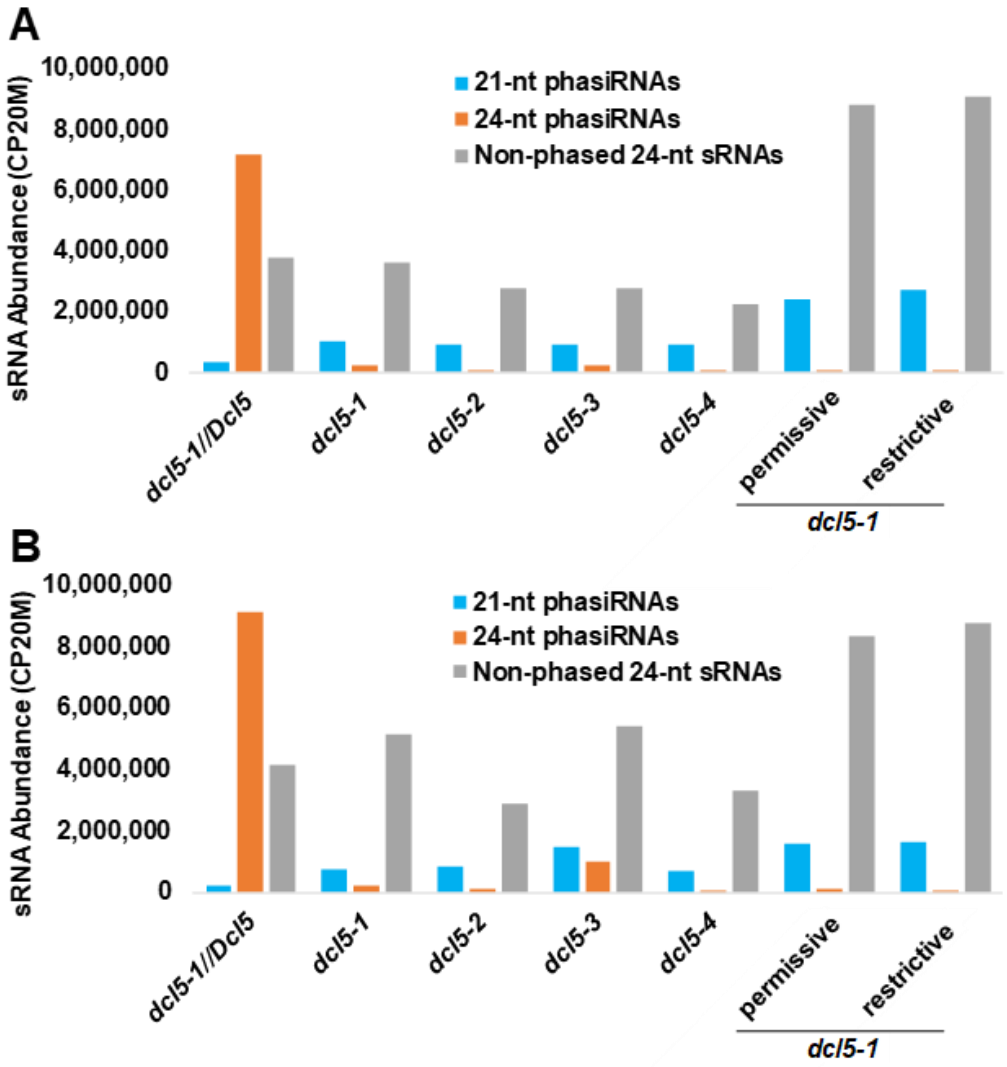

Figure 3.7: Among sRNA types, only abundances of 24-nt phasiRNAs were reduced in *dcl5* anthers.

sRNA abundances (normalized to total reads) from 1.5 mm anthers **(A)** and 2.0 mm anthers **(B)**. Permissive refers to anthers dissected from plants grown at 21° C during

meiosis and the subsequent six days; restrictive refers to anthers from plants grown at 28° C during this interval. The plants and libraries from *dcl5-1* under permissive and restrictive conditions were a separate experiment, and the increased levels of non-phased 24-nt sRNAs likely reflect different growth conditions and experimental variation.

3.5.2.2 Differential expression analysis of *dcl5* transcripts reveals no obvious effects on transcript expression

One of the major interests in this project was to determine the genes that might be impacted by the reduced accumulation of 24-nt phasiRNAs which may be the cause for the sterility observed in these mutants. Thus, we sought to investigate transcript expression in these mutant libraries in comparison to heterozygous controls, which display no mutant phenotype. RNA-seq libraries were trimmed to remove linker adaptor sequences using *Trimmomatic* version 0.32 (Bolger, Usadel, & Lohse, 2014) and reads were mapped to version 4 of the B73 genome using *Tophat* version 2.0.12 (Trapnell, Pachter, & Salzberg, 2009). For differential expression analysis of genes, we assembled the transcripts using the *Cufflinks* package (Trapnell et al., 2010) and raw gene expression levels were quantified using *featureCounts* version 1.5.0 (Smyth, Shi, & Liao, 2013) to generate count tables that were then imported into R for statistical analysis (R Core Team, 2018). Weakly expressed transcripts (fewer than 1 read) were filtered out, and the remaining transcripts were normalized using *DEseq2* to finally identify a set of differentially expressed genes for anthers homozygous for each mutant allele (Love, Huber, & Anders, 2014). Due to a lack of a replicate for the *dcl5-1* restrictive library, we calculated dispersions using the mean of libraries to identify differentially expressed transcripts without a complete set of replicates.

Differential expression analysis found just 23 genes differentially expressed in our 2.0 mm *dcl5-1* anther libraries, but all of these appear to be reflective of arrested

tapetal development (**Table 3.3**) rather than by a direct cause of the 24-nt phasiRNAs. That the thousands of 24-nt phasiRNAs fail to impact mRNA abundances strongly suggests that 24-nt phasiRNAs are distinct from tasiRNAs which do have specific mRNA targets and modulate transcript abundances.

Table 3.3: Differentially expressed genes in 2.0 mm *dcl5-1* anther libraries

AGPv4 Identifier	Gene name	*dcl5-1*//Dcl5-1	*dcl5-1*//Dcl5-1	*dcl5-1*	*dcl5-1*	*dcl5-1*	*dcl5-1*	permissive	restrictive
Zm00001d013101		20	11	214	398	743	761	192	234
Zm00001d016396		447	345	134	71	107	71	19	8
Zm00001d018184	fls1	60	73	366	166	355	394	3665	3206
Zm00001d040166	atm1	13650	16259	10384	7412	8045	6547	4049	1953
Zm00001d023376	alp1	328	296	991	625	831	724	2857	4843
H4C7		9894	6280	4618	3979	5619	4665	1365	419
Zm00001d042930	his2b4	10069	7521	4059	3515	5260	4561	1680	603
Zm00001d024094		1464	1215	467	335	551	375	192	78
Zm00001d050100	his2b2	24083	14330	10097	9429	11263	9182	3212	1226
Zm00001d012275		32838	19894	13693	11231	13834	11778	4339	1752
Zm00001d032070		17595	10883	7794	7876	8965	7413	2520	971
Zm00001d051591	his2b5	9861	7435	4702	4344	5772	4770	2452	857
Zm00001d020580	his2b1	16568	11241	7902	6084	7280	6469	2663	1349
Zm00001d005789	his2b3	8391	5503	3719	3235	3502	3142	1360	675
Zm00001d023801		711	895	334	310	480	314	129	67
Zm00001d007274		278	396	836	734	907	902	1895	2464
Zm00001d039821		7038	6474	4359	3183	4037	3369	1663	860
Zm00001d007084		21114	12476	8612	8564	10133	8948	2985	1827
Zm00001d025687		1165	1209	736	879	800	656	222	149
Zm00001d003140		8	32	150	136	140	218	233	370
Zm00001d027652	tip1	2570	1935	6159	5717	5063	4647	10171	10658
Zm00001d052143		952	1388	733	717	620	661	273	187
Zm00001d017423	orc2	329	363	165	192	165	106	71	48

