Data tell the truth: microRNA-mediated gene networks in plants

Xiaozeng Yang
Wutai, Shanxi, China

B.S., China Agricultural University, 2004
M.S. (Associate), Beijing University of Technology, 2008

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Biology

University of Virginia
May, 2014

_____

_____

_____

_____

_____

Abstract

Temporal and spatial control of transcript abundance for expressed genes is crucial for many biological processes and developmental programs. In eukaryotes, gene expression regulation occurs at many levels. At the post-transcriptional level, microRNAs (miRNAs) are emerging as an important class of sequence-specific, *trans*-acting endogenous small RNA molecules that modulate gene expression. It is well-established that many miRNAs are crucial for diverse plant development processes and responses to environmental challenges. However, three fundamental and inherently related questions wait answers: 1) how many miRNAs are there in plant species; 2) how does regulation from miRNAs cooperate with other gene regulation mechanisms; and 3) how does miRNAs as a group play roles in a specific pathway. This dissertation uses a combination of cutting-edge bioinformatic and experimental approaches to address these fundamental questions at a genomic level.

Ultra deep sampling of small RNA libraries by next generation sequencing has provided rich information on the miRNA transcriptome of various plant species. However, few computational tools have been developed to effectively de-convolute the complex information. I sought to employ the signature distribution of small RNA reads along the miRNA precursor as a model in plants to profile expression of known miRNA genes and to identify novel ones. A freely available package, miRDeep-P, was developed subsequently, which could comprehensively and accurately identify miRNAs from deeply sequenced small RNA libraries. Taking advantage of this method, I have examined miRNAs in 15 plant species and identified thousands of miRNAs. With this unprecedented dataset, several exciting findings were made, including that the number of

miRNAs is strongly correlated with the number of protein-coding genes, and that plant miRNAs in a compact organization suggest several possible miRNA biogenesis mechanisms utilizing existing precursors, and many other clues on miRNA evolution.

Toward an understanding of the cooperation of miRNA-centered network with other gene regulation mechanisms, I investigated and discovered a novel genetic mechanism that links miRNAs with alternative mRNA splicing. By compiling a set of miRNA target genes and going through millions of pieces of RNA sequencing data, it was found for the first time in plants that mRNA isoforms produced by alternative splicing differ in the sequences encoding the miRNA binding sites. In collaboration with a postdoctoral researcher in my lab, it has been functionally shown that these alternative splicing events are relevant in controlling plant development.

Another piece of effort in this dissertation is to perform genetic experiments to test how miRNAs as a group interact with other pathways in response to external stimuli. Systemin-mediated pathway in tomato was selected as it has been largely understood in last decade. Genetic experiments combining with high-throughput data analyses indicate that changes in systemin-mediated pathway could affect the expression of many miRNAs and that a large number of miRNAs, in turn, are involved in this pathway. Examination of phenotypic changes in several transgenic lines further suggests that many miRNAs as a group potentially provides a buffering role when plants meet an intense stimulation.

Acknowledgements

A highly collaborative effort gives birth to this dissertation, and I would like to thank many people for leading me through this process. I am most grateful to my advisor, Dr. Lei Li, whose support has been throughout the whole process. His encouragement has helped me not only solve many complicated issues in my research, also aided me to overcome a mental difficulty during my first two years in my PhD program. I have been always inspired by having a conversation with him when I seriously doubted whether I would get to this point.

I also would like to give my thanks to the members of my research committee, Dr. Michael Timko, Dr. Herman Wijnen, Dr. Martin Wu, Dr. Douglas Taylor, Dr. Stefan Bekiranov and Dr. Benjamin Blackman. At each annual report meeting, they always give me many valuable suggestions. Meanwhile, I really appreciated all of help from past and current labmates, Dr. Chengjun Wu, Dr. Huiyong Zhang, Xin Zhao, Chun Su, Zhang Wang, Matthew Pahl. They have been making the lab a pleasurable and productive environment to work in. Among them, I am especially thankful to Dr. Chengjun Wu, and Dr. Huiyong Zhang for teaching me and collaborating with me in many bench works, and to Zhang for his help with many pieces of phylogenetic analysis. The execution of my dissertation research also depended on many other people in biology depart at UVa. There are too many to name them all, but I would like to specifically thank Dr. Mark Kopeny, Joanne Chaplin, Wendy Crannage. Collaborations also extended to outside of UVa. Dr. Ray Ming, Dr. Xiangfeng Bryan Wang, Dr. Yimiao Tang, Dr. Changpin Zhao, have all made important contributions to individual projects.

During my PhD study, I felt lucky and was very grateful to the institutions that rewarded me fellowships and scholarships to support my research. For instance, I was fortunate enough to receive the dissertation year fellowship from my department and China National Scholarship for Outstanding Students Studying Abroad from China Scholarship Council.

Finally, I would like to thank my family both here and in China. Their unconditional support and encouragement has helped me through my PhD program. My wife, Ying Wang, has been a constant source of love, support and motivation, and she is always being there, both in life and science, which I am deeply grateful. My parents, they have given me so much. They have left me with a debt that I can never repay but will remember forever.

Table of contents

Chapter 1. General introduction

The general introduction embraces the review paper, analyzing the microRNA transcriptome in plants using deep sequencing data[1], and other two parts.

[1]Formatted as a co-authored manuscript and published as:

Yang X, Lei Li. 2012. *Biology*.1:297-310.

One of the most exciting biological finding in recent years is the discovery of many functional small RNA species that regulate diverse spatial and temporal function of the genome (Huttenhofer et al., 2005; Bartel, 2009; Carthew and Sontheimer, 2009; Voinnet, 2009). After the initial discovery of miRNAs in the worm C. elegans (Lee et al., 1993; Wightman et al., 1993), they are emerging as an important class of endogenous gene regulators acting at the post-transcriptional level in both animals and plants. In plants, much of the effort to identify, experimentally validate, and functionally characterize miRNAs has been directed toward the model plant Arabidopsis thaliana. Consequently, dozens of miRNA-target pairs have been identified and studied (Jones-Rhoades et al., 2006; Fahlgren et al., 2007; Alves et al., 2009). It is now well-established that these gene circuits are crucial for many plant development processes as well as responses to environmental challenges (Jones-Rhoades et al., 2006; Garcia, 2008; Voinnet, 2009).

Although hundreds of miRNAs have been predicted in a broad range of plant lineages (Kozomara and Griffiths-Jones, 2011), there are two indications that current miRNA collections in many model plants and important crop species are far from completion. First of all, the numbers of predicted miRNAs among different plant species are conspicuously uneven. As shown in Figure 1, the well-annotated *Arabidopsis* and rice (*Oryza sativa*) genomes contain approximately one miRNA for every 100 protein-coding genes. In other plant species, the relative density of miRNAs is only half or even less than that in *Arabidopsis* and rice (Figure 1). Because it is highly unlikely that these species indeed encode a smaller complement of miRNA genes, the only explanation is that most miRNAs in species other than *Arabidopsis* and rice still await discovery.

In plants, various studies have established that there are about 20 families of conserved miRNAs (Nobuta et al., 2007; Cuperus et al., 2011). In many plant species only miRNAs belonging to conserved families are identified. However, the sequencing of small RNA populations is increasingly revealing miRNAs that are not conserved between species, suggesting a recent evolutionary origin (Rajagopalan et al., 2006; Fahlgren et al., 2007; Molnar et al., 2007; Zhu et al., 2008; Cuperus et al., 2011). In fact, there is increasing evidence that species-specific or subfamily-specific miRNAs are functional constituents of the miRNA-mediated regulatory networks and underscore the dynamic nature of these networks (Lu et al., 2011; Ng et al., 2011; Wang et al., 2011). Thus, it is highly desirable to elucidate the full spectrum of miRNAs in diverse plant lineages to gain a comprehensive understanding of miRNA origin, evolution and function.

Brief overview of miRNA biogenesis in plants

There is no question that any systematic effort to identify miRNAs will dependent on a clear understanding of the miRNA biogenesis pathways. Since miRNA biogenesis is the subject of numerous reviews, we only provide a brief overview here (Figure 2). Like protein-coding genes, miRNAs are encoded by class II genes and transcribed by RNA polymerase II (Pol II) (Tam, 2001; Lee et al., 2004). Although mature miRNAs are typically 20- to 24-nucleotides (nt) in length, their precursor transcripts can be much longer. As shown in Figure 2, after initially transcription by Pol II, splicing and further processing of pri-miRNAs are carried out in nuclear and involve the interactive functions of HYL1 and SE, as well as the cap-binding proteins CBP20 and CBP80 (Laubinger et al., 2008; Yu et al., 2008; Ren et al., 2012).

A characteristic of pri-miRNAs is that they contain internally complementary sequences that fold back to form a hairpin structure, which is called pre-miRNAs (Bartel, 2004; Chen, 2005; Bartel, 2009; Voinnet, 2009). Pri-miRNAs and pre-miRNAs are sequentially processed by Dicer to yield one or several phased miRNA/miRNA* duplexes. Unique to higher plants, pri-miRNA and pre-miRNA processing are both carried out in the nucleus (Papp et al., 2003). The duplexes are stabilized through end methylation catalyzed by HEN1 (Yu et al., 2008) and transported to the cytoplasm by HST1 (Ren et al., 2012). Only the mature miRNA is integrated into the AGO1-containing RNA-induced silencing complex (RISC) whereas the passenger strand, called RNA*, is degraded as a RISC substrate (Khvorova et al., 2003; Schwarz et al., 2003; Bartel, 2009; Motomura et al., 2012). After loading into RISC, miRNAs base pair with their targets and direct either cleavage (Llave et al., 2002; Reinhart et al., 2002) or translational repression (Brodersen et al., 2008) of the target transcripts. Recently, silencing of target genes by miRNA-directed DNA methylation at the target loci has also been reported (Wu et al., 2010).

As a consequence of miRNA maturation, a series of RNA intermediates are generated in addition to the mature miRNAs, which include the stem-loop-structured pre-miRNAs, miRNA* and sliced RNA fragments derived from other parts of the precursors. Detection, quantification, and reconstruction of these RNA intermediates are the goal of essentially all available methods to identify and profile miRNAs.

Comprehensive identification of miRNAs from next-generation sequencing data

Direct sequencing of specifically prepared low-molecular-weight RNA has long been recognized as powerful approach to sample the small RNA species (Lu et al., 2005). Typically, small RNA species in the 18-35 nucleotides range are isolated and ligated to the 5' and 3' RNA adapters. The ligated RNA molecules are reverse-transcribed into cDNA using a primer specific to the 5' adapter and amplified by PCR with two primers that anneal to the ends of the adapters. Quality controlled cDNA libraries are then sequenced (Lu et al., 2005; Fahlgren et al., 2007). In one of the earliest studies, Fahlgren et al. (Fahlgren et al., 2007) sequenced small RNA populations from wild-type *Arabidopsis* as well several mutants defective in miRNA biogenesis using the 454 technology. A total of 48 non-conserved miRNA families were identified by a computational analysis of sequence composition and secondary structure based on knowledge of annotated miRNAs at that time (Fahlgren et al., 2007).

Encouraged by success in *Arabidopsis* (Rajagopalan et al., 2006; Fahlgren et al., 2007), deep sampling of small RNA libraries by next-generation sequencing has become a popular approach to identify miRNAs for functional and evolutionary studies in diverse plant species (Moxon et al., 2008b; Sunkar et al., 2008; Zhu et al., 2008). An advantage of the sequencing methods is their sensitivity in detecting even poorly expressed or species-specific miRNAs (Sunkar et al., 2008). The potential of deep sequencing to provide quantitative information on the expression pattern of known miRNAs has been explored (Fahlgren et al., 2007; Creighton et al., 2009). In addition to validating annotated miRNAs, large numbers of putative new miRNAs have been identified. Following we discuss various recently developed algorithms and programs to profile miRNAs from deep sequencing data, with an emphasis on their applications in plants.

Available tools for analyzing miRNAs from deep sequencing data

A number of computational tools for identifying and profiling miRNAs from deep sequencing data have been developed. With an easy-to-use graphical interface, most of these tools are web-based while a few, such as miRDeep (Friedlander et al., 2008) and miRNA Key (Ronen et al., 2010), are also packaged into a stand-alone version (Table 1). A common module employed by these tools is sequence similarity search to detect miRNAs cross multiple species based on the fact that many miRNAs are evolutionarily conserved. Meanwhile, other algorithms are also introduced to detect new miRNAs based on different models in terms of the pre-miRNA hairpin structures and the duplex of miRNA and miRNA* (Table 1). The challenge now is to separate miRNAs from the pool of other sequenced small RNAs or mRNA degradation products. Further, as most of the miRNA-detecting methods focus exclusively on the mature miRNAs, a drawback is the failure to collect and quantify information on the precursors, which could result in limitations in elucidating the miRNA transcriptome.

miRanalyzer

Utilizing a machine learning algorithm, Hackenberg et al. (Hackenberg et al., 2009) developed miRanalyzer, a web server tool for analyzing results from deep-sequencing experiments on small RNAs. This program requires a simple input file containing a list of unique reads and their copy numbers. Application of this program in seven animal model species (human, mouse, rat, fruit fly, round-worm, zebra fish and dog) not only detected known miRNA sequences annotated in miRBase, but led to prediction of new miRNAs. The core algorithm of miRanalyzer is based on the random

forest classifier and was trained on experimental data, which could accurately predict novel miRNAs with a low false positive rate in animals. Later, miRanalyzer was updated to include a module on miRNA prediction in plants by taking into account differences between plant and animal miRNAs. Currently, 31 genomes, including 6 plant genomes, have been analyzed by the updated miRanalyzer (Hackenberg et al., 2011).

UEA sRNA Toolkit

UEA sRNA Toolkit (Moxon et al., 2008a) combines two integrated parts, miRCat and SiLoCo, to analyze miRNAs using deep sequencing data. miRCat, the package for detecting miRNAs, adopts a number of empirical and published criteria for *bona fide* miRNA loci to mine miRNAs from deeply sequenced small RNA data. In brief, the program accepts a FASTA file of small RNA sequences as input, which are mapped to a plant genome using PatMaN (Prufer et al., 2008) and grouped into discrete loci. Then it obtains miRNA candidates by searching for a two-peak alignment pattern of sequence reads on one strand of the locus and assessing the secondary structures of a series of putative precursor transcripts using RNAfold (Denman, 1993) and randfold (Bonnet et al., 2004). On the other hand, SiLoCo is the tool to compare the miRNA expression between different samples. It weighs each small RNA hit by its repetitiveness in the genome acquired from mapping by PatMaN (Prufer et al., 2008). For each locus, the log2 ratio and the average of the normalized small RNA hit counts are used to calculate the miRNA expression difference (Moxon et al., 2008a).

miRDeep

Maturation of miRNAs from the stem-loop structured pre-miRNAs results in three species of small RNAs: mature miRNA, miRNA* and RNA fragments derived from other parts of the precursors. Typically the mature miRNAs are much stable, which results in uneven abundance of the different small RNA species derived from the same pre-miRNA (Figure 3A). miRDeep, developed by Friedlander et al., employs a novel algorithm based on a probabilistic model of miRNA biogenesis to score compatibility of the nucleotide position and frequency of sequenced small RNA reads with the secondary structure of miRNA precursors (Friedlander et al., 2008). When using miRDeep, the small RNA sequence reads are first aligned to the genome. The genomic DNA bracketing these alignments are extracted and computed for secondary RNA structure. Plausible miRNA precursor sequences based on a model for Dicer-mediated miRNA processing are identified. Finally, miRDeep scores the likelihood of the putative miRNA precursors and outputs a scored list of known and new miRNA precursors and mature miRNAs in the deep-sequencing samples (Friedlander et al., 2008).

miRNAKey

miRNAkey (Ronen et al., 2010) is a software pipeline designed to be used as a base-station for the analysis of deep sequencing data on miRNAs. The package implements a common set of steps generally taken to analyze deep sequencing data including the use of similarity search to detect known or conserved miRNAs as well as adding miRDeep (Friedlander et al., 2008) to predict new miRNAs. This program also includes unique features such as data statistics and multiple mapping levels to generate a comprehensive platform for the analysis of miRNA expression. Based on a statistic

analysis of small RNA sequencing data, it could generate measurement of differentially expressed miRNAs in paired samples by a tabular and graphical output format.

It should be noted that most of the programs listed in Table 1 are originally designed for analyzing miRNAs in animals. Several considerations prevent their direct application to profile miRNAs in plants. First, many available methods highly depend on sequence similarity search to detect known and new miRNAs, which is not sufficient to uncover species-specific miRNAs. In fact, only a minority of annotated miRNAs in plants is conserved between different lineages, suggesting that most unknown miRNAs would not be discovered through sequence similarity search (Cuperus et al., 2011). Second, for those programs that do consider other features their models are usually based on the animal systems. However, it is well studied that some aspects of miRNA biogenesis in plants and animals are critically different. For instance, pre-miRNAs in animals possess a rather uniform length at ~80 nucleotides, which is a key to the success of miRDeep, miRCat, and miRNAKey (Friedlander et al., 2008; Moxon et al., 2008a; Ronen et al., 2010). In plants, the precursor length is longer and more variable. Thus, it is not feasible to simply employ tools developed for animals to detect plant miRNAs. Third, considering the easy-to-use feature, most tools in Table 1 are available as a web-based version, which could easily handle small size sequencing data. However, at present it is standard to generate tens of millions of reads per sample, resulting in increased difficulty of processing the large amount of data on line and reinforcing the desirability of developing plant-specific tools.

miRDeep-P is a program for comprehensive identification of miRNAs in plants

miRDeep-P was modified from miRDeep (Friedlander et al., 2008) to specifically retrieve and quantify miRNA related information from deep sequencing data in plants (Yang and Li, 2011; Yang et al., 2011). Similar to miRDeep, this program maps the small RNA reads to a reference genome and extracts the sequence flanking each anchored read for predicting RNA secondary structure and quantifying the compatibility of the distribution of reads with Dicer-mediated processing. After progressively processing all mapped reads, candidate miRNAs as well pre-miRNAs are scored based on the core miRDeep algorithm (Friedlander et al., 2008) and filtered with plant-specific criteria (Meyers et al., 2008; Yang and Li, 2011; Yang et al., 2011). It thus provides reliable information on the transcription and processing of the pre-miRNAs (Figure 3A). Using training data from both *Arabidopsis* and rice, it was demonstrated that miRDeep-P works effectively for deep sequencing data in plants (Yang et al., 2011). miRDeep-P is freely available as a stand-alone package that runs in a command line environment (Yang and Li, 2011).

miRDeep-P was tested utilizing annotated miRNAs in *Arabidopsis* and available deep sequencing data from three independent small RNA libraries prepared from shoot, root and inflorescence (Yang et al., 2011). By retrieving the signature small RNA distribution from each of the 199 annotated pre-miRNAs (miRBase release 15), the tissue-specific expression pattern of individual miRNAs was determined. In shoot, root and flower, 81, 70 and 55 expressed pre-miRNAs were detected, respectively, indicating that only 40% of the annotated pre-miRNAs are expressed in major organ types. Northern blotting was performed and the results were clearly consistent with expression

determined from the deep sequencing data for miRNAs of single member families. For the multiple member families, the cumulated expression levels combining individual gene-level expression also showed strong agreement with that from the Northern analysis (Figure 3B).

Gene specific expression pattern generated by miRDeep-P revealed transcriptional relationship of paralogous members. For example, the *MIR169* family has 14 members in *Arabidopsis*. According to the neighbor-joining tree constructed from pre-miRNA sequences, this family could be grouped into three major clusters (Figure 3C). The smallest clade only consisting of *MIR169a* was not expressed according to miRDeep-P analysis. By contrast, the clade consisting of *MIR169i/j/k/l/m/n* was expressed simultaneously but only in root and shoot. The only exception for this clade was *MIR169i*, which was not expressed. Meanwhile, the *MIR169b/c/d/e/g/f/h* branch was detected in flower as well (Figure 3C). These results indicate that the tissue specific expression determined from deep sequencing data by miRDeep-P is consistent with the phylogenetic relationship of paralogous *MIR* genes.

Reliable estimation of gene level miRNA expression makes it possible to determine the relationship between miRNA detection rate and the sequencing depth. A simulation approach was taken in which sequence reads in the shoot library were randomly selected to simulate six different sequencing depths. These subsets were processed using miRDeep-P and the number of expressed miRNAs at each sequencing depth calculated. After reiterating this process for three times, the mathematical relationship between the number of detected miRNAs and the number of unique sequence reads was determined by curve fitting based on a logistic function (Figure 3D). This

analysis indicates that, when sampling of the RNA population is unbiased, there would be a finite number of reads needed to reach a saturated detection rate of expressed miRNAs. Further, the maximal number of expressed miRNAs in shoot was estimated to be 94. Accordingly, 1.13 million unique, perfectly mapped sequence reads were required to detect 95% of the expressed miRNAs in this simulation (Figure 3D).

Plant miRNA databases with integrated deep sequencing data

Increasing accumulation and mining of deep sequencing data have resulted in the development of comprehensive databases to facilitate miRNA annotation in a variety of species or experimental systems. The earliest and most comprehensive miRNA database, miRBase, combines deep sequencing data with miRNA annotation in chromalveolata, metazoan, mycetozoa, viridiplantae, and viruses (Griffiths-Jones et al., 2006). On the other hand, genome annotation databases for plant species such as TAIR and TFGD (Fei et al., 2011) also include miRNA loci. The online databases most relevant to plant miRNAs are summarized in Table 2. ASRP, the *Arabidopsis* Small RNA Project Database, is the first database providing a repository for sequences of miRNAs from various *Arabidopsis* genotypes and tissues (Gustafson et al., 2005). PmiRKB, currently focusing on the two model plants *Arabidopsis* and rice was developed to emphasize on single nucleotide polymorphisms regarding miRNAs in these two species supported by deep sequencing data (Meng et al., 2011). miRNEST is a newly released comprehensive miRNA database including miRNA sequences from more than 200 plants, and those annotated miRNAs from some model plants are supported by deep sequencing data from different samples (Szczesniak et al., 2012).

Transcriptome profiling has become indispensable in biology, which now includes not only mRNA but also other regulatory RNA species and intermediates of RNA metabolism. To fully decipher the transcriptome, systems based approaches are highly desirable to integrate the expression profiles with data characterizing other functional elements of the cell. Toward this goal, databases integrating miRNA annotation and deep sequencing data represent an important step forward. Combining sequencing data from various genotypes, diverse tissues, different developmental stages, or in response to different environmental challenges, could elaborate the expression patterns of miRNAs. Further, these databases should prove useful to integrate deep sequencing data of small RNA with other types of high throughput data as those for mRNA transcriptome and degradome. The integrated data, in conjugation with modeling and model testing, will provide important clues to the regulatory networks that ultimately elucidate genome transcription, function, and adaptation.

Beside identification of miRNAs and annotation of miRNA transcriptome, this dissertation also focuses on the following two parts in chapter 5 and appendix 2, respectively.

Interaction between regulations from miRNAs and mRNA alternative splicing

Recent studies indicate that lineage-specific miRNAs have continuously emerged in evolution (Rajagopalan et al., 2006; Zhu et al., 2008). The new miRNAs, once incorporated into the gene regulatory networks through the formation of new miRNA binding sites (MBSs), are thought to generate new gene circuits that expand the known

scope of miRNA-mediated cellular processes (Rajagopalan et al., 2006). For many miRNAs that are deeply conserved across plant lineages (Axtell and Bowman, 2008), they usually regulate homologous targets at identical MBSs in every species in which they are found. Consistently, the MBSs are shown to be subject to strong purifying selection (Guo et al., 2008b). Based on these observations, it has been argued that genetic changes resulting in beneficial miRNA-target interactions are maintained while nonproductive or deleterious changes continue to drift or be purged (Axtell and Bowman, 2008).

We are intrigued to understand how the seemingly rigid miRNA-target interaction, once the miRNA is fully incorporated into the regulatory gene networks, continues to evolve to allow fine-tuning of gene activity and adaptation. In the context of multi-layer gene regulation, it is worth noting that MBSs are present in mRNAs and must be transcribed and processed to be functional. Alternative processing of pre-mRNAs can conceivably eliminate or create functional MBSs in different mRNA isoforms of the same gene. Such alternative splicing events could provide a mechanism to bypass the strict constraint at MBSs while at the meantime maintain the interaction between miRNAs and their targets. Thus, global identification and analysis of alternative splicing events associated with MBSs is highly desirable to further understanding of the biological significance of this form of transcriptomic diversity and the intrinsic complexity of miRNA-target interactions in plants.

In the chapter 5, we performed a series of analyses in the model plant *Arabidopsis thaliana* to elucidate alternatively spliced MBSs as a mechanism for regulating gene expression. Through a genome-wide examination we identified many high confidence

alternative splicing events from annotated gene models and RNA-Seq data that produce mRNA isoforms of the same miRNA target gene differing in the sequences required for miRNA binding. Comparative and functional studies indicate that these events are important for target gene expression and miRNA function. Thus, alternatively spliced MBS represents a plausible and prevalent mechanism for regulating miRNA-target gene circuits in plants.

miRNAs as a group playing a buffer role in systemin-mediated pathway in tomato

Higher plants have evolved defense strategies to protect themselves from wounding by herbivorous insects. Tomato has been widely employed as a model system to study defense responses to herbivore and pathogen attack (Ryan and Pearce, 2003; Bostock, 2005). Briefly, systemin, a 18-amino-acid peptide processed from its 200-amino acid precursor prosystemin (PS) (Ryan and Pearce, 2003), could trigger the entire pathway, and induce the expression of defense genes such as a proteinase inhibitor (*PIN*). A transgenic line of tomato (35S::PS) over-expressing PS could promote the constitutive expression of defense genes and enhance resistance to herbivore and pathogen (Li et al., 2002; Chen et al., 2005). To date, the systemin-mediated signaling pathway has been well established over the last decade and more and more genes encoding proteins involved in this pathway are identified. However, there are no reports that focus on uncovering the role of plant miRNAs in this signaling pathway.

In the appendix 2, we commenced a piece of research that has been trying to understand how miRNAs are involved in this pathway. It is intriguing that the transgenic line 35S::PS re-programs almost half of miRNAs, and that many miRNAs are in turn

engaged in this pathway which is demonstrated by genetic experiments. Further, the comparison of phenotypic changes and entire transcriptome variation indicates that a large number of miRNAs as a group possibly plays a buffering role and deliver the internal signals to different pathways when plants meet the stimulus of wounding.

Figure 1. Comparison of the density of protein-coding and *MIR* genes in six plant lineages.

Vertical axis indicates the density of protein-coding and *MIR* genes per million genomic base pairs. The six plants are Ath (*Arabidopsis thaliana*), Osa (*Oryza sativa*), Aly (*Arabidopsis lyrata*), Ptc (*Populus trichocarpa*), Bdi (*Brachypodium distachyon*) and Ppe (*Prunus persica*). The numbers of protein-coding genes are obtained from TAIR10 (Ath), RGAP6.1 (Osa) (Ouyang et al., 2007) and Phytozome7.0 (Aly, Ptc, Bdi and Ppe) (Goodstein et al., 2012) while all the numbers of *MIR* genes are from miRBase17 (Kozomara and Griffiths-Jones, 2011).

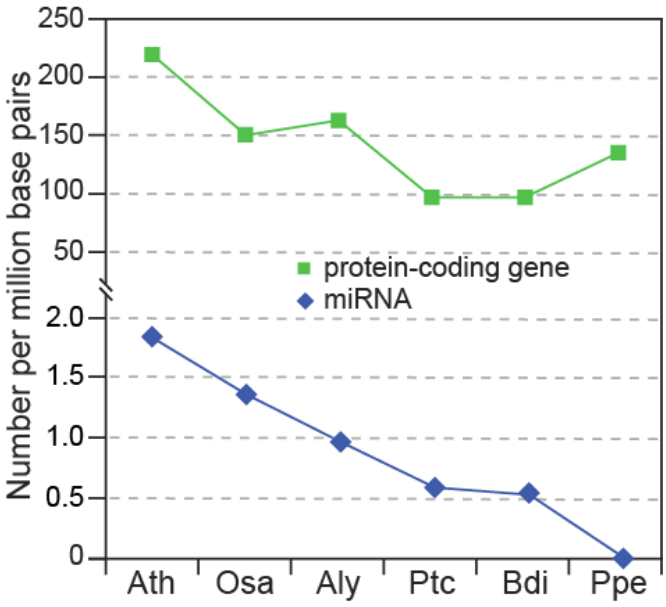

250
200
150
100
50
2.0
1.5
1.0
0.5
0
Number per million base pairs
protein-coding gene
miRNA
Ath
Osa
Aly
Ptc
Bdi
Ppe

Figure 2. Simplified model of miRNA biogenesis in plants.

MIR genes are initially transcribed by Pol II into pri-miRNAs that fold back to form hairpin structure. Splicing and further processing in nuclear involve the interactive functions of HYL1 and SE and of the cap-binding proteins CBP20 and CBP80. Pri-miRNAs and pre-miRNAs are sequentially processed by DCL1 to yield one or several phased miRNA/miRNA* duplexes, which are methylated by HEN1 and transported to the cytoplasm by HST1. The miRNA is selected and incorporated into dedicated AGO1-containing RISC that directs translation inhibition or cleavage of the target mRNA transcript.

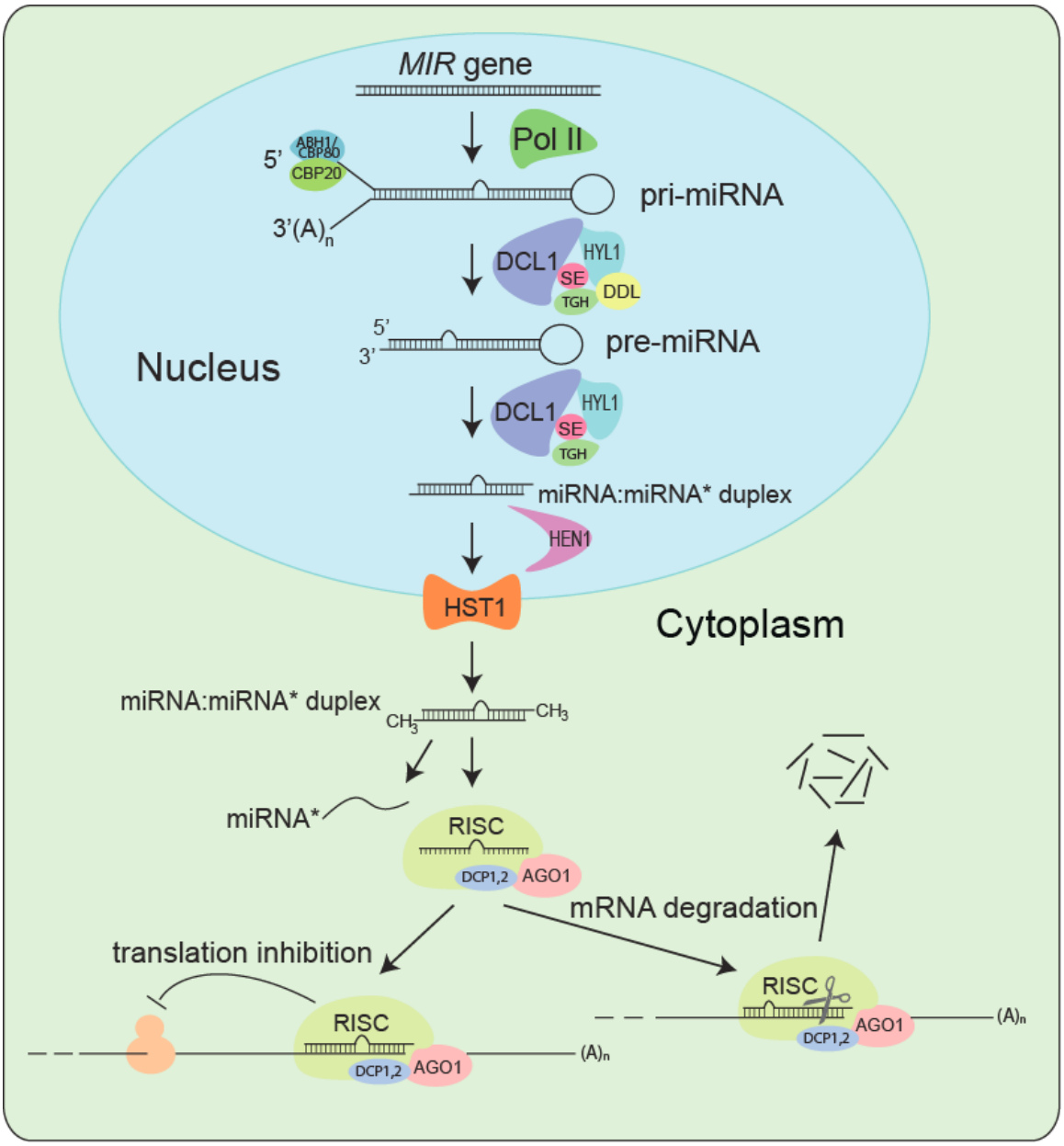

MIR gene
Pol II
5'
ABH1/
CBP80
CBP20
3'(A)n
pri-miRNA
DCL1
HYL1
SE
TGH
DDL
Nucleus
5'
3'
pre-miRNA
DCL1
HYL1
SE
TGH
miRNA:miRNA* duplex
HEN1
HST1
Cytoplasm
miRNA:miRNA* duplex
CH3
CH3
miRNA*
RISC
DCP1,2
AGO1
mRNA degradation
translation inhibition
RISC
DCP1,2
AGO1
(A)n
RISC
DCP1,2
AGO1
(A)n

Figure 3. Using miRDeep-P to identify and profile miRNAs from deep sequencing data.

(A) The core algorithm is based on a miRNA biogenesis model in which the small RNAs derived from a pre-miRNA are considered to have certain probabilities of being sequenced (Friedlander et al., 2008). This model distinguishes an expressed pre-miRNA from a non-expressed pre-miRNA or a genomic locus with the potential to form a hairpin but is not processed by Dicer. (B) Validation of miRDeep-P results in *Arabidopsis* by Northern blotting. RNA blots are shown on the left. Ethidium bromide staining of the 5S/tRNA is used as a loading control. The expression pattern (represented by the green color) of individual genes deduced from the sequencing data is shown on the right with genes with identical pattern combined. (C) Relating miRNA expression to pre-miRNA phylogeny. The annotated pre-miRNAs of the miR169 family in *Arabidopsis* were used to construct a phylogenetic tree. The gene-level expression profile of family members is depicted to the right of the tree. (D) Simulation of the relation between miRNA detection rate and sequencing depth in *Arabidopsis*. Perfectly mapped unique reads from the shoot library were randomly retrieved to create five different simulated sequencing depths. The number of expressed miRNAs was determined at each depth. The scatter plot represents results from three independent simulations and was used for curve fitting. The star sign indicates the actual data from the shoot library. Dashed lines indicate a 95% detection rate of the theoretic maximal number of expressed miRNAs and the corresponding sequencing depth. Adapted from Yang et al. (Yang et al., 2011).

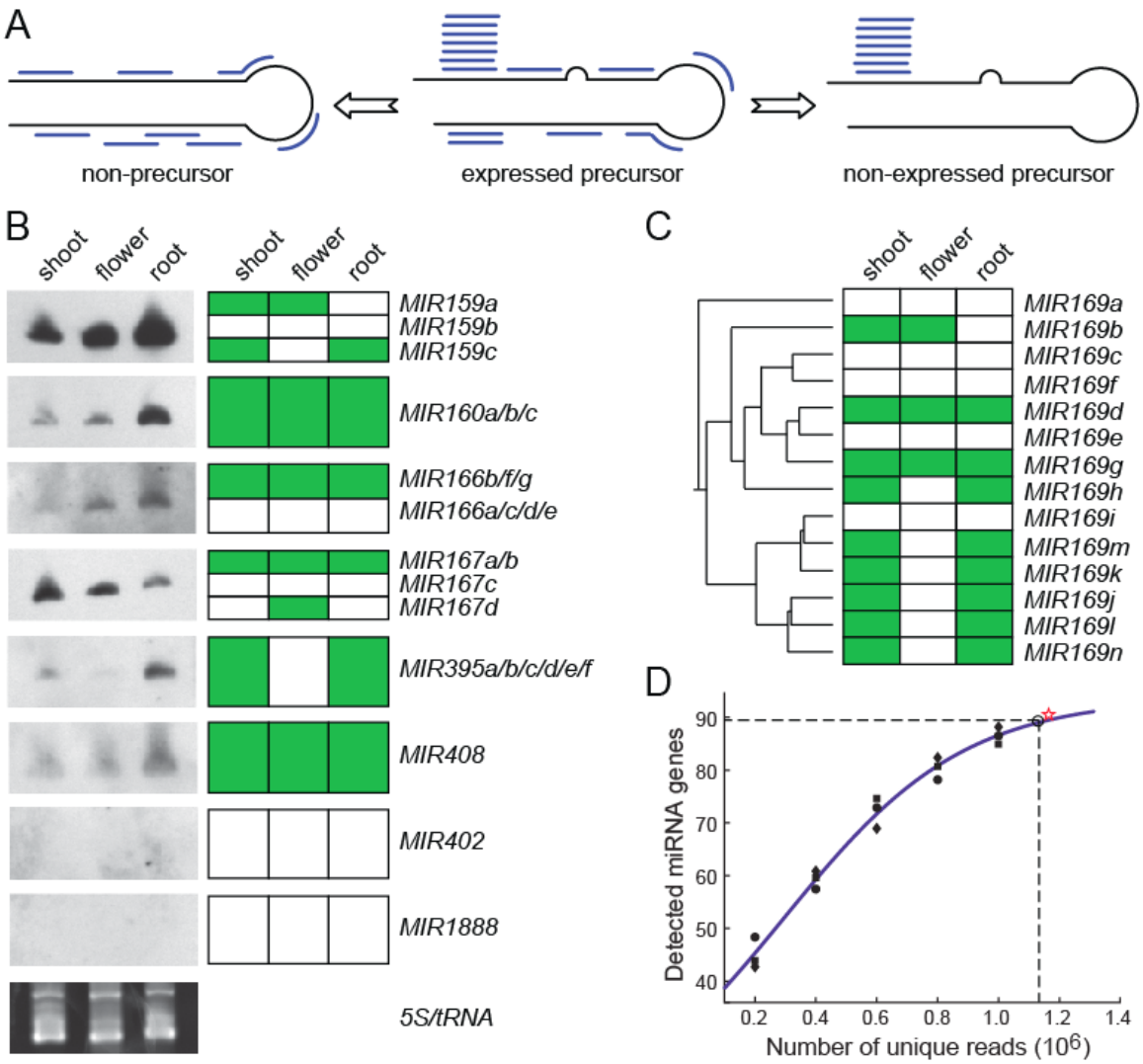

A
non-precursor
expressed precursor
non-expressed precursor
B
shoot
flower
root
MIR159a
MIR159b
MIR159c
MIR160a/b/c
MIR166b/f/g
MIR166a/c/d/e
MIR167a/b
MIR167c
MIR167d
MIR395a/b/c/d/e/f
MIR408
MIR402
MIR1888
5S/tRNA
C
MIR169a
MIR169b
MIR169c
MIR169f
MIR169d
MIR169e
MIR169g
MIR169h
MIR169i
MIR169m
MIR169k
MIR169j
MIR169l
MIR169n
D
Detected miRNA genes
Number of unique reads (10^6)

Table 1. Tools for analyzing deeply sequenced small RNA data

Name	Designed Model	Algorithm	Availability
UEA sRNA Toolkit*	animal & plant	based on criteria of miRNAs	web-based
miRDeep	animal	probabilistic model	stand-alone
miRanalyzer**	animal	machine learning	web-based
SeqBuster [1]	animal	sequence similarity	web-based
DSAP [2]	animal	sequence similarity	web-based
mirTools [3]	animal	sequence similarity & miRDeep	web-based
miRNAKey	animal	sequence similarity	stand-alone
miRNEST	animal & plant	sequence similarity	web-based

*Only the miRCat component is for detecting new miRNAs.

**Updated miRanalyzer could also predict new miRNAs in plants, and it has a new stand-alone version.

[1-3] are cited from (Huang et al., 2010; Pantano et al., 2010; Zhu et al., 2010), respectively.

Table 2. Databases for plant miRNAs with integrated deep sequencing data

Database	Description	URL
miRBase	Including miRNA sequences and annotations in more than 50 plants.	http://www.mirbase.org/
miRNEST	Combining miRNA sequences in more than 200 plant organisms.	http://lemur.amu.edu.pl/share/php/mirnest/home.php
PmiRKB	*Arabidopsis* and Rice miRNA knowledge base.	http://bis.zju.edu.cn/pmirkb/
ASRP	Database of *Arabidopsis* small RNA sequences.	http://asrp.cgrb.oregonstate.edu/

Chapter 2. Analysis of expressed microRNAs in *Arabidopsis*

Global analysis of gene-level microRNA expression in *Arabidopsis* using deep sequencing data[1]

[1]Formatted as a co-authored manuscript and published as:

Yang X, Zhang H, Lei Li. 2011. *Genomics*. 98(1):40-46.

Abstract

MicroRNAs (miRNAs) regulate gene expression at the post-transcriptional level in eukaryotes. Exclusive focus on mature miRNA in most expression profiling efforts has prevented effective measurement of the expression of individual miRNA (*MIR*) genes. Using three sequenced small RNA libraries, we adapted miRDeep, which employs a probabilistic model of miRNA biogenesis, to analyze the miRNA transcriptome in *Arabidopsis*. We determined that less than 40% annotated *MIR* genes are expressed in shoot, root or inflorescence. We found that within paralogous families the expression pattern of individual genes correlates with the phylogenetic distance. Combining novel candidates identified in this study, we deduced the maximal number of expressed *MIR* genes. We further estimated the sequencing depth necessary to reach a near-saturated detection rate by curve fitting simulation. These results demonstrate that signature distribution of small RNA reads along the miRNA precursor is an effective model to profile *MIR* gene expression in *Arabidopsis*.

Introduction

MiRNAs are emerging as an important class of gene regulators functioning at the post-transcriptional level. Although mature miRNAs are only 20- to 24-nucleotides in length, they are processed from much longer primary transcripts known as pri-miRNAs via stem–loop structured intermediates called pre-miRNAs (Bartel, 2004; Chen, 2005; Jones-Rhoades et al., 2006; Voinnet, 2009). In higher plants, pri-miRNA and pre-miRNA processing is carried out in the nucleus mainly by the endonuclease Dicer-like1 (Papp et al., 2003). As a consequence of these processing events, two complementary short RNA molecules are generated. In plants, the duplex is end-methylated (Yang et al., 2006) and then transported to the cytoplasm (Park et al., 2005). Only the mature miRNA is integrated into the RNA-induced silencing complex (RISC) whereas the complementary RNA, RNA*, is degraded as a RISC substrate (Khvorova et al., 2003; Schwarz et al., 2003; Bartel, 2004). Afterwards, miRNAs base-pair with their targets and direct either cleavage (German et al., 2008) or translational repression (Brodersen et al., 2008) of the target transcripts.

Following the initial discovery of miRNAs in the worm *C. elegans* (Lee et al., 1993; Wightman et al., 1993), much attention has focused on the identification of miRNAs and their target genes in various model organisms. Large numbers of miRNAs have been identified in many experimental systems. They impact a substantial portion of the transcriptome. In higher plants, many miRNA–target pairs have been experimentally identified or computationally proposed (Jones-Rhoades et al., 2006; Fahlgren et al., 2007; Alves et al., 2009). It is now well-established that these gene circuits are crucial for plant

development and responses to environmental challenges (Jones-Rhoades et al., 2006; Garcia, 2008; Voinnet, 2009).

MIR genes are transcribed by RNA polymerase II (Tam, 2001; Lee et al., 2004), which means their expression is subject to similar regulatory mechanisms as protein-coding genes. Transcriptional control is thus important for cell- and development-specific function of miRNAs. There has been a surge of interest in the past decade in profiling the expression pattern of miRNAs using various experimental approaches (Wark et al., 2008). Most recently, deep sampling of specifically prepared low-molecular-weight RNA libraries based on next generation sequencing platforms has been used to identify miRNAs in various plant species. In addition to validating annotated miRNAs, sequencing of small RNA populations is increasingly revealing novel miRNAs not conserved across different species (Rajagopalan et al., 2006; Fahlgren et al., 2007; Molnar et al., 2007; Zhu et al., 2008).

An advantage of the sequencing method is its sensitivity in detecting even poorly expressed or species-specific miRNAs (Sunkar et al., 2008). As essentially a random sampling process from a very large population of small RNAs, the potential of deep sequencing to provide quantitative information on the expression pattern of miRNAs has been explored (Fahlgren et al., 2007; Creighton et al., 2009). However, the recovery of sequence reads may not be uniform across the transcriptome in high throughput sequencing efforts (Taub et al., 2010). For example, GC-rich sequences may be over-represented while AT-rich sequences under-represented. The read mapping procedures may also generate regional bias. Because reads that can be mapped to multiple loci are usually discarded, genomic regions with higher degeneracy typically show lower read

coverage. Local DNA or chromatin structure could lead to coverage heterogeneity as well.

Additionally, even with the information-rich deep sequencing data, an exclusive focus on mature miRNAs, the final gene products, could result in limitations in terms of comprehensively profiling the miRNA transcriptome (Friedlander et al., 2008). Besides technical considerations regarding unbiased measurement of miRNA abundance, a major drawback of the miRNA-centric approaches is their failure to take into account the fact that many miRNAs are encoded by gene families (Meyers et al., 2008). Further, miRNA families are known to keep expanding in evolution (Chen and Rajewsky, 2007; Li and Mao, 2007). Different family members can conceivably contain different combinations of *cis*-regulatory elements and contribute to regulatory complexity. In plants, this notion is supported by bioinformatics (Megraw et al., 2006; Zhou et al., 2007) as well as experimental evidence (Xie et al., 2005; Kawashima et al., 2009). New strategies and tools are thus highly desirable in fully utilizing the deep sequencing data to gain insights into the regulation and function of individual *MIR* genes.

Unique to the maturation of miRNA is the cutting of the stem–loop structured pre-miRNA into the miRNA:miRNA* duplex and the subsequent unwinding of the duplex. Processing of the miRNA precursors thus releases three species of small RNAs with various lengths: mature miRNA, miRNA* and fragments of RNA derived from other parts of the stem–loop structure. Typically the mature miRNAs are much more stable, which results in uneven abundance of the different small RNA species derived from the same pre-miRNA. The program miRDeep employs a probabilistic model of miRNA biogenesis to score compatibility of the nucleotide position and frequency of

sequenced small RNA reads with the secondary structure of pre-miRNAs (Friedlander et al., 2008). Here we report our effort to adapt miRDeep to the model plant *Arabidopsis thaliana*. Applying the modified miRDeep program to three sequenced small RNA libraries, we determined the expression status of individual *MIR* genes in a given library (gene-level expression). Our results are useful for further characterizing the miRNA transcriptome in plants.

Materials and methods

Plant materials

Wild type plants used in this work were *Arabidopsis thaliana* ecotype Col-0. The lost-of-function *hen1* mutant was described previously (Chen et al., 2002). Seeds were placed on GM media and incubated at 4 °C in the dark for 4 days. The cold-treated seeds were exposed to continuous white light (170 μmol s^{-1} m^{-2}) at 22 °C for 7 days. Seedlings were then transferred to soil and maintained under continuous light until flowering.

