The Pennsylvania State University

The Graduate School

The Integrative Biosciences Program

FUNCTIONAL AND EVOLUTIONARY GENOMICS OF PLANT SMALL RNAS

A Dissertation in

Integrative Biosciences

by

Zhaorong Ma

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2013

The dissertation of Zhaorong Ma was reviewed and approved* by the following:

Michael J. Axtell
Associate Professor of Biology
Dissertation Advisor
Chair of Committee

Naomi Altman
Professor of Statistics

Claude dePamphilis
Professor of biology

Ross Hardison
T. Ming Chu Professor of Biochemistry and Molecular Biology

Peter J. Hudson
Professor and Department Head
The Huck Institutes of the Life Sciences

*Signatures are on file in the Graduate School

ABSTRACT

In plants, microRNAs (miRNAs) and small interfering RNAs (siRNAs) account for the majority of the small RNA population. They play critical roles in multiple cellular processes through post-transcriptional regulation of RNA targets. Some miRNAs are well conserved among different plant lineages, while others are less conserved. It is not clear whether less-conserved miRNAs have the same functionality as the well conserved ones. Heterochromatic siRNAs are broadly produced in the *Arabidopsis thaliana* genome, sometimes from active “hotspot” loci. It is unknown whether individual heterochromatic siRNA hotspots are retained as hotspots between plant species. In Chapter 2, we compare small RNAs in two closely related species (*Arabidopsis thaliana* and *Arabidopsis lyrata*) and find that less-conserved miRNAs have high rates of divergence in *MIRNA* hairpin structures, mature miRNA sequences, and target complementary sites in the other species. The fidelity of miRNA biogenesis from many less-conserved *MIRNA* hairpins frequently deteriorates in the sister species relative to the species of first discovery. We also observe that heterochromatic siRNA occupied loci have a slight tendency to be retained as heterochromatic siRNA loci between species, but the most active *A. lyrata* heterochromatic siRNA hotspots are generally not syntenic to the most active heterochromatic siRNA hotspots of *A. thaliana*. Altogether, our findings indicate that many *MIRNA*s and most heterochromatic siRNA hotspots are rapidly changing and evolutionarily transient within the *Arabidopsis* genus.

Small RNAs are broadly present in all known plant species, many of which play important regulatory roles. In Chapter 3, we surveyed the small RNA populations from three plants: the tree crop *Theobroma cacao*, oil palm *Elaeis guineensis* Jacq., and the model moss *Physcomitrella patens*. In *Theobroma cacao*, we computationally identified 83 conserved miRNAs and 91 miRNA targets using sequence similarity and secondary structure information. In oil palm (*Elaeis guineensis Jacq.*), we identified 28 expressed miRNA families during flower development by analyzing smallRNAseq data. In *Physcomitrella patens*, *w*e identified a novel family of trans-acting siRNA (ta-siRNA) loci associated with miR156- and miR529-directed slicing by scanning the genome for ta-siRNA-like sRNA accumulation patterns in different genetic background. These studies

as a whole demonstrate that many small RNA species are deeply conserved in the plant kingdom. On the other hand, novel classes of small RNAs can evolve in specific lineages.

Conserved plant microRNAs (miRNAs) modulate important biological processes but little is known about conserved cis-regulatory elements (CREs) surrounding *MIRNA* genes. In Chapter 4, we developed a solution-based targeted genomic enrichment methodology to capture, enrich and sequence flanking genomic regions surrounding conserved *MIRNA* genes with a locked-nucleic acid (LNA)-modified, biotinylated probe complementary to the mature miRNA sequence. Genomic DNA bound by the probe is captured by streptavidin-coated magnetic beads, amplified, sequenced and assembled *de novo* to obtain genomic DNA sequences flanking *MIRNA* locus of interest. We demonstrate the effectiveness of this method in *Arabidopsis thaliana*. We demonstrate the sensitivity and specificity of this enrichment methodology to enrich targeted regions spanning 10-20 kb surrounding known *MIR166* and *MIR165* loci. Assembly of the sequencing reads successfully recovered all targeted loci. While further optimization for larger, more complex genomes is needed, this method may enable determination of flanking genomic DNA sequence surrounding a known core (like a conserved mature miRNA) from multiple species that currently don't have a full genome assembly available.

Altogether, by sequencing data analysis and comparative genomics, these studies contribute to the understanding of the function and evolution of plant small RNAs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am extremely grateful to my thesis advisor, Michael Axtell, for his continuous guidance, encouragement, patience and inspirations during my graduate study. I would like to thank all members of my committee, Naomi Altman, Claude dePamphilis and Ross Hardison, for their support and suggestions on my thesis project.

I would like to thank all members of the Axtell lab. My special thanks goes to Charles Addo-Quaye for his guidance and help on all my computer-science-related questions, Jo Ann Snyder, Ceyda Coruh, Joseph Cho, Cathy Lin, Qikun Liu and Feng Wang for their help with wet-lab techniques, Saima Shahid and Hang Zhang for discussions and inspirations of bioinformatics techniques.

I would like to thank the collaborators of the Axtell lab, Jennifer Ann Harikrishna, Mark Guiltinan, and Wolfgang Frank for giving me the opportunity to work with different types of data from various plant species. I am grateful to have a group of smart and thoughtful classmates in the Bioinformatics and Genomics program, Zhenhai Zhang, Ti-Cheng Chang and Venkatesh Moktali, who helped me through the touch start of my graduate study. I also deeply appreciate my friends Liang Song and Yiliang Ding for sharing their experiences and ideas as senior students.

I would like to thank all my friends and family members for making my life a wonderful experience. I'm indebted to my parents, whose love and pattern for science inspired me to pursue a career in scientific research. I'd like to give special thanks to my husband Tao Zhang and my friend Kunxuan Chen for their encouragement during the process of writing this thesis.

Chapter 1
Introduction

1.1 Functions and evolution of plant small RNAs

1.1.1 Overview of small RNAs

Small RNAs are ~20-30 nucleotide (nt) non-coding RNAs that are prevalently present in metazoans (Kim et al., 2009; Carthew and Sontheimer, 2009) plants (Axtell, 2013a) and unicellular eukaryotes (Zhao et al., 2007; Molnár et al., 2007; Drinnenberg et al., 2009). To date, three major classes of small RNAs have been well characterized: microRNAs (miRNAs), short interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs). The three classes of small RNAs have distinctive their sizes, which are typically 21-22 nts for miRNAs, 21-24 nts for siRNAs and 24-30 nts for piRNAs (Stefani and Slack, 2008; Axtell, 2013a). miRNAs and siRNAs are processed from helical regions of RNA precursors by Dicer or Dicer-like proteins (DCLs) into 20-24-nt-long double-stranded duplexes, whose size range is a signature of Dicer slicing. In contrast, piRNAs derive from single-stranded, presumably non-helical RNA precursors without the requirement of Dicer (Juliano et al., 2011). Unlike the nearly ubiquitous presence of miRNAs and siRNAs in eukaryotes, piRNAs have only been observed in animals (Brennecke et al., 2007; Aravin et al., 2007; Czech et al., 2008; Gan et al., 2011). Despite the foregoing distinctions, miRNAs, siRNAs and piRNAs function in a similar way: associated with Piwi/Argonaute (AGO) family proteins, they recognize target transcripts in a sequence-specific manner, and subsequently regulate gene expression transcriptionally or post-transcriptionally. Small-RNA-directed gene regulation involves repressive chromatin modifications, mRNA destabilization, and translational inhibition, therefore small RNAs function as negative regulators of gene expression (Kim et al., 2009; Carthew and Sontheimer, 2009; Axtell, 2013a).

In plants, miRNAs and siRNAs account for the majority of the small RNA population. They are distinct in the form of their corresponding RNA precursors: miRNAs derive from single-stranded precursors with a hairpin structure, while siRNAs derive from double-stranded RNA (dsRNA) precursors (Axtell, 2013a). The core protein families responsible for small RNA biogenesis and function in plants are RNA-dependent RNA polymerases (RDRs), Dicer-like proteins (DCLs) and Argonautes (AGOs). RDRs produce dsRNAs by synthesizing the second strand from an RNA template, which is an essential step in the siRNA biogenesis pathway (Zong et al., 2009). DCLs, which are RNAse III endonucleases, slice RNA precursors into short double-stranded duplexes, typically 20-

24 nt long with a 3'-2 nt overhang (Margis et al., 2006). AGOs are the downstream effectors of the small RNA pathway. They bind one strand of the double-stranded duplexes and identify target RNAs based on sequence complementarity between the small RNA and the target RNA, directing repressive activities at the target RNA (Vaucheret, 2008; Wee et al., 2012). RDR, DCL, and AGO family proteins, each encoded by multiple paralogous loci in plants, have either redundant or specialized functions in small RNA pathways (Margis et al., 2006; Vaucheret, 2008; Zong et al., 2009). Despite similarities in the their biogenesis and function, miRNAs and siRNAs require distinct set of DCLs and AGOs for their biogenesis and downstream effects on their targets.

1.1.2 microRNAs (miRNAs)

miRNAs are a major class of small RNAs broadly present in metazoans and plants. The first miRNA *lin-4* was discovered two decades ago in the roundworm *Caenorhabditis elegans* (Wightman et al., 1993; Lee et al., 1993). *lin-4* is 22 nt long non-coding RNA deeply conserved in metazoans, and it silences its target genes post-transcriptionally. Later more miRNAs were discovered in the fruit fly *Drsophila melanogaster*, *C. elegans* and *Homo sapiens* (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001; Aravin et al., 2003). In 2002, plant miRNAs were first found in *Arabiodopsis thaliana* and *Oryza sativa* (rice) (Reinhart et al., 2002). Most plant miRNAs found to date are 21 nt in length, require DCL1 for their biogenesis, and require AGO1 for function(Axtell, 2013a).

1.1.2.1 Biogenesis of plant miRNAs

Primary transcripts of miRNA precursors (pri-miRNAs) are transcribed by RNA polymerase II (Pol II) mostly from un-clustered intergenic *MIRNA* loci, capped and polyadenylated (Lee et al., 2004). Part of the pri-miRNA forms a imperfect fold-back structure, which is processed into a hairpin-like stem-loop precursor (pre-miRNA) and further excised into the miRNA/miRNA* duplex by a DCL family protein (mainly DCL1) in nuclear processing centers called D-bodies (Kurihara and Watanabe, 2004; Fang and Spector, 2007). This process is facilitated by proteins including the RNA-binding protein DAWDLE (DDL), C2H2-zinc finger protein SERRATE (SE), the double-stranded RNA-binding protein HYPONASTIC LEAVES1 (HYL1) and nuclear cap-binding complex (CBC) (Kurihara et al., 2006; Gregory et al., 2008; Laubinger et al., 2008; Yu et al., 2008). The miRNA/miRNA* duplexes are methylated at the 3'-ends by the methyltransferase HEN1 to be protected from degradation by the SDN class of exonucleases (Park et al., 2002; Yu et al., 2005) and exported to the cytoplasm by HASTY, the exportin 5 ortholog of plants (Park et al., 2005). The guide strand of the miRNA/miRNA* duplex, i.e. the mature miRNA, is then incorporated into AGO proteins (mainly AGO1) to repress target genes, while the passenger strand called miRNA* is degraded (Voinnet, 2009).

1.1.2.2 Functions of plant miRNAs

miRNA-mediated target gene regulation involves both AGO-catalyzed slicing of the target mRNA and translational repression. Early discoveries revealed that target mRNAs were cleaved by AGO in the center of the regions of near-perfect sequence complementarity to the corresponding miRNAs, resulting in decreased level of the intact target mRNA (Llave et al., 2002; Souret et al., 2004). However, in other studies where protein levels of the miRNA targets were measured together with mRNA levels, a discrepancy was observed that protein levels were affected more strongly than mRNA levels (Aukerman and Sakai, 2003; Chen, 2004; Bari et al., 2006; Gandikota et al., 2007). These studies suggest that plant miRNAs repress targets genes by destabilizing mRNAs as well as inhibiting protein production. Since most miRNAs are incorporated into AGO1, two studies of *ago1* mutants further elucidate the essential role of AGO1 in both miRNA-mediated mRNA slicing and translational repression. Hypomorphic *ago1-27* mutants accumulate near normal levels of target mRNA but higher levels of protein, yet displays similar phenotype as *dcl1* mutants which broadly affects miRNA biogenesis (Brodersen et al., 2008). This indicates that defects in translation repression alone, despite intact slicing activity, causes morphological defects. Slicer-defective mutants of AGO1 cannot complement loss-of-function *ago1* mutants, indicating the essential role of AGO-catalyzed slicing (Carbonell et al., 2012). Taken together, plant miRNAs repress target genes by a blend of both mRNA destabilization and translational inhibition.

1.1.2.3 Complementarity requirements for miRNA targeting

Plant miRNAs exhibit extensive base-pairing to their cognate targets, as is evident in all verified miRNA-target pairs to date (Axtell, 2013a). Mismatches are better tolerated at the first nucleotide of complementarity counting from the 5' of the miRNA, as well as near the 3' end (Mallory et al., 2004; Schwab et al., 2005). The region between the 2nd and 13th nucleotide in the alignment is critical for function, where no more than one mismatches are tolerated. Upon recognition by an miRNA, target mRNA is sliced by AGO between 10th and 11th nucleotide in the alignment (Figure 1.1). In an alternative scenario where a central bulge or mismatches prevent AGO-catalyzed slicing, the target serves as a mimicry or decoy to sequester AGO-bound miRNAs and positively regulates the canonical miRNA targets (Figure 1.1; Franco-Zorrilla et al., 2007).

In animals, seed pairing is used for canonical miRNA/target recognition instead, where base pairing between nucleotides 2-7 of the miRNA is the sole requirement, with optional pairing between nucleotides 1 and 8 (Figure 1.1; Bartel, 2004). In a few cases, 3'-compensatory pairing and centered pairing are also functional in animals (Bartel, 2009; Shin et al., 2010). However, none of the animal-like complementarity patterns are reported in plants (Axtell, 2013a).

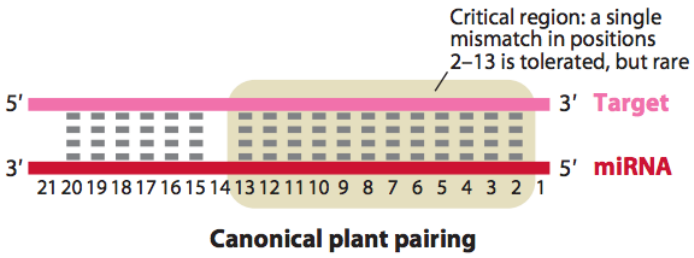
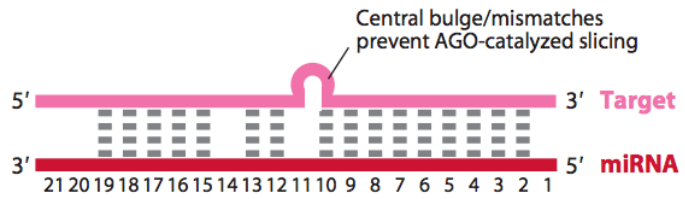
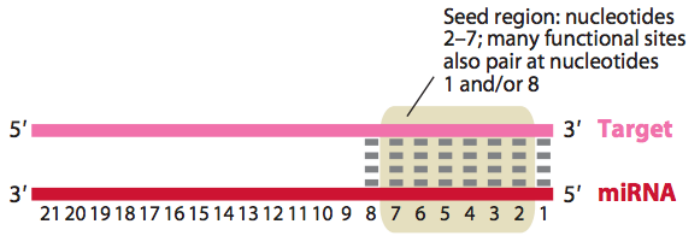
miRNA/target pairing pattern	Prevalence	Mechanisms of target repression
 **Canonical plant pairing**	Common in plants; very rare in animals	Slicing-dependent (and perhaps slicing-independent) reduction in mRNA accumulation, and/or translational repression
 **miRNA target-mimic pairing**	A few examples in plants; in animals, seed pairing is used instead	Competition for and sequestration of active AGO-miRNA complexes
 **Seed pairing (canonical animal pairing)**	No evidence in plants to date; common in animals	Translational repression followed by slicing-independent reduction in mRNA accumulation

Figure 1.1 Patterns of functional microRNA (miRNA) / target complementarity in plants and animals. Adapted from (Axtell, 2013a).

1.1.2.4 Evolution of plant miRNAs

The first miRNAs discovered in plants were found to be conserved between *Arabidopsis thaliana* and *Oryza sativa* and abundantly expressed (Reinhart et al., 2002; Mette et al., 2002; Park et al., 2002; Llave et al., 2002). Later studies revealed that quite a few *MIRNA* families are deeply conserved in plants, which are characterized with abundant expression levels, multiple paralogous loci and easily identifiable targets (Rajagopalan et al., 2006; Fahlgren et al., 2007; Axtell et al., 2007). Well-conserved miRNAs preferentially target transcription factors that regulate development and patterning (Garcia, 2008). Eight families were conserved in all embryophytes (informally, land plants) and ten families are present in all angiosperm (flowering plant) lineages (Cuperus et al., 2011). No shared *MIRNA* families have been identified between embryophytes and the unicellular green alga *Chlamydomonas reinhardtii* (Zhao et al., 2007; Molnár et al., 2007) nor between plants and animals, indicating that *MIRNA* genes may have arisen independently multiple times in eukaryote diversification (Axtell and Bowman, 2008; Cuperus et al., 2011).

Evidence has accumulated that many *MIRNA* families are species-specific or non-conserved (Griffiths-Jones et al., 2008; Cuperus et al., 2011), suggesting that large numbers of *MIRNA* loci are recently spawned. These relatively young miRNAs are lowly expressed, often encoded by a single locus and have few targets (Rajagopalan et al., 2006; Zhang et al., 2006; Fahlgren et al., 2007). How are these new *MIRNA* gens formed? Early studies found that both arms of the fold-back region of *ath-MIR161*, *ath-MIR163* and *ath-MIR822* share extensive sequence similarity to their target genes outside the mature miRNA complementarity sites, likely a result of inverted gene duplication (Allen et al., 2004; Fahlgren et al., 2007). Initial duplication events would generate loci with perfect self-complementarity and produce siRNAs, which is supported by the observations that fold-back regions of young *MIRNAs* are processed not by DCL1 but by DCL4, the Dicer responsible for siRNA production (Rajagopalan et al., 2006; Vazquez et al., 2008; Amor et al., 2009). These loci may accumulate drift mutations, and subsequently are picked up by DCL1-dependent machinery to produce miRNAs. Some may be selected if advantageous, while others will drift further without selection (Voinnet, 2009). An alternative possibility of *MIRNA* origin is spontaneous evolution from small random inverted repeats, which are abundantly present in plant genomes (Jones-Rhoades and Bartel, 2004). This is evident that over half of the young *Arabidopsis MIRNAs* have no sequence identity to other loci (Felippes et al., 2008; Fahlgren et al., 2010).

1.1.3 Small interfering RNAs (siRNAs)

The discovery of siRNAs was preceded by the observations that antisense RNAs expressed in transgenic plants induced gene siliencing (Ecker and Davis, 1986; Napoli et al., 1990). Later it was observed that non-coding regions of viral RNA sequences could enhance the resistance of plants to

viral infection (Covey et al., 1997), indicating a link between viral defense and post transcriptional gene silencing. In 1998, double-stranded RNA (dsRNA) was identified as the causative agent for post transcriptional gene silencing (PTGS), a phenomenon also known as RNA interference (RNAi) (Fire et al., 1998). RNAi is ubiquitous in eukaryotes including plants, metazoans and fungi (Fagard et al., 2000), the only exceptions being *Saccharomyces cerevisiae* (Aravind et al., 2000) and a few known trypanosome pathogens (Smith et al., 2007), likely due to lineage-specific loss. RNAi can be initiated both by exogenous dsRNAs (coming from virus infection or laboratory manipulations) and by endogenous dsRNAs known as small interfering RNAs (siRNAs).

In plants, endogenous siRNAs can be divided into three sub-categories based on their distinct biogenesis and function: heterochromatic siRNAs, secondary siRNAs, and natural antisense transcript siRNAs (NAT-siRNAs).

1.1.3.1 Heterochromatic siRNAs

Heterochromatic siRNAs are derived from intergenic or repetitive genomic regions. They are typically 24 nt in length, require RDR2 and DCL3 for their biogenesis and AGO4-clade AGOs (AGO4, AGO6 and AGO9 in *Arabidopsis*) for their function (Axtell, 2013a). The biogenesis of heterochromatic siRNAs are not fully elucidated. The current understanding is that repetitive intergenic regions are transcribed by RNA polymerase IV (Pol IV), a plant-specific RNA polymerase. The Pol IV transcript serves as a template for RDR2 to produce a dsRNA. DCL3 cleaves the resulting dsRNA into 24-nt siRNA products, and one strand of the siRNA is loaded into AGO4. RNA polymerase V (Pol V), another plant-specific RNA polymerase, produces transcripts that serve as scaffolds for the binding of siRNA-loaded AGO4. Pol V transcription is facilitated by the DDR complex comprising a putative chromatin-interacting ATPase (DRD1), a hinge-domain protein (DMS3), and a single-stranded DNA-binding protein (RDM1) and is independent of siRNA biogenesis. Upon binding to the Pol V transcripts, heterochromatic siRNAs function to trigger the de novo deposition of repressive chromatin modifications at nearby regions, including cytosine methylation at asymmetric CHH (where H=A, C or T) sites and H3K9 histone methylation (Zhong et al., 2012; Wierzbicki et al., 2012).

Heterochromatic siRNA pathways are deeply conserved in plants. 24-nt small RNAs are the most abundant small RNA population in numerous plant species, such as maize and rice, and their accumulation depends on the respective homologous DCL3 proteins (Nobuta et al., 2008; Wu et al., 2010). In the moss *Physcomitrella patens*, 23-24-nt siRNAs accumulate at repetitive intergenic regions, which is dependent on PpDCL3, the *P. patens* DCL3 homolog (Cho et al., 2008). However, in conifers, no 24-nt small RNA population has been identified, nor has a DCL3 homolog, suggesting that heterochromatic siRNAs may be lost or replaced by shorter RNAs (Morin et al., 2008; Dolgosheina et

al., 2008).

1.1.3.2 Secondary siRNAs and *trans*-acting siRNAs (ta-siRNAs)

Secondary siRNAs derive from RDR-sythesized double-stranded RNA precursors whose production is initiated by miRNAs or other secondary siRNAs. The resulting dsRNA is diced by a DCL into secondary siRNAs. Most secondary siRNAs are 21-nt long, requiring RDR6 and DCL4 for their biogenesis (Allen et al., 2005; Yoshikawa et al., 2005; Talmor-Neiman et al., 2006).

Some secondary siRNAs can act in *trans* to silence genes that are distinct from the loci of their origin, thus the name *trans*-acting siRNAs (ta-siRNAs). The loci where the ta-siRNAs derive are called *TAS* genes. Non-coding transcripts from *TAS* loci are recognized by an AGO-bound miRNA and AGO-slicing defines one end of the siRNA production. One fragment of the sliced transcript is converted into dsRNA by RDR6 and subsequently diced into ~ nt ta-siRNAs by DCL4. Many of them are phased, a result of successive DCL cleavage from a determined end of the dsRNA precursor (Allen et al., 2005; Yoshikawa et al., 2005; Talmor-Neiman et al., 2006).

A few ta-siRNAs regulate distinct targets. One well studied example is the miR390-triggered *TAS3*-produced ta-siRNAs that target *Auxin Response Factor 3 (ARF3)* and *ARF4*. The miR390-*TAS3* pathway is ancient and conversed between *Arabidopsis* and *Physcomitrella* (Axtell et al., 2006). In *Arabidospsis*, miR828-triggered *TAS4* ta-siRNAs target *MYB* transcription factor mRNAs (Luo et al., 2012). In some cases, secondary siRNAs can form a cascade to coordinate the repression of a large gene family. In *Arabidopsis*, secondary siRNAs are produced preferentially from a clade of the pentatricopeptide repeat (PPR) family that has recently undergone expansion. The secondary production at these loci are triggered by miR161 and TAS2-derived ta-siRNAs (Chen et al., 2007; Howell et al., 2007a). In multiple species, miR482/miR2118 triggers secondary siRNA biogenesis of the nucleotide-binding site--leucine-rich repeat (NBS-LRR) superfamily genes (Zhai et al., 2011; Li et al., 2012b; Shivaprasad et al., 2012). NBS-LRR disease-resistance genes are up-regulated when viral infections impair small RNA biogenesis and decrease miR482 and secondary siRNA accumulation at the NBS-LRR loci in tomato (Shivaprasad et al., 2012).

1.1.3.3 Natural antisense transcript siRNAs (NAT-siRNAs)

NAT-siRNAs are thought to derive from the double-stranded RNA formed by the hybridization of separately transcribed complementary RNAs. The two complementary RNAs are transcribed from opposite stranded of the same locus, and they produce *cis*-NAT-siRNAs (Borsani et al., 2005; Katiyar-Agarwal et al., 2006; Henz et al., 2007; Ron et al., 2010; Zhang et al., 2012). Hypothetically, the hybridizing RNAs can come from different locus to produce *trans*-NAT-siRNAs, but no *trans*-NAT-

siRNAs have been described in plants.

As NAT-siRNAs arise from hybridized RNAs, their production should not reply on RDR to synthesize the dsRNA precursor. However, in the several cases described to date, *cis*-NAT-siRNAs require either RDR2 or RDR6 for their biogenesis. They also have heterogeneous requirements for DCLs and other factors (Borsani et al., 2005; Katiyar-Agarwal et al., 2006; Ron et al., 2010). Genome-wide analyses shows that *cis*-NAT gene pairs correlated with lower small RNA densities, compared with non-overlapping gene pairs (Henz et al., 2007). In line with the above observation, only 6% and 16% of *Arabidopsis* and rice *cis*-NAT gene pairs were associated with substantial small RNA accumulation (Zhang et al., 2012). Thus, the trigger and machinery of *cis*-NAT-siRNA biogenesis remains elusive.

1.1.4 Interplay between the miRNA and siRNA pathway

The miRNA and siRNA pathway are intertwined in the regulatory networks. Biogenesis of ta-siRNAs, for example, require the interplay of both pathways (Allen et al., 2005; Yoshikawa et al., 2005; Talmor-Neiman et al., 2006). Moreover, the distinction between the miRNA and siRNA pathway is being blurred with the accumulating knowledge of the small RNA population in plants. One example is the discovery of long miRNAs in *Arabidopsis* and rice (Vazquez et al., 2008; Wu et al., 2010; Chellappan et al., 2010). These miRNAs are 24-nt long, require DCL3 for their biogenesis and AGO4 for their function to direct chromatin modifications at their target genes, all of which are signatures of the heterochromatic siRNA pathway. Most of the long miRNAs are evolutionarily young, implying that young *MIRNA* loci may be reprogrammed to enter the siRNA pathway. On the other hand, heterochromatic siRNA loci may also evolve into novel miRNA genes, as is observed in *Arabidopsis* and rice that a number of short nonautonomous DNA-type TEs known as MITEs encode both siRNAs and miRNAs (Piriyapongsa and Jordan, 2008).

1.2 Applications of next-generation sequencing technologies to small RNA research

1.2.1 Overview of next-generation sequencing technologies

The “first-generation” sequencing method, known as Sanger sequencing, was developed by Frederick Sanger using chain-terminating inhibitor dideoxynucleotides to terminate the chain amplification (Sanger et al., 1977). Sanger sequencing technology was extensively used in the human genome project (Collins et al., 2003), but it has high cost and low throughput. The advent of “next-generation” DNA sequencing technologies in the past decade enables rapid production of large amounts of sequence information at low cost (Mardis, 2008a; Simon et al., 2009; Lister et al., 2009).

Early next-generation sequencing technologies include Massively parallel signature sequencing (MPSS) (Brenner et al., 2000) and Polony sequencing (Mitra et al., 2003; Shendure et al., 2005), which pioneered the development of commercially successful next generation sequencing platforms.

Generally, next generation sequencing methods initiate by attaching fragmented DNA to platform-specific adaptors, fixing single DNA fragments to a solid support like a bead or a planar solid surface, then amplifying by emulsion PCR. The colonies are then sequenced in situ and bases are detected using fluorescence scanning or chemiluminescence (Simon et al., 2009; Lister et al., 2009). Three widely adopted next-generation sequencing platforms are Roche 454 System, AB SOLiD system and Illumina Genome Analyzer (GA) / HiSeq System.

1.2.1.1 Roche 454 System

Roche 454 was the earliest next generation system. It uses pyrosequencing technology (Margulies et al., 2005), which detects pyrophosphate released during nucleotide incorporation. DNA libraries are attached to 454-specific adapters, denatured into single-strand, captured by amplification beads and amplified by emulsion PCR. One of dNTP (dATP, dCTP, dGTP, dTTP) is added to the reaction, if complement to the bases of the template strand, equal amount of pyrophosphate (PPi) will be released and induces a serious of reactions that generates visible light (Mardis, 2008b; Liu et al., 2012).

Roche 454 initially could produce 200,000 reads of 100-150 bp in length per run in 2005, giving an output of 20 Mb data. In 2009 its read length could reach 700 bp and output 14 G data per run (Liu et al., 2012). Roche 454 has a competitive speed relative to other platforms: it takes only 10 hours for the sequencing process. However, it has relatively high cost mainly for the reagents (Liu et al., 2012).

1.2.1.2 AB SOLiD System

SOLiD, which stands for Sequencing by Oligo Ligation Detection, uses sequencing-by-ligation method. DNA fragments linked to a universal P1 adapter are attached to the magnetic beads. The ligation step is performed using 8-mer probes fluorescently labeled at the 5' end with a cleavage site between the 5^{th} and 6^{th} nucleotide. In each cycle, a fluorescent signal is emitted when the first two bases from the 3' end of the probe complement the template strand and vanished when the bases 6-8 are cleaved. In the next cycle, new probes will be paired at positions 6 and 7. The entire sequencing step is composed of five rounds, each consisting of 5-7 cycles. The start position in each round will shift by one nucleotide (McKernan et al., 2009).

Unlike most other next-generation sequencing techniques, SOLiD produces sequencing data in colorspace. The colorspace sequence can be decoded to get the nucleotide sequence if the type of

any one nucleotide in the sequence is known. Because each base is read twice in SOLiD sequencing, it has the advantage to reduce the single nucleotide polymorphism (SNP) miscalling. To miscall a SNP, to adjacent colors must be miscalled at the same time. On the other hand, it has the disadvantage that a single color miscall will propagate to the remaining portion of the read when decoded into the nucleotide sequence (McKernan et al., 2009).

SOLiD initially produced 35 bp reads and 3 G data per run. In late 2010, read length was improved to 85 bp and output was 30 G per run. A complete run takes within 7 days. SOLiD has lower cost per base than 454 with a comparable accuracy, but the short read length limits its applications (Mardis, 2008b; Liu et al., 2012).

1.2.1.3 Illumina Genome Analyzer (GA) / HiSeq System

Genome Analyzer (GA) was initially released by Solexa in 2006, which was acquired by Illumina in 2007. The sequencer uses the technology of sequencing by synthesis (SBS). DNA fragments are amplified by primers attached to the surface of the flowcell, a process called bridge amplification. In the sequencing step, four nucleotides (ddATP, ddCTP, ddGTP, ddTTP) with cleavable fluorescent labeling and removable 3' blocking group are added, the nucleotide that complements the template is incorporated, releasing a signal captured by a charge-coupled device (Mardis, 2008b; Liu et al., 2012).

Solexa GA initially produced 35 bp reads with an output of 1 G per run. The latest GAIIx series can output 85 G per run. In 2010, HiSeq 2000 was launched, whose output was 200 G per run initially and improved to 600 G per run. Each run takes within 8 days. HiSeq 2000 has the lowest cost with $0.02 / million bases, compared to 454 and SOLiD (Liu et al., 2012).

1.2.1.4 Other sequencing platforms

Recently, sequencers feature fast turnover and small size were launched, targeting clinical applications and individual labs, including Ion Personal Genome Machine (PGM) by Ion Torrent and MiSeq by Illumina (Liu et al., 2012; Quail et al., 2012).

Third generation sequencing, featuring no PCR and real time signal capture, has recently been developed. Single-molecule real-time (SMRT) by Pacific Bioscience uses modified enzyme to observe in real time the fluorescent signal released by the enzymatic reaction (Flusberg et al., 2010). Nanopore sequencing uses a tiny biopore with diameter in nanoscale. When a single stranded DNA is moved across the nanopore, voltage across the channel is disrupted. The scale of the disruption depends on the size difference of the different deoxyribonucleoside monophosphate (dNMP), and can be measured by standard electrophysiological technique (Branton et al., 2008; Timp et al., 2010).

Third generation sequencing has the advantages of long read length and fast speed, which will enable the sequencing of a whole genome within a day at low cost (Branton et al., 2008).

1.2.2 Small RNA sequencing (small RNA-seq)

1.2.2.1 Traditional small RNA detection method

In the early studies, miRNAs were detected by forward screening (Wightman et al., 1993; Lee et al., 1993) or low-depth cloning (Reinhart et al., 2002; Park et al., 2002; Llave et al., 2002). These techniques are biased to detect abundantly expressed and ancient miRNAs (Axtell and Bartel, 2005).

1.2.2.2 Computational prediction method

Preceding the era of next generating sequencing, computational annotation of small RNAs (mainly miRNAs) was based on comparative genomics to identify conserved sequences in multiple genomes. Twenty-three miRNA candidates, predicted by the conserved presence of both the miRNA candidates and corresponding targets in *Arabidopsis thaliana* and *Oryza sativa*, were experimentally verified (Jones-Rhoades and Bartel, 2004). This homology-based computational prediction method has several limitations. First, the discovered candidates are only predictions, which require experimental confirmation. Second, only conserved small RNAs can be detected in this manner.

1.2.2.3 Small RNA sequencing (small RNA-seq)

In recent years, small RNA discovery and annotation has been empowered by small RNA sequencing (small RNA-seq), which directly takes advantage of the high throughput parallel DNA sequencing technologies. In small RNA-seq experiments, size-fractionated total RNAs are attached to adapters on both ends using RNA ligase, after reverse-transcription and PCR amplification, the resulting cDNA library is sequenced by next generation sequencing technologies. If a reference genome is available, the resulting reads will be mapped to the reference to either discover and annotate novel small RNAs or identify and quantify expressed small RNAs (Fahlgren et al., 2009; Axtell, 2013b). On the other hand, reads can be searched against known small RNAs in other species to identify conserved small RNAs without a reference genome.

1.2.2.4 Applications of small RNA-seq

Small RNA-seq has revealed the complex composition of the small RNA population in plants (Lu et al., 2005; Axtell, 2013a). One notable achievement is the identification and characterization of a class of younger miRNAs, which evaded detection by traditional methods due to their low abundance.

These species-specific or nonconserved *MIRNA* genes have been observed in *Arabidopsis thalina* (Rajagopalan et al., 2006; Lu et al., 2006), *Physcomitrella patens* and *Selaginella moellendorffii* (Axtell et al., 2007), rice (Sunkar et al., 2008; Lu et al., 2008; Heisel et al., 2008; Zhu et al., 2008), *Medicago truncatula* (Szittya et al., 2008; Lelandais-Brière et al., 2009), and *Glycine max* (Subramanian et al., 2008). Small RNA-seq was also used in the study presented in Chapter 2 to detect and quantify expressed miRNAs in two species of the *Arabidopsis* genus (Ma et al., 2010).

Combined with genetics, small RNA-seq is a powerful tool to annotate specific classes of small RNAs based on molecular mechanisms of their biogenesis. For example, 24-nt heterochromatic siRNA population is eliminated in *dcl3* knockout backgrounds in *Arabidopsis*, while in *P. patens*, 22-24 nt small RNAs from repetitive regions fail to accumulate in *Ppdcl3* mutants (Cho et al., 2008), indicating that heterochromatic siRNAs are a conserved class of small RNAs in land plants. In the study presented in Chapter 3.4, small RNAseq data from *Physcomitrella patens* wild type, *dcl3*, *rdr6* and *dcl4* mutants were used to identify novel small RNA loci that exhibit *TAS*-like small RNA accumulation patterns (Arif et al., 2012). Small RNAseq also plays a key role in the identification of secondary siRNAs. As secondary siRNAs derive from successive DCL cleavage from a defined terminus of an RNA precursor, the phasing pattern of small RNA accumulation at a given locus is a distinct signature of secondary siRNAs (Chen et al., 2007; Howell et al., 2007a; Zhai et al., 2011).

Another application of small RNAseq is the profiling of AGO-bound small RNAs by sequencing the small RNAs co-immunoprecipitated with AGO protein complexes. This approach can elucidate preferential association of different types of small RNAs with specific AGO paralogs. Studies using this method found that the 5' nucleotide of small RNA guide strands partially determine the recruited AGO proteins (Takeda et al., 2008; Mi et al., 2008; Montgomery et al., 2008; Czech and Hannon, 2011). In *Arabidopsis*, small RNAs with a 5' uridine (mostly miRNAs) are preferentially loaded by AGO1, while Small RNAs with a 5' adenosine preferentially associate with AGO2 and AGO4. The sorting mechanism likely determines the small RNA functions, as is evidenced that changing the 5' uridine to adenosine of an miRNA resulted in a switch from AGO1 to AGO2 loading that abolished the silencing activity of the miRNA (Mi et al., 2008).

1.2.3 miRNA target detection and degradome sequencing

1.2.3.1 Traditional experimental target detection method

Plant miRNAs direct negative regulation on their target genes by either cleaving the target mRNA or repressing translation. Target mRNA levels are often down-regulated with overexpressed miRNAs. Early studies used microarray experiments to detect down-regulated mRNAs correlated with the overexpression of certain miRNAs (Palatnik et al., 2003; Schwab et al., 2005), and identified

targets of miR156, 159, 164 and 319 in *Arabidopsis*. However, down-regulated mRNAs detected by this method may not be the direct miRNA targets but rather indirectly affected.