3.5.2.3 DCL5 severely disrupts 24-nt phasiRNA processing with modest impact on 24-nt *PHAS* precursors

In *dcl1* mutants, miRNAs with reduced expression have increased precursor miRNA accumulation due to the lack of cleavage facilitated by DCL1 (Thompson et al., 2014). We hypothesized that a similar effect might be observed in these *dcl5* libraries in the *24-PHAS* precursors. Typical RNA-seq transcript analysis would begin with a transcriptome assembly and differential expression analysis, however *PHAS* precursors have no splice junctions and thus can be treated differently. Using the annotated *24-PHAS* loci, I was able to generate a summed hits-normalized abundance of all RNA-seq reads mapping to these loci. By utilizing the clustering program, I clustered all RNA-seq reads that mapped to the 176 *24-PHAS* loci to generate a representative abundance of each *24-PHAS* precursor. With these data, I generated a heatmap of the expression of the *24-PHAS* precursor abundances across each *Dcl5* mutant allele for visual analysis (**Figure 3.8**). Unlike the phasiRNAs, expression of the *24-PHAS* precursors was either slightly increased or unchanged in the *dcl5* alleles. Student's t-test analysis confirms this result with no *PHAS* precursor differentially expressed in any more than a single *dcl5* allele. Therefore, the absence of functional DCL5 severely disrupts 24-nt phasiRNA precursor processing with only a modest impact on their accumulation.

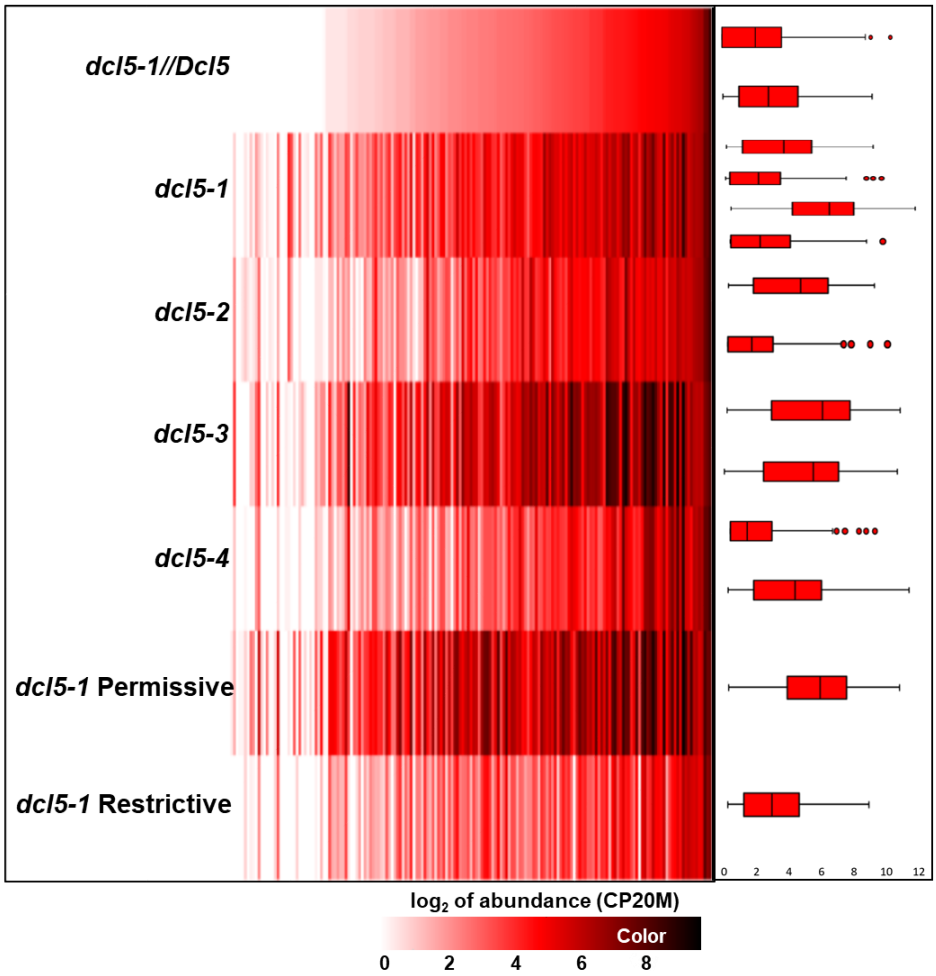

Figure 3.8: Different *dcl5* mutant alleles display altered accumulations of 24-*PHAS* precursors.
The RNA-seq reads from 2.0 mm anthers were mapped to 176 24-nt *PHAS* loci (vertical bars within each row); this analysis was performed for the heterozygous control (top row) and four homozygous *dcl5* mutants, as labeled. Replicates were averaged to create a single abundance for each allele. Shades of white to red to black represent the sum of hits-normalized abundances of RNA-seq reads, converted to $\log_2$ of abundance (see the key at the bottom). To the right, boxplots with the same scale are shown for the same group of *PHAS* loci, for each replicate allele; the center line represents the median, box limits are the upper and lower quartiles, whiskers are the 1.5x interquartile range, and points show the scatter of outliers.

3.5.3 Discussion

This study in maize establishes that DCL5 plays a unique role, having subfunctionalized or neofunctionalized from DCL3 for 24-nt phasiRNA biogenesis in anthers. *Dcl5* transcripts are still expressed in diverse organs, however, protein and critical functions are restricted to anthers. Expression analysis found just 23 differentially expressed genes in our *dcl5-1* mutants, but all of these are likely reflective of arrested tapetal development. That the thousands of 24-nt phasiRNAs fail to impact mRNA abundances significantly indicates that 24-nt phasiRNAs are distinct from tasiRNAs which do have specific mRNA targets and modulate transcript abundances.

Our analysis of temperature-controlled growth of *dcl5-1* mutants identified that while 24-nt phasiRNA accumulation is as reduced as the other *dcl5* alleles in non-temperature regulated environments, fertility is rescued. DCL5 and therefore the 24-nt phasiRNAs are dispensable at cool growing temperatures for maize, however, DCL5 is required for robust fertility at temperatures for optimal yield in the U.S. corn belt, where temperatures are higher. As climate disruption, including more extensive heat waves, and a general warming trend increase field temperatures, buffering maize tapetal development from adverse effects may be possible by enhancing 24-nt phasiRNA functions. The 24-nt phasiRNAs may permit development during other untested environmental challenges. Given the high degree of variation typically observed across maize inbreds for many traits, the *dcl5* mutant phenotype may vary across inbreds. More generally, it is unknown how many crop plants utilize reproductive phasiRNAs and whether their tapetal health and functions could benefit by the introduction or optimization of these or other sRNA pathways (De Storme & Geelen, 2014; Vernoud et al., 2009; Yu et al., 2017).

3.6 Case study: Identification and analysis of *PHAS* clusters in Asparagus

This section includes work that was published in *Genome Research* (citation below). My primary contributions to this work include a variety of analyses specific to asparagus including the search for miRNA triggers of phasiRNA producing loci, and the analysis of 21- and 24-nt phasiRNA producing loci and the genomic features that they overlap with. The text in this case study features only the corresponding analyses that I participated in, the relevant background to these analyses, and the conclusions generated from these analyses.