Data sources

Sequence reads in the shoot (GSM442935), root (GSM442933) and inflorescence (GSM277608) small RNA libraries were downloaded from NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). The three libraries contain 5,003,481, 4,919,514, and 9,777,263 reads, respectively. Annotated miRNA and pre-miRNA sequences for *Arabidopsis* were obtained from miRBase (http://www.mirbase.org/; release 15). *Arabidopsis* genomic sequence was obtained from The Arabidopsis Information Resource (http://www.arabidopsis.org/; version 9). The index files of *Arabidopsis* genome (TAIR9) were downloaded from the Bowtie website (http://bowtie-bio.sourceforge.net/).

Mapping small RNA reads

Identical reads from the same sequenced small RNA libraries were collapsed and the copy number was recorded. We then employed Bowtie (Langmead et al., 2009), which indexes the genome with a Burrows–Wheeler index, to map the collapsed

sequence reads into the indexed *Arabidopsis* genomic sequences. We only retained the reads that were perfectly mapped to the indexed genome (Bowtie parameters -a -v 0). The same strategy was also used to index the miRNA precursor sequences and to map small RNA reads to the indexed precursor sequences.

Modifications to miRDeep

The miRDeep package consists of a series of Perl scripts designed to sequentially process the small RNA sequence reads (Friedlander et al., 2008). These scripts were downloaded from http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/index.html. We tailored several of these scripts for the purpose of this study. First, we replaced BLAST with Bowtie as the mapping tool. Second, the constant -c in filter_alignment.pl, which denotes the maximal mapping loci allowed for each retained read, was set at 15 instead of 5 because the largest *Arabidopsis* miRNA family has 14 members. Third, we changed the default length of pre-miRNA precursor from 110 bp to variable. Forth, we modified the way of scoring the minimum free energy by setting up a max value based on the Gumbel distribution because with increased pre-miRNA length the value of minimum free energy becomes smaller. Finally, to score annotated *MIR* genes, genomic sequences corresponding to the stem–loop structures were extended by 500 bp on both ends. The extended sequences were used as the genomic context for the corresponding pre-miRNAs. To identify novel miRNAs, all reads mapped to the Bowtie-indexed genome were used to extract the 250 bp bracketing sequences to be processed with the modified miRDeep package. Then, the candidates were filtered with

criteria specific to the miRNA:miRNA* duplex in plants (Meyers et al., 2008). All scripts can be downloaded at https://sourceforge.net/projects/mirdp/.

Phylogeny of pre-miRNAs

Of the *Arabidopsis MIR* genes belonging to families with at least three paralogous members, the pre-miRNA sequences were employed to deduce the phylogenetic relationship using ClustalW2 of EMBL-EBI (http://www.ebi.ac.uk/Tools/clustalw2/). The default parameters were used and UPGMA was chosen as the option for clustering.

Estimating the relationship between detected *MIR* genes and sequencing depth

Sequence reads in the shoot library was randomly picked out at five different simulated sequencing depths (i.e. 200,000; 400,000; 600,000; 800,000; and 1,000,000 perfectly mapped reads), which was reiterated three times. Each simulated data was processed using modified miRDeep to detect expressed *MIR* genes. A logistic function was used to fit the resultant datasets:

$$m(t) = \frac{K}{1 + Ce^{-rt}}$$

where *m(t)* is the number of detected *MIR* genes, *t* is the sequencing depth, *K* is the theoretic maximal number of expressed *MIR* genes as $t \to \infty$, *C* is an arbitrary constant, and *r* is a positive parameter. The best fit curve was identified as previously described (Cavallini, 1993) whereby *r* was determined based on minimum error *E*. The following equation was used to calculate the value of *E*:

$$E = \sum_{i=1}^{n} (m(t_i) - m_i)^2$$

where $m(t)_i$ is the number of detected *MIR* genes at each simulated sequencing depth, and m_i is the value of the simulated curve at the same depth. The minimum value of *E* was searched by the function *fminsearch* in MATLAB with an initial estimation as previously described (Cavallini, 1993).

Northern blot analysis

Total RNA was extracted using TRIzol agents (Invitrogen) and treated with RNase-free DNaseI (Invitrogen) as recommended by the manufacturer. For miRNA blotting, 20 μg total RNA was separated on 12% denaturing polyacrylamide gel containing 7 M urea and electrically transferred to Hybond N+ nylon membrane (GE Healthcare Life Sciences). Blots were hybridized overnight with respective miRNA complementary oligo-nucleotides end-labeled with digoxigenin-11- ddUTP (Roche) at 42 °C in hybridization buffer [1% (W/V) BSA, 0.5 M Na_2HPO_4, 15% (V/V) formamide, 1 m MEDTA, and 7% SDS (W/V)] after pre-hybridization for 2 h. Blots were then washed using the DIG Wash and Block Buffer Set and incubated with Anti-Digoxigenin-AP antibody (Roche) as recommended by the manufacturer. Blots were equilibrated in detection buffer for 5 min, and incubated for 15 min with the CDP-Star solution (Roche) applied before exposure to X-ray films.

Results

Modification of miRDeep to score deep sequencing data from *Arabidopsis*

After anchoring small RNA reads to the genome, miRDeep employs a probability model to separate endogenous pre-miRNAs from other stem–loop containing loci based on distribution of the reads (Figure 1A). A key technical consideration for applying this method is the size of the window (N) bracketing the RNA:DNA alignment within which to compute the secondary RNA structure (Figure 1B). For animals, a single default value of N at 110 bp is sufficient (Friedlander et al., 2008) because of the little variation in either the loop sizes or the position of the miRNA:miRNA* duplex within pre-miRNAs (Han et al., 2006). In contrast, the length of pre-miRNAs in plants is much longer and more variable (Meyers et al., 2008). To adapt miRDeep to plant systems, we utilized training data from *Arabidopsis* to experimentally establish the appropriate N value.

The training data consisted of 199 annotated pre-miRNAs and a set of sequenced small RNA libraries prepared from shoot, root and inflorescence (Table 1). All reads pertinent to the pre-miRNAs were retrieved. For each mapped read, local sequence entailed by different N values were used to score the *MIR* gene by miRDeep. Although more expressed *MIR* genes were detected in shoot than in root and inflorescence, we found that the trend line of detection rate as a function of N was the same for all three libraries. As shown in Figure 1C, the slope is steepest when N increases from 100 to 150 bp as many pre-miRNAs in *Arabidopsis* are in this size range. Consistent with the diverse sizes of plant pre-miRNAs, the curves peak at an N of 250 bp and decline slightly thereafter (Figure 1C). We found the problem of using unnecessarily large windows is that it may complicate prediction of the secondary structure and hence obscure the

miRNA:miRNA* duplex for a few pri-miRNAs (Supplemental Figure 1). Therefore, we conclude that an N value of 250 bp is optimal for using miRDeep in *Arabidopsis*.

Determination of gene-level miRNA expression using modified miRDeep

The primary purpose of analyzing the signature read distribution along pre-miRNAs in miRDeep is to identify bone fide miRNAs (Friedlander et al., 2008).We reasoned that this approach should also be useful to determine gene level expression of the *MIR* genes to the extent permitted by divergence of the pre-miRNA sequences as many *MIR* genes encode identical or near-identical mature miRNAs. Toward this goal, we profiled tissue-specific expression of individual *MIR* genes using modified miRDeep. We detected 81, 70 and 54 expressed *MIR* genes from shoot, root and inflorescence, respectively (Table 1). Collectively, 90 annotated *MIR* genes were determined to be expressed in at least one organ type (Supplemental Table 1). These results indicate that expressed *MIR* genes may not exceed 40% of the annotated total at the current sequencing depth in major plant organ types when gene-level expression is considered.

We performed two independent analyses to validate the results on *MIR* gene expression. First, we carried out Northern blotting to profile the expression levels of eight randomly selected miRNAs in shoot, root and inflorescence. We were able to detect expression for six of the miRNAs in at least one organ type while expression of two miRNAs was not detected under our experimental conditions, which is consistent with the miRDeep result (Figure 2). The six miRNAs belong to distinct families containing a single or paralogous members that share identical mature miRNAs. The Northern result is clearly consistent with expression pattern determined by miRDeep for the three miRNAs

representing single member families. For the multiple member families, we found that the combined expression pattern of individual *MIR* genes showed strong agreement with that from the Northern analysis (Figure 2).

Next we examined the relation between the copy number of mature miRNAs in the sequencing data and the expression status determined by miRDeep. From shoot, root and inflorescence, at least one copy was retrieved for 105, 88 and 55 distinct mature miRNAs, respectively, greater than the number of *MIR* genes considered to be expressed (Table 1). We observed that the non-expressed *MIR* genes in general have lower copy number of retrieved mature miRNAs than the expressed (Supplemental Figure 2). Further, we found that many of the non-expressed *MIR* genes with retrieved mature miRNAs belong to multiple member families, indicating that counting the mature miRNAs alone may lead to an overestimation of the number of expressed *MIR* genes. Together our results provide a conservative yet reliable estimation of gene-level miRNA expression in *Arabidopsis*.

Relating *MIR* gene expression to pre-miRNA phylogeny

To further validate the gene-specific expression pattern and to gain insight into the transcriptional relationship of paralogous *MIR* genes, we sought to analyze the phylogeny of the pre-miRNAs. Based on the mature miRNAs, the 199 *Arabidopsis MIR* genes can be grouped in 122 families, with 17 of which including at least three paralogous members. Of the 17 families, 13 include at least two members considered to be expressed by miRDeep. Of these 13 families, we were able to construct the phylogeny for 10 families based on pre-miRNA sequences while the other three were excluded from

this analysis because of the large variation in the length of the pre-miRNAs. We found that the organ-specific expression has co-evolved with the pre-miRNAs for seven of the 10 families (Figure 3; Supplemental Figure 3). Such a strong correlation (hypergeometric test, $P<0.05$) of gene-level expression with the phylogenetic distance between pre-miRNAs thus can be viewed as evidence supporting the expression pattern.

For example, the miR169 family has at least 14 members in *Arabidopsis*. Based on phylogenetic clustering of the pre-miRNAs, the miR169 family can be grouped into three major clades (Figure 3). The smallest clad consisting of *MIR*169a is expressed in none of the three organ types. By contrast, the clad consisting of *MIR*169i/j/k/l/m/n was detected in root and shoot but not in inflorescence. The only exception in this clad is *MIR*169i which is not expressed in either organ. Meanwhile, members from the branch consisting of *MIR*169b/c/d/e/f/ g/h were detected in the floral tissue as well (Figure 3). These results indicate that gene-level expression determined from the sequencing data is consistent with the phylogenetic relationship for most miRNA families and provide useful information to further study transcriptional regulation of paralogous *MIR* genes.

Organ-specific expression of conserved and non-conserved *MIR* genes

Although the largest number of expressed *MIR* genes was found in shoot, normalization against the number of perfectly mapped reads indicates that the size of the miRNA transcriptome in the three organ types is comparable (Table 1). To further analyze the expressed *MIR* genes, we examined their organ-specific expression pattern. This analysis revealed that the 90 expressed *MIR* genes include two groups with distinct expression domains. Group I contains 40 genes that were detected from all three organs.

Group II includes 15 genes (eight, three, and four from shoot, root and inflorescence, respectively) that were uniquely detected in only one of the examined organs (Figure 4A).

To compare Group I and II genes, we searched all 199 genes against those annotated in other plant species. We considered a gene to be conserved if the encoded miRNA could be found in at least one other plant, which is the case for 102 *MIR* genes. The 97 remaining genes encode miRNAs that are only found in *Arabidopsis* and thus deemed non-conserved. We observed that 32 of the 40 (80%) Group I genes, which are expressed in all three organs, are conserved. In contrast, only two of the 15 (13%) Group II genes are conserved (Figure 4B). The relative portion of conserved and non-conserved genes in Groups I and II is thus significantly different (chi square test, $P<0.05$). Consistent with previous reports concerning only the mature miRNAs (Rajagopalan et al., 2006; Fahlgren et al., 2007; Molnar et al., 2007; Zhu et al., 2008), our results suggest that genes encoding conserved miRNAs in *Arabidopsis* tend to be constitutively activated while expression of non-conserved ones is more organ-specific.

Detection of putative novel *MIR* genes

Accurate profiling of gene-level expression suggests that this method can be applied to unambiguously detect novel *MIR* genes. From all small RNA reads mapped to the genome, 9, 10 and 6 candidate *MIR* genes were detected using modified miRDeep from shoot, root and inflorescence, respectively (Table 1). Combined, a total of 18 putative novel *MIR* genes encoding 15 distinct miRNAs were identified (Supplemental Table 2). An example is illustrated in Figure 5A. We randomly selected seven novel

miRNAs and performed Northern blot analysis to confirm their expression in shoot, root and inflorescence. In wild type plants, Northern signals from five predicted miRNAs were detected that are in general consistent with the expression pattern determined by miRDeep (Figure 5B). Further, no signal was detected in the hen1 mutant in which miRNA biogenesis is defective (Yang et al., 2006). Together these results indicate that the novel *MIR* genes encode authentic expressed miRNAs.

Among the putative novel *MIR* genes, 5 (28%) were retrieved fromat least two libraries and 13 (72%) from just one (Supplemental Table 2). This high degree of organ-specificity suggests that they are likely non-conserved (comparing to Figure 4B). Indeed, searching the novel *MIR* genes against all annotated in plants revealed that only *NM14* encodes a documented miRNA — a new member of the miR156 family (Supplemental Figure 3). In addition, *NM08* has been annotated as *MIR*3933 in another deep sequencing effort that appears to be unique to *Arabidopsis* (Chellappan et al., 2010). Collectively, these results indicate that there are a large number of expressed novel *MIR* genes that warrant further investigation.

Simulating the relationship between *MIR* gene detection rate and sequencing depth

Reliable estimation of gene-level expression from deep sequencing data prompted us to examine the relationship between *MIR* gene detection rate and the sequencing depth. To this end, we used a simulation approach in which we randomly selected perfectly mapped sequence reads to create five arbitrary sequencing depths. We used the shoot library because it contains the most mapped reads (Table 1). The five resultant subsets of reads were processed using modified miRDeep and the number of detected

MIR genes at each simulated sequencing depth was calculated. After reiterating this process for three times, we performed curve fitting to determine the mathematical relationship between the number of detected *MIR* genes and the number of perfectly mapped reads (Figure 6). We chose the logistic function because it is considered to be able to best describe the enlargement in a number of detected genes as the simulated sequencing depth increases (Cavallini, 1993).

Successful fitting the simulated data with a specific logistic curve indicates that, assuming sampling of the RNA population is unbiased, there would be a finite number of perfectly mapped reads that is needed to reach a saturated detection rate of expressed *MIR* genes (Figure 6). Meanwhile, the maximal number of expressed *MIR* genes in individual libraries could be mathematically predicted. We estimated that such a maximal number is 94 in shoot and a total of 1.13 million perfectly mapped sequence reads are sufficient to detect 95% (~90) of the expressed *MIR* genes (Figure 6). Interestingly, at the current sequencing depth (1.24 million perfectly mapped reads) in shoot, 90 *MIR* genes (81 annotated plus 9 predicted; Table 1) were indeed detected (Figure 6), attesting to the accuracy of the simulation approach.

Discussion

Although the potential impact of systematic biases of high throughput sequencing has not been assessed for small RNA libraries in plants, we show here that miRDeep is an effective computational tool for profiling *MIR* expression and discovering new miRNAs. When using miRDeep, sequence reads are first aligned to the genome (Friedlander et al., 2008). The DNA sequences from the genomic window bracketing the alignments are then extracted and computed for RNA secondary structure. Plausible miRNA precursor sequences based on the characteristic stem–loop structure required for Dicer-mediated miRNA processing are identified. Finally, miRDeep calculates the likelihood of the observed read distribution along the putative miRNA precursors and outputs a scored list of mature miRNAs as well as the precursors (Friedlander et al., 2008). Given the high conservation of miRNA biogenesis between animals and plants (Bartel, 2004; Arteaga-Vazquez et al., 2006; Bartel, 2009), miRDeep should be useful in processing deep sequencing data generated in plant systems as well.

In contrast to animals in which the miRNA precursors have in general uniform sizes (Han et al., 2006), the length of pre-miRNAs in plants is much longer and more variable (Llave et al., 2002; Reinhart et al., 2002; Meyers et al., 2008). Thus the optimal window size encompassing the RNA:DNA alignment needs to be determined for plants. While smaller windows would exclude detection of legitimate miRNAs with long precursors, larger windows may increase the computational cost and create difficulties in extracting precisely the pre-miRNA information. Utilizing annotated miRNAs and representative deep sequencing data in the model plant *Arabidopsis*, we experimentally determined the optimal window size to be 250 bp (Figure 1). This window size was also

found to be optimal when training data from rice, another model plant, was used (Supplemental Figure 4). Together these results demonstrate that miRDeep can be adapted to plants with specifically modified parameters.

As a novel application of miRDeep, we sought to determine the gene level expression profiles of annotated *MIR* genes in three major organ types: shoot, root and inflorescence. Compared to other miRNA-profiling methods focusing on the gene product, our results are more conservative as paralogous genes encoding the same miRNAs were distinguished. Indeed, less than 40% of the annotated genes were considered to be expressed in any of the three organs using available deep sequencing data (Figure 1C; Table 1). Several lines of evidence indicate that gene-level expression is informative regarding regulation of the *MIR* genes. Although results from Northern blot analysis on selected miRNAs are highly agreeable with the combined expression pattern deduced from all related *MIR* genes (Figure 2), we found that paralogous members of the same family often have a different expression pattern (Supplemental Table 1). When expression domain of individual *MIR* genes within a family was compared, a strong correlation with the phylogenetic distance was observed (Figure 3; Supplemental Figure 3). Comparison of the expression domain further revealed that genes encoding conserved miRNAs are more constitutively expressed than the non-conserved (Figure 4). Taken together, these results provide conservative but accurate gene-level expression information on the miRNA transcriptome in *Arabidopsis*.

Different from the small but abundant miRNA families in animals, plants have fewer but larger families (Li and Mao, 2007; Axtell and Bowman, 2008). Members of plant *MIR* families are often highly similar, suggesting recent expansion via tandem and

segmental duplication events (Li and Mao, 2007). Previous bioinformatic analyses indicate that promoters of family members contain shared as well as unique motifs (Megraw et al., 2006; Zhou et al., 2007). Together with our results on the gene-level expression, these findings collectively argue for a scenario in which the paralogous members have overlapping expression domain dictated by shared *cis*-regulatory elements while regulatory elements unique to specific members drive the expression domain to a different cell type. This notion has been verified by molecular genetic studies. For example, there are six *MIR*395 genes (*MIR*395a–f) in *Arabidopsis*. When the promoter of individual family member was used to drive the expression of the green fluorescent protein (GFP), it was found that family members share the same tissue- and cell-specific patterns of GFP expression while additional expression was observed for some individual members (Kawashima et al., 2009). It would be of great interests to further investigate the expression pattern of miRNA families by increasing the sample sizes. The expanded expression profiles with higher cellular resolution should offer desired information that can be used to trace evolutionary changes in the *cis*-regulatory elements and hence subfunctionalization of the *MIR* genes.

Reliable estimation of gene-level *MIR* expression from deep sequencing data allowed us to use a curve fitting approach to determine the relationship between detection rate and the sequencing depth. We found that a logistic function can be used to perfectly describe this relationship (Figure 6), indicating that there would be a theoretic maximal number of expressed *MIR* genes that can be detected by deep sequencing. We inferred this number to be 94 for shoot and estimated that 1.13 million perfectly mapped reads are sufficient to detect 95% of the expressed genes. Interestingly, the estimated numbers

were validated by the actually detected *MIR* genes at the current sequencing depth (Figure 6; Table 1), indicating that most expressed genes have been discovered. Taken together, results from the current study demonstrate that the gene-level expression information is helpful toward fully elucidating the miRNA transcriptome in plants.

Figure 1. Using miRDeep to score deep sequencing data from *Arabidopsis*.

(A) The core algorithm of miRDeep is based on a model of miRNA biogenesis in which the small RNA products derived from a miRNA precursor are considered to have certain probabilities of being sequenced (Friedlander et al., 2008). This model is used to distinguish a miRNA precursor from a genomic locus with the potential to form an RNA hairpin but is not processed into pre-miRNA. Mapped small RNA reads are depictured as thick lines. (B) Diagram illustrating the selection of genomic regions bracketing the aligned small RNA reads within which to compute the RNA secondary structure. The optimal value of N needs to be determined when adapting miRDeep to *Arabidopsis*. (C) Testing the detection rate of expressed *MIR* genes with different N values. Small RNA reads from deep sequencing libraries prepared from shoot, root and inflorescence were processed in miRDeep when different N values were used. Number of the annotated *MIR* genes in *Arabidopsis* detected at each N value is plotted.

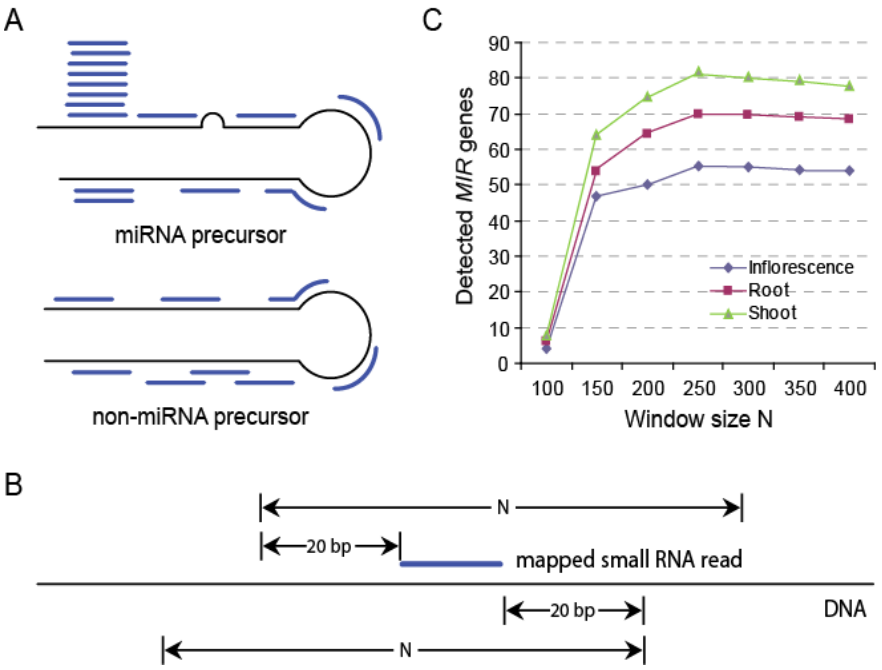

A
miRNA precursor
non-miRNA precursor
C
Detected MIR genes
90
80
70
60
50
40
30
20
10
0
100 150 200 250 300 350 400
Window size N
Inflorescence
Root
Shoot
B
N
20 bp
mapped small RNA read
20 bp
DNA
N

Figure 2. Northern blot analysis of the expression pattern of annotated miRNAs. RNA blots hybridized with probes complementary to the mature miRNAs are shown on the left. Ethidium bromide staining of the low-molecular-weight fraction of total RNA is used as a loading control. The expression pattern of individual genes deduced from the sequencing data is shown on the right with genes with identical pattern combined. The gray boxes represent expression and the blank boxes represent no expression. S, shoot; I, inflorescence; R, root.

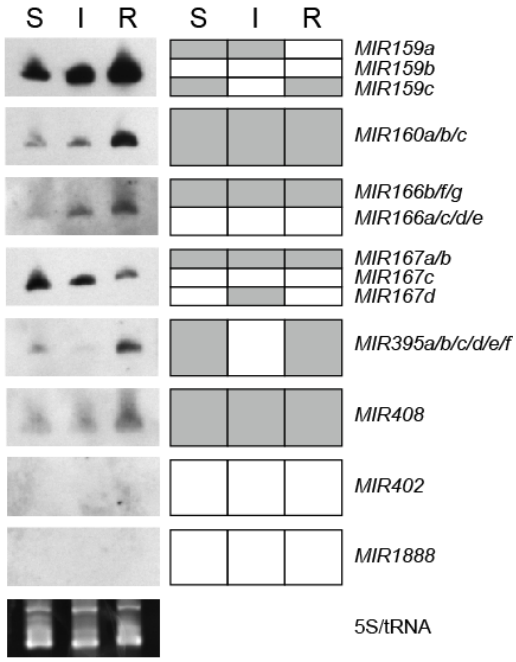

S I R S I R
MIR159a
MIR159b
MIR159c
MIR160a/b/c
MIR166b/f/g
MIR166a/c/d/e
MIR167a/b
MIR167c
MIR167d
MIR395a/b/c/d/e/f
MIR408
MIR402
MIR1888
5S/tRNA

Figure 3. Relating miRNA expression to pre-miRNA phylogeny for the miR169 family.

The annotated pre-miRNA sequences were used to construct a phylogenetic tree. The gene level expression profile of family members is depicted to the right of the tree. A gray box represents expression while a blank box represents no expression. S, shoot; I, inflorescence; R, root.

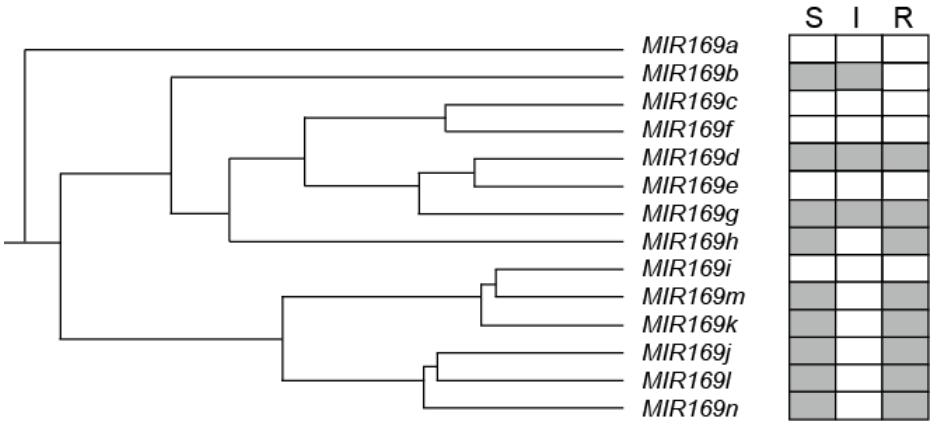

S
I
R
MIR169a
MIR169b
MIR169c
MIR169f
MIR169d
MIR169e
MIR169g
MIR169h
MIR169i
MIR169m
MIR169k
MIR169j
MIR169l
MIR169n

Figure 4. Comparison of the expression domain of annotated *MIR* genes.

(A) Venn diagram showing number of annotated *MIR* genes expressed in shoot, root and inflorescence. (B) Correlation between expression domain distribution and sequence conservation of *MIR* genes. Constitutive genes refer to those that are expressed in all three organs. Organ-specific genes are those only detected in one organ type. Conserved genes refer to those that encode a mature miRNA also found in other plant species. Non-conserved genes encode miRNAs only found in *Arabidopsis*.

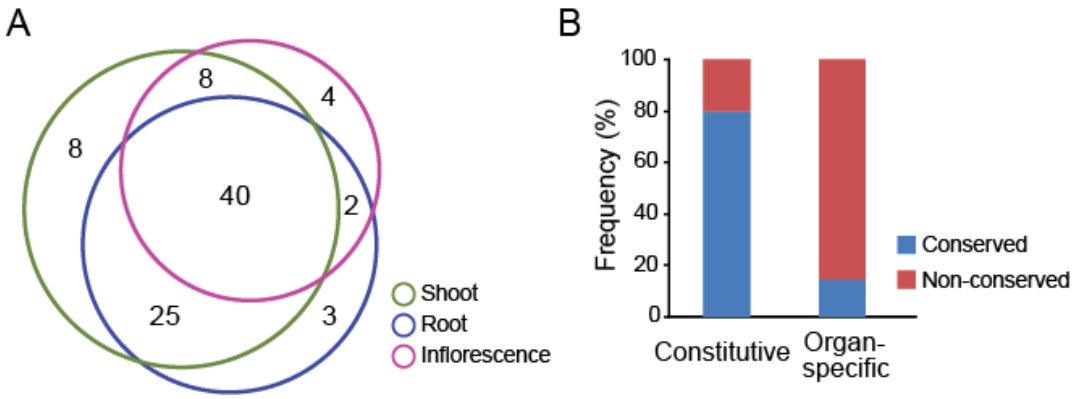
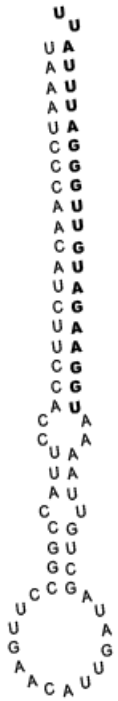

A
8
4
8
40
2
25
3
Shoot
Root
Inflorescence
B
100
80
60
40
20
0
Frequency (%)
Conserved
Non-conserved
Constitutive
Organ-specific

Figure 5. Analysis of putative novel *MIR* genes.

(A) Predicted pre-miRNA secondary structure of the candidate *MIR* gene *NM03*. The mature miRNA is shown in bold letters. (B) Northern blot analysis of selected candidate miRNAs. The expression pattern deduced from the sequencing data is shown on the left where a gray box represents expression and a blank box represents no expression. Northern blots hybridized with probes complementary to the mature miRNAs in wild type as well as *hen1* plants are shown on the right. MiR167 was used as a positive control for *HEN1*-dependent expression. Ethidium bromide staining of the low-molecular-weight fraction of total RNA was used as the loading control. S, shoot; I, inflorescence; R, root.

A



NM03

B

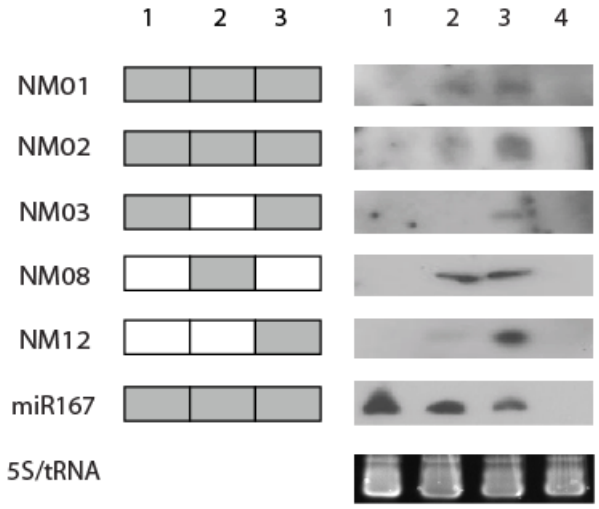


Figure 6. Simulation of the relation between *MIR* gene detection rate and sequencing depth by curve fitting.

Perfectly mapped sequence reads from the shoot library were randomly retrieved to create five different simulated sequencing depths. The number of expressed *MIR* genes was determined at each depth. The scatter plot represents results from three independent simulations and was used for curve fitting. The star sign indicates the actual data from the shoot library. Dashed lines indicate a 95% detection rate of the theoretic maximal number of expressed genes and the corresponding sequencing depth.

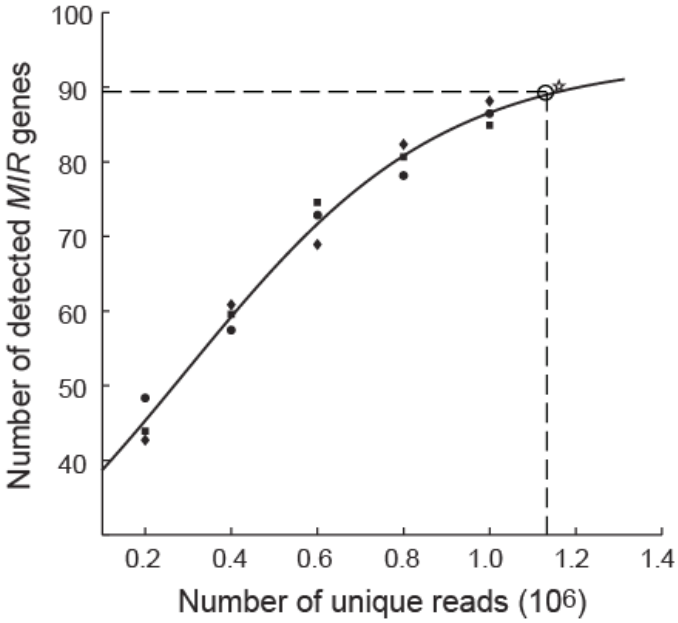

100
90
80
70
60
50
40
Number of detected MIR genes
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Number of unique reads (10^6)

Table 1. Sequence reads and detected miRNAs and *MIR* genes in three independent small RNA libraries from *Arabidopsis*.

Sample	Mapped reads[a]	Recovered miRNAs[b]	Expressed *MIR* genes[c]	Novel *MIR* genes
Shoot	1,235,891	105	81	9
Root	690,294	88	70	10
Inflorescence	589,535	65	54	6

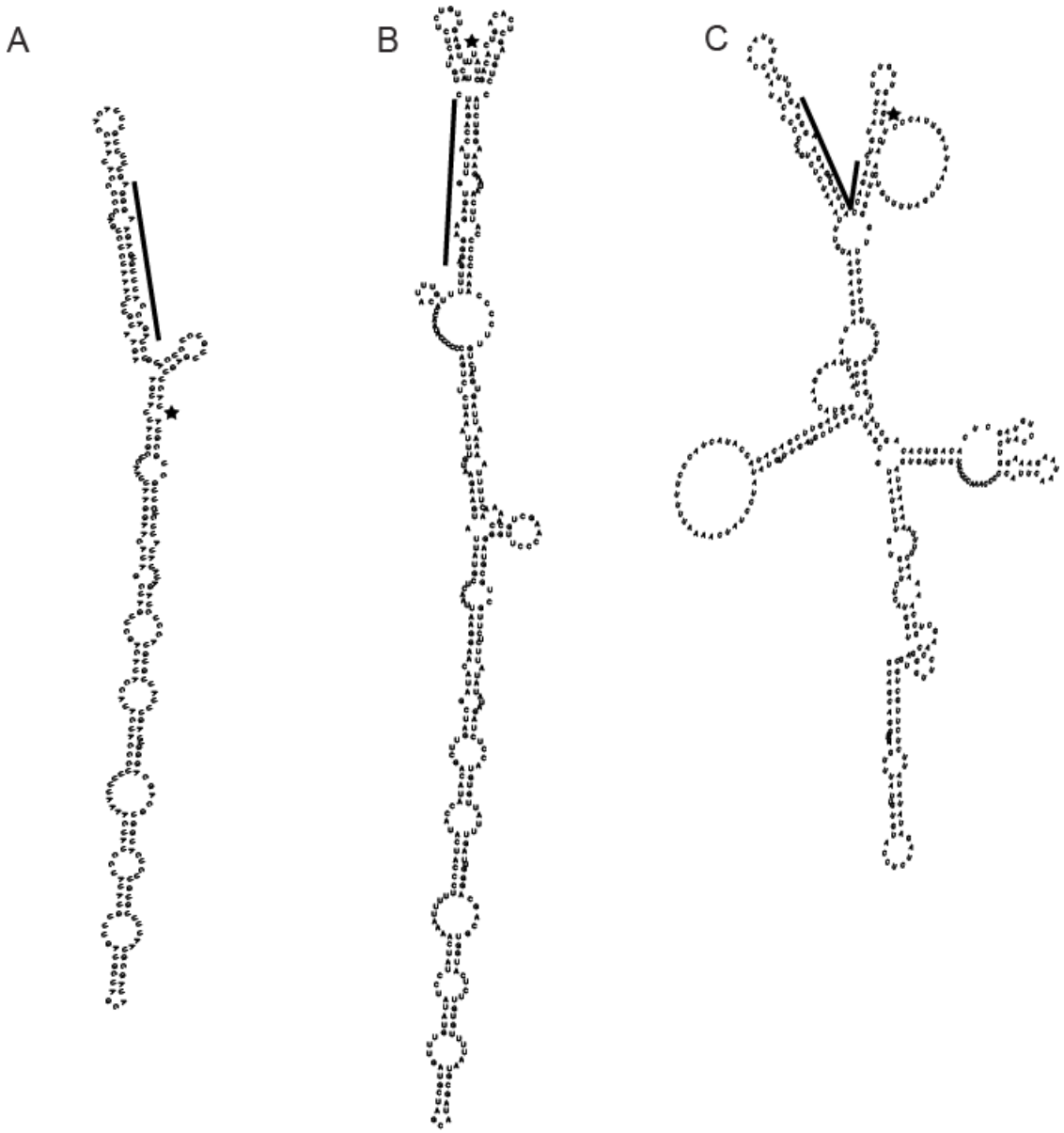
[a]Number of reads that could be perfectly aligned to the *Arabidopsis* genome sequence.

[b]Number of unique miRNAs with at least one matching sequence read retrieved.

[c]Number of the 199 annotated *MIR* genes considered to be expressed.

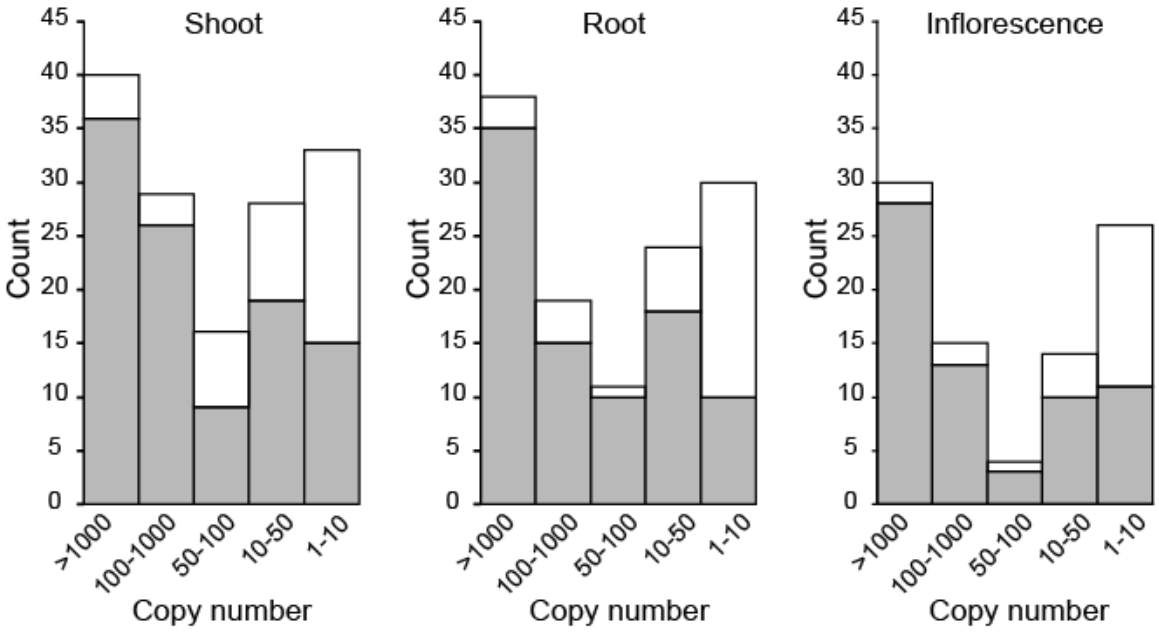
Supplemental Figure 1. Predicted secondary structure of *MIR*824 at different window sizes.

A) Window size of 250 nucleotides. B) Window size of 350 nucleotides. C) Window size of 400 nucleotides. Black bar indicates the position of the mature miRNA. Star indicates 5’ end of the sequence selected. Secondary structure is predicted by RNAfold.


A
B
C

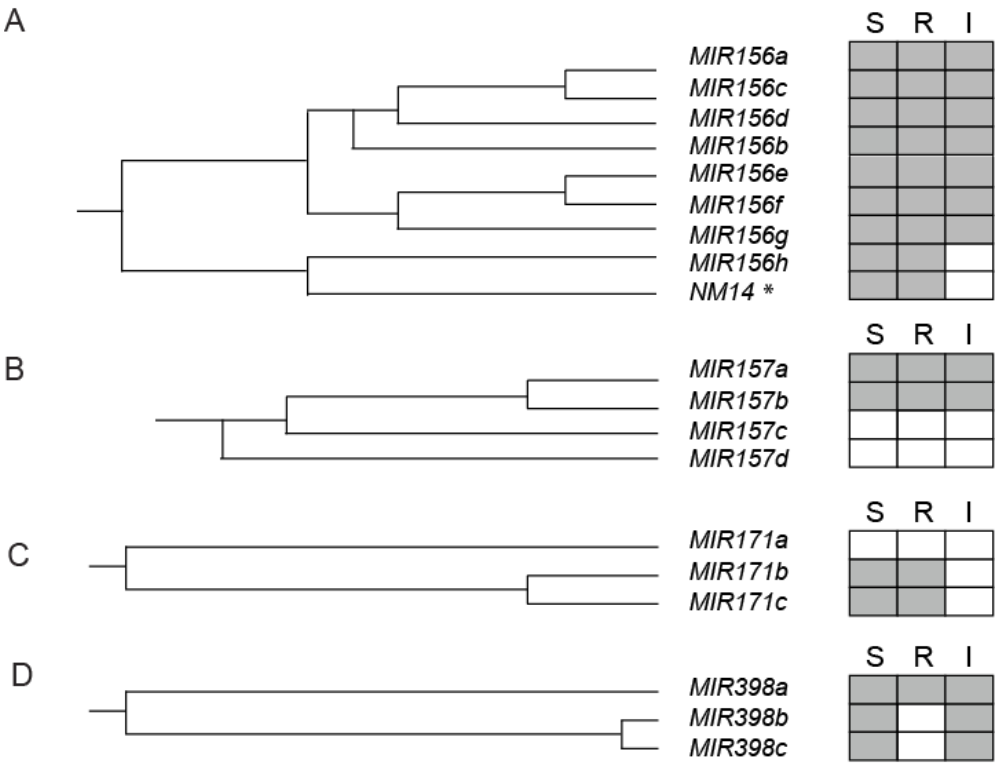
Supplemental Figure 2. Relation between copy number of retrieved mature miRNAs and the expression status of *MIR* genes.

MIR genes with at least one recovered sequence read corresponding to the mature miRNA are divided into five groups according to the frequency at which the mature miRNA was detected in the sequencing data. For each group, the number of expressed and non-expressed *MIR* genes was calculated. The grey bars represent expressed whilst the blank bars represent non-expressed genes.


Shoot
Root
Inflorescence
Count
0
5
10
15
20
25
30
35
40
45
>1000
100-1000
50-100
10-50
1-10
Copy number

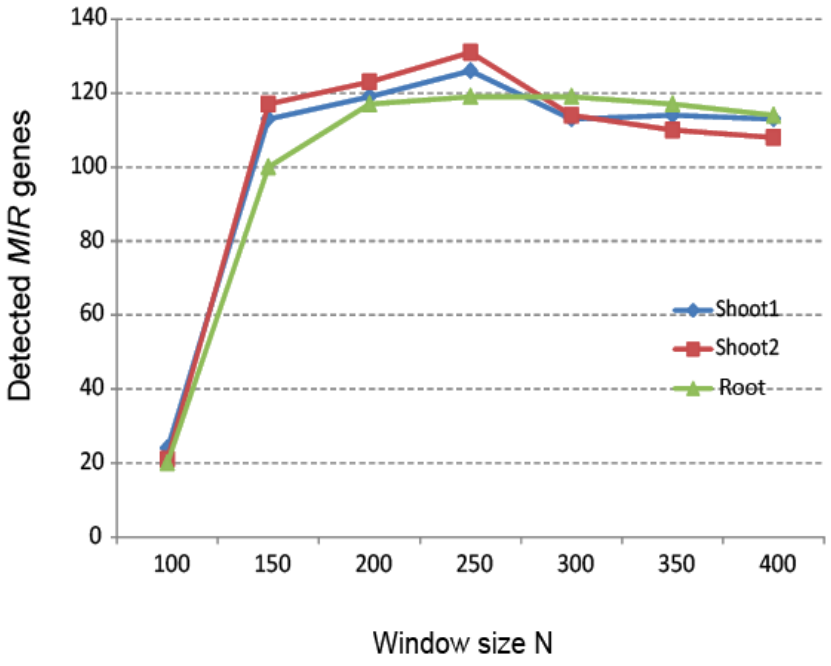
Supplemental Figure 3. Relating *MIR* gene expression to pre-miRNA phylogeny.

The annotated pre-miRNA sequences were used to construct a phylogenetic tree for each miRNA family. The gene-level expression profile of family members is depictured to the right of the tree. A) Relation of miR156 family. B) Relation of miR157 family. C) Relation of miR171 family. D) Relation of miR398 family. A grey box represents expression whilst a blank box represents no expression. S, shoot; R, root; I, inflorescence. Star (*NM14*) indicates the new predicted and validated member of miR156 family.


A
S R I
MIR156a
MIR156c
MIR156d
MIR156b
MIR156e
MIR156f
MIR156g
MIR156h
NM14 *
B
S R I
MIR157a
MIR157b
MIR157c
MIR157d
C
S R I
MIR171a
MIR171b
MIR171c
D
S R I
MIR398a
MIR398b
MIR398c

Supplemental Figure 4. Testing the detection rate of expressed *MIR* genes with different N values in Rice.

Small RNA reads from deep sequencing libraries prepared from root and shoot (two different experiments) were processed in miRDeep when different N values were used.


140
120
100
80
60
40
20
0
Detected MIR genes
100
150
200
250
300
350
400
Window size N
Shoot1
Shoot2
Root

Supplemental Table 1. Expression domain of annotated *MIR* genes in *Arabidopsis*.

MIR gene [a]	Shoot	Root	Inflorescence
MIR156a			
MIR156b			
MIR156c			
MIR156d			
MIR156e			
MIR156f			
MIR156g			
MIR156h			
MIR157a			
MIR157b			
MIR158a			
MIR158b			
MIR159a			
MIR159c			
MIR160a			
MIR160b			
MIR160c			
MIR162a			
MIR162b			
MIR164a			
MIR164b			
MIR165b			
MIR166b			
MIR166f			
MIR166g			
MIR167a			
MIR167b			
MIR167d			
MIR168b			
MIR169b			
MIR169d			
MIR169g			
MIR169h			
MIR169j			
MIR169k			
MIR169l			
MIR169m			
MIR169n			
MIR170			
MIR171b			
MIR171c			
MIR172b			
MIR172c			
MIR172d			
MIR172e			
MIR173			
MIR319b			
MIR390a			
MIR391			
MIR394a			

MIR395a			
MIR395b			
MIR395c			
MIR395d			
MIR395e			
MIR395f			
MIR397b			
MIR398a			
MIR398b			
MIR398c			
MIR399a			
MIR399c			
MIR399f			
MIR400			
MIR408			
MIR773			
MIR775			
MIR776			
MIR777			
MIR779			
MIR781			
MIR823			
MIR824			
MIR825			
MIR827			
MIR829			
MIR830			
MIR833			
MIR837			
MIR840			
MIR842			
MIR843			
MIR859			
MIR863			
MIR864			
MIR866			
MIR867			
MIR869			
MIR2111b			
MIR2936			

[a] Annotated *MIR* genes expressed in at least one of the three examined libraries. Grey color indicates expression.

Supplemental Table 2. Putative novel miRNAs from three sequenced small RNA libraries.