AGO-mediated cleavage of the target mRNA produces a stable, uncapped 3' fragments with a 5' monophosphate. A modified 5' RACE (Rapid amplification of 5′ complementary DNA ends (5′ RACE), 2005) is used to directly detect the end product of AGO-cleaved mRNA in order to confirm the predicted miRNA targets (Llave et al., 2002). Since the 3' fragments of the cleaved mRNA contains a ligation-competent 5' monophosphate rather than a conventional 5' cap, the modified 5' RACE reaction involves ligation of an adapter directly to the 5' end without enzymatic pretreatment (Llave et al., 2002). Microarray experiments following 5' RACE amplified mRNA fragments has been used to detect of cleaved miRNA targets in *Arabidopsis* on the whole transcriptome scale (Jiao et al., 2008; Franco-Zorrilla et al., 2009).

1.2.3.2 Computational prediction of miRNA targets

Plant miRNAs have near perfect complementarity to their target mRNAs, facilitating the computational prediction of putative miRNA targets. The first method implemented to predict targets uses the *Patscan* patten-finding program (Dsouza et al., 1997) to identify mRNA sequences with fewer than four mismatches to the miRNA (Jones-Rhoades et al., 2002). This method regards GU wobbles as mismatches and disallows gaps, which is stringent and may miss bona fide miRNA targets which have gaps and bulges in the pairing region. A refined method allows for gaps and bulges and comprises a scoring method which gives a 0.5 penalty to GU mismatch, 1.0 penalty to non-GU mismatch and 2.0 penalty to bulges. The final score is the sum of all penalties. To correct for higher probabilities of having more mismatches in longer miRNAs, miRNA is scored by a sliding 20-nt window and the minimum score is kept (Jones-Rhoades and Bartel, 2004). The most widely used prediction method was later developed, which has a slightly different scoring scheme, inspired by the profiling of an extensive set of miRNA-mRNA target pairs (Allen et al., 2005). A GU mismatch incurs a penalty of 0.5, all other mismatches, gaps and bulges has a penalty of 1. It was observed that the positions 2-13 of the miRNA is the critical region where mismatches were rare. As a result, the penalty is doubled in the regions of positions 2-13 of the miRNA. This prediction method is implemented as the TargetFinder program (http://carringtonlab.org/resources/targetfinder).

1.2.3.3 Degradome sequencing method

Degradome sequencing method combines the idea of 5' RACE and next generation sequencing to detect sliced miRNA targets. The experimental procedure for degradome sequencing is as follows: total RNA is isolated and polyA+ RNA fraction are purified with oligo-dT agarose beads.

The polyA- fraction can be reserved for small RNA-seq. Next, polyA+ RNAs are ligated to an RNA adapter with a 5' MmeI restriction site using T4 RNA ligase. T4 ligation specifically ligates substrates with a 3' OH and a 5' monophosphate, retaining uncapped polyA+ RNAs. After purification, reverse transcription and second-strand synthesis, a 20-21 nt tag linked to the 5' adapter is produced by MmeI digestion, and is attached to a 3' dsDNA adapter. The resulting “degradome” tags are then sequenced by next generating sequencing technologies (Addo-Quaye et al., 2008; German et al., 2008a).

The sequenced reads require extensive computational analysis. Adapter sequences in the reads are first removed, resulting in ~20 nt long degradome tags. These tags are mapped to the transcriptome. The ~20 nt tag is extended by 15 nt upstream for any matched positions to define a 35 nt query sequence. The query sequences represent the miRNA complementary site. The query sequence is searched against known miRNAs to identify matches at the 5' of the cleavage tag and the 10th nucleotide of the candidate miRNA, which is indicative of an miRNA-target pair. Finally, statistical analysis distinguishes the set of believable sliced targets from false positives (Addo-Quaye et al., 2009a). This computational method is implemented as the CleaveLand software (Addo-Quaye et al., 2009a).

1.2.3.4 Applications of degradome sequencing method

Degradome sequencing was first applied in *Arabidopsis thaliana* (Addo-Quaye et al., 2008; German et al., 2008a), which confirmed many previously known miRNA targets and discovered a few novel ones. Subsequently, degradome sequencing were widely used in multiple plant genomes, including but not limited to rice (Li et al., 2010; Zhou et al., 2010), *P. patens* (Addo-Quaye et al., 2009b), the grapevine *Vitis vinifera* (Pantaleo et al., 2010), cucumber (Mao et al., 2012), *Glycine Max* (Shamimuzzaman and Vodkin, 2012; Hu et al., 2013), wheat (Li et al., 2013), and *Brassica napus* (Xu et al., 2012). In the study presented in Chapter 2, sliced miRNA targets in *Arabidopsis lyrata* were also detected with degradome sequencing (Ma et al., 2010).

1.2.4 Comparative genomics as a powerful tool to study small RNA evolution

With the accumulation of massive genomic data, comparative genomics thrives as a powerful tool to study the relationship of functional genomic elements across different species. The principle is to identify conserved DNA sequences which encode proteins and RNAs responsible for the conserved functions among the species under comparison (Hardison, 2003).

Comparative genomic method has been widely applied to study plant small RNAs. In a pioneer study to computationally identify plant miRNAs, *Arabidopsis* miRNAs were predicted based on the conservation of both the miRNA candidates and corresponding targets in *Arabidopsis thaliana* and *Oryza sativa*, many of which were experimentally verified (Jones-Rhoades and Bartel, 2004).

Comparative analysis of flooding-responsive *MIRNA* genes in maize, *Arabidopsis* and pine reveals that maize *MIRNA* genes display single-nucleotide polymorphisms which expand the set of predicted targets, providing an adaptation mechanism to environmental stresses (Zhang et al., 2008). This study sheds light on the evolutionary fluidity of deeply-conserved miRNAs.

Comparative genomics can be applied to genomes at different phylogenetic distances to address different questions. The aforementioned examples examined relatively large evolutionary distances on the scale of 100 Myr (Chaw et al., 2004) and examined evolution of deeply-conserved miRNAs. On the other hand, comparative genomic studies of species separated by a short evolutionary distance are informative of rapidly evolving small RNA populations, as is evident in the comparative analysis of species-specific miRNAs in the *Drosophila* genus (Berezikov et al., 2010). This study indicates that evolutionarily transient miRNA genes are frequently born and lost, with only a subset being fixed by integration into regulatory networks across drosophilid radiation. Chapter 2 of this dissertation presents a study between two *Arabidopsis* species and reveals similar evolutionary fluidity of less-conserved miRNAs in plants (Ma et al., 2010).

Comparative genomics has broad applications not limited to small RNA research. The idea of finding conserved sequences across multiple species is particularly powerful for the identification of conserved cis-regulatory elements (CREs). For instance, functional elements were discovered by identifying evolutionary signatures in the 12 completely sequenced Drosophila genomes (Stark et al., 2007). Conserved CREs were identified first by comparing human and other mammals (Xie et al., 2005a) and later by 28-way alignments of sequenced vertebrate genomes (Miller et al., 2007). Inspired by the resolving power of sequencing multiple species followed by comparative genomics for conserved regulatory elements discovery, Chapter 4 proposes a novel targeted genomic enrichment methodology which can rapidly capture, enrich and sequence unknown genomic sequences flanking a conserved core sequence in multiple species, thus facilitating conserved CRE discovery near the targeted loci.

1.3 Objectives

I aim to apply sequencing data analysis and comparative genomics to study the function and evolution of small RNAs. This theme runs through the five projects presented in this dissertation:

Chapter 2 compares small RNAs in two closely related species (*Arabidopsis thaliana* and *Arabidopsis lyrata*), in order to characterize the sequence divergence, targeting and processing precision of less-conserved miRNAs relative to more-conserved ones. Heterochromatic siRNA loci are compared between species to estimate the conservation pattern of the global small RNA occupancy

as well as siRNA hotspots.

Chapter 3 presents three separate projects in diverse plant species, all with the objective to identify small RNAs based on conservation patterns. In *Theobroma cacao*, conserved miRNAs are computationally identified. In oil palm (*Elaeis guineensis Jacq.*), expressed miRNA families are identified with small RNA-seq data. In *Physcomitrella patens*, a novel family of ta-siRNA loci is discovered by examining small RNA accumulation patterns similar to known *TAS* loci.

Chapter 4 describes a novel targeted genomic enrichment method that aims to rapidly capture, enrich and sequence flanking genomic DNA surrounding short conserved regions such as mature miRNAs, thus accelerating discovery of cis-regulatory elements surrounding such loci.

Chapter 5 summarizes the conclusions of previous chapters and discusses the prospects of small RNA research.

Chapter 2
Arabidopsis lyrata small RNAs: Transient *MIRNA* and siRNA loci within the *Arabidopsis* genus

2.1 Summary

Plant small RNAs play critical roles in multiple cellular processes through transcriptional and/orpost-transcriptional regulation of RNA targets. 21nt microRNAs (miRNAs) and 24nt Pol IV-dependent siRNAs (p4-siRNAs) are the most abundant types of small RNAs in angiosperms. Some miRNAs are well conserved among different plant lineages, while others are less conserved. It is not clear whether less-conserved miRNAs have the same functionality as the well conserved ones. p4-siRNAs are broadly produced in the *Arabidopsis thaliana* genome, sometimes from active "hotspot" loci. It is unknown whether individual p4-siRNA hotspots are retained as hotspots between plant species. In this study, we compare small RNAs in two closely related species (*Arabidopsis thaliana* and *Arabidopsis lyrata*) and find that less-conserved miRNAs have high rates of divergence in *MIRNA* hairpin structures, mature miRNA sequences, and target complementary sites in the other species. The fidelity of miRNA biogenesis from many less-conserved *MIRNA* hairpins frequently deteriorates in the sister species relative to the species of first discovery. We also observe that p4-siRNA occupied loci have a slight tendency to be retained as p4-siRNA loci between species, but the most active *A. lyrata* p4-siRNA hotspots are generally not syntenic to the most active p4-siRNA hotspots of *A. thaliana*. Altogether, our findings indicate that many *MIRNA*s and most p4-siRNA hotspots are rapidly changing and evolutionarily transient within the *Arabidopsis* genus.

2.2 Introduction

Plant transcriptomes include a multitude of small RNAs produced by the action of Dicer-Like (DCL) proteins. These endogenous small RNAs function as specificity determinants bound to Argonaute (AGO) proteins within complexes which effect transcriptional and/or post-transcriptional regulation of RNA targets. microRNAs (miRNAs) are an abundant subset of the plant small RNA population. They are defined by precise, DCL-catalyzed excision from the helical stems of hairpin-forming single-stranded precursor RNAs (Meyers et al., 2008; Voinnet, 2009). Many plant miRNAs negatively regulate multiple target mRNAs at the post-transcriptional level, promote the formation of

short interfering RNAs (siRNAs) from their RNA targets, and/or interact with naturally occurring target mimics (Mallory and Bouché, 2008). Through these regulatory mechanisms, plant miRNAs are critical for multiple processes, including diverse developmental events, meristem identity, abiotic stress responses, nutrient homeostasis, and pathogen responses.

Plant *MIRNA* loci are most often independent RNA Polymerase II (Pol II)-transcribed units whose expression patterns are individually regulated, and consequently display tissue- or condition-specific accumulation patterns (Xie et al., 2005b; Válóczi et al., 2006; Sieber et al., 2007). Identical or nearly identical mature miRNAs can be encoded by large families of paralogous *MIRNA* loci. The evolution of individual *MIRNA* loci (Warthmann et al., 2008) and patterns of *MIRNA* family expansion and contraction (Maher et al., 2006) can be tracked using comparative genomics. Some plant *MIRNA* families are quite conserved: Over 20 families are expressed in both monocots and eudicots, and at least seven of these families are also expressed in bryophytes (Axtell and Bowman, 2008). Compared to less-conserved miRNAs, well-conserved miRNAs tend to have higher expression levels, more paralogous loci per family, and RNA targets which are easier to computationally predict (using currently understood parameters for miRNA/target interactions in plants) and experimentally verify (chiefly by detecting remnants of AGO-catalyzed target cleavage) (Rajagopalan et al., 2006; Fahlgren et al., 2007; Axtell et al., 2007). These observations have led to the hypothesis that many less-conserved miRNA families may be non-functional and evolutionarily transient (Rajagopalan et al., 2006; Fahlgren et al., 2007; Axtell, 2008; Felippes et al., 2008). A second hypothesis, which explains the difficulty of predicting and validating targets of less conserved miRNAs, suggests that less conserved miRNAs are indeed often functional as target regulators but tend to interact with targets in configurations generally not captured by current target prediction methods and with molecular outcomes which don't often include readily detectable cleavage remnants (Brodersen and Voinnet, 2009). The less conserved miR834 partially conforms to this hypothesis because target regulation occurs without easily detected RNA cleavage (Brodersen et al., 2008). However, in this case, the target site itself was readily predicted by existing methods. A third hypothesis, which explains the lack of conservation and generally low expression levels, posits that less conserved miRNAs often perform regulatory tasks in restricted numbers of cells within a single family or genus. The regulation of *AGL16* transcripts by the less conserved miR824 specifically within the stomatal precursor cells of the *Brassicaceae* provides an example conforming to this idea (Kutter et al., 2007).

In most angiosperm tissues which have been analyzed, the majority of small RNAs are not miRNAs, but instead are 24nt Pol IV-dependent siRNAs (p4-siRNAs) which arise from DCL processing of long, perfectly double-stranded RNA templated by genomic sequences. In *A. thaliana*, most 24nt siRNAs are p4-siRNAs produced and utilized by the Pol IV/Pol V system, which uses them to direct

RNA-directed DNA methylation (RdDM) and repressive histone modifications to target chromatin (Matzke et al., 2009). Production of p4-siRNAs is broadly distributed throughout the *A. thaliana* genome, with concentrations in pericentromeric regions, avoidance of protein-coding loci, and a tendency toward repetitive sequences (Lu et al., 2005; Rajagopalan et al., 2006; Kasschau et al., 2007). Nonetheless, there are clearly "hotspots" of p4-siRNA production from certain loci (Rajagopalan et al., 2006; Kasschau et al., 2007; Zhang et al., 2007; Mosher et al., 2008). Some *A. thaliana* p4-siRNA loci are active in all developmental stages (type II loci) while many others produce p4-siRNAs specifically in floral and reproductive tissues (type I loci; Mosher et al., 2009). Loci marked by cytosine methylation, some of which is likely directed by p4-siRNAs, can vary among *A. thaliana* ecotypes (Vaughn et al., 2007) as do the activities of some p4-siRNA loci (Vaughn et al., 2007; Zhai et al., 2008). However, it is not known if individual p4-siRNA hotspots are frequently retained as hotspots between species.

In this study, we exploit the recent production of a draft nuclear genome sequence for *Arabidopsis lyrata* to examine evolution of plant *MIRNA* and p4-siRNA loci between two congenic Brassicaceae species. We find that many less-conserved miRNA families have high rates of sequence divergence in *MIRNA* hairpin structures, mature miRNA sequences, and in the complementary sites of predicted targets between the two species. High throughput identification of sliced miRNA targets in both *A. lyrata* and *A. thaliana* was generally unsuccessful for targets of less conserved miRNAs. We also observe that the most active p4-siRNA hotspots expressed within *A. lyrata* leaves are not syntenic to the p4-siRNA hotspots in multiple tissues of *A. thaliana*.

2.3 Methods

2.3.1 Small RNA sequencing and data analysis

Total RNA was extracted using Tri-Reagent (Sigma, St. Louis, MO, USA). Illumina sequencing of the *A. lyrata* leaf sample was as follows: Small RNA-enriched fractions were purified from 20% PAGE gel by recovering the 20-30 nucleotide area from total RNA sample. A pre-adenylated 3' adapter (IDT, Coralville, IA) linker 1 (5'-AppCTGTAGGCACCATCAATddC-3') was added using T4 RNA ligase without exogenous ATP. 3'-ligated products were gel eluted and then ligated to a 5' adapter composed of RNA (5'-GUUCAGAGUUCUACAGUCCGACGAUC-3') using T4 RNA ligase with ATP. Gel purification of the ligated product was followed by reverse transcription using an oligo (5'-ATTGATGGTGCCTACAG-3') specific to the 3' linker. The cDNA library was then amplified using a 5' adapter oligo (5'- AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3') and a 3' adapter oligo (5'-CAAGCAGAAGACGGCATACGAATTGATGGTGCCTACAG-3'). The amplified library

was then gel-purified and sequenced using an Illumina genome analyzer by Fasteris, Inc. (Geneva, Switzerland). The *A. lyrata* data (described here) and another library embedded within the raw data were computationally separated by parsing the 3' adapter sequences. Reads between 19 and 26 nts in length were retained for analysis. Construction of *A. lyrata* inflorescence-derived small RNAs was performed using the SOLiD Small RNA Expression Kit per the manufacturer's instructions, followed by sequencing on the SOLiD 2 instrument at the Penn State / Huck Institutes Genomics Core Facility.

2.3.2 Identification of *MIRNA*s in *A. thaliana* and *A. lyrata*

A. thaliana small RNA reads combined from nine publicly available sRNAseq datasets (Table 1) were mapped to the TAIR9 genome, and reads mapped to genome less than 15 times were used to align to 190 annotated *A. thaliana MIRNA* loci retrieved from miRBase 14.0 (Griffiths-Jones et al., 2008). Loci were then filtered based on a conservative interpretation of the updated criteria for plant *MIRNA* annotation (Meyers et al., 2008), which required at least ten raw small RNA sequencing reads matching the hairpin, and a miRNA/miRNA* duplex processing precision greater than 0.25 (calculated as the proportion of the raw read abundance mapping exactly to the mature miRNA and miRNA* out of the total abundance of reads mapping anywhere on the hairpin). *A. lyrata MIRNA* homologs of those *A. thaliana MIRNA* loci that passed the filtering were computationally identified using a microsynteny-based method. For each *A. thaliana MIRNA* considered, the two flanking protein coding loci were retrieved from the TAIR9 annotation set. The top five hits in the *A. lyrata* filtered gene models generated by JGI (http://genome.jgi-psf.org/Araly1/Araly1.home.html) of the *A. thaliana MIRNA* and the flanking loci were identified using BLASTn. Flanking loci and *MIRNA* hits were compared to identify microsyntenic regions. The *MIRNA* hit with maximally preserved synteny (*i.e.* between two hits to the respective *A. thaliana* flanking loci) was identified as the predicted *A. lyrata MIRNA* hairpin homolog. If none of the top five hits of the *MIRNA* maintained the same synteny of *MIRNA* and two flanking genes in *A. lyrata*, a *MIRNA* hit that maintained the synteny with only one of the flanking genes was identified as the predicted *A. lyrata MIRNA* homolog. Meanwhile, *A. lyrata MIRNA* loci were identified *de novo* from the three sRNAseq datasets produced in this study (Table 1). Reads from each dataset were mapped to the *A. lyrata* genome assembly generated by JGI (http://genome.jgi-psf.org/Araly1/Araly1.home.html) and each 300nt flanking genomic region was pre-filtered based on polarity of small RNA accumulation (>=75% from the dominant strand), 21-22mer abundance relative to other size classes (amount of 21-22mers more than double the amount of non 21-22mers), and number of hits for individual small RNAs in the genome (no greater than 15 genome matches) and examined by MIRcheck (Jones-Rhoades and Bartel, 2004). Candidates which survived this pre-screening were then subject to the following additional filters: The most abundant small RNA on the hairpin had to have more than 15 reads in at least one of the three libraries; the processing precision

had to be greater than 0.25; the hairpin had to pass MIRcheck (Jones-Rhoades and Bartel, 2004, -) with the parameters "-mir_bulge",3, "-ass", 2,"-unpair"; miRNA* is expressed, or the mature miRNA had to be the most abundant small RNA in two or more libraries. *A. thaliana* genomic loci syntenic to novel *A. lyrata MIRNA*s were identified as described above.

2.3.3 *MIRNA* divergence analysis

Syntenic *A. thaliana* and *A. lyrata MIRNA* hairpins were divided into five regions: The loop/upper stem, mature miRNA, miRNA*, 5' and 3' regions. Each region was further divided into seven equal-length bins (rounded to the closest integer). Divergence was calculated as the pairwise difference per nucleotide of each bin for each *MIRNA* based on the pairwise alignment using MUSCLE between the *MIRNA* hairpins in both species. Mature miRNAs in both species are identified as the small RNA in the most precisely processed miRNA/miRNA* duplexes in a family calculated from the aforementioned sRNAseq datasets. Thus the mature miRNA sequences were not necessarily the same as annotated miRNAs in miRBase 14.0.

2.3.4 miRNA target prediction and validation

All mature miRNAs derived from hairpins which passed our expression criteria were used to predict targets from the TAIR9 transcriptome (for *A. thaliana*) or from the JGI FM3 transcriptome (*A. lyrata*). Predictions were accomplished with the PERL script “axtell_targetfinder.pl”. This program first uses rmapper-ls (from the SHRiMP package; Rumble et al., 2009) to find a large set of alignments with very low stringency. These initial alignments are then parsed into RNA-RNA alignments, and scored using the scheme of Allen et al. (2005), retaining only those alignments scoring seven or better. Target prediction with this method also includes annotation of the predicted cleavage site as well as randomizations. This program is available as part of the CleaveLand 2.0 package on our lab’s website (http://axtell-lab-psu.weebly.com/cleaveland.html).

Construction of degradome libraries differed considerably from our past efforts (Addo-Quaye et al., 2008, 2009a). Approximately 150ng of polyA+ RNA was used as input to the SOLiD whole transcriptome analysis kit, following the manufacturer’s instruction except that 1) The initial RNAseIII-catalyzed RNA fragmentation was omitted and 2) a large size range was gel-isolated after the RT-PCR. Omitting the RNAseIII fragmentation step restricts the initial adapter ligation to only those RNAs containing 5’-monophosphates. Libraries were sequenced using the P1 (5’) adapter only, resulting in the sequencing of the first 35nts of the inserts which represented the 5’ ends of the original RNAs.

Raw degradome data, in colorspace format, was mapped to the appropriate transcriptomes using rmapper-cs (part of the SHRiMP package, version 1.3.1; Rumble et al., 2009) using the non-default settings: M 35bp,fast -o 10000 -F. Initial mappings were then filtered to retain only the best

scoring alignment(s) for each read with less than six mismatches total, and which had perfect alignments to nucleotides one through six (to allow confident detection of 5' ends) and whose alignments extended at least to position 29 (to exclude any potential miRNAs or siRNAs which might have contaminated the libraries). The filtered map data were compacted into a standard degradome format which gives the number of 5' ends observed at each position in the transcriptome, along with a "peak categorization" score between zero and four. For each miRNA query, axtell_targetfinder.pl was used to predict all targets with alignment scores of seven or less, along with 1,000 identical target predictions for randomly permuted versions of the query miRNA. The significance of any degradome signatures which matched the cleavage sites of a predicted miRNA target assessed by examining the frequencies with which the randomized queries also matched degradome information. Specifically, for each peak category, a cumulative distribution representing the frequency with which the random queries had one or matches at a given miRNA alignment score was calculated. The likelihood that a given cleavage fragment was observed by chance was estimated by retrieving the frequency with which the random queries gave hits of the given peak category at the given alignment score or better; we interpreted this frequency as a p-value. The cutoff for confident target identification was $p <= 0.05$ in both biological replicates. A series of PERL scripts which accomplish these calculations is available from our lab website as the CleaveLand 2.0 package. (http://axtell-lab-psu.weebly.com/cleaveland.html).

2.3.5 Small RNA occupancy calculation and hotspot identification

A whole genome alignment between *A. thaliana* and *A. lyrata* was performed using lastz (http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.01.50/README.lastz-1.01.50.html, Bob Harris and Cathy Riemer, unpublished), and further processed to retain a one-to-one best alignment using chainnet (Kent et al., 2003). The *A. thaliana* genome (TAIR9 release) was divided into 119,184 1kb bins and the *A. lyrata* syntenic regions to 98,357 of the bins were confidently obtained via the whole-genome alignment. Small RNAs from 12 datasets (AT-F1, AT-F2, AT-F3, AT-F4, AT-Sq1, AT-Se1, AT-Se2, AT-L1, AT-L2, AL-L1, AL-F1, and AL-F2; Table 2.1) were mapped to their respective genomes, and repeat-normalized small RNA abundances were tabulated for each confidently aligned bin (98,357 bins). All bins with abundance greater than 10 reads per million calculated by the small RNA abundance of a certain length (21nt or 24nt) minus the abundance of small RNAs of all other lengths were considered "occupied". The overlap of the occupied bins between each pair of the datasets was analyzed, and the normalized overlap (calculated by $\log_2$-transformation of the observed overlap divided by the expected overlap) was reported. Expected overlap was the product of the fractions occupied in the two datasets being compared. For example, given a dataset with 2,000/98,357 (2.03%) bins occupied and a second dataset with 10,000/98,357 (10.17%) bins occupied, the

percentage of bins occupied in both datasets expected by random chance is 0.0203 x 0.1017 = 0.00206 (0.206%; ~203 bins). For hotspot identification, the bins in each dataset were ranked by the abundance of small RNAs of a certain length (21nt or 24nt) minus the abundance of small RNAs of all other lengths. The top 100 ranking bins were considered hotspots. The number of overlapping bins out of the 100 top-ranking bins from every pair of the datasets was calculated.

2.4 Results

2.4.1 Identification and annotation of *A. lyrata* miRNAs

A. lyrata MIRNAs were identified using two complementary methods: Identification of *A. lyrata* genomic regions syntenic to annotated *A. thaliana MIRNAs* and by analysis of sequenced *A. lyrata* small RNA populations. For synteny-based identification, *A. thaliana MIRNA*s annotated in miRBase 14.0 were filtered according to updated criteria for annotation of plant *MIRNA*s (Meyers et al., 2008) in combination with ~1.6 x 10^7 sequenced small RNAs from nine publically available small RNA seq (sRNAseq) datasets from various wild-type tissues (Table 2.1). *A. thaliana* loci which lacked clear evidence for precise excision of an authentic miRNA/miRNA* duplex from a qualifying hairpin structure were discarded, leaving a total of 157 loci. Syntenic *A. lyrata* loci were identified for 144 out of the 157 queries using a micro-synteny based method (Appendix: Supplemental Dataset 2.1). This procedure identified regions of similarity to the *MIRNA* queries which were flanked by upstream and downstream protein-coding loci which were highly similar to the upstream and downstream genes in *A. thaliana* (See Methods). In parallel, three sRNAseq libraries from *A. lyrata* were obtained and randomly sequenced: A library prepared from rosette leaf tissue yielded ~5.8 x 10^5 reads while two biological replicate sRNAseq samples from inflorescences yielded ~2.6 x 10^7 and ~2.2 x 10^7 reads, respectively (Table 2.1). *A. lyrata MIRNA* were identified *de novo* from these small RNA data using a combination of MIRcheck (Jones-Rhoades and Bartel, 2004) and expression-based filters to ensure conformity to current criteria of plant *MIRNA* annotation (Meyers et al., 2008). A total of 155 *A. lyrata MIRNA* loci were identified based on the sRNAseq data (Appendix: Supplemental Dataset 2.1). Many of these loci (107) were identical to the *A. lyrata* loci found by the synteny-based approach. Two of the *A. lyrata* loci were found to be syntenic to annotated *A. thaliana MIRNA* loci which had been missed in the initial homology search because the corresponding *A. thaliana* loci had not met our expression-based criteria for inclusion as queries. Five of the *A. lyrata MIRNA* loci found based on expression were members of known *A. thaliana* miRNA families but which seemed to lack syntenic homologs in *A. thaliana*. Interestingly, these five *A. lyrata* loci were found in two genomic clusters: Two clustered miR395 loci and three clustered miR399 loci. The remaining 41 *A. lyrata MIRNA* loci all produced

mature miRNAs lacking appreciable similarity to previously annotated miRNAs in any species (based on miRBase 14). The micro-synteny method found syntenic *A. thaliana* loci for about half (22) of these 41 new *A. lyrata MIRNA* loci. Four of these 22 *A. thaliana* syntenic regions passed the Meyers et al. (2008) expression criteria for confident annotation as a *MIRNA* based on our reference *A. thaliana* small RNA dataset, but had not been previously noticed in *A. thaliana*. Details on all *A. thaliana* and *A. lyrata MIRNA*s may be found in Appendix: Supplemental Datasets 2.1-2.3.

Our *MIRNA* annotation efforts led to a list of 205 *MIRNA* loci which, using the sRNAseq datasets referenced above, met the Meyers et al. (2008) criteria for annotation of plant *MIRNA*s in either *A. thaliana*, *A. lyrata*, or both (Appendix: Supplemental Dataset 2.1). These loci were classified based on the apparent conservation level of the mature miRNA families: The 26 families (encoded by 99 loci) that were also annotated in one or more non-Brassicaceae species in miRBase 14 were termed “MC” (for “more conserved”). The 104 families (encoded by 106 loci) that were not annotated as present in any non-*Brassicaceae* species in miRBase 14 were termed “LC” (for “less conserved”).

2.4.2 Less conserved *MIRNA*s are often species-specific, weakly expressed, encoded by single loci, and are more likely to produce 22nt RNAs

For most *MIRNA* loci corresponding to MC families, syntenic homologs were identified in both species which, with reference to our sRNAseq data, expressed microRNAs conforming to the Meyers et al. (2008) expression criteria (Figure 2.1A). In contrast, this situation was rare for LC families (Figure 2.1A). Instead, many homologs of LC miRNAs expressed in one species were not found to be expressed in the other species; only a minority of these homologs retained a putative hairpin structure capable of passing MIRcheck (which assesses secondary structures but does not incorporate expression criteria; Figure 2.1A). Most MC miRNA families were encoded by two or more loci, while nearly all LC miRNA families were encoded by just a single locus (Figure 2.1B). As inferred by sRNAseq read coverage, accumulation levels of both MC and LC miRNAs spanned several orders of magnitude in both *A. thaliana* and *A. lyrata* (Figures 2.1C-D). However, as a group, MC miRNA accumulation levels were clearly higher than those for LC miRNA families (Figures 2.1C-D).

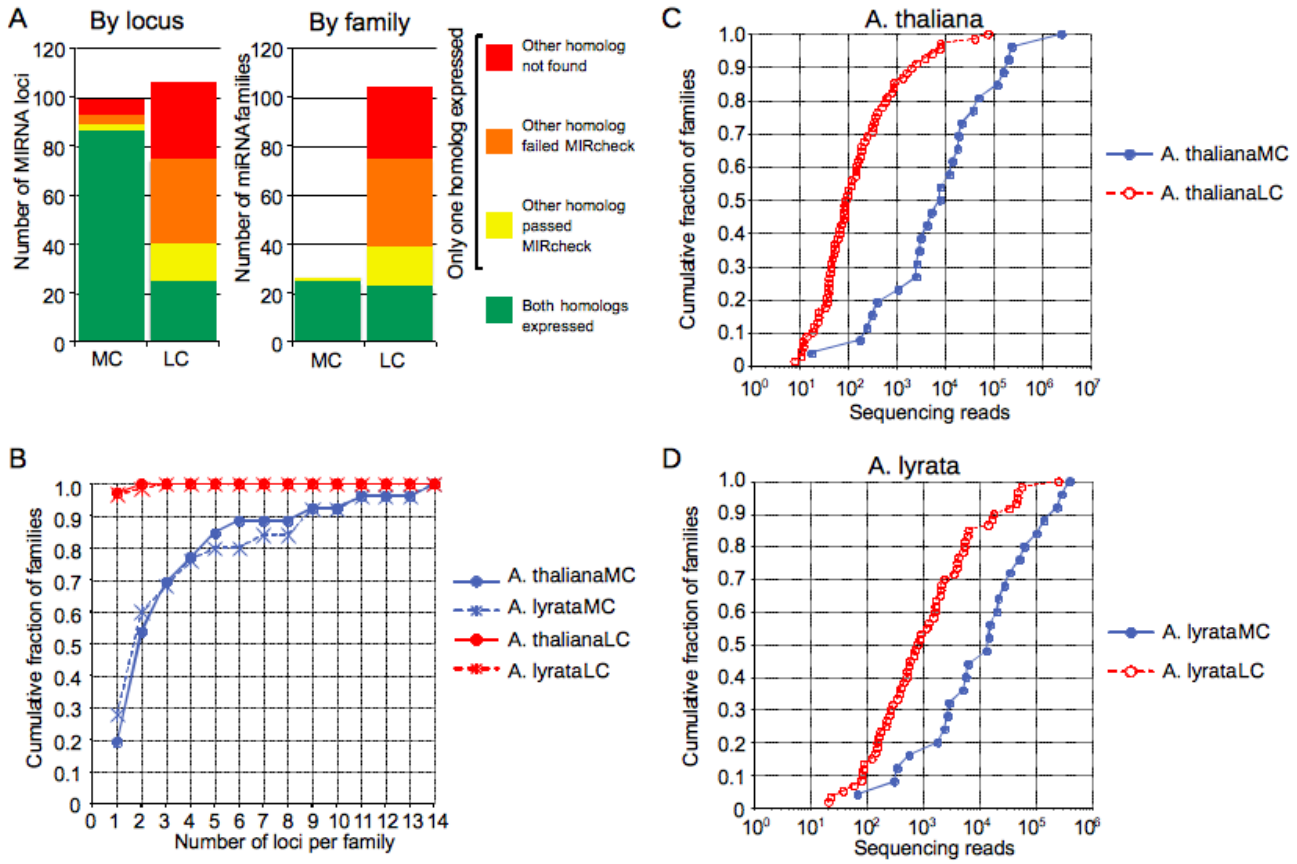


Figure 2.1 Less conserved MIRNAs are often species-specific, weakly expressed, and encoded by single loci.

(A) Identification of homologous *MIRNA* loci (left panel) and families (right panel) in *A. thaliana* and *A. lyrata*. Only loci / families which passed the expression criteria in at least one species were considered. MC: More conserved, LC: Less conserved.

(B) Cumulative distributions of the number of paralogous loci per miRNA family.

(C) Cumulative distributions of the number of sequencing reads per *A. thaliana* miRNA family (based on nine sRNAseq datasets totaling ~1.6 x 10^7 reads; Table 1).

(D) As in C for *A. lyrata* miRNA families (based on ~ 4.8 x 10^7 reads; Table 1).

We next analyzed properties of the observed *MIRNA* hairpin-derived small RNAs themselves (which included annotated mature miRNAs, miRNA*'s, as well as any length and positional variants). Most MC families were dominated by expression of 21nt small RNAs in both *A. thaliana* and *A. lyrata* (Figure 2.2A). However, some MC families were dominated by expression of either 20nt small RNAs or by 22nt small RNAs. The proportion of LC families dominated by 22nt small RNA expression was higher than in MC families in both *A. thaliana* and *A. lyrata*, although 21nt dominant families still accounted for the majority of LC families (Figure 2.2B). Small RNAs with a 5'-U accounted for most small RNAs produced by both MC and LC families in both species (Figures 2.2C-D). However, we noted that the number of families where small RNA expression was not dominated by 5'-U RNAs was higher for *A. lyrata* miRNAs than for *A. thaliana* miRNAs. We do not understand the reason for this result, but we suspect it might be due to differential biases caused the different library construction and sequencing methods used for the different sRNAseq datasets (Table 2.1).

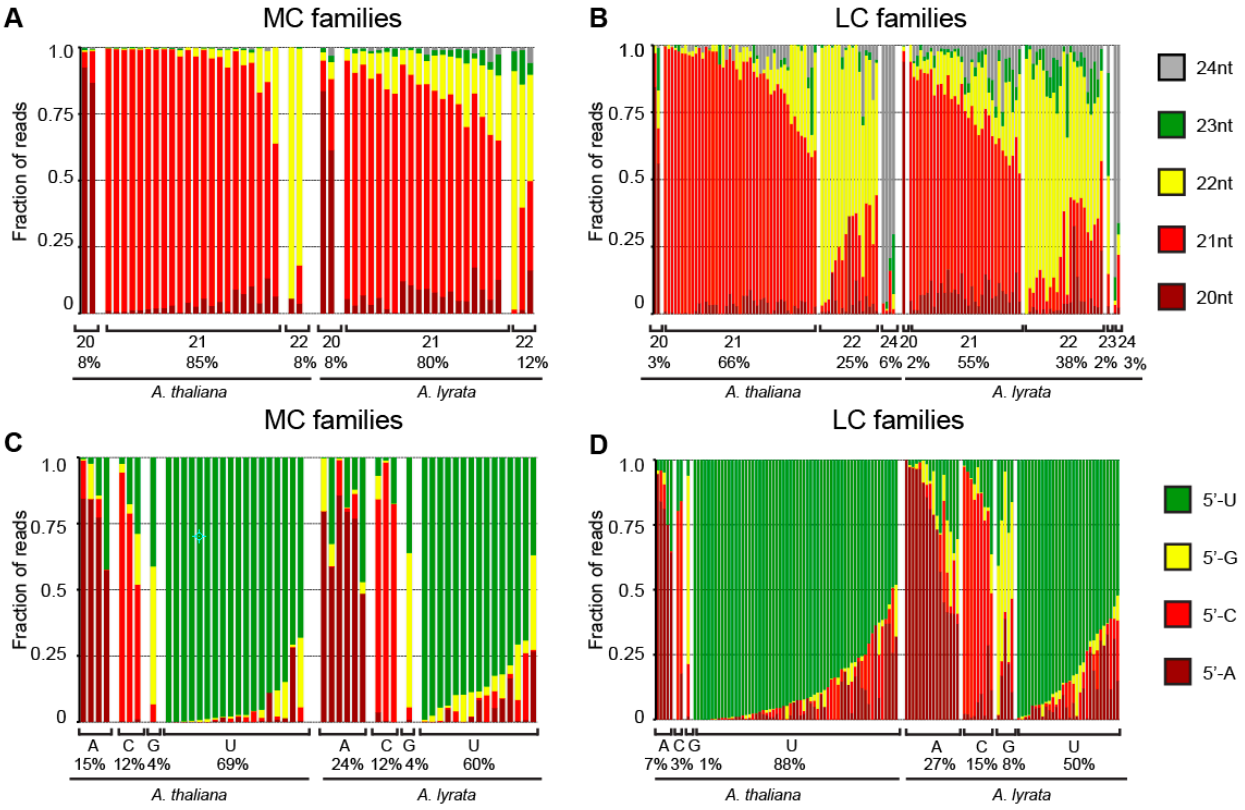


Figure 2.2 Predominant lengths and 5' nucleotides produced by *A. thaliana* and *A. lyrata MIRNA* hairpins.