Kakrana, A., Mathioni, S. M., Huang, K., Hammond, R., Vandivier, L., Patel, P., . . . Meyers, B. C. (2018). Plant 24-nt reproductive phasiRNAs from intramolecular duplex mRNAs in diverse monocots. *Genome Research, 28*(9), 1333-1344. doi:10.1101/gr.228163.117

3.6.1 Background

The functions of plant reproductive phasiRNAs remain unknown, but their role in male reproductive success is evident. For example, a mutant in MEIOSIS ARRESTED AT LEPTOTENE (MEL1), encoding an Argonaute protein in rice, arrests in early meiosis and has an abnormal tapetum and anomalous pollen mother cells (Komiya et al., 2014; Nonomura et al., 2007). MEL1 selectively binds 21-nt phasiRNAs (Komiya et al., 2014), indicating that pre-meiotic phasiRNA are required for male fertility. Variants of two loci generating 21-nt phasiRNAs are the basis for photoperiod and temperature sensitive male sterility in rice (Fan et al., 2016). 24-nt meiotic phasiRNAs also appear to have a role in male fertility (Ono et al., 2018), although we did observe that regulating growth temperatures has recovered male fertility in maize in the absence of 24-nt phasiRNAs (**Section 3.5)** (Teng et al., 2018).

The evolutionary origins of plant reproductive phasiRNAs are also poorly characterized, and little is known of them, outside of their presence in grasses (Poaceae, Poales). In this context, 24-nt meiotic phasiRNAs are particularly interesting as their biogenesis depends on a Dicer (DCL3) homolog, DCL5, originally known as DCL3b, which appears to be specialized for their production, but it is not well described outside of grass genomes (Margis et al., 2006). Also, miR2275, the specialized trigger for these 24-nt phasiRNAs, is poorly characterized outside of the grasses. The 21-nt pre-meiotic phasiRNAs are triggered by members of the miR2118/482 superfamily, which in many other plant genomes triggers the production of phasiRNAs from NB-LRR pathogen-defense genes (Zhai et al., 2011; Y. Zhang, Xia, Kuang, & Meyers, 2016). The evolutionary route by which miR2118 shifted to target non-coding *PHAS* precursors in a highly restricted spatiotemporal manner is not yet known.

3.6.2 MicroRNA triggers and biogenesis components of 21- and 24-nt phasiRNAs in garden asparagus

We investigated 21-nt and 24-nt phasiRNA producing loci for their miRNA triggers. We first did this by predicting and validated PARE supported target sites with *sPARTA* (Kakrana et al., 2014), followed by an exhaustive search for miRNA triggers utilizing *PHASIS*, a phasiRNA prediction tool (Kakrana et al., 2017). Of the 29 predicted 21-*PHAS* loci, we found triggers for 16, 14 of which belonged to miR2118 family members. Target loci included the three reproductive specific 21-*PHAS* loci mentioned above, three NB-LRR genes and four other 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, Jan, Rajagopalan, & Bartel, 2006). The numbers of phased NB-LRRs and TAS3 loci identified in the asparagus genome are low (n=3), relative to other plant genomes in which tens to hundreds of NB-LRRs generating phasiRNAs have been reported (Arikit et al., 2014; Shivaprasad et al., 2012; R. Xia, Xu, Arikit, & Meyers, 2015; Zhai et al., 2011).

Table 3.4: List of asparagus PARE libraries

Code	Title	Total Sequences	Genome Matched Reads	Distinct Genome Matched Reads	Pubmed ID of data source
88F_d	Spears - female	22,588,072	18,009,351	5,119,938	PMID: 29093472
88M_d	Spears - male	25,441,003	20,220,748	5,756,736	PMID: 29093472
88SM_d	Spears - super male	20,141,905	16,066,747	4,435,945	PMID: 29093472
89F_d	Spears - female	21,201,482	16,656,387	4,993,468	PMID: 29093472
89M_d	Spears - male	19,713,775	15,518,328	5,178,869	PMID: 29093472
89SM_d	Spears - super male	22,629,307	18,002,394	5,022,618	PMID: 29093472
103F_d	Spears - female	26,367,163	20,496,055	6,143,486	PMID: 29093472
103M_d	Spears - male	21,334,611	16,554,967	5,335,489	PMID: 29093472
AspM_mid	Male flower mid stage	28,944,491	23,471,321	7,666,998	PMID: 29093472
AspM_ear	Male flower early stage	29,656,483	23,953,176	7,516,117	PMID: 29093472
AspFM_mid	Female flower mid stage	25,144,959	20,389,169	6,740,980	PMID: 29093472
AspFM_ear	Female flower early stage	23,728,164	19,220,543	6,678,019	PMID: 29093472
Asp_shoot	Shoot	21,734,779	17,515,947	5,978,640	PMID: 29093472
Asp_Lf	Cladophylls	24,820,749	20,086,849	6,155,791	PMID: 29093472

For the reproductive 24-*PHAS* loci, our analyses failed to identify a miRNA trigger, including miR2275. This result was highly surprising given that all reported 24-nt phasiRNAs in maize and rice are triggered by miR2775 family members (Johnson et al., 2009; Song, Li, et al., 2012; Zhai et al., 2015). Target analysis in

asparagus for miR2275 did identify 15 intergenic and 10 protein-coding targets, supported by PARE data, but none of these targets corresponded to 24-*PHAS* loci.

Upon manual inspection of the 24-*PHAS* loci, we observed that many showed a substantial strand bias, inconsistent with the RDR6-dependent biogenesis of *PHAS* precursors in grasses (Song, Wang, et al., 2012). Moreover, we also noticed that an overlap of these 24-*PHAS* loci with inverted repeats (**Figure 3.9**A); in other plants like *Arabidopsis*, inverted repeats are processed by DCL4 and DCL1 into 21-nt phasiRNAs and miRNAs, respectively.

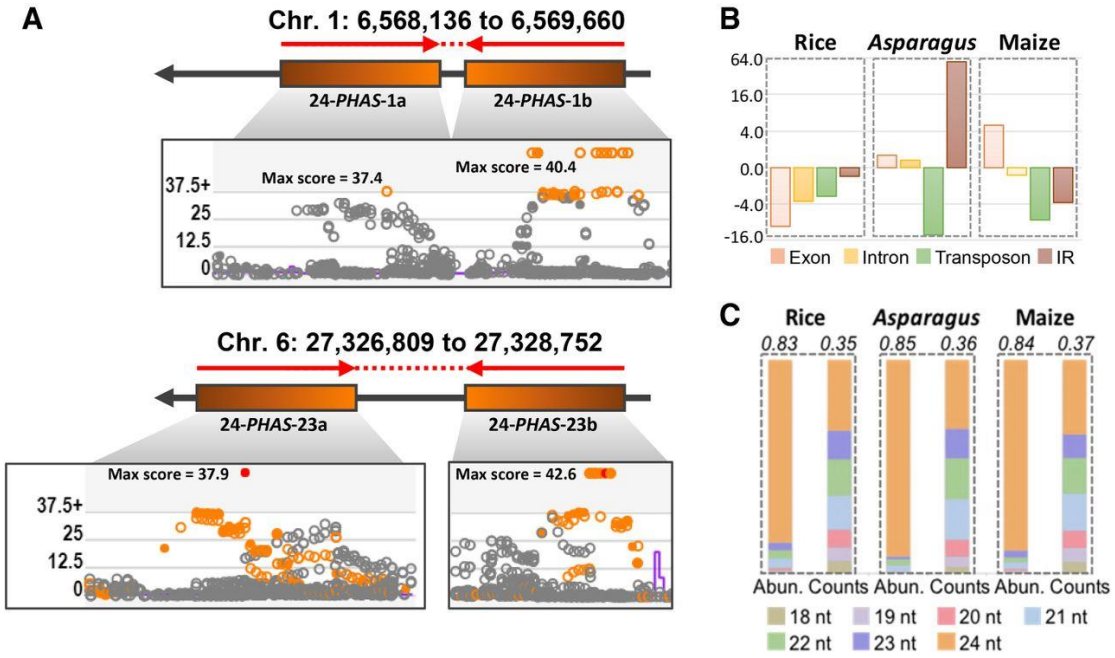

Figure 3.9: Many Asparagus 24-*PHAS* loci are derived from inverted repeats (A) Genomic organization of two representative inverted repeat-type *PHAS* loci, overlapping with the 5′ and 3′ arm of inverted repeats. Phasing scores from sliding window analysis are presented as shown in our custom web viewer. Loci with score >25 were considered as phased loci. (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, versus random chance. (C) Comparison of sRNA abundances and unique sRNA 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.