Family_ID	ID	Sequence	Position	Strand	I	R	S
FNM01	NM01	AGGCTTTTAAGATCTGGTTGCGG	Chr5:12025940..12025963	+			
FNM02	NM02	AGCTCGTTAAGACTAGATGGGGT	Chr5:11310054..11310077	+			
FNM03	NM03	TGGAAGATGCTTTGGGATTTATT	Chr1:11786348..11786370	+			
FNM04	NM04a	AGAAGACCCTTTAAAAACTCTTGT	Chr2:2965508..2965531	-			
	NM04b		Chr2:4098357..4098380	-			
	NM04c		Chr5:11611193..11611216	-			
	NM04d		Chr5:11379285..11379308	+			
FNM05	NM05a	TAAAAGGATTTACAAGGATTTTA	Chr1:13057666..13057688	+			
	NM05b	AAAAGATTTACAAGGATTTTA	Chr3:10296492..10296512	+			
FNM06	NM06	AGATGTGCAATGTGGATGGTCTA	Chr5:21821223..21821245	+			
FNM07	NM07	TTGTACAAATTTAAGTGTACG	Chr1:24554011..24554031	+			
FNM08	NM08	AGAAGCAAAATGACGACTCGG	Chr5:15758111..15758131	+			
FNM09	NM09	GCGGCGACGAAACGAACAGACT/	Chr3:13377098..13377120	+			
FNM10	NM10	ATTCGACAAAGTGAAGGGTTT	Chr5:22322872..22322892	-			
FNM11	NM11	TGTTTTGGATCTTAGATACAC	Chr2:19686960..19686980	+			
FNM12	NM12	ACTCATAAGATCGTGACACGT	Chr4:7844612..7844632	+			
FNM13	NM13	ATCCTTATTGATGATCTCTTAACA	Chr1:30019557..30019580	-			
miR156	NM14	TGACAGAAGAGAGAGAGCAC	Chr2:17589066..17589085	-			

Green color indicates library from which the miRNA was detected. R, root; I, inflorescence; S, shoot. Note that NM14 is a new member of the miR156 family.

Chapter 3. MiRDeep-P

miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants[1]

[1]Formatted as a co-authored manuscript and published as:

Yang X, Lei Li. 2011. *Bioinformatics.* 27(18):2614-2615

Abstract

Motivation: Ultra deep sampling of small RNA libraries by next generation sequencing has provided rich information on the microRNA (miRNA) transcriptome of various plant species. However, few computational tools have been developed to effectively de-convolute the complex information.

Results: We sought to employ the signature distribution of small RNA reads along the miRNA precursor as a model in plants to profile expression of known miRNA genes and to identify novel ones. A freely available package, miRDeep-P, was developed by modifying miRDeep, which is based on a probabilistic model of miRNA biogenesis in animals, with a plant specific scoring system and filtering criteria. We have tested miRDeep-P on eight small RNA libraries derived from three plants. Our results demonstrate miRDeep-P as an effective and easy-to-use tool for characterizing the miRNA transcriptome in plants.

Availability: http://faculty.virginia.edu/lilab/miRDP/

Introduction

miRNAs are an important class of endogenous small RNAs that regulate gene expression at the post-transcription level (Bartel, 2009). There has been a surge of interest in the past decade in identifying miRNAs and profiling their expression pattern using various experimental approaches (Wark et al., 2008). Most recently, deep sequencing of specifically prepared low-molecular weight RNA libraries has been used for both purposes in diverse plant species (Fahlgren et al., 2007; Zhu et al., 2008). A major drawback of these efforts is the exclusive focus on mature miRNAs, the final gene product, and ignorance of sequence information associated with other parts of the miRNA genes. New strategies and tools are thus highly desirable to analyze the increasingly available sequencing data to gain insights into the miRNA transcriptomes.

Although miRNAs are only 20 to 24 nucleotides long, they are processed from longer, stem-loop structured precursors called pre-miRNAs (Bartel, 2009). Maturation of miRNAs releases small RNAs derived from different parts of the stem-loop structure with asymmetric abundance. The program miRDeep employs a probabilistic model of miRNA biogenesis in animals to score compatibility of the nucleotide position and frequency of sequenced small RNA reads with the secondary structure of pre-miRNAs (Friedlander et al., 2008). However, two significant differences in miRNA precursors between animals and plants prevent straightforward adaptation of miRDeep to the plant systems. First, plant pre-miRNAs are much longer with more variable lengths. Second, more miRNAs in plants belong to paralogous families with multiple members encoding identical or near-identical miRNAs (Supplementary Materials 1 and 2). We have demonstrated that miRDeep modified with plant-specific parameters is useful in

analyzing the miRNA transcriptome in the model plant *Arabidopsis* (Yang et al., 2011). Here we describe the improved package, called miRDeep-P, and its applications in plants.

Application description

Workflow of miRDeep-P

Based on ultra deep sampling of small RNA libraries by next generation sequencing, miRDeep-P enables users to explore expression patterns of annotated miRNA genes and discover novel ones. Figure 1 illustrates the workflow of miRDeep-P. To run this application, the reads should be preprocessed by removing adapters, discarding reads shorter than 15 nucleotides and parsing them into FASTA format with their copy number recorded. With correctly formatted input files, miRDeep-P maps the reads to the reference (either genomic or transcriptomic) sequences using Bowtie (Langmead et al., 2009). For a given mapped read, the optimal size of the window from which to extract reference sequences for predicting RNA secondary structure has been shown to be 250 bp (Yang et al., 2011). However, miRDeep-P contains a module for users to empirically determine what window sizes to use in case a set of validated miRNA genes is available (Figure 1). The secondary structures of the extracted reference sequences along with all reads mapped to such sequences are processed by the miRDeep core algorithm (Friedlander et al., 2008) with a plant-specific scoring system (Supplementary Material 3). The output from the core algorithm is then filtered with additional plant-specific criteria based on known characteristics of plant miRNA genes (Meyers et al., 2008). The overall process quantifies the signature distribution of small RNA reads and thereby provides reliable information on the transcription and processing of the pre-miRNAs. miRDeep-P uses such information to effectively profile the miRNA transcriptome (Figure 1).

Identification of new miRNA genes

A major utility of miRDeep-P is to identify miRNA genes in plant species without detailed annotation. As long as there is sufficient read coverage, quantification of the signature small RNA distribution along reference sequences will be effective in revealing expressed pre-miRNAs from deeply sampled small RNA libraries. An advantage of miRDeep-P is that it outputs not only sequences of the putative miRNAs but also the stem-looped precursors and their location in reference sequences, which can be used to distinguish individual miRNA genes. Another advantage of miRDeep-P is that it does not require a priori information on sequence homology to known miRNA genes. This feature should be especially helpful to study the large complements of species-specific miRNA genes in plants (Fahlgren et al., 2007).

Determination of the expression status of individual miRNA genes

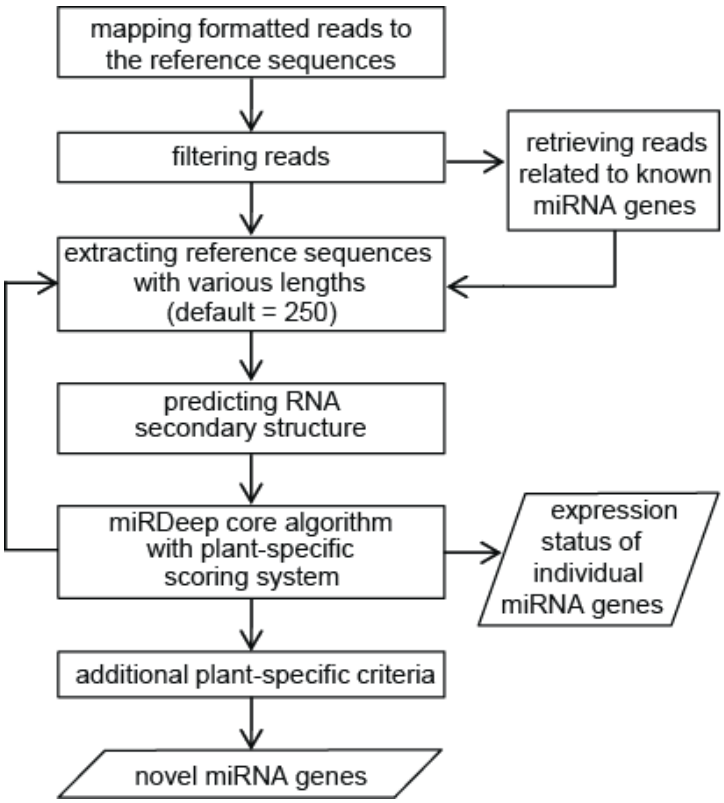
A novel application of miRDeep-P is to assign expression status to individual miRNA genes. Although normalized frequency of the reads matching the miRNAs can be used to estimate the expression level of miRNA genes (Fahlgren et al., 2007), the short length of the reads means cross contamination among paralogous genes due to sequence similarity is potentially an issue. In miRDeep-P, this issue is overcome by quantifying the signature distribution of reads along the entire length of the miRNA precursors. This feature is especially useful to determine the expression status of paralogous miRNA genes that encode identical mature miRNAs. Meanwhile, if multiple libraries prepared from different biological samples (e.g. leaf, root, etc.) are employed, expression profiling of individual miRNA genes can be achieved as well.

Implementation and results

The miRDeep-P package was developed in Perl by combining the core algorithm of miRDeep (Friedlander et al., 2008), the mapping tool Bowtie (Langmead et al., 2009), and the Vienna RNA package for predicting RNA secondary structure (Hofacker, 2003). Current version of miRDeep-P includes 9 Perl scripts, which can be executed sequentially in a command line environment. All scripts have been tested on two Linux platforms, SUSE 10 and Fedora 14, and should work on similar systems that support Perl. The miRDeep-P scripts and user manual can be obtained from http://faculty.virginia.edu/lilab/miRDP/index.html as well as http://sourceforge.net/projects/mirdp/.

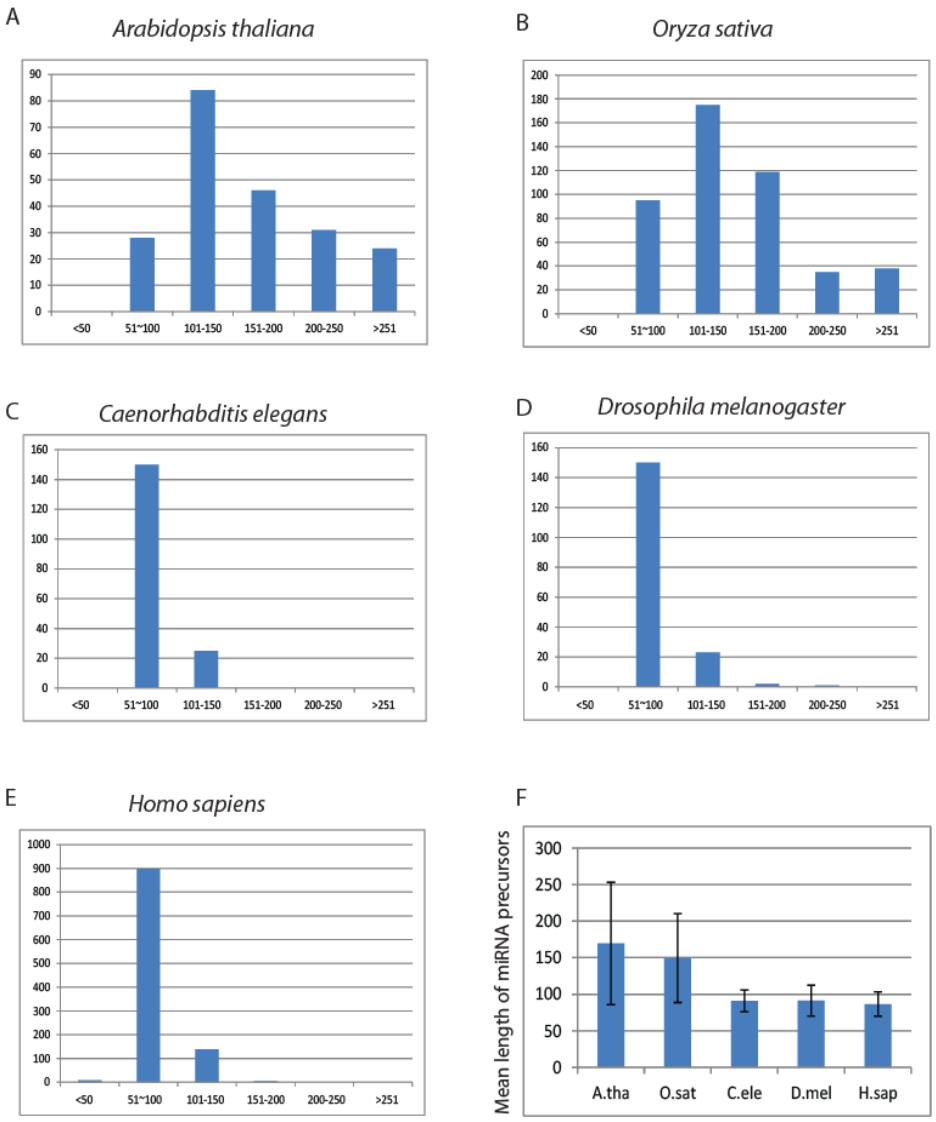
miRDeep-P has been tested using eight small RNA libraries from three plant species, *Arabidopsis*, rice and papaya. Both *Arabidopsis* and rice are well annotated for miRNA genes while there is no annotation in papaya. Based on these tests, it has been shown the optimal window size for extract precursor reference sequences is 250 bp for both dicot and monocot plants (Yang et al., 2011). From the three *Arabidopsis* libraries, a total of 108 expressed (90 annotated and 18 novel) miRNA genes were detected. The two rice libraries yielded 158 annotated and 51 novel miRNA genes. Results from *Arabidopsis* have been successfully validated using other experimental approaches (Yang et al., 2011), demonstrating the reliability of miRDeep-P. From the three papaya libraries, we detected 104 putative expressed miRNA genes of which 56 are conserved in other plant species and 48 are novel, further indicating that miRDeep-P is of broad use in plants.

Figure 1. Diagram of the workflow of miRDeep-P.


mapping formatted reads to the reference sequences
filtering reads
retrieving reads related to known miRNA genes
extracting reference sequences with various lengths (default = 250)
predicting RNA secondary structure
miRDeep core algorithm with plant-specific scoring system
expression status of individual miRNA genes
additional plant-specific criteria
novel miRNA genes

Supplementary Material 1. Distribution of miRNA precursors at length of five model organisms.

(A~E) miRNA precursors distribution of five model organisms at different length. Horizontal axis indicates the length of miRNA precursors. Vertical axis indicates the number of miRNA precursors at specic length. (F) Mean length of miRNA precursors of five model organisms. Black bar indicates standard deviation. Note: original data from miRBase release 16.


A
Arabidopsis thaliana
B
Oryza sativa
C
Caenorhabditis elegans
D
Drosophila melanogaster
E
Homo sapiens
F
Mean length of miRNA precursors
<50
51~100
101-150
151-200
200-250
>251
A.tha
O.sat
C.ele
D.mel
H.sap

Supplementary Material 2. Distribution of miRNA genes from multiple member families of four model organisms.

MIR indicates miRNA gene. Mul-family means this family has more than one *MIR*s. Original data from miRBase release 16.

Organisms	No. of MIRs of Mul-families	No. of total MIRs	Percentage
A.tha	113	213	53.05%
O.sat	325	462	70.35%
C.ele	18	175	10.29%
D.mel	29	176	16.48%

Supplementary Material 3. The plant-specific scoring system of miRDeep-P.

The miRDeep core algorithm scores each potential miRNA precursor based on a number of features related to miRNA biogenesis. Because miRNA precursors are more thermodynamically stable than non-precursor stem-loops, the minimum free energy of the predicted secondary structure also contributes to the score. The probability (log-odds) of each feature contributing to the score is estimated by parameter fitting to known and background miRNA precursors in animals (Friedlander et al., 2008). However, plant miRNA precursors are generally much longer (Llave et al., 2002; Reinhart et al., 2002). With the increase of the length of the miRNA precursors in question, higher scores from the minimum free energy ensue. In practice, this causes inflation of the output score or collapse of the core algorithm altogether in some cases. Our solution to this problem was to set a maximal value of the log-odds score from the minimum free energy based on the Gumbel distribution. When the score from the secondary structure is lower than the maximal value allowed, the original score will be retained. In contrast, when higher scores are encountered, the maximal value will be used instead. In miRDeep-P, this function is realized by two if conditional statements in the subroutine *score_mfe of miRDeep.pl*.

Chapter 4. MicroRNAs in 15 plant species

Comprehensive annotation, profiling, and comparison of the expressed complements of microRNAs in 15 land plant species

Abstract

MicroRNAs (miRNAs) are important post-transcriptional gene regulators that extensively regulate the transcriptome and many biological processes. In the last decade, deep sequencing efforts and functional studies have greatly expanded knowledge about miRNAs in model plants. However, lack of substantial miRNA information in non-model systems and limited comparison cross species have largely constrained comprehensive understanding of miRNA function and evolution in plants. By parsing 78 deeply sequenced small RNA libraries from 15 land plants using the computational tool miRDeep-P, we systematically identified 4,142 expressed miRNAs along with the precursor sequences, including 1,116 annotated miRNAs, 740 conserved but unannotated miRNAs, and 2,286 novel miRNAs. Comparative analysis of this unprecedented large dataset led to several major findings. First, we found that 18% of the identified miRNAs in the examined species locate in close proximity. The compact genome organization of these miRNAs suggests several possible miRNA biogenesis mechanisms utilizing existing precursors. Second, we found that expressed miRNAs exhibit distinct phylogenetic distribution and confirmed that almost two thirds of expressed miRNAs are species-specific. Third, we demonstrated that deeply-conserved miRNAs such as the miR156/157 family have experienced lineage-specific diversification, which further underscores the dynamic nature of miRNA-based regulation. Taken together, comprehensive annotation and comparative analysis revealed prevalent divergence of the miRNA transcriptome in plants due to de novo biogenesis of unique miRNAs and lineage specific duplication of conserved miRNAs. This finding suggests that miRNAs are relevant to genetic novelty and phenotypic diversity in plants.

Introduction

One of the most exciting biological findings in recent years is the discovery that small RNA species regulate diverse spatial and temporal function of the genome, including chromosome segregation, chromatin modification, RNA processing, transcriptional regulation, as well as translational control (Huttenhofer et al., 2005; Bartel, 2009; Carthew and Sontheimer, 2009; Voinnet, 2009). In particular, miRNAs are emerging as an important class of endogenous gene regulators acting at the post-transcriptional level in both animals and plants. Like protein-coding genes, most miRNAs are transcribed by RNA polymerase II (Cai et al., 2004; Lee et al., 2004; Zhao et al., 2013). The primary transcripts known as pri-miRNAs are processed via stem-loop structured intermediates called pre-miRNAs to give rise to mature miRNAs, which are typically 20- to 24-nucleotides (nt) in length (Bartel, 2009; Voinnet, 2009).

Unique to higher plants, pri-miRNA and pre-miRNA processing are both carried out in the nucleus primarily by the endonuclease Dicer-like 1 (DCL1) (Papp et al., 2003; Kurihara and Watanabe, 2004). Following these processing events, a duplex of two complementary short RNA molecules is generated, which is end-methylated (Yang et al., 2006) and then transported to the cytoplasm (Park et al., 2005). Only the mature miRNA is integrated into the RNA-induced silencing complex (RISC) whereas the passenger strand, called miRNA*, is typically degraded as a RISC substrate (Khvorova et al., 2003; Schwarz et al., 2003; Bartel, 2009). After loading into RISC, miRNAs base pair with the complementary sequences within the target miRNA and direct either cleavage (Llave et al., 2002; Reinhart et al., 2002) or translational repression (Brodersen et al., 2008) of the

target transcripts. Gene silencing by miRNA-directed DNA methylation at the target loci has also been reported in plants (Wu et al., 2010).

Since the initial discovery that fortuitously implicated miRNAs in the development of *C. elegans* (Lee et al., 1993; Wightman et al., 1993), there has been an accelerated surge of interest in the past two decades in identifying miRNAs using various experimental approaches. In plants, much of the effort to identify, experimentally validate, and functionally characterize miRNAs has been concentrated on a few model plants. It is well-established that miRNAs play important roles in plant development and response to environmental challenges (Voinnet, 2009). Meanwhile, encouraged by success of next generation sequencing technologies, deep sampling of specifically prepared low-molecular-weight RNA libraries has become a popular approach to identify miRNAs for functional and evolutionary studies in diverse plants (Rajagopalan et al., 2006; Friedlander et al., 2008; Zhu et al., 2008; Xie et al., 2010; Zhu et al., 2012). In addition to validating annotated miRNAs, large numbers of putative new miRNAs have been identified from these efforts.

Although annotation for hundreds of miRNAs from approximately 46 plant species is available in miRBase (Kozomara and Griffiths-Jones, 2011), the number of plant lineages with extensive miRNA annotation is still small. Many putative miRNAs identified from accumulating deep sequencing data lack precursor information (Xie et al., 2010; Zhu et al., 2012). Further, in many plant species only miRNAs belonging to the most conserved families are identified while it is known from well-annotated species that the majority of miRNAs are recently evolved and species-specific (Barakat et al., 2007; Fahlgren et al., 2007). In fact, newly accumulated evidence suggests that species-specific

or family-specific miRNAs are functional constituents of the miRNA-mediated regulatory networks and underscore the dynamic nature of these networks (Lu et al., 2011; Ng et al., 2011; Wang et al., 2011). Thus, it is highly desirable to elucidate the full spectrum of miRNAs in diverse plant lineages using new strategies and tools to gain a comprehensive understanding of miRNA origin, evolution, and function in plants.

miRDeep-P is a computational program developed on a model of miRNA biogenesis to retrieve miRNA related information from deep sequencing data in plants (Yang and Li, 2011; Yang et al., 2011). After mapping the reads to a reference genome, this program extracts the sequence flanking each anchored read for predicting RNA secondary structure and quantifying the compatibility of the distribution of small RNA reads with Dicer-mediated processing. After progressively processing all mapped reads, candidate miRNAs are scored based on a probabilistic model (Friedlander et al., 2008) and filtered with plant-specific criteria (Yang and Li, 2011; Yang et al., 2011). We have previously demonstrated that miRDeep-P works effectively for deep sequencing data in *Arabidopsis* and rice to retrieve known and novel miRNAs as well as the corresponding pre-miRNAs (Yang and Li, 2011; Yang et al., 2011). Subsequently, this method was used to identify miRNAs in a variety of other plant species including papaya, peach (Colaiacovo et al., 2012), Chinese tulip tree (Wang et al., 2012), sugarcane (Gentile et al., 2013) and potato (Zhang et al., 2013).

Towards a comprehensive understanding of plant miRNAs, we carried out an effort centered on miRDeep-P to parse 78 deeply sequenced small RNA libraries totaling 330 million reads from 15 representative land plant species. We identified 4,142 miRNAs along with the precursor sequences that include 1,116 annotated miRNAs, 740 conserved

but unannotated miRNAs, and 2,286 novel miRNAs. We show that this unprecedented large dataset likely represent expressed and functional miRNAs, bringing us one step closer to elucidating the full spectrum of plant miRNAs. Compared to previous comparative works in plants (Barakat et al., 2007; Sunkar et al., 2008), by focusing on expressed miRNAs with accompanying precursors and expanding the taxonomic representation with extensive miRNA information, the work reported here provided new perspectives on understanding miRNA function and evolution in plants.

Results and discussion

Identification of miRNAs from deep sequencing data in 15 plant species

Toward a comprehensive annotation and comparison of miRNAs in land plants, we selected 15 species (Figure 1) for which annotated or draft genome sequences and deep sequencing data of low-molecular-weight RNA are available. These species are phylogenetic representatives of the land plants that include dicotyledons, monocotyledons as well as a moss (*Physcomitrella patens*) (Figure 1). Because most studies on plant miRNAs have been carried out in *Arabidopsis* (*Arabidopsis thaliana*) and rice (*Oryza sativa*), both species are selected in this study. We also chose several species in which no miRNA annotation is yet available in miRBase. These include tomato (*Solanum lycopersicum*), monkey-flower (*Mimulus guttatus*), peach (*Prunus persica*), switchgrass (*Panicum virgatum*), and orange (*Citrus sinensis*).

To effectively and accurately identify candidate miRNAs and pre-miRNAs from deeply sampled small RNA libraries in plants, we developed a computational pipeline centered on the program miRDeep-P (Yang and Li, 2011; Yang et al., 2011) (Figure S1). Tested in both *Arabidopsis* and rice, it was previously found that 250 nucleotides (nt) represent the optimal length for exacting a local sequence flanking each read mapped to the genome to predict the secondary RNA structure and quantify read distribution (Yang and Li, 2011; Yang et al., 2011). First, we examined 10 species for which a significant number of miRNAs have been annotated (Figure 2), including representatives of moss, monocots, and dicots, to test this parameter for miRDeep-P. For each mapped read, local genomic sequences of different lengths were used to detect expressed miRNAs by miRDeep-P in these species. The number of detected miRNAs as a function of the

sequence length was plotted and examined (Figure 2). Although not all annotated miRNAs were detected due to tissue specificity of miRNA expression and unsaturated sequencing depth, the curves exhibit essentially the same trend and all plateau around 250 nt in all species (Figure 2). Considering both coverage and efficiency, we concluded that 250 nt is optimal for applying miRDeep-P in diverse plant species.

Due to distinctions of the miRNA biogenesis pathway, pre-miRNAs in animals and plants exhibit conspicuous differences. Animal pre-miRNAs generally have a uniform size (~ 70 nt) because of little variation in the loop sizes and fixed position of the miRNA:miRNA* duplex within pre-miRNAs (Han et al., 2006; Yang and Li, 2011). By contrast, the length of pre-miRNAs in plants is considerably longer and more variable (Meyers et al., 2008). The implications of result shown in Figure 2 are twofold. Firstly, it demonstrate that a 250-nt sequence context is sufficient to detect the core region of the miRNA precursors in diverse plants and thus the feasibility of broadly applying miRDeep-P. Secondly, it provides the platform for correlating secondary structure with reads distribution to virtually reconstitute the miRNA precursors, which should be especially useful for annotating miRNAs in underrepresented plant species.

For the 15 species, we collected and parsed 78 deeply sequenced small RNA libraries totaling 330 million reads (Figure 1; Table S1). The identified putative pre-miRNAs were further filtered with regard to the minimal length including the miRNA/miRNA* duplex and the encircled loop region. Previous molecular studies revealed that small loop may create structural constraints that hinder *DCL1* processing during miRNA biogenesis both in animals (Mateos et al., 2010; Tsutsumi et al., 2011), and that the distance between the mature miRNA and the miRNA* within *bona fide* pre-

miRNAs is normally no less than 14 nt (Friedlander et al., 2008). Given that the length of mature miRNA and miRNA* is up to 24 nt, 62 nt is thus considered the necessary length for legitimate precursors harboring the miRNA/miRNA* duplex and the loop region. In effect, miRDeep-P was used to identify pre-miRNAs in the size range of 62-250 nt with qualified secondary structure and read distribution.

In total, we identified from the deep sequencing data 4,142 expressed miRNAs together with the corresponding precursor sequences for the 15 examined plant species (Figure 1; Table S2). Searching against all annotated plant miRNAs in miRBase (release 17.0; Kozomara and Griffiths-Jones, 2011), we found that our dataset included 1,116 annotated miRNAs, 740 conserved but unannotated miRNAs, and 2,286 novel miRNAs (Figure 1). This unprecedented large dataset of miRNAs with the accompanying precursor sequences in diverse species helps to fill several major gaps in our knowledge of plant miRNAs, which are elaborated below.

Retrieval of precursors facilitates cross-species comparison of plant miRNAs

For the 4,142 expressed miRNAs in 15 species, we identified a total of 3,861 precursors. It should be noted that the precursors were outnumbered by the mature miRNAs because there were cases where more than one miRNA could be derived from a single precursor (see below). Having the precursors in our dataset increased confidence in the identified miRNAs, revealed genome organization of the miRNA genes, and facilitated comparison across plant species. Previous sequence and phylogenetic analyses based on the mature miRNAs revealed about 21 families that appear to be deeply conserved in angiosperms (Barakat et al., 2007; Axtell and Bowman, 2008; Cuperus et

al., 2011). Except miR159 and miR319, the average length of pre-miRNAs for the conserved families was about 80 nt (Figure 3A). Compared between dicots and monocots, several noticeable exceptions were found. For example, the average pre-miRNA length of miR168 in dicots was significantly longer than that in monocots. By contrast, the average pre-miRNA length of miR408 was much longer in monocots (Figure 3A). Further, compared to the 21 conserved families, average precursor length of the novel miRNAs increased by about 15 nt (Figure 3B), suggesting a general trend of decreasing precursor length as the miRNAs continue to evolve in plants.

MiR408 is annotated in 18 plant species in the miRBase, which places it among the most conserved miRNA families in land plants. Computational and experimental evidence indicates that miR408 targets genes encoding copper-containing proteins that include three members of the Laccase family and *Plantacyanin* (Yamasaki et al., 2007; Abdel-Ghany and Pilon, 2008; Yamasaki et al., 2009; Zhang and Li, 2013). In this study, we confirmed expression of miR408 in nine species and identified miR408 from six species in which this miRNA has not been annotated, including switchgrass, soybean (*Glycine max*), Medicago (*Medicago truncatula*), peach, tomato, and monkey-flower (Figure 3C). Additionally, we conducted similarity searches against the genomic sequences of the 15 examined species. Even when two mismatches were allowed, no sequence homologous to mature miR408 was found besides the identified miR408 loci, indicating that the miR408 sequence is under strong purifying selection.

We determined the stem-loop secondary structure of all pre-miR408s and also mapped the reads distribution along the primary sequence of the precursors (Figure 3C). Comparison of these two profiles revealed that expansion of the loop region caused

increased pre-miR408 length in monocots (Figure 3C). Further, the complementarity of the stem region encompassing mature miR408 is only similar among closely related species. For example, this region in monocots contains four small bubbles caused by four mismatches. Interestingly, the relative abundance of miR408 and miR408* is highly variable. In both dicots and monocots there are species in which miR408* is more abundant than the mature miRNA (Figure 3C). Taken together, these results suggest that while the mature miRNA of the conserved miR408 family is under strong selection, the precursors are continuingly evolving in diverse plant lineages.

For 13 of the 15 examined species, miR408 is encoded only by a single gene. Exceptions are switchgrass and soybean. In soybean, four miR408 loci were identified, which could be divided into the Gma-miR408a/b and Gma-miR408c/d groups based on comparison of the sequences and secondary structures of pre-miRNAs (Figure 3C). In switchgrass, although only three miR408 isoforms were discovered, searching the genomic assembly revealed an additional locus located at the very end of a contig that is highly similar to the identified pre-miR408s. Thus it is likely that switchgrass also encodes four miR408 loci as is the case in soybean. Interestingly, both soybean and switchgrass have the equivalent of a tetraploid genome. Taken together, these results suggest that there are independent miR408 duplications associated with polyploidy in either the monocot or dicot clade.

Genome organization of neighboring miRNAs suggests possible mechanisms for miRNA biogenesis

When aligned to the genome, we found that many of the identified miRNAs cluster together or locate in close proximity with annotated miRNAs. Closer inspection revealed that these miRNAs, which account for 18% of the identified miRNAs in the examined plant species (Tables S2 and 3), are organized following four distinct scenarios (Figure 4A). In scenario I, multiple miRNAs are derived from tightly clustered pre-miRNAs, which are widespread in the examined species and account for approximately 6% of all miRNAs (Table S3). In *Arabidopsis* and rice, it was noted that some of the most conserved miRNA families are so organized (Li and Mao, 2007). Co-transcription of clustered miRNAs as a single unit was experimentally demonstrated in *Arabidopsis* (Merchan et al., 2009). We found that six conserved miRNA families, miR156, miR166, miR167, miR169, miR395, and miR399, all include members organized in this fashion in various plant species. As an example, the Ath-*MIR*395b and Ath-*MIR*395c cluster in *Arabidopsis* together with the reads distribution is illustrated in Figure 4B. High sequence similarity between these clustered miRNAs indicates that they are likely derived from tandem duplications (Li and Mao, 2007). In addition, we identified many clustered pre-miRNAs that produce non-homologous miRNAs, which might be functionally related (Merchan et al., 2009) or have correlatively evolved (Zhang et al., 2011).

In scenario II, two miRNAs are separately processed from a shared precursor (Figure 4A). Collectively 6% miRNAs in all examined species are potentially generated following this scenario (Table S3). Known examples include miR159 and miR319. These two families possess precursors that can be sequentially processed to produce distinct miRNAs from different parts of the stem region (Addo-Quaye et al., 2009; Bologna et al., 2009). At least in the case of miR159a, it was shown that the additional miRNA

processed from the same precursor is conserved in different plant species and exhibits expression pattern different from that of miR159a (Contreras-Cubas et al., 2012). We found that the mechanism of utilizing miR159 and miR319 precursors, which are longest among the conserved miRNA families (Figure 3A), to produce multiple miRNAs is conserved in different plant species. Further, we found that many of the previously unknown miRNAs are produced in this way. The example shown in Figure 4C depicts that Osa-miR820a and a newly identified miRNA in rice are derived from the same pre-miRNA. Sequence inspection indicates that the precursor possesses features similar to those described for noncanonical miRNA processing as in the case of miR159 and miR319 (Addo-Quaye et al., 2009; Bologna et al., 2009). Further, analysis of copy number and reads distribution revealed that Osa-miR820a and the new miRNA are both highly expressed (Figure 4C). Taken together, these results indicate that scenario II represents a common mechanism for miRNA biogenesis in plants.

In our dataset, approximately 6% of miRNAs are generated following scenario III (Table S3), which differs from scenario II in that the two miRNAs processed from the same precursor are overlapping and thus mutually exclusive (Figure 4A). An example involving Osa-miR1868 and a newly identified miRNA in rice is depicts in Figure 4D. The abundant reads correspond uniquely to each of the two miRNA/miRNA* duplexes indicate both are expressed (Figure 4D). Further, to be sure that reads corresponding to the new miRNA are indeed unique, we searched the rice genome and found that only this locus matches the reads. It should be stressed that different miRNAs derived from shared precursors in scenario III are offset by at least 4 nt, which may completely change their target specificity given that plant miRNAs typically share high complementarity over its

entirety to their targets. Therefore, we considered these distinct miRNAs that compete for the same precursors.

Scenario IV is relatively rare in which two miRNAs are derived from two precursors transcribed in opposite direction on complementary DNA strands at the same locus (Figure 4A). When both precursors are separately transcribed and processed, two identical or near identical miRNAs are produced. In the 15 examined plants, approximately 1% of miRNAs belongs to scenario IV and includes both conserved miRNAs as well as miRNAs unique to individual species (Table S3). As an example Ath-miR781 is shown in Figure 4E. This miRNA is encoded by two genes transcribed from complementary strands of the same locus. Reads distribution indicates that the two precursors are separately processed to produce an identical mature miRNA but two slightly different miRNA* (Figure 4E). Sequence analysis further revealed that this locus represents a highly eroded inverted duplication, indicating that only the regions corresponding to the mature miRNA and miRNA* sequences are under strong selection.

There are two prerequisites for successfully producing a mature miRNA from a genome locus. One is that the locus encodes a stem-loop structure compatible with Dicer processing and the other is that the locus is transcribed. Several models regarding de novo biogenesis of new miRNAs in plants have been proposed that involve inverted duplication of various origins to suffice the two prerequisites (Felippes et al., 2008; Voinnet, 2009). Results presented above indicate that utilization of existing miRNA precursors, therefore bypassing either or both requirements, is also a mechanism for generating new miRNAs. For example, multiple miRNAs processed from one primary

transcript (model I in Figure 4) would eliminate the need for a new promoter to drive precursor expression.

Both models II and III represent cases in which more than one miRNA is produced from the same precursor (Figure 4). Production of miRNAs under these scenarios is consistent with our biochemical knowledge of miRNA maturation. In addition to conventional miRNA processing in which the miRNA/miRNA* duplex is first chopped out of the stem-loop precursor (Mateos et al., 2010; Werner et al., 2010), it has been shown in the case of miR159 and miR319 that the common precursor is first sliced in the loop region and then two duplexes are sequentially released to generate two independent and functional miRNAs (Addo-Quaye et al., 2009; Bologna et al., 2009; Contreras-Cubas et al., 2012). Thus, the biochemical foundation exists in plants for producing miRNAs according to scenarios II.

Significant nucleotide difference in mature miRNAs due to shifted processing was first noticed when miR171 and miR172 were compared between fern and *Arabidopsis* (Axtell and Bartel, 2005). Lu et al. (Lu et al., 2007) also found that the miR156 sequence of loblolly pine has shifted 3-4 nucleotides toward the 3’ end relative to the homologs in other plant species. These findings indicate that different species might select mature miRNAs from different positions of the stem region of conserved precursors. Most recently, Manavella et al. (Manavella et al., 2012) demonstrated that the *CPL1/2* genes play a role in determining what positions within the precursor are used for mature miRNAs selection in *Arabidopsis*. Without *CPL1*, which regulates dephosphorylation of *HYL1*, multiple miRNA/miRNA* duplexes with shifted positions would be processed from the same precursor (Manavella et al., 2012). An implication of this finding is that

other regions of the pre-miRNA compete with the canonical miRNA for Dicer-mediated processing. While guarding cellular mechanisms are in place to favor the canonical miRNA, it is tempting to speculate that relaxation of these mechanisms or mutations in the precursors could have the potential to generate new miRNAs with shifted nucleotides. It should be noted that producing overlapping miRNAs from the same precursor not only diversifies the miRNA portfolios but also could represent a mechanism for regulating the abundance of the miRNAs as they are mutually exclusive. Further inspection of miRNAs derived from these scenarios thus offers new perspectives for tracing miRNA biogenesis and evolution.

Conserved miRNAs exhibit distinct phylogenetic distribution

Another major outcome of our effort was the addition of significant numbers of expressed miRNAs to plant species in which miRNAs are underrepresented (Figure S2A). For example, no miRNA has been identified in monkey-flower. Though miRNAs have been analyzed based on sequence similarity and small RNA sequencing in switchgrass and peach (Xie et al., 2010; Zhu et al., 2012), none has been annotated in miRBase presumably due to lack of reliable precursor information. For these three species, we were able to identify 1,033 miRNAs (Figure 1) including miRNAs from all the 21 conserved families (Figure 5). By significantly expanding the miRNA reservoir in plants, we made an intriguing observation that the density of miRNAs is strongly correlated ($R2 > 0.9$) with the density of protein-coding genes in the examined plant species (Figure S2B). On average, we found there are one miRNA for approximately every 100 protein-coding genes in plants (Figure S2B). Interestingly, similar

miRNA/protein-coding gene ratios have been reported in several animals as well (Lai et al., 2003; Lim et al., 2003a; Lim et al., 2003b).

Based on detailed comparison across the 15 species, we observed that 1,543 (37%) of the 4,142 expressed miRNAs could be found in at least two plants. These are referred to as conserved miRNAs hereafter. Consistent with previously reported comparisons of smaller scale (Barakat et al., 2007; Sunkar et al., 2008), we found that the conserved miRNAs have distinct phylogenetic distribution. We detected expressed members of seven miRNA families (miR156, miR160, miR166, miR171, miR319, miR390, and miR408) from all examined species including flowering plants and moss (Figure 5), suggesting that these miRNAs are fundamental to embryophytes. Meanwhile, we found that 14 miRNA families (miR159, miR162, miR164, miR167, miR168, miR169, miR172, miR393, miR394, miR395, miR396, miR397, miR398, and miR399) were expressed in all species other than moss (Figure 5), suggesting that these miRNAs originated at a later time in evolution after the split of moss from the ancestors for angiosperms.

Contrary to the 21 families of widely conserved miRNAs, there are many expressed miRNAs exhibit much narrower phylogenetic distribution. For example, four miRNA families (miR157, miR403, miR530, and miR2111) were inconsecutively expressed in the eudicot species but none of the examined monocots (Figure 5). Conversely, five miRNA families (miR528, miR1432, miR1878, miR2275, and miR5072) were detected specifically in monocot species (Figure 5). In addition, we found many miRNAs that appear to be family-specific. For instance, we detected eight miRNAs (e.g. miR1507, miR1509, and miR1510) specifically in legume species including soybean

and *Medicago* but none other plants (Figure 5). Similarly, a total of 29 miRNA families (e.g. miR158, miR165, and miR173) were only detected in Brassicaceae including *A. thaliana* and *A. lyrata* (Figure 5).

To validate the finding shown in Figure 5, we isolated RNA from nine plant species including five dicots and four monocots. We then selected eight representative miRNAs with different degree of conservation and performed RNA blotting analysis. As shown in Figure 6, expression pattern of these miRNAs in the tested plant species is fully consistent with results deduced from the sequencing data. Together these results confirmed that there are conserved miRNAs with narrow phylogenetic distribution.

Expressed miRNAs specific to individual species

In contrast to the miRNA families discussed above, we identified a large collection of miRNAs that were only found in a single species. In fact, of the 4,142 expressed miRNAs, 2,599 (63%) were species-specific (Figure 7A). In moss, 71% of miRNAs were found only in this species. Among the 14 flowering species, the proportion of species-specific miRNAs ranged from 37% in *Arabidopsis* to 77% in switchgrass (Figure 7A). The species-specific miRNAs were generally thought to be young in evolution and have very low expression (Rajagopalan et al., 2006; Fahlgren et al., 2007). However, we noticed frequently that the species-specific miRNAs were represented by hundreds even thousands of reads in the sequencing data. This observation prompted us to examine the expression level of miRNAs using their normalized reads frequency. Overall, the top 6% highly expressed species-specific miRNAs exhibited expression level higher than the median of conserved miRNAs (Figure 7B), which is consistently the case

for individual species a well (Figure 7B). This result indicates that at least a portion of the young miRNAs in plants is abundantly expressed.

We chose *Arabidopsis*, which is the model plant with miRNAs best annotated, to further characterize the species-specific miRNAs. In *Arabidopsis*, we identified a total of 40 species-specific miRNAs belonging to 36 families (Table S4). All these candidate miRNAs possess features of canonical miRNAs. For example, all identified precursors could fold into classic stem-loop structures (Figure 7C) and produced abundant reads (>10/million) corresponding to the mature miRNAs, which all started with either A or U at the 5' end (Mi et al., 2008). In addition, for 32 precursors reads corresponding to the miRNA* were recovered (Table S4). Finally, we selected six novel miRNAs and performed Northern blot analysis to confirm their expression. In wild type plants, Northern signals were detected for all miRNAs (Figure 7D). By contrast, no signal was detected in the *ago1* (Morel et al., 2002) or *hen1* (Boutet et al., 2003) mutant in which miRNA biogenesis is defective (Figure 7D). Together these results indicate that the identified species-specific miRNAs in *Arabidopsis* are authentic and expressed.

Species-specific young miRNAs were first reported in the model plants *Arabidopsis* and rice (Rajagopalan et al., 2006; Fahlgren et al., 2007; Zhu et al., 2008). Recently, accumulating evidence suggests that species-specific or family-specific miRNAs in plants indeed carry functions in the miRNA-mediated gene networks and underscore the dynamic nature of these networks (Lu et al., 2011; Ng et al., 2011; Wang et al., 2011). In this work, we discovered that almost two thirds of the identified miRNAs from the 15 examined plants were species-specific. Further, these miRNAs were detected from various tissue types or developmental stages with some exhibited seemingly high

expression levels. Thus, our results confirm that species-specific miRNAs represent a general feature in plants rather than an exception. While further genetic and molecular studies are necessary to comprehensively explore species-specific miRNAs, knowledge from gain-of-function analysis (Wang et al., 2011) should prove fruitful in elucidating the contribution of species-specific miRNAs to genetic novelty and evolutionary adaptation in diverse plant species.

Dynamic diversification of conserved miRNAs

Comprehensively cataloging expressed miRNA precursors allowed comparative analysis of the composition of miRNA families across plant lineages. This analysis revealed that even conserved miRNA families may be continuingly diversifying in plants (Table S5). For example, several families such as miR156/157, miR159, miR168 and miR319 include isoforms specifically present in the dicot or monocot lineages (Table S5). We selected the miR156/157 superfamily as an example to illustrate this point. The miR156 and miR157 subfamilies are among the first batch of conserved miRNAs discovered in plants (Llave et al., 2002; Reinhart et al., 2002; Arazi et al., 2005). As the sequences of mature miR156 and miR157 are highly similar, they are often considered as the same family. From the 15 species we identified 127 precursors encoding miR156. Comparison of the encoded miRNAs revealed a total of 12 isoforms of mature miR156 (Figure 8A). Encoded by 93 precursors, isoform 1 was clearly the predominant one and found in the genome of all 15 species (Figure 8B). By contrast, all other isoforms showed much more limited phylogenetic distribution (Figure 8B). For example, isoform 9, which differs from isoform 1 by one nucleotide at position 14 and represents the second most

prevalent isoform, was detected in *Arabidopsis*, soybean, and all examined monocots (Figure 8A). On the other hand, isoforms 2, 3, 4, 6, 8, and 12 were only detected from a single species (Figure 8A).

Although miR157 is highly similar to miR156, we were able to discover two features that could unambiguously distinguish the two subfamilies. First, mature miR156 is 20-nucleotide-long in all examined species judging from both sequencing reads and Northern blotting analysis (Figure 8A; Reinhart et al., 2002). By contrast, mature miR157 is 21-nucleotide-long by the same analyses with an additional nucleotide added to the 5' end (Figure 8A). Second, miR157 has a signature sequence "GAAGAUAGAGAGC" that is not found in any of the miR156 isoform (Figure 8A). Based on these two features, we identified 33 precursors specific for miR157. Comparison across the examined plant species revealed that there are five isoforms for miR157 with isoform 1 having the broadest phylogenetic distribution while the other four isoforms only detected in a single species (Figure 8A). Interestingly, while miR156 is conserved in all examined species, miR157 is only present in the dicots (Figure 8A). In fact, when searching against the genome of the four monocot plants, the miR157 sequences were not found, indicating that miR157 has specifically evolved in dicot plants.

Combined with expression profiling from the sequencing data, our analysis led to an exciting finding that different plant species exhibit distinct expression profiles of the miR156/157 family. The most conspicuous pattern is that in moss and monotots only miR156 was expressed (Figure 8B). Composition of the miR156/157 transcriptome in the dicot clade varied in different species. Interestingly, only in the legume species was miR156 prevailingly expressed. By contrast, miR157 was predominantly expressed in

most species while there are species in which both miR156 and miR157 contributed to the miRNA pool at comparably levels (Figure 8B). Further, for either miR156 or miR157, only one isoform, usually isoform 1 of miR156 and isoform 1 of miR157, was predominantly expressed (Figure 8B). Orange (*Citrus sinensis*) represents a notable exception in which isoform 2 of miR157 was the mainly expressed isoform (Figure 8B).

Members of the miR156/157 family target the conserved SPL family of transcription factors (Rhoades et al., 2002; Wu and Poethig, 2006; Wang et al., 2009; Wu et al., 2009; Yang et al., 2012). To investigate the functional implications of miR156/157 diversification, we identified the miR156/157 binding sites within the *SPL* genes. Alignment of the consensus sequences for miR156/157 and the binding sites revealed that miR157 mainly differ from miR156 at positions 11 and 14 (Figure 8C). Interestingly, a change from “G” to “U” at position 11 would significantly undermine the effect of miR157 as this position is at the cleavage site (Llave et al., 2002; Reinhart et al., 2002). Yet the occurrence of “A” at position 14 would clearly strengthen base-pairing between miR157 and the targeting site (Figure 8C). These observations suggest that miR156 and miR157 have the ability to differentially target different members of the *SPL* family (Figure 8D). An intriguing example is the tomato *CNR* gene, which is likely orthologous to *SPL3* of *Arabidopsis* and functions in fruit ripening (Manning et al., 2006). A one nucleotide deletion in the binding site made this gene incompatible with miR156 but nonetheless an excellent target for miR157 (Figure 8D).