(A) Proportions of sRNAseq reads of the indicated lengths from more conserved (MC) families. Families are grouped according to the most abundant small RNA length, as indicated below the chart.

(B) As in A for less conserved (LC) families.

(C) As in A for 5' nucleotides of reads from MC families.

(D) As in A for 5' nucleotides of reads from LC families.

Table 2.1 sRNAseq and degradome datasets.

Name	Species / Tissue	Type	Sequencing Instrument	Number of Reads	Unique reads	References	Accessions [NCBI GEO]
AL-L1-sRNA	*A. lyrata* / rosette leaves	sRNAseq	Illumina / SBS	583,895[a]	382,520[a]	This study	GSM451894
AL-F1-sRNA	*A. lyrata* / inflorescences	sRNAseq	SOLiD	26,192,231[a]	61,287,122[d]	This study	GSM512644
AL-F2-sRNA	*A. lyrata* / inflorescences	sRNAseq	SOLiD	21,620,398[a]	52,184,197[d]	This study	GSM512645
AT-deg1	*A. thaliana* / inflorescences	Degradome	SOLiD	9,031,213[b]	671,981[e]	This study	GSM512878
AT-deg2	*A. thaliana* / inflorescences	Degradome	SOLiD	9,612,258[b]	878,714[e]	This study	GSM512879
AL-deg1	*A. lyrata* / inflorescences	Degradome	SOLiD	7,087,227[b]	739,992[e]	This study	GSM512880
AL-deg2	*A. lyrata* / inflorescences	Degradome	SOLiD	15,665,420[b]	743,889[e]	This study	GSM512881
AT-F1-sRNA	*A. thaliana* / inflorescences	sRNAseq	Roche/454	205,649[c]	100,658[c]	Rajagopalan et al. (2006)	GSM118372
AT-F2-sRNA	*A. thaliana* / inflorescences	sRNAseq	Roche/454	78,596[c]	57,966[c]	Kasschau et al. (2007)	GSM154336
AT-F3-sRNA	*A. thaliana* / inflorescences	sRNAseq	Illumina / SBS	7,686,781[c]	2,841,896[c]	Lister et al. (2008)	GSM227608
AT-F4-sRNA	*A. thaliana* / inflorescences	sRNAseq	Illumina / SBS	7,576,080[c]	1,482,150[c]	Montgomery et al. (2008)	GSM342999, GSM343000, GSM343001
AT-Sq1-sRNA	*A. thaliana* / siliques	sRNAseq	Roche/454	305,764[c]	141,539[c]	Rajagopalan et al. (2006)	GSM118375
AT-Se1-sRNA	*A. thaliana* / seedlings	sRNAseq	Roche/454	188,954[c]	77,937[c]	Rajagopalan et al. (2006)	GSM118374
AT-Se2-sRNA	*A. thaliana* / seedlings	sRNAseq	Roche/454	22,467[c]	12,718[c]	Kasschau et al. (2007)	GSM154375
AT-L1-sRNA	*A. thaliana* / rosette leaves	sRNAseq	Roche/454	186,899[c]	67,663[c]	Rajagopalan et al. (2006)	GSM118373
AT-L2-sRNA	*A. thaliana* / rosette leaves	sRNAseq	Roche/454	15,833[c]	8,112[c]	Kasschau et al. (2007)	GSM154370

a. Mapped to genome

b. Mapped to the sense strand of the transcriptome

c. Present within accession

d. Number of distinct genomic positions matched by one or more reads. Because SOLiD data were mapped allowing for errors, it is not possible to tally the number of unique reads that were mapped.

e. Number of distinct 5' ends matched by one or more reads. Because SOLiD data were mapped allowing for errors, it is not possible to tally the number of unique reads that were mapped.

2.4.3 High levels of *MIRNA* hairpin and mature miRNA sequence divergence between *A. thaliana* and *A. lyrata*

We next compared the sequences of syntenic *A. thaliana* and *A. lyrata MIRNA* hairpins. Analysis of sequence divergence was limited only to syntenic pairs for which both members passed the Meyers et al. (2008) expression criteria or pairs in which one member failed the expression criteria but still had a putative hairpin capable of passing MIRcheck in the expression-negative species (*e.g.* the green and yellow regions of Figure 2.1A). Sequence divergence between these 129 syntenic *A. thaliana* and *A. lyrata* loci (89 from MC families and 40 from LC families) was calculated by scoring each position of all pair-wise alignments. Hairpins were divided into five regions (Figure 2.3A); regions were scaled and divided into seven bins each to account for variations in length among the loci. As expected for conserved, functional *MIRNA* hairpins (Ehrenreich and Purugganan, 2008; Warthmann et al., 2008), divergence was lowest within the mature miRNA itself for both the MC and the LC groups (Figures 2.3B-C); this likely reflects purifying selection on the mature miRNA sequences to maintain complementarity with target mRNAs. The miRNA*'s also showed low levels of divergence, most likely reflecting the requirement to maintain base pairs with the constrained miRNAs in the context of the stem-loop secondary structure. In contrast, the 5', loop, and 3' regions were relatively unconstrained (Figures 2.3B-C). Although the divergence profiles of MC and LC *MIRNA*s were qualitatively similar, there was clearly more divergence in mature miRNA sequences among the LC families, even after discarding loci whose predicted secondary structures were highly aberrant in one of the two species (Figures 2.3B-C). This indicates that the mature miRNA sequences of LC families often have less constraint in mature miRNA and miRNA* sequences over short evolutionary distances. The mature MC miRNAs had a slight tendency towards higher divergence at their 3' ends, although diversity was low throughout the MC miRNAs (Figure 2.3D). In contrast, mature LC miRNAs showed high levels of divergence at all sequence positions (Figure 2.3E).

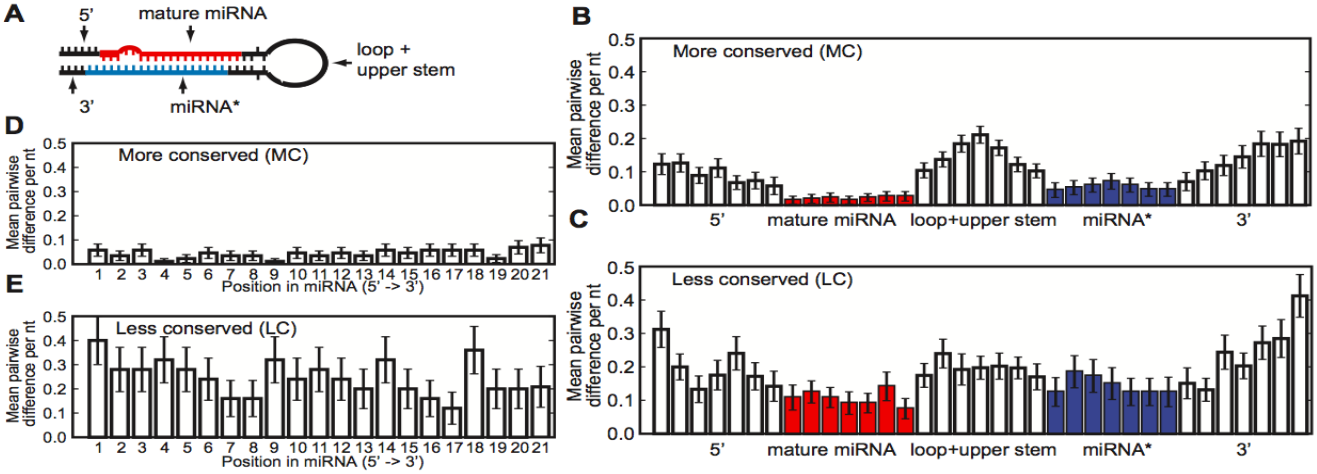


Figure 2.3 Less conserved miRNAs diverge more between *A. thaliana* and *A. lyrata* than do more conserved miRNAs.

(A) A sketch showing the five regions of *MIRNA* hairpins which were analyzed. For convenience, the mature miRNA is shown on the 5' arm although in reality it can be either on the 5' arm or 3' arm.

(B) Average sequence divergence between more conserved *A. thaliana* and *A. lyrata MIRNA* hairpins. Both 5'-arm and 3'-arm mature miRNAs were tallied and displayed together. Bars indicate the standard errors of the means (SEMs).

(C) As in B for less conserved *MIRNA*s.

(D) As in B for each nucleotide position within mature miRNAs from more conserved families.

(E) As in D for less conserved families.

2.4.4 Imprecise and inconsistent processing of less conserved *MIRNA*s

Plant *MIRNA*s are defined by precise processing of a single-stranded stem-loop precursor RNA to release one or more specific miRNA/miRNA* duplexes (Ambros et al., 2003; Meyers et al., 2008). In practice, stem-loop derived small RNAs fall into a continuous spectrum of processing precisions, from very imprecisely processed inverted repeats (whose products are often classified as a form of endogenous siRNA; Lu et al., 2006; Zhang et al., 2007) to canonical *MIRNA*s producing almost exclusively a single miRNA/miRNA* duplex. Some less conserved *MIRNA*s are imprecisely processed in *A. thaliana*, and this inaccuracy is sometimes correlated with their reliance upon Dicer-Like 4 (DCL4) instead of DCL1 for processing (Rajagopalan et al., 2006). To examine *MIRNA* processing precision, the ~4.8 x 10^7 *A. lyrata* sRNAseq reads from our three libraries (Table 2.1) were mapped to the *A. lyrata MIRNA* hairpins. In parallel, the ~1.6 x 10^7 publicly available *A. thaliana* sRNAseq reads (all nine *A. thaliana* sRNAseq libraries in Table 2.1) from several wild-type tissues were mapped to the *A. thaliana MIRNA* hairpins. We defined processing precision at each locus as the abundance of reads corresponding exactly to the mature miRNA or miRNA* divided by the total abundance of all reads mapping to the hairpin. Thus, values close to one indicate very high precision, while values close to zero indicate processing which produces only small amounts of any given miRNA/miRNA* duplex. To compare how *MIRNA* processing precision differed between *A. thaliana* and *A. lyrata* homologs, we considered only those syntenic pairs for which both members passed the Meyers et al. (2008) expression criteria or pairs in which one member failed the expression criteria but still had a putative hairpin which both passed MIRcheck and expressed at least ten sRNAseq reads. The processing precisions of hairpins from MC miRNA families were typically high in both species (Figure 2.4A). Some processing precisions from LC hairpins also had similar precisions in both species (Figure 2.4B). However, several of the LC hairpins were processed quite imprecisely in *A. lyrata*; some of these were also imprecisely processed in *A. thaliana*, where most of them were first described, while others had higher precisions in *A. thaliana* (Figure 2.4B). We conclude that many LC *MIRNA*s are processed very imprecisely, especially outside of the species in which they were first observed.

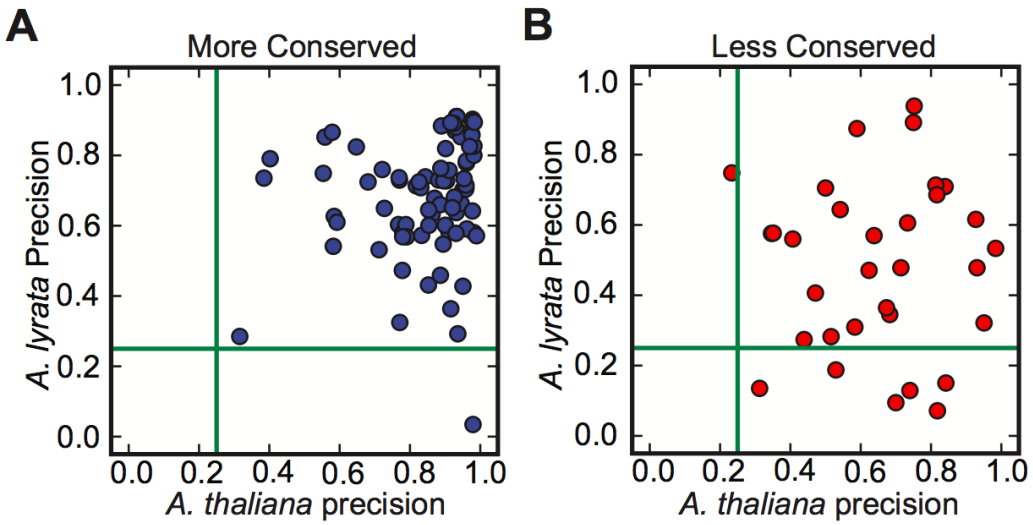


Figure 2.4 Less conserved *MIRNA*s tend to be processed imprecisely.

(A) Scatterplot of *A. lyrata* vs. *A. thaliana MIRNA* processing precisions for more conserved *MIRNA*s. Green lines show the precision value of 0.25 which we used as a cutoff for determining miRNA-like expression patterns (Meyers et al., 2008).

(B) As in *A* for less conserved *MIRNA*s.

2.4.5 High levels of miRNA target divergence between *A. thaliana* and *A. lyrata*

We next examined predicted targets of *A. thaliana* and *A. lyrata* miRNAs. Targets were predicted only for miRNAs whose precursors passed the Meyers et al. (2008) expression criteria. Potential miRNA target sites were scored according to the criteria of Allen et al. (2005); higher scores indicated less confidence in the predictions. In both species, lower-scoring (and thus higher confidence) targets were more frequently predicted for MC families, and less frequently predicted for LC families (Figure 2.5A). Based on previous results (Allen et al., 2005, 2; Rajagopalan et al., 2006), we used a score of three as the upper limit for confident target prediction (Figure 2.5A). All target predictions meeting this cutoff for *A. lyrata* miRNAs are given in Appendix: Supplemental Dataset 2.4. The overlap in confident target predictions (score <= 3) between *A. thaliana* and *A. lyrata* was examined. Importantly, this analysis was limited to the subset of miRNA families which had at least one hairpin which passed the Meyers et al. (2008) expression criteria in both species; in other words, we examined miRNA target overlap only for miRNA families that actually existed in both species. Syntenic homologs were frequently predicted targets of MC miRNA families, while this was rarely the case for the predicted targets of LC miRNA families (Figure 2.5B). We conclude that target predictions using standard criteria are more unreliable and more inconsistent between species for LC miRNA families than for MC families.

In order to experimentally determine miRNA targets, four “degradome” libraries were prepared: Two biological replicates from *A. thaliana* inflorescences and two biological replicates from *A. lyrata* inflorescences; each library consisted of ~1 x 10^7 reads which mapped to the sense strand of one or more annotated transcripts (Table 2.1). Degradome sequencing (synonymous with Parallel Analysis of RNA Ends [PARE] and Genome-wide Mapping of Uncapped and Cleaved Transcripts [GMUCT]) determines the 5’ ends of RNAs with a 5’-monophospate (German et al., 2008b; Addo-Quaye et al., 2008; Gregory et al., 2008). This RNA population includes the sliced remnants of many miRNA-targeted transcripts. Sliced miRNA targets were identified from these data using an updated version of the CleaveLand software (Addo-Quaye et al., 2009a) which calculates empirically estimated p-values for each possible sliced miRNA target (See Methods). Targets with a p-value <= 0.05 in both biological replicates were considered verified; all verified targets along with supporting information are found in Appendix: Supplemental Datasets 5-8. Nearly all verified targets in both species were those of MC miRNA families (Figure 2.5C), and many of these were syntenic homologs which were validated in both species (Figure 2.5D). Taken together, these observations suggested that many of the putative targets of LC miRNA families were inconsistent between *A. thaliana* and *A. lyrata*, and difficult to verify by looking for evidence of slicing.

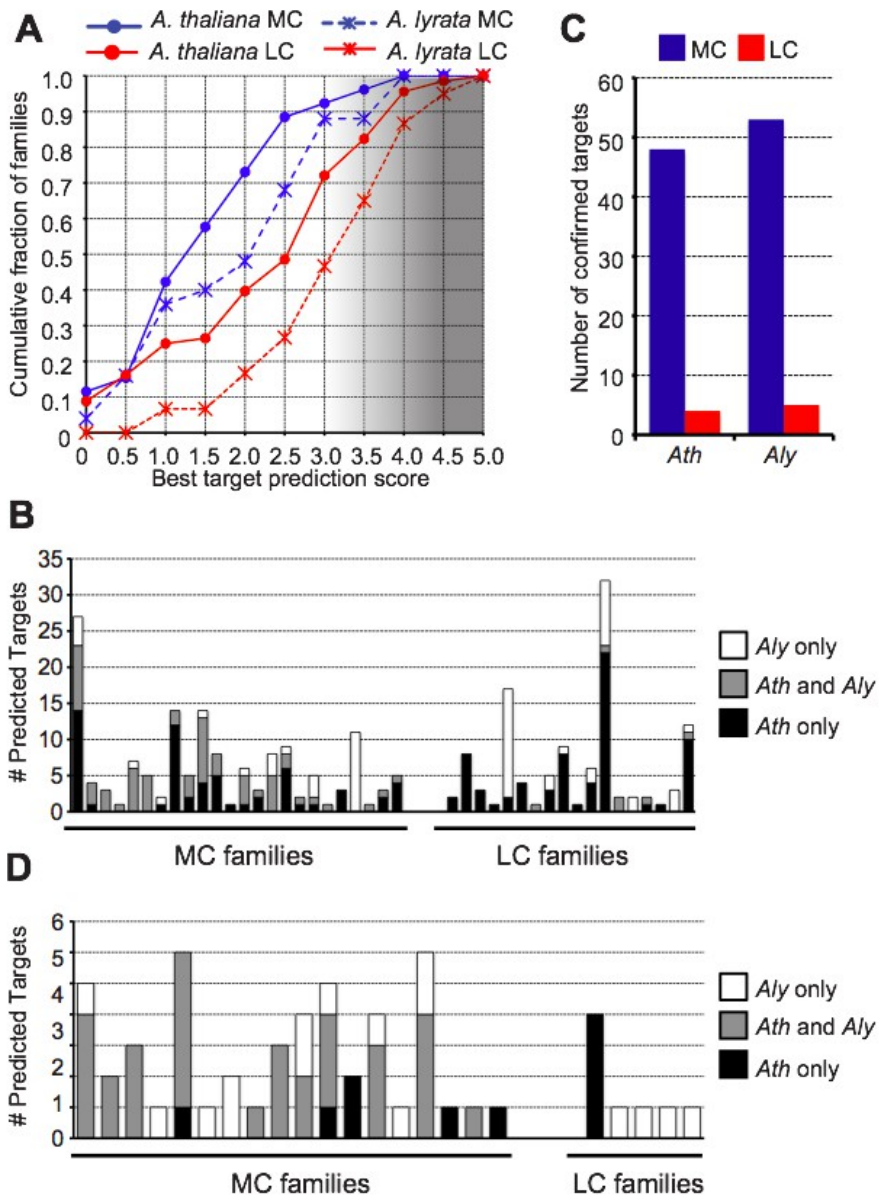


Figure 2.5 Targets of less conserved miRNAs are difficult to identify and inconsistent between *A. thaliana* and *A. lyrata*.

(A) Cumulative distributions of the number of miRNA families with the indicated target prediction scores. The lowest scoring prediction for each family was used. MC: More conserved, LC: Less conserved. Shaded region indicates low confidence predictions (score > 3).

(B) miRNA target predictions by family. The number of predicted targets found only in *A. thaliana* (*Ath*), only in *A. lyrata* (*Aly*), or syntenic homologs predicted in both species are shown. Families without any predicted targets in either species are omitted, as are families which were only expressed

in a single species.

(C) Sliced targets confidently found by degradome sequencing. Sliced targets were those which were found in both biological replicate degradome libraries for the given species.

(D) As in B for degradome-confirmed targets.

2.4.6 Pol IV siRNA occupancy and hotspots differ between *A. thaliana* and *A. lyrata*

We next turned our attention to comparisons between *A. thaliana* and *A. lyrata* 24nt siRNAs, most of which are likely due to the action of the Pol IV/Pol V pathway. A whole genome alignment between *A. thaliana* and *A. lyrata* was performed. The *A. thaliana* genome was divided into 1,000nt bins (119,184 in total) and *A. lyrata* genomic regions aligned to each of the bins were identified. In total, 82.5% of the bins (98,357) were confidently paired with an *A. lyrata* syntenic region; the rest were ambiguous, largely due to small-scale inversions (data not shown).

Repeat-normalized small RNA abundances from nine publicly available *A. thaliana* sRNAseq datasets (Table 2.1) were tabulated for all *A. thaliana* bins. Importantly, these datasets represented diverse wild-type tissues (inflorescences, leaves, seedlings, and siliques) and were produced with multiple technologies (Roche/454 pyrosequencing and Illumina sequencing-by synthesis) by different investigators. Abundances derived from our three *A. lyrata* sRNAseq samples were also calculated for all aligned *A. lyrata* bins. Occupancies of small RNAs of a given size were determined by applying a minimum threshold which corrects for loci that produce multiple sizes of small RNAs (See Methods). Only bins confidently aligned between *A. thaliana* and *A. lyrata* were used for analysis. The number of bins occupied by 24nt RNAs varied from 1,378 to 11,431 (1.4% to 11.6% of 98,357 bins); variations in occupancy were directly correlated with the sequencing depths of the samples. The number of bins which were co-occupied in each pairwise combination of all datasets was calculated (O: Observed overlap). Additionally, the number of co-occupied bins expected by random chance (E: Expected overlap) was also calculated for each pairwise combination. In this scheme, positive values of O/E indicate more co-occupancy than would be expected by chance. Importantly, this metric can be compared across samples of differing sequencing depths. The observed co-occupancies for 24nt RNAs between different *A. thaliana* samples was always greater than expected by chance alone, typically between four- and eight-fold higher (Figure 2.6A). These data indicate that 24nt RNA producing loci are somewhat consistent across different tissues of *A. thaliana* whose small RNAs were sampled at different depths and using different sequencing methodologies. Comparisons to the 24nt RNA accumulation pattern of *A. lyrata* also showed that co-occupancy of syntenic bins occurred more often than expected by chance. However, the enrichments relative to chance alone were very modest, with all O/E ratios less than two-fold for all interspecies comparisons (Figure 2.6A). As a control, the same analysis was performed for 21nt-occupied bins from each sample, with the expectation that because many of these bins contain conserved *MIRNA*s or *TAS* loci, they would be consistently occupied both in the various *A. thaliana* samples and in *A. lyrata*. Indeed, the number of 21nt RNA co-occupied bins in *A. thaliana* intraspecies comparisons and in the *A. lyrata*-*A. thaliana* interspecies comparisons greatly exceeded the values expected by chance alone, and also exceeded the values

seen for 24nt RNA co-occupancy (Figures 2.6A-B). We conclude that only a small percentage of the genomic regions producing 24nt small RNAs in one species also produce 24nt RNAs from the syntenic region in a closely related species. This contrasts with 21nt RNA expressing loci, which are more consistently found to be 21nt expressers in both two species.

Next we examined whether the most active "hotspots" of small RNA expression behaved similarly to the global patterns for all occupied loci. The top 100 24nt-expressing 1kb bins, ranked in order of the abundance of mapped 24mers, were obtained for each *A. thaliana* and *A. lyrata* dataset within the 98,357 confidently aligned 1kb regions. The raw observed overlap between all of the datasets was analyzed. The top 100 21nt-expressing bins from each sample were analyzed in the same way as a control. Similar to the global patterns of occupancy, the consistency of 24nt RNA hotspots between different *A. thaliana* samples was generally lower than that for 21nt hotspots (Figures 2.6C-D). In all *A. thaliana*-only pair-wise combinations, between one and 63 24nt hotspots were in the top 100 in both samples, with a mean of 19.0, while values for 21nt hotspot overlap between *A. thaliana* samples ranged from 32 to 76 with a mean of 54.4. Some of the variation in *A. thaliana* 24nt RNA hotspots was due to tissue specific expression patterns. Particularly striking were the highly consistent 24nt RNA hotspots from the inflorescence and silique samples (Figure 2.6C). These are likely to be type I p4-siRNA loci, defined by their specific expression in reproductive tissues (Mosher et al., 2009). In contrast, 24nt siRNA hotspots from *A. thaliana* leaves were not as consistent, with 14 out of 100 overlapping between the two *A. thaliana* leaf samples examined (Figure 2.6C). The leaf hotspots are likely type II p4-siRNA loci, which are defined by their broad expression patterns, particularly in non-reproductive tissues (Mosher et al., 2009). Other probable sources of the within-species variation in small RNA hotspots include sampling error due to non-saturating sequencing depths, variations in small RNA library construction and particularly in the level of contamination by degraded RNA fragments, and artifacts arising from differing sequencing technologies. Pair-wise comparisons of all nine *A. thaliana* hotspot lists to all three *A. lyrata* hotspot lists revealed very low overlap in the top 100 24nt RNA hotspots; with between zero and two shared syntenic bins (mean = 0.52; Figure 2.6C). This low to non-existent overlap in 24nt RNA hotspots was much lower than the 14 out of 100 overlap seen when comparing two *A. thaliana* leaf samples, but higher than the values expected from random chance (~$1E^{-6}$). In contrast, many more of the top 100 21nt hotspots overlapped between *A. thaliana* and *A. lyrata*; values ranged from 21 to 34 with a mean of 28.3 (Figure 2.6D). Most of these shared 21nt hotspots were abundant miRNAs or *trans*-acting siRNAs (data not shown).

It is possible that the failure to identify more 24nt RNA hotspots shared between *A. thaliana* and *A. lyrata* was because syntenic hotspots disproportionately fell into genomic regions which were

not confidently aligned. To test this, we examined both co-occupancy and the top 100 small RNA hotspots from all 119,184 1kb bins, including the ambiguously aligned bins, for the nine *A. thaliana* datasets. Both for global occupancy, and for hotspots, 24nt expressing regions were indeed more likely to fall into non-aligned bins than were the control 21nt hotspots (Figure 2.6E-F). In all cases, the percentages of 24nt occupied bins or hotspots exceeded the genome-wide percentage of all non-alignable bins. In contrast, the 21nt occupied bins or hotspots fell into ambiguously aligned bins with consistently lower frequencies that were roughly centered upon the genome-wide value (Figures 2.6E-F). Thus, *A. thaliana* 24nt RNA hotspots are indeed more likely to arise from genomic regions difficult to align with *A. lyrata*. However, it should be noted that this effect is modest; most *A. thaliana* 24nt RNA hotspots and occupied loci were in confidently aligned bins but almost none of these were also 24nt RNA hotspots in *A. lyrata*. Like their *A. thaliana* counterparts, *A. lyrata* 24nt small RNAs tended to emanate from regions of the genome devoid of annotated protein-coding capacity (Figure 2.6G). Altogether, we observe a slight tendency of 24nt RNA expressing loci to be retained as 24nt expressers between species, but our data provide little evidence for retention of individual 24nt RNA hotspots between *A. thaliana* and *A. lyrata*. This contrasts strongly with the most active 21nt expressing loci, which are often highly expressed in both species.

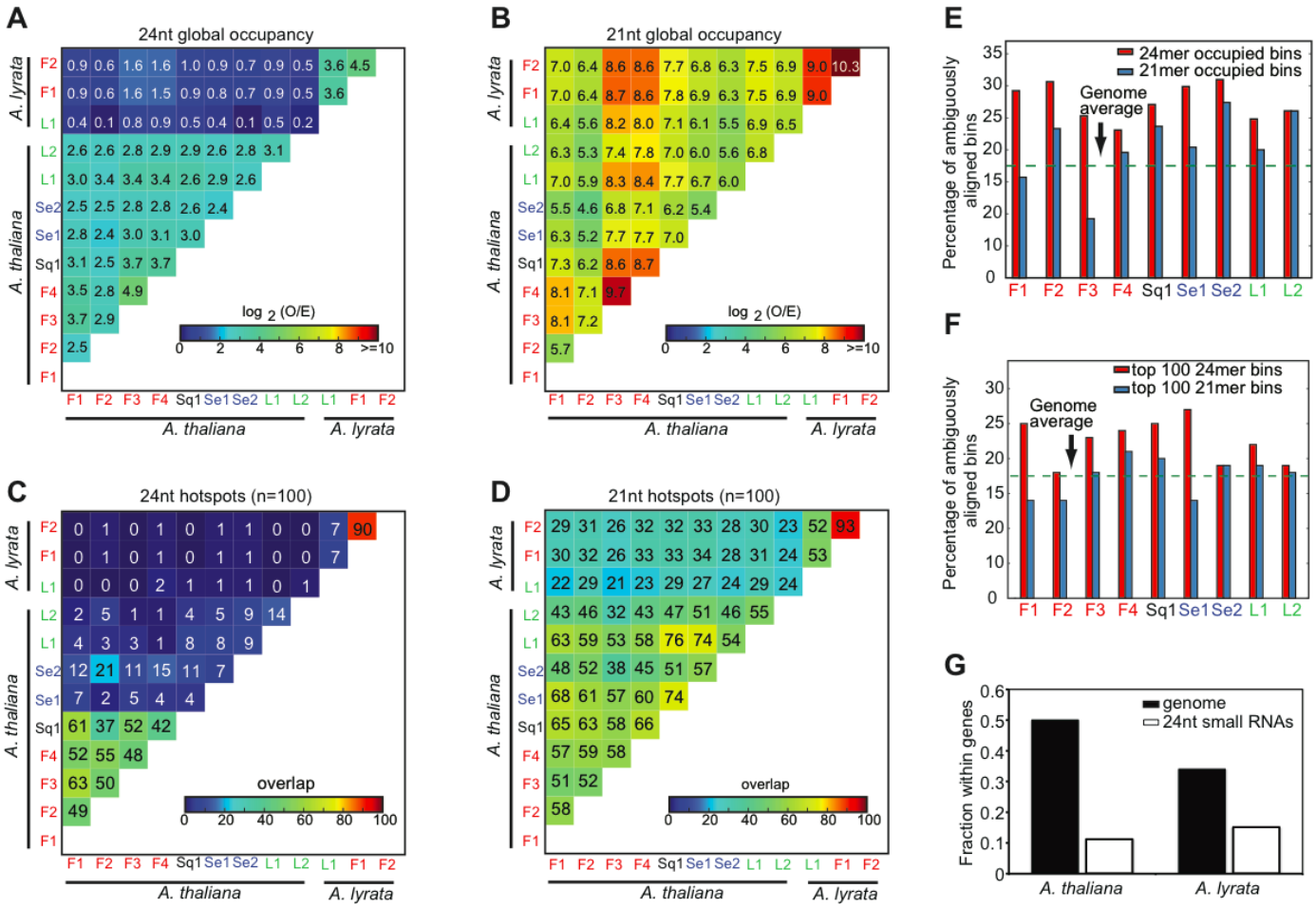


Figure 2.6 24nt RNA expression and hotspots frequently differ between *A. thaliana* and *A. lyrata*.

(A) Log_2 ratios of observed to expected overlaps between 24nt small RNA occupied 1kb bins for all pairwise comparisons between various *A. thaliana* and *A. lyrata* small RNA samples. F1-F4: floral, Sq1: silique, Se1-Se2: seedlings, L1-L2: rosette leaves.

(B) As in A for 21nt small RNA occupied bins.

(C) Overlaps between the top 100 24nt small RNA expressing 1kb bins for all pairwise comparisons between various *A. thaliana* and *A. lyrata* small RNA samples.

(D) As in C for the top 100 21nt small RNA expressing loci.

(E) Percentages of *A. thaliana* 21nt and 24nt small RNA occupied bins which were ambiguously aligned in the *A. thaliana-A. lyrata* whole genome alignment.

(F) As in E for the top 100 *A. thaliana* 21nt and 24nt small RNA hotspots.

(G) Fraction of 24nt small RNAs which mapped to annotated genes in *A. thaliana* and *A. lyrata*. *A. thaliana* sRNAseq data was the combination of all nine sRNAseq libraries, and *A. lyrata* sRNAseq data was the combination of all three sRNAseq libraries (Table 1).

2.5 Discussion

2.5.1 Emergence or degeneration of *MIRNA*s at the species level

The notion that many plant *MIRNA*s are lineage-specific has been clearly supported by comparisons of *MIRNA* inventories between different plant families (Rajagopalan et al., 2006; Fahlgren et al., 2007). *MIRNA* emergence must be fairly rapid, as there are several examples of *MIRNA*s which probably arose specifically in the species *A. thaliana*, as judged by their absence in the closest relative *A. lyrata* (Felippes et al., 2008). Our sampling of *A. lyrata* small RNA expression also indicated that there are many *A. lyrata*-specific *MIRNA*s which arose after the divergence of the *A. thaliana* lineage. We found that, as a group, homologs of "young" *A. thaliana* or *A. lyrata MIRNA* loci frequently do not conform to classical ideas of *MIRNA* biogenesis and function. Specifically, syntenic homologs of young *MIRNA*s in the closest related species 1) frequently lack the capacity to form a *MIRNA*-like stem-loop, 2) have high divergence rates in mature miRNA sequences, 3) tend to lose complementarity with homologs of the known/predicted targets, and 4) are processed with diminished accuracy in the sister species. Mis-annotations of *MIRNA* loci were unlikely to have confounded these results, as we restricted our analyses to a subset of annotated *MIRNA*s for which there is unambiguous experimental evidence for miRNA biogenesis in at least one of the two species being analyzed. Thus, we conclude that many of the less conserved *MIRNA*s either degenerated in one species from a functional common ancestor or were specifically refined from a non-functional ancestor in a given lineage. Either option entails relatively rapid changes in *MIRNA* sequences, processing accuracies, and target repertoires.

At least two hypotheses are consistent with the differences in properties between young *MIRNA* loci and their syntenic homologs in closely related species. The first is that these homologs are performing biologically meaningful regulatory roles in both species. Because of the non-canonical sequence conservation, diminished target-site conservation, and reduced processing accuracies, this hypothesis necessitates that the young *MIRNA*s exert biological effects in a manner which is quite different than for canonical *MIRNA*s. For instance, it could be that these young miRNAs interact with targets with pairing geometries not captured by commonly used prediction methods (Brodersen and Voinnet, 2009). Such pairing configurations could potentially be more tolerant of positional and sequence heterogeneity (although this is not the case for the "seed" targeting which has been extensively documented in animals). Another possibility is that some of these young *MIRNA*s could exert regulatory roles on host transcripts in *cis* simply by being processed by DCL proteins, similar to the suspected functions of *A. thaliana MIR838* and *Physcomitrella patens MIR1047* (Rajagopalan et

al., 2006; Axtell et al., 2007). A second hypothesis which cannot be excluded with currently available data is that many homologs of young *MIRNA*s are simply degenerate and do not function in any biologically relevant role in the sister species. This would in turn imply that these young *MIRNA*s are truly species-specific, in that they exist and function only in a single species but not in its closest relative.

Our analysis highlighted broad trends which differentiated more conserved and less conserved *MIRNA*s between *A. thaliana* and *A. lyrata*. However, it is important to point out that not all less conserved *MIRNA*s followed the overall trends. Some homologs of less conserved *MIRNA*s had conserved mature miRNA sequences within *MIRNA*-like hairpins, maintained high levels of complementarity to their targets, and were processed accurately in both species. Thus, there are certainly some *MIRNA*s which are restricted to the Brassicaceae but also have the properties of more conserved, canonical *MIRNA* loci.

2.5.2 Pol IV siRNA hotspots can be evolutionarily transient

P4-siRNAs, which are 24nt in length, are hypothesized to counteract productive Pol II transcription of intergenic regions by directing chromatin modifications to Pol V transcribed areas (Wierzbicki et al., 2009). Much of the genome is involved in p4-siRNA production at low levels, and there are clear “hotspots” of production which account for a disproportionate amount of p4-siRNA production (Zhang et al., 2007). We find that global expression patterns of 24nt RNAs, which we presume to be p4-siRNA loci, are relatively consistent across different tissue samples derived from *A. thaliana*, with pair-wise overlaps between different samples consistently four- to eight-fold higher than expected by chance alone. In contrast, 24nt RNA hotspots are consistent between different inflorescence and silique samples of *A. thaliana*, but often differ between vegetative (leaves and seedlings) and reproductive (inflorescences and siliques) tissues of *A. thaliana*. These data imply that type I p4-siRNAs (defined by their absence from vegetative tissues) have more reproducible hotspots than do type II p4-siRNAs (defined by their presence in vegetative tissues). Despite the higher intra-species variability in the vegetative p4-siRNA hotspots, there are a significant number which are reproducibly active in different samples: For instance, 14 out of the top 100 p4-siRNA hotspots are shared between two independently derived *A. thaliana* leaf samples (Figure 2.5C). However, there is essentially no significant overlap between the top 100 p4-siRNA loci expressed in *A. lyrata* leaves or inflorescences and the top 100 p4-siRNA loci from any *A. thaliana* tissue. Similarly, there is only a slight trend toward overlap in the global patterns of p4-siRNA accumulation between *A. lyrata* and *A. thaliana* regardless of expression level. Thus, the loci which produce p4-siRNAs often differ between *A. thaliana* and *A. lyrata*.

The significance of most individual p4-siRNA hotspots within plant genomes is unknown. In

Drosophila melanogaster, Piwi-interacting RNAs (piRNAs) have some functional analogies to plant p4-siRNAs. Like plant p4-siRNAs, fly piRNAs also function to silence transposable element expression by the use of Watson-Crick interactions between the small RNA and target (Aravin et al., 2007). In germline cells and surrounding somatic support cells, piRNA expression disproportionately emanates from just a few master regulator loci, the most prominent of which is *flamenco* (Brennecke et al., 2007; Lau et al., 2009; Malone et al., 2009). The *flamenco* locus is an ~180kb region dominated by transposon fragments which are arranged almost exclusively in a single orientation (Brennecke et al., 2007). The abundant piRNAs produced from the *flamenco* locus function to initiate post-transcriptional silencing of active transposon copies elsewhere in the genome. The *flamenco* piRNA hotspot is critical for suppression of *gypsy* retroelements in the female germline (Prud'homme et al., 1995). Importantly, *flamenco* is conserved with respect to high production of piRNAs and single-stranded transposon orientation in both *D. yakuba* and *D. erecta* (Malone et al., 2009). By analogy with *flamenco*, one potential function for plant p4-siRNA hotspots could be as master loci which produce siRNAs with the capacity to silence expression of many unlinked transposons with sequence similarity *in trans*. However, unlike the *flamenco* analogy, we did not find evidence for maintenance of high expression for any p4-siRNA hotspots between two closely related plant species. This implies that highly active individual p4-siRNA loci can, like less-conserved *MIRNA*s, be evolutionarily transient in plants.