These 24-*PHAS* loci were processed by the locus-genomic feature clustering script to determine if there was any bias in the genomic features that these 24-*PHAS* loci overlap with (exons, introns, TE-related regions, and inverted repeats). Genome-wide, 90% of asparagus 24-*PHAS* loci corresponded with inverted repeats. We performed this same analysis utilizing annotated maize and rice 24-*PHAS* loci to determine what the overlap might look like in these genomes. These results were then compared utilizing an enrichment or depletion analysis of these 24-*PHAS* loci overlapping with genomic features compared to random chance for asparagus, maize, and rice (formula in **Section 3.6.2.1**). Both maize and rice had few 24-*PHAS* loci overlapping with inverted repeats compared to asparagus (**Figure 3.9B**). In all three species, there existed few 24-*PHAS* loci in TE-related regions and displayed similar distributions of sizes (**Figure 3.9C**), suggesting similar efficiencies of Dicer processing.

3.6.2.1 Computing degree of overlap between two genomic features

The enrichment or depletion of overlap between sRNA generating locations like miRNA or *PHAS* loci with 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 a locus (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

3.6.3 Discussion

Reproductive phasiRNAs are the least-characterized small RNA pathway in plants. First described in rice (Johnson et al., 2009), large-scale genetics, imaging and biochemical studies from rice and maize have defined their pattern of accumulation in developmental time and space. And yet despite their seemingly clear functional importance, they are still poorly understood (Fei et al., 2013; Komiya et al., 2014; Ono et al., 2018; Song, Wang, et al., 2012; Zhai et al., 2015).

The work described in this case study is just a small sampling of the results generated for this project. Our investigation of asparagus small RNAs identified the production of reproductive phasiRNAs from intramolecular duplex mRNAs. The work in this case study describes the identification of a previously unobserved biogenesis pathway for the production of 24-nt meiotic phasiRNAs in asparagus. This process generates inverted repeat-derived phasiRNAs from mRNAs lacking a miR2275 target site. The grass 24-nt reproductive phasiRNAs and the asparagus phased inverted repeat derived siRNAs appear analogous based on similar spatial localization and temporal dynamics. The precise nature of the triggering mechanism for the inverted repeat derived siRNAs that guides their processing, such as a structural motif, has yet to be determined. In conclusion, our characterization of reproductive phasiRNAs in diverse monocots is consistent with a broad role of these pathways in germinal development.

Chapter 4

DISCUSSION AND CONCLUSIONS

The Human Genome Project was initiated in 1990 by the NIH and DOE. The first drafts of the human genome were completed in 2001 (Lander et al., 2001; Venter et al., 2001), and the project was declared complete in April 2003 with a total cost of nearly $3 billion. Since the work of these researchers, sequencing technologies have improved greatly. Today, high-throughput sequencing technologies have drastically reduced both the financial cost of sequencing a genome and the time required to sequence that genome. This evolution in sequencing has resulted in the production of massive amounts of data that is generated which has brought on the need for various bioinformatics algorithms and tools to assist in the analysis of this data. At the time of writing, the Meyers lab has 425 expression databases comprised of primarily of sRNA-seq, PARE, and RNA-seq libraries. These databases are comprised of a total of 6806 total sequencing libraries. These data were primarily generated for single analyses, but these datasets can be reused in other analyses to make findings in a completely unrelated area. To do this however, we must create versatile tools that deal with the constantly increasing volume of data that we as a scientific community generate. I have spent the last 7 years of my graduate career working to develop novel algorithms and tools that can do just that. Additionally, however, I have tried to develop these tools with the intent of being utilized beyond the cases of my specific interests. This allows these tools to be used in a wide variety of analyses that I may have not even imagined when I initially wrote them. I have also developed these tools

with extensive documentation to ensure the intentions of my code is clear to future developers who can both maintain and update these tools. In this dissertation, I discuss multiple projects in which I attempted to utilize computational tools to generate novel biological findings in a variety of plant systems.

4.1 Development of a novel miRNA prediction and validation pipeline

miRNAs are perhaps the most well studied class of small RNAs in plants. miRNAs regulate gene expression via posttranscriptional gene silencing, and they have been found to regulate important processes such as development, growth, and stress response. This role in gene regulation makes miRNAs highly interesting, but the ability to accurately identify these is a complicated process. There is significant diversity in sRNA sequencing libraries with hundreds of thousands to millions of unique sRNA sequences in any given library, and annotated miRNAs only represent a few hundred of those sequences. And while miRNAs are typically thought of as being 21-22 nucleotides in length, there is some variability here, and there are other classes of sRNAs that exist at this length as well. This makes the annotation of miRNAs difficult, but by studying miRNAs that have been identified over the past decade, researchers have been able to identify several rules that can be utilized to computational predict miRNAs in plants with a high level of accuracy.

Utilizing these rules, I developed *miRador* to quickly and accurately predict the presence of miRNAs in sRNA libraries in a variety of plant species. I developed this tool to be both computationally quick while also being memory efficient to allow for it to be utilized in a variety of computational environments. I developed an annotation script and integrated it directly within *miRador* to provide users with insights of the classification of its predicted miRNAs. I also exported this tool to be

utilized for the predictions of two of the most commonly utilized miRNA prediction tools for plants, *ShortStack* and *mirDeep-P2* to provide users with proper miRNA annotations of the predictions that these tools. I compared the results of *miRador* to *ShortStack* and *miRdeep-P2* to determine the quality of its predictions in comparison to these tools. Finally, I utilized *sPARTA*, a target prediction and validation tool that I co-wrote, in an attempt to validate several novel miRNAs in maize. With *miRador*, researchers will be able to identify miRNAs that are present in their datasets with a high level of confidence. Coupling these predictions with *sPARTA* will reduce false positives even further. If a user should choose to do so, they may replace the utility of *miRador* with another miRNA prediction program while still utilizing the annotation component of *miRador*. While my bias for *miRador* does exist, I cannot ignore that both *ShortStack* and *mirDeep-P2* produce high quality results. My ultimate goal has always been for miRNAs be to quickly and accurately identified while producing few false positives, and I believe that all three tools do a good job at fulfilling this goal.

4.1.1 Enhancements to *miRador*

The development of *miRador* was inspired by the publication of plant miRNA annotation criteria that were developed to prevent further false annotations of miRNAs while also simultaneous preventing true miRNAs from being classified as false negatives. While I believe *miRador* made significant strides in upholding these hopes, I do believe that there are enhancements that could be made to *miRador* to increase its predictive power. Both of these ideas would utilize machine learning techniques with the intent of improving the miRNA identification power of *miRador*.