Taken together, these results suggest that after the split of monocot and dicot species, miR157 was evolved specifically in the dicot lineages from miR156 and subsequently dominated the expression in some lineages such that different plant species

now possess signature expression profiles of the miR156/157 family. Analysis of the target genes indicates that lineage-specific diversification of the miR156/157 family could be mechanism for determining target specificity. Interestingly, we previously found that miR156-targeted *SPL4* is alternatively spliced in *Arabidopsis* such that multiple transcripts with or without the miR156 binding site are generated (Yang et al., 2012). We also found that increasing the expression level of the *SPL4* transcripts with or without the miR156 binding site by a transgenic means resulted in distinguishable phenotypes (Yang et al., 2012). Thus, fine-tuning miRNA-target circuits might provide sufficient selection pressure leading to diversification of even conserved miRNA families.

Conclusions

By parsing a large number of sequenced small RNA libraries, we systematically identified 4,142 miRNAs along with 3,861 precursor sequences in 15 land plant species (Figures 1-3). Comprehensive retrieval and comparison of expressed miRNAs with precursors in diverse plants provided new insights into miRNA evolution and function. First, this unprecedented dataset revealed that plant miRNAs have distinct phylogenetic distribution and that species-specific miRNAs account for large proportions (37% to 77%) of the miRNA transcriptome (Figures 5-7). This finding confirmed previous results that young miRNAs are emerging and disappearing at a high frequency (Fahlgren et al., 2010; Ma et al., 2010). Second, genome organization and sequencing reads distribution of neighboring miRNAs hinted new paradigms for miRNA biogenesis (Figure 4). In particular, two overlapping miRNAs processed from the same precursor (scenario III in Figure 4) suggest both a model for deriving new miRNAs from existing precursors and a potential mechanism for controlling miRNA maturation, which is substantiated by the recent characterization of *CPL1/2* function in *Arabidopsis* (Manavella et al., 2012). Third, even in highly conserved miRNA families, new members are emerging in a lineage-specific manner, which also increases the dynamic nature of miRNA-based regulation (Figure 8). Taken together, our results indicate that the whole spectrum of miRNAs in plants is subject to evolutionary changes so that new miRNA isoforms or genes are continuingly appearing. Incorporation of the new miRNAs into the regulatory networks may constitute a yet to be fully understood driving force for generating evolutionary novelty and phenotypic diversity in plants (Fahlgren et al., 2010; Ma et al., 2010).

Materials and methods

Data sources

Annotated whole genome sequences and gene models of the 15 plants used in this study were obtained from the following sources: release 10 of The *Arabidopsis* Information Resource database (http://www.arabidopsis.org/) for *Arabidopsis thaliana*, release 6.1 of Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/; Ouyang et al., 2007) for *Oryza sativa*, release 2.40 of the Tomato Genome Database (http://mips.helmholtz-muenchen.de/plant/tomato/; The tomato genome sequence provides insights into fleshy fruit evolution, 2012) for *Solanum lycopersicum*, and release 7.0 of Phytozome (http://www.phytozome.net/; Goodstein et al., 2012) for *Arabidopsis lyrata*, *Brachypodium distachyon*, *Citrus sinensis*, *Glycine max*, *Medicago truncatula*, *Mimulus guttatus*, *Panicum virgatum*, *Physcomitrella patens*, *Populus trichocarpa*, *Prunus persica*, *Vitis vinifera*, and *Zea mays*. Complete information of the 78 deeply sequenced small RNA libraries was downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; Table S1). Annotated miRNAs of the 15 plant species, if available, were downloaded from the miRBase (release 17.0; http://www.mirbase.org/; Kozomara and Griffiths-Jones, 2011). All plant tRNAs used to remove false prediction were downloaded from the Genomic tRNA Database (http://gtrnadb.ucsc.edu/; Lowe and Eddy, 1997)(Lowe and Eddy, 1997).

Plant materials

Wild type *Arabidopsis thaliana* used in this work was ecotype Col-0. The loss-of-function mutants ago1 and hen1 were described previously (Morel et al., 2002; Boutet et

al., 2003). For germination, wild type, ago1, and hen1 seeds were placed on GM media, incubated at 4°C in the dark for 4 days, and then exposed to continuous white light (170 μmole sec-1 m-2) at 22°C. Seedlings were transferred to soil and maintained under continuous light until harvested. Seeds of *Solanum lycopersicum*, *Solanum tuberosum*, *Glycine max*, *Vigna unguiculata*, *Oryza sativa*, *Triticum aestivum*, *Zea mays*, and *Sorghum bicolor* were acquired from the Germplasm Resources Information Network (http://www.ars-grin.gov/npgs/). The seeds were placed on water saturated filter paper and incubated at 29°C in the dark in temperature-controlled incubators until germination. After transferred to soil, seedlings were allowed to grow in the greenhouse for two weeks before harvesting. All plant materials were frozen in liquid nitrogen and stored at -80°C until used for RNA extraction.

Process of retrieving annotated miRNAs and detecting new miRNAs

A computational pipeline centered on miRDeep-P (Yang and Li, 2011) was employed to retrieve known miRNAs and detect new miRNAs from deep sequencing data (Figure S1). For each sequenced small RNA library, reads were first filtered by length and only those longer than 15 nt and shorter than 28 nt were kept. Identical reads were then collapsed into Fasta format with an ID including copy numbers (Yang and Li, 2011). The formatted reads were used as input for processing by miRDeep-P with such specific parameters: 1) 250 nt was used as the length for extracting reference sequences to predict RNA secondary structure and measure reads distribution (Yang and Li, 2011), 2) species-specific filter parameter, which refers to the default value set for the parameter -c in the Perl script *filter_alignments.pl* of miRDeep-P, was employed and listed in Table

S1, and 3) 62 nt was used as the minimum length to filter identified pre-miRNAs. As tRNAs could potentially fold into stem-loop structure, false predictions were removed by a BLAST search between putative pre-miRNAs and all plant tRNAs from GtRNAdb (Lowe and Eddy, 1997) with the cutoff of length ⩾80% and identity ⩾90%. As for each species except Glycine max, multiple small RNA libraries were employed, all retrieved miRNAs and pre-miRNAs were merged within a species with redundant items removed.

After compiling a miRNA dataset for each species, a similarity search was conducted using both mature miRNA (criterion1: two mismatches allowed) and precursor (criterion2: cutoff of length ⩾90% and identity ⩾95%) sequences against all annotated plant miRNAs in miRBase (release 17). Homologous sequences were identified in the search. After this step, two datasets were generated for each plant species. The first dataset included retrieved known miRNAs (both criteria met). The second dataset included miRNAs (only criterion 1 met) conserved with known miRNAs but have not been annotated (Figure S1). These non-conserved items were further filtered by a series of constraints to ensure high confidence. First, they were filtered with specific criteria regarding the miRNA:miRNA* duplex established in plants (Meyers et al., 2008; Yang and Li, 2011). Second, all new items overlapping with exons of protein-coding genes on the same DNA strand were filtered out. Third, the number of reads corresponding to the new miRNAs was set at 10 reads per million (RPM) as a filtering criterion for expression (Breakfield et al., 2012).

Quantification of miRNA expression level

From each sequenced small RNA library, number of reads corresponding to a given mature miRNA (offset by 1 nt either end along the precursor was allowed) was calculated. Then, the miRNA copy number was normalized against the library size with the following formula:

$$\frac{miRNA\ copy\ No}{TR} \cdot 10^6$$

in which *TR* is the total number of reads from a libraries that are shorter than 28 nt and longer than 15 bp (Table S1). For results shown in Figure 7, the mean of normalized expression values (in RPM) from all libraries was calculated as the expression level of a given miRNA. Logarithm-transformed data (base 10) were then used.

Genome organization of neighboring miRNAs

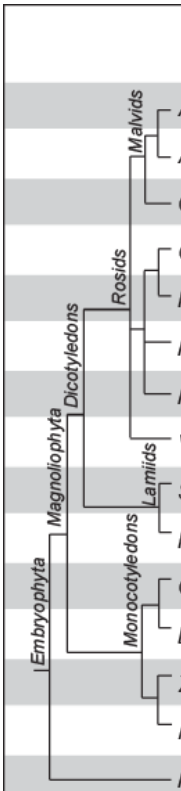
As an output of miRDeep-P, coordinates miRNAs and pre-miRNAs on chromosome or scaffold of the 15 plant species were recorded. Copy number of reads corresponding to mature miRNAs was normalized to RPM and the mean of RPM from all small RNA libraries from a given species calculated. As a stringent cutoff for analyzing neighboring miRNAs, miRNAs with average level less than 10 RPM were all discarded. Remaining miRNAs with non-overlapped pre-miRNAs were searched and pre-miRNAs located within 3 Kb on the same chromosomes or scaffolds were identified as clustered miRNAs (model I) following the procedure and cutoff previously described (Merchan et al., 2009). Neighboring miRNAs with shared or overlapping precursors were then further screened. For models II and III, to reduce false positives, the two neighboring duplexes of miRNA and miRNA* must have a combined level greater than 100 RPM. Further, paired miRNAs in model III must be offset by no less than 4 nt. To identify miRNAs in model

IV, two overlapping pre-miRNAs located on opposite DNA strands were screened. Only those cases in which a mature miRNA overlapped for at least 5 nt with the miRNA* on the opposite strand were selected.

Northern blot analysis

Total RNA was extracted using the TRIzol agents (Invitrogen) and treated with RNase-free DNaseI (Invitrogen) as recommended by the manufacturer. Twenty μg total RNA was used for miRNA blotting as previously described (Yang et al., 2011). Briefly, gel-separated low-molecular-weight RNA was electrically transferred to Hybond N+ nylon membrane (GE Healthcare Life Sciences). After pre-hybridization for 2 hours, blots were hybridized overnight with respective miRNA-complementary oligonucleotides end-labeled with digoxigenin-11-ddUTP (Roche) at 42°C, washed with the DIG Wash and Block Buffer Set, and then incubated with Anti-Digoxigenin-AP antibody (Roche) as recommended by the manufacturer. Blots were equilibrated in detection buffer for 5 minutes and incubated in the CDP-Star solution (Roche) for 15 minutes prior to exposure to X-ray films.

Figure 1. Comprehensive identification of expressed miRNAs in 15 land plants. Shown on far left is the species tree illustrating the phylogenetic relationship of the 15 examined species. For each species, a three-letter abbreviation is given in parentheses and the number of sequenced libraries indicated. Detected miRNAs are divided into annotated miRNAs (according to miRBase release 17), conserved but unannotated miRNAs, and novel miRNAs detected in this study only. The sums of the three groups of miRNAs in the 15 species are given at the bottom.



Species	Library	Annotated	Conserved/ unannoated	Novel	Total
Arabidopsis thaliana (Ath)	11	129	5	40	174
Arabidopsis lyrata (Aly)	4	104	6	51	161
Citrus sinensis (Csi)	5	26	46	141	213
Glycine max (Gma)	1	57	124	103	284
Medicago truncatula (Mtr)	3	75	73	137	285
Prunus persica (Ppe)	2	0	75	84	159
Populus trichocarpa (Ptc)	3	117	15	86	218
Vitis vinifera (Vvi)	4	96	26	53	175
Solanum lycopersicum (Sly)	12	24	76	271	371
Mimulus guttatus (Mgu)	3	0	61	173	234
Oryza sativa (Osa)	8	172	19	268	459
Brachypodium distachyon (Bdi)	2	64	37	101	202
Zea mays (Zma)	7	98	20	252	370
Panicum virgatum (Pvi)	3	0	145	495	640
Physcomitrella patens (Ppt)	9	154	12	31	197
Total	**78**	**1,116**	**740**	**2,286**	**4,142**

Figure 2. Testing the retrieval rate of annotated miRNAs by miRDeep-P in diverse plants.

To investigate what window sizes are optimal for miRDeep-P to extract the sequence flanking each anchored read for predicting RNA secondary structure and quantifying the compatibility of reads distribution with Dicer-mediated processing in different plants, 10 species with well-annotated miRNAs were chosen for testing. One sequenced small RNA library for each species was processed by miRDeep-P to retrieve annotated miRNAs when different window sizes were used. The number of detected miRNAs in each species was then plotted against the window sizes. For *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Glycine max*, *Medicago truncatula*, *Populus trichocarpa*, *Vitis vinifera*, *Oryza sativa*, *Brachypodium distachyon*, *Zea mays*, and *Physcomitrella patens*, the libraries GSM442935, GSM518430, GSM621344, GSM643815, GSM458929, GSM717876, GSM489087, GSM506621, GSM306488, and GSM313212 were used, respectively.

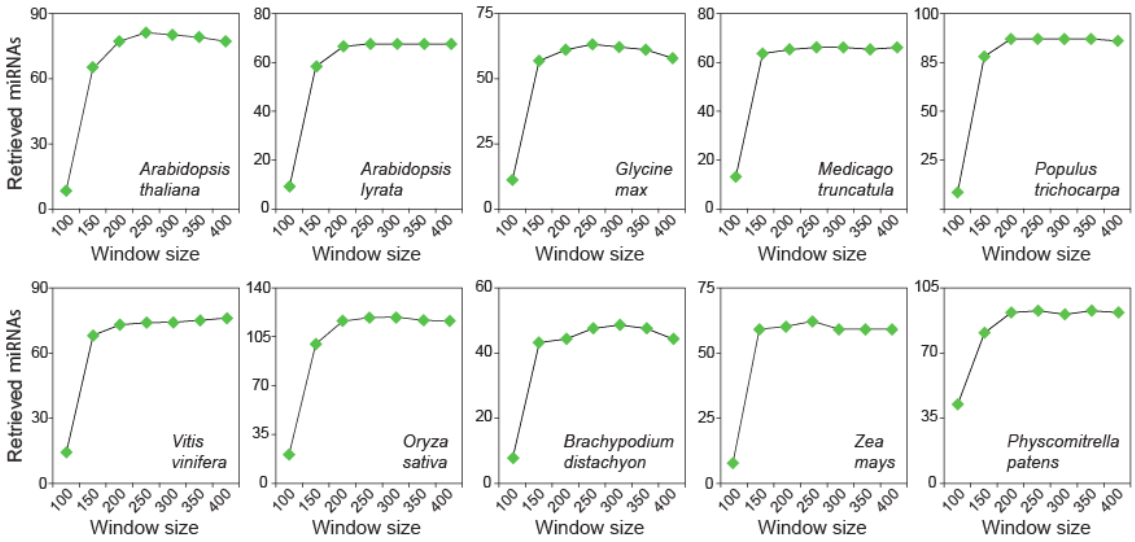

Retrieved miRNAs
Window size
Arabidopsis thaliana
Arabidopsis lyrata
Glycine max
Medicago truncatula
Populus trichocarpa
Vitis vinifera
Oryza sativa
Brachypodium distachyon
Zea mays
Physcomitrella patens

Figure 3. Retrieval of precursor sequences facilitates cross-species comparison.

A, Average pre-miRNA length of all identified miRNAs in the 21 conserved families was calculated and compared between the 10 dicot and 4 monocot plants. Error bars represent standard errors. B, Comparison of the average pre-miRNA length of the 21 conserved families and all other miRNAs. * $p < 0.0001$ by t test. C, Analysis of miR408 precursors in 15 examined plant species. The left panel shows the phylogenetic relationship of the 20 miR408 precursors and their predicted secondary structures, which were generated using RNAfold (Schuster et al., 1994). Precursors labeled in bold were specifically uncovered in this study. Red lines along the secondary structures indicate the location of mature miRNAs. In the right panel, reads distribution in 220-bp windows encompassing the miR408 precursors in each species is recapitulated from deep sequencing data. The vertical axis of the plots shows relative frequency of reads at a given position. Red vertical lines represent reads corresponding to mature miR408 whereas blue lines indicate reads mapped to other regions. For each species, the number of total reads (TR) perfectly mapped to the precursors from all analyzed small RNA libraries is indicated.

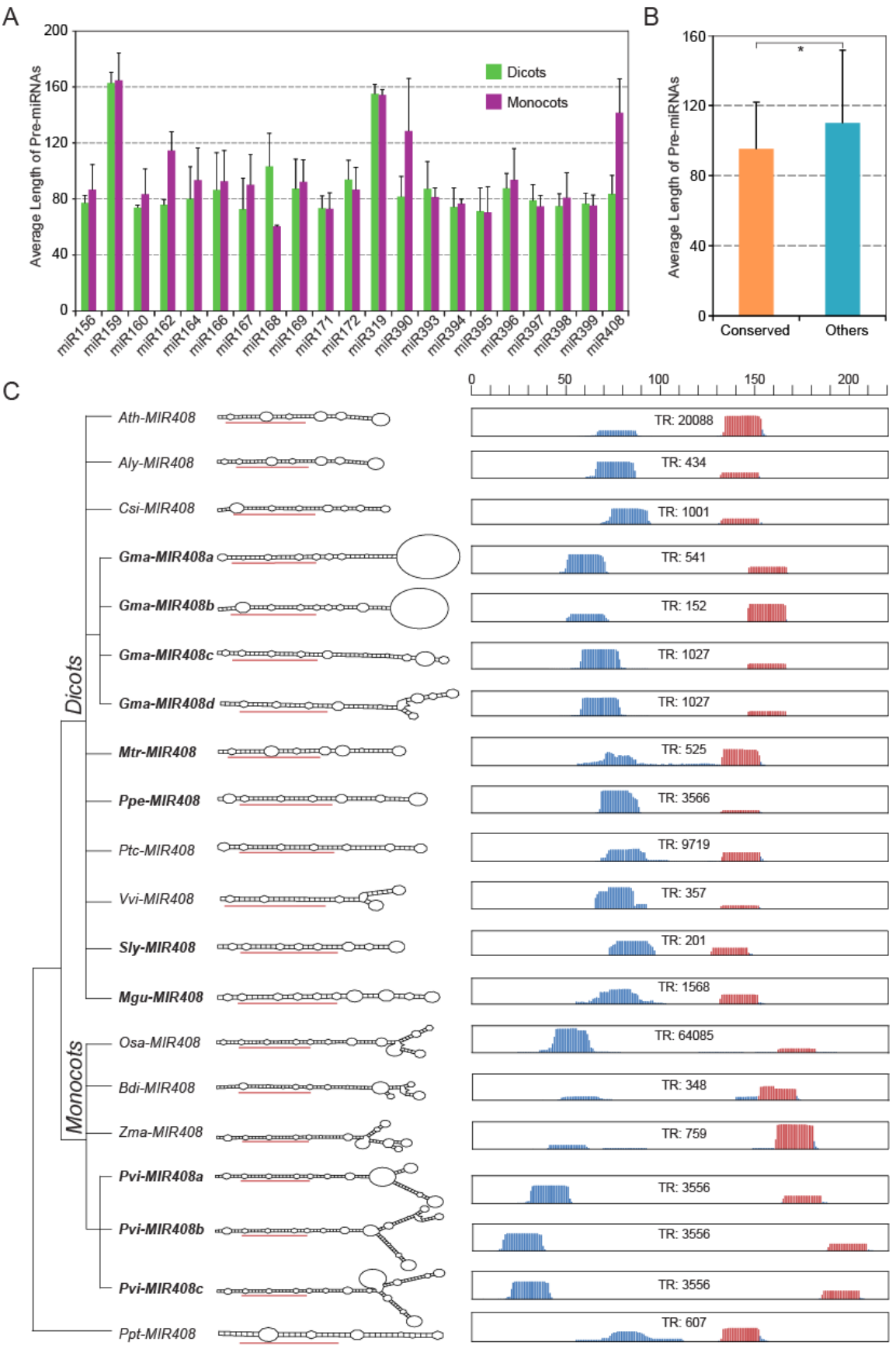

A
Average Length of Pre-miRNAs
200
160
120
80
40
0
Dicots
Monocots
miR156
miR159
miR160
miR162
miR164
miR166
miR167
miR168
miR169
miR171
miR172
miR319
miR390
miR393
miR394
miR395
miR396
miR397
miR398
miR399
miR408
B
Average Length of Pre-miRNAs
160
120
80
40
0
*
Conserved
Others
C
0
50
100
150
200
Ath-MIR408
Aly-MIR408
Csi-MIR408
Gma-MIR408a
Gma-MIR408b
Gma-MIR408c
Gma-MIR408d
Mtr-MIR408
Ppe-MIR408
Ptc-MIR408
Vvi-MIR408
Sly-MIR408
Mgu-MIR408
Osa-MIR408
Bdi-MIR408
Zma-MIR408
Pvi-MIR408a
Pvi-MIR408b
Pvi-MIR408c
Ppt-MIR408
Dicots
Monocots
TR: 20088
TR: 434
TR: 1001
TR: 541
TR: 152
TR: 1027
TR: 1027
TR: 525
TR: 3566
TR: 9719
TR: 357
TR: 201
TR: 1568
TR: 64085
TR: 348
TR: 759
TR: 3556
TR: 3556
TR: 3556
TR: 607

Figure 4. Compact genome organization of neighboring miRNAs suggests new biogenesis models.

A, Four scenarios for generating multiple miRNAs from single or co-transcribed precursors. Model I, clustered miRNA precursors (< 3kb apart) (Merchan et al., 2009) are co-transcribed and separately processed; model II, two different miRNA:miRNA* duplexes are separately spliced from the same transcribed precursor; model III, two miRNA:miRNA* duplexes compete for processing from the same precursor such that two overlapping but different miRNAs are produced in a mutually exclusive way; and model IV, two miRNA precursors are separately transcribed from opposite strands at the same locus and processed. Filled boxes in red and blue indicate two mature miRNAs while unfilled boxes represent miRNA*. B, Ath-miR395 as an example for model I in which a polycistronic precursor leads to generation of both Ath-miR395b and Ath-miR395c in *Arabidopsis*. The left panel shows the secondary structure of the precursor with thick lines in red and black indicate mature miRNAs and miRNA*. The right panel recapitulates reads distribution along the precursor. Mature, reads corresponding to the mature miRNA; star, reads corresponding to miRNA*. Consecutive dots (left panel) and dash lines (right panel) indicate ellipsis of nucleotides. C, Example of model II in which Osa-miR820a and a newly identified miRNA, Osa-miRN27a, are produced from the same precursor in rice. D, Example of model III in which the precursor of Osa-miR1868 is also used to generate a second miRNA, Osa-miRN179, in a mutually exclusive way. E, Ath-miR781 and Ath-miR781N as an example for model IV in which the two identical miRNAs are generated from precursors that are transcribed and processed from the opposite strands at the same locus.

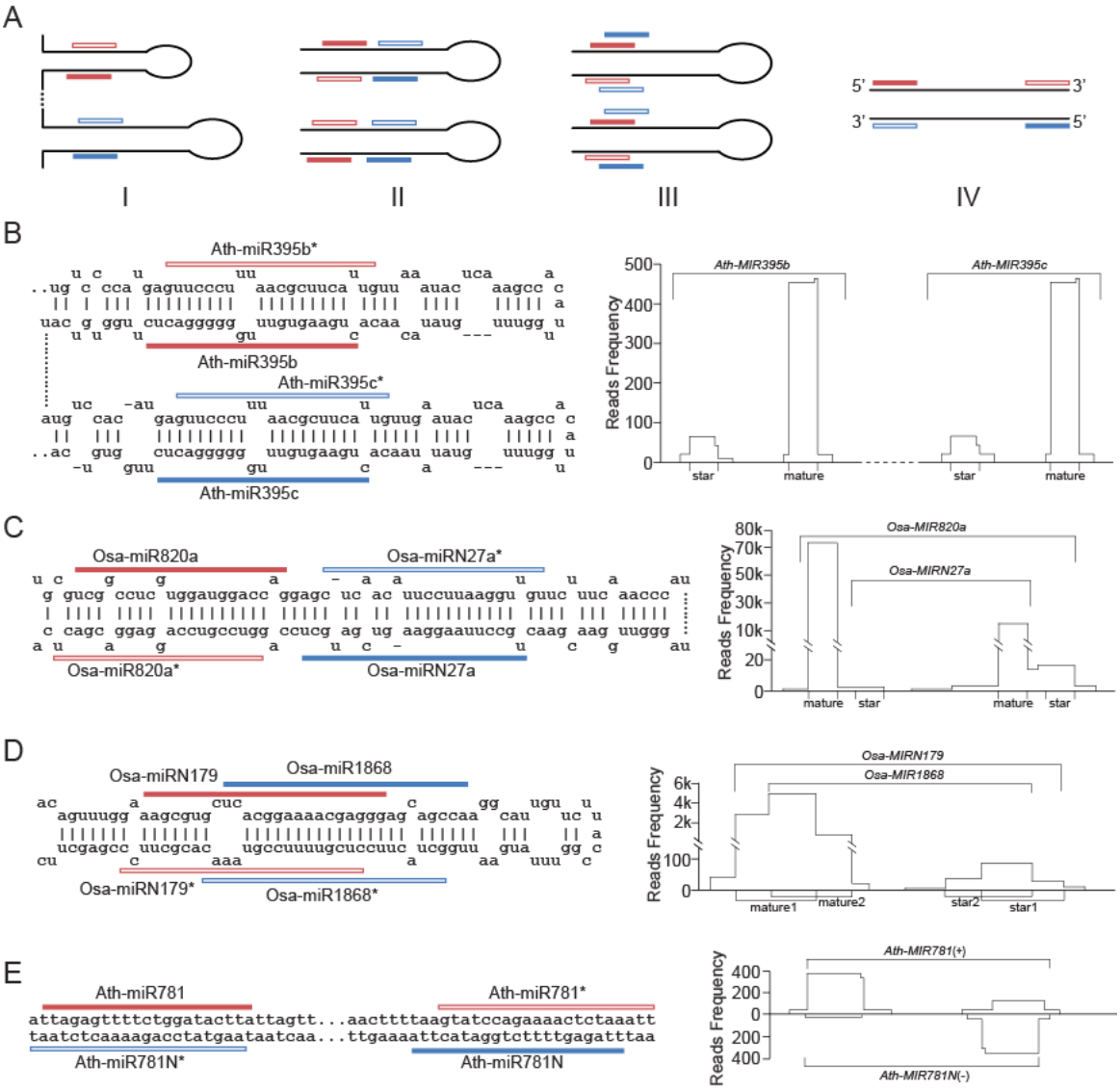

A
I
II
III
IV
5'
3'
3'
5'
B
Ath-miR395b*
Ath-miR395b
Ath-miR395c*
Ath-miR395c
Reads Frequency
Ath-MIR395b
Ath-MIR395c
star
mature
star
mature
C
Osa-miR820a
Osa-miRN27a*
Osa-miR820a*
Osa-miRN27a
Reads Frequency
Osa-MIR820a
Osa-MIRN27a
mature
star
mature
star
D
Osa-miRN179
Osa-miR1868
Osa-miRN179*
Osa-miR1868*
Reads Frequency
Osa-MIRN179
Osa-MIR1868
mature1
mature2
star2
star1
E
Ath-miR781
Ath-miR781*
Ath-miR781N*
Ath-miR781N
Reads Frequency
Ath-MIR781(+)
Ath-MIR781N(-)

Figure 5. Phylogenetic distribution of expressed conserved miRNAs.

In the 15 examined plant species (rows), miRNA families (columns) were analyzed for phylogenetic distribution. A box is highlighted if at least one member of a given miRNA family was detected from deep sequencing data in a particular species. Groups of miRNA families are shaded by dark to light green based on taxonomic distribution.

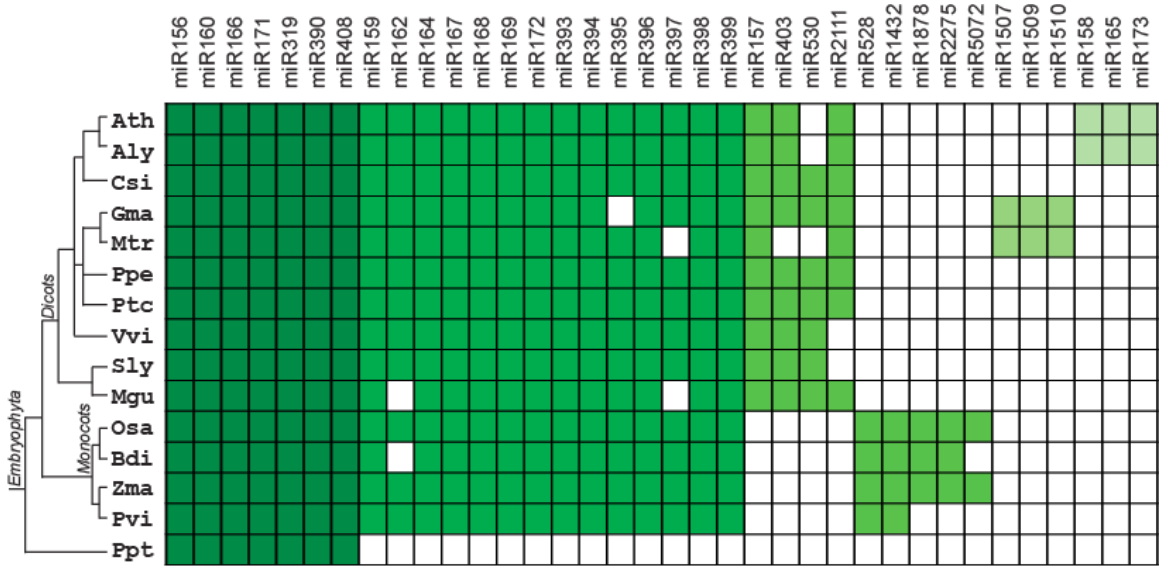

miR156
miR160
miR166
miR171
miR319
miR390
miR408
miR159
miR162
miR164
miR167
miR168
miR169
miR172
miR393
miR394
miR395
miR396
miR397
miR398
miR399
miR157
miR403
miR530
miR2111
miR528
miR1432
miR1878
miR2275
miR5072
miR1507
miR1509
miR1510
miR158
miR165
miR173
Ath
Aly
Csi
Gma
Mtr
Ppe
Ptc
Vvi
Sly
Mgu
Osa
Bdi
Zma
Pvi
Ppt
Dicots
Monocots
Embryophyta

Figure 6. Northern blotting validation of lineage-specific miRNAs.

RNA blotting analysis was performed using eight representative miRNAs with different degree of conservation in plants. Analyzed plant species include *Arabidopsis* (*Arabidopsis thaliana*), tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), soybean (*Glycine max*), cowpea (*Vigna unguiculata*), rice (*Oryza sativa*), wheat (*Triticum aestivum*), maize (*Zea mays*), and sorghum (*Sorghum bicolor*). U6 snRNA was used as the loading control.

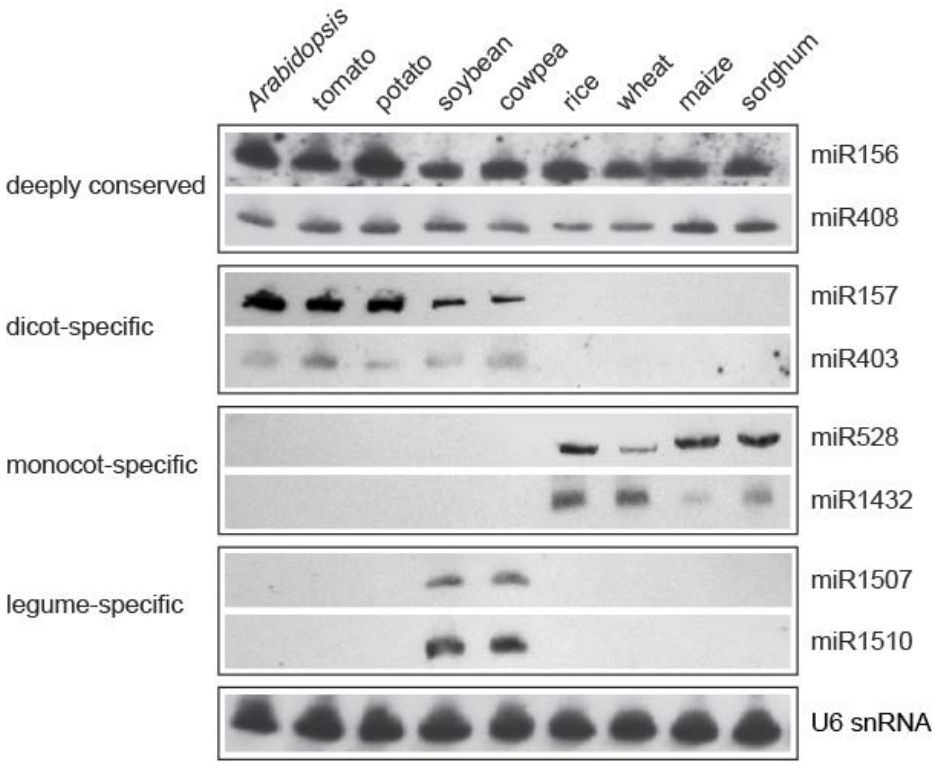

Arabidopsis
tomato
potato
soybean
cowpea
rice
wheat
maize
sorghum
deeply conserved
miR156
miR408
dicot-specific
miR157
miR403
monocot-specific
miR528
miR1432
legume-specific
miR1507
miR1510
U6 snRNA

Figure 7. Analysis of species-specific miRNAs.

A, Species-specific miRNAs in the 15 plant species. Based on sequence comparison, 63% of the 4,142 identified miRNAs can only be found in a single species and considered species-specific, which is shown on the left as a large pie graph. For each species, the portion of conserved (found in at least two species) and species-specific miRNAs is displayed as small pie charts on the right. B, Relative expression levels of species-specific miRNAs. Large box plot on the left displays expression distribution of all conserved miRNAs in 15 plant species. The stars indicate expression levels of the top 5% most-abundant species-specific miRNAs. The same analysis is extended to each specific species and shown on the right. C, Predicted secondary structure for the precursors of six species-specific miRNAs identified in *Arabidopsis thaliana*. Horizontal lines along the secondary structures indicate location of the mature miRNAs. D, Northern blot analysis of species-specific miRNAs shown in (C) in wild type as well as *ago1-27* and *hen1-1* plants. MiR408 was used as a positive control for *HEN1* and *AGO1* dependent accumulation. U6 snRNA was used as the loading control.

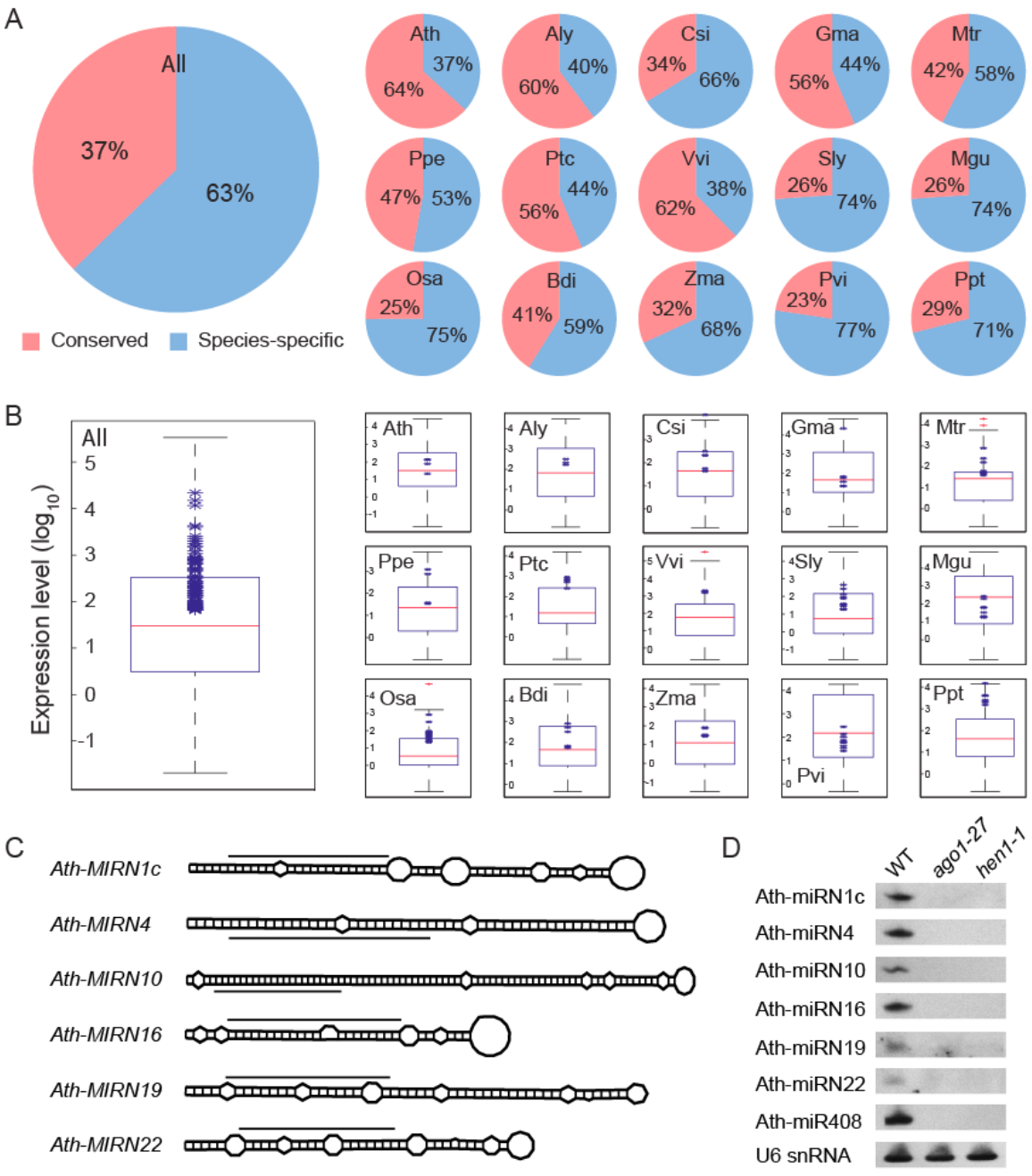
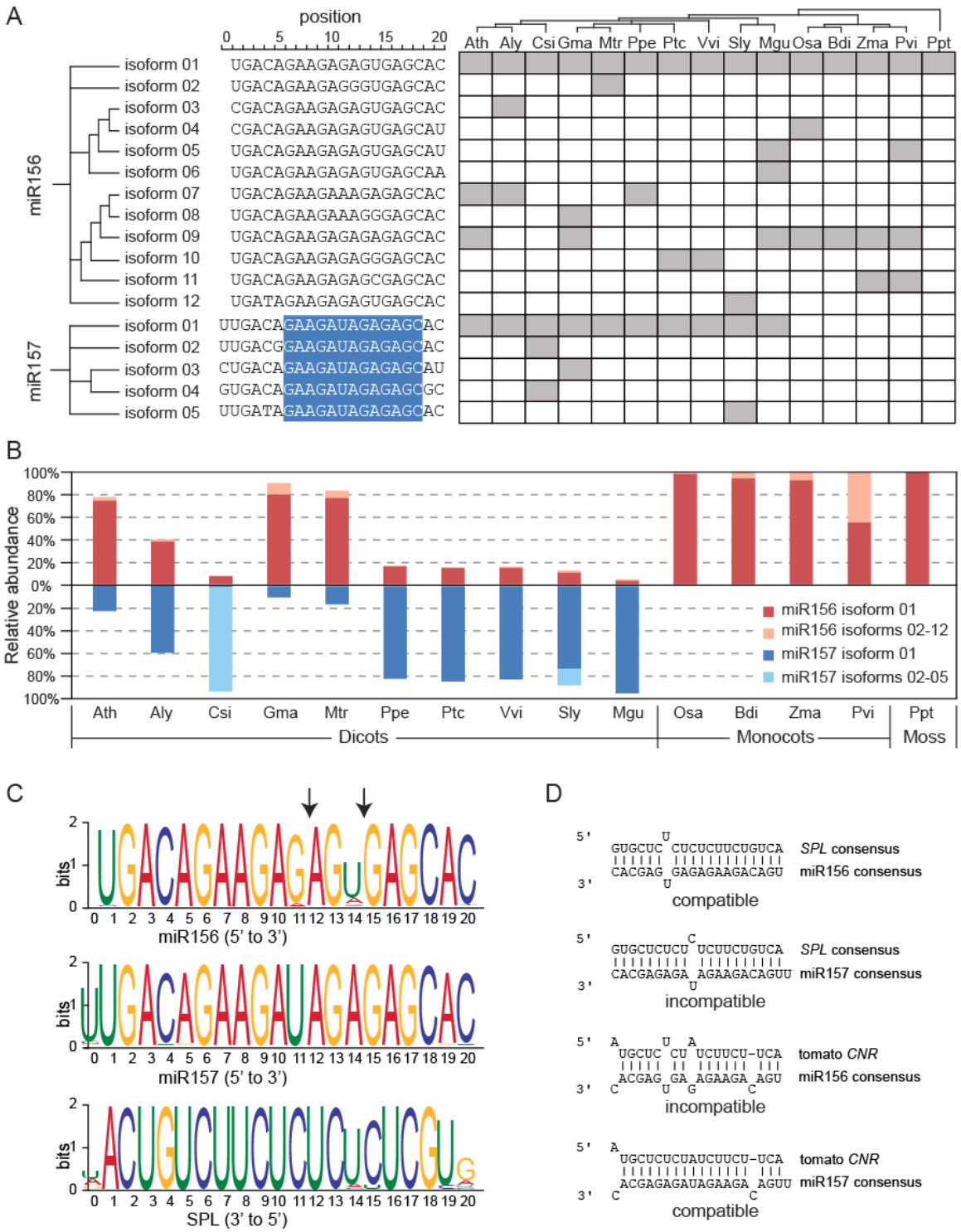

A
All
37%
63%
Conserved
Species-specific
Ath
37%
64%
Aly
40%
60%
Csi
34%
66%
Gma
44%
56%
Mtr
42%
58%
Ppe
47%
53%
Ptc
44%
56%
Vvi
38%
62%
Sly
26%
74%
Mgu
26%
74%
Osa
25%
75%
Bdi
41%
59%
Zma
32%
68%
Pvi
23%
77%
Ppt
29%
71%
B
All
Expression level (log10)
Ath
Aly
Csi
Gma
Mtr
Ppe
Ptc
Vvi
Sly
Mgu
Osa
Bdi
Zma
Pvi
Ppt
C
Ath-MIRN1c
Ath-MIRN4
Ath-MIRN10
Ath-MIRN16
Ath-MIRN19
Ath-MIRN22
D
WT
ago1-27
hen1-1
Ath-miRN1c
Ath-miRN4
Ath-miRN10
Ath-miRN16
Ath-miRN19
Ath-miRN22
Ath-miR408
U6 snRNA

Figure 8. Dynamic diversification of miR156 and miR157.

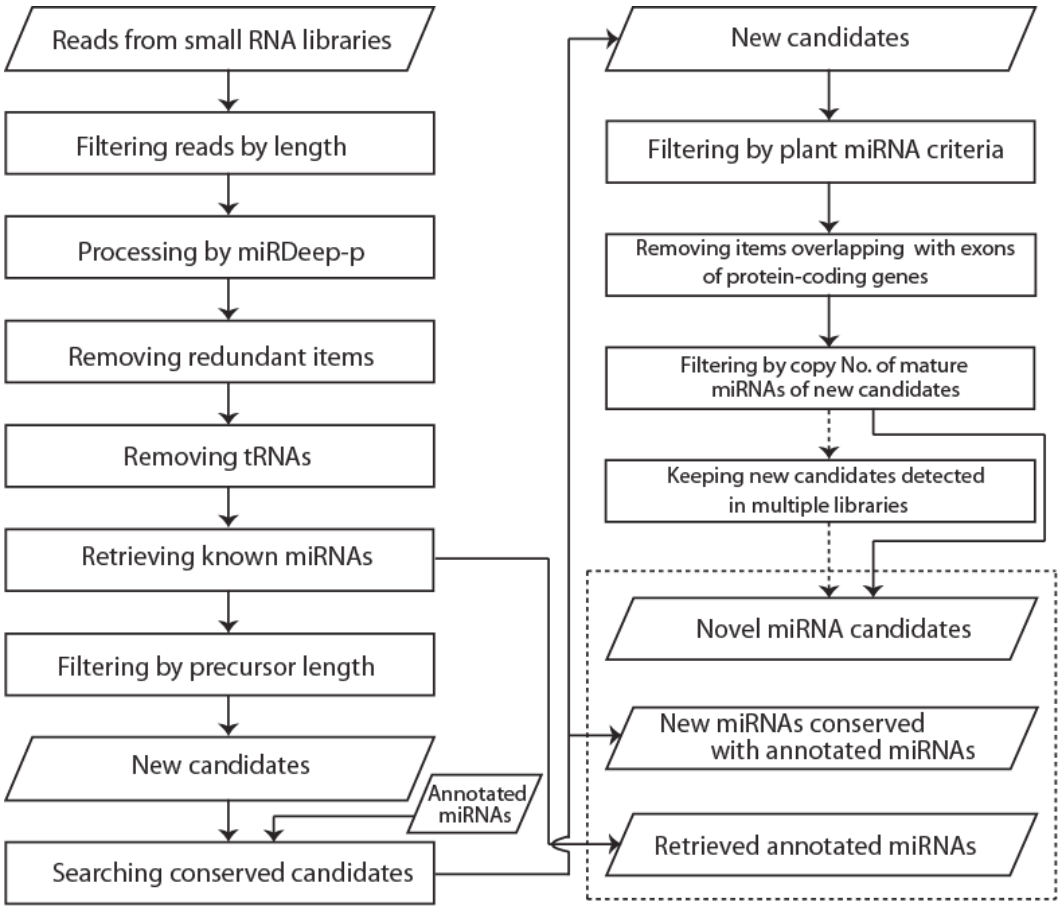
A, Identification and distribution of miR156 and miR157 isoforms in 15 land plants. The left panel shows 12 miR156 and five miR157 isoforms identified from 127 and 33 precursors, respectively. The isoforms are aligned in the 5' to 3' direction and the position of each nucleotide is indicated. Conserved region unique to miR157 is highlighted in blue. The right panel illustrates distribution of miR156 and miR157 isoforms in the 15 examined plant species. A box is shaded in gray if a particular isoform is present in a given species. B, Relative contribution of miR156 and miR157 isoforms to the final transcript pool in 15 plants. For each species, the total number of reads corresponding to miR156 and miR157 isoforms was calculated and normalized. Percentage of reads corresponding to miR156 isoform 1 (red), miR156 isoforms 2-12 (light red), miR157 isoform 1 (blue), and miR157 isoforms 2-5 (light blue) is displayed. C, Sequence logos for miR156, miR157, and the miRNA binding site of SPL genes. A total of 127 miR156 and 33 miR157 from 15 plants were employed to generate the respective logos using the MEME Suite (Bailey et al., 2009). The X axis indicates the nucleotide positions in the 5' to 3' direction. The Y axis represents information content in bits and the height of each letter is proportional to its prevalence at a given position. For the miRNA binding site, 34 SPL genes from *Arabidopsis*, rice, and tomato were used to generate the logo in the 3' to 5' direction. Arrows indicate the two positions most variable between miR156 and miR157. D, Predicted RNA duplex between the consensus sequences of miR156, miR157, the miRNA binding site of the SPL family, and the tomato CNR gene. Base paring is indicated by a short vertical line. Compatible, the

miRNA is predicted to target the mRNA according to the criteria set forth by Schwab et al. (2005). Incompatible, the miRNA is predicted not to target the mRNA.