Accession Numbers

Newly generated *A. lyrata* small RNA data has been deposited at NCBI GEO (GSE18077 and GSE20442). *A. lyrata* and *A. thaliana* degradome data have also been deposited at NCBI GEO (GSE20451). Accession numbers for all datasets used in this study are listed in Table 2.1.

*The work in Chapter 2 is published (Ma et al., 2010) and is reproduced here with minor modifications.

Chapter 3
Small RNAs in other plants

3.1 Summary

Small RNAs are broadly present in all known plant species, many of which play important regulatory roles. Here we surveyed the small RNA populations from three plants: the tree crop *Theobroma cacao*, oil palm *Elaeis guineensis* Jacq., and the model moss *Physcomitrella patens*. In *Theobroma cacao*, we computationally identified 83 conserved miRNAs and 91 miRNA targets using sequence similarity and secondary structure information. In oil palm (*Elaeis guineensis Jacq.*), we identified 28 expressed miRNA families during flower development by analyzing smallRNAseq data. In *Physcomitrella patens*, *w*e identified a novel family of trans-acting siRNA (ta-siRNA) loci associated with miR156- and miR529-directed slicing by scanning the genome for ta-siRNA-like sRNA accumulation patterns in different genetic background. These studies as a whole demonstrate that many small RNA species are deeply conserved in the plant kingdom. On the other hand, novel classes of small RNAs can evolve in specific lineages.

3.2 *Theobroma cacao* miRNAs

3.2.1 Introduction

The cocoa tree *Theobroma cacao* is an important tree crop native to the South American rainforests, which is the source of the main ingredients of chocolate. The Criollo cocoa variety was cultivated by the Maya over 1500 years ago (Motamayor et al., 2002), and is now one of the two varieties that provide fine flavor chocolate. The Criollo variety is highly homozygous for its diploid genome (2n = 20), thus suitable for a high-quality genome assembly. As part of the international effort that sequenced and analyzed the genome of *Theobroma cacao*, we annotated the conserved miRNAs in the genome and predicted the conserved miRNA targets.

3.2.2 Methods

3.2.2.1 *Theobroma cacao* miRNA annotation

Sequences of mature plant miRNAs were retrieved from miRBase release 14 (Griffiths-Jones et al., 2008) and used as queries to search the *T. cacao* genome assembly using BLASTN. Hits with

no more than one mismatch from a query were expanded to 150 nt upstream and 150 nt downstream and examined by MIRcheck (Jones-Rhoades and Bartel, 2004). miRNA candidates that were on the same arm of the hairpin as the known family members and passed MIRcheck with the parameters "-mir_bulge",3, "-ass", 2,"-unpair" were collapsed to retain a single miRNA for a given hairpin if length variants or position variants are present. The decision of which variant to retain was made as follows: for length variants, if the miRNA family was expressed in *A. thaliana* (Ma et al., 2010), then the miRNA variant with the length of the most abundantly expressed miRNA was kept; if not, a 21-mer was favored. For positional variants, the miRNA variant with the greatest number of similar miRNA sequences in miRBase was retained.

3.2.2.2 *Theobroma cacao* microRNA target prediction

miRNA targets were predicted with the PERL script “axtell_targetfinder.pl” from the CleaveLand 2 package (http://axtell-lab-psu.weebly.com/cleaveland.html). Randomization using 8300 randomly-shuffled miRNAs (100 times of the total number of real miRNAs) was also done using the same script. An average of 0.7 targets were predicted at a complementarity score of three for each randomized miRNA; this noise estimation increased to over two at a score of four. Therefore, a cutoff score of three was used for target prediction (higher target prediction scores indicate less complementarity).

Predicted targets were used as queries to search *A. thaliana* (TAIR9) and *Oryza sativa* mRNAs, and the gene annotation of top BLASTX hit with an E value <= 10E-4 in either species (if available) was used to indicate the potential function of the predicted targets in *T. cacao*. GO annotation was performed with homologous *A. thaliana* genes of the predicted targets to search for GO term enrichment (Berardini et al., 2004). P values are calculated based on hypergeometric tests.

3.2.3 Results

A total of 83 *T. cacao* microRNAs (miRNAs) from 25 families were computationally predicted based on sequence similarity with known miRNAs in miRBase release 14 (Table 3.1, Appendix: Supplemental Dataset 3.1 & 3.2). The miRNA population size is reasonable compared to the number of miRNAs in other plant genomes in miRBase (Figure 3.1), although our tally of *T. cacao* miRNAs is certainly an underestimate, as we were limited to identification by homology.

Table 3.1 miRNA families found in *Theobroma cacao.*

miRNA family	Number of paralogous loci	Number of plant species where also found	List of species
156	7	17	aqc, ath, bdi, bna, ghr, gma, mtr, osa, ppt, pta, ptc, sbi, sly, smo, sof, vvi, zma
160	3	15	aqc, ath, bdi, bra, gma, mtr, osa, ppt, ptc, sbi, sly, smo, tae, vvi, zma
162	1	10	ath, cpa, ghr, gma, mtr, osa, ptc, sly, vvi, zma
164	3	11	ath, bna, bra, gma, mtr, osa, ptc, sbi, tae, vvi, zma
166	4	17	aqc, ath, bdi, bna, ghr, gma, mtr, osa, ppt, pta, ptc, pvu, sbi, sly, smo, vvi, zma
167	3	17	aqc, ath, bdi, bna, bra, gma, lja, mtr, osa, ppt, ptc, sbi, sly, sof, tae, vvi, zma
168	1	11	aqc, ath, bna, gma, mtr, osa, ptc, sbi, sof, vvi, zma
169	14	13	aqc, ath, bdi, bna, ghb, gma, mtr, osa, ptc, sbi, sly, vvi, zma
171	8	18	aqc, ath, bdi, bna, bol, bra, gma, mtr, osa, ppt, pta, ptc, sbi, sly, smo, tae, vvi, zma
172	5	13	aqc, ath, bdi, bol, bra, gma, mtr, osa, ptc, sbi, sly, vvi, zma
319	1	14	aqc, ath, gma, mtr, osa, ppt, pta, ptc, pvu, sbi, sly, smo, vvi, zma
390	2	11	ath, bna, ghr, gma, mtr, osa, ppt, pta, ptc, sbi, vvi
393	2	9	ath, bna, gma, mtr, osa, ptc, sbi, vvi, zma
394	2	6	ath, osa, ptc, sbi, vvi, zma
395	2	10	aqc, ath, mtr, osa, ppt, ptc, sbi, sly, vvi, zma
396	5	15	aqc, ath, bna, ghr, gma, lja, mtr, osa, pta, ptc, sbi, smo, sof, vvi, zma
397	1	8	ath, bdi, bna, osa, ptc, sbi, sly, vvi
398	2	9	aqc, ath, bol, gma, mtr, osa, pta, ptc, vvi
399	9	14	aqc, ath, bdi, bna, ghr, mtr, osa, ptc, pvu, sbi, sly, tae, vvi, zma
403	2	3	ath, ptc, vvi
529	1	4	aqc, osa, ppt, sbi
530	2	3	aqc, osa, ptc
535	1	4	aqc, osa, ppt, vvi
827	1	3	ath, osa, ptc
2111	1	2	ath, bna

aqc: *Aquilegia coerulea*, ath: *Arabidopsis thaliana*, bdi: *Brachypodium distachyon*, bna: *Brassica napus*, bol: *Brassica oleracea*, bra: *Brassica rapa*, cpa: *Carica papaya*, ghb: *Gossypium herbecium*, ghr: *Gossypium hirsutum*, gma: *Glycine max*, lja: *Lotus japonicus*, mtr: *Medicago truncatula*, osa: *Oryza sativa*, ppt: *Physcomitrella patens*, pta: *Pinus taeda*, ptc: *Populus trichocarpa*, pvu: *Phaseolus vulgaris*, sbi: *Sorghum bicolor*, sly: *Solanum lycopersicum*, smo: *Selaginella moellendorffii*, sof: *Saccharum officinarum*, tae: *Triticum aestivum*, vvi: *Vitis vinifera*, zma: *Zea mays*

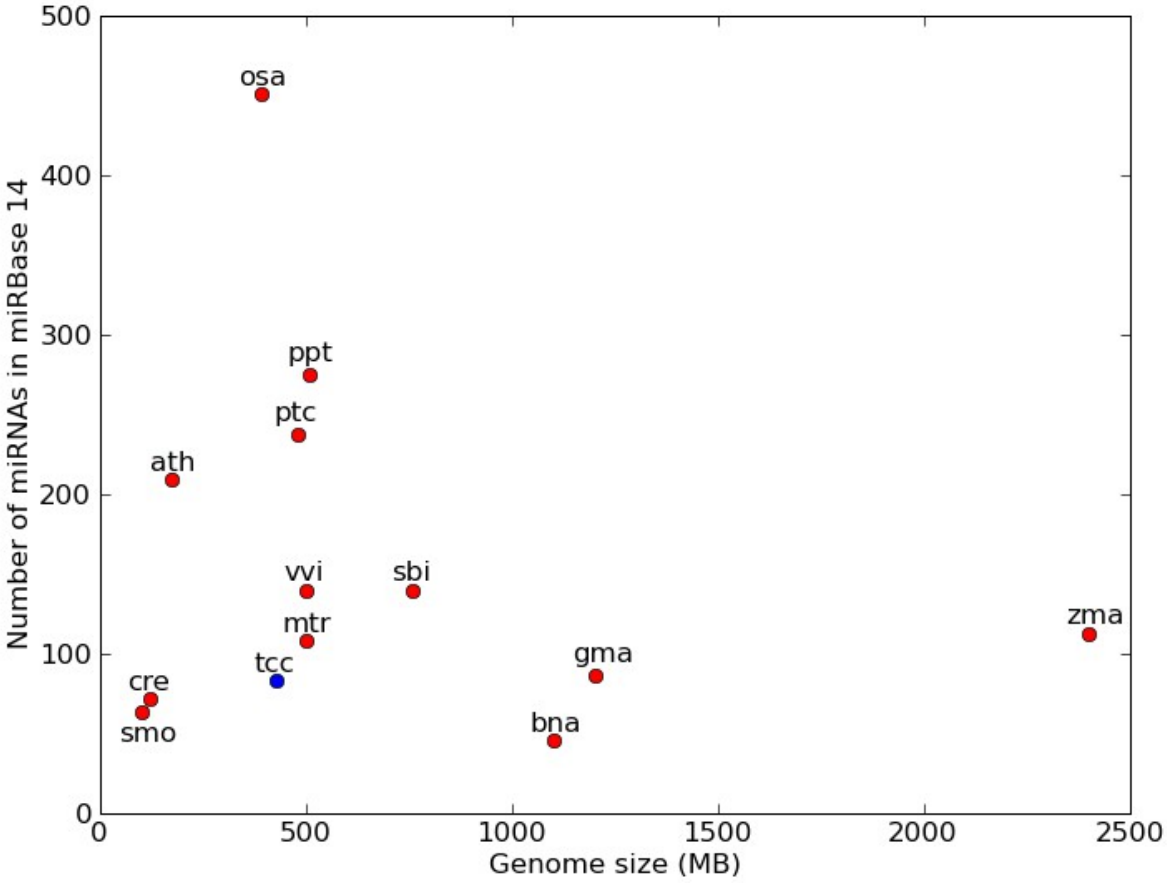


Figure 3.1 Number of miRNAs in each plant species in miRBase 14 and *Theobroma Cacao* with the corresponding genome size. All plant species with more than 45 miRNAs in miRBase 14 and a known genome size (NCBI http://www.ncbi.nlm.nih.gov/genomeprj/?term=txid33090[Organism%3Aexp]) are plotted. No obvious correlation of miRNA population and genome size is observed. The high variation in the number of miRNAs in different species are mostly due to different discovering methods (prediction only versus experimental confirmation) or different level of stringency of the prediction. ath: *Arabidopsis thaliana*, bna: *Brassica napus*, cre: *Chlamydomonas reinhardtii*, gma: *Glycine max*, mtr: *Medicago truncatula*, osa: *Oryza sativa*, ppt: *Physcomitrella patens*, ptc: *Populus trichocarpa*, sbi: *Sorghum bicolor*, smo: *Selaginella moellendorffii*, tcc: *Theobroma Cacao*, vvi: *Vitis vinifera*, zma: *Zea mays*

Because 25 T. cacao miRNA families were encoded by 83 loci, the number of paralogous loci per family was examined. Compared with *A. thaliana*, the cumulative distribution of *T. cacao* miRNAs was similar to the more-conserved (MC) subset of the *A. thaliana* miRNAs (annotated outside of the Brassicaceae family in miRBase release 14), but quite different from the miRNA population in *A. thaliana* (Figure 3.2). This is expected, because the miRNA prediction method finds only miRNAs conserved between *T. cacao* and another species.

89 targets of 19 miRNA families were predicted (Appendix: Supplemental Dataset 3.3), and 85 top hits (68 unique proteins) in *A. thaliana* and 83 top hits (66 unique) in *O. sativa* were identified by BLASTX search above the cutoff of E value <= 10E-4. Because GO annotation was not yet available for *T. cacao*, GO annotation for the 68 unique *A. thaliana* genes homologous to the predicted *T. cacao* targets was used to search for GO term enrichment (Berardini et al., 2004). The terms “nucleus”, “transcription factor activity” and “developmental processes” were the most significantly enriched terms in each GO category (Table 3.2) for *T. cacao* miRNA target homologs in *A. thaliana* compared to the entire *A. thaliana* genome. These results are consistent with previous findings (Axtell and Bowman, 2008) that many conserved miRNA targets are transcription factors involved in developmental processes.

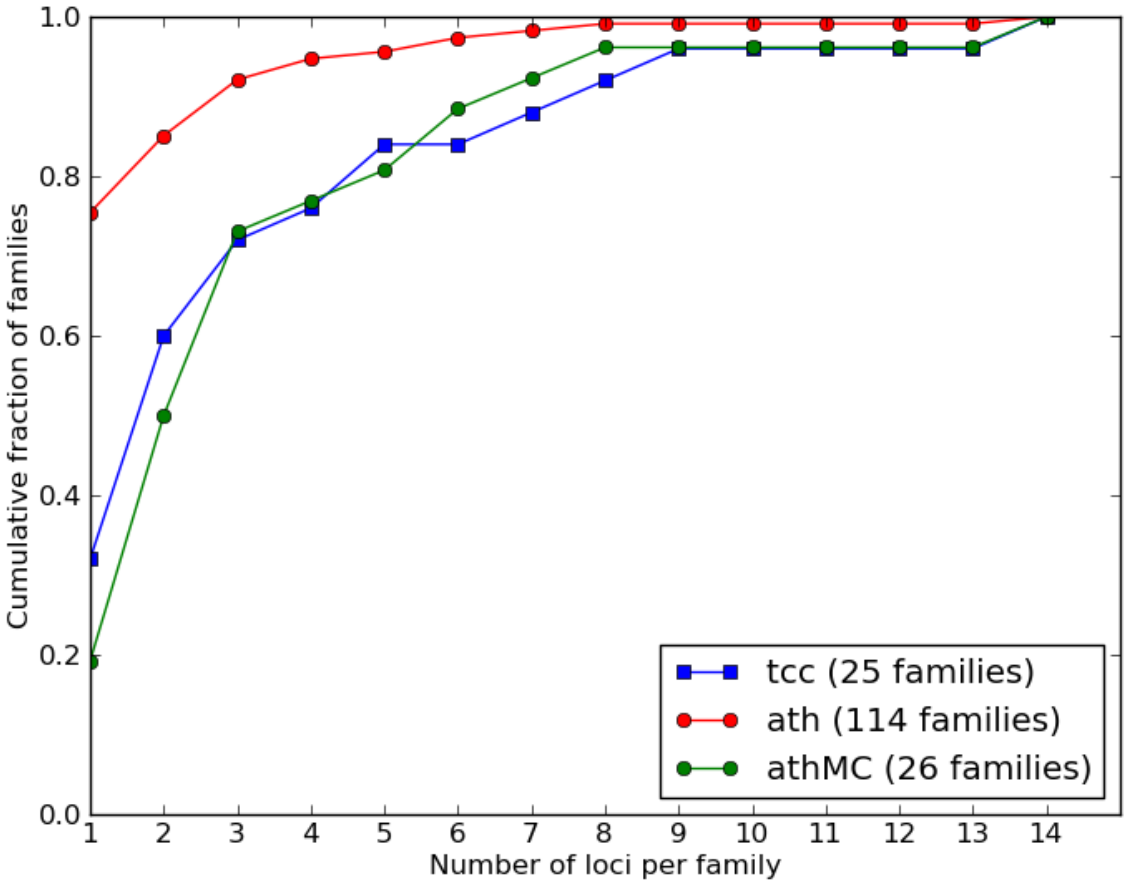


Figure 3.2 Cumulative distributions of the number of loci per miRNA family in *T. cacao* and *A. thaliana*. tcc: *T. cacao*, ath: *A. thaliana*, athMC: *A. thaliana* more conserved families.

Table 3.2 Gene ontology (GO) annotation of cacao miRNA target homologs in *A. thaliana*. Terms in bold indicates the most significant enrichment in each GO category

Term	Ontology category	Gene number of cacao miRNA target homologs in *A. thaliana* (total 67)	Gene number in *A. thaliana* genome (total 34278)	P-value
Nucleus	Cellular component	24	2609	6.97E-12
Extracellular	Cellular component	7	441	2.37E-06
Transcription factor activity	Molecular function	23	1679	3.93E-15
Other enzyme activity	Molecular function	17	3345	5.30E-05
DNA or RNA binding	Molecular function	14	2714	1.93E-04
Transporter activity	Molecular function	7	1242	2.85E-03
Developmental processes	Biological process	24	2006	2.01E-14
Transcription	Biological process	20	1709	5.67E-12
Other cellular processes	Biological process	43	10140	1.08E-09
Other metabolic processes	Biological process	41	9410	1.77E-09
Other biological processes	Biological process	12	1913	7.32E-005

3.2.4 Conclusions

MicroRNAs (miRNAs) are short noncoding RNAs that regulate target genes transcriptionally or post-transcriptionally. Many of them play important roles in development and stress responses. A total of 83 T. cacao miRNAs from 25 families were computationally predicted based on sequence similarity to known plant miRNAs in miRBase 14. Ninety-one *T. cacao* miRNA targets were predicted. Most predicted targets were homologous to known miRNA targets in other plant species, but there was a profound bias toward putative transcription factors compared to the other species (Table 3.2), suggesting that miRNAs are major regulators of gene expression in *T. cacao*.

* The work presented in Chapter 3.2 has been published (Argout et al., 2011) and is reproduced with minor modifications.

3.3 Expressed miRNAs in oil palm (*Elaeis guineensis Jacq.*) during flower development

3.3.1 Introduction

Oil palm (*Elaeis guineensis* Jacq.) is a perennial monocot and an important economic crop (Price et al., 2007). Because of unavailable genomic data, long life cycle and inaccessibility of the floral meristems of the oil palm plant, it is difficult to study miRNAs that are expressed during oil palm flower development. In this study, we analyzed smallRNAseq data from inflorescence, emerging flower and mature flower of oil palm, and identified 28 families of conserved miRNAs are expressed during flower development.

3.3.2 Methods

SmallRNAseq data from inflorescence (IF), emerging flower (EF) and mature flower (F) of oil palm were kindly provided by J. A. Harikrishna from the University of Malaya. A detailed description of the three developmental stages is presented in (Mehrpooyan et al., 2012). Sequencing reads were pre-processed by BGI to remove adapters and remove low quality reads. A previously curated list of highly confident microRNAs (miRNAs) from *Arabidpsis thaliana*, rice and *Physcomitrella patens* (by examining processing precision using small RNA seq data in respective genomes) were combined into a non-redundant fasta file (Appendix: Supplemental Dataset 3.4) to serve as the reference to map the sequencing reads. The reason for not using all plant miRNAs in miRBase (as was done in Chapter 3.2) is that we do not have a reference genome in this study, therefore no secondary structure information can be used to distinguish authentic miRNAs from bogus ones. Thus we started with a highly confident set of real miRNAs for oil palm miRNA identification.

Sequencing reads were mapped to the high-confidence miRNA reference by Bowtie 0.12.7 (Langmead et al., 2009) allowing one mismatch (the parameter was set as “-v 1”). Mapped reads were then examined for sizes and only sizes in the range of 20-24 nt were retained. Each mapped read was assigned to a single miRNA family based on the top matching reference miRNA. In the occasional cases where the read was mapped to two families with equal quality (namely, miR156/157, miR165/166 and miR159/319), it was arbitrarily assigned to the miRNA family with smaller numbers (namely, miR156, miR165 and miR159). As a result, multiple sequence variants can be assigned to a given miRNA family, therefore, the most abundant variant in each of the three datasets were examined. If at least two datasets agree on the most abundant variant, the specific variant is considered the real miRNA. It should be noted that other variants are likely real miRNAs too, maybe from paralogous loci in the genome. However, without a reference genome, we prefer to tolerate more

false negatives rather than increasing false positives.

3.3.3 Results and Discussion

We found that 28 miRNA families were expressed in oil palm flowers (Table 3.3). Most of the cases (19 out of 28), all three datasets agree on the most abundant miRNA variant, indicating that these 19 miRNAs are consistently expressed through flower development. Despite the consistent expression, the raw abundances of these 19 miRNAs vary greatly in the three datasets. However, the raw abundance cannot be applied to quantify the expression levels, because without replicates, normalization between the datasets is difficult. In 9 cases, not all datasets agree on the most abundant variant. Two possible explanations exist. One is that other paralogous miRNAs are also expressed and the expression varies during flower development. The other explanation is that the miRNA variant abundantly expressed in two stages of the flower development is down-regulated in the third development stage. With the limitation of the data available, unfortunately we cannot distinguish the two.

It is worth noting that our method to identify expressed miRNAs will intrinsically miss miRNAs that are only expressed in one stage during flower development. With the available data, we cannot assume a read from a single dataset matching a known miRNA is originated from an *MIRNA* locus. Doing so will incur many false positives. This limitation can be resolved if replicates of each development stage are performed.

Table 3.3 Expressed miRNA families and the most abundant miRNA variant in each family during oil palm flower development.

miRNA family	Most abundant miRNA variant	# datasets present*	Raw abundance IF**	Raw abundance EF**	Raw abundance F**
156	UGACAGAAGAGAGUGAGCAC	3	10716	437	8534
157	UUGACAGAAGAUAGAGAGCAC	3	4517	652	4448
159	UUUGGAUUGAAGGGAGCUCUA	3	3425	481	3083
160	UGCCUGGCUCCCUGUAUGCCA	3	527	133	10
162	UCGAUAAACCUCUGCAUCCGG	3	951	978	1787
165	UCGGACCAGGCUUCAUUCCCC	3	2111	46689	16062
166	UCGGACCAGGCUUCAUUCCUC	3	164	2733	1061
168	UCGCUUGGUGCAGGUCGGGAA	3	9088	11133	56093
170	UGAUUGAGCCGUGCCAAUAUC	3	501	321	314
171	UUGAGCCGCGCCAAUAUCACU	3	40	3	2
172	AGAAUCUUGAUGAUGCUGCAU	3	4141	2170	13784
390	AAGCUCAGGAGGGAUAGCGCC	3	421	525	64
396	UUCCACAGCUUUCUUGAACUU	3	87	48	323
444	UGCAGUUGCUGCCUCAAGCUU	3	13	100	25
529	AGAAGAGAGAGAGUACAGCCU	3	919	141	1319
535	UGACAACGAGAGAGAGCACGC	3	32376	27909	141276
827	UUAGAUGACCAUCAGCAAACG	3	2938	132	664
894	CGUUUCACGUCGGGUUCACC	3	2	13	94
2118	UUGCCGAUGCCUCCCAUUCCA	3	708	185	273
158	UCCCAAAUGUAGACAAAGCA	2 (EF= F)	0	73	2
164	UGGAGAAGCAGGGCACGUGCA	2 (IF= F)	5014	513	2650
167	UGAAGCUGCCAGCAUGAUCUA	2 (IF=EF)	4815	26186	7
169	CAGCCAAGGAUGACUUGCCGG	2 (IF=EF)	301	696	66
319	UUGGACUGAAGGGAGCUCCCU	2 (EF= F)	23	232	34
393	UCCAAAGGGAUCGCAUUGAU	2 (IF= F)	3	5	7
394	UUGGCAUUCUGUCCACCUCC	2 (IF=EF)	118	20	0
395	CUGAAGUGUUUGGGGGAACUC	2 (EF= F)	2	1	1
397	UCAUCGAGUGCAGCGUUGAUG	2 (IF=EF)	8	90	143

* Number of datasets in which the miRNA variant is the most abundant

** IF: inflorescence; EF: emerging flower; F: mature flower

3.4 Identification of additional *DCL4/RDR6* dependent *TAS* loci in *Physcomitrella patens*

3.4.1 Introduction

Trans-acting siRNAs (ta-siRNAs) silence genes that are distinct from the loci of their own production (thus the name trans-acting). Biogenesis of ta-siRNAs relies on both the miRNA and siRNA pathways (Allen et al., 2005; Talmor-Neiman et al., 2006). Non-coding transcripts produced from loci known as *TAS* genes are first cleaved by an miRNA which defines one end of the siRNA production, then converted into dsRNAs by the RNA-dependent RNA polymerase RDR6, and subsequently diced into ~ 21 nt ta-siRNAs by DCL4, a functionally conserved Dicer-like enzyme between *Arabidopsis* and *Physcomitrella* (Allen et al., 2005; Talmor-Neiman et al., 2006).

Four *TAS* gene families have been identified (*TAS1-4*) in *Arabidopsis*. *TAS1* and *TAS2* transcripts are cleaved by miR173, *TAS3* by miR390 and *TAS4* by miR828 (Allen et al., 2005; Rajagopalan et al., 2006). Recently, a fifth *TAS* family (*TAS5*), cleaved by miR482, was identified in tomato (Li et al., 2012a). In *Physcomitrella*, four loci of the *TAS3* family *TAS3a-d* have been identified previously, all of which are cleaved at two distinct sites by miR390 spanning the region of tasiRNA production (Talmor-Neiman et al., 2006; Axtell et al., 2006). ta-siRNAs produced from these loci negatively regulate transcription factors with an N-terminal AP2 domain, and also Auxin Response Factors (ARFs) by slicing their mRNAs (Talmor-Neiman et al., 2006; Axtell et al., 2007).

In *dcl4* mutants, the size profile of the siRNAs produced at the *TAS* loci shift from 21nt-dominance to 23-24nt dominance (Howell et al., 2007b; Liu et al., 2007), while in *rdr6* mutants, nearly all siRNA accumulation at the *TAS* loci is lost (Talmor-Neiman et al., 2006). In this study, we used available *Physcomitrella* small RNA-seq data to search for additional small RNA loci showing similar changes in siRNA accumulation in ∆*PpDCL4* and ∆*PpRDR6* plants.

3.4.2 Methods

Small RNA-seq data from *Physcomitrella* wild type (GSM115095, GSM115096, GSM115097, GSM313212, and GSM313213), *dcl3* (GSM313214, GSM313215) and *rdr6* (GSM313216, GSM313217) was previously described (Axtell et al., 2006; Cho et al., 2008) and a small RNA-seq library from *Physcomitrella dcl4* mutant was created (GSM459911) and used in the analysis. 20-24nt small RNAs from each of the ten libraries were mapped to the *Physcomitrella* genome (v1.1) using Bowtie (Langmead et al., 2009), v0.12.7, with settings "-v 1 -a --best --strata -f -p 7 -k 100" as well as "-C" for color space data. Therefore, mappings with either zero or one mismatch were accepted, and for reads that had more than 100 optimal mapping positions, only the first 100 were reported. Reads

mapping to rRNA, tRNA, or *MIRNA* hairpin positions were then discarded. The total small RNA abundance in 200nt non-overlapping bins was calculated genome-wide, and scaled to reads per million to facilitate comparisons between different datasets. The top 1000 most abundant bins in the wild type datasets were then screened to find bins meeting the following criteria: 1) Symmetry (defined as the proportion of small RNAs mapped to one or the other strand) of 0.25-0.75, 2) *rdr6* total < 0.2 x wild type total, 3) Percentage 21-22nt RNAs in wild type and *dcl3* > 70%, and 4) Percentage 23-24nt RNAs in *dcl4* > 50%. These parameters were inferred from the characteristics of the previously known *PpTAS3a*-d loci. The three novel bins which passed this filter were then further examined by search for miRNA complementary sites (based on BLASTN) within a centered 1000nt window. The final annotations of these three novel loci, *PpTAS3e*, *PpTAS3f*, and *PpTAS6a* were based on manual curation and the positions of the miRNA complementary sites. *PpTAS6b* did not pass the initial computational filter but was identified serendipitously based upon its proximity to *PpTAS3f*. Similarly, *PpTAS6c* lacked corresponding siRNAs in our dataset but was identified based on proximity to *PpTAS3d* and the presence of a miR156 target sites. The siRNA population may be generated from *PpTAS6c* under different physiological conditions or different tissue types. Degradome sequencing data from *Physcomitrella* has been previously described (Addo-Quaye et al., 2009b). CleaveLand 2 (Ma et al., 2010) was used for degradome analysis. For siRNA target predictions we used axtell_targetfinder.pl (Ma et al., 2010). Phylogenetic analysis of *PpTAS3a-f* utilized a MUSCLE alignment (Edgar, 2004) of the regions between the miR390 complementary sites as input to the Maximum Likelihood method in MEGA5 based on the Tamura-Nei model (Tamura and Nei, 1993). The bootstrap consensus tree, based on 500 replicates, was shown (Figure 3.9). A .bed file has been created (Appendix: Supplemental Dataset 3.6) which indicates genomic locations of *Physcomitrella TAS* loci (genome assembly version 1.1) as well as important features within the *TAS* loci.

3.4.3 Results

The four previously known *PpTAS3* loci were discovered based upon phased patterns of siRNA accumulation and the presence of miR390 complementary sites (Axtell et al., 2006; Talmor-Neiman et al., 2006). As described above, siRNA accumulation patterns at these four loci change in distinctive patterns in Δ*PpDCL4* and Δ*PpRDR6* plants; Δ*PpDCL4* mutants shift the size profile of the siRNAs from a 21nt-dominated pattern to a 23-24nt dominated pattern, while nearly all siRNA accumulation is lost in Δ*PpRDR6* mutants. We therefore used existing *Physcomitrella* smallRNAseq data to search for additional small RNA loci showing similar changes in siRNA accumulation in Δ*PpDCL4* and Δ*PpRDR6* plants. Initially, seven loci were identified, four of which overlapped the previously described *PpTAS3a-d* (Figure 3.3-3.6, Table 3.4). Two of the three novel loci overlapped previously described hotspots of non-miRNA 21nt-dominated small RNA expression (*Pp21SR6* and

Pp21SR39; Cho et al., 2008). The third novel locus as well as *Pp21SR6* possessed dual miR390 complementary regions and thus we named them *PpTAS3e* and *PpTAS3f*, respectively. Degradome data (Addo-Quaye et al., 2009b) demonstrated that the four miR390 sites within *PpTAS3e* and *PpTAS3f* were sliced (Table 3.4, Figure 3.7-3.8). Two distinct *PpTAS3* ta-siRNA populations have previously been shown to be capable of directing the slicing of targets *in trans*: One population targets *AP2*-domain transcripts, while the other directs slicing of *ARF*-domain transcript (Talmor-Neiman et al., 2006; Axtell et al., 2007; Khraiwesh et al., 2010). *PpTAS3e* could produce a ta-siARF with reasonable base-pairing to one of the known ta-siARF target sites, while *PpTAS3f* could produce both ta-siAP2 and ta-siARF (Table 3.4, Figure 3.7-3.8, Appendix: Supplemental Dataset 3.5). The *PpTAS3e* siRNA population was typical in that it was largely contained between, and in phase with the two miR390 complementary regions (Figure 3.7). In contrast, the *PpTAS3f* siRNA population extended beyond the region bounded by the dual miR390 complementary sites, and was not well-phased (Figure 3.8).

The final novel *DCL4/RDR6*-dependent siRNA locus that we found in our initial computational screen produced siRNAs from a region bounded by upstream miR156 and miR529 complementary sites and a downstream miR156 site (Figure 3.3A). Degradome data support slicing at all three complementary sites and the siRNA population is largely in a 21nt phase with the 3' miR156 cleavage site (Figure 3.3A). This locus, previously described as *Pp21SR39* but heretofore unrecognized as a *TAS* locus (Cho et al., 2008), did not share detectable similarity with any previously described families of plant *TAS* loci. Therefore we renamed it *PpTAS6a*. Curiously, *PpTAS6a* is located within ~0.7kb of *PpTAS3a*, and their respective miRNA complementary sites are located on the same strand (Figure 3.3A). Inspection of the genomic region surrounding *PpTAS3f* revealed a second nearby cluster of mainly 21nt siRNAs in the wild type. We found that this region also contained a high confidence miR156 complementary site downstream of the siRNA generating region which showed degradome-based evidence of slicing (Figure 3.8). In addition, upstream miR156 and miR529 sites could also be observed, albeit with extensive mismatches and no evidence of slicing. Based on this fact, we named this locus *PpTAS6b*, although there was no detectable sequence similarity outside of the pattern of miR156 and miR529 complementary sites between *PpTAS6a* and *PpTAS6b*. Like its neighbor *PpTAS3f*, siRNA production from *PpTAS6b* was atypical in that it was not in phase with the miRNA-directed cleavage site(s) and extended for a long distance beyond the complementary sites (Figure 3.8). Finally, we detected sequence similarity between *PpTAS6a* and a region neighboring *PpTAS3d.* Although we could find no evidence of small RNA production from this region, we did observe a pattern of miR156 and miR529 complementary sites similar to the other *TAS6* loci with some weak evidence of slicing; therefore, we have provisionally named this locus *PpTAS6c* (Figure 3.6). Phylogenetic analysis demonstrated that the six *PpTAS3* loci fall into two distinct clades: *PpTAS3a, -d,*

and *-f* form one clade, while *PpTAS3b*, *-c*, and *-e* form the other clade (Figure 3.9). All three members of the *PpTAS3a/d/f* clade have upstream flanking regions which produce miR156 and/or miR529-sliced RNAs, in at least two cases (*PpTAS3a/PpTAS6a* and *PpTAS3f/PpTAS6b*) corresponding with the production of phased *DCL4/RDR6* dependent siRNAs. Despite the fact that *PpTAS6a* and *PpTAS6b* do not share sequence similarity in their siRNA generating regions, their common arrangement proximal to *PpTAS3* loci and their shared presence of sliced miR156 sites strongly suggests a common ancestry which we believe justifies their classification into the same family. We have yet to observe siRNA accumulation from *PpTAS6c*, so this annotation should be regarded as provisional.

We used all 248 *PpTAS6a* and *PpTAS6b*-derived siRNAs from our wild type sequencing data as queries for ta-siRNA target predictions using a standard scoring matrix (Allen et al., 2005). 150 target sites, from 129 distinct transcripts, were predicted with alignment scores of 3.5 or less (Appendix: Supplemental Dataset 3.5). At the present time, we have low confidence in these predictions, as the cohort of transcripts did not seem to have a coherent functional theme, and none of these predictions were supported by robust evidence of slicing in the available degradome data. We did identify robust degradome data supporting the slicing of a zinc-finger domain transcript by a *PpTAS6a*-derived siRNA with a higher alignment score (5.5; Figure 3.3B). While this alignment score is relatively high, we note that the siRNA-target mismatches are concentrated on the 3' end of the complementary region and that the degradome data strongly support the *in vivo* accumulation of the predicted slicing remnant. We conclude that at least one *PpTAS6a* ta-siRNA, which we name ta-siZNF, is functional in slicing a zinc-finger domain transcript *in trans*, and that other *trans* targets may exist for other *PpTAS6*-derived siRNAs.