The first area in which I believe *miRador* could be improved is in its method of identifying a candidate miRNA as true or false. In fact, this very binary nature of

classification of candidate miRNAs is the concept that I believe could be improved. Rather than utilizing elimination filters of miRNAs through the filters established by Blake Meyers and Mike Axtell, I believe that we can examine the known miRNA:miRNA* duplexes and their precursors in miRBase to provide further insights as to how similar candidate miRNAs are to known miRNAs. If these features could be plotted into a probability distribution, we could then assign a confidence value to the predicted miRNAs. Such a change would effectively eliminate the filtering criteria that this tool was originally developed to model, but this modeling method would allow for a level of fluidity with miRNA prediction that remains unseen in miRNA prediction tools today.

The second modification that I believe would greatly improve *miRador* would be the use of machine learning techniques to identify additional features that could be utilized to predict miRNAs. Like the previous suggestion, this change would help non-canonical miRNAs appear in the output of *miRador*, and likely with greater confidence than before. By adding more features, it is possible that we can improve the predictive power of *miRador* to greater levels while potentially identifying novel miRNAs that do not fit our current models for miRNA prediction.

4.2 Utilizing clustering techniques to characterize 24-nt phasiRNA producing loci and the impact of those 24-nt phasiRNAs

siRNAs are the most abundant class of sRNAs, and they encapsulate some very interesting, but very poorly understood sub-classes of siRNAs. siRNAs are often found to be clustered in regions of the genome and their silencing effects are often found to be the cumulative effect on a locus rather than the result of an individual siRNA. I have been involved in numerous analyses in which we tried to characterize

siRNAs impact on an organism, extending beyond just my work presented here in this Ph.D dissertation. In the analyses presented in this dissertation, however, I was focused on studying phasiRNAs and the loci from which they are derived. I developed a clustering script that was designed to function on a specific set of loci and attempt to characterize common genomic features in which these loci overlap with. Through the use of MySQL, however, I was able to extend this tool's utility beyond its initial use with little modifications to be used in two published projects, and several others.

The first of these projects was an analysis to characterize the effect that 24-nt phasiRNAs have on maize anthers. Four *dcl5* mutant alleles identified to be male sterile. Through the use of clustering, I was able to quantify the loss of accumulation of 24-nt phasiRNAs as in each of these mutant alleles as compared to heterozygous controls. Through the use of similar clustering techniques, I quantified the abundance of *24-PHAS* precursors to identify if these precursors were present at higher abundance levels than normal plants, as would be observed in plants with impaired miRNA biogenesis. In this analysis, however, we only saw a modest impact on *24-PHAS* accumulation while 24-nt phasiRNAs were severely impacted. Surprisingly, under cooler growth temperatures, we found that phasiRNA accumulation is still prevented, however, we find that male fertility remains. Thus, we hypothesize that 24-nt phasiRNAs are required for male fertility at warmer temperatures, those typically seen in the U.S. corn belt. Finally, RNA-seq transcript analysis found few genes that were differentially expressed suggesting that 24-nt phasiRNAs are distinct from tasiRNAs which do have specific mRNA targets and modulate transcript abundances directly.

The second project involved another analysis of 24-nt phasiRNAs, but the nature of this analysis different from the first case study. In this analysis, we utilized *sPARTA* to identify miRNA triggers of computationally identified *24-PHAS* loci in asparagus. This analysis, however, failed to identify a trigger from any known miRNA. Upon manual analysis of these loci, we identified that many of these loci overlapped with inverted repeats. By modifying the clustering program used in the first case study, I was able to produce a quantifiable value for each predicted *24-PHAS* locus as the number of nucleotides that overlapped with every genomic feature. By performing this analysis with maize, rice, and asparagus *24-PHAS* precursors, I was able to identify a significant bias of these precursors in asparagus with inverted repeats which are hardly present in maize and rice. The results of this analysis showed that asparagus generates inverted repeat-derived phasiRNAs from mRNAs lacking a miR2275 target site.

4.3 Deployment of tools

4.3.1 Deployment of miRNA analytical tools

miRador and *sPARTA* have both been made available for download on Github, a web hosting service for Git, a version control tool. This method for deployment makes the code easily accessible by all potential users, while also making any future updates easily propagated to current users. They are both available via GNU Public License (v3) allowing them to be utilized in a variety of ways to researchers across the world. *miRador* is accessible on my personal Github page: https://github.com/rkweku/miRador. *sPARTA* is accessible on Atul Kakrana's Github page, https://github.com/atulkakrana/sPARTA.

4.3.2 Deployment of clustering and differential expression analysis tools

The analytical tools described in **Chapter 3** are available on Github, though these are housed on a private Github repository because their function depends on the proprietary database schema utilized by the Meyers lab, as mentioned previously.

4.4 Summary

This dissertation presents several new tools and pipelines that can be utilized to study small RNAs in all plants. The works presented in **Chapter 2** detail a novel miRNA prediction tool and pipeline to identify new miRNAs. This tool was developed with reusability in mind, and thus it was developed to be quick and efficient so that it may be scaled to larger genomes, and larger datasets with little compromise. It also attempts to ensure high quality predictions are made by utilizing strict replication requirements that many tools do not utilize. In addition to the miRNA prediction tool, I also detail a miRNA-target prediction and validation tool that I co-developed with the intent of using its predictions to validate the action of novel candidate miRNAs. The works presented in **Chapter 3** detail two biological case studies in which we sought to answer specific biological questions with a clustering tool that I developed to utilize the Meyers Lab database system. While this tool is thus specific to Meyers Lab members only, it has significant modularity that allows it to be utilized in a wide variety of analyses. I have constantly tried to develop tools with reusability in mind. Thus, I hope that the tools presented in this dissertation will continue to be used by other scientists in the future. I have presented future directions that can be taken with these tools, and I have developed these tools with significant documentation to assist future developers with the modifications of these tools after I leave.

REFERENCES

Addo-Quaye, C., Miller, W., & Axtell, M. J. (2009). CleaveLand: A Pipeline for Using Degradome Data to Find Cleaved Small RNA Targets. *Bioinformatics, 25*(1), 130-131. doi:10.1093/bioinformatics/btn604

Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J. E., White, J., . . . Chandler, V. L. (2006). An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature, 442*(7100), 295-298. doi:10.1038/nature04884

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research, 25*(17), 3389-3402. doi:10.1093/nar/25.17.3389

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., . . . Tuschl, T. (2003). A uniform system for microRNA annotation. *RNA, 9*(3), 277-279. doi:10.1261/rna.2183803

Arikit, S., Xia, R., Kakrana, A., Huang, K., Zhai, J., Yan, Z., . . . Meyers, B. C. (2014). An Atlas of Soybean Small RNAs Identifies Phased siRNAs from Hundreds of Coding Genes. *The Plant Cell, 26*(12), 4584-4601. doi:10.1105/tpc.114.131847

Arteaga-Vazquez, M., Sidorenko, L., Rabanal, F. A., Shrivistava, R., & Nobuta, K. (2010). RNA-mediated *trans*-communication can establish paramutation at the *b1* locus in maize. *Proc Natl Acad Sci U S A, 107*(29), 12986-12991.

Axtell, M. J. (2013). Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol, 64*, 137-159. doi:10.1146/annurev-arplant-050312-120043

Axtell, M. J., Jan, C., Rajagopalan, R., & Bartel, D. P. (2006). A Two-Hit Trigger for siRNA Biogenesis in Plants. *Cell, 127*(3), 565-577. doi:https://doi.org/10.1016/j.cell.2006.09.032

Axtell, M. J., & Meyers, B. C. (2018). Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. *The Plant Cell, 30*(2), 272-284. doi:10.1105/tpc.17.00851

Barbour, J.-E. R., Liao, I. T., Stonaker, J. L., Lim, J. P., Lee, C. C., Parkinson, S. E., . . . Hollick, J. B. (2012). *Rrequired to maintain repression2* is a novel protein that facilitates locus-specific paramutation in maize. *Plant Cell, 24*(5), 1761-1775.