A
position
0 5 10 15 20
Ath Aly Csi Gma Mtr Ppe Ptc Vvi Sly Mgu Osa Bdi Zma Pvi Ppt
miR156
isoform 01 UGACAGAAGAGAGUGAGCAC
isoform 02 UGACAGAAGAGGGUGAGCAC
isoform 03 CGACAGAAGAGAGUGAGCAC
isoform 04 CGACAGAAGAGAGUGAGCAU
isoform 05 UGACAGAAGAGAGUGAGCAU
isoform 06 UGACAGAAGAGAGUGAGCAA
isoform 07 UGACAGAAGAAAGAGAGCAC
isoform 08 UGACAGAAGAAAGGGAGCAC
isoform 09 UGACAGAAGAGAGAGAGCAC
isoform 10 UGACAGAAGAGAGGGAGCAC
isoform 11 UGACAGAAGAGAGCGAGCAC
isoform 12 UGATAGAAGAGAGUGAGCAC
miR157
isoform 01 UUGACAGAAGAUAGAGAGCAC
isoform 02 UUGACGGAAGAUAGAGAGCAC
isoform 03 CUGACAGAAGAUAGAGAGCAU
isoform 04 GUGACAGAAGAUAGAGAGCGC
isoform 05 UUGATAGAAGAUAGAGAGCAC
B
Relative abundance
100% 80% 60% 40% 20% 0% 20% 40% 60% 80% 100%
miR156 isoform 01
miR156 isoforms 02-12
miR157 isoform 01
miR157 isoforms 02-05
Ath Aly Csi Gma Mtr Ppe Ptc Vvi Sly Mgu Osa Bdi Zma Pvi Ppt
Dicots
Monocots
Moss
C
bits
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
miR156 (5' to 3')
miR157 (5' to 3')
SPL (3' to 5')
D
5' U
GUGCUC CUCUCUUCUGUCA SPL consensus
CACGAG GAGAGAAGACAGU miR156 consensus
3' U
compatible
5' C
GUGCUCUCU UCUUCUGUCA SPL consensus
CACGAGAGA AGAAGACAGUU miR157 consensus
3' U
incompatible
5' A U A
UGCUC CU UCUUCU-UCA tomato CNR
ACGAG GA AGAAGA AGU miR156 consensus
3' C U G C
incompatible
5' A
UGCUCUCUAUCUUCU-UCA tomato CNR
ACGAGAGAUAGAAGA AGUU miR157 consensus
3' C C
compatible

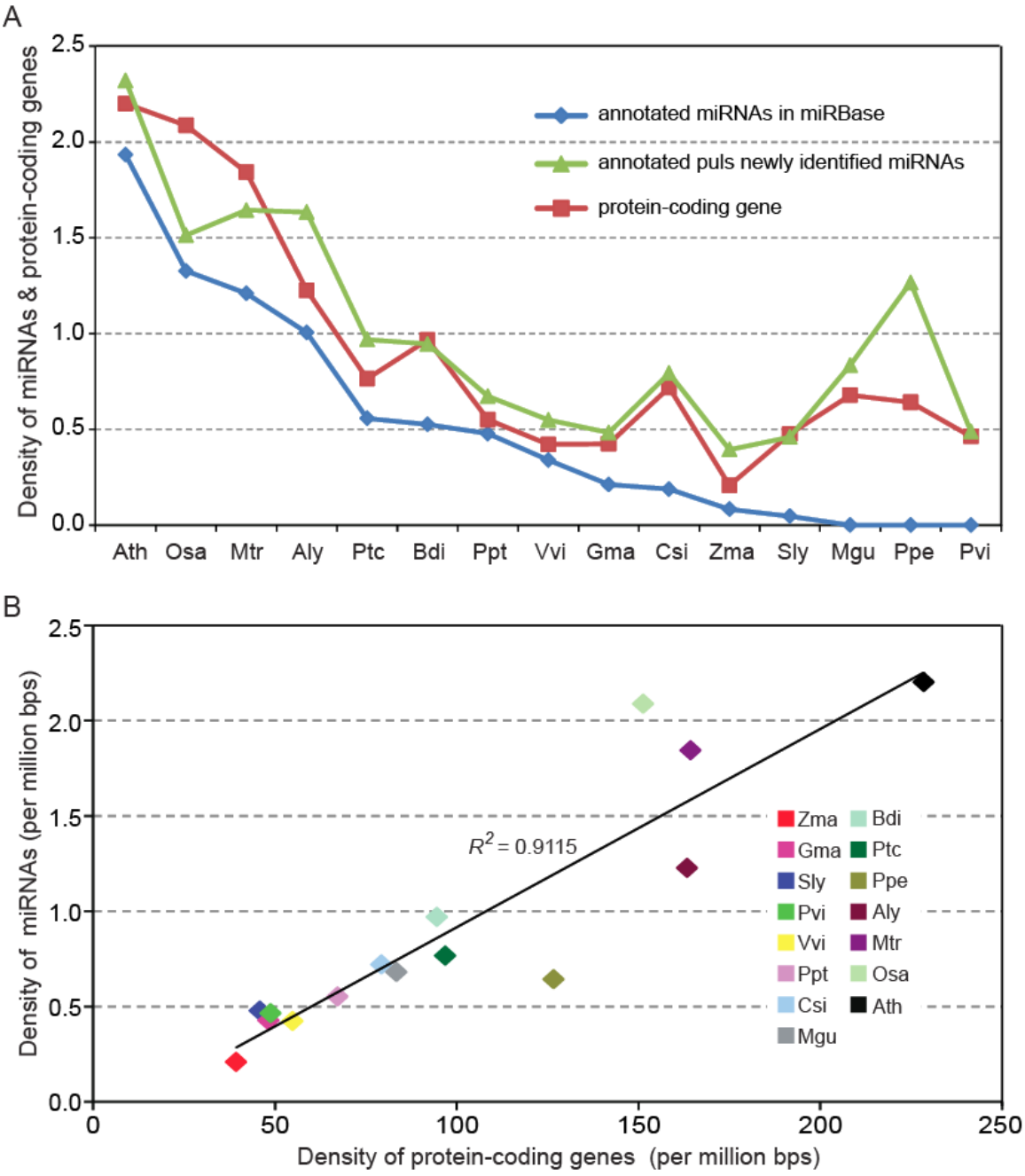
Supplemental Figure 1. Workflow of detecting miRNAs from deeply sequenced small RNA libraries.

Schematic of the workflow for detecting expressed miRNAs and their precursors from deeply sequenced small RNA libraries in plants based on the program miRDeep-P. Points shown in parallelograms indicate input or output for data-processing steps shown in rectangles. Steps linked by dotted arrows are optional for complex datasets, which were only employed in the current study for analyzing data from *Oryza sativa*, *Solanum lycopersicum*, and *Zea mays*. The three parallelograms included in the dotted box display the final output including retrieved annotated miRNAs, new but conserved miRNAs, and novel miRNAs.


Reads from small RNA libraries
Filtering reads by length
Processing by miRDeep-p
Removing redundant items
Removing tRNAs
Retrieving known miRNAs
Filtering by precursor length
New candidates
Annotated miRNAs
Searching conserved candidates
New candidates
Filtering by plant miRNA criteria
Removing items overlapping with exons of protein-coding genes
Filtering by copy No. of mature miRNAs of new candidates
Keeping new candidates detected in multiple libraries
Novel miRNA candidates
New miRNAs conserved with annotated miRNAs
Retrieved annotated miRNAs

Supplemental Figure 2. Comparison of the density of miRNAs and protein-coding genes in 15 land plants.

A, Density of miRNAs and protein-coding genes in the 15 examined land plants. Density of miRNAs was normalized to per million base pairs of the genome. For protein-coding genes, density was normalized to 0.01 million base pairs. For miRNAs, both the annotation of miRBase release 17 and the dataset combining annotated miRNAs and miRNAs identified in this work were used for comparison. B, Linear correlation of density of protein-coding genes and miRNAs in plants. The density of miRNAs in the combined dataset was plotted against the density of protein-coding genes in the 15 plant species. Pearson Correlation was used to calculate the coefficient of determination, which changed from 0.91 to 0.77 when only annotated miRNAs were used.


A
Density of miRNAs & protein-coding genes
2.5
2.0
1.5
1.0
0.5
0.0
annotated miRNAs in miRBase
annotated puls newly identified miRNAs
protein-coding gene
Ath Osa Mtr Aly Ptc Bdi Ppt Vvi Gma Csi Zma Sly Mgu Ppe Pvi
B
Density of miRNAs (per million bps)
2.5
2.0
1.5
1.0
0.5
0.0
R^2 = 0.9115
Zma Bdi
Gma Ptc
Sly Ppe
Pvi Aly
Vvi Mtr
Ppt Osa
Csi Ath
Mgu
0 50 100 150 200 250
Density of protein-coding genes (per million bps)

Supplemental Table 1. Summary of small RNA libraries used in this study.

Species	Library[1]	TR[2]	UR[3]	MR[4]	Tissue	FP[5]
Arabidopsis lyrata	GSM451894	583,895	382,520	223,347		25
	GSM518429	4,034,948	888,688	888,686		25
	GSM518430	4,836,368	1,628,622	1,628,613		25
	GSM518431	4,201,180	1,363,401	1,363,395		25
Arabidopsis thaliana	GSM442934	3,780,910	1,250,440	1,235,891		15
	GSM442935	3,682,981	1,257,646	1,235,891		15
	GSM707678	4,964,909	2,655,273	1,945,934		15
	GSM707679	4,998,773	2,318,668	1,609,949		15
	GSM707680	3,474,472	1,482,148	989,583		15
	GSM707681	5,406,787	1,494,499	917,958		15
	GSM456944	5,860,234	1,789,157	1,121,679		15
	GSM456945	6,719,875	2,161,753	1,378,281		15
	GSM342999	2,637,986	802,493	802,471		15
	GSM442932	3,616,457	607,213	601,276		15
	GSM442933	3,466,596	696,968	690,294		15
Brachypodium distachyon	GSM506620	3,992,373	2,135,020	1,191,806		25
	GSM506621	5,087,539	2,293,122	1,567,195		25
Citrus sinensis	GSM455228	4,598,696	2,787,921	1,200,883		25
	GSM455230	6,775,323	3,416,419	1,570,173		25
	GSM712528	3,198,491	1,634,083	827,514		25
	GSM712529	2,770,871	1,515,486	618,390		25
	GSM712530	2,409,984	1,053,181	562,685		25
Glycine max	GSM621344	5,908,211	2,145,586	1,496,438		50
Mimulus guttatus	GSM717696	3,663,627	557,124	232,087		25
	GSM717697	3,937,426	2,093,391	1,438,085		25
	GSM717698	4,586,237	2,202,495	1,109,414		25
Medicago truncatula	GSM346592	872,048	426,608	205,155		25
	GSM346593	3,076,823	1,238,984	576,219		25
	GSM643815	8,089,116	3,851,320	2,065,443		25
	GSM643816	8,149,617	3,756,582	2,131,317		25
Oryza sativa	GSM693279	10,034,728	4,581,492	3,872,986		25
	GSM693280	9,283,494	3,028,914	2,539,047		25
	GSM686039	12,806,715	3,499,831	2,871,523		25
	GSM686040	13,270,975	3,641,248	2,894,520		25
	GSM489087	4,278,651	1,827,520	1,236,196		25
	GSM455965	4,029,462	1,991,942	1,991,925		25
	GSM278571	5,741,313	2,668,254	1,755,301		25
	GSM278572	5,313,959	2,668,254	1,755,301		25
Panicum virgatum	GSM717866	3,183,269	926,069	542,359		50
	GSM717867	3,149,080	1,950,028	1,264,764		50

	GSM717868	2,601,939	716,495	395,433		50
Physcomitrella patens	GSM313212	925,746	337,199	122,572		25
	GSM313213	972,386	348,472	128,740		25
	GSM313214	1,322,773	300,562	66,175		25
	GSM313215	1,088,114	274,910	55,771		25
	GSM313216	982,730	332,000	107,996		25
	GSM313217	932,138	313,649	103,908		25
	GSM115095	140,013	214,957	82,906		25
	GSM115096	200,526	214,957	82,906		25
	GSM115097	220,520	214,957	82,906		25
Prunus persica	GSM944971				root	25
	GSM944972				leaf	25
	GSM944973				flower	25
	GSM944974				fruit	25
Populus trichocarpa	GSM465746	7,862,905	2,565,896	1,093,169	leaf	25
	GSM465747	12,476,493	10,014,495	1,156,476	leaf	25
	GSM717875	4,472,811	1,410,957	728,067		25
	GSM717876	4,558,622	2,122,939	1,154,615		25
	GSM717877	3,972,826	1,039,594	416,244		25
Solanum lycopersicum	GSM304985	390,987	222,268	127,080		50
	GSM304986	159,821	100,111	60,058		50
	GSM452712	4,866,044	1,918,529	1,326,550		50
	GSM452713	4,199,438	1,856,022	1,231,114		50
	GSM452714	2,716,181	1,254,562	758,398		50
	GSM452715	2,024,256	944,578	542,008		50
	GSM452716	3,340,902	1,549,707	914,480		50
	GSM452717	5,207,307	2,093,422	1,420,047		50
	GSM452718	4,695,137	1,169,884	769,327		50
	GSM452719	5,148,196	1,420,351	963,688		50
	GSM452720	1,381,036	388,686	259,431		50
	GSM452721	3,157,377	915,093	606,874		50
Vitis vinifera	GSM458927	3,346,481	216,139	89,743		25
	GSM458928	4,521,156	554,445	200,504		25
	GSM458929	5,289,281	688,069	317,233		25
	GSM458930	3,933,002	489,094	179,467		25
Zea mays	GSM306487	5,775,485	3,354,592	1,807,921		50
	GSM306488	7,273,605	2,232,702	1,214,934		50
	GSM381716	4,406,055	1,639,984	1,102,715		50
	GSM381738	3,960,345	709,440	431,597		50
	GSM433620	3,796,402	1,247,806	825,167		50
	GSM433621	4,644,825	2,603,191	1,961,661		50
	GSM433622	4,207,601	2,463,445	1,947,645		50

[1] GEO Accession number for each library.

[2] Number of total sequence reads in a given libraries. All reads with length shorter than 28 nt and longer than 15 bp were included.

[3] Number of unique sequence reads in a given libraries. This number was obtained after collapsing identical reads in each library.

[4] Number of unique reads mapped perfectly to the corresponding genome sequences.

[5] Filter parameter, which refers to the default value set for the parameter -c in the Perl script filter_alignment.pl of miRDeep-P.

Supplemental Table 2. Identified miRNAs in the 15 examined plant species.

Since this supplemental table including more than 4,000 line, it could be downloaded from http://people.virginia.edu/~xy2y/Research%20interests.html under ongoing projects, uncovering evolutionary tracks of miRNAs in plants.

Supplemental Table 3. Neighboring miRNAs with compact genome organization in the 15 examined plants.

Species	Model I*	Model II	Model III	Model IV	Total
Arabidopsis thaliana	20 (11%)	12 (7%)	14 (8%)	6 (3%)	52 (30%)
Arabidopsis lyrata	8 (5%)	12 (7%)	8 (5%)	2 (1%)	30 (19%)
Citrus sinensis	12 (6%)	14 (7%)	16 (8%)	6 (3%)	48 (23%)
Glycine max	18 (5%)	14 (5%)	20 (7%)	2 (1%)	54 (19%)
Medicago truncatula	6 (2%)	16 (6%)	16 (6%)	2 (1%)	40 (14%)
Prunus persica	10 (6%)	10 (6%)	10 (6%)	2 (1%)	32 (20%)
Populus trichocarpa	21 (10%)	14 (6%)	14 (6%)	4 (2%)	53 (24%)
Vitis vinifera	12 (7%)	14 (8%)	4 (2%)	0 (1%)	30 (17%)
Solanum lycopersicum	20 (5%)	28 (8%)	28 (8%)	2 (1%)	78 (21%)
Mimulus guttatus	10 (4%)	18 (8%)	16 (7%)	4 (2%)	48 (21%)
Oryza sativa	24 (5%)	20 (4%)	24 (5%)	4 (1%)	72 (16%)
Brachypodium distachyon	23 (11%)	12 (6%)	14 (7%)	2 (1%)	51 (25%)
Zea mays	21 (6%)	26 (7%)	26 (7%)	4 (1%)	77 (21%)
Panicum virgatum	22 (3%)	16 (3%)	16 (3%)	2 (0.3%)	56 (9%)
Physcomitrella patens	20 (10%)	10 (5%)	14 (7%)	2 (1%)	46 (23%)
Total	247 (6%)	236 (6%)	240 (6%)	44 (1%)	767 (18%)

* See Figure 4 in main text for description of the four models.

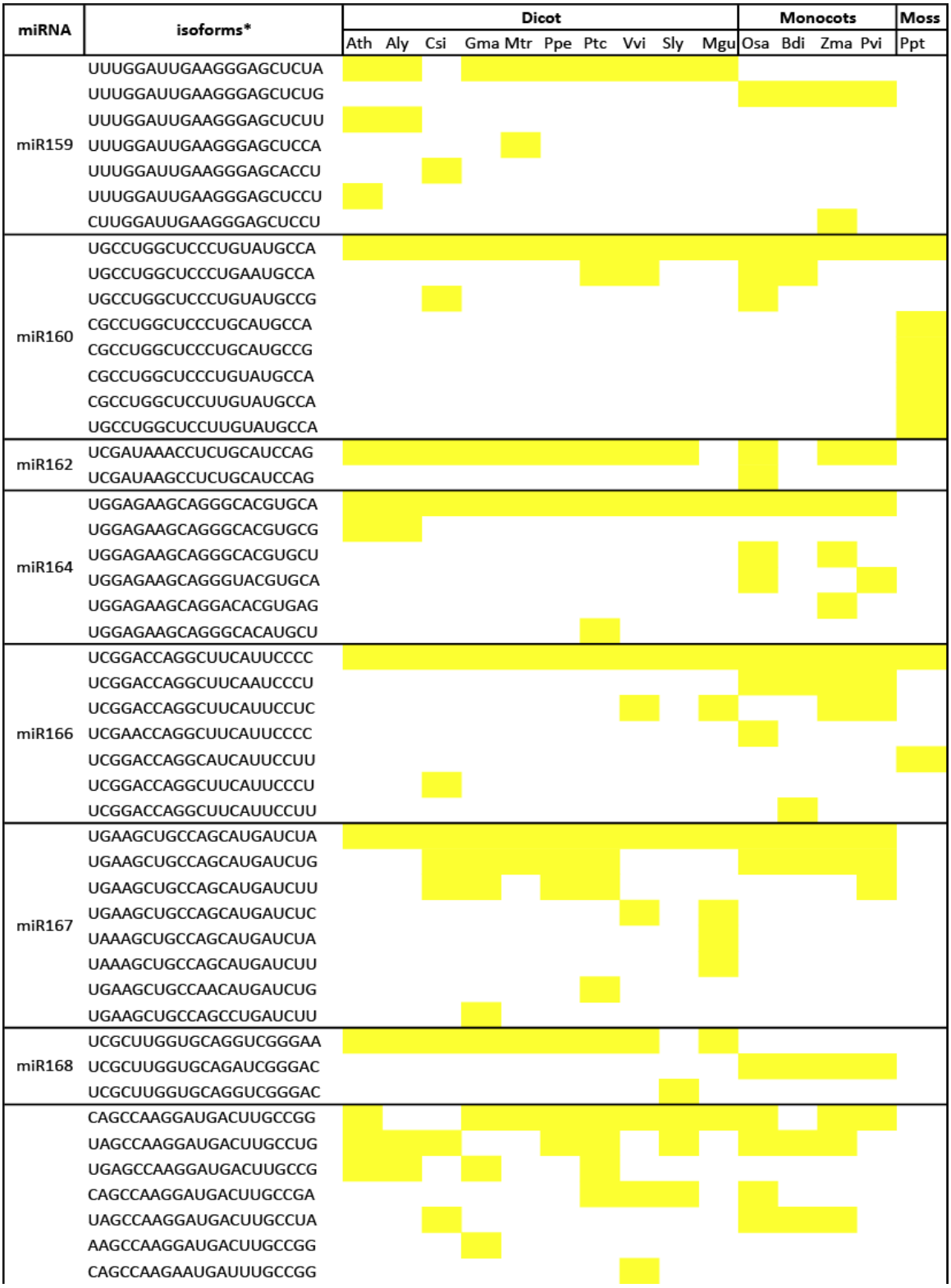
Supplemental Table 4. New species-specific miRNAs identified in *Arabidopsis thaliana*.

miRNA ID	Chr	Strand	Pre-miRNA coordinate	Mature sequence	Star*	Location
Ath-MIRN1a	Chr2	-	4098355..4098461	AGAAGACCCUUUAAAAACUCUUGU	Y	intergenic
Ath-MIRN1b	Chr1	-	22165454..22165532		Y	intergenic
Ath-MIRN1c	Chr2	-	2965506..2965624		Y	intergenic
Ath-MIRN2	Chr5	-	10845900..10846014	AGAGACAAAAAUAGAGAUCUCUAU	Y	intergenic
Ath-MIRN3	Chr1	+	18686860..18686965	AGCUAAGGAUUUGCAUUCUCA	Y	intron
Ath-MIRN4	Chr5	+	12025941..12026047	AGGCUUUUAAGAUCUGGUUGCGGU	N	intergenic
Ath-MIRN5	Chr5	-	14026971..14027045	AGGUAGUUUUCUGGCAGAUUUUA	Y	intergenic
Ath-MIRN6	Chr4	+	16864125..16864334	AUACGCUUGUAGUUUGAUUGUUA	Y	intergenic
Ath-MIRN7	Chr5	+	9394474..9394587	AUAGAAUUGUGGUAAGCUAGAAAU	Y	intergenic
Ath-MIRN8	Chr1	-	30019414..30019578	AUCCUUAUUGAUGAUCUCUUAACA	Y	intergenic
Ath-MIRN9	Chr1	-	13990675..13990752	AUCUUGGUCUUAGUCAGGUGG	N	intergenic
Ath-MIRN10	Chr5	-	22322727..22322890	AUUCGACAAAGUGAAGGGUUU	Y	intergenic
Ath-MIRN11	Chr2	-	10963712..10963792	UAGUAAACUGAGAAAACAGUU	Y	intron
Ath-MIRN12	Chr3	-	8984589..8984676	UCCCAGAUAGAAUCUCUUCUA	Y	intergenic
Ath-MIRN13	Chr2	-	5251126..5251199	UCUGGAAUUAGUAGUUCAGACAUU	Y	TE gene
Ath-MIRN14	Chr4	-	7161871..7161948	UGACUGCAUUAACUUGAUCGU	Y	intron
Ath-MIRN15a	Chr3	+	15088743..15088863	UGAGACGUACGUUAAGUGAUUCUC	Y	intergenic
Ath-MIRN15b	Chr3	+	16519764..16519884		Y	intergenic
Ath-MIRN15c	Chr4	-	4232401..4232523		Y	intergenic
Ath-MIRN16	Chr1	+	11786296..11786371	UGGAAGAUGCUUUGGGAUUUAUU	Y	5' UTR
Ath-MIRN17	Chr2	+	19686961..19687039	UGUUUUGGAUCUUAGAUACAC	Y	intron
Ath-MIRN18	Chr3	-	11747724..11747816	UUAGUUGACGGAAUUGUGGCG	N	pseudogene
Ath-MIRN19	Chr4	+	2179443..2179544	UUCCUUCUGAAAACUAAAAUU	N	intergenic
Ath-MIRN20	Chr5	+	13868432..13868508	UUCGGGUCGAUUCGGUUAUUUUA	N	intergenic
Ath-MIRN21	Chr5	-	8594675..8594738	UUCUGUAGACCCAUCAUACCUU	Y	intergenic
Ath-MIRN22	Chr2	+	14100021..14100100	UUGCUUAAAGAUUUUCUAUGU	Y	intron
Ath-MIRN23	Chr4	+	13295926..13296101	UUUGAUUUUACUGUAUAACUUG	Y	other_RNA
Ath-MIRN24	Chr2	+	11159720..11159789	ACCAGCUCGAAGAAGCUUAGCU	Y	intergenic**
Ath-MIRN25	Chr5	-	4479445..4479537	UGGGACUGCCUAAGCUAAAGG	Y	intron**
Ath-MIRN26	Chr4	+	7846689..7846813	AUGCAAUUUAGAAUAUGUUUU	Y	other_RNA**
Ath-MIRN27	Chr5	-	15891590..15891855	UCCAUACGGUAGAAGAAAGUAU	Y	intergenic**
Ath-MIRN28	Chr1	+	24553940..24554022	UUUGUACAACAUUUUUAGUGU	Y	intron**
Ath-MIRN29	Chr5	+	10943905..10944021	UUUUGUGGUUUUCUUGGUUAA	Y	intergenic**
Ath-MIRN30	Chr5	+	5169999..5170126	AUGAAUUUGGAUCUAAUUGAG	Y	intergenic**
Ath-MIRN31	Chr1	+	18686868..18686957	AUUUGCAUUCUCAUGAUUGAG	Y	intron**
Ath-MIRN32	Chr1	-	30019435..30019557	ACAUUACUCUUAGUUAAGAGAUAU	Y	intergenic**
Ath-MIRN33	Chr5	-	22322764..22322853	UCUUGCAAGGUUCAAGAACGGAUC	Y	intergenic**
Ath-MIRN34	Chr4	-	7161879..7161940	UUAACUUGAUCGUUGGUUGUU	Y	intron**
Ath-MIRN35	Chr4	+	13295943..13296184	ACUUGUAUUUACUGUAAUAACAAC	Y	other_RNA**
Ath-MIRN36	Chr4	+	16864140..16864319	UUUAGUGUUUAAGAUUAUACGCUU	Y	intergenic**

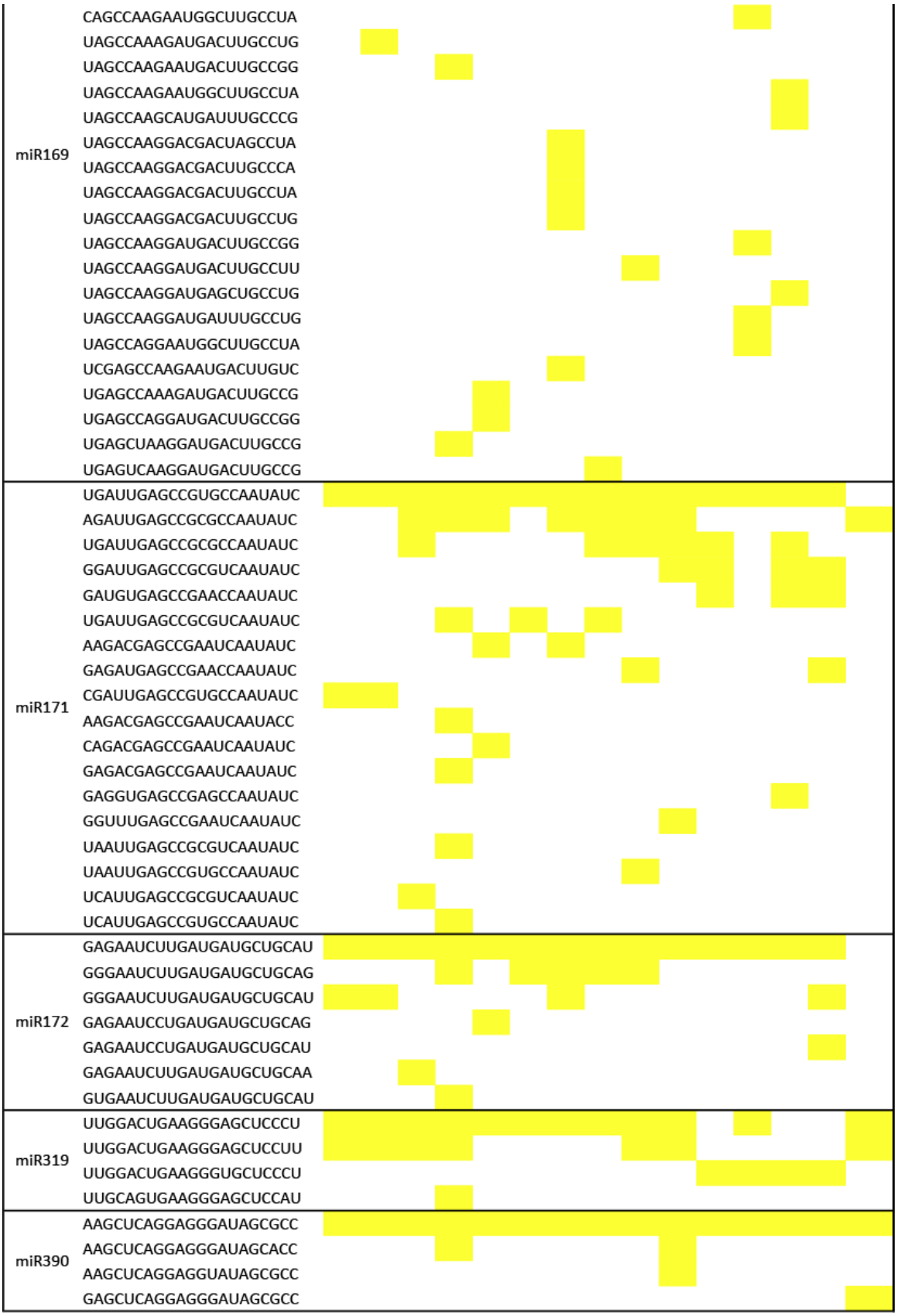
* Y and N denote whether or not sequence reads corresponding to the predicted miRNA star sequences were found in the examined small RNA libraries.

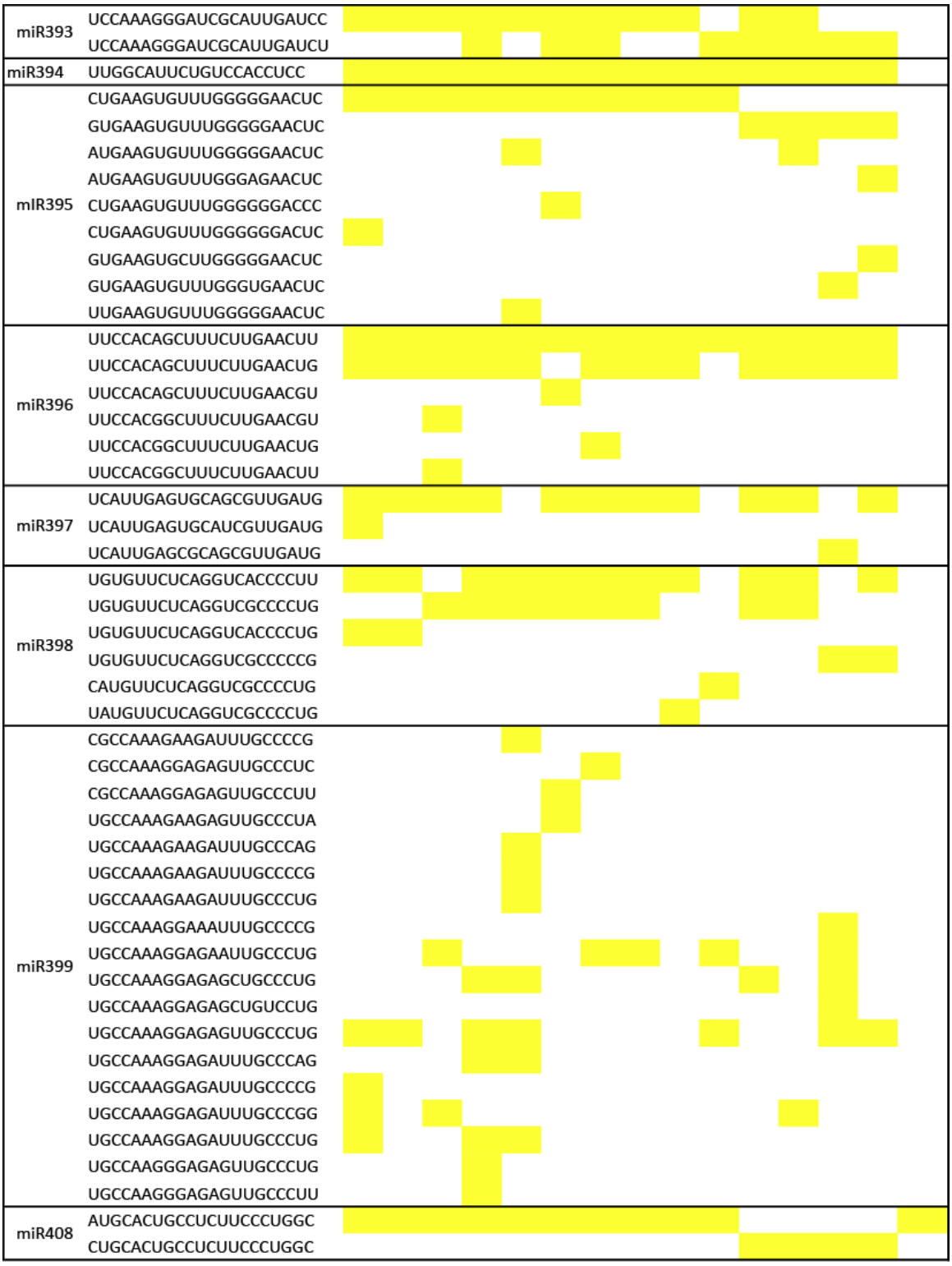
** These miRNAs have shared or overlapping precursors with other annotated or new miRNAs.

Supplemental Table 5. Distribution of isoforms of the 20 conserved miRNA families in 15 land plants.



miRNA	isoforms*	Dicot										Monocots				Moss
		Ath	Aly	Csi	Gma	Mtr	Ppe	Ptc	Vvi	Sly	Mgu	Osa	Bdi	Zma	Pvi	Ppt
miR159	UUUGGAUUGAAGGGAGCUCUA															
	UUUGGAUUGAAGGGAGCUCUG															
	UUUGGAUUGAAGGGAGCUCUU															
	UUUGGAUUGAAGGGAGCUCCA															
	UUUGGAUUGAAGGGAGCACCU															
	UUUGGAUUGAAGGGAGCUCCU															
	CUUGGAUUGAAGGGAGCUCCU															
miR160	UGCCUGGCUCCCUGUAUGCCA															
	UGCCUGGCUCCCUGAAUGCCA															
	UGCCUGGCUCCCUGUAUGCCG															
	CGCCUGGCUCCCUGCAUGCCA															
	CGCCUGGCUCCCUGCAUGCCG															
	CGCCUGGCUCCCUGUAUGCCA															
	CGCCUGGCUCCUUGUAUGCCA															
	UGCCUGGCUCCUUGUAUGCCA															
miR162	UCGAUAAACCUCUGCAUCCAG															
	UCGAUAAGCCUCUGCAUCCAG															
miR164	UGGAGAAGCAGGGCACGUGCA															
	UGGAGAAGCAGGGCACGUGCG															
	UGGAGAAGCAGGGCACGUGCU															
	UGGAGAAGCAGGGUACGUGCA															
	UGGAGAAGCAGGACACGUGAG															
	UGGAGAAGCAGGGCACAUGCU															
miR166	UCGGACCAGGCUUCAUUCCCC															
	UCGGACCAGGCUUCAAUCCCU															
	UCGGACCAGGCUUCAUUCCUC															
	UCGAACCAGGCUUCAUUCCCC															
	UCGGACCAGGCAUCAUUCCUU															
	UCGGACCAGGCUUCAUUCCCU															
	UCGGACCAGGCUUCAUUCCUU															
miR167	UGAAGCUGCCAGCAUGAUCUA															
	UGAAGCUGCCAGCAUGAUCUG															
	UGAAGCUGCCAGCAUGAUCUU															
	UGAAGCUGCCAGCAUGAUCUC															
	UAAAGCUGCCAGCAUGAUCUA															
	UAAAGCUGCCAGCAUGAUCUU															
	UGAAGCUGCCAACAUGAUCUG															
	UGAAGCUGCCAGCCUGAUCUU															
miR168	UCGCUUGGUGCAGGUCGGGAA															
	UCGCUUGGUGCAGAUCGGGAC															
	UCGCUUGGUGCAGGUCGGGAC															
	CAGCCAAGGAUGACUUGCCGG															
	UAGCCAAGGAUGACUUGCCUG															
	UGAGCCAAGGAUGACUUGCCG															
	CAGCCAAGGAUGACUUGCCGA															
	UAGCCAAGGAUGACUUGCCUA															
	AAGCCAAGGAUGACUUGCCGG															
	CAGCCAAGAAUGAUUUGCCGG															


CAGCCAAGAAUGGCUUGCCUA
UAGCCAAAGAUGACUUGCCUG
UAGCCAAGAAUGACUUGCCGG
UAGCCAAGAAUGGCUUGCCUA
UAGCCAAGCAUGAUUUGCCCG
UAGCCAAGGACGACUAGCCUA
UAGCCAAGGACGACUUGCCCA
UAGCCAAGGACGACUUGCCUA
UAGCCAAGGACGACUUGCCUG
UAGCCAAGGAUGACUUGCCGG
UAGCCAAGGAUGACUUGCCUU
UAGCCAAGGAUGAGCUGCCUG
UAGCCAAGGAUGAUUUGCCUG
UAGCCAGGAAUGGCUUGCCUA
UCGAGCCAAGAAUGACUUGUC
UGAGCCAAAGAUGACUUGCCG
UGAGCCAGGAUGACUUGCCGG
UGAGCUAAGGAUGACUUGCCG
UGAGUCAAGGAUGACUUGCCG
miR169
UGAUUGAGCCGUGCCAAUAUC
AGAUUGAGCCGCGCCAAUAUC
UGAUUGAGCCGCGCCAAUAUC
GGAUUGAGCCGCGUCAAUAUC
GAUGUGAGCCGAACCAAUAUC
UGAUUGAGCCGCGUCAAUAUC
AAGACGAGCCGAAUCAAUAUC
GAGAUGAGCCGAACCAAUAUC
CGAUUGAGCCGUGCCAAUAUC
AAGACGAGCCGAAUCAAUACC
CAGACGAGCCGAAUCAAUAUC
GAGACGAGCCGAAUCAAUAUC
GAGGUGAGCCGAGCCAAUAUC
GGUUUGAGCCGAAUCAAUAUC
UAAUUGAGCCGCGUCAAUAUC
UAAUUGAGCCGUGCCAAUAUC
UCAUUGAGCCGCGUCAAUAUC
UCAUUGAGCCGUGCCAAUAUC
miR171
GAGAAUCUUGAUGAUGCUGCAU
GGGAAUCUUGAUGAUGCUGCAG
GGGAAUCUUGAUGAUGCUGCAU
GAGAAUCCUGAUGAUGCUGCAG
GAGAAUCCUGAUGAUGCUGCAU
GAGAAUCUUGAUGAUGCUGCAA
GUGAAUCUUGAUGAUGCUGCAU
miR172
UUGGACUGAAGGGAGCUCCCU
UUGGACUGAAGGGAGCUCCUU
UUGGACUGAAGGGUGCUCCCU
UUGCAGUGAAGGGAGCUCCAU
miR319
AAGCUCAGGAGGGAUAGCGCC
AAGCUCAGGAGGGAUAGCACC
AAGCUCAGGAGGUAUAGCGCC
GAGCUCAGGAGGGAUAGCGCC
miR390



* Conserved miRNA families in addition to miR156/157 are analyzed as shown in Figure 8 in the main text. Distribution of the isoforms in the 15 examined plant species is indicated where a box shaded in yellow marks the presence of a particular isoform in a given species.

Chapter 5. MicroRNAs and alternative mRNA splicing

Alternative mRNA processing increases the complexity of microRNA-based gene regulation in *Arabidopsis*[1]

[1]Formatted as a co-authored manuscript and published as:

Yang X, Zhang H, Lei Li. 2012. *Plant J.* 70(3):421-431

Abstract

MicroRNAs (miRNAs) represent an important class of sequence-specific, *trans*-acting endogenous small RNA molecules for modulating gene expression at the post-transcription level. They function through binding to partial complementary *cis*-regulatory sites, called miRNA binding sites (MBSs), in their target mRNAs. Motivated by two recent observations from studying plant genomes, namely that alternative splicing is a common phenomenon and that miRNA regulates a significant portion of the transcriptome, we hypothesize that there is a possible mechanism for gene regulation that involves both processes. In the current effort, we performed a systemic search in the model plant *Arabidopsis thaliana* using annotated gene models as well as publically available high-throughput RNA-sequencing (RNA-Seq) data with a total of 570 million reads. For the 354 high confidence MBSs we compiled in *Arabidopsis*, we identified at least 44 (12.4%) impacted by alternative splicing such that mRNA isoforms of the same miRNA target gene differ in sequences encoding the MBS. Through simulation, we found the frequency of alternative splicing at MBS is significantly higher than other regions. Comparative and functional analyses further indicate that the alternative splicing events are important for target gene expression and miRNA action. Together our results revealed that alternatively spliced MBS is a plausible mechanism for attenuating miRNA-mediated gene regulation.

Introduction

Temporal and spatial control of transcript abundance for the expressed genes is crucial for many biological processes and developmental programs. In eukaryotes, gene expression regulation is orchestrated at both the transcription and post-transcription levels. Concurrent with transcription, the primary transcripts are recognized and processed by the spliceosome to splice out introns and join exons. Alternative processing of the primary transcripts through the combination of different splice junctions can produce multiple mRNA isoforms from a single gene. Alternative splicing is increasingly being recognized as a major cellular mechanism for generating transcriptome plasticity and proteome diversity in plants (Reddy, 2007). For example, 20-30% of the transcripts are found to be alternatively spliced in both *Arabidopsis* and rice through large scale EST-genome alignments (Campbell et al., 2006; Wang and Brendel, 2006). Most recently, deep sampling of the transcriptome using high-throughput RNA-Seq has indicated that at least 40% of intron-containing genes in *Arabidopsis* are alternatively spliced (Filichkin et al., 2010).

At the post-transcription level, miRNAs are emerging as an important class of sequence-specific, *trans*-acting endogenous small RNA molecules for modulating gene expression (Bartel, 2009; Voinnet, 2009). The 20-to-24 nucleotides (nt) long mature miRNAs are processed from much longer primary transcripts via stem-loop-structured intermediates (Bartel, 2009). In plants, miRNA processing is carried out in the nucleus mainly by the endonuclease Dicer-like 1 (DCL1; Papp et al., 2003). The mature miRNA is exported to the cytoplasm and integrated into the RNA-induced silencing complex (RISC) where the miRNA is used as a guide to recognize the mRNA targets through base

pairing with the MBS (Bartel, 2009). Interaction between the miRNA and its targets leads to repression of the target genes through cleavage (Llave et al., 2002; Reinhart et al., 2002) and translational inhibition (Brodersen et al., 2008) of the target transcripts, or miRNA-directed DNA methylation at the target loci (Wu et al., 2010). Currently, hundreds of miRNAs have been annotated in a broad spectrum of plant lineages (Kozomara and Griffiths-Jones, 2011) and dozens have been studied experimentally. It is well-established that many miRNAs are crucial for diverse plant development processes and responses to environmental challenges (Jones-Rhoades et al., 2006; Garcia, 2008; Voinnet, 2009).

Recent studies indicate that lineage-specific miRNAs have continuously emerged in evolution (Rajagopalan et al., 2006; Fahlgren et al., 2007; Molnar et al., 2007; Zhu et al., 2008; Yang et al., 2011). The new miRNAs, once incorporated into the gene regulatory networks through the formation of new MBSs, are thought to generate new gene circuits that expand the known scope of miRNA-mediated cellular processes (Rajagopalan et al., 2006; Fahlgren et al., 2007). For the many miRNAs that are deeply conserved across plant lineages (Axtell and Bowman, 2008), they usually regulate homologous targets at identical MBSs in every species in which they are found. Consistently, the MBSs are shown to be subject to strong purifying selection (Guo et al., 2008b). Based on these observations, it has been argued that genetic changes resulting in beneficial miRNA-target interactions are maintained while nonproductive or deleterious changes continue to drift or be purged (Chen and Rajewsky, 2006; Axtell and Bowman, 2008).

We are intrigued to understand how the seemingly rigid miRNA-target interaction, once the miRNA is fully incorporated into the regulatory gene networks, continues to evolve to allow fine-tuning of gene activity and adaptation. In the context of multi-layer gene regulation, it is worth noting that MBSs are present in mRNAs and must be transcribed and processed to be functional. Alternative processing of pre-mRNAs can conceivably eliminate or create functional MBSs in different mRNA isoforms of the same gene. Such alternative splicing events could provide a mechanism to bypass the strict constraint at MBSs while at the meantime maintain the interaction between miRNAs and their targets. Thus, global identification and analysis of alternative splicing events associated with MBSs is highly desirable to further understanding of the biological significance of this form of transcriptomic diversity and the intrinsic complexity of miRNA-target interactions in plants.

In the current effort, we performed a series of analyses in the model plant *Arabidopsis thaliana* to elucidate alternatively spliced MBSs as a mechanism for regulating gene expression. Through a genome-wide examination we identified 44 high confidence alternative splicing events from annotated gene models and RNA-Seq data that produce mRNA isoforms of the same miRNA target gene differing in the sequences required for miRNA binding. Comparative and functional studies indicate that these events are important for target gene expression and miRNA function. Together our results revealed that alternatively spliced MBS represents a plausible and prevalent mechanism for regulating miRNA-target gene circuits in plants.

Results

Compilation of high confidence MBSs

To conduct a genome wide survey of alternative splicing events that impact miRNA-based gene regulation in *Arabidopsis*, we sought to compile a comprehensive list of valid MBSs. Because plant miRNAs have near-perfect complementarity to their binding sites within the target mRNAs (Schwab et al., 2005; Bartel, 2009), they typically induce target cleavage and produce relatively stable RNA intermediates with a 5' phosphate (Llave et al., 2002). Cloning these cleavage intermediates based on the 5' Rapid Amplification of cDNA Ends (5' RACE) technique has been widely used to validate miRNA targeting and to identify the corresponding MBS in plants. We first obtained 109 individually validated MBSs from The Arabidopsis Information Resource (TAIR). We next utilized 510 putative MBSs identified from high-throughput sequencing of the 5' ends of polyadenylated RNAs that are predicted to be compatible with miRNA-mediated mRNA decay (German et al., 2008). Using a recent algorithm specifically designed for predicting plant miRNA targets (Alves et al., 2009), we filtered out MBSs not capable of forming stable RNA duplexes with the known miRNAs and collected 349 high confidence MBSs. By combining the two sets, we identified a total of 354 non-redundant MBSs for downstream analyses (Figure 1a; Table S1).

Identification of alternative splicing events at MBS

Conceptually, there are four possible models for alternative splicing to create a functional MBS in one mRNA isoform but not the other (Figure 1b). This can be achieved in Model I by the use of tandem splice donor or acceptor sites near the MBS, in

Model II by retention or removal of an intron either encompassing or encompassed in the MBS, in Model III by differential initiation or termination, and in Model IV by an exon cassette or mutually excluding exons (Figure 1b). Based on these models, we performed a systemic search using two sets of data to identify alternative splicing events relevant to the 354 MBSs of *Arabidopsis*. The first dataset contains 5,885 alternatively spliced genes obtained from TAIR that produce 13,594 gene models. Mapping both the MBSs and the alternate gene models to the genome revealed alternative splicing events that impact 14 MBSs. The second dataset involves publically available high-throughput RNA-Seq data generated from 12 libraries with over 570 million reads (Table S2). Utilizing this dataset, we identified 30 additional MBS impacted by alternative splicing, bringing the total cases to 44 (Figure 1a). The corresponding miRNAs are from both the most conserved families (e.g. miR156) and those have so far only been found in *Arabidopsis* (e.g. miR864; Table S1). Thus, current data indicates that alternatively spliced MBS is an experimentally supported phenomenon affecting a significant portion (12.4%) of the known miRNA target genes in *Arabidopsis*.