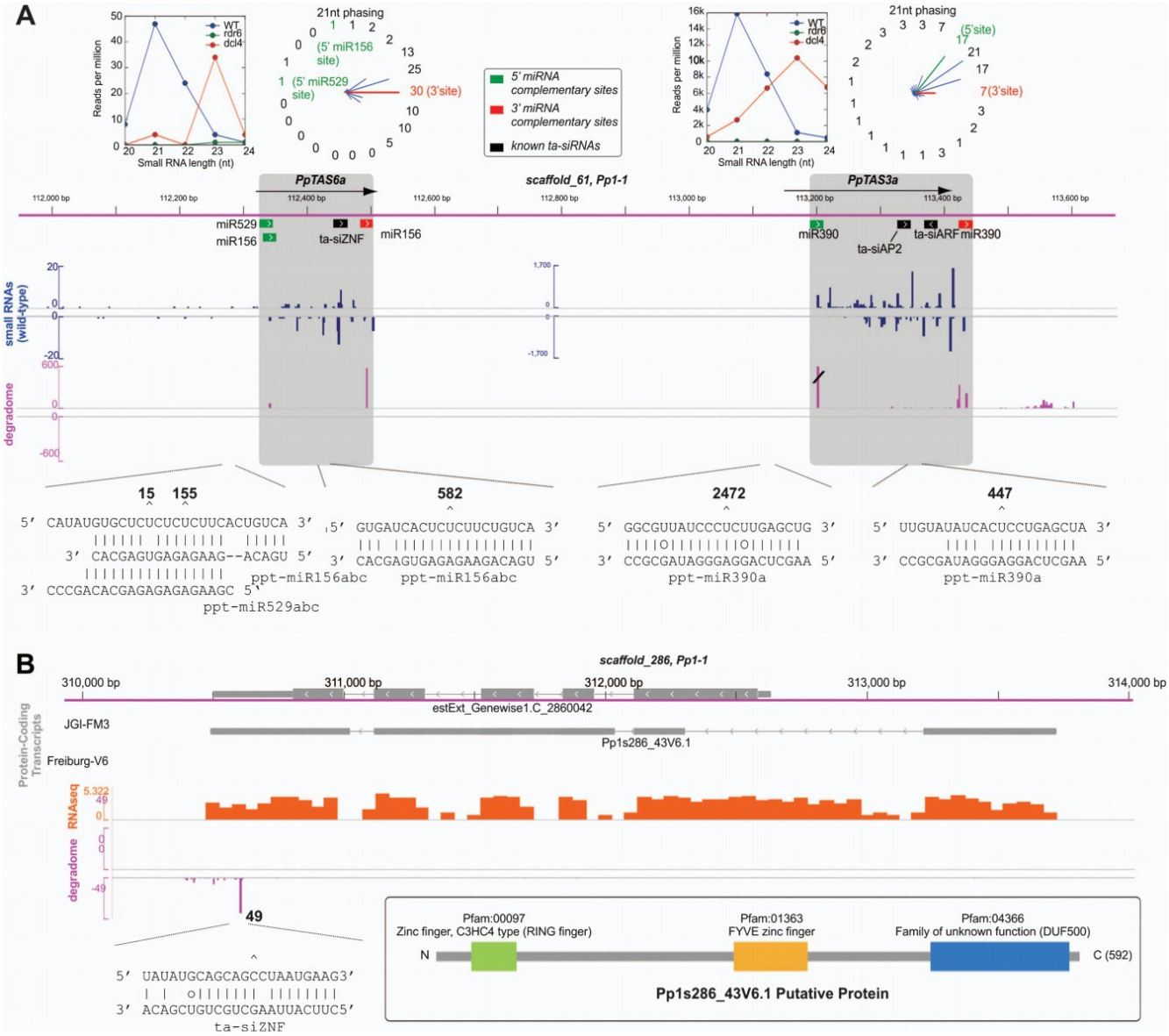


Figure 3.3 Neighboring miR156- and miR390-sliced *TAS* loci.

(A) An annotated genomic snapshot of *PpTAS3a* and *PpTAS6a*. Shaded regions indicate boundaries of indicated *TAS* loci, with locations and strand orientations of miRNA complementary sites and known functional ta-siRNAs shown. Insets above show the small RNA size distribution in the indicated genotypes, as well as the "phasing" distributions of wild type small RNAs in 21nt bins (numbers on periphery indicate percentage of siRNAs in each bin). Browser track small RNA data (blue) and degradome data (magenta) shows the 5' end positions from wild type samples, with positive values

indicating Watson-strand mapped reads, and negative values indicating Crick-strand mapped reads. Insets below show miRNAs aligned with miRNA complementary sites. Numbers above alignments are the number of degradome-derived 5’ ends mapped to the tenth nucleotide of the alignment. (B) Evidence for trans-acting slicing directed by the *PpTAS6a*-derived ta-siZNF. Annotated genomic snapshot showing protein-coding transcripts (gray), polyA+ RNAseq data (not strand-specific; (Zemach et al*.,* 2010)), and degradome data. Inset below-left shows ta-siZNF/target alignment, with the number of degradome reads at the tenth nucleotide of the alignment indicated. Inset below-right shows schematic of the protein (from Phytozome).

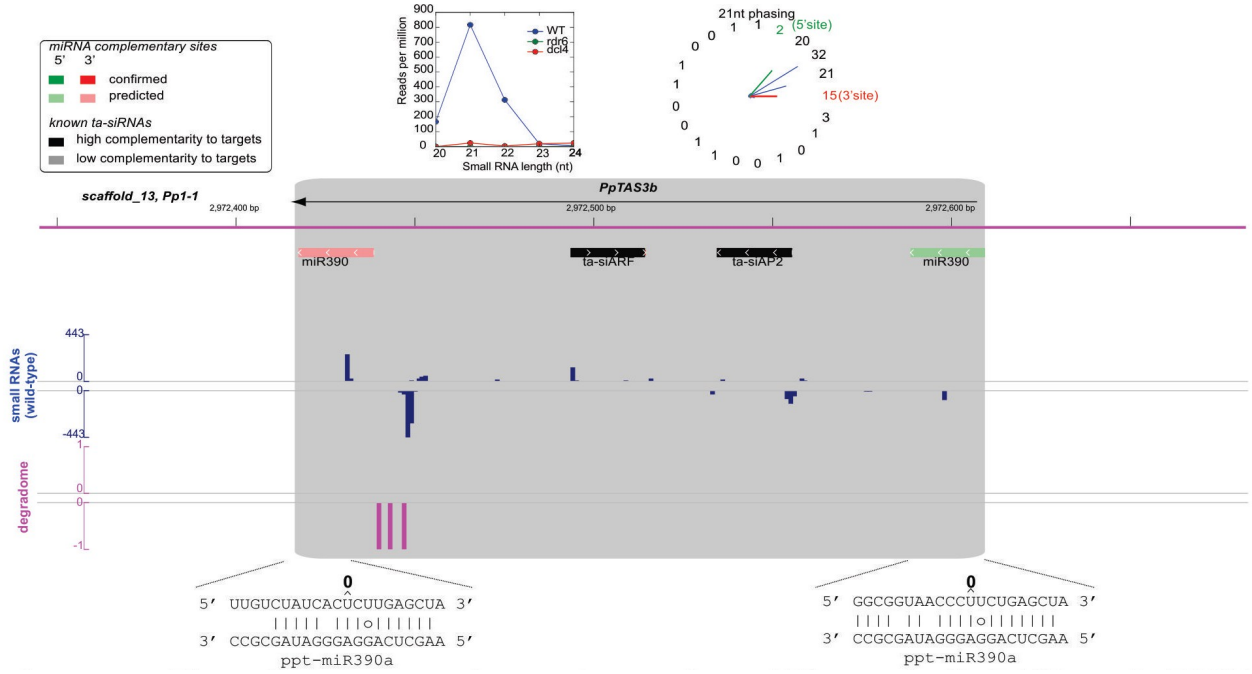


Figure 3.4 Annotated genomic snapshots of *PpTAS3b*.

Shaded regions indicate boundaries of indicated TAS loci, with locations and strand orientations of miRNA complementary sites and known functional tasiRNAs shown. Insets above show the small RNA size distribution in the indicated genotypes, as well as the 'phasing' distributions of wild-type small RNAs in 21nt bins (numbers on periphery indicate percentage of siRNAs in each bin). Browser track small RNA data (blue) and degradome data (magenta) shows the 5' end positions from wild-type samples, with positive values indicating Watson-strand mapped reads, and negative values indicating Crick-strand mapped reads. Insets below show miRNAs aligned with miRNA complementary sites. Asterisks next to degradome data peaks correspond with asterisk-marked positions in the inset miRNA/target alignments.

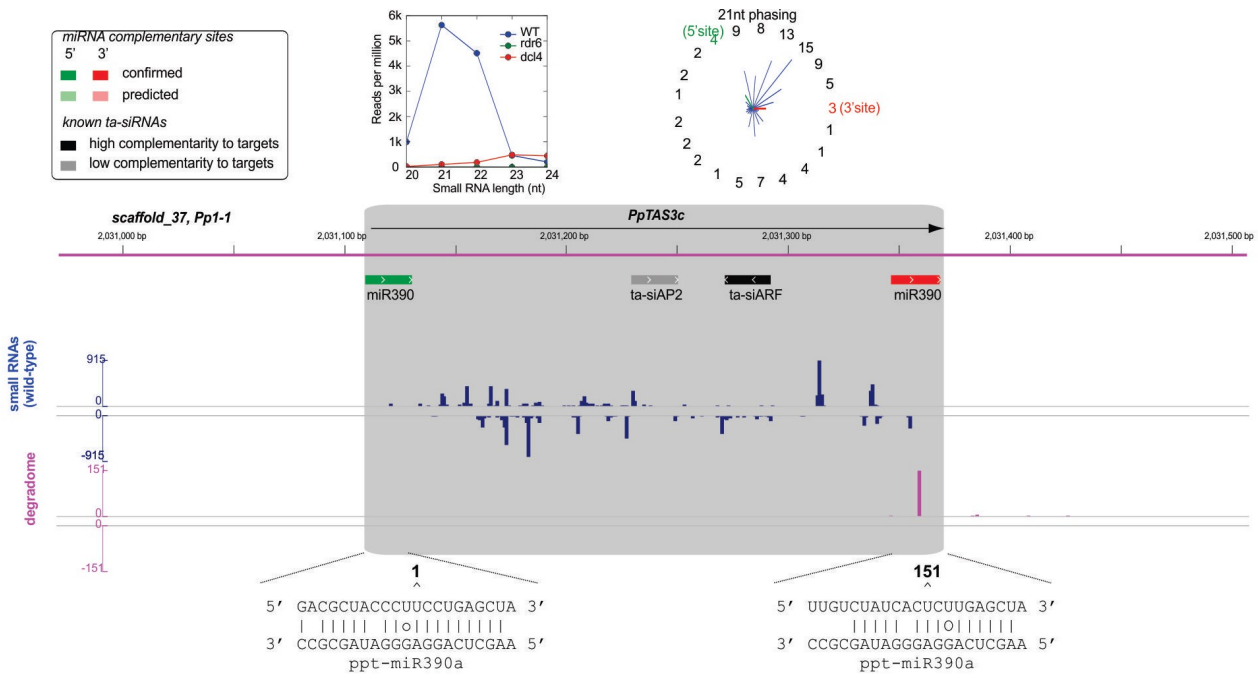


Figure 3.5 Annotated genomic snapshots of *PpTAS3c*.

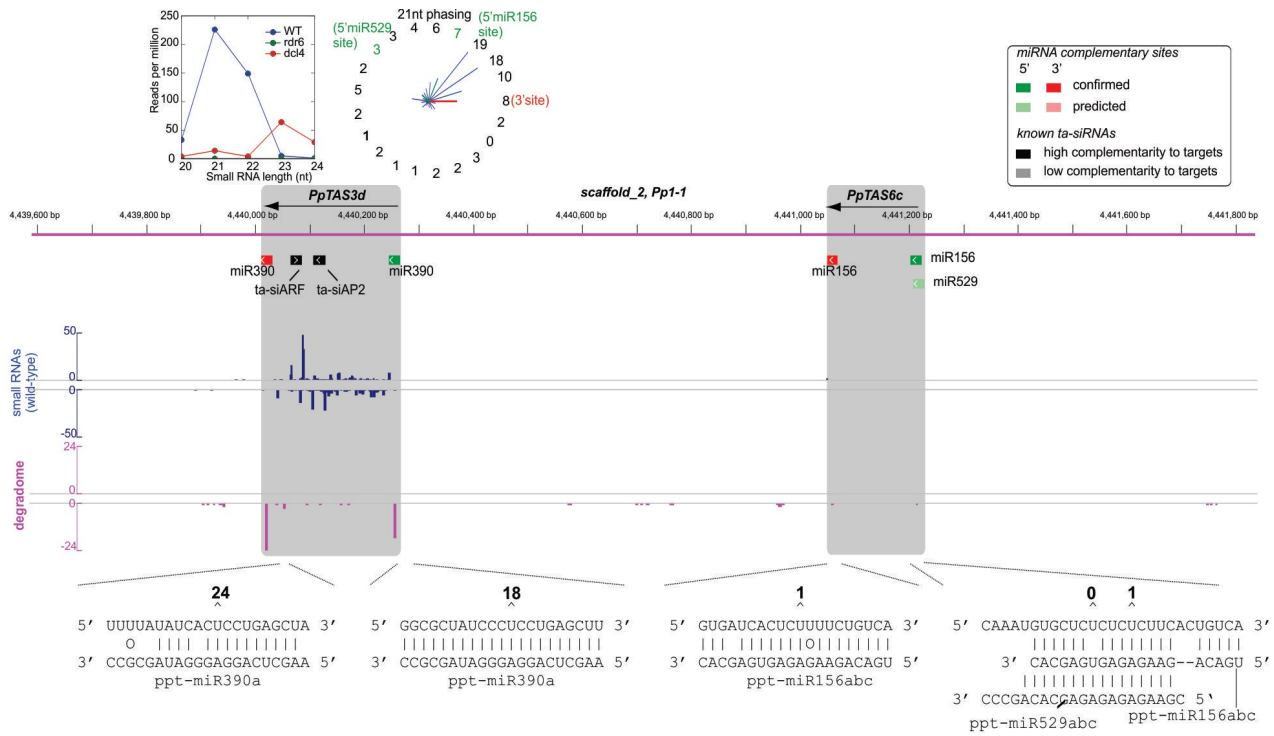


Figure 3.6 Annotated genomic snapshots of *PpTAS3d* / *PpTAS6c*.

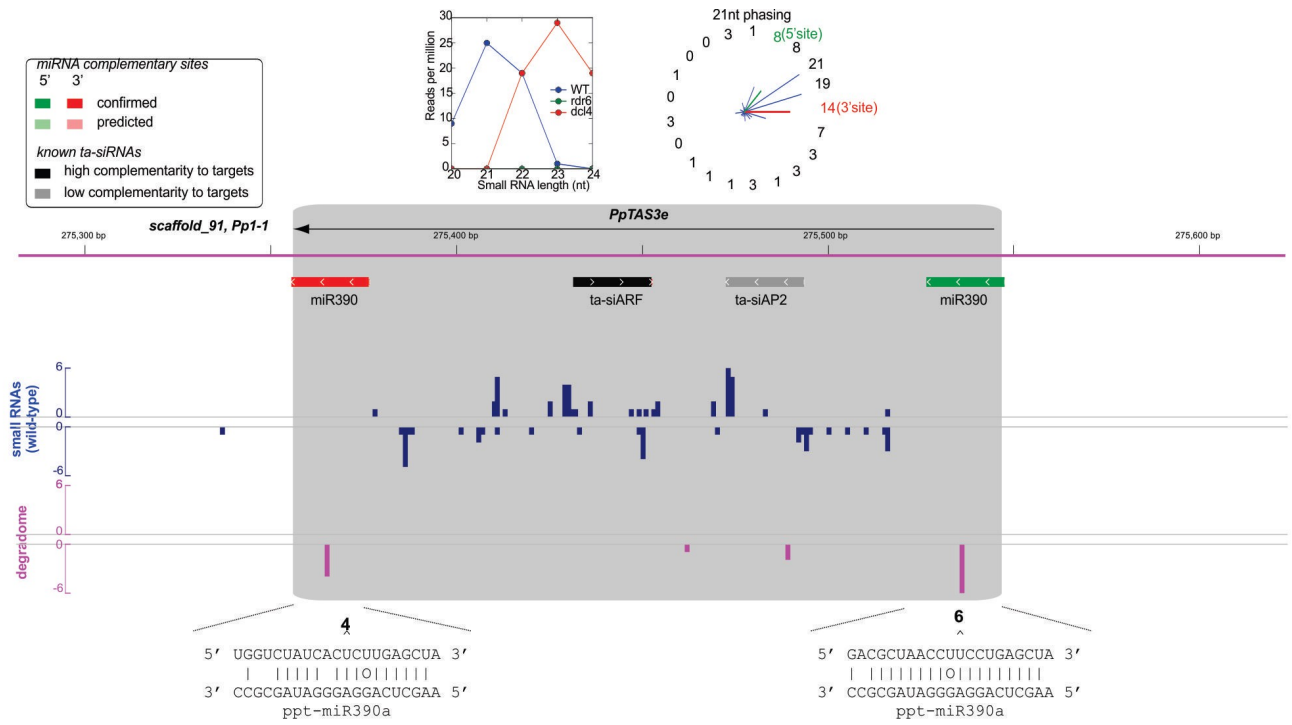


Figure 3.7 Annotated genomic snapshots of *PpTAS3e*.

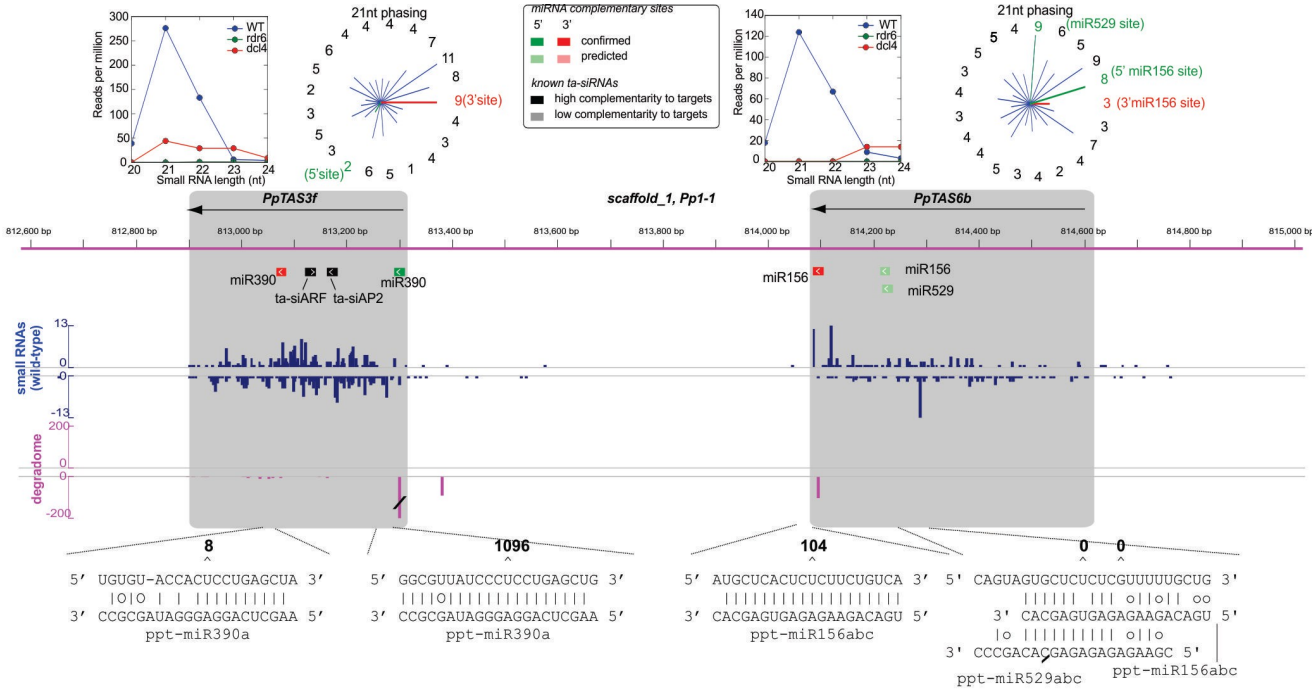


Figure 3.8 Annotated genomic snapshots of *PpTAS3f* / *PpTAS6b*.

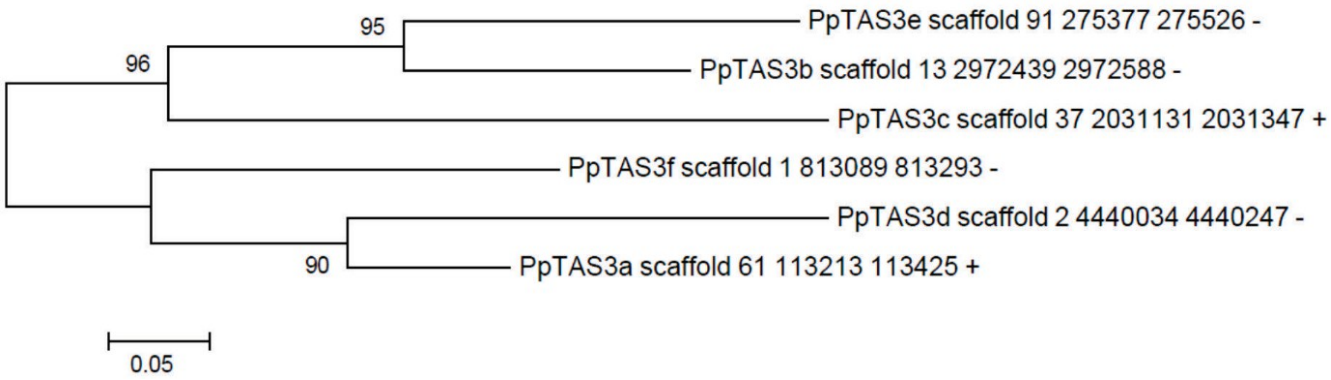


Figure 3.9 A MUSCLE alignment of the regions between the miR390 complementary sites was input to the Maximum Likelihood method in MEGA5 based on the Tamura-Nei model. The bootstrap consensus tree, based on 500 replicates, was shown. Units are substitutions per informative site. Nodes with bootstrap values less than 100% have bootstrap percentages shown.

Table 3.4 Summary of *Physcomitrella DCL4/RDR6*-dependent *TAS* loci.

Locus	5' miR390 site	3' miR390 site	5' miR529 site	5' miR156 site	3' miR156 site	ta-siAP2	ta-siARF	ta-siZNF
PpTAS3a / PpTAS6a	S	S	S	S	S	+	+	+
PpTAS3b	P	P	N/A	N/A	N/A	+	+	N/A
PpTAS3c	S	S	N/A	N/A	N/A	-	+	N/A
PpTAS3d / PpTAS6c	S	S	P	S	S	+	+	absent
PpTAS3e	S	S	N/A	N/A	N/A	-	+	N/A
PpTAS3f / PpTAS6b	S	S	P	P	S	+	+	absent

S: Degradome-based evidence of slicing

P: Predicted miRNA target site w/o degradome evidence

+: ta-siRNA pairs well with one or more target

-: ta-siRNA does not pair well (score >=6) with any targets

N/A: Not applicable

absent: site or ta-siRNA not found within locus

3.4.4 Discussion

In the *Arabidopsis* (Howell et al., 2007b) and *Physcomitrella* (Axtell et al., 2006) genomes only a limited number of endogenous *DCL4/RDR6*-dependent small RNA loci have been identified. In this study, we analyzed of *DCL4/RDR6*-dependent loci in *Physcomitrella* and revealed several interesting findings. All of the *Physcomitrella DCL4/RDR6*-dependent siRNAs appear to derive from primary transcripts sliced at least once, and often several times, by a miRNA. We describe a new family of *TAS* loci that are associated with miR156- and miR529-directed slicing. miR156 is an ancient plant miRNA that has a cohort of conserved, protein-coding targets independent of *TAS* loci. In all plant species examined, including *Physcomitrella*, miR156 targets mRNAs encoding *Squamosa Promoter Binding Like* (*SBP/SPL*) transcription factors (Poethig, 2009). In angiosperms, miR156-mediated regulation of *SBP/SPL* targets is critical for many distinct aspects of developmental timing (Wu and Poethig, 2006; Wu et al., 2009; Nodine and Bartel, 2010). However, miR156-associated *TAS* loci have not been previously described in any plant species. miR529 is related in sequence to miR156, and also widely conserved. At least one *PpTAS6*-derived siRNA appears to be a true ta-siRNA, in that we could identify a sliced zinc-finger domain target. The *PpTAS6* family is also unique in its very close proximity to miR390-targeted *PpTAS3* loci; the three *TAS3/TAS6* pairs are separated by only ~0.7kb, and their miRNA complementary sites are all on the same strands. This suggests that these *PpTAS6*/*PpTAS3* pairs could share single common primary transcripts, and that miR156-, miR529-, and miR390-mediated activities may be inter-related in *Physcomitrella*.

* The work presented in Chapter 3.4 has been published (Arif et al., 2012) and is reproduced with minor modifications.

Chapter 4
A novel targeted genomic enrichment method enables assembly of unknown genomic regions flanking a known core sequence

4.1 Summary

Conserved plant microRNAs (miRNAs) modulate important biological processes but little is known about conserved cis-regulatory elements (CREs) surrounding *MIRNA* genes. We developed a solution-based targeted genomic enrichment methodology to capture, enrich and sequence flanking genomic regions surrounding conserved *MIRNA* genes with a locked-nucleic acid (LNA)-modified, biotinylated probe complementary to the mature miRNA sequence. Genomic DNA bound by the probe is captured by streptavidin-coated magnetic beads, amplified, sequenced and assembled *de novo* to obtain genomic DNA sequences flanking *MIRNA* locus of interest. We demonstrate the effectiveness of this method in *Arabidopsis thaliana*. We demonstrate the sensitivity and specificity of this enrichment methodology to enrich targeted regions spanning 10-20 kb surrounding known *MIR166* and *MIR165* loci. Assembly of the sequencing reads successfully recovered all targeted loci. While further optimization for larger, more complex genomes is needed, this method may enable determination of flanking genomic DNA sequence surrounding a known core (like a conserved mature miRNA) from multiple species that currently don't have a full genome assembly available.

4.2 Introduction

microRNAs (miRNAs) originate from primary transcripts called pri-miRNAs that are transcribed by RNA polymerase II. In plants, after two separate cleavage by the Dicer-like 1 (DCL1) protein, pri-miRNAs are processed into 20-24 nt mature miRNAs and then incorporated into RNA-induced silencing complexes (RISCs) which serve to negatively regulate target mRNAs (Voinnet, 2009). Conserved plant miRNAs modulate important biological processes including development, immune responses, nutrient homeostasis and hormone responses (Axtell and Bowman, 2008; Voinnet, 2009; Cuperus et al., 2011). The spatial and temporal control of miRNA accumulation needs to be fine tuned in order for plants to respond to ever-changing environmental and intracellular signals. This fine-tuning can be done either at the transcriptional level of *MIRNA* genes or the post-transcriptional level. In animals, post-transcriptional regulation of miRNA expression functions either via signaling pathways

centered on the Microprocessor (the protein complex processing pri-miRNAs) or interaction between RNA-binding proteins and cis-regulatory sequences on the terminal loop of miRNA precursors (Newman and Hammond, 2010). However, besides the presence of core promoters and an over-representation of motifs related to development, stress responses, and hormonal control (Xie et al., 2005b; Megraw et al., 2006), the regulation at the transcriptional level of *MIRNA* genes is still mostly unknown.

Control of gene expression is partly conveyed by specific DNA sequences termed cis-acting elements or cis-regulatory elements (CREs) recruiting transcription factors (TFs) or repressors (Barberis et al., 1987; Inostroza, 1992; Lee and Young, 2000). Conserved CREs have been discovered by sequencing multiple species followed by comparative genomics (Ettwiller et al., 2005; Eddy, 2005; Xie et al., 2005a; Stark et al., 2007; Miller et al., 2007). However, even with the advances in next generation sequencing technologies, sequencing and assembling multiple plant genomes is still beyond the resources of a typical lab. If the flanking genomic sequences of interest can be captured specifically in multiple species, identification of CREs need not require complete genome assemblies. To select and enrich the flanking genomic sequences surrounding *MIRNA* genes, we could exploit the fact that conserved *MIRNA*s always have nearly identical sequences in the 20-24 nt mature miRNA region in multiple plant species (Axtell and Bowman, 2008; Cuperus et al., 2011). A methodology which captures long, unknown genomic DNA sequences flanking a short known core sequence, the mature miRNA in this case, could be used to efficiently isolate the flanking DNA of interest from species that lack a reference genome assembly.

The idea of enriching and sequencing specific genomic regions of interest has been widely implemented. Earlier strategies for targeted genomic enrichment include polymerase chain reaction (PCR) (Barnes, 1994), molecular inversion probes (MIPs) (Porreca et al., 2007; Krishnakumar et al., 2008) and microarray capture (Albert et al., 2007; Okou et al., 2007; Hodges et al., 2007; D'Ascenzo et al., 2009; Fu et al., 2010). However, PCR requires the knowledge of two primer sequences flanking the region of interest, thus it is impossible to obtain unknown sequences flanking a single known core sequence. PCR also tends to lack robustness for sequences longer than 10kb (Mamanova et al., 2010). MIPs uses a single-stranded oligonucleotide consisting of a common linker flanked by target-specific sequences to anneal to the target DNA, followed by "gap-filling" between the target-specific sequences with a DNA polymerase, and finally amplifies by PCR with primers directed at the common linker (Porreca et al., 2007; Krishnakumar et al., 2008). MIPs also require two known sequences, and the capture uniformity is relatively poor (Mamanova et al., 2010; Teer et al., 2010). Microarray hybrid capture, using probes against sequences of interest (Albert et al., 2007; Okou et al., 2007; Hodges et al., 2007; D'Ascenzo et al., 2009; Fu et al., 2010), is inefficient for capturing extremely long sequences

flanking a short known sequence (Hodges et al., 2007), and requires a vast excess of samples over probes, which is laborious to obtain. To overcome many of the above shortcomings, solution-based target enrichment methods have been developed, which apply similar principles as microarray-based capture using specific probes designed to the targeted regions of interest. Solution-based target enrichment uses an excess of probes over genomic DNA, which drives the hybridization further to completion with a smaller amount of genomic DNA than microarray-based capture (Mamanova et al., 2010). Also, solution-based capture can be performed in micro-centrifuge tubes or 96-well plates, which is easily scalable compared to microarray capture. To date, the major application of solution-based capture is exon targeting followed by SNP finding (Gnirke et al., 2009; Bamshad et al., 2011). However, the current application of solution-based capture uses long RNA probes of several hundred bases in length to cover the full lengths of exons, and the design of the probes requires a fully sequenced reference genome, or at least the exon sequences of interest.

We developed a novel solution-based targeted enrichment methodology to rapidly capture, enrich and sequence a large, unknown genomic region flanking a small known target of interest. In this study, we tested the strategy with a 21 nt probe against the miR166 mature sequence in *Arabidopsis thaliana*, and found that this methodology was highly specific and sensitive to enrich regions flanking the targeted loci. *de novo* assembly of the reads sequenced from the enriched sample successfully assembled all targeted loci into long contigs. We propose that the successful development of this method may enable us to easily obtain flanking genomic DNA surrounding short conserved regions (like mature miRNAs) in multiple plant taxa that lack complete genome assemblies, and in turn accelerate discovery of CREs surrounding such loci.

4.3 Methods

4.3.1 Targeted genomic enrichment experiment in *Arabidopsis*

Genomic DNA was extracted from wild-type *Arabidopsis thaliana* Col-0 leaves using Nucleon PhytoPure Genomic DNA Extraction Kits (GE Healthcare). 100 ug genomic DNA (100 ul @ 1 ug/ 1ul), 300 ul Hybridization Buffer P5 (Invitrogen) and 1 pmole LNA-biotinylated capture probe (1 ul @ 1uM) were placed in a 1.7 ml centrifuge tube and boiled for 5 minutes to denature the genomic DNA. The mix was placed in 45C for 30 minutes for hybridization. 20 ul streptavidin beads from the RiboMinus Plant Kit (Invitrogen) were prepared per the manufacturer's protocol. After 30 minutes of hybridization, the hybridization mix was added to the beads and incubated at 45C for 15 minutes with occasional (every 2-3 minutes) gentle mixing by inversion. Beads were captured with a magnetic stand and were washed 3 times each for 2 minutes with 500 ul 0.1X SSC incubated at 45C. Captured DNA was eluted

for one minute with 500 ul nanopure water at 90C twice. DNA was mixed with 1/10 volume 3M sodium acetate pH 5.2, 20 ug glycogen (1 ul @ 20 ug/ul) and 3 volumes of 95% ethanol, vortexed for 30 seconds, and then placed at -20C overnight for ethanol precipitation. DNA was centrifuged at maximum speed at 4C for 20 minutes to spin down pellets. Pellets were washed by 75% ethanol and centrifuged at maximum speed at 4C for 5 minutes and air dried at 4C. DNA was then resuspended in a minimal volume (4-8 ul) of nanopure water. DNA was linearly amplified with the illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare) per the manufacturer's protocol, with the only modification that the incubation time was increased from 1.5 hours to 2 hours.

4.3.2 Quantitative real-time PCR and data analysis

Real-time PCR was performed using a QuantiTect SYBR Green PCR kit (Qiagen) on a StepOne Real-Time PCR System (Applied Biosystems). Primers were designed to amplify regions with varying distances from one of the targeted loci, *MIR166a*. *Actin1* was used as a control, as *Actin1* is far from any of the targeted loci, therefore should not be enriched. For each primer set, two samples of captured and then linear-amplified DNA (captured DNA for short), as well as two samples of the diluted original genomic DNA were loaded. At the same time, serial dilution of the extracted genomic DNA was used. The method to calculate the normalized fold of enrichment is as follows. First, Ct values from the serial dilution experiments were used to calculate the linear relationship between Ct and log(Dilution) as Ct = A*log(Dilution) + b. Second, average Cts were taken for captured DNA and genomic DNA samples respectively, and the “pseudo dilution” values were calculated from the average Cts, A and b. Dividing the “pseudo dilution” value of the captured DNA by that of the genomic DNA resulted in the relative concentration of the specific targeted region in the captured DNA sample. Finally, the relative concentration of the targeted region was normalized by that of *Actin1* to calculate the normalized fold of enrichment.

4.3.3 Paired-end sequencing and reference-based mapping and analysis

An *Arabidopsis* sample prepared with the targeted genomic enrichment methodology was fragmented into ~ 400 bp and paired-end sequenced on an Illumina GAIIx sequencer. Paired-end reads were mapped to the *Arabidopsis thaliana* reference genome (TAIR10) using Bowtie 0.12.7 (Langmead et al., 2009) with parameters “-v2 -X500”. Non-uniquely mapped reads (12.93% of all mapped reads) were identified and one mapped location was randomly kept. Mapped reads were assigned to 1kb-sized bins of the nuclear genome based on the midpoint of the mapping positions. Reads mapping to the chloroplast or mitochondria were discarded for the following analysis. Read coverage for each bin was defined simply as the number of reads assigned to that bin. Normalized coverage for each bin was simply the read coverage of that bin divided by the nuclear genome

average of the read coverage per bin. Enrichment is implied when the normalized coverage is above 1. Threshold of normalized coverage by which a bin was considered “enriched” was determined by performance analysis, and a normalized coverage of 10 was chosen by balancing sensitivity and specificity. Enriched bins were merged if within 10 bins apart and extended 10 bins to each side to define the surrounding regions of enriched bins. The Pearson correlation coefficient r was calculated to examine the linear dependence between |x| and log(y) where x is the distance to the most highly enriched bin in the 21 kb region centered on that bin and y is the normalized coverage. To assess the tolerance of mismatches between the probe and potential targets, the reference genome was scanned to identify sequences with different mismatch patterns. Then the normalized coverage of the bin where the sequence fell into was used to evaluate the effect mismatches had on enrichment, assuming the sequence was responsible for the enrichment.

4.3.4 *de novo* assembly of the paired-end reads and assembly quality evaluation

Random samples from all the paired-end sequenced reads were generated by accepting each pair of reads at a given probability. For example, to generate 1% of the total reads, the acceptance probability is 0.01. Sampled reads were then *de novo* assembled with the Velvet assembler (Zerbino and Birney, 2008). Parameters used for velveth were “31 -shortPaired -fastq” and parameters for velvetg were “exp_cov 20 ins_length 400 ins_length_sd 100”. However, we observed that changing “exp_cov” to “40” did not affect the assembly result. Assembled contigs were searched for complementary sequences to miR166 with BLASTn. All contigs harboring a miR166 matching sequence, together with all contigs long than 1000 bp, were BLASTed against the reference genome to identify the origin. Assembly quality of contigs from targeted loci were evaluated by first generating global alignment between the contig and corresponding sequence in the reference genome using EMBOSS needle (Rice et al., 2000) and then counting the number of mismatches, short gaps (defined as indels <= 5 bp long) and long gaps (defined as indels > 5 bp long).

4.4 Results

4.4.1 Enrichment of an ~20 kb region flanking *Arabidopsis MIR166a*

The enrichment methodology is outlined as follows (Figure 4.1A): Genomic DNA is hybridized with a biotinylated locked nucleic acid (LNA)-modified capture probe. Targeted genomic fragments paired with the probe are retained by binding to paramagnetic, streptavidin coated-beads while unbound fragments are washed away. Then the targeted fragments are eluted in hot water, subject to linear amplification by the DNA polymerase Φ29 and subsequently fragmented, sequenced and

assembled.

A pilot enrichment experiment was performed with *Arabidopsis* genomic DNA and a 21 nt, biotinylated LNA capture probe complementary to the mature miR166 DNA sequence. The relative fold of enrichment of the targeted loci compared to a control region was determined with quantitative real-time PCR (qPCR) performed on the enriched and Φ29-amplified DNA. The pilot experiment successfully yielded enrichment in a region of ~20 kb flanking the *MIR166a* locus, with a peak enrichment above 1,000-fold (Figure 4.1B).

To optimize the enrichment protocol to increase final DNA yield, input genomic DNA concentrations, washing conditions, and linear amplification times were varied, and relative fold of enrichment at *MIR166a* was determined by qPCR. The optimized protocol is described in Methods. We find that elongating ϕ29 amplification time to 2 hours or more increases the final quantity of DNA without affecting enrichment (Supplementary Figure 4.S1) and increasing the input amount of genomic DNA in the hybridization step of the targeted genomic enrichment by two-fold increases the final yield of enriched DNA product by four-fold (Figure 4.1D) without lowering enrichment (Figure 4.1C), while using an even larger amount of the input DNA does not further increase the total DNA yield (Figure 4.1D). Overall, with 100 ug genomic DNA input in the targeted enrichment followed by 2-hour ϕ29 amplification results in over 10 ug enriched DNA, enough for a high-throughput sequencing run which typically requires ~1 ug DNA.