Bolger, A. M., Usadel, B., & Lohse, M. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Borges, F., & Martienssen, R. A. (2015). The expanding world of small RNAs in plants. *Nat Rev Mol Cell Biol, 16*(12), 727-741. doi:10.1038/nrm4085

Brink, R. A. (1956). A genetic change associated with the *R* locus in maize which is directed and potentially reversible. *Genetics, 41*, 872-889.

Brousse, C., Liu, Q., Beauclair, L., Deremetz, A., Axtell, M. J., & Bouché, N. (2014). A non-canonical plant microRNA target site. *Nucleic acids research, 42*(8), 5270-5279. doi:10.1093/nar/gku157

Budak, H., & Akpinar, B. A. (2015). Plant miRNAs: biogenesis, organization and origins. *Funct Integr Genomics, 15*(5), 523-531. doi:10.1007/s10142-015-0451-2

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*(1), 421. doi:10.1186/1471-2105-10-421

Chandler, V., & Alleman, M. (2008). Paramutation: Epigenetic Instructions Passed Across Generations. *Genetics, 178*(4), 1839-1844.

Chávez Montes, R. A., Rosas-Cárdenas, d. F. F., De Paoli, E., Accerbi, M., Rymarquis, L. A., Mahalingam, G., . . . de Folter, S. (2014). Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. *Nat Commun, 5*, 3722.

Daxinger, L., Kanno, T., Bucher, E., van der Winden, J., Naumann, U., Matzke, A. J. M., & Matzke, M. (2009). A stepwise pathway for biogenesis of 24-nt secondary siRNAs and spreading of DNA methylation. *The EMBO journal, 28*(1), 48-57. doi:10.1038/emboj.2008.260

De Storme, N., & Geelen, D. (2014). The impact of environmental stress on male reproductive development in plants: biological processes and molecular mechanisms. *Plant, cell & environment, 37*(1), 1-18. doi:10.1111/pce.12142

Dezulian, T., Palatnik, J. F., Huson, D., & Weigel, D. (2005). Conservation and divergence of microRNA families in plants. *Genome Biol, 6*(11), 13. doi:10.1186/gb-2005-6-11-p13

Dorweiler, J. E., Carey, C. C., Kubo, K. M., Hollick, J. B., Kermicle, J. L., & Chandler, V. L. (2000). *mediator of paramutation1* is required for establishment and maintenance of paramutation at multiple maize loci. *Plant Cell, 12*(11), 2101-2118. doi:10.1105/tpc.12.11.2101

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

Dukowic-Schulze, S., Liu, C., & Chen, C. (2017). Not just gene expression: 3D implications of chromatin modifications during sexual plant reproduction. *Plant Cell Rep, 37*(1), 1-6. doi:10.1007/s00299-017-2222-0

Erhard, K. F. J., Stonaker, J. L., Parkinson, S. E., Lim, J. P., Hale, C. J., & Hollick, J. B. (2009). RNA polymerase IV functions in paramutation in *Zea mays*. *Science, 323*(5918), 1201-1206.

Fahlgren, N., & Carrington, J. C. (2010). miRNA Target Prediction in Plants. In B. C. Meyers & P. J. Green (Eds.), *Plant MicroRNAs: Methods and Protocols* (pp. 51-57). Totowa, NJ: Humana Press.

Fan, Y., Yang, J., Mathioni, S. M., Yu, J., Shen, J., Yang, X., . . . Zhang, Q. (2016). *PMS1T*, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. *Proceedings of the National Academy of Sciences, 113*(52), 15144-15149. doi:10.1073/pnas.1619159114

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

Folkes, L., Moxon, S., Woolfenden, H. C., Stocks, M. B., Szittya, G., Dalmay, T., & Moulton, V. (2012). PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome Sequencing. *Nucleic Acids Res, 40*(13), e103. doi:10.1093/nar/gks277

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., & Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology, 26*, 407. doi:10.1038/nbt1394
https://www.nature.com/articles/nbt1394#supplementary-information

Gabriel, J. M., Narain, A., Talbot, J.-E. R. B., Liao, I. T., Simon, S. A., Kong, G., . . . Hollick, J. B. (2015). *dicer-like3 is required for paramutation, small RNA biogenesis, and normal development in Zea mays*. Paper presented at the Maize Genetics Conference.

Gasciolli, V., Mallory, A. C., Bartel, D. P., & Vaucheret, H. (2005). Partially Redundant Functions of Arabidopsis DICER-like Enzymes and a Role for DCL4 in Producing trans-Acting siRNAs. *Current Biology, 15*(16), 1494-1500. doi:https://doi.org/10.1016/j.cub.2005.07.024

Gebert, D., Hewel, C., & Rosenkranz, D. (2017). Unitas: the universal tool for annotation of small RNAs. *BMC Genomics, 18*(1), 644. doi:10.1186/s12864-017-4031-9

German, M. A., Luo, S., Schroth, G., Meyers, B. C., & Green, P. J. (2009). Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat Protoc, 4*(3), 356-362.

Groszmann, M., Greaves, I. K., Albertyn, Z. I., Scofield, G. N., Peacock, W. J., & Dennis, E. S. (2011). Changes in 24-nt siRNA levels in Arabidopsis hybrids suggest an epigenetic contribution to hybrid vigor. *Proceedings of the National*

Academy of Sciences of the United States of America, 108(6), 2617-2622. doi:10.1073/pnas.1019217108

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

Hale, C. J., Erhard, K. F., Lisch, D., & Hollick, J. B. (2009). Production and processing of siRNA precursor transcripts from the highly repetitive maize genome. *PLoS Genet, 5*(8), e1000598. doi:10.1371/journal.pgen.1000598

Hale, C. J., Stonaker, J. L., Gross, S. M., & Hollick, J. B. (2007). A novel Snf2 protein maintains *trans*-generational regulatory states established by paramutation in maize. *PLoS Biol, 5*(10), e275.

Hollick, J. B. (2016). Paramutation and related phenomena in diverse species. *Nat Rev Genet, 18*(1), 5-23. doi:10.1038/nrg.2016.115

Hollick, J. B., & Chandler, V. L. (2001). Genetic factors required to maintain repression of a paramutagenic maize *pl1* allele. *Genetics, 157*(1), 369-378.

Hollick, J. B., Kermicle, J. L., & Parkinson, S. E. (2005). *Rmr6* maintains meiotic inheritance of paramutant states in *Zea mays*. *Genetics, 171*(2), 725-740. doi:10.1534/genetics.105.045260

Howell, M. D., Fahlgren, N., Chapman, E. J., Cumbie, J. S., Sullivan, C. M., Givan, S. A., . . . Carrington, J. C. (2007). Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in *Arabidopsis* reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell, 19*(3), 926-942. doi:10.1105/tpc.107.050062

Hunter, C., Willmann, M. R., Wu, G., Yoshikawa, M., Gutirrez-Nava, M. d. l. L., & Poethig, S. R. (2006). *Trans*-acting siRNA-mediated repression of ETTIN and ARF4 regulates heteroblasty in *Arabidopsis*. *Development, 133*(15), 2973-2981. doi:10.1242/dev.02491

Huntzinger, E., & Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet, 12*(2), 99-110.

Jeong, D.-H., Park, S., Zhai, J., Gurazada, S. G. R., De Paoli, E., Meyers, B. C., & Green, P. J. (2011). Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. *Plant Cell, 23*(12), 4185.