We use the *DCL1* gene (At1g01040) as an example to illustrate the impact of pre-mRNA splicing on a functional MBS. The annotated *DCL1* gene model (*DCL1-1*) contains 20 exons and possesses a functional binding site for miR162 that splits into exons 12 and 13 (Figure 2a, b; Xie et al., 2003). Analysis of RNA-Seq data revealed that intron 12 is spliced in multiple ways (Figure 2a). Removal of the 84-nt-intron generates *DCL1-1* (Figure 2b). Presence of specific junction reads reveals another isoform (*DCL1-2*) that is derived from utilizing a wobble splice donor site within intron 12 and differs from *DCL1-1* by a GUU trinucleotide (Figure 2a-c). The sequencing reads also supports

retention of the entire intron, which results in *DCL1-3* (Figure 2b). Additionally, we performed RT-PCR to clone a portion of the *DCL1* transcript (Figure 2c). By sequencing 19 independent clones (Figure 2d), we discovered a shorter *DCL1* isoform that is derived from an in-frame cryptic splice acceptor site located within exon 13 and 69 nt downstream of the known splice site of intron 12 (Figure 2b). Inspection of all four *DCL1* gene models indicates that only *DCL1-1* forms an mRNA:miRNA duplex that is compatible with miR162 targeting (Figure 2e). In previous studies (Xie et al., 2003; German et al., 2008), the 5' ends of cleaved *DCL1* mRNA as a result of miRNA-guided slicing were all mapped to *DCL1-1* (Figure 2e). Together these results indicate that pre-mRNA processing generates sequence polymorphisms around the miR162 binding site in *DCL1* such that only *DCL1-1* possesses the functional MBS.

Alternative splicing of the MBS in non-coding *TAS* genes

Most plant MBSs locate in the open reading frame (ORF), which creates a constraint on splicing. To evaluate the effect of alternative splicing on MBS without the constraint of an ORF, we examined the *TAS* genes that produce non-coding transcripts. Eight *TAS* genes have been identified in *Arabidopsis* that, after miRNA-directed cleavage of the primary transcript, each produces a series of *trans*-acting siRNA (Allen et al., 2005; Chen et al., 2007). RNA-Seq data indicates that at least for *TAS1A* and *TAS2*, intron removal results in the exclusion of the MBS, indicating that MBSs in non-coding genes are also subject to regulation by alternative splicing.

TAS1A (At2g27400) encodes a 930 nt transcript (*TAS1A-1*) that contains a functional binding site for miR173 (Yoshikawa et al., 2005; Montgomery et al., 2008).

RNA-Seq data conclusively show that this gene is subject to alternative splicing (Figure 3a). Based on specific junction reads, we propose that a 572 nt and a 594 intron can be spliced out from *TAS1A-1*, which results in *TAS1A-2* and *TAS1A-3*, respectively (Figure 3b). Either *TAS1A-2* or *TAS1A-3* no longer contains the MBS (Figure 3c). Indeed, scanning the sequences of all three *TAS1A* isoforms revealed no plausible binding site for miR173 except for the original MBS in *TAS1A-1* (Figure 3d). In contrast to the protein-coding gene *DCL1* (Figure 2d), RT-PCR and sequence analysis revealed that *TAS1A-2*, the isoform without the MBS, is predominant in all the organ types examined (Figure 3e). Similar results were also obtained when *TAS2* was analyzed (Figure S1). These results indicate that in the absence of an ORF, alternative splicing appears to be a major determinant for the relative abundance of transcripts competent for miRNA-based regulation.

Frequency of alternative splicing increases at MBS

We reasoned that increased frequency of alternative splicing at MBS, relative to other transcribed regions of the genome, would be indicative of selection and hence functional importance of this form of gene regulation. To test this notion, we first randomly extracted 354 (the same as the number of MBSs) 21-nt-long (typical length of MBS) mRNA segments in *Arabidopsis* and identified splicing events coincide with these sequences. For this analysis, we did not consider intron retention for the RNA-Seq data (corresponding to 3 cases of alternatively spliced MBSs) because of potential ambiguity (Figure S2). After reiterating the process for 1,000 times, we found the simulated frequency of splicing events impacting the random RNA segments forms a normal

distribution with the mean and standard deviation being 27.1 and 3.96, respectively. The observed number of alternatively spliced MBSs (41) is significantly ($P < 0.001$) greater than the mean of the simulated data (Figure 4a).

To rule out the possibility that elevated alternative splicing frequency at MBS is caused by unusually high background of alternative splicing of the miRNA target genes, we compared MBSs against the flanking mRNA regions. After aligning all MBSs, we used a sliding window approach to establish a trend line of the frequency of alternative splicing in the regions surrounding MBS. This analysis clearly shows a decrease of splicing frequency in the regions outside of MBS (Figure 4b), indicating that the observed high frequency of alternative splicing at MBS is specific to the binding sites. We further analyzed the occurrence of the dinucleotides GT (the almost invariant sequence of splice donor site) and AG (the almost invariant sequence of splice acceptor site) in the upstream and downstream regions of MBS, respectively. This analysis indicates that frequency of the potential splice sites tracks that of observed splicing events very well (Figure 4c). Together these results demonstrate an elevated splicing frequency at MBS in *Arabidopsis*, which correlates with the occurrence of potential splice sites immediately next to the MBSs.

Increased frequency of alternative splicing at the MBS suggests that the splicing events may not be conserved in different species due to divergence of sequences outside the MBS. To test this prediction, we obtained and compared *DCL1* sequence from 21 plant species. Although exon 13 is highly conserved, the in-frame alternative acceptor site is conserved only in the Brassicaceae and Rosaceae, but not any other families (Figure S3). By contrast, the wobble donor site in intron 12 is only conserved in the

Brassicaceae and Fabaceae (Figure S3). The sizes of intron 12 in all other species, unlike that in *Arabidopsis* (84nt), are not multiples of three (not shown). It is unlikely these introns will be retained because of frame shift. Together these results indicate that specific alternative splicing event at MBS may have phylogenetic distribution that is more limited than the MBSs themselves.

Functional implication of an alternatively spliced MBS

The miR156-regulated *SPL4* gene encodes a transcription factor in the *SPL* (SQUAMOSA Promoter Binding Protein-Like) gene family (Guo et al., 2008a). Function of this gene circuit in regulating vegetative to reproductive development has been well characterized in *Arabidopsis* (Wu and Poethig, 2006; Wang et al., 2009; Wu et al., 2009). We therefore selected *SPL4* as an example to demonstrate the role of alternatively spliced MBS in gene expression and function. There are two annotated gene models from the *SPL4* locus. At1g53160.1 (SPL4-1) has a two-exon structure and contains an experimentally confirmed miR156 binding site located within its 3' untranslated region (Figure 5a, b; Wu and Poethig 2006). Another annotated gene model At1g53160.2 (SPL4-3) contains three exons and differs from *SPL4-1* immediately downstream of the stop codon. The two mRNA isoforms encode an identical protein but the MBS is spliced out in *SPL4-3* (Figure 5a, b). In sequencing the PCR amplicons derived from *SPL4* mRNAs, we found a new isoform in which an upstream donor site for the last intron is used. Not only does this splicing event eliminate the MBS, but alter the N terminal portion (last 18 amino acids) of the protein (Figure 5a, b). We name the new gene model represented by this mRNA isoform *SPL4-2*. We then scanned the three gene models for

possible miR156 binding sites and found only the original one in *SPL4-1* (Figure 5c). Thus, alternative splicing produces multiple *SPL4* mRNA isoforms with or without the binding site for miR156.

To examine the influence of alternative splicing on the regulation of *SPL4*, we compared expression pattern of the three mRNA isoforms with that of miR156 in various organ types. RT-PCR analysis revealed presence of the three isoform-specific products in the seven examined organ types, albeit at varying abundance. We found *SPL4-1* is the predominant isoform as its expression level is higher than that of *SPL4-2* and *SPL4-3* in essentially all the examined samples (Figure 5d). Consistent with previous reports (Wu and Poethig, 2006), expression level of *SPL4-1* is highest in adult tissues (e.g. rosette and cauline leaf) and lowest in seedling and root. Such a pattern is clearly anti-parallel to that of miR156 which is high in seedling and root and low in adult leaves (Figure 5e). The only exception is silique in which miR156 is not expressed. By contrast, the other isoforms (especially *SPL4-3*) do not exhibit such an anti-parallel expression pattern with miR156. We further compared wild type and transgenic plants in which miR156 expression is driven by the constitutive CaMV 35S promoter. In the 35S::miR156 plants, miR156 expression level is elevated in rosette leaf (Wu and Poethig, 2006). However, we found that only *SPL4-1* is down-regulated but not *SPL4-2* or *SPL4-3* (Figure 5f). These results indicate that alternative spliced *SPL4* transcripts with or without the miR156 binding site are regulated differently by miR156. We therefore conclude that alternative splicing at the MBS constitutes a mechanism for controlling the development-specific expression levels of the *SPL4* gene.

SPL4 and its paralog *SPL3* both contain a functional miR156 binding site (Figure S4a, b; Wu and Poethig 2006). However, only *SPL4*, but not *SPL3*, has been found capable of producing alternative mRNAs that differentiate in the miR156 binding site. We reasoned that if alternative splicing is important for the regulation of *SPL4*, its expression pattern should deviate further from that of *SPL3* in conditions where miR156 is expressed. To test this hypothesis, we reconstructed the expression profiles of *SPL3* and *SPL4* during the vegetative-to-reproductive phase based on published microarray data (Balasubramanian et al., 2006). Because miR156 level decreases in this phase (Wu and Poethig, 2006; Wang et al., 2009; Wu et al., 2009), the steady state level of *SPL3* and *SPL4* both increases (Figure S4c). Importantly, it has been shown that miR156 exhibit circadian regulation (Hazen et al., 2009). However, only *SPL3*, but not *SPL4*, shows a moderate clock-dependent expression pattern (Figure S4c), indicating that the endogenous *SPL4* isoforms without the miR156 binding site contribute to the distinctive expression pattern of the *SPL4* gene.

Because of genetic redundancy among the *SPL* genes, loss-of-function mutation in *SPL4* does not results in obvious development defects (Wu and Poethig, 2006). We therefore chose to demonstrate the importance of the alternative transcripts by a gain-of-function approach. We created transgenic lines over-expressing the three *SPL4* isoforms driven by the CaMV 35S promoter (Figure 6a) and assayed their development phenotypes (Figure 6b). Over-expressing *SPL4-1* has no significant effect on timing of flowering compared to wild type (Figure 6b, c) but at flowering causes slight reduction in the number of adult leaves (leaves with abaxial trichome; Figure 6d). By contrast, the 35S::SPL4-2 and 35S::SPL4-3 plants in which the over-expressed *SPL4* transcripts all

lack the MBS, have accelerated rate of flowering induction (Figure 6b, c) and significantly fewer adult leaves at bolting (Figure 6d). Thus, constitutive expression of the *SPL4* transcripts with or without the miR156 binding site resulted in distinguishable phenotypes.

Discussion

As *trans*-acting gene regulators, miRNAs function through binding to the complementary MBS in their target mRNAs (Bartel, 2009). MBSs are present in mRNA molecules and need to be transcribed from DNA. In the current study, we show in the model plant *Arabidopsis* mRNA splicing can generate transcript isoforms that differentiate in the MBS. This process impacts a substantial portion of the MBSs (Figure 1a; Table S1) and involves all major types of alternative RNA processing (Figure 1b; Reddy 2007). Two observations led us to conclude that the reported cases of alternatively spliced MBSs are an underestimation. First, the RNA-Seq data used in the current study were derived from limited tissue types and physiological conditions (Table S2) such that low expression genes or rare isoforms may not be sufficiently covered (e.g. Figure S2). Second, different *DCL1* isoforms were recovered from direct amplicon sequencing and whole genome RNA-Seq (Figure 2d), indicating that neither analysis is exhaustive. As miRNAs and miRNA target genes are increasingly being discovered in plants, alternative splicing of mRNA sequences encoding functional MBSs thus has the potential to significantly enhance the regulatory complexity of miRNA-mediated gene networks.

A series of intriguing questions ensue following the initial discovery. First, is alternatively spliced MBS a regulated event or simply coincidence between two separate molecular processes? MBSs are parts of the mRNA molecules so it is perhaps not surprising that some are alternatively spliced by chance. However, our results show that the frequency of alternative splicing events encompassing MBS is significantly higher than other portions of the transcriptome (Figure 4a). Within the miRNA target genes, the frequency of alternative splicing specially increases at the MBS (Figure 4b). Importantly,

we show the elevated frequency of alternative splicing tracks the occurrence of potential splice sites surrounding the MBS (Figure 4c). These results collectively suggest a scenario in which the presence of MBS causes changes to local sequences which in turn increases the possibility of alternative mRNA splicing. Future experiments aimed at detecting the biochemical interactions between the nucleus-located RISC and components of the spliceosome (Bayne et al., 2008; Ohrt et al., 2008) should help to fully elucidate this potential new genetic mechanism.

Second, will alternatively spliced MBSs influence the regulation of the corresponding miRNA target genes? Because plant MBSs often locate in the coding regions, it is important to distinguish the impact on the encoded proteins from the MBSs when the influence of alternative splicing to the miRNA target genes is concerned. This was one of the reasons prompted us to choose *SPL4* of which two mRNA isoforms (*SPL4-1* and *SPL4-3*) encode an identical protein but are differentiated by the presence or absence of the miR156 binding site (Figure 5). We found that alternatively spliced *SPL4* transcripts with or without the miR156 binding site are regulated differently by miR156 at the post-transcription level in a development-specific manner (Figure5; Figure S4). We also found that increasing the expression level of the *SPL4* transcripts with or without the miR156 binding site by a transgenic means resulted in distinguishable phenotypes (Figure 6). Together with the finding that at least two *TAS* genes are alternatively spliced at the MBS (Figure 2; Figure S1), our results indicate that alternatively spliced MBS constitutes a mechanism to specifically attenuate miRNA-based regulation of the target genes.

Third, what is the influence of alternatively spliced MBSs on the function of miRNAs? In addition to birth and death of miRNA genes, MBSs conceivably could serve as an important determinant for the evolutionary dynamics of miRNA-target interactions. On the one hand, the short length of MBS may have given them the opportunity to mutate and change specificity rapidly in evolution. On the other hand, many MBSs are deeply conserved across plant lineages (Axtell and Bowman, 2008), suggesting that they are under strong purifying selection. An investigation of paralogous gene families targeted by miRNAs in rice supports this view (Guo et al., 2008a). Conserved MBSs in human are also found under stronger purifying selection than surrounding sequences (Chen and Rajewsky, 2006). Compared to mutation in either the MBS or the miRNA, alternatively processing of MBS at the mRNA level provides a more flexible mechanism for a miRNA target gene to attenuate the rigid miRNA regulation while at the same time maintain its competence for such regulation. The additional layer of regulation conferred by alternative splicing could link spliceosome activity to the regulation of certain miRNA-target interactions. A predicted advantage of this mechanism is that coordinated regulation by miRNAs and splicing of the target genes could allow the plants to integrate multiple input signals to fine tune target transcript abundance for the precise and timely execution of developmental programs.

Finally, what is the implication of alternative splicing in the evolution of miRNA-target gene circuits? Given that many plant MBSs locate in the coding regions, the selection at the MBS should have substantially influenced the evolution of miRNA target genes and hence the biological processes involving these genes. Our work shows that one of such influences is increased occurrence of mRNA splicing sites near the MBS to

generate transcript isoforms that bypass miRNA regulation (Figure 4). Consistent with reports from cross-species EST alignments (Wang et al., 2008), we show that the alternative splicing events impacting MBS could be family- or species-specific (Figure S3). This means that mRNA processing could bring divergence to the miRNA-mediated gene networks beyond the presence of new miRNA genes across plant lineages. Our finding further implies that genetic changes in other parts the gene where the selective constraint might be weaker (e.g. synonymous mutations that create/eliminate functional splice sites) could be used to modulate the selected MBSs. The many possible ways (e.g. in-frame cryptic splice sites) to impact MBSs by alternative splicing suggest that this mechanism may constitute a force for differentially regulating orthologous genes that underlie numerous developmental processes. The divergence brought about by this force can conceivably provide a genetic basis for the great physiological and ecological diversity among plants.

Taken together, our results indicate that alternatively spliced MBS is a plausible and prevalent mechanism for regulating gene expression in *Arabidopsis*. Further identification of alternatively spliced MBS in other plants is therefore important to fully elucidate the biological significance and intrinsic complexity of this form of gene regulation. Such inquiries will likely provide new insights into the gene networks that integrate both transcriptional and post-transcriptional regulations. Perhaps more importantly, these efforts should create opportunities to integrate different genetic components both within the genome (e.g. mRNA processing and miRNA function) and within the genes (MBSs and sequence changes that create/eliminate splice sites) to

provide holistic view of gene expression programs that underpin development and responses to environmental challenges in diverse plant species.

Experimental procedures

Data Sources

Annotated whole genome, cDNA, intron sequences and other gene model features of *Arabidopsis thaliana* used in this study correspond to release 10 from TAIR and were downloaded from ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/. The 12 RNA-Seq libraries employed to verify annotated alternative gene models and to detect new splicing events were download from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; Table S2). 5' RACE-validated miRNA targets (set I) were obtained from TAIR at ftp://ftp.arabidopsis.org/home/tair/Genes/SmallRNAs/. Putative miRNA targets identified through high-throughput sequencing of 5'RACE products (set II; German et al., 2008) were obtained at http://mpss.udel.edu/at_pare/. The miRNA dataset of *Arabidopsis* was downloaded from the miRBase (release 16.0; http://www.mirbase.org/). BLAST search for homologous *DCL1* sequences was performed using the Nucleotide collection, EST and Genomic survey sequences databases in NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

MBS filtration and compilation

Set II miRNA target genes were assessed using a method previously described (Alves et al., 2009) to identify high confidence MBSs. Briefly, the program RNAhybrid (Rehmsmeier et al., 2004; Kruger and Rehmsmeier, 2006) was used to predict the energetically plausible miRNA:mRNA duplexes with plant-specific constrains: (i) perfect base pairing of the miRNA:mRNA duplex from nucleotide 8 to 12 counting from the 5' end of the miRNA; (ii) loops and bulges no longer than one nucleotide long in either

strand; (iii) end overhangs no longer than two nucleotides in size; and (iv) G:U wobble base pairs not treated as mismatches, but contributing less favorable to the overall free energy. The putative MBSs were further filtered using the ratio between minimum free energy (mfe) of the identified miRNA:mRNA duplexes to the minimum duplex energy (mde) of the miRNA when bound by perfectly match targets as calculated in RNAhybrid. A cutoff value of mfe/mde at 0.70 was applied according to Alves et al. (2009). 349 of the 510 set II MBSs were retained as a result of the filtering steps. These MBSs were combined with the 109 set I MBS to reach a non-redundant collection of 354 high confidence MBSs.

Identification of new splicing events from RNA-Seq data

The RNA-Seq data utilized in this study include 12 independent libraries with more than 570 million trimmed and filtered reads (Table S2). To identify new junction reads for the annotated gene models, the spliced read mapping tool TopHat, which is a read-mapping algorithm designed to align RNA-Seq reads to a reference genome without relying on known splice sites (Trapnell et al., 2009), was employed. Using the built-in short read aligner Bowtie (Langmead et al., 2009) in TopHat, more than 350 million reads were mapped to *Arabidopsis* genome, yielding an average coverage depth of 105 reads per base for each DNA strand. Five parameters were set to facilitate the mapping of reads to the annotated *Arabidopsis* gene models: i) maximal intron length allowed was set to 2000 bases as the vast majority of introns (99.5%) of annotated *Arabidopsis* gene models are smaller than 1200 bases; ii) maximal mismatches allowed in read alignment was two; iii) no mismatch was allowed in an spliced alignment; iv) minimal number of

matched bases of a junction read at either end was set at eight; and v) only the classic splicing motifs GT-AG and GC-AG were used. For intron retention, we only considered introns that were completely covered by mapped sequencing reads.

Comparison of alternative splicing frequency at MBS and other genome regions

To compare MBS with other transcribed regions, a pool of 50,000 21-nt-long sequences was randomly generated from all gene models in TAIR 10 annotation, excluding pseudogenes. These short sequences were searched against all annotated alternate gene models and the RNA-Seq data to identify those that are impacted by alternative splicing. A subset of 354 short sequences was then randomly picked from the pool and the number of sequences impacted by alternative splicing calculated. After reiterating the process for 1000 times, a histogram of the frequency of alternatively spliced sequences was determined. A curve fitting approach was used to determine the best-fitted normal distribution curve following the equation:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ, the mean value, was determined to be 27.1 and σ, the standard deviation, 3.96.

To compare MBS to other portions of the miRNA target transcripts, the transcript sequences were stacked and aligned at the MBS. Starting from the MBS, sliding windows (window size = 21 nt, step = 10 nt) were applied at both directions. In each window, the number of sequences impacted by alternative splicing was obtained as percentage of the total sequences in that window. In each window, occurrence of the dinucleotides GT and AG was also recorded. The numbers were then normalized against the window at symmetric positions on the opposite side of the MBSs.

Plant materials

Arabidopsis plants used in this study were in Columbia background. Seeds were grown on GM media and incubated at 4°C in the dark for 4 days after which exposed to continuous white light (~150 μmole sec-1 m-2) at 22°C. Seedlings were transferred to soil and maintained under continuous light until plant organs at various development stages were collected as indicated in the text. Plant age was measured from the time seeds were exposed to light. Flowering time represents the time takes for the first open flower to appear. To generate transgenic plants over-expressing *SPL4* isoforms, the *SPL4-1*, *SPL4-2*, and *SPL4-3* cDNA was PCR-amplified with the Pfusion DNA polymerase (New England Biolabs) using primers listed in Table S3. PCR products were cloned into the 35S-pKANNIBAL vector between the XhoI and SpeI restrict sites and sequenced. Plants were transformed with the sequence-confirmed constructs using the standard floral dipping method. Transformants were selected on BASTA-containing (20mg/L) media. T3 generation plants homozygous for individual transgenes were identified by PCR analysis of genomic DNA and used for all experiments.

RNA analyses

Total RNA was extracted using the TRIzol agents (Invitrogen) and treated with RNase-free DNaseI. For RT-PCR analysis, total RNA was reverse transcribed using the SuperScript II reverse transcriptase (Invitrogen). The resultant cDNA was PCR amplified using the Phusion DNA Polymerase. PCR products were cloned into the pCR4Blunt-TOPO vector (Invitrogen) and sequenced. qPCR analysis of the reversed transcribed cDNA was carried out using the ABI 7500 system and the Power SYBR Green PCR

master mix (Applied Biosystems). Actin7 was used as an endogenous control to normalize the relative expression level. At least three independent experiments were performed for each amplicon and the data were analyzed using the ABI 7500 System SDS software (Applied Biosystems). Primer sequences are listed in Table S3. Low-molecular-weight RNA blotting was performed as previously described (Yang et al., 2011; Zhang et al., 2011). Probe sequences are listed in Table S3.

Figure 1. Identification of alternatively splicing MBSs in the model plant *Arabidopsis*

(a) Numbers of experimentally supported MBSs identified in the *Arabidopsis* genome (All MBSs) and alternatively spliced MBSs are shown. (b) Models for identifying alternatively spliced MBSs. Based on mRNA to genome alignment, MBSs are differentially spliced in Model I by use of tandem splice donor or acceptor sites, in Model II by retention or removal of an intron, in Model III by differential initiation or termination, and in Model IV by an exon cassette or mutually excluding exons.

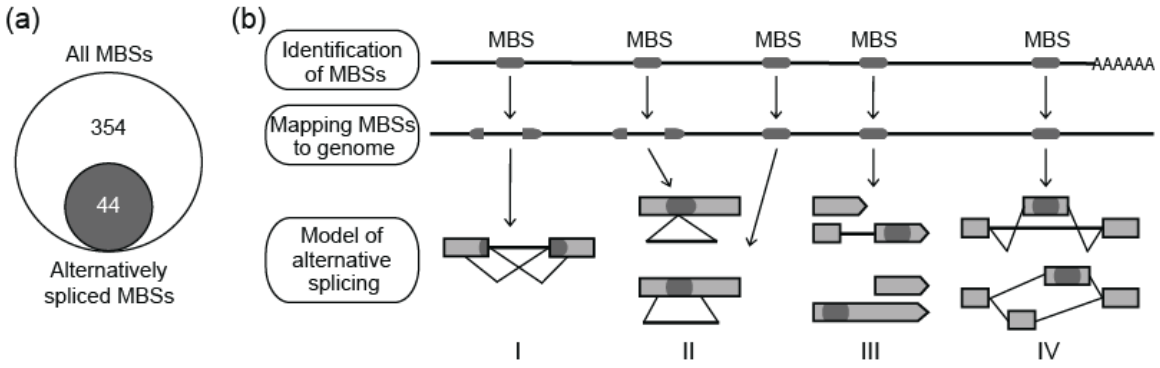
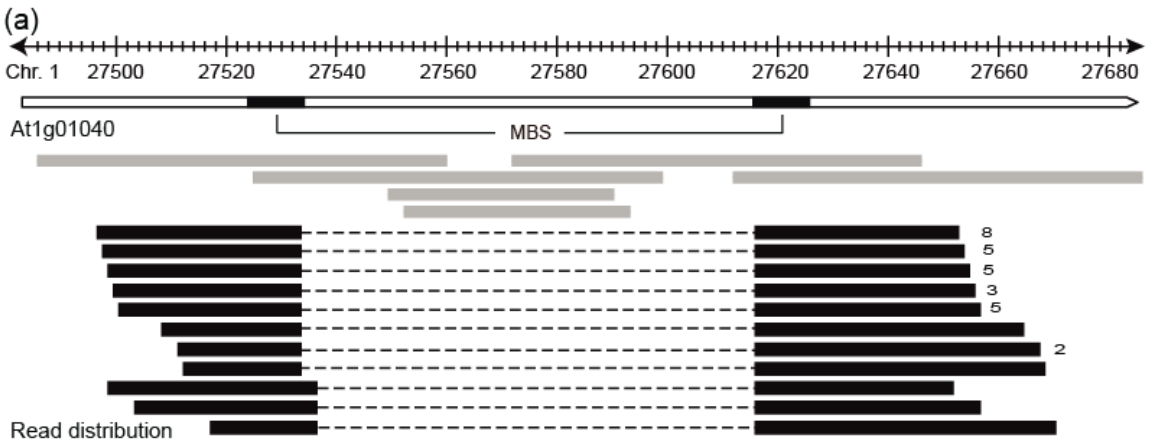
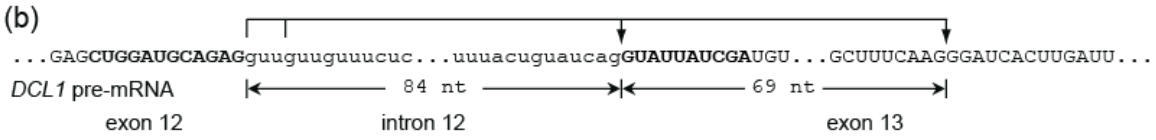
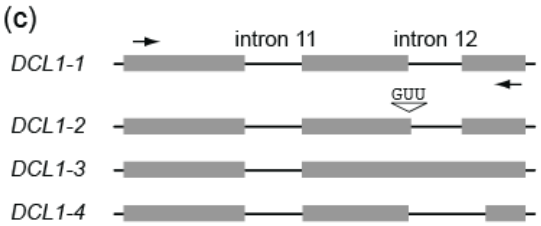

(a)
All MBSs
354
44
Alternatively spliced MBSs
(b)
Identification of MBSs
Mapping MBSs to genome
Model of alternative splicing
MBS
MBS
MBS
MBS
MBS
AAAAAA
I
II
III
IV

Figure 2. Alternative splicing of the miR162 binding site in *DCL1*

(a) RNA-Seq reads mapped to the genomic region surrounding the MBS. Chromosome coordinates are shown on top. Reads mapped to a single locus and those split into two loci are depicted as grey and black horizontal bars, respectively. Copy number of reads retrieved more than once is indicated. (b) Proposed splicing models of intron 12 in the *DCL1* pre-mRNA. Exon and intron sequence of the annotated *DCL1* gene model are shown in upper and lower case letters, respectively. MBS is shown in bold upper case letters. The two splice donor sites, which are separated by 3 nt, and the two acceptor sites, which are separated by 69 nt, are indicated by the arrows. (c) Partial gene structure of the four identified *DCL1* isoforms. Exons and introns are depicted as boxes and horizontal lines, respectively. Positions of the primers used for obtaining amplicons flanking the MBS are indicated by the horizontal arrows. (d) Comparison of the results from RNA-Seq and amplicon sequencing. *DCL1* amplicons are obtained by RT-PCR analysis of seedling using primers indicated in *C*. (e) Predicted RNA duplex between miR162 and the four *DCL1* isoforms. Base paring is indicated by a short vertical line and G:U wobble pairing by the "○" sign. Mapped 5' ends of *DCL1* mRNA resulted from miR162-guided cleavage are indicated by the vertical arrows. Numbers are combined from Xie *et al.* (2003) and German *et al.* (2008).







(d)

	DCL1-1	DCL1-2	DCL1-3	DCL1-4
number of clone	15		3	1
supportive reads	√	√	√	

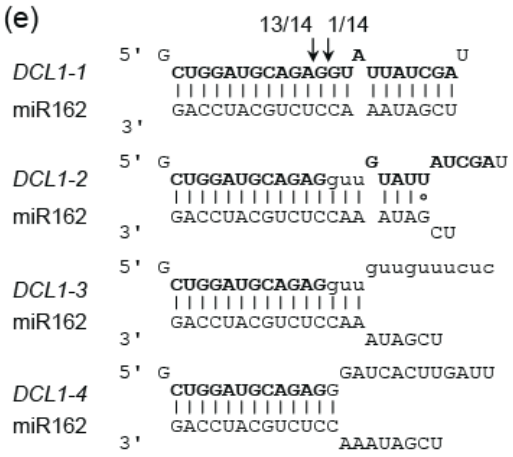


Figure 3. Alternative splicing of the miR173 binding site in *TAS1A*

(a) RNA-Seq reads mapped to the genomic region surrounding the MBS. Chromosome coordinates are shown on top. Reads mapped to a single locus and those split into two loci are depicted as grey and black horizontal bars, respectively. Copy number of reads retrieved more than once is indicated. (b) Proposed splicing model of *TAS1A*. MBS is shown in upper case letters. The optional splice donor site and the two acceptor sites are indicated by the arrows. (c) Gene structure of the three identified *TAS1A* isoforms. Exons and introns are depicted as boxes and horizontal lines, respectively. Positions of the primers used for obtaining amplicons flanking the MBS are indicated by the horizontal arrows. (d) Predicted RNA duplex of miR173 and *TAS1A-1*. Base paring is indicated by a short vertical line and G:U wobble pairing by the "○" sign. (e) Developmental expression profiles of *TAS1A-1* and *TAS1A-2* for which the amplicons were obtained by RT-PCR analysis using primers indicated in *C*. Sequencing indicates *TAS1A-3* was not amplified. *Actin7* was used as the loading control.

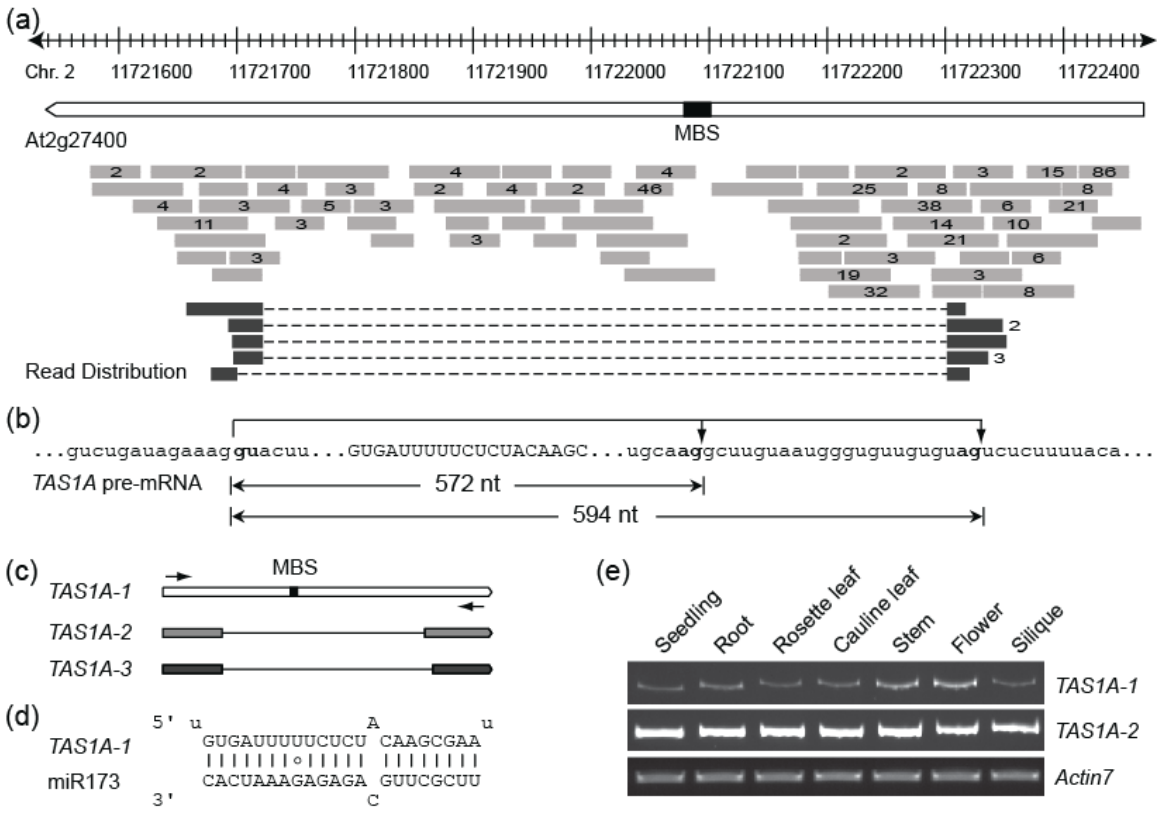

(a)
Chr. 2
11721600
11721700
11721800
11721900
11722000
11722100
11722200
11722300
11722400
At2g27400
MBS
Read Distribution
(b)
...gucugauagaaagguacuu...GUGAUUUUUCUCUACAAGC...ugcaaggcuuguaauggguguugugguagucucuuuuaca...
TAS1A pre-mRNA
572 nt
594 nt
(c)
MBS
TAS1A-1
TAS1A-2
TAS1A-3
(d)
5' u A u
TAS1A-1 GUGAUUUUUCUCU CAAGCGAA
miR173 CACUAAAGAGAGA GUUCGCUU
3' C
(e)
Seedling
Root
Rosette leaf
Cauline leaf
Stem
Flower
Silique
TAS1A-1
TAS1A-2
Actin7

Figure 4. Frequency of alternative splicing increases near MBS

(a) Simulation of the number of alternative splicing events impacting random 21-nt mRNA segments. Grey bars represent the histogram of the simulated data that consists of 1,000 subsets of 354 randomly selected mRNA segments. The X axis shows the number of mRNA segments impacted by alternative splicing and the Y axis shows the frequency of observing such a number. Solid black line is the best fitted curve following a normal distribution. The arrow indicates the observed number of alternatively spliced MBSs, which is greater than the simulated mean by 3.51 times the standard deviation ($P<0.001$). (b) Trend line of the frequency of alternative splicing flanking the MBS. After aligning all MBSs, the frequency of alternative splicing in the surrounding genomic regions is calculated using a sliding window approach. The X axis indicates the position of a given window relative to the MBS whereas the Y axis indicates observed frequency of alternative splicing for that window. (c) Trend line of the frequency of potential splice sites flanking the MBS. After aligning all MBSs, the frequency of GT and AG is respectively calculated for the upstream and downstream region of MBS with varying distance from the MBS. The X axis indicates the distance of the flanking sequences whereas the Y axis indicates observed frequency of the dinucleotides.

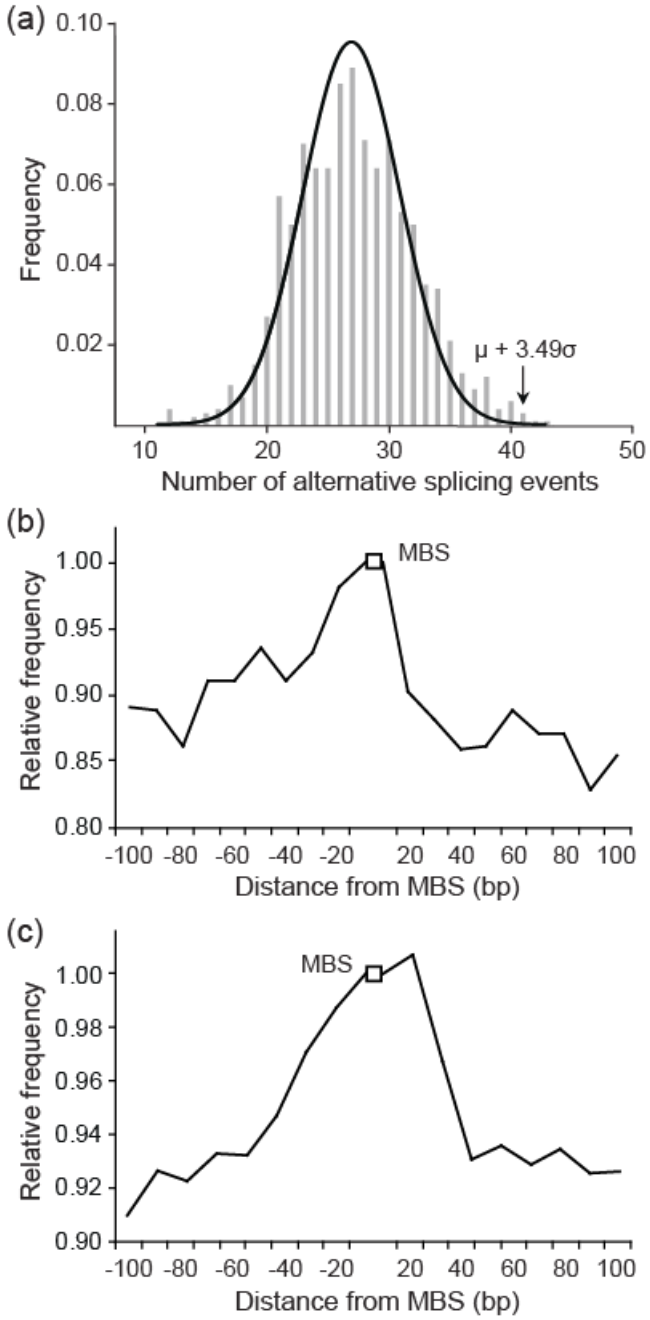

(a)
0.10
0.08
0.06
0.04
0.02
Frequency
μ + 3.49σ
10
20
30
40
50
Number of alternative splicing events
(b)
1.00
0.95
0.90
0.85
0.80
Relative frequency
MBS
-100 -80 -60 -40 -20 20 40 60 80 100
Distance from MBS (bp)
(c)
1.00
0.98
0.96
0.94
0.92
0.90
Relative frequency
MBS
-100 -80 -60 -40 -20 20 40 60 80 100
Distance from MBS (bp)

Figure 5. Alternative splicing of the miR156 binding site influences *SPL4* expression

(a) Proposed splicing model of *SPL4* spanning the MBS, which is shown in upper case letters. The optional splice donor site and the two acceptor sites are indicated by the arrows. (b) Gene structure of the three identified *SPL4* isoforms. Exons and introns are depicted by boxes and horizontal lines, respectively. Untranslated regions, coding regions and the miR156 binding site are shaded as indicated. Positions of the primers used for obtaining amplicons flanking the MBS are indicated by the horizontal arrows. (c) Proposed miRNA:mRNA hybrid between miR156 and *SPL4-1*. (d) RT-PCR analysis of the developmental expression profiles of *SPL4-1*, *SPL4-2*, and *SPL4-3*. Amplicons were obtained using primers indicated in *B*. *Actin7* was used as the loading control. (e) Northern blot analysis of the expression profiles of miR156. U6snRNA was used as the loading control. (f) RT-PCR analysis of the relative expression levels of *SPL4* isoforms in seedlings and rosette leaf of adult plants. WT, wild type plants. *35S::miR156*, transgenic plants in which expression of miR156 is driven by the constitutive CaMV 35S promoter.

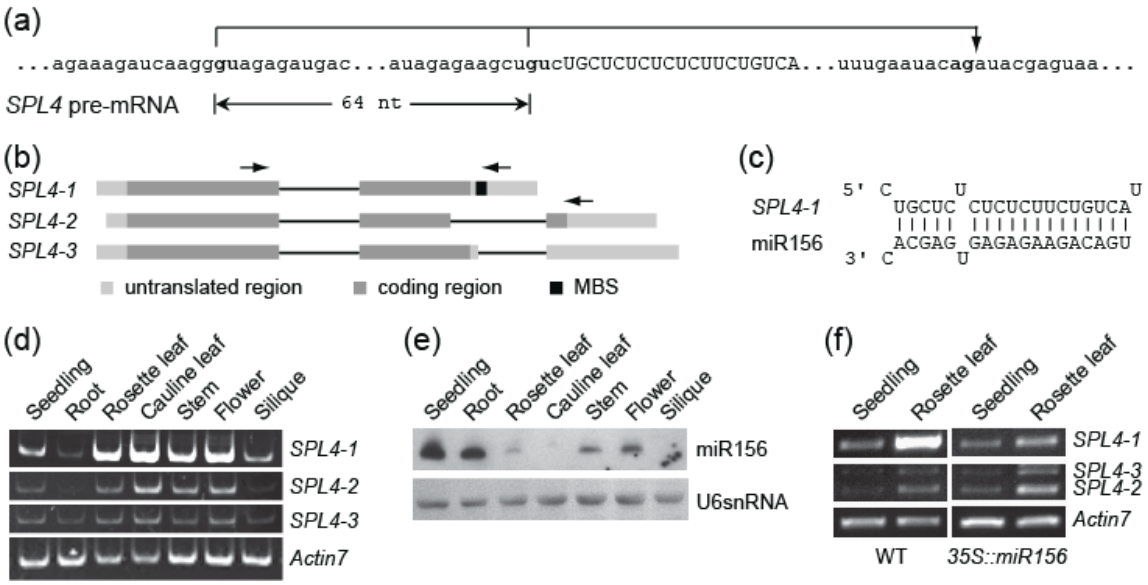

(a)
...agaaagaucaaggguagagaugac...auagagaagcugucUGCUCUCUCUCUUCUGUCA...uuugaauacagauacgaguaa...
SPL4 pre-mRNA
64 nt
(b)
SPL4-1
SPL4-2
SPL4-3
untranslated region
coding region
MBS
(c)
5' C U U
SPL4-1 UGCUC CUCUCUUCUGUCA
miR156 ACGAG GAGAGAAGACAGU
3' C U
(d)
Seedling
Root
Rosette leaf
Cauline leaf
Stem
Flower
Silique
SPL4-1
SPL4-2
SPL4-3
Actin7
(e)
Seedling
Root
Rosette leaf
Cauline leaf
Stem
Flower
Silique
miR156
U6snRNA
(f)
Seedling
Rosette leaf
Seedling
Rosette leaf
SPL4-1
SPL4-3
SPL4-2
Actin7
WT
35S::miR156

Figure 6. Functional analysis of the alternatively spliced miR156 binding site in *SPL4*

(a) RT-PCR analysis of the relative transcript abundance of the three *SPL4* isoforms in WT and transgenic plants over-expressing individual *SPL4* isoforms. (b) Morphology of WT and the transgenic plants at the bolting stage. (c) Flowering time (days after planting) and (d) the number of leaves with abaxial trichome at bolting in WT and transgenic plants expressing *SPL4* with (*35S::SPL4-1*) or without the miR156 binding site (*35S::SPL4-2* and *35S::SPL4-3*). Error bars represent standard deviation (n=12).

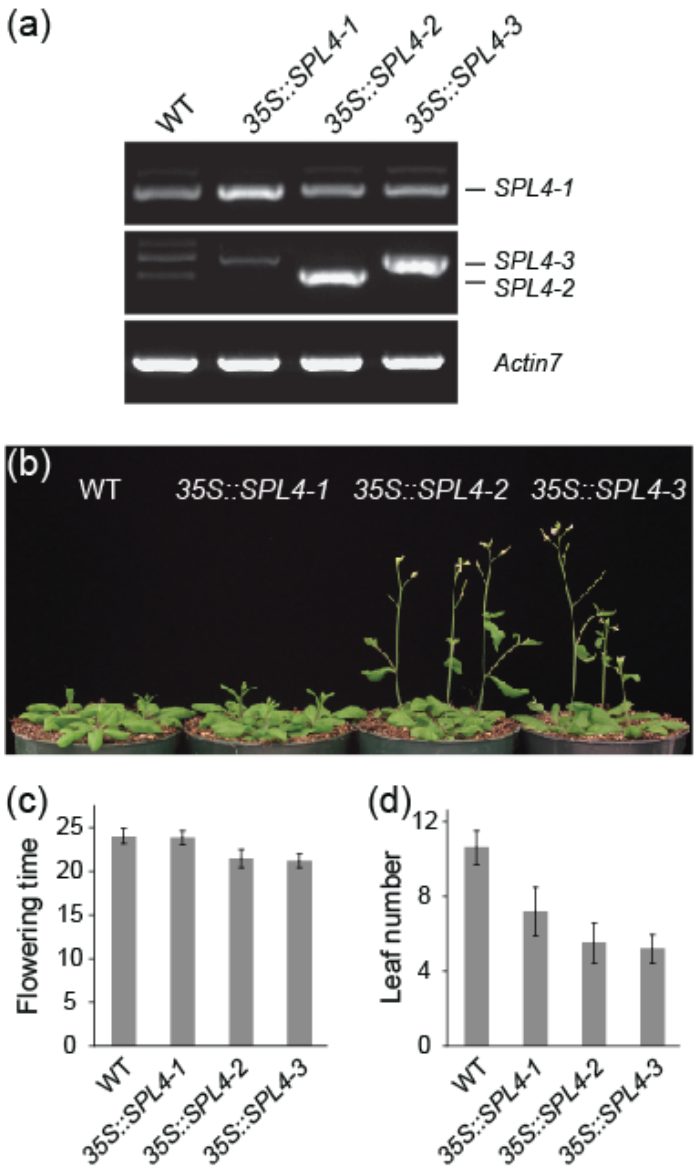

(a)
WT
35S::SPL4-1
35S::SPL4-2
35S::SPL4-3
SPL4-1
SPL4-3
SPL4-2
Actin7
(b)
WT
35S::SPL4-1
35S::SPL4-2
35S::SPL4-3
(c)
Flowering time
25
20
15
10
5
0
WT
35S::SPL4-1
35S::SPL4-2
35S::SPL4-3
(d)
Leaf number
12
8
4
0
WT
35S::SPL4-1
35S::SPL4-2
35S::SPL4-3

Figure S1: Alternative splicing of the miR173 binding site in *TAS2*

(a) RNA-Seq reads mapped to the genomic region surrounding the MBS. Chromosome coordinates are shown on top. Reads mapped to a single locus and those split to two loci are depicted as grey and black horizontal bars, respectively. Copy number of reads retrieved more than once is indicated. (b) Proposed splicing model of *TAS2*. MBS is shown in upper case letters. The optional splice donor site and the two acceptor sites are indicated by the arrows. (c) Gene structure of the three identified *TAS2* isoforms. Exons and introns are depicted as boxes and horizontal lines, respectively. Positions of the primers used for obtaining amplicons flanking the MBS are indicated by the horizontal arrows. (d) Predicted RNA duplex of miR173 and *TAS2-1*.Base paring is indicated by a short vertical line and G:U wobble pairing by the “○” sign. (e) Developmental expression profiles of *TAS2-1* and *TAS2-2*. *TAS2* amplicons are obtained by RT-PCR analysis using primers indicated in (c). Sequencing indicates *TAS2-3* is not amplified. *Actin7* is used as the loading control.