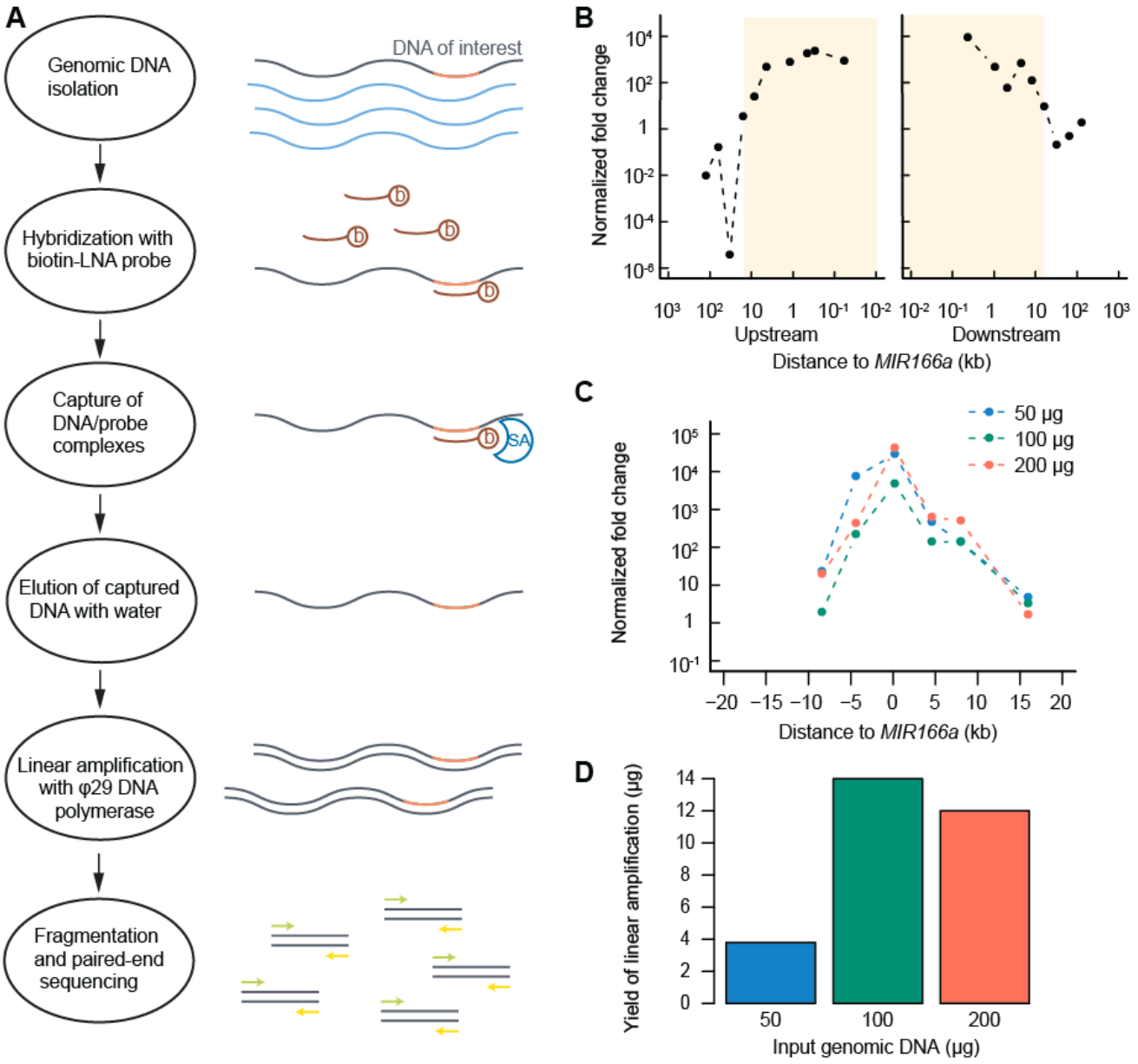


Figure 4.1 Pilot targeted enrichment experiment in *Arabidopsis* shows enrichment near a targeted locus. (A) Schematic overview of targeted genome enrichment method. b: Biotin, SA: Streptavidin. **(**B) Quantitative real-time PCR (qPCR) of enriched DNA with designed primers surrounding the *MIR166a* locus. Normalized fold change relative to Act1 (as a control) after enrichment is shown. Shaded box indicates the region with a normalized fold change above 1. (C) Amount of input genomic DNA (gDNA) does not affect the fold of the enrichment. Normalized fold change relative to Act1 after enrichment is shown with varying amount of gDNA. (D) Amount of gDNA affects the yield of the enrichment. Yield after enrichment is shown, as is measured by Qubit® Fluorometer.

4.4.2 Successful enrichment at all *MIR166* and *MIR165* loci

An *Arabidopsis* genomic DNA sample prepared with the optimized targeted enrichment protocol was fragmented into ~ 400 bp and sequenced on one lane of an Illumina GAIIx sequencer. The goal of sequencing the enriched sample was two fold: first, the sequencing reads were mapped back to the reference genome to evaluate the performance of the targeted enrichment methodology; second, the reads were *de novo* assembled with the Velvet assembly software (Zerbino and Birney, 2008) and parameters of the assembler tuned to optimize assembly quality.

We obtained ~25 million pairs of 76-nucleotide (nt) paired-end reads, of which ~18 were mapped to the *Arabidopsis* genome (Table 4.1). 65.1% of the mapped reads mapped to the nuclear genome, 32.4% to the plastid genome and 2.5% to the mitochondrial genome.

Mapped reads were then tallied into 1kb-sized bins and read coverage of each bin was calculated. The average read coverage per bin for the nuclear genome is 98 reads, while the average coverage per bin is 36,433 for chloroplast and 748 for mitochondria. The deep coverage of the organellar genomes is expected based on their high copy numbers relative to the nuclear genome and their small sizes. To achieve the first goal of evaluating the enrichment methodology, bins from organellar genomes were discarded, keeping only bins in the nuclear genome. Coverage of each bin was normalized to the nuclear genome average (termed normalized coverage). Intuitively, enrichment is implied when the normalized coverage is above 1. There are seven *MIR166* loci with perfect matches to the probe, and two *MIR165* loci with a single mismatch to the probe (miR165 and miR166 are highly similar miRNA families; Figure 4.2A). Enrichment was observed in a ~10 kb region flanking all targeted loci (Figure 4.2B). A peak enrichment of 100-fold or more was evident for the seven *MIR166* loci in the genome with full complementarity to the capture probe, while a slightly lower peak of enrichment was evident for both *MIR165* loci in the genome which have one mismatch to the probe (Figure 4.2A-B). As a control, three *MIR164* loci which have no complementarity to the probe were analyzed and indeed showed no evidence of enrichment (Figure 4.2C-D).

In order to estimate the size of the enriched regions, Student's t tests are performed to test the hypothesis that the mean normalized coverage of bins with increasing distances from one target site is not different from 1. Normalized coverages of bins that are within 9 kb from any one of the target sites are different from 1 with statistical significance ($p < 0.05$), indicating that the size of the enriched regions is about 19 kb on average (totaling 135 kb for the eight targeted loci, *MIR166c* and *MIR166d* considered as a single locus as they are just two bins apart). Bins that are within 5 kb of the targets have an mean normalized coverage different from 1 with $p < 0.01$, corresponding to a size of 11 kb significantly enriched regions (totaling 79 bins for 8 targeted loci; Figure 4.2E).

Table 4.1 Genome mapping result of sequencing reads.

Genome	Number of mapped reads	Percentage out of all mapped reads	Percentage out of all reads (total: 24,859,558)
All	17,548,544	100	70.59
Nuclear genome	11,426,625	65.11	45.97
Plastid genome	5,676,213	32.35	22.83
Mitochondrial genome	445,706	2.54	1.79

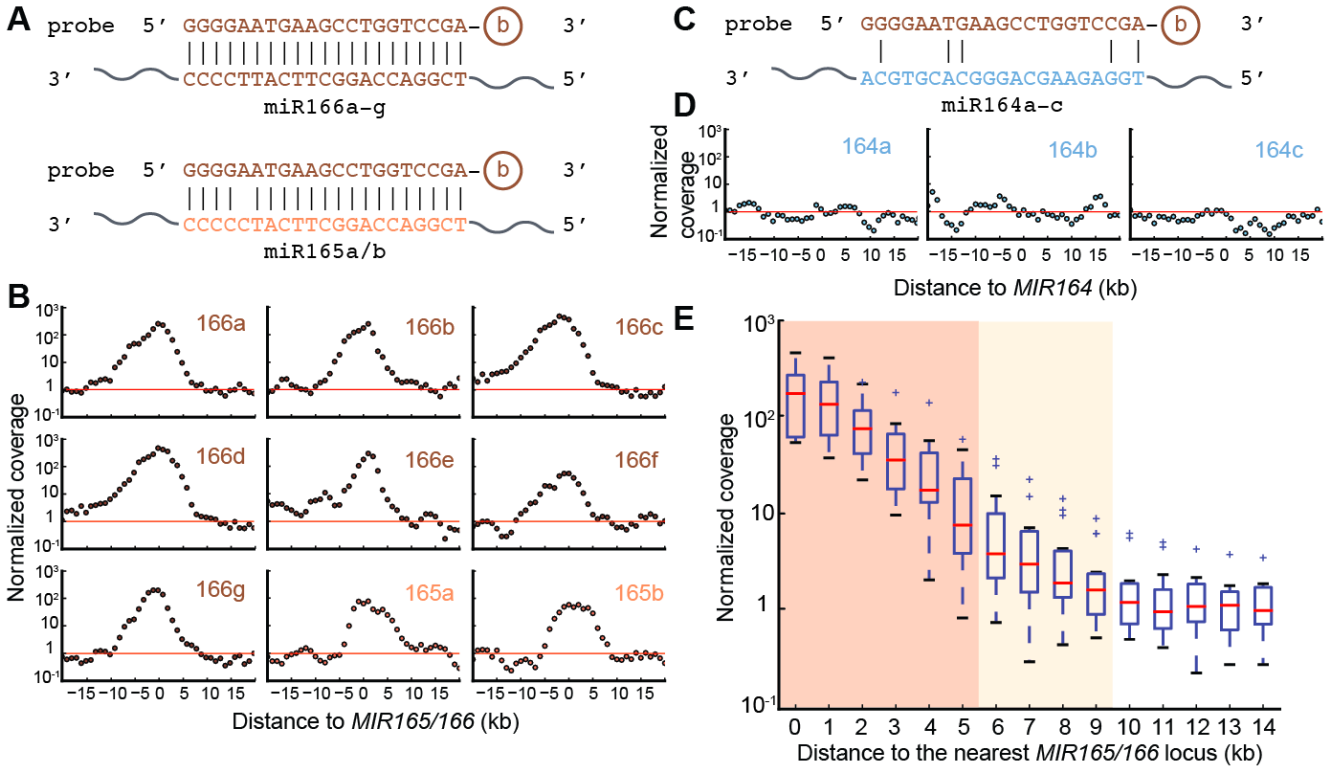


Figure 4.2 Enrichment in a 10 kb region flanking the targeted *MIRNA* loci. (A) Sequence alignments between capture probe and miR166/miR165 respectively. b: Biotin. (B) Normalized coverage at each 1kb-sized bin flanking the indicated *MIRNA* loci. Red horizontal line indicates the genome average of the normalized coverage, which equals 1. (C) As in (A) for miR164, which is not targeted by the probe. (D) As in (B) for miR164, which is not targeted by the probe. (E) Regions of +/- 9 kb flanking the target sites are enriched. Box plot shows fold of enrichment of bins with increasing distance to the target sites. This is a tallied view of the nine individual targeted loci shown in (B). Dark shade denotes $p < 0.01$ with Student's t test against a normalized coverage of 1. Light shade denotes $p < 0.05$.

4.4.3 Enrichment is both sensitive and specific

Next, the enrichment pattern was assessed across the genome, focusing on all “enriched” regions regardless of whether or not they were *MIR166* or *MIR165* loci. In order to determine the threshold of normalized coverage above which a bin is defined as “enriched”, the sensitivity and specificity of the enriched bins at different thresholds were evaluated. The 79 bins within 5 kb away from any target sites are assumed as positives. All other bins (totaling 119,070 bins) in the nuclear genome are considered negatives. Thus, a true positive is defined as a bin above the threshold of normalized coverage and within 5 kb from any target sites, while a false positive is a bin above the threshold but outside the +/- 5 kb window. A true negative is a bin below the threshold and are outside the +/- 5kb region, while a false negative is a bin within the +/- 5 kb region but below the threshold. By decreasing the threshold of normalized coverage of each bin, sensitivity is increased while specificity is decreased as expected (Table 4.2, Figure 4.3). The same analysis was performed with the 135 within-9kb bins as true positives as well (Table 4.2, Figure 4.3). Sensitivity of the latter is not as high as the former at each threshold of normalized coverage, which is partly because the set of within-9kb bins is less stringent (although the mean enrichment of all these bins is statistically significant, many bins in this set are in fact not enriched). We chose a normalized coverage of ten as the threshold of enrichment for further analysis, which maintained both sensitivity and specificity at relatively high levels. It is worth noting that at the chosen threshold, the false discovery rate is quite high (153/221), however, false positives (i.e. enriched regions not close to the targeted loci) are not a major concern for downstream analysis, because false positives, when later assembled into contigs, will lack the sequence targeted by the probe (i.e. mature miR165/166), as is shown later in Figure 4.6.

Table 4.2 Performance analysis of varying threshold of normalized coverage to determine enriched regions.

Within 5kb from target site as true positives

Threshold (normalized coverage)	True positive (TP)	False negative (FN)	False positive (FP)	True negative (TN)	Sensitivity (TP/ (TP+FN))	Specificity (TN/ (TN+FP))
100	22	57	7	119,063	0.2785	0.9999
50	36	43	17	119,053	0.4557	0.9999
20	56	23	61	119,009	0.7089	0.9995
10*	68	11	153	118,917	0.8608	0.9987
5	72	7	636	118,434	0.9114	0.9947
2	76	3	8,104	110,966	0.9620	0.9319
1	78	1	37,556	81,514	0.9873	0.6846
0.5	79	0	83,131	35,939	1.0000	0.3018
0.2	79	0	112,713	6,357	1.0000	0.0534

* Threshold chosen for further analysis.

Within 9kb from target site as true positives

Threshold (normalized coverage)	True positive (TP)	False negative (FN)	False positive (FP)	True negative (TN)	Sensitivity (TP/ (TP+FN))	Specificity (TN/ (TN+FP))
100	22	113	7	119,007	0.1630	0.9999
50	36	99	17	118,997	0.2667	0.9999
20	59	76	58	118,956	0.4370	0.9995
10*	76	59	145	118,869	0.5630	0.9988
5	89	46	619	118,395	0.6593	0.9948
2	107	28	8,073	110,941	0.7926	0.9322
1	124	11	37,510	81,504	0.9185	0.6848
1	132	3	83,078	35,936	0.9778	0.3019
0	135	0	112,657	6,357	1.0000	0.0534

* Threshold chosen for further analysis.

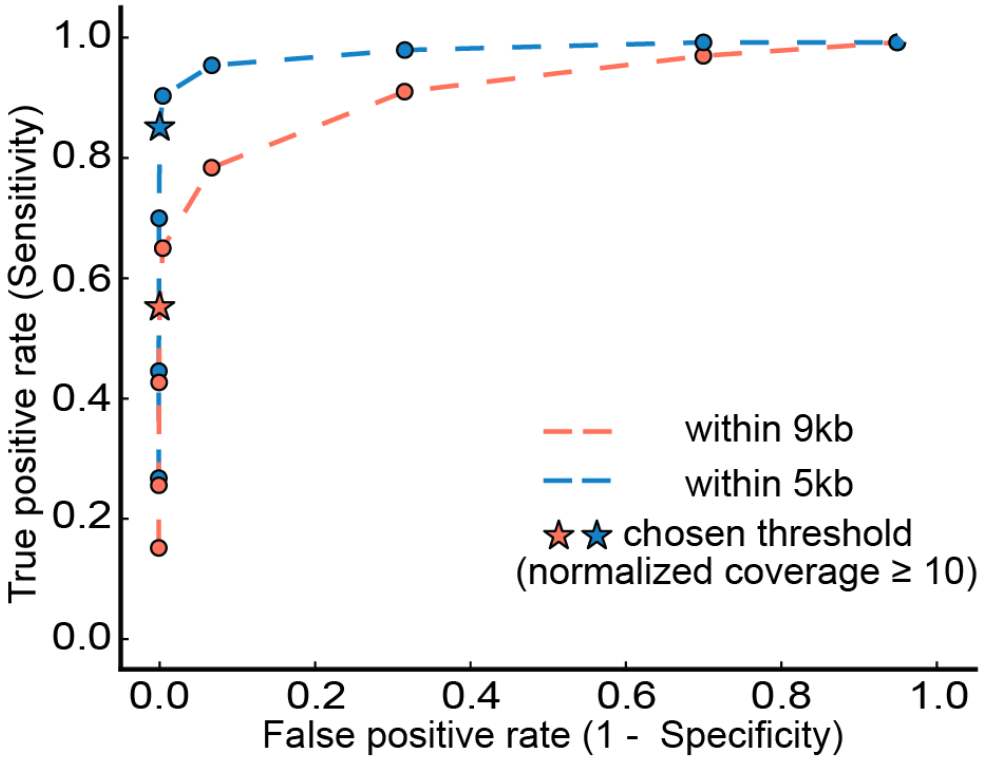


Figure 4.3 Performance analysis to determine enriched regions. Receiver operating characteristic (ROC) curves are shown with varying thresholds of normalized fold change, using within-9kb or within-5kb bins from target sites, respectively, as positives.

4.4.4 Targeted regions can be discriminated from sporadically enriched loci

In order to examine the pattern of enriched genomic regions, bins with a normalized coverage above 10 were merged if within 10 kb apart, and extended 10 kb on each side to examine the genomic landscape surrounding the enriched regions. After merging and extending, a total of 64 highly enriched regions were generated (Supplementary Figure 4.S2), including all eight *MIR165/166* loci (*MIR166c* and *MIR166d* are closely linked on chromosome five, and as such are merged into a single locus in this analysis). When observing the landscape of adjacent bins centered on a highly enriched bin, *MIR165/166* flanking regions all exhibit a bell shape, reflecting lower enrichment further away from the probe binding site (Figure 4.4A-B, Supplementary Figure 4.S2, shaded panels), while other enriched regions generally show only one or two highly enriched bins flanked by regions with a coverage close to the background level, likely due to random amplification during sequencing or unannotated copy number variation relative to the reference genome assembly (Figure 4.4C, Supplementary Figure 4.S2, unshaded panels). In order to distinguish targeted regions from non-targeted regions based on the enrichment pattern in the surrounding regions of highly enriched bins, the Pearson correlation coefficient r was calculated to examine the linear dependence between |x| and log(y) where x is the distance to the most highly enriched bin in the 21 kb region centered on that bin and y is the normalized coverage (Figure 4.4E, Supplementary Table 4.S1). The hypothesis is that if the region is centered on a real target site, enrichment should decrease exponentially as it moves further away from the target site. On the other hand, if the region is not targeted, no such correlation should be observed. To test this hypothesis, sensitivity and specificity was assessed with varying thresholds of r as the classifier of targeted and non-targeted regions (Figure 4.4F). As expected, sensitivity increases while specificity decreases as the threshold of r increases (i.e. becomes less negative, indicating a weaker linear relationship). We chose $r < -0.9$ as the threshold to distinguish non-targeted from targeted regions. With this threshold, seven out of the eight enriched regions flanking *MIR165/166* loci are recovered (Figure 4.4E-F, Supplementary Figure 4.S2), the only exception being *MIR166e* locus (Figure 4.4B), possibly due to the secondary non-specific peak near the targeted locus confounding the linear dependence pattern. All other regions have $r > -0.90$ (a typical example is shown in Figure 4C) except one: enriched locus 6 with genome coordinates chr1: 10314k-10344k (Figure 4.4D, Supplementary Table 4.S1). Overall, a Pearson correlation test with threshold of $r < -0.90$ results in a sensitivity of 7/8 and specificity of 55/56, which is a sensitive and specific classifier of targeted and non-targeted loci. The above analysis demonstrates that the targeted enrichment methodology is highly specific to enrich a relatively long region flanking the targeted loci.

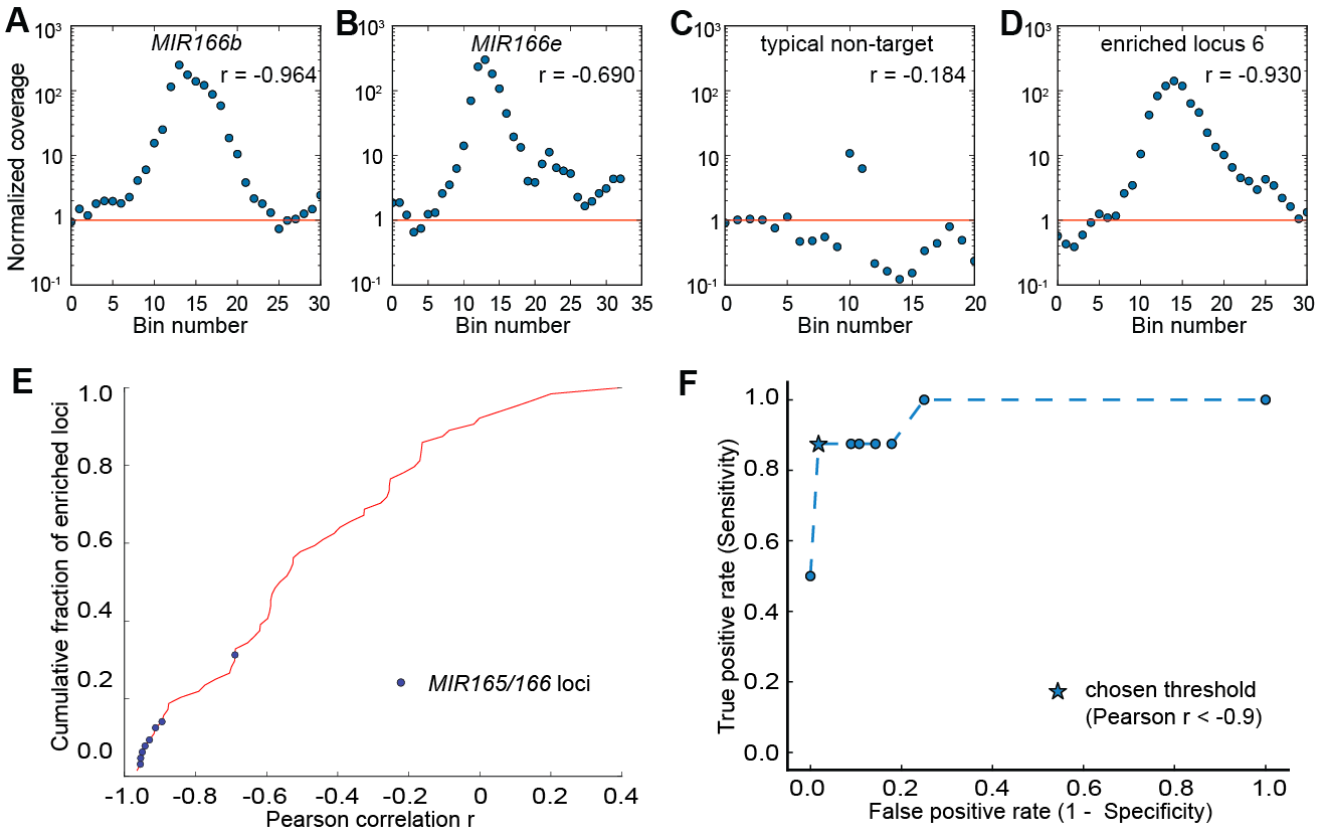


Figure 4.4 Targeted regions have a distinctive enrichment pattern. (A-D) Each panel shows the normalized coverage at each 1kb-sized bin centered on a highly enriched bin. Pearson correlation r of |x| and log(y) is shown, where x is the distance to the most highly enriched bin in the region and y is the normalized coverage. Red line indicates the genome average of the normalized coverage, which equals 1. See Supplementary Figure 4.S2 for full details. (A) A typical region surrounding a *MIR165/166* targeted locus (*MIR166b*). (B) Region surrounding *MIR166e* targeted locus. (C) A typical region surrounding a non-targeted locus. (D) Region surrounding enriched locus 6, which is not a *MIR166* or *MIR165* locus. (E) Cumulative distribution of the Pearson correlation r for all 64 highly enriched regions. Blue dots indicate targeted *MIR165/166* loci. (F) Performance analysis to determine the optimized threshold of r to classify targeted and non-targeted regions. ROC curve is shown with varying threshold of r. Star-shaped dot indicates the chosen threshold of r = -0.9.

4.4.5 Enrichment requires a high amount of probe complementarity

We next analyzed how mismatches between potential targets and the probe affect enrichment. As slight sequence variation exists even for deeply conserved plant miRNAs, it is important to know how much sequence variation in the targeted sites can be tolerated. Therefore, normalized coverage at genomic loci with zero to five mismatches to the capture probe was examined, disallowing insertions or deletions (indels). All the loci with zero or one mismatches are *MIR165* or *MIR166* loci, and Student's t test revealed that the mean normalized coverage of loci with perfect complementarity and with one mismatch were both significantly different from the null hypothesis of 1 with p-values < 0.01 and < 0.05, respectively (Figure 4.5A). No locus in the genome had exactly two mismatches to our probe. Genomic loci with three, four or five mismatches to the probe showed no enrichment, as the normalized coverage was not statistically different from the genome average. None of the 56 false-positive enriched loci (Supplementary Figure 4.S2) had potential probe complementarity sites with between zero and four mismatches, emphasizing that the reasons for sporadically enriched loci are likely not due to probe hybridization. This demonstrates that our strategy is generally specific to loci with zero, one, and perhaps two mismatches to the probe.

We next examined in closer detail enriched locus 6, which was the sole enriched locus that showed a robust bell curve of enrichment despite not being a *MIR166* or *MIR165* locus (Figure 4.4D). Enriched locus 6 resides in the intergenic region between *AT1G29540.1* (unknown protein) and *AT1G29550.1* (eukaryotic initiation factor 4E protein). This enriched locus has no sequence similarity to the *MIR165/166* flanking regions (+/- 5kb), nor does it exhibit similarity to rRNA sequences, thus ruling out simple explanations for its enrichment. We did identify a rather poor complementary site with a 5' A-A mismatch, and a central 2 nt bulge (Figure 4.5B). However, this is unlikely to be responsible for the enrichment of locus 6: Out of the six genomic loci which had 1 or 2 mismatches and 1 or 2 indels to the probe, enriched locus 6 was the only one with significant enrichment (Figure 4.5A). Our *de novo* sequencing confirmed the sequence at this site was identical to the reference genome, ruling out the possibility of an un-annotated indel that created a perfect probe complementarity site. We currently do not understand the reason why this locus was enriched. However, it is the single exception to the general rule that robust enrichment requires high complementarity to the probe.

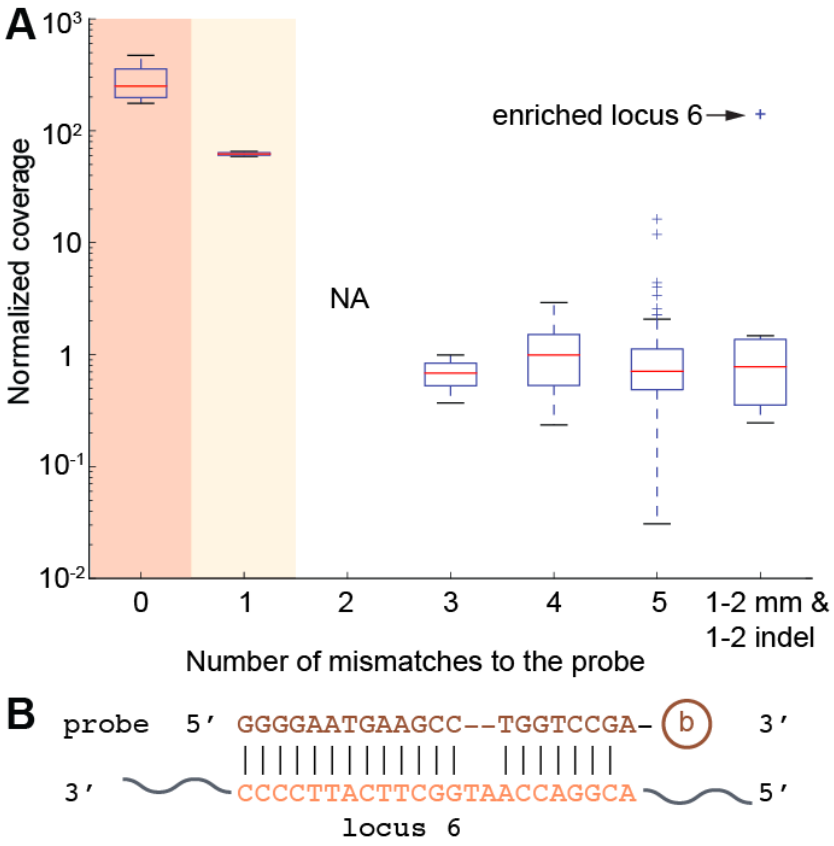


Figure 4.5 Enrichment is highly specific for loci with zero or one mismatch. (A) Box plot shows normalized coverage of loci with different mismatches to the probe. Last box shows genomic loci which are similar to locus 6, with 1 or 2 mismatches and 1 or 2 insertions and deletions in the alignment to the probe. Dark shade denotes $p < 0.01$ with Student's t test against a normalized coverage of 1. Light shade denotes $p < 0.05$. (B) Sequence alignment between capture probe and enriched locus 6.

4.4.6 *de novo* assembly accurately recovers genomic sequences flanking targeted loci

Reads were *de novo* assembled with the Velvet assembler (Zerbino and Birney, 2008), in order to test the feasibility to recover flanking sequences of the targeted loci in the absense of a reference sequence. Assembly proceeded using 1% of the total paired-end reads, which were randomly selected. All contigs greater than 1 kb in length and having sequence complementary to the capture probe (identified by BLASTn against the miR166 sequence) were indeed *MIR165/166* flanking regions (identified by BLASTn against the genome) (Figure 4.6). Seven out of the eight *MIR165/166* loci are recovered in the assembled contigs, missing only *MIR166c/MIR166d*. This is likely due to the fact that *MIR166c/d* locus has the highest enrichment among all targeted loci, resulted from an additive effect of two target sites (Supplementary Figure 4.S2, 4th panel). We hypothesized that different coverage may affect the the assembly result. Therefore we varied the number of reads fed into Velvet from 0.25% to 4% of the total reads (approximately 62k to 994k reads), resulting in a coverage per nt ranging from 5 to 80 at the assembled contigs, as was estimated by Velvet (Table 4.3). Indeed, the number of *MIR165/166* loci recovered in the assembled contigs changes with varying read coverage. Specifically, at the lower extreme of 5 reads per nt, the two *MIR165* loci, whose enrichment level were the lowest among all targets due to one mismatch to the capture probe, are missing in the assembled contigs. At the upper extreme of 80 reads per nt, none of the targeted loci are recovered, likely because at such a high coverage, the enriched loci were treated as repetitive regions by Velvet (Martin and Wang, 2011). At the intermediate coverage levels, for example, 10 reads per nt, all targets but *MIR165b* (lowest enrichment, Supplementary Figure 4.S2, 2nd panel) are recovered (Table 4.4). At 20 reads per nt, all but *MIR166c/d* (hightest enrichment, Supplementary Figure 4.S2, 4th panel) are recovered (Table 4.4). Therefore, by combining the assembly result at both coverage levels, all targeted regions can be assembled. Overall, Velvet is sensitive to the local read coverage near the targeted loci. However, by tuning the read coverage to the range of 10-20, we can assemble all the targeted loci.

Next, we evaluated the quality of the contigs matching the *MIR165/166* loci assembled from 1% and 0.5% of the total reads respectively. Size of the contigs ranges from 1639 bp to 11652 bp, with a median of 5499 bp (Table 4.4). Undetermined nucleotide Ns in the contigs (originated from Ns in the reads) account for about one third of the total differences between the contigs and the referene genome (Table 4.4). After removing all alignment positions with an N in the contigs, the percentage of mismatches to the reference genome is low, ranging from 0% to 1.56%, with a median of 0.17%. The percentage of gaps (single or multiple indels) is relatively high, ranging from 5.17% to 21.65%, with a median of 12.45%. However, most of the differences are caused by gaps larger than 5 nt (Table 4.4). The presence of large gaps in the assembly should not significantly affect the downstream analysis, if

we apply this methodology to sample multiple plant genomes in order to study conserved CREs of *MIRNA*s. Since CREs are generally short (Lescot et al., 2002), large gaps will only appear as missing information, rather than errors and noise that confound short motif identification.

Taken together, we demonstrated in *Arabidopsis* the feasibility to specifically and sensitively enrich targeted regions with the proposed solution-based targeted genomic enrichment methodology and *de novo* assemble large genomic sequences flanking the target sites with relatively low error rates.

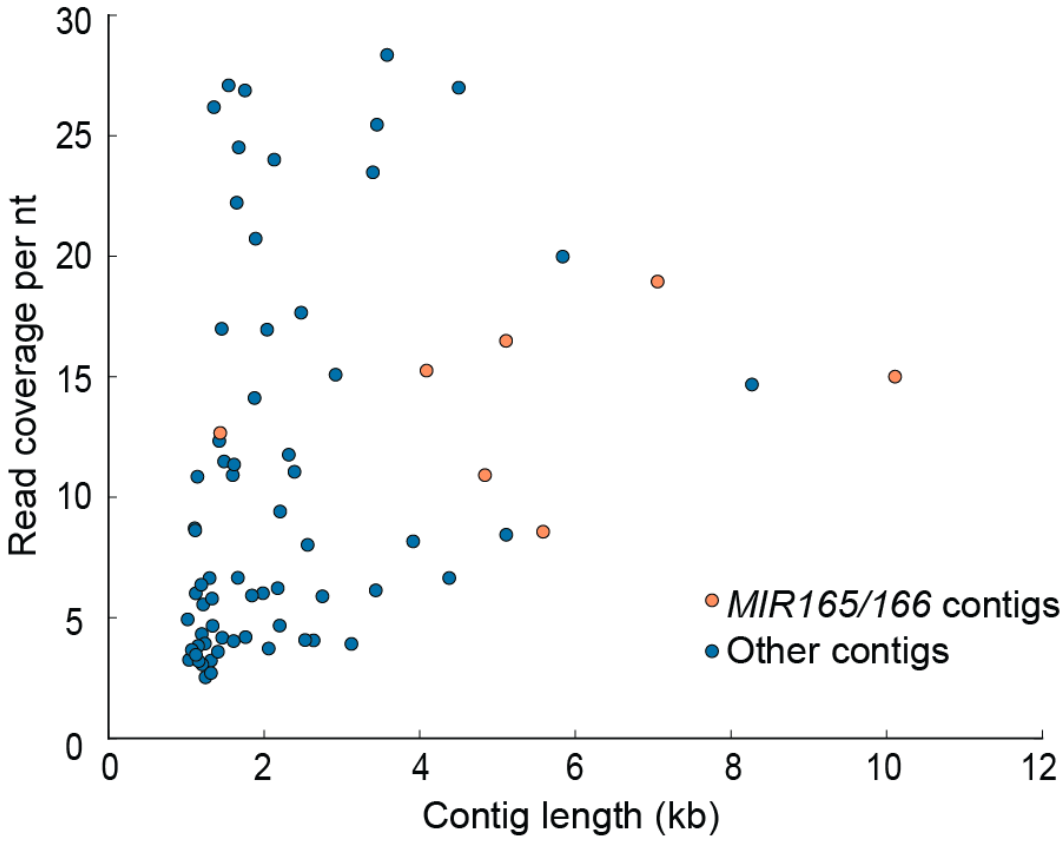


Figure 4.6 *de novo* assembly of enriched *MIR166* and *MIR165* loci. Coverage and length of contigs longer than 1 kb assembled by Velvet, using 1% of the paired-end data, are shown. Brown dots: Contigs hosting a sequence targeted by the capture probe. Blue dots: Other contigs with no complementarity to the capture probe.

Table 4.3 Velvet assembly result is sensitive to read coverage.

Percentage of total reads used in assembly	Number of reads	Coverage per nt	Number of *MIRNA* containing contigs (length > 1000 bp)	*MIRNA*s recovered
4.0%	993,695	80	0	None
2.0%	497,844	40	2	*MIR166f* and *MIR165b*
1.0%*	248,489	20	7	All but *MIR166c/d*
0.5%*	124,550	10	8**	All but *MIR165b*
0.25%*	62,474	5	6	All but *MIR165a*, *MIR165b*

* At these levels of read coverage, three independent read sampling and assembly experiments were performed. All results were consistent.

** *MIR166c* and *MIR166d* were assembled into separate contigs, despite their ~ 2kb distance. See Table 4.4 for details.

Table 4.4 Quality of assembled *MIR165/166* contigs.

Matching locus	Contig length	# mis-matches *	% mis-matches *	# nt in gaps*	% nt in gaps *	% nt in gaps > 5 nt**	# mis-matches due to N	% mis-matches due to N	# nt in gaps due to N	% nt in gaps due to N	Assembled from % of total reads
MIR166a	11,652	22	0	1,636	14	98	27	0	1,248	11	1
MIR166b	5,394	1	0	284	5	100	0	0	84	2	1
MIR166e	4,758	74	2	762	16	88	2	0	378	8	1
MIR166f	6,355	0	0	770	12	100	0	0	569	9	1
MIR166g	7,442	7	0	385	5	98	0	0	167	2	1
MIR165a	5,456	17	0	621	11	98	10	0	413	8	1
MIR165b	1,639	6	0	204	12	97	0	0	0	0	1
MIR166a	5,499	57	1	733	13	93	5	0	311	6	0.5
MIR166b	7,398	16	0	430	6	97	11	0	170	2	0.5
MIR166c	3,007	5	0	592	20	99	6	0	342	11	0.5
MIR166d	6,861	7	0	570	8	97	2	0	351	5	0.5
MIR166e	7,225	12	0	1,328	18	99	1	0	739	10	0.5
MIR166f	1,889	1	0	312	17	100	0	0	112	6	0.5
MIR166g	6,096	7	0	1,320	22	100	0	0	604	10	0.5
MIR165a	4,925	6	0	491	10	98	2	0	279	6	0.5

* These calculations exclude alignment positions where the nucleotide of the contig is N.