Johnson, C., Kasprzewska, A., Tennessen, K., Fernandes, J., Nan, G. L., Walbot, V., . . . Bowman, L. H. (2009). Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome Res, 19*(8), 1429-1440. doi:10.1101/gr.089854.108

Kakrana, A., Hammond, R., Patel, P., Nakano, M., & Meyers, B. C. (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), 2973-2981. doi:10.1093/nar/gku693

Kakrana, A., Li, P., Patel, P., Hammond, R., Anand, D., Mathioni, S., & Meyers, B. (2017). PHASIS: a computational suite for de novo discovery and

characterization of phased, siRNA-generating loci and their miRNA triggers. *bioRxiv*.
Kerpedjiev, P., Hofacker, I. L., & Hammer, S. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics, 31*(20), 3377-3379. doi:10.1093/bioinformatics/btv372
Komiya, R., Ohyanagi, H., Niihama, M., Watanabe, T., Nakano, M., Kurata, N., & Nonomura, K.-I. (2014). Rice germline-specific Argonaute MEL1 protein binds to phasiRNAs generated from more than 700 lincRNAs. *The Plant Journal, 78*(3), 385-397. doi:doi:10.1111/tpj.12483
Kuang, Z., Wang, Y., Li, L., & Yang, X. (2018). miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants. doi:10.1093/bioinformatics/bty972
Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . The Wellcome, T. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*, 357. doi:10.1038/nmeth.1923
https://www.nature.com/articles/nmeth.1923#supplementary-information
Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(25).
Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet, 11*(3), 204-220. doi:10.1038/nrg2719
Li, S., Le, B., Ma, X., Li, S., You, C., Yu, Y., . . . Chen, X. (2016). Biogenesis of phased siRNAs on membrane-bound polysomes in *Arabidopsis*. *eLife, 5*, e22750. doi:10.7554/elife.22750
Li, Y., Varala, K., Moose, S. P., & Hudson, M. E. (2012). The inheritance pattern of 24 nt siRNA clusters in arabidopsis hybrids is influenced by proximity to transposable elements. *PloS one, 7*(10), e47043-e47043. doi:10.1371/journal.pone.0047043
Liu, B., Chen, Z., Song, X., Liu, C., Cui, X., Zhao, X., . . . Cao, X. (2007). *Oryza sativa Dicer-like4* reveals a key role for small interfering RNA silencing in plant development. *Plant Cell, 19*(9), 2705-2718. doi:10.1105/tpc.107.052209
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology, 15*(12), 550-550. doi:10.1186/s13059-014-0550-8
Luo, J., & Hall, B. D. (2007). A multistep process gave rise to RNA polymerase IV of land plants. *J Mol Evol, 64*(1), 101-112.
Maher, C., Timmermans, M., Stein, L., & Ware, D. (2004). *Identifying microRNAs in plant genomes.* Paper presented at the Proc IEEE Comput Syst Bioinform Conf.

Margis, R., Fusaro, A. F., Smith, N. A., Curtin, S. J., Watson, J. M., Finnegan, E. J., & Waterhouse, P. M. (2006). The evolution and diversification of dicers in plants. *FEBS Lett, 580*(10), 2442-2450. doi:10.1016/j.febslet.2006.03.072

Matzke, M. A., & Mosher, R. A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet, 15*(6), 394-408. doi:10.1038/nrg3683

Meng, Y., Shao, C., Wang, H., & Chen, M. (2012). Are all the miRBase-registered microRNAs true? *RNA Biology, 9*(3), 249-253. doi:10.4161/rna.19230

Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., . . . Zhu, J.-K. (2008). Criteria for Annotation of Plant MicroRNAs. *The Plant Cell, 20*(12), 3186-3190. doi:10.1105/tpc.108.064311

Nobuta, K., Lu, C., Shrivastava, R., & Pillay, M. a. (2008). Distinct size distribution of endogeneous siRNAs in maize: evidence from deep sequencing in the *mop1-1* mutant. *Proc Natl Acad Sci U S A, 105*(39), 14958-14963. doi:10.1073/pnas.0808066105

Nonomura, K.-I. (2018). Small RNA pathways responsible for non-cell-autonomous regulation of plant reproduction. *Plant Reproduction, 31*(1), 21-29. doi:10.1007/s00497-018-0321-x

Nonomura, K.-I., Morohoshi, A., Nakano, M., Eiguchi, M., Miyao, A., Hirochika, H., & Kurata, N. (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), 2583-2594. doi:10.1105/tpc.107.053199

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

Parent, J.-S., Bouteiller, N., Elmayan, T., & Vaucheret, H. (2015). Respective contributions of Arabidopsis DCL2 and DCL4 to RNA silencing. *The Plant Journal, 81*(2), 223-232. doi:doi:10.1111/tpj.12720

Park, M. Y., Wu, G., Gonzalez-sulser, A., Vaucheret, H., & Poethig, R. S. (2005). Nuclear processing and export of microRNAs in *Arabidopsis*. *Proc Natl Acad Sci U S A, 102*(10), 3691-3696.

Pikaard, C. S., Haag, J. R., Ream, T., & Wierzbicki, A. T. (2008). Roles of RNA polymerase IV in gene silencing. *Trends Plant Sci, 13*(7), 390-397.

Pikaard, C. S., & Tucker, S. (2009). RNA-silencing enzymes Pol IV and Pol V in maize: more than one flavor? *PLoS Genet, 5*(11), e1000736. doi:10.1371/journal.pgen.1000736

R Core Team. (2018). R: A language and environment for statistical computing. Available online at https://www.R-project.org/. R Foundation for Statistical Computing.

Rice, P., Longden, I., & Alan, B. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet, 16*(6), 276-277.

Rogers, K., & Chen, X. (2013). Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell, 25*(7), 2383-2399. doi:10.1105/tpc.113.113159

Shahid, S., & Axtell, M. J. (2014). Identification and annotation of small RNA genes using ShortStack. *Methods, 67*(1), 20-27.

Shivaprasad, P. V., Chen, H.-M., Patel, K., Bond, D. M., Santos, B. A. C. M., & Baulcombe, D. C. (2012). A MicroRNA Superfamily Regulates Nucleotide Binding Site–Leucine-Rich Repeats and Other mRNAs. *The Plant Cell, 24*(3), 859-874. doi:10.1105/tpc.111.095380

Sidorenko, L., Dorweiler, J. E., Cigan, A. M., Arteaga-Vazquez, M., Vyas, M., Kermicle, J., . . . Chandler, V. L. (2009). A dominant mutation in *mediator of paramutation2*, one of three second-largest subunits of a plant-specific RNA polymerase, disrupts multiple siRNA silencing processes. *PLoS Genet, 5*(11), e1000725. doi:10.1371/journal.pgen.1000725

Sloan, A. E., Sidorenko, L., & McGinnis, K. M. (2014). Diverse gene-silencing mechanisms with distinct requirements for RNA polymerase subunits in *Zea mays*. *Genetics, 198*(3), 1031-1042.

Smyth, G. K., Shi, W., & Liao, Y. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics, 30*(7), 923-930. doi:10.1093/bioinformatics/btt656

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

Song, X., Wang, D., Ma, L., Chen, Z., Li, P., Cui, X., . . . Cao, X. (2012). Rice RNA-dependent RNA polymerase 6 acts in small RNA biogenesis and spikelet development. *The Plant Journal, 71*(3), 378-389. doi:doi:10.1111/j.1365-313X.2012.05001.x

Stonaker, J. L., Lim, J. P., Erhard, K. F., & Hollick, J. B. (2009). Diversity of Pol IV function is defined by mutations at the maize *rmr7* locus. *PLoS Genet, 5*(11), e1000706. doi:10.1371/journal.pgen.1000706

Sun, F., Guo, G., Du, J., Guo, W., Peng, H., Ni, Z., . . . Yao, Y. (2014). Whole-genome discovery of miRNAs and their targets in wheat (Triticum aestivum L.). *BMC plant biology, 14*, 142-142. doi:10.1186/1471-2229-14-142

Tamim, S., Cai, Z., Mathioni, S., Zhai, J., Teng, C., Zhang, Q., & Meyers, B. C. (2018). The basis of accumulation differences in plant 21-nt reproductive phasiRNAs, and their *cis*-directed activity. *bioRxiv*.