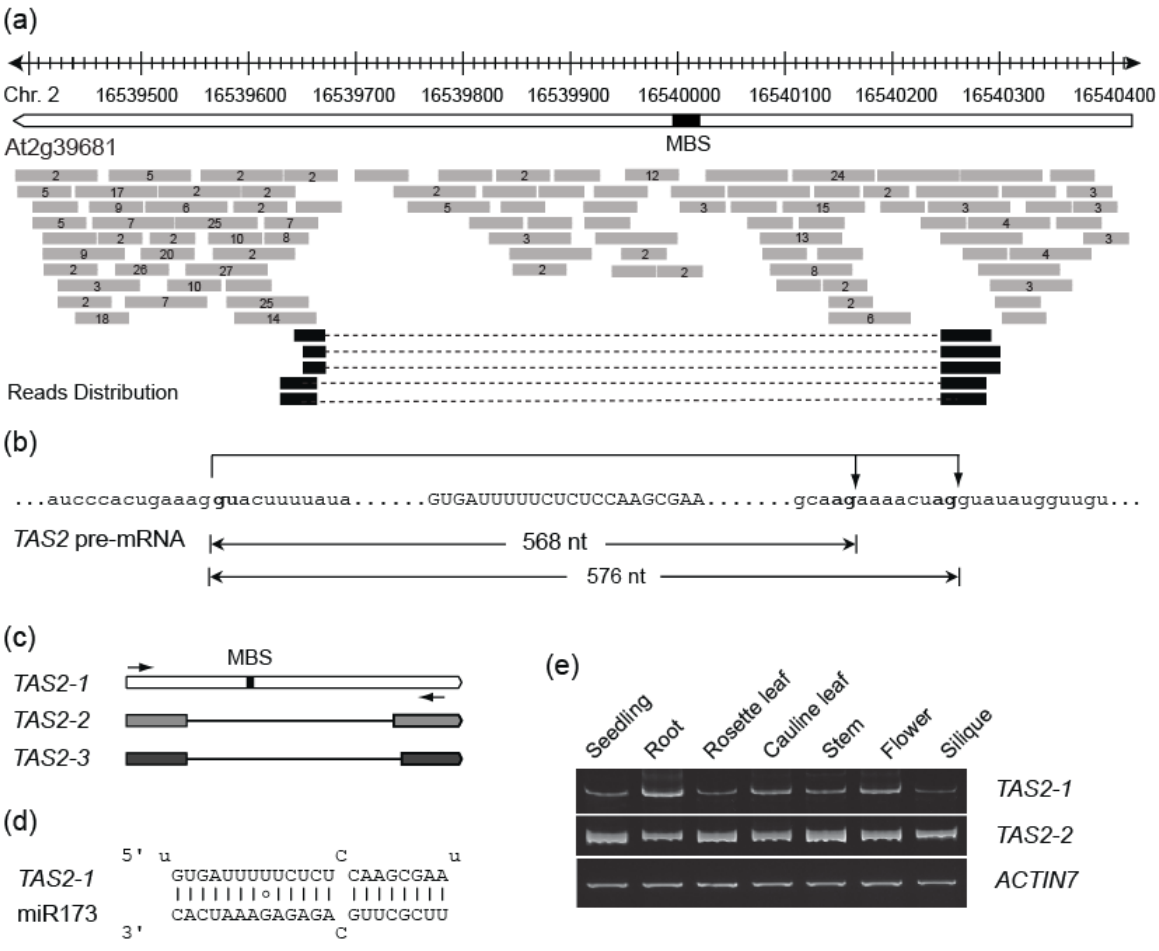

(a)
Chr. 2
16539500
16539600
16539700
16539800
16539900
16540000
16540100
16540200
16540300
16540400
At2g39681
MBS
Reads Distribution
(b)
...aucccacugaaagguacuuuuaua.....GUGAUUUUUCUCUCCAAGCGAA.....gcaagaaaacuagguauaugguugu...
TAS2 pre-mRNA
568 nt
576 nt
(c)
MBS
TAS2-1
TAS2-2
TAS2-3
(d)
5' u C u
TAS2-1 GUGAUUUUUCUCU CAAGCGAA
miR173 CACUAAAGAGAGA GUUCGCUU
3' C
(e)
Seedling
Root
Rosette leaf
Cauline leaf
Stem
Flower
Silique
TAS2-1
TAS2-2
ACTIN7

Figure S2: MBSs impacted by potential intron retention events

(a) RNA-seq reads, depicted as horizontal black bars, mapped to the At5g60690 locus unambiguously support inclusion of the intron 4 sequence in the mRNA. Retention of this intron will disrupt the MBS for miR156/miR157 that splits into exon 4 and exon 5 in the other alternatively spliced mRNA isoform. We identified three cases like this for the 354 examined MBSs. (b) Reads belong to the At3g09220 locus are mapped to intron 1 as well as the boundaries between intron 1 and two neighboring exons, which suggests that at least parts of intron 1 might be present in an alternatively spliced mRNA isoforms. Yet the lack of complete coverage makes it difficult to assess the impact of alternative splicing on the MBS for miR857. We found at least three similar cases for the 354 examined MBSs.

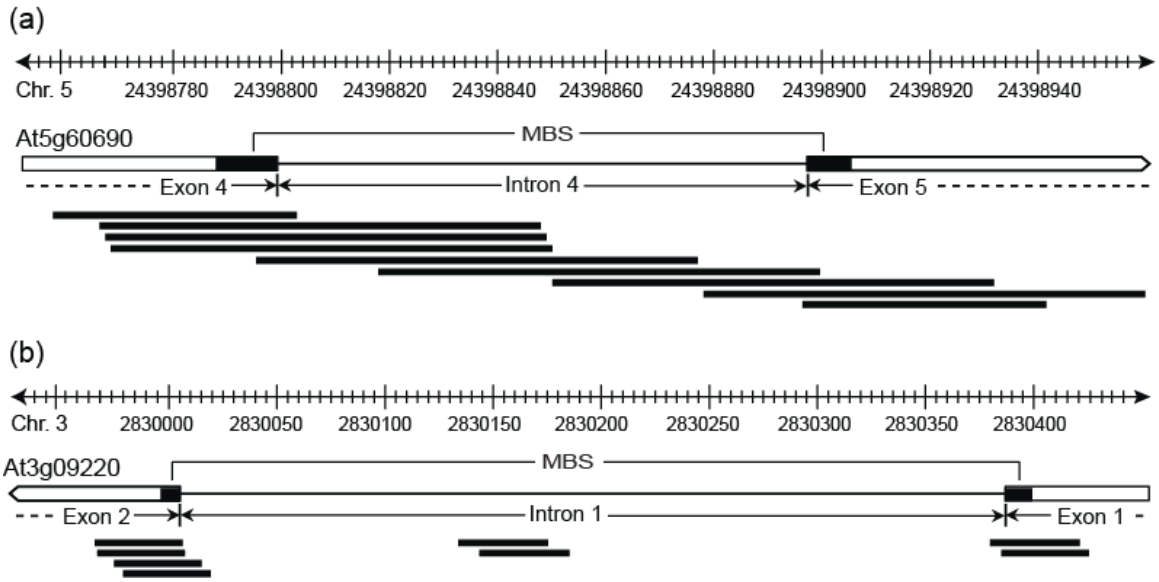

(a)
Chr. 5
24398780
24398800
24398820
24398840
24398860
24398880
24398900
24398920
24398940
At5g60690
MBS
Exon 4
Intron 4
Exon 5
(b)
Chr. 3
2830000
2830050
2830100
2830150
2830200
2830250
2830300
2830350
2830400
At3g09220
MBS
Exon 2
Intron 1
Exon 1

Figure S3: Cross-species sequence comparison of *DCL1* splicing sites at the MBS

(a) Alignment of part of exon13. Invariant nucleotides are shaded in gray and invariant amino acids are indicated below the alignment. The cryptic acceptor site in *Arabidopsis* is indicated by the arrow. Conserved sites in other species are highlighted. (b) Alignment of the junction region of exon12 and intron 12. Invariant nucleotides are shaded in gray. The wobble donor site in *Arabidopsis* is indicated by the arrow. Conserved sites in other species are highlighted.

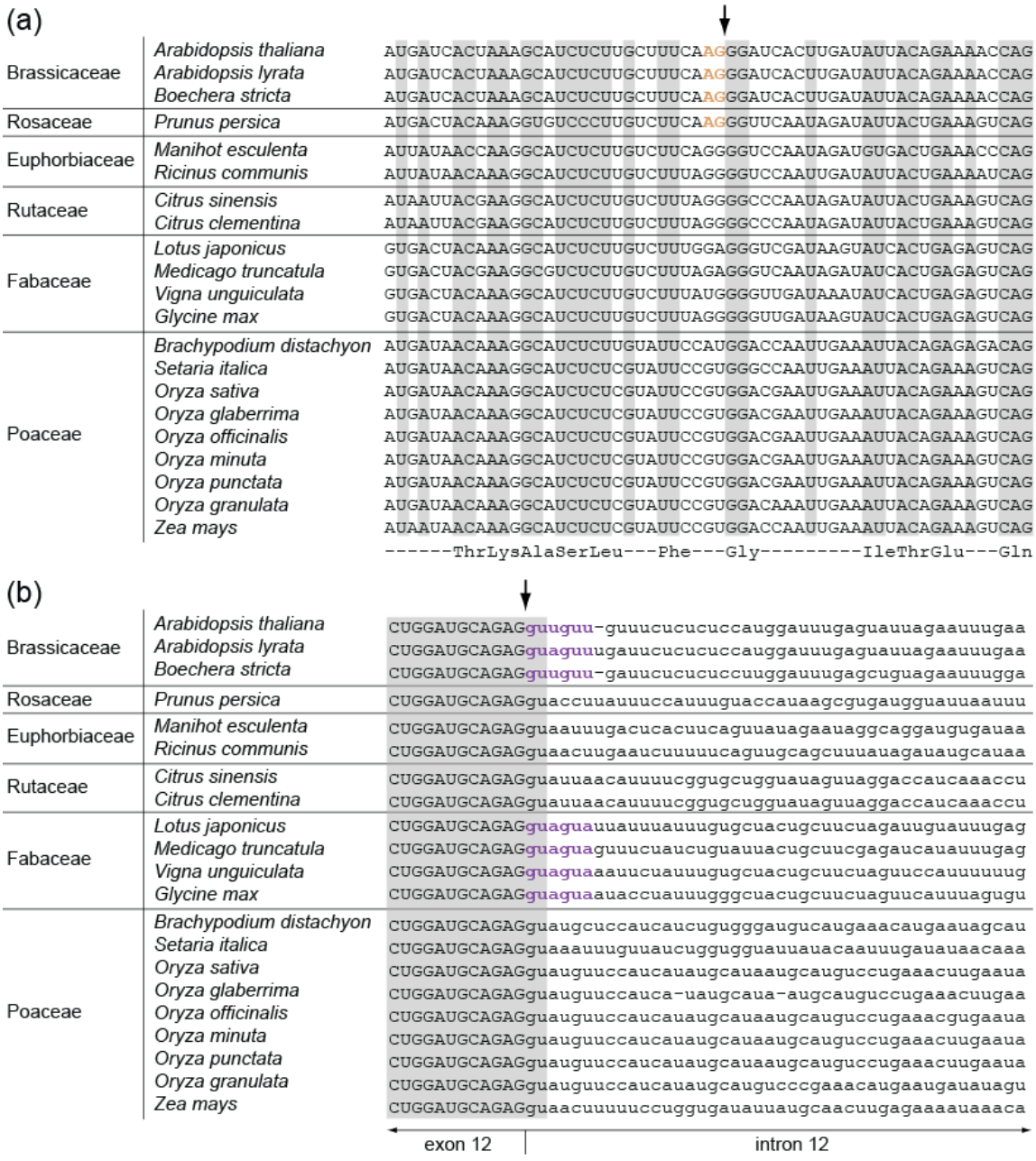

(a)
Brassicaceae
Arabidopsis thaliana
Arabidopsis lyrata
Boechera stricta
Rosaceae
Prunus persica
Euphorbiaceae
Manihot esculenta
Ricinus communis
Rutaceae
Citrus sinensis
Citrus clementina
Fabaceae
Lotus japonicus
Medicago truncatula
Vigna unguiculata
Glycine max
Poaceae
Brachypodium distachyon
Setaria italica
Oryza sativa
Oryza glaberrima
Oryza officinalis
Oryza minuta
Oryza punctata
Oryza granulata
Zea mays
------ThrLysAlaSerLeu---Phe---Gly---------IleThrGlu---Gln
(b)
exon 12
intron 12

Figure S4: Alternatively spliced MBS regulates the expression of *SPL4*

(a) Clustering of several SPL4-related genes in *Arabidopsis* by Maximum-likelihood phylogenetic analysis reconstructed using Molphy. Scale bar corresponds to 0.1 amino acid substitution per residue. (b) Comparison of the gene structure of *SPL4* and *SPL3*. Exons and introns are depicted by boxes and horizontal lines, respectively. The coding regions and the miR156 binding site are shaded in grey and black, respectively. (c) Clock-related *SPL4* and *SPL3* expression pattern reconstructed from the microarray data reported in Balasubramanian et al. (PloS Genet 2, 981). Only *SPL3* exhibits moderate rhythmic expression in Day 5 and Day 9.

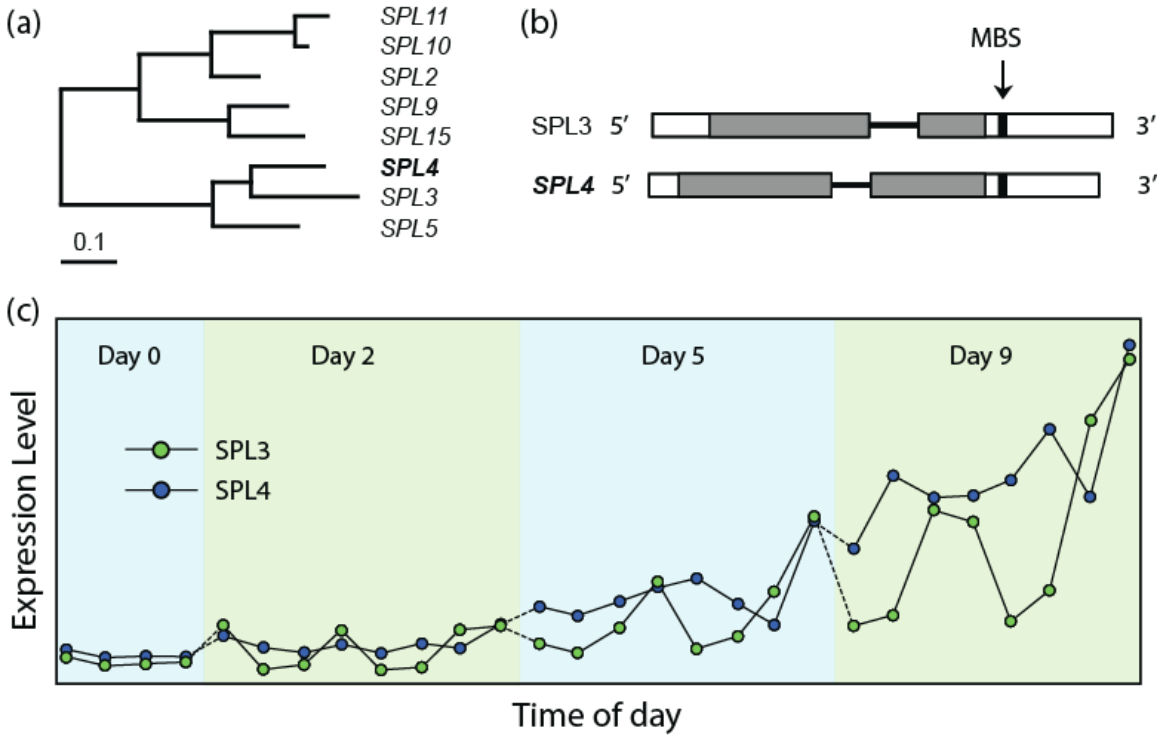

(a)
SPL11
SPL10
SPL2
SPL9
SPL15
SPL4
SPL3
SPL5
0.1
(b)
MBS
SPL3 5′
3′
SPL4 5′
3′
(c)
Day 0
Day 2
Day 5
Day 9
SPL3
SPL4
Expression Level
Time of day

Table S1. MBSs and alternatively spliced MBSs used in this study

miRNA	Target	Evidence for MBS [a]	Evidence for splicing [b]	Model for splicing [c]	Target Function
miR156/miR157	AT1G27360	II			SPL11
	AT1G69170	II			SPL like
	AT2G42200	II			SPL9
	AT3G57920	II			SPL15
	AT5G50570	II			SPL13A
	AT5G50670	II			SPL13B
	AT1G27370	I			SPL10
	AT1G53160	I	I	II	SPL4
	AT2G33810	I			SPL3
	AT3G15270	I			SPL5
	AT5G43270	I			SPL2
	AT5G08620	II			STRS2
miR158	AT1G64100	II			PPR protein
	AT2G03210	II			FUT2
	AT2G03220	II			FT1
	AT3G03580	II			TPR like
miR159/miR319	AT2G26950	II			AtMYB104
	AT2G26960	II			AtMYB81
	AT2G32460	II			MYB101I
	AT2G34010	II			unknown protein
	AT4G26930	II			MYB97
	AT5G55020	II	II	II	ATMYB120
	AT1G30210	I			TCP24
	AT1G53230	I			TCP3
	AT2G31070	I			TCP10
	AT3G11440	I			ATMYB65
	AT3G15030	I			TCP4
	AT4G18390	I			TCP2
	AT5G06100	I			MYB33
miR160	AT1G77850	I			ARF17
	AT2G28350	I			ARF10
	AT4G30080	I			ARF16
miR161	AT1G06580	I			PPR protein
	AT1G62910	I			PPR protein
	AT1G63080	I			PPR protein
	AT1G63130	I			TPR-like
	AT1G63150	I			TPR-like
	AT1G63230	II	II	II	TPR-like
	AT1G63400	I			PPR protein
	AT1G63630	II	I	II	TPR-like

	AT5G41170	I			PPR-like
miR162	AT1G01040	I	I	I/II	DCL1
miR163	AT1G66690	I			methyltransferases
	AT1G66700	I			PXMT1
	AT1G66720	I			methyltransferases
	AT3G44860	I			FAMT
	AT3G44870	II			methyltransferases
miR164	AT1G56010	I			NAC1
	AT3G15170	I			CUC1
	AT5G07680	I			ANAC080
	AT5G39610	II			ATNAC2
	AT5G53950	I			CUC2
	AT5G61430	I			ANAC100
miR165/miR166	AT4G32880	II			ATHB-8
	AT1G30490	I	I	II	PHV
	AT1G52150	I			ATHB-15
	AT2G34710	I			PHB
	AT5G60690	I	I	II	REV
miR167	AT1G30330	I			ARF6
	AT5G37020	I			ARF8
miR168	AT1G48410	I			AGO1
miR169	AT1G17590	I			NF-YA8
	AT1G54160	I			NFYA5
	AT1G70700	II			TIFY7
	AT1G72830	I			HAP2C
	AT3G05690	I			UNE8
	AT3G20910	I			NF-YA9
	AT5G06510	I			NF-YA10
	AT5G12840	II			HAP2A
miR170/miR171	AT2G45160	I			HAM1
	AT3G60630	I			HAM2
	AT4G00150	I			HAM3
miR172	AT2G28550	I			RAP2.7
	AT2G39250	II			SNZ
	AT3G14770	II	II	I	SWEET2
	AT3G54990	II	I	III	SMZ
	AT4G36920	I			AP2
	AT5G60120	I			TOE2
	AT5G67180	I			TOE3
miR173	AT1G50055	I			TAS1B
	AT2G27400	I	II	II	TAS1A
	AT2G39675	I			TAS1C
	AT2G39681	I	II	II	TAS2
miR390/miR391	AT3G17185	I			TASIR-ARF

	AT3G17185	I			TASIR-ARF
	AT5G49615	I			TAS3b
	AT5G49615	I			TAS3b
	AT5G57735	I			TASIR-ARF
	AT5G57735	I			TASIR-ARF
miR393	AT1G12820	I			AFB3
	AT3G23690	I			bHLH
	AT3G26810	I			AFB2
	AT3G62980	I			TIR1
	AT4G03190	I			GRH1
miR394	AT1G27340	I			Galactose oxidase
miR395	AT3G22890	I			APS1
	AT4G14680	II			APS3
	AT5G10180	I			AST68
	AT5G13630	II			GUN5
	AT5G43780	I			APS4
miR396	AT1G63000	II			NRS/ER
	AT2G22840	I			AtGRF1
	AT2G36400	I			AtGRF3
	AT2G40760	II			phosphatase
	AT2G45480	I			AtGRF9
	AT3G52910	II			AtGRF4
	AT4G24150	I			AtGRF8
	AT4G37740	I			AtGRF2
	AT5G24660	II			LSU2
	AT5G53660	I			AtGRF7
miR397	AT2G29130	I			LAC2
	AT2G38080	I			IRX12
	AT3G60250	II			CKB3
	AT5G12250	II			TUB6
	AT5G60020	I			LAC17
miR398	AT1G08830	I			CSD1
	AT2G28190	I			CSD2
	AT3G15640	I			Rubredoxin-like
miR399	AT2G33770	I			UBC24
	AT2G33770	I			UBC24
	AT3G54700	II			PHT1
	AT5G43280	II			ATDCI1
miR400	AT1G06580	II	II	II	PPR
	AT1G62670	II			RPF2
	AT1G62720	II			PPR-like
	AT1G62910	II			PPR
	AT1G62930	II			TPR-like
	AT1G63130	II			TPR-like

	AT1G63150	II			TPR-like
	AT1G63400	II			PPR
	AT4G19440	II			TPR-like
miR401	AT2G06095	II			unknown protein
miR402	AT4G34060	I			DML3
miR403	AT1G31280	I			AGO2
	AT4G21510	II			F-box protein
miR404	AT1G07650	II	I	III	Leucine-rich kinase
miR408	AT2G02850	II			ARPN
	AT2G30210	I			LAC3
	AT2G47020	II			Peptide CR factor 1
miR413	AT5G02830	II			TPR-like
miR414	AT1G06950	II			ATTIC110
	AT1G07990	II			SIT4 phosphatase-associated
	AT1G13220	II	I	III	LINC2
	AT1G14460	II			ATPase protein
	AT1G15280	II			eIF4AIII binding
	AT1G21560	II			unknown protein
	AT1G25370	II			unknown protein
	AT1G28670	II			ARAB-1
	AT1G32490	II			ESP3
	AT1G47970	II	II	I/II	unknown protein
	AT1G50410	II			SNF2 domain-containing
	AT1G54440	II			transferase
	AT1G55440	II			Cysteine/Histidine-rich
	AT1G58210	II			EMB1674
	AT1G60220	II			OTS1
	AT1G60870	II			MEE9
	AT1G63800	II			UBC5
	AT1G63980	II			D111/G-patch containing
	AT1G64880	II	II	II	Ribosomal protein
	AT1G66620	II			TRAF-like
	AT1G68440	II			unknown protein
	AT1G69070	II	II	II	unknown protein
	AT1G75150	II			unknown protein
	AT1G75500	II	I	II	WAT1
	AT1G80000	II			eIF4AIII binding
	AT2G05110	II			TE gene
	AT2G11910	II	II	II	unknown protein
	AT2G16600	II			ROC3
	AT2G16900	II			phospholipase-like
	AT2G18220	II			Noc2p family
	AT2G19480	II	II	I	NAP1
	AT2G20280	II			Zinc finger protein

	AT2G22080	II	II	I/II	unknown protein
	AT2G24500	II	II	I	FZF
	AT2G24550	II			unknown protein
	AT2G27990	II			BLH8
	AT2G31660	II			SAD2
	AT2G33250	II			unknown protein
	AT2G33400	II			unknown protein
	AT2G33430	II			DAL1
	AT2G39020	II			NAT protein
	AT2G42040	II			unknown protein
	AT3G06130	II			Heavy metal transport
	AT3G11100	II			DNA binding TF
	AT3G11200	II			AL2
	AT3G11810	II	II	II	unknown protein
	AT3G14900	II	II	II	unknown protein
	AT3G17160	II	II	I	unknown protein
	AT3G18100	II			MYB4R1
	AT3G18390	II			EMB1865
	AT3G18520	II			HDA15
	AT3G19910	II			RING/U-box protein
	AT3G21060	II			Transducin/WD40 protein
	AT3G24490	II	II	II	dehydrogenase TF
	AT3G24760	II			oxidase/kelch protein
	AT3G28030	II			UVH3
	AT3G28850	II			Glutaredoxin protein
	AT3G29075	II			glycine-rich protein
	AT3G42790	II			AL3
	AT3G48430	II			REF6
	AT3G49140	II	II	IV	PPR protein
	AT3G49470	II			NACA2
	AT3G50550	II			unknown protein
	AT3G52120	II			SWAP containing protein
	AT3G56150	II			EIF3C
	AT3G58110	II			unknown protein
	AT4G00270	II			DNA-binding TF
	AT4G00830	II			RNA-binding protein
	AT4G02220	II			zinc finger protein
	AT4G02460	II			PMS1
	AT4G04614	II			unknown protein
	AT4G17940	II	II	I	TPR-like
	AT4G19100	II			PAM68
	AT4G22350	II			hydrolases protein
	AT4G23040	II			Ubiquitin-like
	AT4G26110	II			NAP1

	AT4G26600	II			methyltransferases
	AT4G27310	II			zinc finger protein
	AT4G27320	II			ATPHOS34
	AT4G27500	II	II	I	PPI1
	AT4G29520	II			unknown protein
	AT4G30720	II			FAD/NAD(P)-binding
	AT4G31610	II			REM1
	AT4G32610	II			copper ion binding
	AT4G33240	II			FAB1A
	AT4G37210	II	II	IV	TPR-like
	AT4G38480	II			Transducin/WD40
	AT5G01260	II			Carbohydrate-binding-like
	AT5G03545	II			AT4
	AT5G04810	II			PPR containing
	AT5G05550	II			DNA binding TF
	AT5G06770	II			KH domain-containing
	AT5G09860	II			THO1
	AT5G10070	II			RNase L inhibitor related
	AT5G11270	II	II	II	OCP3
	AT5G11600	II			unknown protein
	AT5G12110	II			Glutathione S-transferase
	AT5G14050	II			Transducin/WD40 like
	AT5G15140	II			mutarotase-like
	AT5G15800	II	I	I	SEP1
	AT5G18570	II			EMB269
	AT5G22010	II			AtRFC1
	AT5G22320	II			LRR protein
	AT5G22640	II			emb1211
	AT5G23420	II	II	I	HMGB6
	AT5G23950	II			lipid-binding
	AT5G24500	II			unknown protein
	AT5G37780	II			CAM1
	AT5G39740	II			RPL5B
	AT5G40340	II	II	II	Tudor/PWWP/MBT
	AT5G40690	II			EF-Hand 1
	AT5G41340	II			UBC4
	AT5G42280	II	II	I	Cysteine/Histidine-rich
	AT5G44200	II			CBP20
	AT5G46030	II			unknown protein
	AT5G47040	II	II	I	LON2
	AT5G49400	II			CCHC-type
	AT5G53000	II			TAP46
	AT5G53220	II			unknown protein
	AT5G53440	II			unknown protein

	AT5G55920	II			OLI2
	AT5G58000	II			Reticulon protein
	AT5G62260	II			DNA-binding
	AT5G64420	II			DNA polymerase
	AT5G64920	II			CIP8
miR419	AT2G36290	II			alpha/beta-Hydrolases
miR447	AT5G60760	I			P-loop containing
miR472	AT1G12290	II			Disease resistance
	AT1G51480	I			Disease resistance
	AT4G14610	II			pseudogene
	AT5G43740	I			Disease resistance
miR773	AT4G08990	II			methyltransferase
	AT4G14140	I			DMT2
miR774	AT3G17490	II			F-box associated
	AT3G19890	I			F-box protein
miR775	AT1G53290	I			Galactosyltransferase
miR776	AT5G62310	II			IRE
miR778	AT2G22740	I			SUVH6
	AT2G35160	II			SUVH5
miR780	AT5G41610	I			ATCHX18
miR781	AT1G44900	II			MCM2
	AT5G23480	II			Plus-3;GYF
miR822	AT2G02620	II			Cysteine/Histidine-rich
	AT2G13900	II			Cysteine/Histidine-rich
	AT5G02330	II			Cysteine/Histidine-rich
	AT5G02350	II			Cysteine/Histidine-rich
miR823	AT1G69770	I			CMT3
miR824	AT3G14560	II			unknown protein
	AT3G57230	I			AGL16
miR825	AT2G41870	II			Remorin
miR826	AT4G03060	II			AOP2
miR827	AT1G02860	I			NLA
miR828	AT1G66370	I			MYB113
	AT3G60980	II			TPR-like
	AT5G52600	II			AtMYB82
	TAS4	I			TAS4
miR831	AT3G12190	II			Frigida-like
miR832	AT2G46960	II	I	III	CYP709B1
miR833	AT1G77650	II			F-box associated
miR834	AT3G03870	II			unknown protein
	AT3G10010	II			DML2
	AT3G48050	II			BAH domain
	AT4G00930	II			CIP4.1
	AT5G05840	II			unknown protein

	AT5G24510	II			60S acidic ribosomal
miR835	AT1G49560	II			Homeodomain-like
miR837	AT1G16800	II			P-loop containing
	AT1G23020	II			FRO3
	AT1G52500	II			FPG-2
	AT1G52590	II			oxidoreductase DCC
	AT1G68845	II			unknown protein
	AT4G00850	II			GIF3
	AT5G18590	II			Galactose oxidase
miR838	AT1G12850	II			Phosphoglycerate mutase
	AT1G72360	II			HRE1
	AT3G12380	II			ARP5
	AT5G21930	II			PAA2
miR841	AT1G70470	II			unknown protein
	AT2G38810	II	II	II	HTA8
	AT4G13570	II	I	IV	HTA4
miR842	AT1G52100	II			Mannose-binding
	AT1G52120	II			Mannose-binding
	AT2G37340	II			RSZ33
	AT5G38550	I			Mannose-binding
miR843	AT2G22810	II			ACS4
	AT3G13830	II			F-box domains-containing
miR844	AT5G51270	I			U-box domain-containing
miR846	AT2G25980	II			Mannose-binding lectin
	AT5G49850	I			Mannose-binding lectin
	AT5G49870	II			Mannose-binding lectin
miR847	AT1G49120	II			Integrase-type DNA-binding
	AT1G53340	II			Cysteine/Histidine-rich C1
miR854	AT5G05100	II	II	II	R3H binding
miR856	AT2G46800	II			ZAT
	AT5G41610	I	I	II	ATCHX18
miR857	AT3G09220	I			LAC7
miR858	AT1G06180	II			ATMYB13
	AT2G47460	I			MYB12
	AT3G01720	II			unknown protein
	AT3G08500	I			MYB83
	AT5G35550	II	II	I	TT2
miR859	AT3G17265	II			F-box containing
	AT3G49510	I			F-box
	AT5G36200	II			F-box containing
miR860	AT5G26030	II	I	II	FC1
miR863	AT1G62710	II			BETA-VPE
	AT4G13495	II			other RNA
miR864	AT4G25210	II	II	II	DNA-binding TF

miR865	AT2G07750	II	DEA(D/H)-box RNA helicase
	AT2G34900	II	GTE01
	AT3G52280	II	GTE6
	AT5G42240	II	scpl42
miR866	AT2G41540	II	GPDHC1
	AT4G21400	II	CRK28

[a] Set I MBS are 5' RACE-validated while Set II are identified from high throughput sequencing of the 5' ends of polyadenylated products predicted to be compatible with miRNA-mediated mRNA decay.

[b] Evidence for alternative splicing is either I, annotated gene models, or II, RNA-Seq data.

[c] Models for identifying alternatively spliced MBS. Based on mRNA to genome alignment, MBS are differentially spliced in Model I by use of tandem splice donor or acceptor sites, in Model II by retention or removal of an intron, in Model III by differential initiation or termination, and in Model IV by an exon cassette or mutually excluding exons.

Table S2. RNA-Seq libraires used in this study

Tissue Type	Library ID	Reads Length	Reads [a]	Mapped reads [b]
Whole Seedling	GSM613465	41	21142970	19017098
	GSM613466	41	25981473	23643374
Immature Flower	GSM284751	50	24512218	6667118
Embryo	GSM623879	36	21195414	2598224
	GSM623880	36	22327176	1883869
	GSM607726	76	54655897	37222697
	GSM607724	76	34520973	18122968
Endosperm	GSM607723	76	111006572	79132506
	GSM607725	76	29996870	16609679
	GSM607727	76	57340220	39556103
	GSM607728	76	87776825	59870713
	GSM607729	76	81239523	55242355
Total			571696131	359566704

[a] Number of trimmed and filtered reads in each library.

[b] Number of reads mapped to the *Arabidopsis* genome.

Table S3. Primers and probes used in this study

Name	Sequence (5' to 3')	Purpose
a-miR156	GTGCTCACTCTCTTCTGTCA	Hybridization to miR156
U6	CTCGATTTATGCGTGTCATCCTTGC	Hybridization to U6snRNA
DCL1-F	GGCAATGAGCTGGATGCAGA	Forward, *DCL1* cDNA cloning
DCL1-R	CAGGTAGGCCTTTGCAGGAT	Reverse, *DCL1* cDNA cloning
SPL4-F	GGCTCGAGTTTCCTTCCTCCAACAA	Forward, *SPL4* cDNA cloning
SPL4-R1	GCACTAGTATAGTCTTAGCGTTTGCA	Reverse, *SPL4-1*cDNA cloning
SPL4-R2	CGACTAGTGATGGACCCTGAAAGGT	Reverse, *SPL4-2* and *SPL4-3* cloning
SPL4-Q5	CTCAGGACTTAACCAACGCTT	Forward, *SPL4* quantification
SPL4-Q3S	GAGAGCAGACAGCTTCTCTAT	Reverse, *SPL4-1* quantification
SPL4-Q3L	CTCCTTCGTGGCTCTGAAACT	Reverse, *SPL4-2* and *SPL4-3* quantification
TAS1AF	GGCTAAGCCTGACGTCATAT	Forward, *TAS1A* cDNA cloning
TAS1AR	GCTGACTTGCTTCATGTAGA	Reverse, *TAS1A* cDNA cloning
TAS2F	CCAAGCTCTGCAAAGAGAT	Forward, *TAS2* cDNA cloning
TAS2R	CGCTCTTTAATGTGCTTCAC	Reverse, *TAS2* cDNA cloning
Actin7F	GGTGTCATGGTTGGTATGGGTC	Forward, *Actin7* cDNA cloning
Actin7R	CCTCTGTGAGTAGAACTGGGTGC	Reverse, *Actin7* cDNA cloning

Appendix 1. A user's manual for miRDeep-P

1. OVERVIEW

1.1 BACKGROUND

MiRNAs are an important class of endogenous small RNAs that regulate gene expression at the post-transcription level (Bartel, 2009). There has been a surge of interest in the past decade in identifying miRNAs and profiling their expression pattern using various experimental approaches (Wark et al., 2008). Most recently, deep sequencing of specifically prepared low-molecular weight RNA libraries has been used for both purposes in diverse plant species (Fahlgren et al., 2007; Zhu et al., 2008). A major drawback of these efforts is the exclusive focus on mature miRNA, the final gene product, and ignorance of sequence information associated with other parts of the miRNA genes. New strategies and/or tools are thus highly desirable to analyze the increasingly available sequencing data to gain insights into the miRNA transcriptomes. The development of miRDeep-P, miRDP for short, was motivated by this need.

1.2 SUMMARY OF *MIRDP* FUNCTION

Based on ultra deep sampling of small RNA libraries by next generation sequencing, miRDP has two main purposes. First, miRDP can be used to identify miRNA genes in plant species, even for those without detailed annotation. Second, miRDP is designed to assign expression status to individual miRNA genes, which is critical as more miRNAs in plants belong to paralogous families with multiple members encoding identical or near-identical miRNAs.

1.3 IMPLEMENTATION AND ALGORITHM

MiRDP is documented by Perl (Perl 5.8 or later versions) and makes use of fundamental packages from Perl library. All the scripts have been tested on two Linux platforms, SUSE 10 and Fedora 14, and should work on similar systems that support Perl.
The core algorithm of miRDP was developed by modifying miRDeep (Friedlander et al., 2008), which is based on a probabilistic model of miRNA biogenesis in animals, with a plant-specific scoring system and filtering criteria.

1.4 LICENSE AND AVAILABILITY

MiRDP is freely available under a GNU Public License (Version 3) at:
http://faculty.virginia.edu/lilab/miRDP/index.html
and http://sourceforge.net/projects/mirdp/
The miRDeep-P scripts, demos and user manual can be obtained from both web sites.

1.5 CONVENTIONS AND RECOMMENDATIONS

All command lines, filenames and directory names are in *italic.* The command lines are separated by a blank line. The line starting with # is an interpretation of the following command line.
Two attached demo packages, *Known_M* (to explore expression patterns of known miRNA genes in *Arabidopsis* (Yang et al., 2011)) and *New_M* (to detect new miRNA genes in *Arabidopsis*), which include the files by which users can reproduce every step and gain familiarity with miRDP. Please note that some of the intermediate files of the *New_M* (e.g. the bowtie aligned result) are not included due to the size of these files. The users are recommended to generate these files by themselves.

2. INSTALLATION

Several dependencies are required to run miRDP. First, the Bowtie package can be downloaded from the site: http://bowtie-bio.sourceforge.net/index.shtml. Second, the Vienna package should be downloaded from the site: http://www.tbi.univie.ac.at/~ivo/RNA/. Third, you should set paths to include the location of the downloaded miRDP scripts, as well as the directories where you have the bowtie, and Vienna executables. This is done by adding the lines:

export PATH=~/location:$PATH

to the *.bashrc* or equivalent file ('location' designates the desired location in the file system).

3. PREPROCESSING THE READS

Before reads are mapped to the genome, they must be preprocessed. First, the deep sequencing reads should have the adapters removed from 5' and 3' ends (if present). Second, deep sequencing reads shorter than 15 nt should be discarded, since they will otherwise flood the mapping output. Third, the deep sequencing reads must be parsed into FASTA format. Forth, redundancy should be removed such that reads with identical sequence are represented with a single FASTA entry. Therefore, each sequence identifier must end with a '_x' and an integer, with the integer indicating the number of times the exact sequence was retrieved in the deep sequencing dataset. Finally, all of the FASTA ids should be unique. One way to ensure this is to include a running number in the id. For reference, see the file, *AtShoot.fa*, in the *Known_M* package. The following are several examples:

>ShootPi1000000_x3
AGAGAGTTCTTACATAAGATCTA
>ShootPi1000001_x1
TCCTTGATTTGATGCAACTAAAT
>ShootPi1000002_x3
TCTTGATTCTATGGGTGGTGGTG
>ShootPi1000003_x1
TAAGATCTCAAAGGAATTAGCAT

4. RETRIEVING EXPRESSION PATTERNS OF KNOWN MIRNA GENES AND OBTAINING THE OPTIMAL LENGTH OF PRECURSORS

One main function of miRDP is designed to assign expression status to individual miRNA genes. The following steps show how to get the expression patterns of known miRNA genes in *Arabidopsis*.

First, the annotated precursors of miRNAs are extended based on genomic sequences and general feature information by the script, *fetch_extended_precursors.pl*.

fetch_extended_precursors.pl TAIR9_genome.fa ath.gff >annotated_miRNA_extended.fa

For *Arabidopsis*, we add 500 nt at both sides of annotated precursors of miRNAs. Note that the sequence ids in *ath-miR.gff* must be identical to those in *TAIR9_genome.fa*.

ath.gff downloaded from miRBase (release15) in the GFF (General Feature Format) format, including annotated miRNA general information.
Second, retrieve annotated miRNA precursors at a specific length by the following commands.
#indexing *annotated_miRNA_extended.fa* for a bowtie mapping search (please check bowtie manual for how to use bowtie-build).

mkdir bowtie-index

bowtie-build –f annotated_miRNA_extended.fa bowtie-index/annotated_miRNA_extended

#mapping reads to *annotated_miRNA_extended.fa* (only keep the perfect alignments (full length, 100% identity) and all the valid alignments. Please check bowtie manual for how to use bowtie.

bowtie –a –v 0 bowtie-index/annotated_miRNA_extended –f AtShoot.fa >AtShoot_extende.aln

#converting bowtie format to blast parsed format (a miRDeep format).

convert_bowtie_to_blast.pl AtShoot_extended.aln AtShoot.fa annotated_miRNA_extended.fa > AtShoot_extended.bst

#excising candidate precursors at a specific length (the third option of the script, *excise_candidate.pl*, is the length of candidate precursors, here we choose 250 as an example)

excise_candidate.pl annotated_miRNA_extended.fa AtShoot_extended.bst 250 >precursors_250.fa

#indexing the candidate precursors and predicting their secondary structure (the secondary structures of the potential precursors are predicted using RNAfold and -noPS means that no graphical output is produced).

bowtie-build –f precursors_250.fa bowtie-index/precursors_250

cat precursors_250.fa|RNAfold –noPS >precursors_250_structure

#aligning reads to candidate precursors and achieving signatures.

bowtie –a –v 0 bowtie-index/precursors_250 –f AtShoot.fa >AtShoot_250.aln

convert_bowtie_to_blast.pl AtShoot_250.aln AtShoot.fa precursors_250.fa >AtShoot_250.bst

note that it is necessary for the prediction to do this sorting.

sort +3 -25 AtShoot_250.bst >AtShoot_250_signature

#retrieving annotated miRNAs

miRDP.pl AtShoot_250_signature precursors_250_structure >AtShoot_250_prediction

At different lengths of candidate precursors, the retrieved number of miRNAs is different. In *Arabidopsis*, our result shows at 250 bp most miRNA genes could be retrieved. Thus, 250 is the optimal length for detecting miRNAs in *Arabidopsis* (Yang et al., 2011). As an option, different lengths, e.g. 100, 150, 200, 250, 300, 350 and 400 bp, can be tested for retrieval of annotated miRNAs in other plant species with the same commands as above. Meanwhile, if multiple libraries prepared from different biological samples are employed, expression profiling of individual miRNA genes can be achieved.

5. DETECTING NEW *MIRNA*S

5-1. ALIGNING THE READS TO THE REFERENCE GENOME

This step is to align the deep sequencing reads to the genomic (or transcriptomic if preferred) sequences. It is computationally demanding when the reference genome or small RNA library is large. The indexed TAIR genomic file can be downloaded from the bowtie website (http://bowtie-bio.sourceforge.net/index.shtml) or indexed by the users.

bowtie-build –f TAIR9_genome.fa bowtie-index/TAIR9_genome

Then, align the deep sequencing reads to the indexed genome. Only keep the perfect alignments and all the valid alignments.

bowtie –a –v 0 bowtie-index/TAIR9_genome –f AtShoot.fa >AtShoot.aln

After that, convert it into blast format.

convert_bowtie_to_blast.pl AtShoot.aln AtShoot.fa TAIR9_genome.fa >AtShoot.bst

5-2. FILTERING UBIQUITOUS ALIGNMENTS

The reads are now filtered such that only perfect alignments (full length, 100% identity) are retained. Then, filter the reads which are mapped to multiple positions in the genome. For *Arabidopsis*, 15 is set as the cutoff since the largest miRNA family, miR169, has 14 members.

filter_alignments.pl AtShoot.bst -c 15 >AtShoot_filter15.bst

For other species, different cutoffs, based on the known family sizes or other empirical considerations such as genome sizes, might be selected.

5-3. FILTERING ALIGNMENTS BY ANNOTATION

The reads which mapped to exons or other non-coding RNAs, including rRNA, snRNA, snoRNA, and tRNA, are filtered by:

overlap.pl AtShoot_filter15.bst ncRNA_CDS.gff -b >id_overlap_ncRNA_CDS

Only alignments where the read ids are not included in the *id_overlap_ncRNA_CDS* file are retained. –g designates that lines where the query read ids are included in the *id_overlap_ncRNA_CDS* file are discarded:

alignedselected.pl AtShoot_filter15.bst -g id_overlap_ncRNA_CDS >AtShoot_filter15_ncRNA_CDS.bst

Reads can also be filtered such that only reads that have one or more remaining alignments are kept. -b designates that the output should be FASTA entries and not alignments:

filter_alignments.pl AtShoot_filter15_ncRNA_CDS.bst -b AtShoot.fa > AtShoot_filtered.fa

5-4. PREDICTING MICRORNAS

Using the remaining alignments as guidelines, the potential precursor sequences are excised from the genome. This step is time-consuming, especially when the reference genome is large.

excise_candidate.pl TAIR9_genome.fa AtShoot_filter15_ncRNA_CDS.bst 250 >AtShoot_precursors.fa

The secondary structures of the potential precursors are predicted using RNAfold. -noPS means that no graphical output is produced.

cat AtShoot_precursors.fa | RNAfold -noPS > AtShoot_structures

The signatures are generated by aligning the remaining reads to the potential precursors.

bowtie-build -f AtShoot_precursors.fa bowtie-index/ AtShoot_precursors

bowtie -a -v 0 bowtie-index/ AtShoot_precursors -f AtShoot_filtered.fa > AtShoot_precursors.aln

convert_bowtie_to_blast.pl AtShoot_precursors.aln AtShoot_filtered.fa AtShoot_precursors.fa > AtShoot_precursors.bst

#Note that it is necessary for the prediction to do this sorting.

sort +3 -25 AtShoot_precursors.bst >AtShoot_signatures

Predictions are made:

miRDeep.pl AtShoot_signatures AtShoot_structures > AtShoot_predictions

Note that there are several parameters of miRDeep that can be custom adjusted. See the miRDeep reference.

5-5. REMOVING REDUDANT PREDICTED MICRORNAS AND FILTERING PREDICTED ONES BY PLANT CRITERIA

One issue that miRDeep has not dealt with is that in some cases, there are redundant predicted items if the precursors are extracted by mapped reads that are closely located at the same chromosome loci. Meanwhile, recently updated criteria of plant miRNAs are considered critical in identifying new species- or tissue-specific miRNAs (Meyers et al.

2008). The script, *Rm_redundancy_meet_plant.pl*, can remove redundant items and filter the items out that do not meet the criteria of plant miRNAs.

rm_redundancy_meet_plant.pl chromosome_length AtShoot_precursors.fa AtShoot_predictions AtShoot_nr_prediction AtShoot_filter_P_prediction

The file, *chromosome_length*, includes the information on chromosome length. The file, *AtShoot_nr_prediction*, is the output file which contains non-redundant predicted miRNA information. The file, *AtShoot_filter_P_prediction*, is also the output file, which contains predicted miRNAs that meet the criteria of plant miRNAs. For the details of the format of the two output files, please see Section 6 of this manual.