** Percentage of nucleotides in gaps > 5nt out of all nucleotides in gaps.

4.4.7 Trial enrichment experiments in *Zea mays* were unsuccessful

Given the success of the targeted enrichment method in *Arabidopsis*, we would like to investigate its potential application to large, complex genomes. The targeted enrichment experiment with the protocol optimized in *Arabidopsis* was performed to enrich *MIR165/166* loci in *Zea mays* (maize), whose genome is highly repetitive and 20 times the size of *Arabidopsis* genome. However, we failed to observe any significant enrichment in any of the targeted loci compared to control regions. Experimental conditions have been explored again to try to accommodate the difficulty of enrichment in a large, complex genome, including increasing hybridization temperature, increasing the amount of input gDNA, varying the probe-to-gDNA ratio, and applying a second round of enrichment. Unfortunately, none of the above attempts succeeded in enriching the targeted regions. An enrichment experiment performed with both *Arabidopsis* and maize in parallel rules out technical errors as the reason for the failure in maize, since over ~1000 fold of enrichment is observed for an *Arabidopsis MIR166* locus, while enrichment is barely seen for two maize *MIR166* loci (Table 4.5). Therefore, further optimization of the enrichment procedure will be required to extend this methodology into species with more complex and/or unknown genomes.

Table 4.5 Quantitative real-time PCR results from an enrichment experiment in both *Arabidopsis* (Ath) and maize (Zma). Ath *Act1*, Zma *Actin* and Zma *GAPDH* serve as controls.

Distance to nearest targeted locus	PCR-amplified region in Ath	Median fold of enrichment relative to Ath Act1	PCR-amplified region in Zma	Median fold of enrichment relative to Zma GAPDH
~1k	*MIR166a* Close	7466.3	*MIR166c* Close	18.9
~4k	*MIR166a* Far	1976.9	*MIR166c* Far	11.1
~1k	/	/	*MIR166m* Close	1.7
> 200k	*Act1*	1	*Actin*	2.3
> 200k	/	/	*GAPDH*	1

4.5 Discussion

4.5.1 A novel solution-based targeted genomic enrichment method successfully enriched large regions flanking targeted loci in *Arabidopsis*

We have shown the potential application of a novel solution-based targeted genomic enrichment method to enrich large flanking regions surrounding a known core sequence. Pilot experiments in *Arabidopsis* demonstrate the high specificity and sensitivity of this method to enrich sequences of interest. Successful *de novo* assembly of the sequencing reads into contigs covering the targeted loci indicated the feasibility to assemble the enriched regions in species with unknown genomes. This targeted genomic enrichment methodology is novel in several ways: First, it is the only existing enrichment method that relies solely on the knowledge of a short conserved core sequence. This method is especially suitable to study CREs of plant *MIRNAs*, because for deeply conserved loci, the ~ 21 nt mature miRNA sequences are almost identical in multiple plant species (Axtell and Bowman, 2008; Cuperus et al., 2011), while other regions of the primary transcripts are variable, and CREs are generally unknown. Since the capture probe can only be as long as the conserved sequence, i.e. 21 nt long in this project, a locked-nucleic acid (LNA)-modified probe is used to increase the thermostability of the probe-DNA-hybrid. Second, it aims to capture and enrich large genomic regions, evidently several kilobases long (Figure 4.1B, Figure 4.2E, Figure 4.6). In order to achieve this goal, DNA extraction is performed with care to reduce physical shearing, and genomic DNA is not fragmented before capture. Third, unlike most other enrichment methods which require a reference genome for mapping and identification (Mamanova et al., 2010), this method aims to identify unknown sequences flanking a known core, therefore de novo assembly is required. This requirement poses new challenge to the downstream data analysis. Finally, this method is designed to be applied to multiple species at the same time, in order to extract conservation information from multiple sequence alignments of the enriched regions. Other targeted enrichment methods are generally designed for a single genome (Mamanova et al., 2010; Bamshad et al., 2011).

4.5.2 Assembly does not require large numbers of reads

The *de novo* assembly results indicate that a small fraction of the reads generated from one lane of an Illumina GAIIx system is sufficient to assemble all targeted regions (Table 4.3), on the order of ~10^5 reads. This suggests that we could potentially bar-code a hundred samples in one sequencing run, or even more on higher-throughput instruments. One caveat in using the Velvet assembler is that its assembly result is sensitive to the read coverage (Table 4.3). We found that a coverage of 10-20 reads per nt at the targeted loci worked best. The coverage at the targeted loci can be roughly

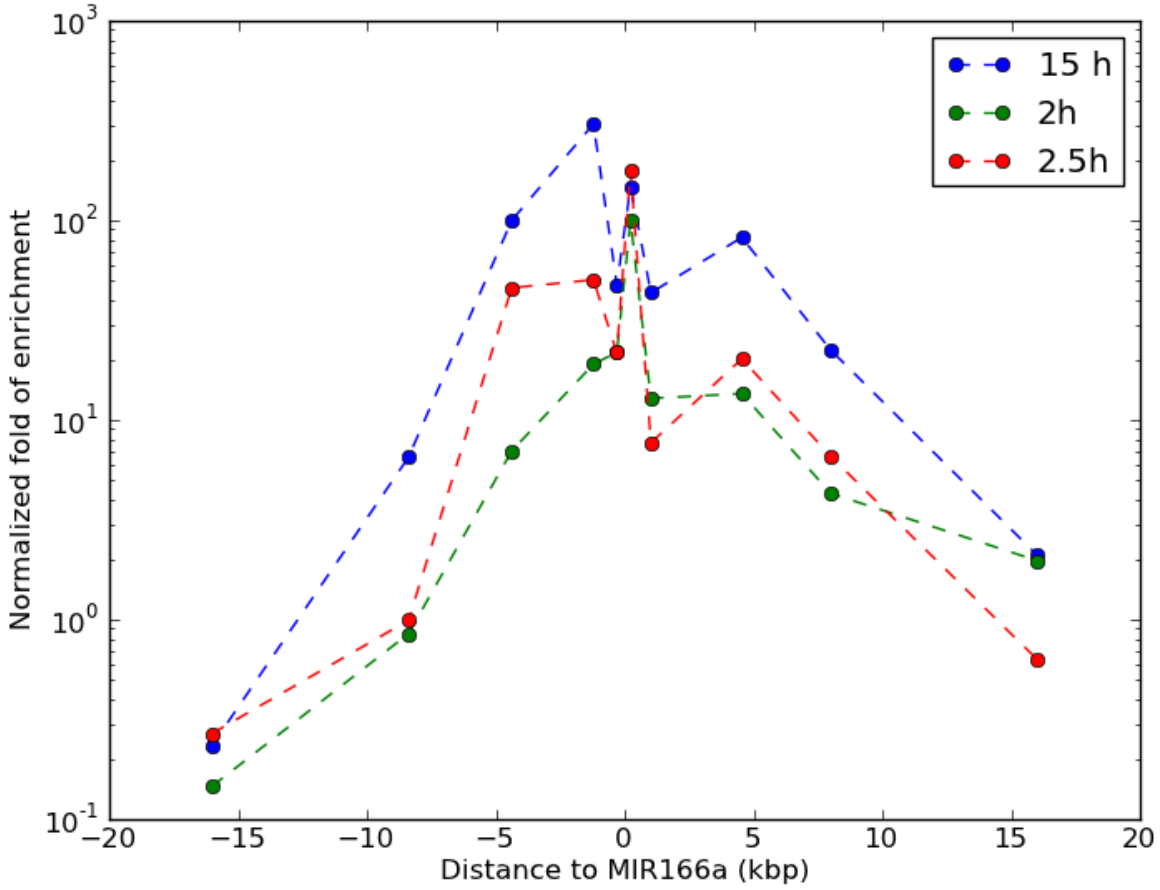
estimated even in an unknown genome as follows: 1. Count the number of reads x matching the target sequence, e.g. miR166. 2. Normalize x by x*insert_size/(2*read_length). This is the number of reads having the target sequence if the entire region has been sequenced. 3. Divide x by the number of paralogous loci in the genome, resulting in y. This information may be available in miRBase (Griffiths-Jones et al., 2008). 4. We observe that with the distance increasing from 0kb to 10kb, the normalized coverage reduced to approximately 1/100 (Figure 4.2B). Therefore we can estimate that if the assembled contig size is 2kb, then the normalized coverage near the edge of the contig is y/10 = z. 5. Adjust z to be in the range 10-20 by using a fraction of the total reads. For example, with the *Arabidopsis* paired-end dataset, x = 26,338. After step 2, x = 69,311, y ~= 10,000, and z = 1,000. To achieve 10 reads per nt, we need to sample 10/z = 1% of reads. This way we can have an educated guess how many reads to start with for an assembly in an unknown genome.

Room for improvement exists in the assembly stage, including pre-assembly error correction and using transcriptome assemblers. Pre-assembly error correction by detection and removing low frequency k-mers have been shown to increase assembly quality (Martin and Wang, 2011; Salzberg et al., 2012). Removing low complexity reads in the pre-processing may reduce the error caused by Ns in the assembled contigs (Table 4.4). Transcriptome assemblers, which take account of the large variations in sequencing depth, may be able to resolve the issue of Velvet favoring regions of a narrow range of coverage (Martin and Wang, 2011). However, using transcriptome assemblers to assemble genomic DNA may introduce unnecessary overheads, such as assembling regions of low coverage at the cost of large memory requirements, computational cost to consider strand information and splicing variants, which are not relevant for genomic DNA. Adapting transcriptome assemblers is not the aim of this study, but is worth investigating in the follow-up experiments.

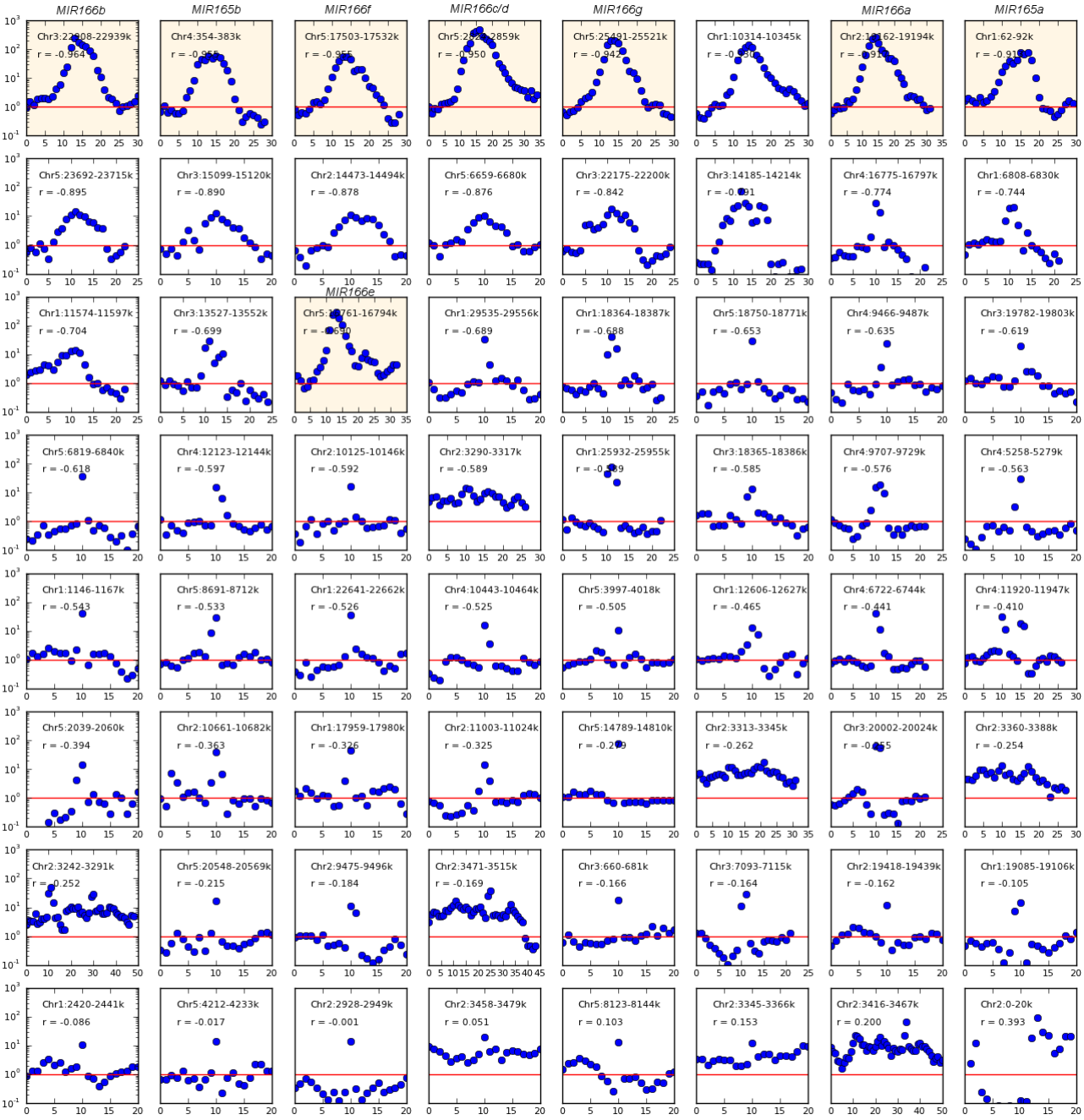
4.5.3 Methodological improvements are necessary for application in unknown genomes

Our attempt to enrich targeted regions in maize failed, despite varying multiple experimental parameters. We think the failure is likely due to the highly repetitive nature of the maize genome. We observed that ~2 kb away from the target sites, the probability of finding unique 20mers reduced significantly. Therefore, it is possible that the targeted loci are indeed captured, but the repetitive sequences flanking the the targeted loci hybridize with other repetitive sequences in the genome, and are captured and enriched together with the targeted loci. In the worst scenario, this could approaching the capture of the entire genome. One experiment to test this hypothesis is to shear the DNA into ~ 2 kb fragments before performing the targeted enrichment experiment. Alternatively, the failure of enrichment may be due to the fact that the maize genome is 20 times as large as *Arabidopsis* genome, and 11 distinct *MIR165/166* loci are present in *Z. mays* compared to eight in *Arabidopsis*, so the potential targeted sites are diluted to one fifteenth in *Z. mays*, possibly rendering

the capture infeasible. However, our effort of increasing the input genomic DNA concentration and varying the probe-to-DNA ratio did not help. It is also possible that some intrinsic property of the maize genome hinders the hybridization between the probe and the target. In any case, more efforts need to be spent before a general protocol can be developed in order to enrich sequences of interests in genomes with different size and complexity.



Supplementary Figure 4.S1 ϕ29 amplification time does not significantly affect the normalized fold of the enrichment. Quantitative real-time PCR (qPCR) shows that the normalized fold change relative to Act1 after enrichment with different ϕ29 amplification time at different distances flanking a targeted locus *MIR166a*.



Supplementary Figure 4.S2 Normalized fold change in highly enriched regions and surrounding bins. Each panel shows the normalized fold change at each 1kb-sized bin centered on a highly enriched region. Genomic coordinates of the region and the Pearson correlation r are shown. Red line indicates the genome average of the normalized coverage, which equals 1. Shaded panels are regions surrounding the 8 *MIR165/166* loci.

Supplementary Table 4.S1 Pearson correlation coefficient r of |x|* and log(y)** of highly enriched regions is a good classifier of targeted and non-targeted loci.

Locus number***	Genome coordinates	Pearson correlation r	*MIR165/166* loci
1	3:22908-22939k	-0.964	Y
2	4:354-383k	-0.955	Y
3	5:17503-17532k	-0.955	Y
4	5:2824-2859k	-0.950	Y
5	5:25491-25521k	-0.942	Y
6	1:10314-10345k	-0.930	N
7	2:19162-19194k	-0.917	Y
8	1:62-92k	-0.913	Y
9	5:23692-23715k	-0.895	N
10	3:15099-15120k	-0.890	N
11	2:14473-14494k	-0.878	N
12	5:6659-6680k	-0.876	N
13	3:22175-22200k	-0.842	N
14	3:14185-14214k	-0.791	N
15	4:16775-16797k	-0.774	N
16	1:6808-6830k	-0.744	N
17	1:11574-11597k	-0.704	N
18	3:13527-13552k	-0.699	N
19	5:16761-16794k	-0.690	Y
20	1:29535-29556k	-0.689	N
21	1:18364-18387k	-0.688	N
22	5:18750-18771k	-0.653	N
23	4:9466-9487k	-0.635	N
24	3:19782-19803k	-0.619	N
25	5:6819-6840k	-0.618	N
26	4:12123-12144k	-0.597	N
27	2:10125-10146k	-0.592	N
28	2:3290-3317k	-0.589	N
29	1:25932-25955k	-0.589	N
30	3:18365-18386k	-0.585	N
31	4:9707-9729k	-0.576	N
32	4:5258-5279k	-0.563	N
33	1:1146-1167k	-0.543	N
34	5:8691-8712k	-0.533	N
35	1:22641-22662k	-0.526	N
36	4:10443-10464k	-0.525	N
37	5:3997-4018k	-0.505	N
38	1:12606-12627k	-0.465	N
39	4:6722-6744k	-0.441	N
40	4:11920-11947k	-0.410	N
41	5:2039-2060k	-0.394	N
42	2:10661-10682k	-0.363	N
43	1:17959-17980k	-0.326	N
44	2:11003-11024k	-0.325	N
45	5:14789-14810k	-0.279	N

46	2:3313-3345k	-0.262	N
47	3:20002-20024k	-0.255	N
48	2:3360-3388k	-0.254	N
49	2:3242-3291k	-0.252	N
50	5:20548-20569k	-0.215	N
51	2:9475-9496k	-0.184	N
52	2:3471-3515k	-0.169	N
53	3:660-681k	-0.166	N
54	3:7093-7115k	-0.164	N
55	2:19418-19439k	-0.162	N
56	1:19085-19106k	-0.105	N
57	1:2420-2441k	-0.086	N
58	5:4212-4233k	-0.017	N
59	2:2928-2949k	-0.001	N
60	2:3458-3479k	0.051	N
61	5:8123-8144k	0.103	N
62	2:3345-3366k	0.153	N
63	2:3416-3467k	0.200	N
64	2:0-20k	0.393	N

* x is the distance to the most highly enriched bin in the 21 kb region centered on that bin.

** y is the normalized coverage.

*** See Supplementary Figure 4.S2 for plots of all loci.

Chapter 5
Summary and prospects

5.1 Summary

5.1.1 Transient plant *MIRNA* and siRNA loci

In Chapter 2, by comparative analysis of small RNAs in two closely related species *Arabidopsis thaliana* and *Arabidopsis lyrata*, we find that less-conserved miRNAs have high rates of sequence divergence, have few and divergent targets between species, and are processed imprecisely in the other species. These findings suggest that young *MIRNAs* are evolutionarily transient, and their formation seems to be a neutral process. These observations are in line with multiple studies. First, nucleotide divergence patterns between *A. thaliana* and *A. lyrata* orthologous *MIRNA* genes were muchi higher in the mature miRNA region for less-conserved miRNAs, consistent with neutral evolution. Second, about half of *A. thaliana* targets of conserved miRNAs accumulate at higher levels in miRNA biogenesis mutants (Ronemus et al., 2006), whereas putative targets of nonconserved miRNAs are largely unaffected, indicating that most newly evolved miRNAs are not integrated in regulatory circuits (Fahlgren et al., 2007). Indeed, we find that predicted and validated targets are rare and highly species-specific for less-conserved miRNAs. Third, we find that *Arabidopsis-specific* miRNAs are expressed at lower levels relative to more-conserved miRNAs and their expression is often limited in one species. Recent data shows that the expression level bias is also evident at the primary *MIRNA* transcript level (Laubinger et al., 2010). This could be explained either by highly tissue-specific expression pattern, or by the lack of regulatory elements that deliver robust expression. Young miRNAs also tend to be processed imprecisely, which further questions a relevant regulatory role of young miRNAs.

We also observe that heterochromatic siRNA loci have a slight tendency to be retained between species, but the most active *A. lyrata* heterochromatic siRNA hot spots are generally not syntenic to the most active siRNA hot spots of *A. thaliana*. This indicates that despite the deep conservation of the heterochromatic siRNA pathway (Cho et al., 2008), most heterochromatic siRNA hot spots are rapidly changing and evolutionarily transient within the *Arabidopsis* genus, likely a mechanism in response to the rapid changes in transposon positions. Admittedly, the identification method of siRNA loci used in this study is quite crude. More studies and better siRNA annotation techniques are needed to clarify this issue. For example, ShortStack, a recently-developed comprehensive small RNA annotation software package, can integrate small RNA size distributions,

repetitiveness, strandedness, hairpin-association, and phasing information to annotate and quantify heterochromatic siRNA loci (Axtell, 2013b).

5.1.2 Resolving power of integrated sequencing data analysis

In Chapter 3, we surveyed three plant genomes to characterize specific small RNA population. It is worth noting that in the three projects, we have increasing amount of information. In the first *Theobroma cacao* project, the only available information is the draft genome. What we can do to characterize small RNAs is thus limited to computational prediction of conserved *MIRNAs*. Without small RNA-seq data, it remains uncertain whether these conserved *MIRNAs* are actually expressed in *T. cacao*. In the second oil palm project, we have small RNA-seq data but not the reference genome. This time, we are able to identify expressed conserved miRNA families, but without a reference genome, it is impossible to determine whether these miRNAs are encoded by a single locus or multiple paralogous loci. In the third *Physcomitrella patens* project, we have at our disposal data from multiple small RNA-seq libraries in different genetic backgrounds, a draft genome assembly, as well as degradome sequencing data. By integrating all this information, we identified a novel family of trans-acting siRNA (ta-siRNA) loci associated with miR156- and miR529-directed slicing, and elucidated a regulatory cascade initiated by miRNA-directed cleavage of a *TAS6* locus that produces ta-siRNAs, one of which triggers slicing of a zinc-finger domain transcript.

Taken together, the three projects contrast each other to demonstrate the power of integrated sequencing data analysis. This is consistent with the concept of value of information theory, which shows that additional information has a non-negative value (Howard, 1966). Simply put, it means something is better than nothing. Fortunately, with the advancement of sequencing technologies, sequencing a small RNA library, or even a genome, will soon become a daily routine (Mardis, 2008a; Branton et al., 2008; Simon et al., 2009; Lister et al., 2009).

5.1.3 Opportunities and challenges in the sequencing era

Next generation sequencing provides unprecedented opportunities for all biological and medical researches (Mardis, 2008a; Simon et al., 2009; Lister et al., 2009), including but not limited to small RNA researches. Sequencing technologies transform traditional research method into high-throughput, genome-wide and single-base resolution techniques. Notable examples are 5' RACE-based degradome sequencing (Addo-Quaye et al., 2008; German et al., 2008a), chromatin-immunoprecipitation-based ChIP-Seq (Bernstein et al., 2005; Johnson et al., 2007; Robertson et al., 2007) and CLIP-Seq (Licatalosi et al., 2008; Chi et al., 2009), just to name a few. Novel experimental

methods enabled by cheap highly parallel sequencing are constantly emerging.

In Chapter 4, we developed a solution-based targeted genomic enrichment methodology to capture, enrich, and sequence flanking genomic regions surrounding conserved *MIRNA* genes. This method may enable the determination of flanking genomic DNA sequences surrounding a known core from multiple species that lack complete genome assemblies, and in turn accelerate discovery of CREs surrounding such loci.

Wide adoption of the next generation sequencing technologies presents new challenges related to large-scale data manipulation, processing and analysis (Fahlgren et al., 2009; Schadt et al., 2010). Standadized data analysis guidelines should be developed for well established method, as has been done for RNA sequencing (RNAseq) (http://encodeproject.org/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf) and ChIP-seq (Landt et al., 2012). Also, limitation and biases of specific methodology needs to be considered during experimental design and data analysis, such as ligation biases in small RNA-seq experiments (Jayaprakash et al., 2011; Sorefan et al., 2012).

5.2 Prospects

5.2.1 Characterization of diverse small RNA population in plants

A decade has passed since the first description of endogenous small RNAs in plants (Reinhart et al., 2002; Park et al., 2002; Llave et al., 2002). With the help of sequencing technologies, we now understand that a diverse classes of small RNAs exist in plants. However, the current knowledge of the small RNA population is largely gained from researches in a few model organisms. Studies in other plants are limited by the availability of small RNA-seq data and/or whole genome assemblies (see Chapter 3.2 and 3.3). In the next decade of plant small RNA researches, sequencing technologies are likely to deepen our understanding of the diversity and complexity of small RNA population in two ways:

First, studies of small RNAs in non-model plants will likely reveal novel types or functional pathways of small RNAs. The recent discovery of miR482/miR2118-triggered secondary siRNA biogenesis of disease-resistance genes of the NBS-LRR superfamily in tomato (Zhai et al., 2011; Shivaprasad et al., 2012), tobacco and potato (Li et al., 2012b) is a good example. These studies reveal a link between viral/bacterial infections and increases in NBS-LRR mRNA accumulation, mediated by reduced miR482/miR2118 trigger and reduced production of secondary siRNAs.

Second, characterization of cell-type-specific small RNAs will uncover small RNAs whose expression is restricted to discrete time or space. Recent studies of small RNAs in pollen, ovules and

developing seeds point out the importance of this direction (Slotkin et al., 2009; Mosher et al., 2009; Olmedo-Monfil et al., 2010).

5.2.2 Elucidation of small RNA biogenesis pathways

The biogenesis pathway of miRNAs has been extensively studied. However, posttranscriptional control of *MIRNAs* conveyed by regulation of the miRNA biogenesis pathway is not fully characterized in plants, whereas in animals the paradigm is well established (Newman and Hammond, 2010). A few studies in plants point out the possibility that pri-miRNAs may be differentially processed for exquisite spatio–temporal control. Maize mature miR166a are not detected in the tip of the shoot apical meristem, despite the accumulation of the pri-miRNA, indicating that miR166 accumulation is partly controlled from intricate transcriptional regulation of its precursor loci (Nogueira et al., 2009). A few *A. thaliana* pri-miRNAs accumulate in tissue-specific manner in wild-type, for example, *ath-MIR172b* transcripts were present in inflorescence and absent in seedlings, but in *dcl1* mutants they were detected in both tissues (Laubinger et al., 2010). It remains unclear the mechanism and prevalence of such differential processing at the post-transcriptional level of *MIRNAs*.

Our understanding of the biogenesis pathway of heterochromatic siRNAs has advanced with the identification of Pol V transcripts (Wierzbicki et al., 2009) and the recent characterization of genome-wide Pol V occupancy in *Arabidopsis* (Zhong et al., 2012; Wierzbicki et al., 2012). However, a number of questions remain to be answered. Not all Pol V-occupied sites correlate with asymmetric DNA methylation and heterochromatic siRNA accumulation, indicating that Pol V occupancy is independent of siRNA production. Thus, what controls Pol V occupancy? Does Pol V occupancy correlate with transcription of the Pol V transcripts? How is Pol V transcription initiation regulated? Elucidating the mechanisms of Pol IV and Pol V occupancy and initiation is an important goal for future research.

5.2.3 Conservation and diversification of small RNA regulatory networks

miR156 and miR390 are two ancient plant miRNAs conserved between *Arabidopsis* and *Physcomitrella*, and are involved in vegetative phase transitions in both species (Wu and Poethig, 2006; Axtell et al., 2006, 2007). In Chapter 3.3, we find in *Physcomitrella* that a new family of *TAS* loci *TAS6*, which are associated with miR156-directed slicing, are in close proximity to miR390-targeted *TAS3* loci. Thus, it suggests that these *PpTAS6*/*PpTAS3* pairs could share single common primary transcripts, and that miR156- and miR390-mediated activities may be inter-related in *Physcomitrella*. Indeed, a subsequent study confirmed that *TAS6a* and *TAS3a* share a single primary transcript (Cho et al., 2012). It is interesting that in *Arabidopsis*, miR156 and miR390 regulate developmental

transitions independently (Fahlgren et al., 2006; Wu and Poethig, 2006). However, in *Physcomitrella*, they converge on the same pathway by targeting a single *TAS* precursor, which may regulate the optimal timing of phase transition by fine-tuning the levels of tasiRNAs and their targets. The miR390-tasiRNA pathway is conserved in both species to repress vegetative developmental transitions, while miR156 has opposite roles: it represses developmental transitions in *Arabidopsis* and promotes them in *Physcomitrella* (Wu and Poethig, 2006; Cho et al., 2012). Future researches are needed to fully elucidate the differences in the pathways regulating developmental transitions in different plant lineages.

Most conserved plant *MIRNAs* are encoded by multiple paralogous loci, and are often transcribed independently (Reinhart et al., 2002; Zhou et al., 2007). Thus, expression of paralogous *MIRNAs* can be regulated differentially to convey a fine-tuned spatial, temporal, and developmental control. In *Arabidopsis*, four *MIR167* genes are expressed in distinct floral organ domains, driven by their respective promoters (Wu et al., 2006). Computational analysis has identified an enrichment of cis-regulatory elements involved in development, stress responses and hormonal control in a group of conserved *MIRNA* loci (Megraw et al., 2006). However, it remains unclear whether cis-regulatory elements of conserved *MIRNAs* are evolutionarily conserved. The targeted genomic enrichment methodology we developed In Chapter 4 can potentially determine flanking genomic DNA sequences surrounding a known core from multiple species, and discover conserved CREs surrounding such loci. Orthologous *MIRNA* loci across species can be identified by the conservation of flanking genes, therefore the distinctive regulation of each paralogous locus can be sorted out. Alternatively, with the growing number of available plant genome assemblies (Goodstein et al., 2011), we will be able to study conserved CREs of plant *MIRNAs* on various evolutionary scales using whole genome sequences in the near future.

Appendix
List of Supplemental Datasets

Supplemental Dataset 2.1 *A. thaliana* and *A. lyrata MIRNA* loci examined in this study.

Supplemental Dataset 2.2 Expression details of *A. thaliana MIRNA* loci which passed the Meyers et al. (2008) criteria.

Supplemental Dataset 2.3 Expression details of *A. lyrata MIRNA* loci which passed the Meyers et al. (2008) criteria.

Supplemental Dataset 2.4 Predicted targets of *A. lyrata* miRNAs.

Supplemental Dataset 2.5 Degradome information for *A. thaliana* sliced targets confidently identified in both *A. thaliana* biological replicates; data from the AT-deg1 sample.

Supplemental Dataset 2.6 Degradome information for *A. thaliana* sliced targets confidently identified in both *A. thaliana* biological replicates; data from the AT-deg2 sample.

Supplemental Dataset 2.7 Degradome information for *A. lyrata* sliced targets confidently identified in both *A. lyrata* biological replicates; data from the AL-deg1 sample.

Supplemental Dataset 2.8 Degradome information for *A. lyrata* sliced targets confidently identified in both *A. lyrata* biological replicates; data from the AL-deg2 sample.

Supplemental Dataset 3.1 Newly identified Theobroma cacao MIRNA loci, scaffold coordinates, and the predicted secondary structure by Mfold with the corresponding miRNAs in lower case letters

Supplemental Dataset 3.2 Scaffold coordinates of Theobroma cacao microRNAs in GFF format

Supplemental Dataset 3.3 The alignment between the target gene and a microRNA in the family, complimentarity score, the top BLASTx hit in Arabidopsis thaliana and Oryza sativa are shown.

Supplemental Dataset 3.4 Fasta file of the curated high-confidence miRNAs from *Arabidpsis thaliana*, rice and *Physcomitrella patens*.

Supplemental Dataset 3.5 Target predictions for *PpTAS6*-derived siRNAs. Targets were predicted using the JGI FM3 transcript set with axtell_targetfinder.pl (Ma *et al.,* 2010). Alignment scores of 3.5 or less are shown.

Supplemental Dataset 3.6 A .bed file indicating genomic locations of *Physcomitrella TAS* loci (genome assembly version 1.1) as well as important features within the *TAS* loci.

Note: Supplemental Datasets are provided in case the readers need to look into specific details. These are large datasets that are not supposed to be read line by line, but rather for search purposes. They are accessible at:
http://axtelldata.bio.psu.edu/data/Ma_Dissertation_SupDatasets

References

Addo-Quaye, C., Eshoo, T.W., Bartel, D.P., and Axtell, M.J. (2008). Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome. Curr. Biol. **18**: 758–762.

Addo-Quaye, C., Miller, W., and Axtell, M.J. (2009a). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics **25**: 130–131.

Addo-Quaye, C., Snyder, J.A., Park, Y.B., Li, Y.-F., Sunkar, R., and Axtell, M.J. (2009b). Sliced microRNA targets and precise loop-first processing of MIR319 hairpins revealed by analysis of the Physcomitrella patens degradome. RNA **15**: 2112–2121.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., Weinstock, G.M., and Gibbs, R.A. (2007). Direct selection of human genomic loci by microarray hybridization. Nat. Meth. **4**: 903–905.

Allen, E., Xie, Z., Gustafson, A.M., and Carrington, J.C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell **121**: 207–221.

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.-H., Spatafora, J.W., and Carrington, J.C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat. Genet. **36**: 1282–1290.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G., and Tuschl, T. (2003). A uniform system for microRNA annotation. RNA **9**: 277–279.

Amor, B.B., Wirth, S., Merchan, F., Laporte, P., d' Aubenton-Carafa, Y., Hirsch, J., Maizel, A., Mallory, A., Lucas, A., Deragon, J.M., Vaucheret, H., Thermes, C., and Crespi, M. (2009). Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. Genome Res. **19**: 57–69.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The Small RNA Profile during Drosophila melanogaster Development. Dev. Cell **5**: 337–350.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally Regulated piRNA Clusters Implicate MILI in Transposon Control. Science **316**: 744–747.

Aravind, L., Watanabe, H., Lipman, D.J., and Koonin, E.V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc. Natl. Acad. Sci. U. S. A. **97**: 11319–11324.

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J.F., Sabot, F., Kudrna, D., et al. (2011). The genome of Theobroma cacao. Nat. Genet. **43**: 101–108.

Arif, M.A., Fattash, I., Ma, Z., Cho, S.H., Beike, A.K., Reski, R., Axtell, M.J., and Frank, W. (2012). DICER-LIKE3 Activity in Physcomitrella patens DICER-LIKE4 Mutants Causes Severe

Developmental Dysfunction and Sterility. Mol. Plant **5**: 1281–1294.

Aukerman, M.J. and Sakai, H. (2003). Regulation of Flowering Time and Floral Organ Identity by a MicroRNA and Its APETALA2-Like Target Genes. Plant Cell Online **15**: 2730–2741.

Axtell, M.J. (2013a). Classification and Comparison of Small RNAs from Plants. Annu. Rev. Plant Biol. **64**: 137–159.

Axtell, M.J. (2008). Evolution of microRNAs and their targets: Are all microRNAs biologically relevant? Biochim. Biophys. Acta Bba - Gene Regul. Mech. **1779**: 725–734.

Axtell, M.J. (2013b). ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA. **19**: 740-751.

Axtell, M.J. and Bartel, D.P. (2005). Antiquity of MicroRNAs and Their Targets in Land Plants. Plant Cell Online **17**: 1658–1673.

Axtell, M.J. and Bowman, J.L. (2008). Evolution of plant microRNAs and their targets. Trends Plant Sci. **13**: 343–349.

Axtell, M.J., Jan, C., Rajagopalan, R., and Bartel, D.P. (2006). A Two-Hit Trigger for siRNA Biogenesis in Plants. Cell **127**: 565–577.

Axtell, M.J., Snyder, J.A., and Bartel, D.P. (2007). Common Functions for Diverse Small RNAs of Land Plants. Plant Cell Online **19**: 1750–1769.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. **12**: 745–755.

Barberis, A., Superti-Furga, G., and Busslinger, M. (1987). Mutually exclusive interaction of the CCAAT-binding factor and of a displacement protein with overlapping sequences of a histone gene promoter. Cell **50**: 347–359.

Bari, R., Pant, B.D., Stitt, M., and Scheible, W.-R. (2006). PHO2, MicroRNA399, and PHR1 Define a Phosphate-Signaling Pathway in Plants. Plant Physiol. **141**: 988–999.

Barnes, W.M. (1994). PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc. Natl. Acad. Sci. U. S. A. **91**: 2216–2220.

Bartel, D.P. (2004). MicroRNAs genomics, biogenesis, mechanism, and function. Cell **116**: 281–297.

Bartel, D.P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. Cell **136**: 215–233.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D., and Rhee, S.Y. (2004). Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies. Plant Physiol. **135**: 745–755.

Berezikov, E., Liu, N., Flynt, A.S., Hodges, E., Rooks, M., Hannon, G.J., and Lai, E.C. (2010). Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. Nat. Genet. **42**: 6–9.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., Gingeras, T.R., Schreiber, S.L., and Lander, E.S. (2005). Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. Cell **120**: 169–181.

Borsani, O., Zhu, J., Verslues, P.E., Sunkar, R., and Zhu, J.-K. (2005). Endogenous siRNAs Derived from a Pair of Natural cis-Antisense Transcripts Regulate Salt Tolerance in Arabidopsis. Cell **123**: 1279–1291.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., Krstic, P.S., Lindsay, S., Ling, X.S., Mastrangelo, C.H., Meller, A., Oliver, J.S., Pershin, Y.V., Ramsey, J.M., Riehn, R., et al. (2008). The potential and challenges of nanopore sequencing. Nat. Biotechnol. **26**: 1146–1153.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell **128**: 1089–1103.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., et al. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. **18**: 630–634.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread Translational Inhibition by Plant miRNAs and siRNAs. Science **320**: 1185–1190.

Brodersen, P. and Voinnet, O. (2009). Revisiting the principles of microRNA target recognition and mode of action. Nat. Rev. Mol. Cell Biol. **10**: 141–148.

Bruce Wightman, Ilho Ha, and Gary Ruvkun (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell **75**: 855–862.

Carbonell, A., Fahlgren, N., Garcia-Ruiz, H., Gilbert, K.B., Montgomery, T.A., Nguyen, T., Cuperus, J.T., and Carrington, J.C. (2012). Functional Analysis of Three Arabidopsis ARGONAUTES Using Slicer-Defective Mutants. Plant Cell Online **24**: 3613–3629.

Carthew, R.W. and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell **136**: 642–655.