Taylor, R. S., Tarver, J. E., Foroozani, A., & Donoghue, P. C. J. (2017). MicroRNA annotation of plant genomes − Do it right or not at all. *BioEssays, 39*(2), 1600113. doi:10.1002/bies.201600113

Taylor, R. S., Tarver, J. E., Hiscock, S. J., & Donoghue, P. C. J. (2014). Evolutionary history of plant microRNAs. *Trends in Plant Science, 19*(3), 175-182. doi:https://doi.org/10.1016/j.tplants.2013.11.008

Teng, C., Zhang, H., Hammond, R., Huang, K., Meyers, B., & Walbot, V. (2018). *Dicer-like 5* deficiency confers temperature-sensitive male sterility in maize. *bioRxiv*, 498410. doi:10.1101/498410

Thieme, C. J., Gramzow, L., Lobbes, D., & Theißen, G. (2011). SplamiR—prediction of spliced miRNAs in plants. *Bioinformatics, 27*(9), 1215-1223. doi:10.1093/bioinformatics/btr132

Thompson, B. E., Basham, C., Hammond, R., Ding, Q., Kakrana, A., Lee, T.-F., . . . Hake, S. (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), 4702-4717. doi:10.1105/tpc.114.132670

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England), 25*(9), 1105-1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology, 28*(5), 511-515. doi:10.1038/nbt.1621

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The Sequence of the Human Genome. *Science, 291*(5507), 1304-1351. doi:10.1126/science.1058040

Vernoud, V., Laigle, G., Rozier, F., Meeley, R. B., Perez, P., & Rogowsky, P. M. (2009). The HD-ZIP IV transcription factor OCL4 is necessary for trichome patterning and anther development in maize. *The Plant Journal, 59*(6), 883-894. doi:doi:10.1111/j.1365-313X.2009.03916.x

Wang, F., & Axtell, M. J. (2017). *AGO4* is specifically required for heterochromatic siRNA accumulation at Pol V-dependent loci in *Arabidopsis thaliana*. *Plant J, 136*(4), 669-687. doi:10.1111/tpj.13463

Wu, Z. (2014). *Discovery and Evolutionary Analyses of Abiotic Stress-Regulated MicroRNAs in Crops (Doctoral dissertation).* University of California Riverside, Retrieved from https://search.proquest.com/docview/1564217915

Xia, R., Xu, J., Arikit, S., & Meyers, B. C. (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, 32*(11), 2905-2918. doi:10.1093/molbev/msv164

Xia, R., Xu, J., & Meyers, B. C. (2017). The emergence, evolution, and diversification of the *miR390-TAS3-ARF* pathway in land plants. *Plant Cell, 29*(6), 1232-1247. doi:10.1105/tpc.17.00185

Xia, Z., Zhao, Z., Li, M., Chen, L., Jiao, Z., Wu, Y., . . . Fan, Z. (2018). Identification of miRNAs and their targets in maize in response to Sugarcane mosaic virus infection. *Plant Physiology and Biochemistry, 125*, 143-152. doi:https://doi.org/10.1016/j.plaphy.2018.01.031

Xie, Z., Johansen, L. K., Gustafson, A. M., Kasschau, K. D., Lellis, A. D., Zilberman, D., . . . Carrington, J. C. (2004). Genetic and Functional Diversification of

Small RNA Pathways in Plants. *PLOS Biology, 2*(5), e104. doi:10.1371/journal.pbio.0020104

Yang, X., & Li, L. (2011). miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics, 27*(18), 2614-2615. doi:10.1093/bioinformatics/btr430

Yifhar, T., Pekker, I., Peled, D., Friedlander, G., Pistunov, A., Sabban, M., . . . Eshed, Y. (2012). Failure of the tomato *trans*-acting short interfering RNA program to regulate AUXIN RESPONSE FACTOR3 and ARF4 underlies the wiry leaf syndrome. *Plant Cell*, *24*(9), 3575-3589. doi:10.1105/tpc.112.100222

Yu, J., Han, J., Kim, Y.-J., Song, M., Yang, Z., He, Y., . . . Zhang, D. (2017). Two rice receptor-like kinases maintain male fertility under changing temperatures. *Proceedings of the National Academy of Sciences, 114*(46), 12327-12332. doi:10.1073/pnas.1705189114

Zhai, J., Jeong, D.-H., De Paoli, E., Park, S., Rosen, B. D., Li, Y., . . . Meyers, B. C. (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), 2540-2553. doi:10.1101/gad.177527.111

Zhai, J., Zhang, H., Arikit, S., Huang, K., Nan, G.-L., Walbot, V., & Meyers, B. C. (2015). Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proc Natl Acad Sci U S A, 112*(10), 3146-3151. doi:10.1073/pnas.1418918112

Zhang, L., Chia, J.-M., Kumari, S., Stein, J. C., Liu, Z., Narechania, A., . . . Ware, D. (2009). A genome-wide characterization of microRNA genes in maize. *PLoS Genet, 5*(11), e1000716.

Zhang, Y., Xia, R., Kuang, H., & Meyers, B. C. (2016). The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them. *Molecular Biology and Evolution, 33*(10), 2692-2705. doi:10.1093/molbev/msw154

Zhao, X.-Y., Wang, X.-Y., Zhao, L., Zhang, X.-M., Chen, S.-Y., Ma, P.-F., . . . Guo, Z.-H. (2015). Investigating the MicroRNAomes of Two Developmental Phases of Dendrocalamus latiflorus (Poaceae: Bambusoideae) Inflorescences. *Plant Molecular Biology Reporter, 33*(4), 1141-1155. doi:10.1007/s11105-014-0808-z

Zheng, Y., Li, Y. F., Sunkar, R., & Zhang, W. (2012). SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic Acids Res, 40*(4), e28. doi:10.1093/nar/gkr1092

Appendix

PERMISSIONS

Permission for published works in Chapter 1.

This Agreement between University of Delaware -- Reza Hammond ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.	
License Number	4626080953070
License date	Jul 11, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Maize Small RNAs as Seeds of Change and Stability in Gene Expression and Genome Stability
Licensed Content Author	Reza Hammond, Chong Teng, Blake C. Meyers
Licensed Content Date	Jan 1, 2018
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	yes
Title	Bioinformatics Tools to Classify and Characterize Plant Small RNAs
Institution name	University of Delaware
Expected presentation date	Jul 2019

Permission for published works in Chapter 2.6

This Agreement between University of Delaware -- Reza Hammond ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	4626090520582
License date	Jul 11, 2019
Licensed Content Publisher	Oxford University Press
Licensed Content Publication	Nucleic Acids Research
Licensed Content Title	*sPARTA*: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software
Licensed Content Author	Kakrana, Atul; Hammond, Reza
Licensed Content Date	Aug 12, 2014
Licensed Content Volume	42
Licensed Content Issue	18
Type of Use	Thesis/Dissertation
Requestor type	Author of this OUP content
Format	Print and electronic
Portion	Text Extract
Number of pages requested	12
Will you be translating?	No
Title	Bioinformatics Tools to Classify and Characterize Plant Small RNAs
Institution name	University of Delaware
Expected presentation date	Jul 2019

Permissions for published works in 3.5

Permissions for published works in 3.6