6. THE MIRDP SOFTWARE PACKAGE

The miRDP package consists of nine documented Perl scripts that should be run sequentially by the user. Of the nine scripts, three, *filter_alignments.pl*, *overlap.pl,* and *alignedselected.pl,* are inherited from miRDeep (Friedlander et al., 2008). The other scripts are either novel or have been modified from the original miRDeep version. Functions of the nine scripts are described in the following:

a. *fetch_extended_precursors.pl* fetches the extended precursors from reference sequences based on the location information of annotated miRNAs. The gff file could be downloaded from miRBase (http://www.miRBase.org).

b. *convert_bowtie_to_blast.pl* changes the bowtie format into blast parsed format. Blastparsed format is a custom tabular separated format derived from standard NCBI blast output format.

c. *convert_SAM_to_blast.pl* changes the SAM format into blast parsed format. Note that when use aligners which could generate SAM format file, all perfect valid alignments should be kept. When using bowtie, for instance, the option –a and –v 0 indicate reporting all perfect valid alignment.

d. *filter_alignments.pl* filters the alignments of deep sequencing reads to a genome. It filters partial alignments as well as multi-aligned reads (user-specified frequency cutoff). The basic input is a file in blast parsed format.

e. *overlap.pl* can be used (user-specified) to remove reads that align to the genome in positions that overlap with selected annotation tracks provided by the user (for example known rRNAs, tRNAs, etc). The basic input is a file in blast parsed format and an annotation file in standard gff format. In fact, just some aspects of the GFF format, including seq name, start, end, and strand, are used. (For details on GFF format, please check http://genome.ucsc.edu/FAQ/FAQformat.html#format3)

f. *alignedselected.pl* cleans the ids overlapping with ncRNA or exons.

g. *excise_candidate.pl* cuts out potential precursor sequences from a reference sequence using aligned reads as guidelines. The basic input is a file in blastparsed format and a FASTA file. The basic output is also in FASTA format.

h. *miRDP.pl* needs two input files, signature file and structure file, which is modified from the core miRDeep algorithm by changing the scoring system with plant specific parameters.

i. *compare_annotated_retrieved.pl* exactly picks out the predicted miRNAs which are the annotated ones.

j. *rm_redundant_meet_plant.pl* needs three input files: chromosome_length, precursors and original_prediction generated by *miRDP.pl*. It generates two output files, non-redundant predicted file and predicted file filtered by plant criteria. The tab-delimited files contain columns that indicate chromosome id, strand direction, reads id, precursor id, mature miRNA location, precursor location, mature sequences, and precursor sequences.

7. ISSUES USING MIRDP

7.1 REDUNDANCY AND MIRNA*

In some cases, the output miRNAs from miRDP may differ from the known miRNAs. We found that this is mainly due to one of two reasons: heterogeneity of the mature miRNAs or the relative abundance of miRNA and miRNA*. We found that this does not

impact the optimal length selection of precursors and the profiling of known miRNA genes. However, if users desire to exactly extract the annotated miRNAs, the optional script *compare_annotated_retrieved.pl* could be used to assist for this purpose.

compare_annotated_retrieved.pl ath_miR_info precursors_250.fa AtShoot_250_prediction

ath_miR_info is a file including miRNA precursor and mature miRNA location information from the annotation.

7.2 COMPUTATIONAL TIME

There are three steps that might be time-consuming when use miRDP, especially when the size of small RNA library and the genome are both large.

a. the process of mapping reads to genome sequences (command line likes: *bowtie-build –f TAIR9_genome.fa bowtie-index/TAIR9_genome; bowtie –a –v 0 bowtie-index/TAIR9_genome –f AtShoot.fa >AtShoot.aln*). To save time, you may want to download bowtie index files from the bowtie website (http://bowtie-bio.sourceforge.net/index.shtml) if the genome sequences of the species you are working with have been indexed. Otherwise, you should index reference sequences by yourself. Please keep the index file for a while till you have finish your project since you might need to re-index your genome.

b. the process of obtaining candidate precursors(command line likes: *excise_candidate.pl TAIR9_genome.fa AtShoot_filter15_ncRNA_CDS.bst 250 >AtShoot_precursors.fa*). This step might take hours of computational time on a modest-sized cluster. A good strategy is to divide the xxx.bst file into smaller sub-groups based on chromosomes or contigs. Running the divided sub-groups simultaneously could significantly shorten the computational time.

c. the process of exploring secondary structure of candidate precursors(command line likes: *cat AtShoot_precursors.fa | RNAfold -noPS > AtShoot_structures*). If the number of candidate precursors is huge, the strategy of dividing them into smaller sub-groups and running the sub-groups at the same time can be used.

Appendix 2. MicroRNA roles in systemin-mediated pathway

MiRNA buffer roles in systemin-mediated pathway in tomato

Introduction

Higher plants have evolved defense strategies to protect themselves from wounding by herbivorous insects. Tomato has been widely employed as a model system to study defense responses to herbivore and pathogen attack (Ryan and Pearce, 2003; Bostock, 2005). Briefly, systemin, a 18-amino-acid peptide processed from its 200-amino acid precursor prosystemin (PS) (Ryan and Pearce, 2003), could trigger the entire pathway, and induce the expression of defense genes such as a proteinase inhibitor (*PIN*). A transgenic line of tomato (35S::PS) that over-expresses PS could promote the constitutive expression of defense genes and enhance resistance to herbivore and pathogen (Li et al., 2002; Chen et al., 2005). To date, the systemin-mediated signaling pathway has been established over the last decade and more and more genes encoding proteins involved in this pathway are identified. However, there are not reports which focus on uncovering the role of plant miRNAs in this signaling pathway. In the project, we commenced a piece of research that has been trying to understand how miRNAs are involved in this pathway.

Preliminary results

Toward an understanding of miRNAs role in systemin-mediated pathway, we first sequenced small RNA libraries from 14-day leaf from wild type plants (WT) and plants overexpressing prosystemin (PS), respectively. Intriguingly, compared to libraries in WT, these in PS possess much more reads at 21 nt (Figure 1A), and after careful examination, it is found that around one third of increased 21 nt reads in PS are corresponding to mature miRNAs (Figure 1B), and that around half of annotated miRNAs are expressed differently in WT and PS. To further confirm this, northern blot is performed and greatly supports the result from small RNA libraries (Figure 1C). To validate consequence accounted for miRNAs' change in PS, we used degradome sequencing to detect miRNA targets, and around 700 good candidates are achieved (Figure 2A-D). Furthermore, we randomly selected several target candidates and performed RT-PCR to detect their expression in WT and PS. Interestingly, their expression is correlated with the expression of miRNAs such that higher miRNA expression is corresponding to lower target expression (Figure 2E). Taken together, PS could effect the expression of a large group of miRNAs and their targets.

To examine how these miRNAs are involved in systemin-mediated pathway, we selected several of them and simulated their changes in PS. miR157 is one of the highest expressed miRNAs, and its expression increased one third in PS. We found that there are 5 miR157 loci cross the entire chromosomes in tomato (Figure 3A,C), and they could produce two mature isoforms with one nucleotide difference (Figure 3B). Given the number of reads corresponding to these two different isoforms in small RNA libraries, the isoform1 dominates miR157 expression (Figure 3D). In fact, we detected that only

MIR157c are highly expressed among the five miR157 loci when we performed RT-PCR to test the expression of them (Figure 3E). *MIR157d* is selected to be overexpressed in WT background to mimic the changes in PS. Strikingly, it is found that the lineage overexpressing *MIR157d* (OE157d) have several phenotypic changes that are similar with PS (Figure 4). For instance, OE157d increases the seed size (Figure 4A,B), shortens root length during seed germination (Figure 4C-E), and delays flowering (Figure 4F-H) as PS dose, suggesting that miR157 is involved in systemin-mediated pathway indirectly. Then, RNA-seq is used to compare the transcriptome of the three lineages, WT, PS and OE157d (Figure 5). Around 2,000 genes are differently expressed among these three lineages (Figure 5A), and a big truck of these genes are either up-regulated or down-regulated in both PS and OE157d (Figure 5B). We further confirmed that two defense genes are induced in OE157d when we examined four, indicating that miR157 could participate systemin-mediated pathway in terms of changing the expression of defense genes.

In sum, given the experiments performed above, it is shown that PS could impact many miRNAs' expression, and that many miRNAs in turn are involved in systemin-mediated pathway. The example of miR157 illustrates that the change of one miRNA could partially mimic some phenotypic changes introduced in PS, and miRNAs could effect the expression of a large number of genes including defense genes to take part in systemin-mediated pathway, indicating that miRNAs potentially play a buffer role in systemin-mediated pathway.

Figure 1. Prosystemin (PS) disturbing the expression of miRNAs

(A) Profile of deeply sequenced small RNA libraries in wild type (WT) and 35S::PS (PS). Arrow indicates the increased number of reads in 21 nt. (B) Content of increased reads in PS. 36% reads are corresponding to miRNAs. (C) Northern blots in WT and PS. U6 is loaded as a control.

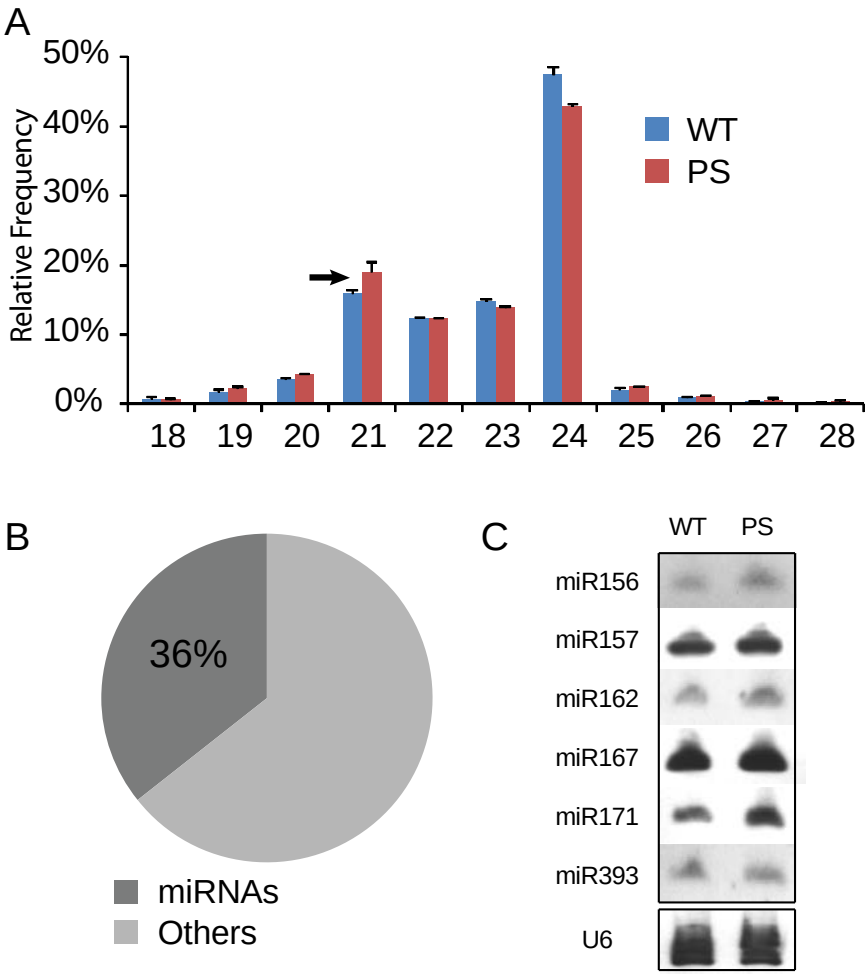

A
50%
40%
30%
20%
10%
0%
Relative Frequency
WT
PS
18
19
20
21
22
23
24
25
26
27
28
B
36%
miRNAs
Others
C
WT
PS
miR156
miR157
miR162
miR167
miR171
miR393
U6

Figure 2. PS impacting the expression of miRNA targets

(A)~(D) Validation of miRNA targets from degradome sequencing. Red dots indicate the position of miRNA cleavage site. (E) The expression of miRNA targets by RT-PCT.

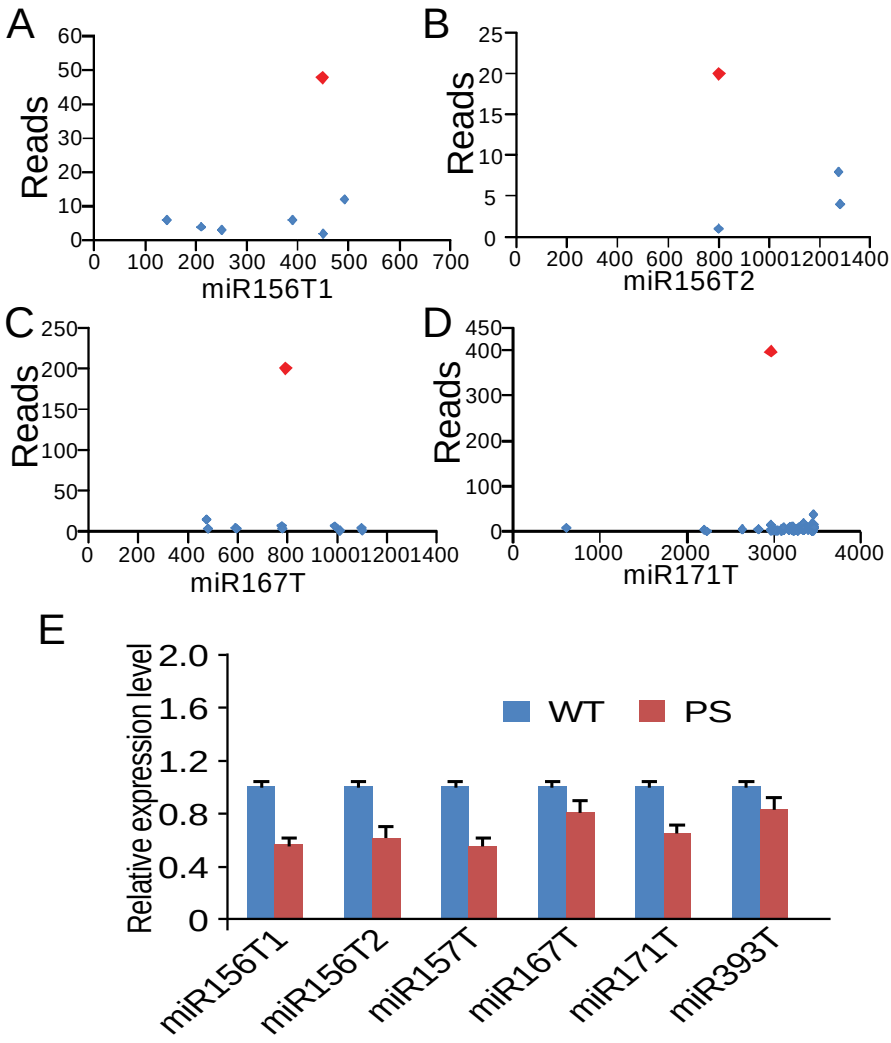

A
Reads
miR156T1
B
Reads
miR156T2
C
Reads
miR167T
D
Reads
miR171T
E
Relative expression level
WT
PS
miR156T1
miR156T2
miR157T
miR167T
miR171T
miR393T

Figure 3. Summary of miR157 loci in tomato

(A) Chromosome location of five miR157 loci. *MIR157a* and *MIR157e* are clustered. (B) Sequences of mature miRNA from five *MIR157*s. Two isoform are generated by these five miR157 genes with one nucleotide difference. (C) Secondary structure of five pre-miR157s. (D) Expression of two isoforms. Isoform1 dominates the expression. (E) Relative expression of five miR157 genes. *MIR157c* dominates the expression.

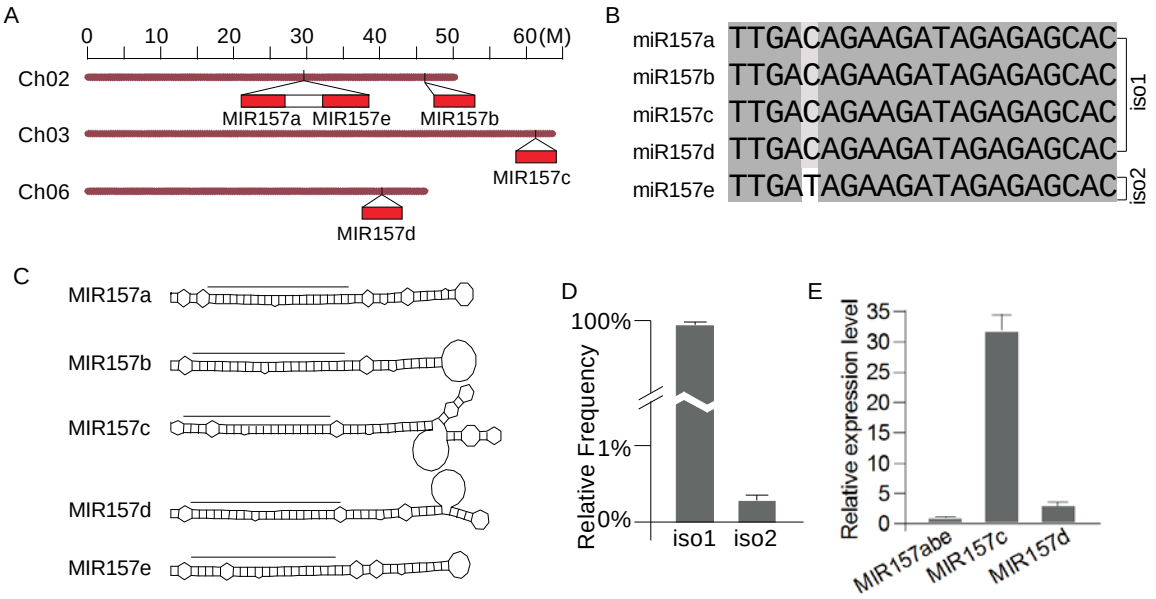

A
0
10
20
30
40
50
60(M)
Ch02
MIR157a
MIR157e
MIR157b
Ch03
MIR157c
Ch06
MIR157d
B
miR157a TTGACAGAAGATAGAGAGCAC
miR157b TTGACAGAAGATAGAGAGCAC
miR157c TTGACAGAAGATAGAGAGCAC
miR157d TTGACAGAAGATAGAGAGCAC
miR157e TTGATAGAAGATAGAGAGCAC
iso1
iso2
C
MIR157a
MIR157b
MIR157c
MIR157d
MIR157e
D
100%
1%
0%
Relative Frequency
iso1
iso2
E
Relative expression level
35
30
25
20
15
10
5
0
MIR157abe
MIR157c
MIR157d

Figure 4. Phenotypes in PS mimicked by over-expression MIR157d (OE157d) in WT.

(A) OE157d increasing the seed size partially as PS. (B) Weight of one hundred seeds in each lineage. Values are the mean and SD of three biological replicates. (C) OE157d shortening root length as PS. (D) Seed germination time in three lineage. Twenty are counted in each lineage. (E) Root length in 105 hours after germination in three lineage. Twenty are counted in each lineage. (F) OE157d delaying flowering as PS. (G) Flower budding time in three lineage. Twenty are counted in each lineage. (H) Flower blooming time in three lineage. Twenty are counted in each. *P<0.05, **P<0.01.

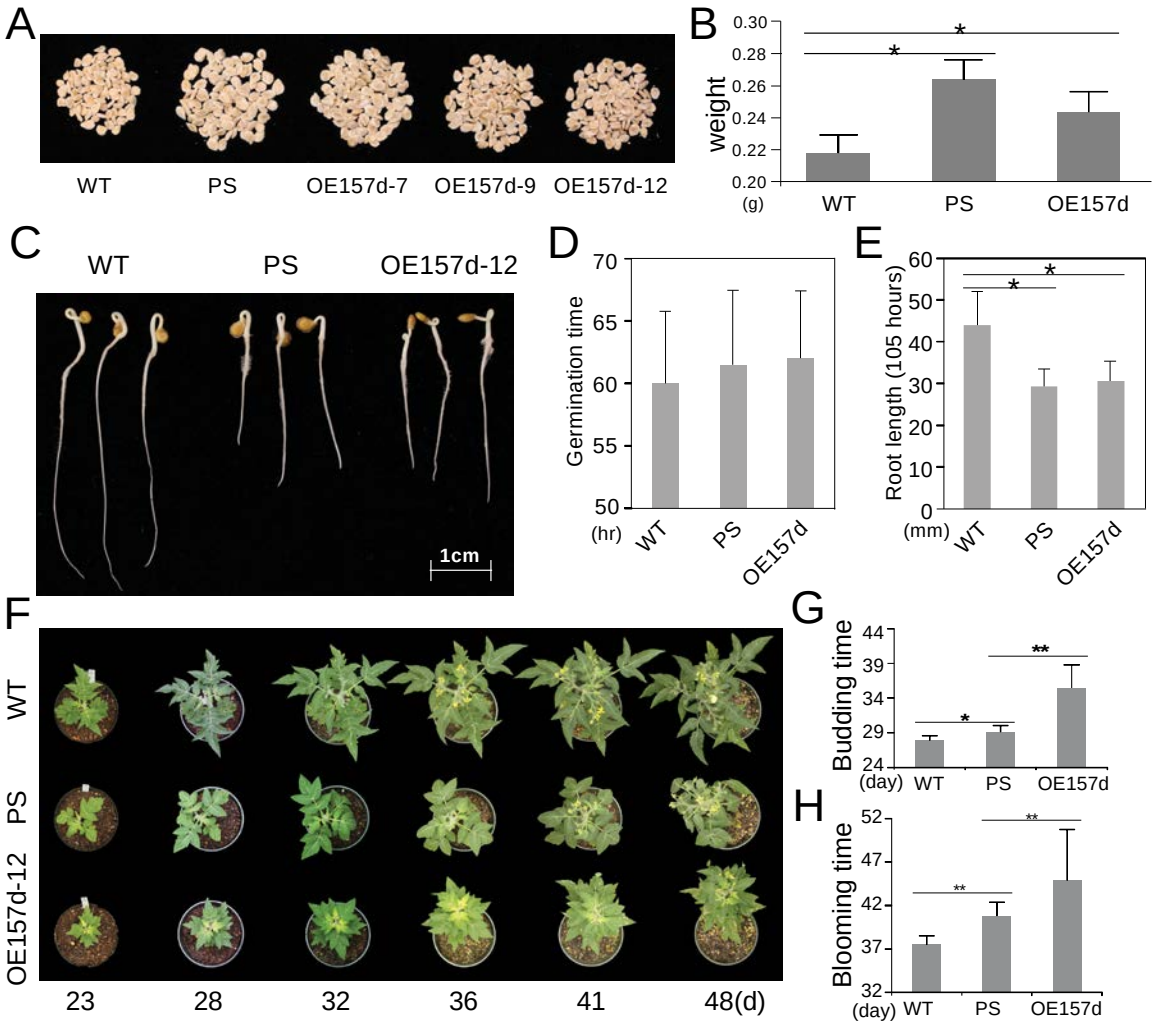
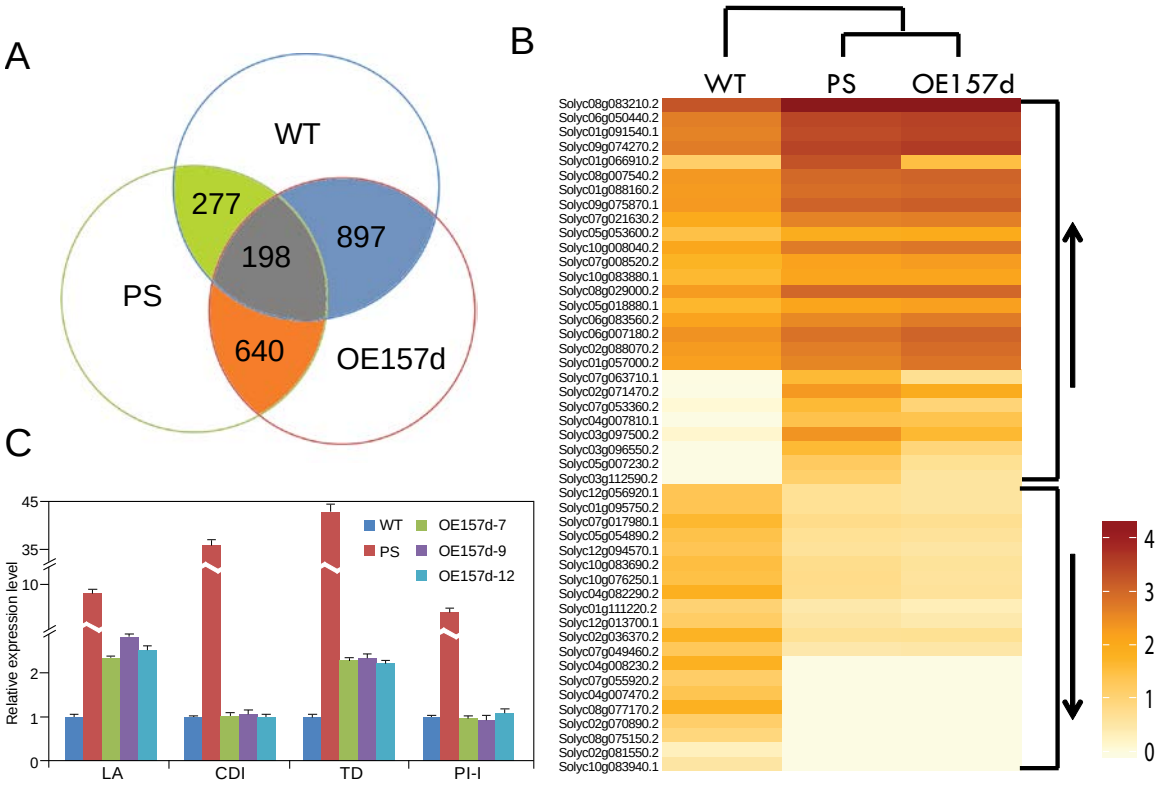

A
WT
PS
OE157d-7
OE157d-9
OE157d-12
B
weight
(g)
0.30
0.28
0.26
0.24
0.22
0.20
WT
PS
OE157d
C
WT
PS
OE157d-12
1cm
D
Germination time
(hr)
70
65
60
55
50
WT
PS
OE157d
E
Root length (105 hours)
(mm)
60
50
40
30
20
10
0
WT
PS
OE157d
F
WT
PS
OE157d-12
23
28
32
36
41
48(d)
G
Budding time
(day)
44
39
34
29
24
WT
PS
OE157d
H
Blooming time
(day)
52
47
42
37
32
WT
PS
OE157d

Figure 5. transcriptome comparison among WT, PS and OE157d

(A) Differently expressed genes among three lineage. (B) A part of heat map representing genes that have similar expression changes in PS and OE157d when compared to WT. (C) OE157d inducing defense genes' expression validated by RT-PCT.


A
WT
PS
OE157d
277
198
897
640
B
WT
PS
OE157d
Solyc08g083210.2
Solyc06g050440.2
Solyc01g091540.1
Solyc09g074270.2
Solyc01g066910.2
Solyc08g007540.2
Solyc01g088160.2
Solyc09g075870.1
Solyc07g021630.2
Solyc05g053600.2
Solyc10g008040.2
Solyc07g008520.2
Solyc10g083880.1
Solyc08g029000.2
Solyc05g018880.1
Solyc06g083560.2
Solyc06g007180.2
Solyc02g088070.2
Solyc01g057000.2
Solyc07g063710.1
Solyc02g071470.2
Solyc07g053360.2
Solyc04g007810.1
Solyc03g097500.2
Solyc03g096550.2
Solyc05g007230.2
Solyc03g112590.2
Solyc12g056920.1
Solyc01g095750.2
Solyc07g017980.1
Solyc05g054890.2
Solyc12g094570.1
Solyc10g083690.2
Solyc10g076250.1
Solyc04g082290.2
Solyc01g111220.2
Solyc12g013700.1
Solyc02g036370.2
Solyc07g049460.2
Solyc04g008230.2
Solyc07g055920.2
Solyc04g007470.2
Solyc08g077170.2
Solyc02g070890.2
Solyc08g075150.2
Solyc02g081550.2
Solyc10g083940.1
4
3
2
1
0
C
Relative expression level
45
35
10
2
1
0
WT
PS
OE157d-7
OE157d-9
OE157d-12
LA
CDI
TD
PI-I

Preliminary conclusions

It is intriguing that the transgenic line 35S::PS reprograms almost half of annotated miRNAs, and that many miRNAs are in turn engaged in this pathway which is demonstrated by over-expressing *MIR157d* in WT. Further, the comparison of phenotypic changes and entire transcriptome variation indicates that a large number of miRNAs as a group possibly play a buffer role and deliver the internal signals to different pathways when plants meet stimulus of wounding.

References

Abdel-Ghany, S.E., and Pilon, M. (2008). MicroRNA-mediated systemic down-regulation of copper protein expression in response to low copper availability in Arabidopsis. J Biol Chem **283,** 15932-15945.

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

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.

Alves, L., Jr., Niemeier, S., Hauenschild, A., Rehmsmeier, M., and Merkle, T. (2009). Comprehensive prediction of novel microRNA targets in Arabidopsis thaliana. Nucleic Acids Res **37,** 4010-4021.

Arazi, T., Talmor-Neiman, M., Stav, R., Riese, M., Huijser, P., and Baulcombe, D.C. (2005). Cloning and characterization of micro-RNAs from moss. Plant J **43,** 837-848.

Arteaga-Vazquez, M., Caballero-Perez, J., and Vielle-Calzada, J.P. (2006). A family of microRNAs present in plants and animals. Plant Cell **18,** 3355-3369.

Axtell, M.J., and Bartel, D.P. (2005). Antiquity of microRNAs and their targets in land plants. Plant Cell **17,** 1658-1673.

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

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res **37,** W202-208.

Balasubramanian, S., Sureshkumar, S., Lempe, J., and Weigel, D. (2006). Potent induction of Arabidopsis thaliana flowering by elevated growth temperature. PLoS Genet **2,** e106.

Barakat, A., Wall, K., Leebens-Mack, J., Wang, Y.J., Carlson, J.E., and Depamphilis, C.W. (2007). Large-scale identification of microRNAs from a basal eudicot (Eschscholzia californica) and conservation in flowering plants. Plant J **51,** 991-1003.

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.

Bayne, E.H., Portoso, M., Kagansky, A., Kos-Braun, I.C., Urano, T., Ekwall, K., Alves, F., Rappsilber, J., and Allshire, R.C. (2008). Splicing factors facilitate RNAi-directed silencing in fission yeast. Science **322,** 602-606.

Bologna, N.G., Mateos, J.L., Bresso, E.G., and Palatnik, J.F. (2009). A loop-to-base processing mechanism underlies the biogenesis of plant microRNAs miR319 and miR159. EMBO J **28,** 3646-3656.

Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics **20,** 2911-2917.

Bostock, R.M. (2005). Signal crosstalk and induced resistance: straddling the line between cost and benefit. Annu Rev Phytopathol **43,** 545-580.

Boutet, S., Vazquez, F., Liu, J., Beclin, C., Fagard, M., Gratias, A., Morel, J.B., Crete, P., Chen, X., and Vaucheret, H. (2003). Arabidopsis HEN1: a genetic link between endogenous miRNA controlling development and siRNA controlling transgene silencing and virus resistance. Curr Biol **13,** 843-848.

Breakfield, N.W., Corcoran, D.L., Petricka, J.J., Shen, J., Sae-Seaw, J., Rubio-Somoza, I., Weigel, D., Ohler, U., and Benfey, P.N. (2012). High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis. Genome Res **22,** 163-176.

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.

Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA **10,** 1957-1966.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics **7,** 327.

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

Cavallini, F. (1993). Fitting a Logistic Curve to Data. Coll. Math. J. **24,** 7.

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., Wilkerson, C.G., Kuchar, J.A., Phinney, B.S., and Howe, G.A. (2005). Jasmonate-inducible plant enzymes degrade essential amino acids in the herbivore midgut. Proc Natl Acad Sci U S A **102,** 19237-19242.

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, K., and Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet **38,** 1452-1456.

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet **8,** 93-103.

Chen, X. (2005). MicroRNA biogenesis and function in plants. FEBS Lett **579,** 5923-5931.

Chen, X., Liu, J., Cheng, Y., and Jia, D. (2002). HEN1 functions pleiotropically in Arabidopsis development and acts in C function in the flower. Development **129,** 1085-1094.

Colaiacovo, M., Bernardo, L., Centomani, I., Crosatti, C., Giusti, L., Orru, L., Tacconi, G., Lamontanara, A., Cattivelli, L., and Faccioli, P. (2012). A Survey of MicroRNA Length Variants Contributing to miRNome Complexity in Peach (Prunus Persica L.). Front Plant Sci **3,** 165.

Contreras-Cubas, C., Rabanal, F.A., Arenas-Huertero, C., Ortiz, M.A., Covarrubias, A.A., and Reyes, J.L. (2012). The Phaseolus vulgaris miR159a

precursor encodes a second differentially expressed microRNA. Plant Mol Biol **80,** 103-115.

Creighton, C.J., Reid, J.G., and Gunaratne, P.H. (2009). Expression profiling of microRNAs by deep sequencing. Brief Bioinform **10,** 490-497.

Cuperus, J.T., Fahlgren, N., and Carrington, J.C. (2011). Evolution and functional diversification of MIRNA genes. Plant Cell **23,** 431-442.

Denman, R.B. (1993). Using RNAFOLD to predict the activity of small catalytic RNAs. Biotechniques **15,** 1090-1095.

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., Dangl, J.L., and Carrington, J.C. (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PLoS One **2,** e219.

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 **22,** 1074-1089.

Fei, Z., Joung, J.G., Tang, X., Zheng, Y., Huang, M., Lee, J.M., McQuinn, R., Tieman, D.M., Alba, R., Klee, H.J., and Giovannoni, J.J. (2011). Tomato Functional Genomics Database: a comprehensive resource and analysis package for tomato functional genomics. Nucleic Acids Res **39,** D1156-1163.

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

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K., and Mockler, T.C. (2010). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res **20,** 45-58.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol **26,** 407-415.

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

Gentile, A., Ferreira, T.H., Mattos, R.S., Dias, L.I., Hoshino, A.A., Carneiro, M.S., Souza, G.M., Calsa, T., Jr., Nogueira, R.M., Endres, L., and Menossi, M. (2013). Effects of drought on the microtranscriptome of field-grown sugarcane plants. Planta **237,** 783-798.

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. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol **26,** 941-946.

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. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40,** D1178-1186.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res **34,** D140-144.

Guo, A.Y., Zhu, Q.H., Gu, X., Ge, S., Yang, J., and Luo, J. (2008a). Genome-wide identification and evolutionary analysis of the plant specific SBP-box transcription factor family. Gene **418,** 1-8.

Guo, X., Gui, Y., Wang, Y., Zhu, Q.H., Helliwell, C., and Fan, L. (2008b). Selection and mutation on microRNA target sequences during rice evolution. BMC Genomics **9,** 454.

Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C., and Kasschau, K.D. (2005). ASRP: the Arabidopsis Small RNA Project Database. Nucleic Acids Res **33,** D637-640.

Hackenberg, M., Rodriguez-Ezpeleta, N., and Aransay, A.M. (2011). miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res **39,** W132-138.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J.M., and Aransay, A.M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res **37,** W68-76.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell **125,** 887-901.

Hazen, S.P., Naef, F., Quisel, T., Gendron, J.M., Chen, H., Ecker, J.R., Borevitz, J.O., and Kay, S.A. (2009). Exploring the transcriptional landscape of plant circadian rhythms using genome tiling arrays. Genome Biol **10,** R17.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res **31,** 3429-3431.

Huang, P.J., Liu, Y.C., Lee, C.C., Lin, W.C., Gan, R.R., Lyu, P.C., and Tang, P. (2010). DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Res **38,** W385-391.

Huttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? Trends Genet **21,** 289-297.

Jones-Rhoades, M.W., Bartel, D.P., and Bartel, B. (2006). MicroRNAS and their regulatory roles in plants. Annu Rev Plant Biol **57,** 19-53.

Kawashima, C.G., Yoshimoto, N., Maruyama-Nakashita, A., Tsuchiya, Y.N., Saito, K., Takahashi, H., and Dalmay, T. (2009). Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different cell types. Plant J **57,** 313-321.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell **115,** 209-216.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res **39,** D152-157.

Kruger, J., and Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res **34,** W451-454.

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.

Lai, E.C., Tomancak, P., Williams, R.W., and Rubin, G.M. (2003). Computational identification of Drosophila microRNA genes. Genome Biol **4,** R42.

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.

Laubinger, S., Sachsenberg, T., Zeller, G., Busch, W., Lohmann, J.U., Ratsch, 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.

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

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.

Li, A., and Mao, L. (2007). Evolution of plant microRNA gene families. Cell Res **17,** 212-218.

Li, C., Williams, M.M., Loh, Y.T., Lee, G.I., and Howe, G.A. (2002). Resistance of cultivated tomato to cell content-feeding herbivores is regulated by the octadecanoid-signaling pathway. Plant Physiol **130,** 494-503.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003a). Vertebrate microRNA genes. Science **299,** 1540.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003b). The microRNAs of Caenorhabditis elegans. Genes Dev **17,** 991-1008.

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.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25,** 955-964.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R., and Golub, T.R. (2005). MicroRNA expression profiles classify human cancers. Nature **435,** 834-838.

Lu, S., Yang, C., and Chiang, V.L. (2011). Conservation and diversity of microRNA-associated copper-regulatory networks in Populus trichocarpa. J Integr Plant Biol **53,** 879-891.

Lu, S., Sun, Y.H., Amerson, H., and Chiang, V.L. (2007). MicroRNAs in loblolly pine (Pinus taeda L.) and their association with fusiform rust gall development. Plant J **51,** 1077-1098.

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.

Manavella, P.A., Hagmann, J., Ott, F., Laubinger, S., Franz, M., Macek, B., and Weigel, D. (2012). Fast-forward genetics identifies plant CPL phosphatases as regulators of miRNA processing factor HYL1. Cell **151,** 859-870.

Manning, K., Tor, M., Poole, M., Hong, Y., Thompson, A.J., King, G.J., Giovannoni, J.J., and Seymour, G.B. (2006). A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet **38,** 948-952.

Mateos, J.L., Bologna, N.G., Chorostecki, U., and Palatnik, J.F. (2010). Identification of microRNA processing determinants by random mutagenesis of Arabidopsis MIR172a precursor. Curr Biol **20,** 49-54.

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.

Meng, Y., Gou, L., Chen, D., Mao, C., Jin, Y., Wu, P., and Chen, M. (2011). PmiRKB: a plant microRNA knowledge base. Nucleic Acids Res **39,** D181-187.

Merchan, F., Boualem, A., Crespi, M., and Frugier, F. (2009). Plant polycistronic precursors containing non-homologous microRNAs target transcripts encoding functionally related proteins. Genome Biol **10,** R136.

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., and Zhu, J.K. (2008). Criteria for annotation of plant MicroRNAs. Plant Cell **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.

Molnar, 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., Yoo, S.J., Fahlgren, N., Gilbert, S.D., Howell, M.D., Sullivan, C.M., Alexander, A., Nguyen, G., Allen, E., Ahn, J.H., and Carrington, J.C. (2008). AGO1-miR173 complex initiates phased siRNA formation in plants. Proc Natl Acad Sci U S A **105,** 20055-20062.

Morel, J.B., Godon, C., Mourrain, P., Beclin, C., Boutet, S., Feuerbach, F., Proux, F., and Vaucheret, H. (2002). Fertile hypomorphic ARGONAUTE (ago1) mutants impaired in post-transcriptional gene silencing and virus resistance. Plant Cell **14,** 629-639.

Motomura, K., Le, Q.T., Kumakura, N., Fukaya, T., Takeda, A., and Watanabe, Y. (2012). The role of decapping proteins in the miRNA accumulation in Arabidopsis thaliana. RNA Biol **9**.

Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D.J., and Moulton, V. (2008a). A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics **24,** 2252-2253.

Moxon, S., Jing, R., Szittya, G., Schwach, F., Rusholme Pilcher, R.L., Moulton, V., and Dalmay, T. (2008b). Deep sequencing of tomato short RNAs identifies

microRNAs targeting genes involved in fruit ripening. Genome Res **18,** 1602-1609.

Ng, D.W., Zhang, C., Miller, M., Palmer, G., Whiteley, M., Tholl, D., and Chen, Z.J. (2011). cis- and trans-Regulation of miR163 and target genes confers natural variation of secondary metabolites in two Arabidopsis species and their allopolyploids. Plant Cell **23,** 1729-1740.

Nobuta, K., Venu, R.C., Lu, C., Belo, A., Vemaraju, K., Kulkarni, K., Wang, W., Pillay, M., Green, P.J., Wang, G.L., and Meyers, B.C. (2007). An expression atlas of rice mRNAs and small RNAs. Nat Biotechnol **25,** 473-477.

Ohrt, T., Mutze, J., Staroske, W., Weinmann, L., Hock, J., Crell, K., Meister, G., and Schwille, P. (2008). Fluorescence correlation spectroscopy and fluorescence cross-correlation spectroscopy reveal the cytoplasmic origination of loaded nuclear RISC in vivo in human cells. Nucleic Acids Res **36,** 6439-6449.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., and Buell, C.R. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res **35,** D883-887.

Pantano, L., Estivill, X., and Marti, E. (2010). SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. Nucleic Acids Res **38,** e34.

Papp, I., Mette, M.F., Aufsatz, W., Daxinger, L., Schauer, S.E., Ray, A., van der Winden, J., Matzke, M., and Matzke, A.J. (2003). Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. Plant Physiol **132,** 1382-1390.

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.

Prufer, K., Stenzel, U., Dannemann, M., Green, R.E., Lachmann, M., and Kelso, J. (2008). PatMaN: rapid alignment of short sequences to large databases. Bioinformatics **24,** 1530-1531.

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,** 3407-3425.

Reddy, A.S. (2007). Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev Plant Biol **58,** 267-294.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. RNA **10,** 1507-1517.

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

Ren, G., Xie, M., Dou, Y., Zhang, S., Zhang, C., and Yu, B. (2012). Regulation of miRNA abundance by RNA binding protein TOUGH in Arabidopsis. Proc Natl Acad Sci U S A **109,** 12817-12821.

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.

Ronen, R., Gan, I., Modai, S., Sukacheov, A., Dror, G., Halperin, E., and Shomron, N. (2010). miRNAkey: a software for microRNA deep sequencing analysis. Bioinformatics **26,** 2615-2616.

Ryan, C.A., and Pearce, G. (2003). Systemins: a functionally defined family of peptide signals that regulate defensive genes in Solanaceae species. Proc Natl Acad Sci U S A **100 Suppl 2,** 14577-14580.

Schuster, P., Fontana, W., Stadler, P.F., and Hofacker, I.L. (1994). From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci **255,** 279-284.

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.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell **115,** 199-208.

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.

Szczesniak, M.W., Deorowicz, S., Gapski, J., Kaczynski, L., and Makalowska, I. (2012). miRNEST database: an integrative approach in microRNA search and annotation. Nucleic Acids Res **40,** D198-204.

Tam, W. (2001). Identification and characterization of human BIC, a gene on chromosome 21 that encodes a noncoding RNA. Gene **274,** 157-167.

Taub, M.A., Corrada Bravo, H., and Irizarry, R.A. (2010). Overcoming bias and systematic errors in next generation sequencing data. Genome Med **2,** 87.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25,** 1105-1111.

Tsutsumi, A., Kawamata, T., Izumi, N., Seitz, H., and Tomari, Y. (2011). Recognition of the pre-miRNA structure by Drosophila Dicer-1. Nat Struct Mol Biol **18,** 1153-1158.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. Cell **136,** 669-687.

Wang, B.B., and Brendel, V. (2006). Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci U S A **103,** 7175-7180.

Wang, B.B., O'Toole, M., Brendel, V., and Young, N.D. (2008). Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. BMC Plant Biol **8,** 17.

Wang, J.W., Czech, B., and Weigel, D. (2009). miR156-regulated SPL transcription factors define an endogenous flowering pathway in Arabidopsis thaliana. Cell **138,** 738-749.

Wang, K., Li, M., Gao, F., Li, S., Zhu, Y., and Yang, P. (2012). Identification of conserved and novel microRNAs from Liriodendron chinense floral tissues. PLoS One **7,** e44696.

Wang, Y., Itaya, A., Zhong, X., Wu, Y., Zhang, J., van der Knaap, E., Olmstead, R., Qi, Y., and Ding, B. (2011). Function and Evolution of a MicroRNA that

regulates a Ca2+-ATPase and triggers the formation of phased small interfering RNAs in tomato reproductive growth. Plant Cell **23,** 3185-3203.

Wark, A.W., Lee, H.J., and Corn, R.M. (2008). Multiplexed detection methods for profiling microRNA expression in biological samples. Angew Chem Int Ed Engl **47,** 644-652.

Werner, S., Wollmann, H., Schneeberger, K., and Weigel, D. (2010). Structure determinants for accurate processing of miR172a in Arabidopsis thaliana. Curr Biol **20,** 42-48.

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

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, 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, 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.

Xie, F., Frazier, T.P., and Zhang, B. (2010). Identification and characterization of microRNAs and their targets in the bioenergy plant switchgrass (Panicum virgatum). Planta **232,** 417-434.

Xie, Z., Kasschau, K.D., and Carrington, J.C. (2003). Negative feedback regulation of Dicer-Like1 in Arabidopsis by microRNA-guided mRNA degradation. Curr Biol **13,** 784-789.

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

Yamasaki, H., Hayashi, M., Fukazawa, M., Kobayashi, Y., and Shikanai, T. (2009). SQUAMOSA Promoter Binding Protein-Like7 Is a Central Regulator for Copper Homeostasis in Arabidopsis. Plant Cell **21,** 347-361.

Yamasaki, H., Abdel-Ghany, S.E., Cohu, C.M., Kobayashi, Y., Shikanai, T., and Pilon, M. (2007). Regulation of copper homeostasis by micro-RNA in Arabidopsis. J Biol Chem **282,** 16369-16378.

Yang, X., and Li, L. (2011). miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics **27,** 2614-2615.

Yang, X., Zhang, H., and Li, L. (2011). Global analysis of gene-level microRNA expression in Arabidopsis using deep sequencing data. Genomics **98,** 40-46.

Yang, X., Zhang, H., and Li, L. (2012). Alternative mRNA processing increases the complexity of microRNA-based gene regulation in Arabidopsis. Plant J **70,** 421-431.

Yang, Z., Ebright, Y.W., Yu, B., and Chen, X. (2006). HEN1 recognizes 21-24 nt small RNA duplexes and deposits a methyl group onto the 2' OH of the 3' terminal nucleotide. Nucleic Acids Res **34,** 667-675.

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.

Zhang, H., and Li, L. (2013). SQUAMOSA promoter binding protein-like7 regulated microRNA408 is required for vegetative development in Arabidopsis. Plant J **74,** 98-109.

Zhang, L., Zheng, Y., Jagadeeswaran, G., Li, Y., Gowdu, K., and Sunkar, R. (2011). Identification and temporal expression analysis of conserved and novel microRNAs in Sorghum. Genomics **98,** 460-468.

Zhang, R., Marshall, D., Bryan, G.J., and Hornyik, C. (2013). Identification and characterization of miRNA transcriptome in potato by high-throughput sequencing. PLoS One **8,** e57233.

Zhao, X., Zhang, H., and Li, L. (2013). Identification and analysis of the proximal promoters of microRNA genes in Arabidopsis. Genomics **101,** 187-194.

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, E., Zhao, F., Xu, G., Hou, H., Zhou, L., Li, X., Sun, Z., and Wu, J. (2010). mirTools: microRNA profiling and discovery based on high-throughput sequencing. Nucleic Acids Res **38,** W392-397.

Zhu, H., Xia, R., Zhao, B., An, Y.Q., Dardick, C.D., Callahan, A.M., and Liu, Z. (2012). Unique expression, processing regulation, and regulatory network of peach (Prunus persica) miRNAs. BMC Plant Biol **12,** 149.

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.