Chaw, S.-M., Chang, C.-C., Chen, H.-L., and Li, W.-H. (2004). Dating the Monocot–Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes. J. Mol. Evol. **58**: 424–441.

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

Chen, H.-M., Li, Y.-H., and Wu, S.-H. (2007). Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. **104**: 3318–3323.

Chen, X. (2004). A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development. Science **303**: 2022–2025.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature **460**: 479–486.

Cho, S.H., Addo-Quaye, C., Coruh, C., Arif, M.A., Ma, Z., Frank, W., and Axtell, M.J. (2008). Physcomitrella patens DCL3 Is Required for 22–24 nt siRNA Accumulation, Suppression of Retrotransposon-Derived Transcripts, and Normal Development. Plos Genet. **4**: e1000314.

Cho, S.H., Coruh, C., and Axtell, M.J. (2012). miR156 and miR390 Regulate tasiRNA Accumulation and Developmental Timing in Physcomitrella patens. Plant Cell Online **24**: 4837–4849.

Collins, F.S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. Science **300**: 286–290.

Covey, S.N., Al-Kaff, N.S., Lángara, A., and Turner, D.S. (1997). Plants combat infection by gene silencing. Nature **385**: 781–782.

Cuperus, J.T., Fahlgren, N., and Carrington, J.C. (2011). Evolution and Functional Diversification of MIRNA Genes. Plant Cell Online **23**: 431–442.

Czech, B. and Hannon, G.J. (2011). Small RNA sorting: matchmaking for Argonautes. Nat. Rev. Genet. **12**: 19–31.

Czech, B., Malone, C.D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J.A., Sachidanandam, R., Hannon, G.J., and Brennecke, J. (2008). An endogenous small interfering RNA pathway in Drosophila. Nature **453**: 798–802.

D'Ascenzo, M., Meacham, C., Kitzman, J., Middle, C., Knight, J., Winer, R., Kukricar, M., Richmond, T., Albert, T.J., Czechanski, A., Donahue, L.R., Affourtit, J., Jeddeloh, J.A., and Reinholdt, L. (2009). Mutation discovery in the mouse using genetically guided array capture and resequencing. Mamm. Genome Off. J. Int. Mamm. Genome Soc. **20**: 424–436.

Dolgosheina, E.V., Morin, R.D., Aksay, G., Sahinalp, S.C., Magrini, V., Mardis, E.R., Mattsson, J., and Unrau, P.J. (2008). Conifers have a unique small RNA silencing signature. RNA **14**: 1508–1515.

Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R., and Bartel, D.P. (2009). RNAi in Budding Yeast. Science **326**: 544–550.

Dsouza, M., Larsen, N., and Overbeek, R. (1997). Searching for patterns in genomic data. Trends Genet. **13**: 497–498.

Ecker, J.R. and Davis, R.W. (1986). Inhibition of gene expression in plant cells by expression of antisense RNA. Proc. Natl. Acad. Sci. U. S. A. **83**: 5372–5376.

Eddy, S.R. (2005). A Model of the Statistical Power of Comparative Genome Sequence Analysis. Plos Biol. **3**: e10.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**: 1792 –1797.

Ehrenreich, I.M. and Purugganan, M.D. (2008). Sequence Variation of MicroRNAs and Their Binding Sites in Arabidopsis. Plant Physiol. **146**: 1974–1982.

Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J., and Birney, E. (2005). The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. Genome Biol. **6**: R104–R104.

Fagard, M., Boutet, S., Morel, J.-B., Bellini, C., and Vaucheret, H. (2000). AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. Proc. Natl. Acad. Sci. U. S. A. **97**: 11650–11654.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., and Dangl, J.L. (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. Plos one **2**: 219.

Fahlgren, N., Jogdeo, S., Kasschau, K.D., Sullivan, C.M., Chapman, E.J., Laubinger, S., Smith, L.M., Dasenko, M., Givan, S.A., Weigel, D., and Carrington, J.C. (2010). MicroRNA Gene Evolution in Arabidopsis lyrata and Arabidopsis thaliana. Plant Cell Online **22**: 1074–1089.

Fahlgren, N., Montgomery, T.A., Howell, M.D., Allen, E., Dvorak, S.K., Alexander, A.L., and Carrington, J.C. (2006). Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA Affects Developmental Timing and Patterning in Arabidopsis. Curr. Biol. **16**: 939–944.

Fahlgren, N., Sullivan, C.M., Kasschau, K.D., Chapman, E.J., Cumbie, J.S., Montgomery, T.A., Gilbert, S.D., Dasenko, M., Backman, T.W.H., Givan, S.A., and Carrington, J.C. (2009). Computational and analytical framework for small RNA profiling by high-throughput sequencing. RNA **15**: 992–1002.

Fang, Y. and Spector, D.L. (2007). Identification of Nuclear Dicing Bodies Containing Proteins for MicroRNA Biogenesis in Living Arabidopsis Plants. Curr. Biol. **17**: 818–823.

Felippes, F.F. de, Schneeberger, K., Dezulian, T., Huson, D.H., and Weigel, D. (2008). Evolution of Arabidopsis thaliana microRNAs from random sequences. RNA **14**: 2455–2459.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature **391**: 806–811.

Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods **7**: 461–465.

Franco-Zorrilla, J.M., Del Toro, F.J., Godoy, M., Pérez-Pérez, J., López-Vidriero, I., Oliveros, J.C., García-Casado, G., Llave, C., and Solano, R. (2009). Genome-wide identification of small RNA targets based on target enrichment and microarray hybridizations. Plant J. **59**: 840–850.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J.A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. Nat. Genet. **39**: 1033–1037.

Fu, Y., Springer, N.M., Gerhardt, D.J., Ying, K., Yeh, C.-T., Wu, W., Swanson-Wagner, R.,

D'Ascenzo, M., Millard, T., Freeberg, L., Aoyama, N., Kitzman, J., Burgess, D., Richmond, T., Albert, T.J., Barbazuk, W.B., Jeddeloh, J.A., and Schnable, P.S. (2010). Repeat subtraction-mediated sequence capture from a complex genome. Plant J. Cell Mol. Biol. **62**: 898–909.

Gan, H., Lin, X., Zhang, Z., Zhang, W., Liao, S., Wang, L., and Han, C. (2011). piRNA profiling during specific stages of mouse spermatogenesis. RNA **17**: 1191–1203.

Gandikota, M., Birkenbihl, R.P., Höhmann, S., Cardon, G.H., Saedler, H., and Huijser, P. (2007). The miRNA156/157 recognition element in the 3′ UTR of the Arabidopsis SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings. Plant J. **49**: 683–693.

Garcia, D. (2008). A miRacle in plant development: Role of microRNAs in cell differentiation and patterning. Semin. Cell Dev. Biol. **19**: 586–595.

German, M.A., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L.A., Nobuta, K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B.C., and Green, P.J. (2008a). Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends. Nat. Biotechnol. **26**: 941–946.

German, M.A., Pillay, M., Jeong, D.H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L.A., Nobuta, K., and German, R. (2008b). Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends. Nat. Biotechnol. **26**: 941–946.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S., and Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. **27**: 182–189.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2011). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. **40**: D1178–D1186.

Gregory, B.D., O'Malley, R.C., Lister, R., Urich, M.A., Tonti-Filippini, J., Chen, H., Millar, A.H., and Ecker, J.R. (2008). A Link between RNA Metabolism and Silencing Affecting Arabidopsis Development. Dev. Cell **14**: 854–866.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. Nucl Acids Res **36**: D154–158.

Hardison, R.C. (2003). Comparative Genomics. Plos Biol. **1**: e58.

Heisel, S.E., Zhang, Y., Allen, E., Guo, L., Reynolds, T.L., Yang, X., Kovalic, D., and Roberts, J.K. (2008). Characterization of Unique Small RNA Populations from Rice Grain. Plos One **3**: e2871.

Henz, S.R., Cumbie, J.S., Kasschau, K.D., Lohmann, J.U., Carrington, J.C., Weigel, D., and Schmid, M. (2007). Distinct Expression Patterns of Natural Antisense Transcripts in Arabidopsis. Plant Physiol. **144**: 1247–1255.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., and McCombie, W.R. (2007). Genome-wide in situ exon

capture for selective resequencing. Nat. Genet. **39**: 1522–1527.

Howard, R.A. (1966). Information Value Theory. Ieee Trans. Syst. Sci. Cybern. **2**: 22–26.

Howell, M.D., Fahlgren, N., Chapman, E.J., Cumbie, J.S., Sullivan, C.M., Givan, S.A., Kasschau, K.D., and Carrington, J.C. (2007a). Genome-Wide Analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 Pathway in Arabidopsis Reveals Dependency on miRNA- and tasiRNA-Directed Targeting. Plant Cell Online **19**: 926–942.

Howell, M.D., Fahlgren, N., Chapman, E.J., Cumbie, J.S., Sullivan, C.M., Givan, S.A., Kasschau, K.D., and Carrington, J.C. (2007b). Genome-Wide Analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 Pathway in Arabidopsis Reveals Dependency on miRNA- and tasiRNA-Directed Targeting. Plant Cell Online **19**: 926–942.

Hu, Z., Jiang, Q., Ni, Z., Chen, R., Xu, S., and Zhang, H. (2013). Analyses of a Glycine max Degradome Library Identify microRNA Targets and MicroRNAs that Trigger Secondary SiRNA Biogenesis. J. Integr. Plant Biol. **55**: 160–176.

Inostroza, A. (1992). Dr1, a TATA-binding protein-associated phosphoprotein and inhibitor of class II gene transcription. Cell **70**: 477–489.

Jayaprakash, A.D., Jabado, O., Brown, B.D., and Sachidanandam, R. (2011). Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. **39**: e141–e141.

Jiao, Y., Riechmann, J.L., and Meyerowitz, E.M. (2008). Transcriptome-Wide Analysis of Uncapped mRNAs in Arabidopsis Reveals Regulation of mRNA Degradation. Plant Cell Online **20**: 2571–2585.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science **316**: 1497–1502.

Jones-Rhoades, M.W. and Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol. Cell **14**: 787–799.

Jones-Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of Plant MicroRNA Targets. Cell **110**: 513–520.

Juliano, C., Wang, J., and Lin, H. (2011). Uniting Germline and Stem Cells: The Function of Piwi Proteins and the piRNA Pathway in Diverse Organisms. Annu. Rev. Genet. **45**: 447–469.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-Wide Profiling and Analysis of Arabidopsis siRNAs. Plos Biol. **5**: e57.

Katiyar-Agarwal, S., Morgan, R., Dahlbeck, D., Borsani, O., Villegas, A., Zhu, J.-K., Staskawicz, B.J., and Jin, H. (2006). A pathogen-inducible endogenous siRNA in plant immunity. Proc. Natl. Acad. Sci. U. S. A. **103**: 18002–18007.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. U. S. A. **100**: 11484–11489.

Khraiwesh, B., Arif, M.A., Seumel, G.I., Ossowski, S., Weigel, D., Reski, R., and Frank, W. (2010). Transcriptional Control of Gene Expression by MicroRNAs. Cell **140**: 111–122.

Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nat. Rev. Mol. Cell Biol. **10**: 126–139.

Krishnakumar, S., Zheng, J., Wilhelmy, J., Faham, M., Mindrinos, M., and Davis, R. (2008). A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc. Natl. Acad. Sci. U. S. A. **105**: 9296–9301.

Kurihara, Y., Takashi, Y., and Watanabe, Y. (2006). The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. RNA **12**: 206–212.

Kurihara, Y. and Watanabe, Y. (2004). Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. Proc. Natl. Acad. Sci. U. S. A. **101**: 12753–12758.

Kutter, C., Schöb, H., Stadler, M., Meins, F., and Si-Ammour, A. (2007). MicroRNA-Mediated Regulation of Stomatal Development in Arabidopsis. Plant Cell Online **19**: 2417–2429.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of Novel Genes Coding for Small Expressed RNAs. Science **294**: 853–858.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. **22**: 1813–1831.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**: R25.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. Science **294**: 858–862.

Lau, N.C., Robine, N., Martin, R., Chung, W.-J., Niki, Y., Berezikov, E., and Lai, E.C. (2009). Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome Res. **19**: 1776–1785.

Laubinger, S., Sachsenberg, T., Zeller, G., Busch, W., Lohmann, J.U., Rätsch, G., and Weigel, D. (2008). Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. **105**: 8795–8800.

Laubinger, S., Zeller, G., Henz, S.R., Buechel, S., Sachsenberg, T., Wang, J.-W., Rätsch, G., and Weigel, D. (2010). Global effects of the small RNA biogenesis machinery on the Arabidopsis thaliana transcriptome. Proc. Natl. Acad. Sci. U. S. A. **107**: 17466–17473.

Lee, R.C. and Ambros, V. (2001). An Extensive Class of Small RNAs in Caenorhabditis elegans. Science **294**: 862–864.

Lee, T.I. and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. Annu. Rev. Genet.

34: 77–137.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. Embo J. **23**: 4051–4060.

Lelandais-Brière, C., Naya, L., Sallet, E., Calenge, F., Frugier, F., Hartmann, C., Gouzy, J., and Crespi, M. (2009). Genome-Wide Medicago truncatula Small RNA Analysis Revealed Novel MicroRNAs and Isoforms Differentially Regulated in Roots and Nodules. Plant Cell Online **21**: 2780–2796.

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouzé, P., and Rombauts, S. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. **30**: 325 –327.

Li, F., Orban, R., and Baker, B. (2012a). SoMART: a web server for plant miRNA, tasiRNA and target gene analysis. Plant J. **70**: 891–901.

Li, F., Pignatta, D., Bendix, C., Brunkard, J.O., Cohn, M.M., Tung, J., Sun, H., Kumar, P., and Baker, B. (2012b). MicroRNA regulation of plant innate immune receptors. Proc. Natl. Acad. Sci. U. S. A. **109**: 1790–1795.

Li, Y.-F., Zheng, Y., Addo-Quaye, C., Zhang, L., Saini, A., Jagadeeswaran, G., Axtell, M.J., Zhang, W., and Sunkar, R. (2010). Transcriptome-wide identification of microRNA targets in rice. Plant J. **62**: 742–759.

Li, Y.-F., Zheng, Y., Jagadeeswaran, G., and Sunkar, R. (2013). Characterization of small RNAs and their target genes in wheat seedlings using sequencing-based approaches. Plant Sci. **203–204**: 17–24.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., and Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature **456**: 464–469.

Lister, R., Gregory, B.D., and Ecker, J.R. (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr. Opin. Plant Biol. **12**: 107–118.

Liu, B., Chen, Z., Song, X., Liu, C., Cui, X., Zhao, X., Fang, J., Xu, W., Zhang, H., Wang, X., Chu, C., Deng, X., Xue, Y., and Cao, X. (2007). Oryza sativa Dicer-like4 Reveals a Key Role for Small Interfering RNA Silencing in Plant Development. Plant Cell Online **19**: 2705–2718.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. Biomed Res. Int. **2012**.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of Scarecrow-like mRNA Targets Directed by a Class of Arabidopsis miRNA. Science **297**: 2053–2056.

Lu, C., Jeong, D.-H., Kulkarni, K., Pillay, M., Nobuta, K., German, R., Thatcher, S.R., Maher, C., Zhang, L., Ware, D., Liu, B., Cao, X., Meyers, B.C., and Green, P.J. (2008). Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). Proc. Natl. Acad. Sci. U. S. A. **105**: 4951–4956.

Lu, C., Kulkarni, K., Souret, F.F., MuthuValliappan, R., Tej, S.S., Poethig, R.S., Henderson, I.R.,

Jacobsen, S.E., Wang, W., Green, P.J., and Meyers, B.C. (2006). MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. Genome Res. **16**: 1276–1288.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C., and Green, P.J. (2005). Elucidation of the Small RNA Component of the Transcriptome. Science **309**: 1567–1569.

Luo, Q.-J., Mittal, A., Jia, F., and Rock, C.D. (2012). An autoregulatory feedback loop involving PAP1 and TAS4 in response to sugars in Arabidopsis. Plant Mol. Biol. **80**: 117–129.

Ma, Z., Coruh, C., and Axtell, M.J. (2010). Arabidopsis lyrata Small RNAs: Transient MIRNA and Small Interfering RNA Loci within the Arabidopsis Genus. Plant Cell **22**: 1090–1103.

Maher, C., Stein, L., and Ware, D. (2006). Evolution of Arabidopsis microRNA families through duplication events. Genome Res. **16**: 510–519.

Mallory, A.C. and Bouché, N. (2008). MicroRNA-directed regulation: to cleave or not to cleave. Trends Plant Sci. **13**: 359–367.

Mallory, A.C., Reinhart, B.J., Jones-Rhoades, M.W., Tang, G., Zamore, P.D., Barton, M.K., and Bartel, D.P. (2004). MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5′ region. Embo J. **23**: 3356–3364.

Malone, C.D., Brennecke, J., Dus, M., Stark, A., McCombie, W.R., Sachidanandam, R., and Hannon, G.J. (2009). Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. Cell **137**: 522–535.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., and Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. Nat. Methods **7**: 111–118.

Mao, W., Li, Z., Xia, X., Li, Y., and Yu, J. (2012). A Combined Approach of High-Throughput Sequencing and Degradome Analysis Reveals Tissue Specific Expression of MicroRNAs and Their Targets in Cucumber. Plos One **7**: e33040.

Mardis, E.R. (2008a). Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet. **9**: 387–402.

Mardis, E.R. (2008b). The impact of next-generation sequencing technology on genetics. Trends Genet. **24**: 133–141.

Margis, R., Fusaro, A.F., Smith, N.A., Curtin, S.J., Watson, J.M., Finnegan, E.J., and Waterhouse, P.M. (2006). The evolution and diversification of Dicers in plants. Febs Lett. **580**: 2442–2450.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**: 376–380.

Martin, J.A. and Wang, Z. (2011). Next-generation transcriptome assembly. Nat. Rev. Genet. **12**: 671–682.

McKernan, K.J., Peckham, H.E., Costa, G., McLaughlin, S., Tsung, E., Fu, Y., Clouser, C., Dunkan, C., Ichikawa, J., Lee, C., Zhang, Z., Sheridan, A., Fu, H., Ranade, S., Dimilanta, E., Sokolsky, T., Zhang, L., Hendrickson, C., Li, B., Kotler, L., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. Genome Res. **19**: 1527–1541.

Megraw, M., Baev, V., Rusinov, V., Jensen, S.T., Kalantidis, K., and Hatzigeorgiou, A.G. (2006). MicroRNA Promoter Element Discovery in Arabidopsis. RNA **12**: 1612–1619.

Mehrpooyan, F., Othman, R.Y., and Harikrishna, J.A. (2012). Tissue and temporal expression of miR172 paralogs and the AP2-like target in oil palm (Elaeis guineensis Jacq.). Tree Genet. Genomes **8**: 1331–1343.

Mette, M.F., Winden, J. van der, Matzke, M., and Matzke, A.J.M. (2002). Short RNAs Can Identify New Candidate Transposable Element Families in Arabidopsis. Plant Physiol. **130**: 6–9.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J., Griffiths-Jones, S., Jacobsen, S.E., Mallory, A.C., Martienssen, R.A., Poethig, R.S., Qi, Y., Vaucheret, H., Voinnet, O., Watanabe, Y., Weigel, D., et al. (2008). Criteria for Annotation of Plant MicroRNAs. Plant Cell Online **20**: 3186–3190.

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., Chen, S., Hannon, G.J., and Qi, Y. (2008). Sorting of Small RNAs into Arabidopsis Argonaute Complexes Is Directed by the 5′ Terminal Nucleotide. Cell **133**: 116–127.

Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S.L., Nekrutenko, A., Giardine, B., Harris, R.S., Tyekucheva, S., Diekhans, M., Pringle, T.H., Murphy, W.J., Lesk, A., Weinstock, G.M., et al. (2007). 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. **17**: 1797–1808.

Mitra, R.D., Shendure, J., Olejnik, J., Edyta-Krzymanska-Olejnik, and Church, G.M. (2003). Fluorescent in situ sequencing on polymerase colonies. Anal. Biochem. **320**: 55–65.

Molnár, A., Schwach, F., Studholme, D.J., Thuenemann, E.C., and Baulcombe, D.C. (2007). miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. Nature **447**: 1126–1129.

Montgomery, T.A., Howell, M.D., Cuperus, J.T., Li, D., Hansen, J.E., Alexander, A.L., Chapman, E.J., Fahlgren, N., Allen, E., and Carrington, J.C. (2008). Specificity of ARGONAUTE7-miR390 Interaction and Dual Functionality in TAS3 Trans-Acting siRNA Formation. Cell **133**: 128–141.

Morin, R.D., Aksay, G., Dolgosheina, E., Ebhardt, H.A., Magrini, V., Mardis, E.R., Sahinalp, S.C., and Unrau, P.J. (2008). Comparative analysis of the small RNA transcriptomes of Pinus contorta and Oryza sativa. Genome Res. **18**: 571–584.

Mosher, R.A., Melnyk, C.W., Kelly, K.A., Dunn, R.M., Studholme, D.J., and Baulcombe, D.C. (2009). Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. Nature **460**: 283–286.

Mosher, R.A., Schwach, F., Studholme, D., and Baulcombe, D.C. (2008). PolIVb influences RNA-

directed DNA methylation independently of its role in siRNA biogenesis. Proc. Natl. Acad. Sci. U. S. A. **105**: 3145–3150.

Motamayor, J.C., Risterucci, A.M., Lopez, P.A., Ortiz, C.F., Moreno, A., and Lanaud, C. (2002). Cacao domestication I: the origin of the cacao cultivated by the Mayas. Heredity **89**: 380–386.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. Plant Cell **2**: 279–289.

Newman, M.A. and Hammond, S.M. (2010). Emerging paradigms of regulated microRNA processing. Genes Dev. **24**: 1086 –1092.

Nobuta, K., Lu, C., Shrivastava, R., Pillay, M., Paoli, E.D., Accerbi, M., Arteaga-Vazquez, M., Sidorenko, L., Jeong, D.-H., Yen, Y., Green, P.J., Chandler, V.L., and Meyers, B.C. (2008). Distinct size distribution of endogenous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. Proc. Natl. Acad. Sci. U. S. A. **105**: 14958–14963.

Nodine, M.D. and Bartel, D.P. (2010). MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. Genes Dev. **24**: 2678–2692.

Nogueira, F.T.S., Chitwood, D.H., Madi, S., Ohtsu, K., Schnable, P.S., Scanlon, M.J., and Timmermans, M.C.P. (2009). Regulation of Small RNA Accumulation in the Maize Shoot Apex. Plos Genet. **5**: e1000320.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., and Zwick, M.E. (2007). Microarray-based genomic selection for high-throughput resequencing. Nat. Meth. **4**: 907–909.

Olmedo-Monfil, V., Durán-Figueroa, N., Arteaga-Vázquez, M., Demesa-Arévalo, E., Autran, D., Grimanelli, D., Slotkin, R.K., Martienssen, R.A., and Vielle-Calzada, J.-P. (2010). Control of female gamete formation by a small RNA pathway in Arabidopsis. Nature **464**: 628–632.

Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. Nature **425**: 257–263.

Pantaleo, V., Szittya, G., Moxon, S., Miozzi, L., Moulton, V., Dalmay, T., and Burgyan, J. (2010). Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. Plant J. **62**: 960–976.

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

Park, W., Li, J., Song, R., Messing, J., and Chen, X. (2002). CARPEL FACTORY, a Dicer Homolog, and HEN1, a Novel Protein, Act in microRNA Metabolism in Arabidopsis thaliana. Curr. Biol. **12**: 1484–1495.

Piriyapongsa, J. and Jordan, I.K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. RNA **14**: 814–821.

Poethig, R.S. (2009). Small RNAs and developmental timing in plants. Curr. Opin. Genet. Dev. **19**: 374–378.

Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., Gao, Y., Church, G.M., and Shendure, J. (2007). Multiplex amplification of large sets of human exons. Nat. Meth. **4**: 931–936.

Price Z, Mayes S, Billotte N, Hafeez F, Dumortier F, and MacDonald D (2007). Oil palm. In: Kole C (ed) Genome mapping and molecular breeding in plants, Volume 6 Technical Crops. In (Springer, Berlin), pp. 93–108.

Prud'homme, N., Gans, M., Masson, M., Terzian, C., and Bucheton, A. (1995). Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. Genetics **139**: 697–711.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. Bmc Genomics **13**: 341.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev. **20**: 3407.

Rapid amplification of 5′ complementary DNA ends (5′ RACE) (2005). Nat. Meth. **2**: 629–630.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. Genes Dev. **16**: 1616–1626.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. **16**: 276–277.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O.L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Meth. **4**: 651–657.

Ron, M., Saez, M.A., Williams, L.E., Fletcher, J.C., and McCormick, S. (2010). Proper regulation of a sperm-specific cis-nat-siRNA is essential for double fertilization in Arabidopsis. Genes Dev. **24**: 1010–1021.

Ronemus, M., Vaughn, M.W., and Martienssen, R.A. (2006). MicroRNA-Targeted and Small Interfering RNA–Mediated mRNA Degradation Is Regulated by Argonaute, Dicer, and RNA-Dependent RNA Polymerase in Arabidopsis. Plant Cell Online **18**: 1559–1574.

Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell **75**: 843–854.

Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A., and Brudno, M. (2009). SHRiMP: Accurate Mapping of Short Color-space Reads. Plos Comput. Biol. **5**: e1000386.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., Marçais, G., Pop, M., and Yorke, J.A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. **22**: 557–567.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. **74**: 5463–5467.

Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., and Nolan, G.P. (2010). Computational solutions to large-scale data management and analysis. Nat. Rev. Genet. **11**: 647–657.

Schwab, R., Palatnik, J.F., Riester, M., Schommer, C., Schmid, M., and Weigel, D. (2005). Specific effects of microRNAs on the plant transcriptome. Dev. Cell **8**: 517–527.

Shamimuzzaman, M. and Vodkin, L. (2012). Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing. Bmc Genomics **13**: 310.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science **309**: 1728–1732.

Shin, C., Nam, J.-W., Farh, K.K.-H., Chiang, H.R., Shkumatava, A., and Bartel, D.P. (2010). Expanding the MicroRNA Targeting Code: Functional Sites with Centered Pairing. Mol. Cell **38**: 789–802.

Shivaprasad, P.V., Chen, H.-M., Patel, K., Bond, D.M., Santos, B.A.C.M., and Baulcombe, D.C. (2012). A MicroRNA Superfamily Regulates Nucleotide Binding Site–Leucine-Rich Repeats and Other mRNAs. Plant Cell Online.

Sieber, P., Wellmer, F., Gheyselinck, J., Riechmann, J.L., and Meyerowitz, E.M. (2007). Redundancy and specialization among plant microRNAs: role of the MIR164 family in developmental robustness. Development **134**: 1051–1060.

Simon, S.A., Zhai, J., Nandety, R.S., McCormick, K.P., Zeng, J., Mejia, D., and Meyers, B.C. (2009). Short-Read Sequencing Technologies for Transcriptional Analyses. Annu. Rev. Plant Biol. **60**: 305–333.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdžić, M., Becker, J.D., Feijó, J.A., and Martienssen, R.A. (2009). Epigenetic Reprogramming and Small RNA Silencing of Transposable Elements in Pollen. Cell **136**: 461–472.

Smith, D.F., Peacock, C.S., and Cruz, A.K. (2007). Comparative genomics: From genotype to disease phenotype in the leishmaniases. Int. J. Parasitol. **37**: 1173–1186.

Sorefan, K., Pais, H., Hall, A.E., Kozomara, A., Griffiths-Jones, S., Moulton, V., and Dalmay, T. (2012). Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence **3**: 4.

Souret, F.F., Kastenmayer, J.P., and Green, P.J. (2004). AtXRN4 Degrades mRNA in Arabidopsis and Its Substrates Include Selected miRNA Targets. Mol. Cell **15**: 173–183.

Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., Ruby, J.G., Brennecke, J., Hodges, E., Hinrichs, A.S., Caspi, A., Paten, B., Park, S.-W., Han, M.V., Maeder, M.L., Polansky, B.J., et al. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature **450**: 219–232.

Stefani, G. and Slack, F.J. (2008). Small non-coding RNAs in animal development. Nat. Rev. Mol. Cell Biol. **9**: 219–230.

Subramanian, S., Fu, Y., Sunkar, R., Barbazuk, W.B., Zhu, J.-K., and Yu, O. (2008). Novel and nodulation-regulated microRNAs in soybean roots. Bmc Genomics **9**: 160.

Sunkar, R., Zhou, X., Zheng, Y., Zhang, W., and Zhu, J.-K. (2008). Identification of novel and candidate miRNAs in rice by high throughput sequencing. Bmc Plant Biol. **8**: 25.

Szittya, G., Moxon, S., Santos, D.M., Jing, R., Fevereiro, M.P., Moulton, V., and Dalmay, T. (2008). High-throughput sequencing of Medicago truncatula short RNAs identifies eight new miRNA families. Bmc Genomics **9**: 593.

Takeda, A., Iwasaki, S., Watanabe, T., Utsumi, M., and Watanabe, Y. (2008). The Mechanism Selecting the Guide Strand from Small RNA Duplexes is Different Among Argonaute Proteins. Plant Cell Physiol. **49**: 493–500.

Talmor-Neiman, M., Stav, R., Klipcan, L., Buxdorf, K., Baulcombe, D.C., and Arazi, T. (2006). Identification of trans-acting siRNAs in moss and an RNA-dependent RNA polymerase required for their biogenesis. Plant J. **48**: 511–521.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**: 512–526.

Teer, J.K., Bonnycastle, L.L., Chines, P.S., Hansen, N.F., Aoyama, N., Swift, A.J., Abaan, H.O., Albert, T.J., Margulies, E.H., Green, E.D., Collins, F.S., Mullikin, J.C., and Biesecker, L.G. (2010). Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome Res. **20**: 1420–1431.

Timp, W., Mirsaidov, U.M., Wang, D., Comer, J., Aksimentiev, A., and Timp, G. (2010). Nanopore Sequencing: Electrical Measurements of the Code of Life. Ieee Trans. Nanotechnol. **9**: 281–294.

Válóczi, A., Várallyay, É., Kauppinen, S., Burgyán, J., and Havelda, Z. (2006). Spatio-temporal accumulation of microRNAs is highly coordinated in developing plant tissues. Plant J. **47**: 140–151.

Vaucheret, H. (2008). Plant ARGONAUTES. Trends Plant Sci. **13**: 350–358.

Vaughn, M.W., Tanurdzic, M., Lippman, Z., Jiang, H., Carrasquillo, R., Rabinowicz, P.D., Dedhia, N., McCombie, W.R., Agier, N., Bulski, A., Colot, V., Doerge, R.., and Martienssen, R.A. (2007). Epigenetic Natural Variation in Arabidopsis thaliana. Plos Biol. **5**: e174.

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

Voinnet, O. (2009). Origin, Biogenesis, and Activity of Plant MicroRNAs. Cell **136**: 669–687.

Warthmann, N., Das, S., Lanz, C., and Weigel, D. (2008). Comparative Analysis of the MIR319a MicroRNA Locus in Arabidopsis and Related Brassicaceae. Mol. Biol. Evol. **25**: 892–902.

Wee, L.M., Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2012). Argonaute Divides Its RNA Guide into Domains with Distinct Functions and RNA-Binding Properties. Cell **151**: 1055–1067.

Wierzbicki, A.T., Cocklin, R., Mayampurath, A., Lister, R., Rowley, M.J., Gregory, B.D., Ecker,

J.R., Tang, H., and Pikaard, C.S. (2012). Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. Genes Dev. **26**: 1825–1836.

Wierzbicki, A.T., Ream, T.S., Haag, J.R., and Pikaard, C.S. (2009). RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nat. Genet. **41**: 630–634.

Wu, G., Park, M.Y., Conway, S.R., Wang, J.-W., Weigel, D., and Poethig, R.S. (2009). The Sequential Action of miR156 and miR172 Regulates Developmental Timing in Arabidopsis. Cell **138**: 750–759.

Wu, G. and Poethig, R.S. (2006). Temporal regulation of shoot development in Arabidopsis thaliana by miR156 and its target SPL3. Development **133**: 3539–3547.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. (2010). DNA Methylation Mediated by a MicroRNA Pathway. Mol. Cell **38**: 465–475.

Wu, M.-F., Tian, Q., and Reed, J.W. (2006). Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. Development **133**: 4211–4218.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005a). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature **434**: 338–345.

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., and Carrington, J.C. (2005b). Expression of Arabidopsis MIRNA Genes. Plant Physiol. **138**: 2145–2154.

Xu, M.Y., Dong, Y., Zhang, Q.X., Zhang, L., Luo, Y.Z., Sun, J., Fan, Y.L., and Wang, L. (2012). Identification of miRNAs and their targets from Brassica napus by high-throughput sequencing and degradome analysis. Bmc Genomics **13**: 421.

Yoshikawa, M., Peragine, A., Park, M.Y., and Poethig, R.S. (2005). A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes Dev. **19**: 2164–2175.

Yu, B., Bi, L., Zheng, B., Ji, L., Chevalier, D., Agarwal, M., Ramachandran, V., Li, W., Lagrange, T., Walker, J.C., and Chen, X. (2008). The FHA domain proteins DAWDLE in Arabidopsis and SNIP1 in humans act in small RNA biogenesis. Proc. Natl. Acad. Sci. U. S. A. **105**: 10073–10078.

Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R.W., Steward, R., and Chen, X. (2005). Methylation as a Crucial Step in Plant microRNA Biogenesis. Science **307**: 932–935.

Zerbino, D.R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**: 821–829.

Zhai, J., Jeong, D.-H., De Paoli, E., Park, S., Rosen, B.D., Li, Y., González, A.J., Yan, Z., Kitto, S.L., Grusak, M.A., Jackson, S.A., Stacey, G., Cook, D.R., Green, P.J., Sherrier, D.J., and 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 Dev. **25**: 2540–2553.

Zhai, J., Liu, J., Liu, B., Li, P., Meyers, B.C., Chen, X., and Cao, X. (2008). Small RNA-Directed

Epigenetic Natural Variation in Arabidopsis thaliana. Plos Genet. **4**: e1000056.

Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P., and Anderson, T.A. (2006). Conservation and divergence of plant microRNA genes. Plant J. **46**: 243–259.

Zhang, X., Henderson, I.R., Lu, C., Green, P.J., and Jacobsen, S.E. (2007). Role of RNA polymerase IV in plant small RNA metabolism. Proc. Natl. Acad. Sci. U. S. A. **104**: 4536–4541.

Zhang, X., Xia, J., Lii, Y.E., Barrera-Figueroa, B.E., Zhou, X., Gao, S., Lu, L., Niu, D., Chen, Z., Leung, C., Wong, T., Zhang, H., Guo, J., Li, Y., Liu, R., Liang, W., Zhu, J.-K., Zhang, W., and Jin, H. (2012). Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. Genome Biol. **13**: R20.

Zhang, Z., Wei, L., Zou, X., Tao, Y., Liu, Z., and Zheng, Y. (2008). Submergence-responsive MicroRNAs are Potentially Involved in the Regulation of Morphological and Metabolic Adaptations in Maize Root Cells. Ann. Bot. **102**: 509–519.

Zhao, T., Li, G., Mi, S., Li, S., Hannon, G.J., Wang, X.-J., and Qi, Y. (2007). A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. Genes Dev. **21**: 1190–1203.

Zhong, X., Hale, C.J., Law, J.A., Johnson, L.M., Feng, S., Tu, A., and Jacobsen, S.E. (2012). DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. Nat. Struct. Mol. Biol. **19**: 870–875.

Zhou, M., Gu, L., Li, P., Song, X., Wei, L., Chen, Z., and Cao, X. (2010). Degradome sequencing reveals endogenous small RNA targets in rice (Oryza sativa L. ssp. indica). Front. Biol. **5**: 67–90.

Zhou, X., Ruan, J., Wang, G., and Zhang, W. (2007). Characterization and Identification of MicroRNA Core Promoters in Four Model Species. Plos Comput. Biol. **3**: e37.

Zhu, Q.-H., Spriggs, A., Matthew, L., Fan, L., Kennedy, G., Gubler, F., and Helliwell, C. (2008). A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. Genome Res. **18**: 1456–1465.

Zong, J., Yao, X., Yin, J., Zhang, D., and Ma, H. (2009). Evolution of the RNA-dependent RNA polymerase (RdRP) genes: Duplications and possible losses before and after the divergence of major eukaryotic groups. Gene **447**: 29–39.

VITA

Zhaorong Ma mazhaorong@gmail.com

Education

The Pennsylvania State University University Park, PA, USA

Ph.D. Integrative Biosciences (Bioinformatics and Genomics option)

Fudan University Shanghai, China

B.S. Biological Science

Publications

Ma Z, Axtell MJ. A novel targeted genomic enrichment method enables assembly of unknown genomic regions flanking a known core sequence. Manuscript in preparation.

Arif MA, Fattash I, **Ma Z**, Cho SH, Beike AK, Reski R, Axtell MJ, Frank W (2012). DICER-LIKE3 Activity in *Physcomitrella patens* DICER-LIKE4 Mutants Causes Severe Developmental Dysfunction and Sterility. Molecular Plant 5 (6): 1281-1294

Argout X, ..., **Ma Z** (author 37 out of 61), ..., Lanaud C (2011). The genome of *Theobroma cacao*. Nature Genetics 43: 101- 108. Epub 2010

Ma Z, Coruh C, Axtell MJ (2010). *Arabidopsis lyrata* small RNAs: Transient *MIRNA* and siRNA loci within the *Arabidopsis* genus. Plant Cell 22: 1090-1103

Sun R, Fu X, Guo F, **Ma Z**, Goulbourne C, Jiang M, Li Y, Xie Y, Mao Y (2009). A strategy for meta- analysis of short time series microarray datasets. Frontiers in bioscience: a journal and virtual library 14: 4058-4070

Cho SH, Addo-Quaye C, Coruh C, Asif MA, **Ma Z**, Frank W, Axtell MJ (2008). *Physcomitrella patens* DCL3 is required for 22-24nt siRNA accumulation, suppression of retrotransposon-derived transcripts, and normal development. PLoS Genetics 4: e1000314